Conditioned Simulation of Ground Motion Time Series using Gaussian Process Regression

Aidin Tamhidi[1,*], Nicolas Kuehn[1,†], S. Farid Ghahari[1,‡], Ertugrul Taciroglu[1,§], Yousef Bozorgnia[1,**]

[1]*Civil and Environmental Engineering Department, University of California, Los Angeles, CA 90095, USA*

ABSTRACT

Ground motion time series are critical elements of earthquake engineering for performance analysis of seismic regions' built environment. At present, the number of available instruments to record the free-field ground motions in the US is generally sparse. Therefore, ground motion estimation methods are used to obtain input motion estimates at locations where there is no available instrumentation. In this study, the ground motion time series are constructed using a Gaussian Process regression, which models the Fourier spectrum's real and imaginary parts as random Gaussian variables. The proposed model's training and validation are carried out using the physics-based simulated ground motions of the 1906 San Francisco Earthquake. The evaluation of the model's performance is also carried out using the simulated magnitude 7.0 Hayward fault earthquake and the ground motions recorded in the 2019 magnitude 7.1 Ridgecrest Earthquake sequence within the Los Angeles area. All evaluations imply that the trained Gaussian Process regression model can estimate the ground motion time series properly. It is also observed that the trained Gaussian Process regression model has decent performance on the long-period ground motion estimation due to the ground motion directivity pulses. The results also illustrate that the stations' prediction either at the boundary edges or outside of the network might not be as accurate as other stations' estimations.

KEYWORDS: Conditioned Simulation of Ground-Motions; Gaussian Process Regression, Spatial Variation of Ground-Motions, Regional Seismic Assessment

1. INTRODUCTION

In the last few decades, thanks to the increase in recording stations, available recorded earthquake ground motions have increased. However, the current sensor networks of ground motion are still too sparse. As an illustration, there are about 2000 stations to record the free-field ground motions around the California state (Southern California Earthquake Data Center). Therefore, an estimation of either ground motion intensity measure (GMIM) or the entire ground motion time series is required to evaluate the damage state and performance level of a site-specific structure where there is no available recording instrumentation. To do so, ShakeCast, ShakeMap, and USGS "Did You Feel It?" platform could be employed to provide an assessment of the shaking level and GMIM at a specific location after an event (Wald et al. 2008; Fraser, Wald, and Lin 2008; Worden et al. 2018; Lin et al. 2018; Wald et al. 2012). Also, several existing studies aimed to estimate the GMIMs at the unobserved locations using the existing observed GMIMs (Baker and Chen 2020; Otake et al. 2020; Sun et al. 2018; Worden et al. 2018). However, for the nonlinear analysis of the structural response, the entire ground-motion time

* Ph.D. Candidate. Corresponding author: Aidin Tamhidi, E-mail: aidintamhidi@ucla.edu
† Research Scientist
‡ Research Scientist
§ Professor
** Professor

series at a specific location is needed to quantify the damage state of a desired site-specific structure or facility. Therefore, it is crucial to develop a model that can generate the entire ground motion time series using the sparsely recorded ground motions after an event. The generated motions at the unobserved stations must be consistent with the spatial variation of ground motion, which refers to the changes in the amplitude, phase, and frequency content of the recorded motions over an area (Zerva and Zervas 2002; Zerva 2009). This phenomenon can have a considerable effect, especially on distributed lifeline structures (Adanur et al. 2016; Jayaram and Baker 2009; Zerva, Falamarz-Sheikhabadi, and Poul 2018; Todorovska, Ding, and Trifunac 2017; Tian et al. 2016). A combination of available recorded ground motion dataset with the pre-developed structural models for a region can provide an immediate estimation of the slight, moderate, and extensive damages for the buildings, infrastructures, and lifeline structures.

There is an extensive list of publications that address the topic of conditioned ground motion simulation (Rodda and Basu 2019; 2018; Huang and Wang 2017; Wu et al. 2016; Alimoradi et al. 2015; Konakli and Der Kiureghian 2012; Kameda and Morikawa 1992). The majority of the traditional conditioned simulation of ground motion is based on deploying Cross Spectral Densities (CSD) and Auto Spectral Densities (ASD) to determine the covariance between the Fourier series coefficients among neighboring stations' motions (Der Kiureghian 1996; Konakli and Der Kiureghian 2012; Rodda and Basu 2018). The conditioned ground motion simulation results depend on the spatial variability of the motions captured by CSD and ASD. The ASD and CSD are commonly determined using empirical coherency functions. These functions' coefficients are required to be set empirically using data-driven methods. Moreover, a detailed description of the site properties and wave passage might be needed for generating the motions, which could be computationally expensive and time-consuming, especially when an ensemble of ultra-dense stations is desired.

In this study, a Gaussian Process (GP) regression model (also known as Kriging) is employed to generate the ground motion at a target station where there is no available recording instrumentation. This GP regression model spatially interpolates the real and imaginary parts of the observed neighboring motions' frequency content to establish the ground motion time series at the target location. A *Matern* covariance function is used to develop such a GP regressor. This function allocates the correlation between motions' frequency content based on their geographical separation distance and their site condition difference. In other words, the covariance function rather than the empirical coherency functions capture the spatial variability of the ground motions.

In Section 2 of this paper, the GP regression and the way it works are described. Section 3 demonstrates which GP regression models are implemented and how they estimate the entire ground motion time series at the target unobserved stations. In Section 3, it is also explained how the hyper-parameters of the covariance function are tuned. Section 4 elaborates the hyper-parameters' tuning procedure using the physics-based simulation of the 1906 San Francisco earthquake ground motions, which are developed by Aagaard et al. in 2008 (Aagaard et al. 2008). Section 5 illustrates the evaluation of the trained GP regressor performance for predicting the unobserved stations' motions. Finally, the major findings of this study are summarized in Section 6.

2. THEORETICAL BACKGROUND

Suppose the ground motion acceleration time series at location s, $a_s(t)$, is constructed of N discrete data points, $a_s(t_i)$, at equal time intervals Δt. The $a_s(t_i)$ can then be expressed using its Discrete Fourier Transform (DFT) coefficients A_k as (Oppenheim, Willsky, and Nawab 1997)

$$a_s(t_i) = \sum_{k=0}^{N-1} A_k e^{j\omega_k t_i}$$
$$A_k = \frac{1}{N}\sum_{i=0}^{N-1} a_s(t_i)[\cos(\omega_k t_i) + j \cdot \sin(\omega_k t_i)] = \mathcal{R}e_k + j \cdot \mathcal{I}m_k \quad (1)$$

in which $\mathcal{R}e_k$ and $\mathcal{I}m_k$ are the real and imaginary parts of the DFT coefficients of acceleration time series $a_s(t)$ at the k^{th} frequency. These real and imaginary parts are modeled as Gaussian random variables which are estimated using GP regression for the unobserved location.

2.1. Gaussian Process Regression

The GP regression is a powerful supervised learning method which has had many applications among various research areas including ground-motion estimation, post-earthquake damage assessment, and seismic fragility assessment to name but a few (Tamhidi et al. 2019; 2020; Sajedi and Liang 2019; Sheibani, Ou, and Zhe 2020; Landwehr et al. 2016; Sun et al. 2018; Gentile and Galasso 2020). A Gaussian Process is a collection of indexed random variables such that every finite subset is distributed according to a multivariate normal distribution. Loosely speaking, one can understand a Gaussian Process as a multivariate normal distribution for infinitely many random variables. A Gaussian process can be understood as a distribution over functions $f(\boldsymbol{x}) \in \mathrm{R}$

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')) \quad (2)$$

which reads as "The function value $f(\boldsymbol{x})$ at input location $\boldsymbol{x}$ is drawn from a GP with mean function $m(\boldsymbol{x})$ and covariance function $k(\boldsymbol{x},\boldsymbol{x}')$". The GP is entirely defined by its mean and covariance functions.

$$m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})]$$
$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}') - m(\boldsymbol{x}'))] \quad (3)$$

The covariance function $k(\boldsymbol{x}, \boldsymbol{x}')$ encodes a notion of similarity of function values between data points. In Bayesian non-parametrics, a GP is often used to specify a prior distribution over possible functions. It is assumed that the real and imaginary parts of the DFT coefficients at different frequencies are a function of the location, and place a GP prior on these functions.

Let's denote the ensemble of observed stations' input matrix by $\boldsymbol{X}$, each row of which includes an input vector of one observed station. It is desired to fit a GP regression over the observed stations, $\boldsymbol{X}$, and their corresponding observed Gaussian random variable, $\boldsymbol{f}$ (either real or imaginary part.)

It is needed to define a prior mean and covariance functions for the GP prior distribution to start the GP fitting procedure. It is common to assume a zero mean function for the prior distribution. This prior distribution is then converted to a posterior distribution with updated mean and covariance functions based on the observations. Therefore, the posterior mean is not restricted to be zero (Rasmussen and Williams 2006). Consider the unobserved stations' input matrix by $\boldsymbol{X}_*$. The joint prior Gaussian distribution of the GP at $\boldsymbol{X}$ and $\boldsymbol{X}_*$ can then be specified by

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \sigma_y^2 \boldsymbol{I} & \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*) \\ \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*) \end{bmatrix}\right) \quad (4)$$

where σ_y is the observations' noise and $\boldsymbol{I}$ is an identity matrix with the same size of $\boldsymbol{X}$ and $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_*$ are the prior mean function vectors at the locations $\boldsymbol{X}$ and $\boldsymbol{X}_*$, respectively. In this study, the existing (observed) ground motions are considered as noise-free observations ($\sigma_y = 0$). The GP fitting process and 1906 San

Francisco physics-based simulated ground motions' features are discussed in detail in Sections 3 and 4, respectively.

The GP regression's output and its smoothness depend on the employed covariance function. One of the commonly used covariance functions is the *Matern* function given by Equation (5). In Equation (5), Γ is the Gamma function, $\Gamma(\mathrm{n}) = (n-1)!$, K_ν is a modified Bessel function (Abramowitz and Irene 1972), ν is a positive parameter which controls the smoothness of the output function, σ_f is the signal variance that governs how uncertain the GP regression's output is for a given input, and $\boldsymbol{r}$ is the distance between the input vectors, $\boldsymbol{x}$ and $\boldsymbol{x'}$ given by Equation (6). In Equation (6), θ_i is a positive normalizing factor (also known as length-scale) related to the i^{th} dimension of the input vector and d is the dimension of the input vector. The θ_i specifies the rate of decay for the correlation function along the dimension i. In other words, higher values for θ_i results in a higher decay of correlation by increasing the distance along the i^{th} dimension.

$$k_{Matern}(r) = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} r\right)^{\nu} K_\nu(\sqrt{2\nu} r) \tag{5}$$

$$r = \sqrt{\sum_{i=1}^{d} \theta_i^2 (x_i - x'_i)^2} \tag{6}$$

In this study, the *Matern* function with $\nu = 1.5$ is used. The $\nu = 1.5$ is one of the most interesting values for machine learning (Rasmussen and Williams 2006). The corresponding covariance function is given in Equation (7).

$$k_{\nu=1.5}(r) = \sigma_f^2 (1 + \sqrt{3} r) exp(-\sqrt{3} r) \tag{7}$$

It is possible to show that the posterior distribution of the GP at the unobserved locations, $\boldsymbol{f}_*$, is given by Equation (8) (Rasmussen and Williams 2006).

$$\boldsymbol{f}_* | \boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{f} \sim \mathcal{N}(K(\boldsymbol{X}_*, \boldsymbol{X}) K(\boldsymbol{X}, \boldsymbol{X})^{-1} \boldsymbol{f}, K(\boldsymbol{X}_*, \boldsymbol{X}_*) - K(\boldsymbol{X}_*, \boldsymbol{X}) K(\boldsymbol{X}, \boldsymbol{X})^{-1} K(\boldsymbol{X}, \boldsymbol{X}_*)) \tag{8}$$

This section presented a brief overview of the GP regression and its formulations. For more detailed description, one can refer to the related literature books and researches (Li and Sudjianto 2005; Rasmussen and Williams 2006).

3. PROPOSED METHOD

The hyper-parameters of the model (θ_i) need to be tuned in order to implement the GP regression. A standard method to find the optimum hyper-parameters is to maximize the log marginal likelihood of the n observations using Equation (9) (Rasmussen and Williams 2006). In Equation (9), $(\boldsymbol{f} - \boldsymbol{m}(\boldsymbol{x}))^{\mathrm{T}}$ stands for the transpose of the vector $(\boldsymbol{f} - \mathbf{m}(\boldsymbol{x}))$ and $|\boldsymbol{K_{XX}}|$ is the determinant of the matrix $\boldsymbol{K_{XX}}$.

$$\log p(\boldsymbol{f} | \boldsymbol{X}, \boldsymbol{\theta}) = -\frac{1}{2} (\boldsymbol{f} - \boldsymbol{m}(\boldsymbol{X}))^{\mathrm{T}} \boldsymbol{K_{XX}}^{-1} (\boldsymbol{f} - \mathbf{m}(\boldsymbol{X})) - \frac{1}{2} \log |\boldsymbol{K_{XX}}| - \frac{n}{2} \log 2\pi \tag{9}$$

One of the issues with the maximum likelihood estimates (MLEs) is that the optimized hyper-parameters of the covariance function have a considerable variance near their optimum solution. The reason for the latter is that the likelihood function is almost flat close to its extermum, especially when there is a sparse number of observations (Li and Sudjianto 2005). To tackle this issue, one can use the penalized log-likelihood estimates rather than the MLEs. Equation (10) shows the penalized log-likelihood, $Q(\theta)$, formulation.

$$Q(\theta) = -\frac{1}{2}\left(\boldsymbol{f} - \boldsymbol{m}(\boldsymbol{X})\right)^{\mathrm{T}} \boldsymbol{K_{XX}}^{-1}\left(\boldsymbol{f} - \mathbf{m}(\boldsymbol{X})\right) - \frac{1}{2}\log|\boldsymbol{K_{XX}}| - n\sum_{i=1}^{d} p_\lambda(\theta_i) \tag{10}$$

In Equation (10), $p_\lambda(\theta_i)$ is a non-negative penalty function which is a function of i^{th} dimension's length-scale. The λ is the regularization factor which needs to be tuned using the data-driven methods. There are several choices of the penalty functions to be employed in Equation (10) such as L_1 ($p_\lambda(\theta) = \lambda|\theta|$), the least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), and smoothly clipped absolute deviation (SCAD) (Fan and Li 2001). In this study, the SCAD penalty function is implemented. This selection is made based on comparing the various models' performance in prediction. The SCAD penalty function is given by

$$p_\lambda(\theta) = \begin{cases} \lambda\theta & \theta \le \lambda \\ -\dfrac{\lambda^2 + \theta^2 - 2a\lambda\theta}{2(a-1)} & \lambda < \theta \le a\lambda \\ \dfrac{\lambda^2(a+1)}{2} & a\lambda < \theta \end{cases} \tag{11}$$

where a is a constant which is assumed to be 3.7 based on Fan and Li. It is shown that the performance of the model cannot be considerably improved if the parameter a is selected employing data-driven methods.

The hyper-parameters of the covariance function (θ_i) are obtained by maximizing the penalized log-likelihood given in Equation (10). Then, the posterior mean function and signal variance, σ_f, are estimated by having the optimum hyper-parameters (Li and Sudjianto 2005). Eventually, one can predict the real and imaginary parts of each frequency at the unobserved station using the posterior distribution given in Equation (8). Then, the motion time series can be retrieved using Equation (1).

The optimum λ can be estimated using a cross-validation procedure by splitting the observed ground motion dataset into randomly chosen training and validation subsets. Then, the λ, which results in the lowest average error between the predicted motion's response spectrum and the observed one at the same locations, is chosen as the optimum regularization factor. This procedure is discussed in detail in Section 4.

The homogeneity assumption is valid if the recording stations are all located on a fairly uniform site condition (Zerva and Zervas 2002). All the GP's stochastic descriptors depend on just the geographical separation distance between the stations if there is homogeneity. In this study, two types of input vectors for the stations are investigated: **type 1)** The 3D Cartesian components of each station (after converting geographic coordinates, longitude, and latitude into the Cartesian ones), $\boldsymbol{x} = \{x_1, x_2, x_3\}$ is considered as the input vector. In fact, $\{x_1, x_2, x_3\}$ is the Cartesian coordinates of the station on the Earth's surface, and **type 2)** The 3D Cartesian components stacked up with $\log(V_{s_{30}})$ as the 4th component, $\boldsymbol{x} = \{x_1, x_2, x_3, \log(V_{s_{30}})\}$. The former one must be employed for the regions with a fairly uniform soil condition (since it is insensitive to the soil condition) where the homogeneity assumption is valid, while the latter can be used where there is a considerable soil condition variability. One can extend the latter input vector to include various attributes of the locations such as $Z_{1.0}$ (depth to V_s = 1 km/s), $Z_{2.5}$ (depth to V_s = 2.5 km/s), R_{JB} (closest distance to the surface projection of coseismic rupture). The type 2 model's input vector attributes are all normalized, such that each of them has zero mean and unit standard deviation.

In this study, all θ_i are considered to be the same. In other words, one length-scale is used to normalize all attributes within an input vector. Such a covariance function is called isotropic. As an alternative, anisotropic covariance functions where each attribute has its own specific length-scale also can be used,

currently under investigation by the authors. Although all the attributes are embedded within one input vector given to a covariance function in this study, it is possible to use a combination of the covariance functions (multiplication, summation, etc.) each of which is given by a subset of attributes. As an example, a multiplication of two *Matern* functions could be used such that

$$k_{GP}(\boldsymbol{x}, \boldsymbol{x}') = k_1(r_1) \times k_2(r_2)$$

$$r_1 = \sqrt{\sum_{i=1}^{3} \theta_1^2 (x_i - x'_i)^2}\,, \qquad r_2 = \sqrt{\theta_2^2 (\log(V_{s_{30}}) - \log(V'_{s_{30}}))^2} \tag{12}$$

where $k_{GP}(\boldsymbol{x}, \boldsymbol{x}')$ denotes the covariance function of the GP between two stations $\boldsymbol{x}$ and $\boldsymbol{x}'$, k_1 and k_2 are *Matern* functions given by Equation (7) (in general, they can be any kinds of covariance functions such as Squared Exponential, etc.), and r_1 and r_2 are the separation distance of the two stations along the first three attributes (Cartesian coordinates) and the last one, $\log(V_{s_{30}})$, respectively. It is worth noting that each of the k_1 and k_2 shown in Equation (12) is isotropic but not the k_{GP}. Equation (12) can be extended when there is a higher dimension of input vectors (such as considering $Z_{1.0}$, $Z_{2.5}$, and R_{JB}.)

4. MODEL TRAINING

4.1. 1906 San Francisco Physics-Based Simulated Motions

It is needed to obtain the optimum regularization factor, λ, to implement the GP regression model. To do so, the 1906 San Francisco physics-based simulated ground motions, generated by Aagard et al. are employed (Aagaard et al. 2008). Five different ground motion modeling groups, 1) Aagaard, 2) Graves, 3) Harmsen et al. 4) Larsen et al., and 5) Petersson et al studied to generate the 1906 San Francisco earthquake motions. The first four groups verified their wave-propagation codes using the 1989 Loma Prieta earthquake to evaluate how well their methods can generate ground motions consistent with the observed shaking intensities (Aagaard et al. 2008). In this study, the simulated ground motions generated by Graves wave propagation code is used. . Graves simulated the 1906 San Francisco ground motions at 40,700 stations on a 1.5 km × 1.5 km grid along two orthogonal directions, East-West (EW) and North-South (NS).

Table 1 illustrates various features of the ground-motion simulations deployed by Graves to generate the 1906 San Francisco earthquake motions. It is worth noting that although the minimum V_s for the simulation process was considered 760 m/s, the site corrections across all periods are applied to account for the nonlinear site effects for the locations with shear wave velocity lower than 760 m/s. In addition, the simulated ground motions are constructed deterministically at long periods (T > 1 sec) and stochastically at short periods (0.1 sec < T < 1 sec) (Aagaard et al. 2008). Graves simulated the 1906 San Francisco ground motions at 40,700 stations on a 1.5 km × 1.5 km grid along two orthogonal directions, East-West (EW) and North-South (NS).

Table 1. Graves Wave-Propagation codes and Domain

Domain			Resolution		Features			
Length (km)	Width (km)	Maximum depth	Bandwidth	Minimum V_s	Topography	Water	Material Properties	Attenuation
555	162	45	T > 1.0 sec	760 m/s	Bulldozed	Sediment filled	USGS 05.1.0	Graves

4.2. Hyper-Parameter's Tuning

The 1906 San Francisco simulated motions are employed to tune the regularization factor for the two GP models having two different types of input vectors discussed in Section 3. To do so, a 20 km × 65 km rectangular region is chosen for training the type 1 GP model. Then, all the stations which are simulated for the same V_s value (560 m/s) are picked within that region. For type 2, two regions, each of which with a considerable variation of site condition, are chosen for training purposes. There are 396 and 215 chosen stations for model type 1 and model type 2, respectively. Figure 1 illustrates the bounding box for the Graves simulated ground motions as well as the study regions for both type 1 and type 2 models.

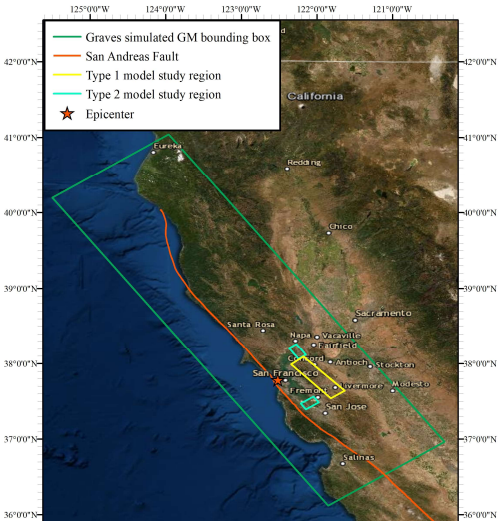


Figure 1. Graves' 1906 San Francisco simulated ground motion bounding box and the study regions corresponding to the type 1 and type 2 of the GP regression models

The GP regression model for each type of input vectors is trained (to find the optimum λ) within its corresponding study region. To do so, for each model, about 80% of the stations are randomly chosen for the training set, while the remaining 20% are considered for the test set. The distribution of the training and test set stations for both models are shown in Figure 2. From now on, the two study regions for the type 2 model are called Palo Alto and South Napa for more convenience (Figure 2.b and Figure 2.c). A cross-validation (CV) procedure with five separate folds is implemented for each training set to select the optimum regularization factor, λ, for the corresponding model. The optimum λ for each direction is picked such that it results in the lowest normalized root mean square error (NRMSE) between the observed and predicted motion's linear response spectrum (5% damping PSA). The NRMSE is computed for all periods by

$$NRMSE = \sqrt{\frac{1}{L}\sum_{i=1}^{L}\frac{(PSA_i - \widehat{PSA}_i)^2}{\widehat{PSA}_i^{\,2}}} \tag{13}$$

where L is the number of periods included in the PSA and PSA_i and $\widehat{PSA}_i$ are the predicted motion's PSA at the i^{th} period and its observed value at the test station, respectively. Eventually, there are two optimum λ for two orthogonal directions for each model. Table 2 illustrates the tuned λ for two orthogonal directions, Fault-Normal (FN) and Fault-Parallel (FP), for each model. There are two optimum λ corresponding to the two study regions for model type 2. Table 3 also illustrates the NRMSE

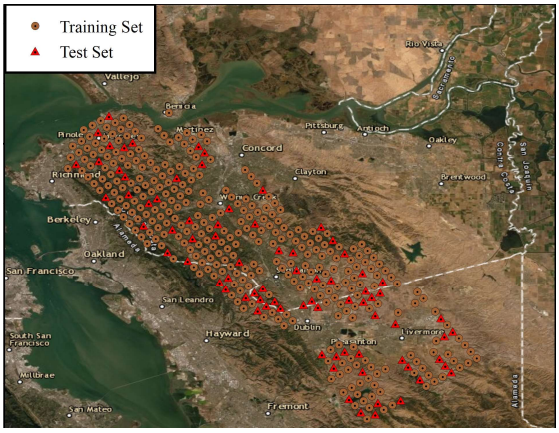
results of the CV procedure for each model along both FN and FP directions. The results shown in Table 3 are the average NRMSE among all folds' predictions for the optimum λ value given in Table 2.

Table 2. optimized regularization factor, λ, for the type 1 and type 2 models

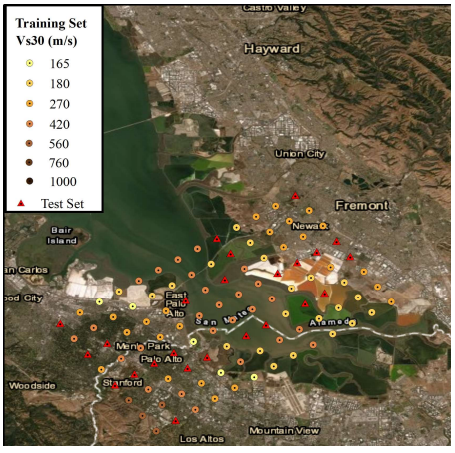
type 1		type 2	
FN	FP	FN	FP
1.2	1.2	0.7	0.7

Table 3. CV procedure NRMSE for model type 1 and type 2

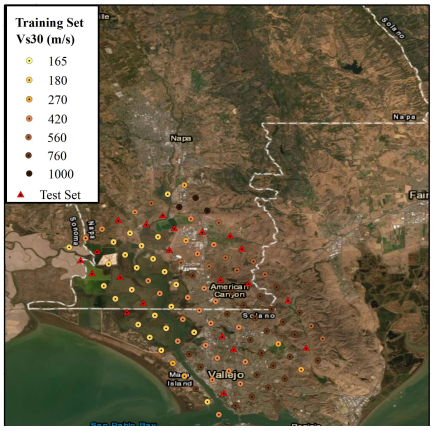
type 2 (λ = 0.7)				type 1 (λ = 1.2)	
Palo Alto		South Napa			
FN	FP	FN	FP	FN	FP
0.36	0.38	0.39	0.39	0.24	0.24



(a)



(b)

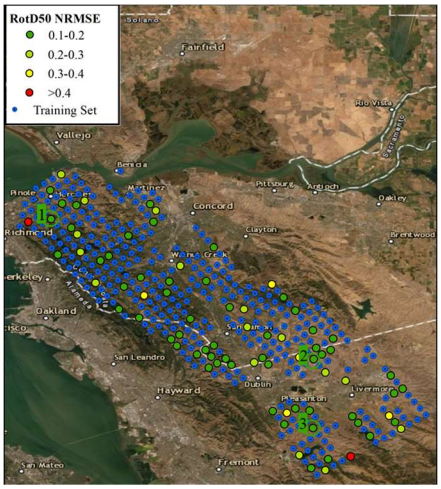


(c)

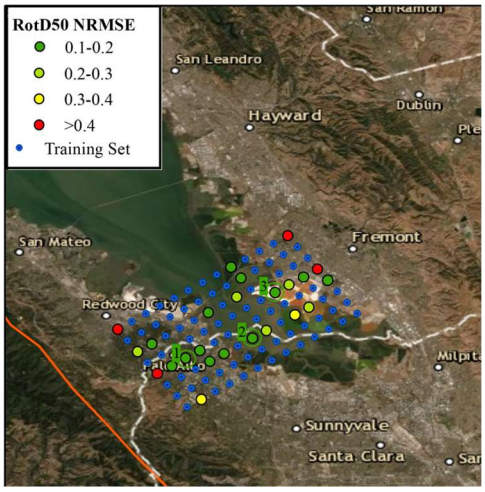
Figure 2. Distribution of the training and test set for the a) type 1, b) type 2 Palo Alto, and c) type 2 South Napa study regions

5. MODEL VALIDATION

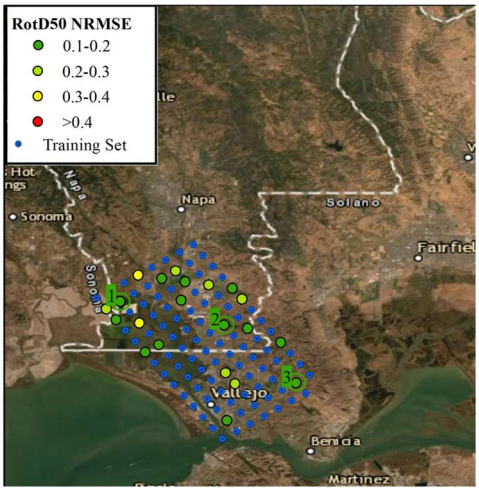
It is desired to evaluate the performance of the trained GP regression models employing the test set within each region. To do so, all the training set (circular points in Figure 2) within each region are considered as observed stations for the corresponding GP regression model. Then, each test station (triangular points in Figure 2) ground motion is predicted using the trained GP regression model and the observed motions in that region. Table 4 illustrates the average test set NRMSE between the predicted and observed motions' linear response spectrum (for 5% damping) for each model. Figure 3 demonstrates the distribution of the NRMSE between the predicted and observed motions' RotD50 spectrum (for 5% damping) for both models (Boore 2010). In Figure 3, there are three test stations picked for each study region. The prediction results of these chosen stations for RotD50, velocity time series, and Fourier Amplitude Spectrum (FAS) are shown for model type 1, model type 2 in Palo Alto, and South Napa regions in the Figures Figure 4, Figure 5, and Figure 6, respectively.



(a)



(b)



(c)

Figure 3. The distribution of the test set NRMSE between the predicted and observed motions' RotD50 spectrum (5% damping) for the model a) type 1, b) type 2 in Palo Alto, and c) type 2 in South Napa study regions

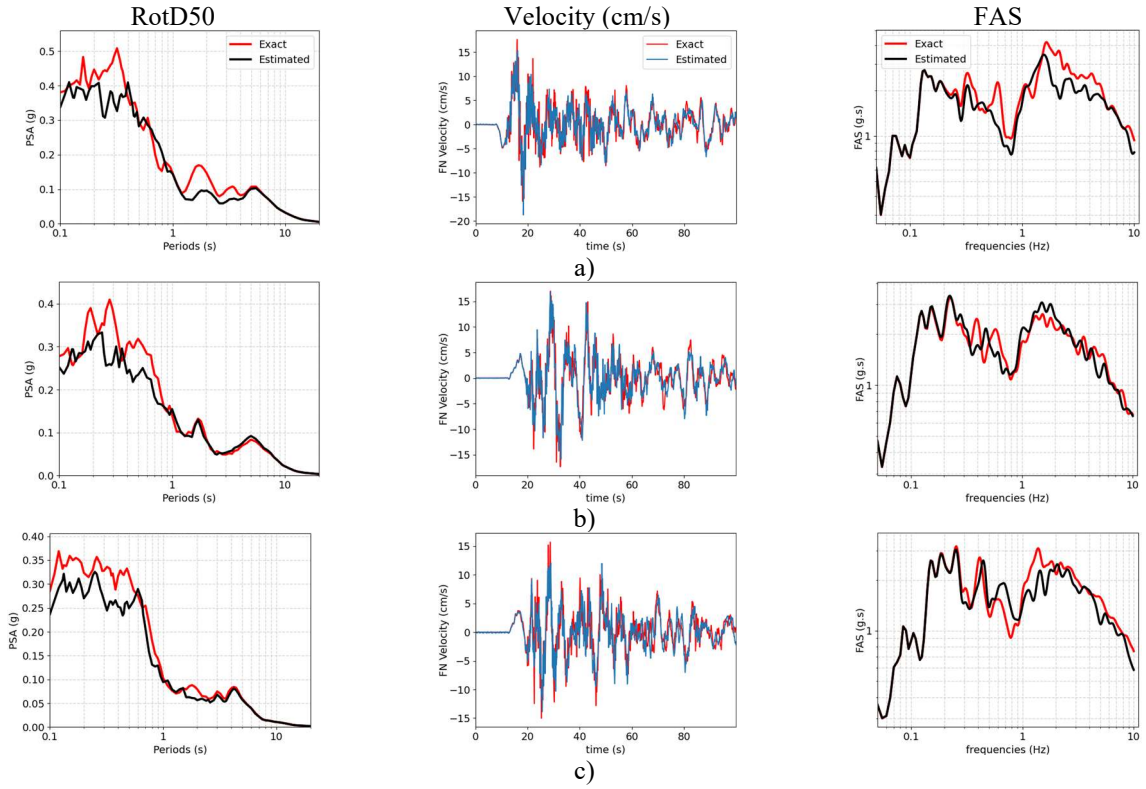


Figure 4. The RotD50, velocity time series, and FAS of the predicted as well as observed motions (Fault-Normal) for the chosen test points a) No. 1, b) No. 2, and c) No. 3 within the model type 1 study region

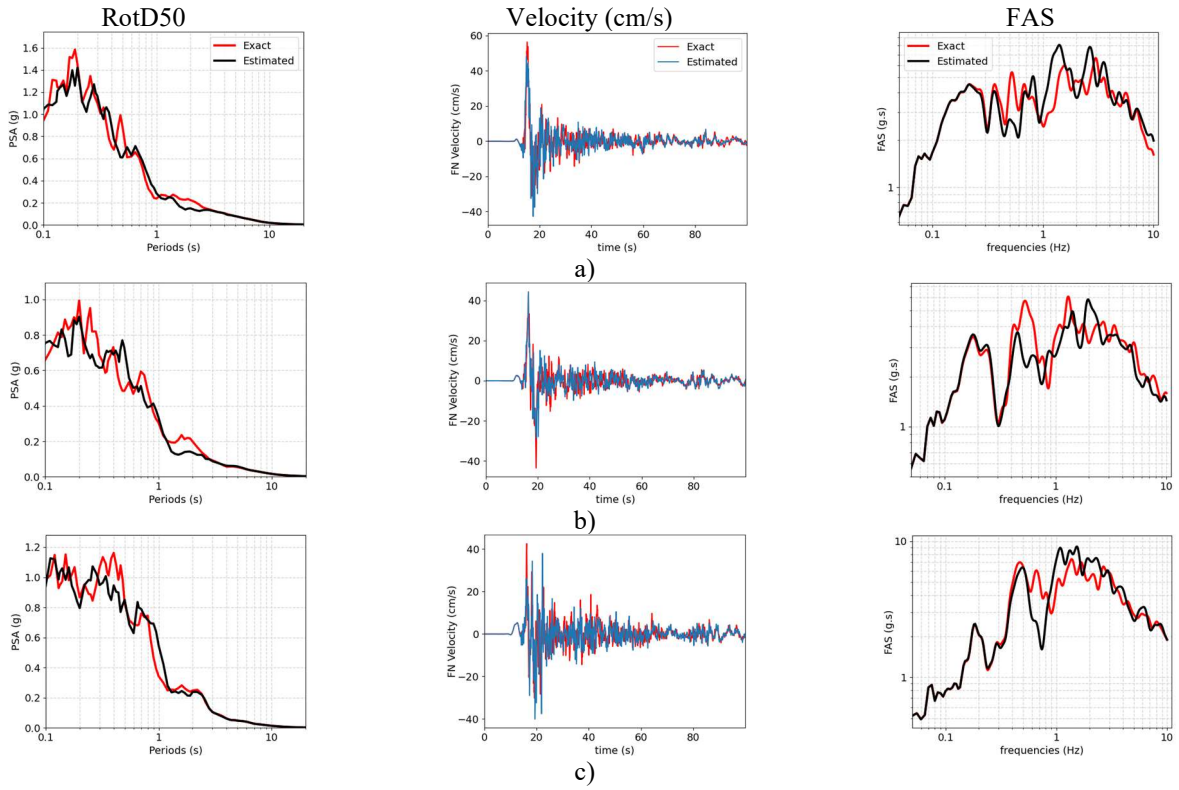


Figure 5. The RotD50, velocity time series, and FAS of the predicted as well as observed motions (Fault-Normal) for the test points a) No. 1, b) No. 2, and c) No. 3 within the model type 2 Palo Alto study region

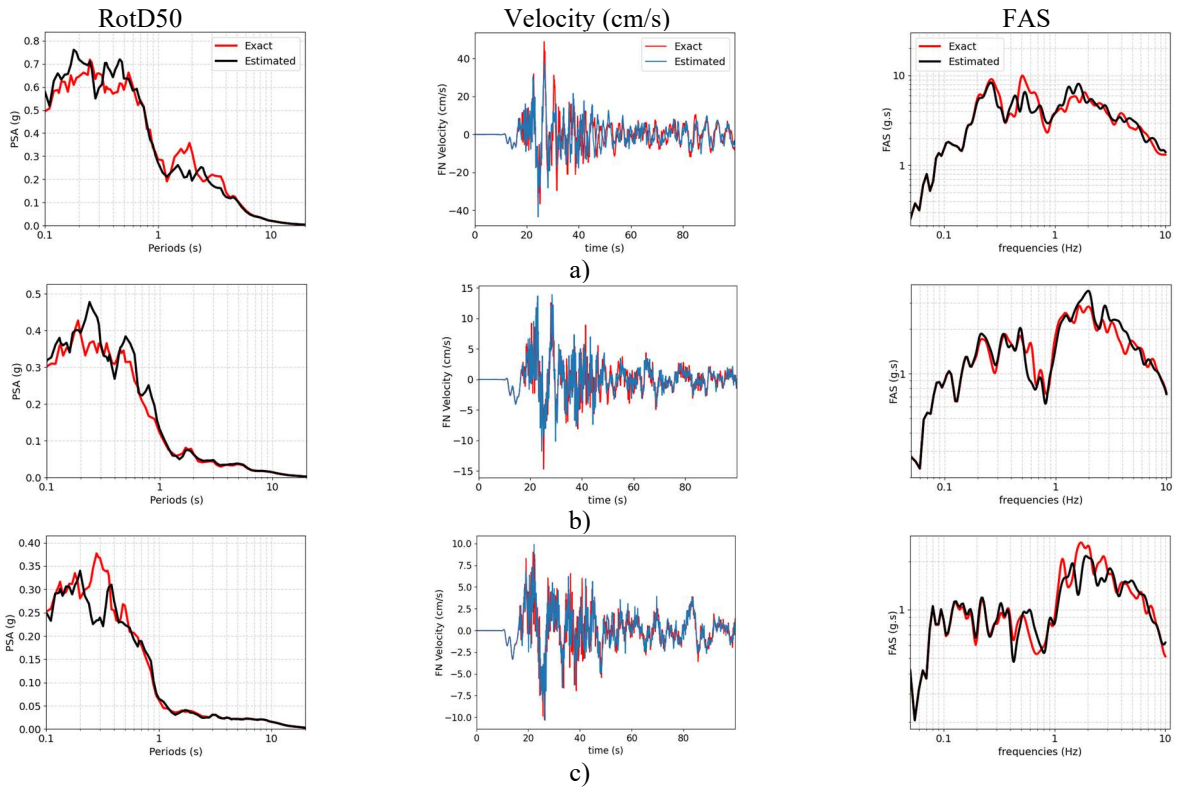


Figure 6. The RotD50, velocity time series, and FAS of the predicted as well as observed motions (Fault-Normal) for the test points a) No. 1, b) No. 2, and c) No. 3 within the model type 2 South Napa study region

Table 4. Test Set RotD50 NRMSE for model type 1 and type 2

type 2				type 1	
Palo Alto		South Napa			
FN	FP	FN	FP	FN	FP
0.34	0.38	0.23	0.26	0.23	0.23

Figure 3 demonstrates that the GP regression model is able to estimate the ground motion at most of the unobserved stations decently. Although, its estimation can be less accurate for the stations located at the boundary edges of the network (as is shown in Figure 3.a and Figure 3.b). By comparing the results of Figure 3.b and Figure 3.c, one can recognize that the estimation accuracy for the stations that are far away from the fault might be higher than that of those close to the fault. The latter could be due to the usage of an isotropic covariance model that allocates a uniform correlation to the surrounding stations based on their separation distance. The GP regression model's prediction can be improved by employing an anisotropic covariance structure for the regions close to the fault, which is currently under development by the authors. It is also noticeable in Figure 4 through Figure 6 that RotD50 estimation's error is lower for the long-period structures (> 1.0 sec) while the difference between the predicted and observed RotD50 increases for the short-period structures (< 1.0 sec.) The reason is that the short period content of the simulated motions is constructed stochastically, which results in a lower correlation among the short period contents of the neighboring motions. Thus, the GP regression model estimation for the short period content of the motion might be less accurate than longer period ones. In general, the motions are less correlated to each other at the higher-frequencies and long geographical separation distance. This phenomenon is recognizable from the existing developed lagged coherency models (Liao and Zerva

2006; Rodda and Basu 2018; Abrahamson, Schneider, and Stepp 1991). The lagged coherency, as a representative of the correlation between the motions, drops with increasing the frequency.

Eventually, it is desired to evaluate the current trained models' performance on other earthquake datasets, including simulated and real ones, such as magnitude 7.0 Hayward fault physics-based simulated motions (Rodgers et al. 2019) and 2019 Ridgecrest. Authors developed estimations similar to Figure 4 through Figure 6 for magnitude 7.1 2019 Ridgecrest ground motion dataset recorded by Community Seismic Network (CSN) within the Los Angeles using the trained GP regression model (Tamhidi et al. 2020; Clayton et al. 2020). The preliminary results illustrate that the current GP regression model can capture the spatial variation of the ground motions and is able to predict the motion time series with an acceptable error for most of the test stations for the magnitude 7.1 2019 Ridgecrest earthquake dataset. In addition, the estimated ground motions for the Hayward fault simulated motion dataset illustrated goodness of fit and promising accuracy.

6. CONCLUSION

A novel approach to estimate the entire ground motion time series at an unobserved station using its observed surrounding motions was developed. Most of the current conditioned simulation of ground motion methods are established based on the Cross Spectral Density and empirical coherency functions. These empirically-tuned models could be computationally expensive specially when an ultra-dense network of target (unobserved) stations' motions are required instantaneously after a severe earthquake. This computational cost is due to the various features needed from each target location to establish the coherency model.

In this study, the GP regression was employed to estimate the entire motion time series at the unobserved target stations. To do so, the GP regression covariance structure and its hyper-parameters were tuned. Two GP regression models were developed for the homogeneous (relatively uniform site condition) regions and the regions with a considerable site condition variation. Both models demonstrated acceptable accuracy for the estimation of the ground motion, as well as its response spectral ordinates and FAS for the 1906 San Francisco and magnitude 7.0 Hayward fault physics-based simulated ground motions as well as the 2019 M7.1 Ridgecrest Earthquake sequence recorded ground motions within its useable bandwidth. The trained Gaussian Process regression model were able to accurately estimate the long-period ground motion due to the ground motion directivity pulses. The estimation of the motions for the stations located either at the edges of the network, where there is no uniform distribution of the observed motions around the target station or at the regions with fewer observations might not be as accurate as other stations' prediction. It is worth noting that the latter might differ from one earthquake dataset to another one.

The expansion of the established GP regression models by considering other site attributes such as $Z_{1.0}$, $Z_{2.5}$, and R_{JB} as well as combining the covariance functions is introduced. In addition, it is necessary to quantify the uncertainty of the estimated motion at the target stations. To do so, it is required to randomly sample the target station's motion's amplitude at each frequency using the posterior distribution of the real and imaginary parts. The latter is under development by the authors.

ACKNOWLEDGMENTS

This study was partially supported by the University of California, Los Angeles (UCLA) Graduate Division Fellowship to the first author, which is gratefully acknowledged. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the University of California, Los Angeles. The authors would like to thank Dr. Robert Graves, Dr. Arthur Rodgers, and Dr. Monica Kohler for providing their ground-motion datasets. We would like to thank Dr. Tadahiro Kishida for his assistance in organizing the dataset. We benefitted from constructive discussions with Dr. Silvia Mazzoni as well.

REFERENCES

Aagaard, Brad T., Thomas M. Brocher, David Dolenc, Douglas Dreger, Robert W. Graves, Stephen Harmsen, Stephen Hartzell et al. "Ground-Motion Modeling of the 1906 San Francisco Earthquake, Part II: Ground-Motion Estimates for the 1906 Earthquake and Scenario EventsGround-Motion Modeling of the 1906 San Francisco Earthquake, Part II." *Bulletin of the Seismological Society of America* 98, no. 2 (2008): 1012-1046.

Aagaard, Brad T., Thomas M. Brocher, David Dolenc, Douglas Dreger, Robert W. Graves, Stephen Harmsen, Stephen Hartzell, Shawn Larsen, and Mary Lou Zoback. "Ground-motion modeling of the 1906 San Francisco earthquake, Part I: Validation using the 1989 Loma Prieta earthquake." *Bulletin of the Seismological Society of America* 98, no. 2 (2008): 989-1011.

Abrahamson, N. A., Schneider J. F., and Stepp. J. C. "Empirical spatial coherency functions for application to soil-structure interaction analyses." Earthquake spectra 7, no. 1 (1991): 1-27.

Abramowitz, Milton, and Irene A. Stegun. "Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables." National Bureau of Standards Applied Mathematics Series 55. Tenth Printing. (1972).

Adanur, Suleyman, Ahmet C. Altunisik, Kurtulus Soyluk, A. Aydin Dumanoglu, and Alemdar Bayraktar. "Contribution of local site-effect on the seismic response of suspension bridges to spatially varying ground motions." *Earthquakes and Structures* 10, no. 5 (2016): 1233-1251.

Alimoradi, Arzhang, and James L. Beck. "Machine-learning methods for earthquake ground motion analysis and simulation." Journal of Engineering Mechanics 141, no. 4 (2015): 04014147.

Baker, Jack W., and Yilin Chen. "Ground motion spatial correlation fitting methods and estimation uncertainty." Earthquake Engineering & Structural Dynamics (2020).

Boore, David M. "Orientation-independent, nongeometric-mean measures of seismic intensity from two horizontal components of motion." *Bulletin of the Seismological Society of America* 100, no. 4 (2010): 1830-1835.

Clayton, Robert W., Monica Kohler, Richard Guy, Julian Bunn, Thomas Heaton, and Mani Chandy. "CSN-LAUSD network: A dense accelerometer network in Los Angeles Schools." Seismological Research Letters 91, no. 2A (2020): 622-630.

Der Kiureghian, A. "A coherency model for spatially varying ground motions." *Earthquake engineering & structural dynamics* 25, no. 1 (1996): 99-111.

Fan, Jianqing, and Runze Li. "Variable selection via nonconcave penalized likelihood and its oracle properties." Journal of the American statistical Association 96, no. 456 (2001): 1348-1360.

Fraser, William A., David J. Wald, and Kuo-Wan Lin. "Using ShakeMap and ShakeCast to Prioritize Post-Earthquake Dam Inspections." In Geotechnical Earthquake Engineering and Soil Dynamics IV, pp. 1-10. 2008.

Gentile, Roberto, and Carmine Galasso. "Gaussian process regression for seismic fragility assessment of building portfolios." Structural Safety 87 (2020): 101980.

Huang, Duruo, and Gang Wang. "Energy-compatible and spectrum-compatible (ECSC) ground motion simulation using wavelet packets." Earthquake Engineering & Structural Dynamics 46, no. 11 (2017): 1855-1873.

Jayaram, Nirmal, and Jack W. Baker. "Correlation model for spatially distributed ground-motion intensities." *Earthquake Engineering & Structural Dynamics* 38, no. 15 (2009): 1687-1708.

Kameda, Hiroyuki, and Hitoshi Morikawa. "An interpolating stochastic process for simulation of conditional random fields." *Probabilistic Engineering Mechanics* 7, no. 4 (1992): 243-254.

Konakli, Katerina, and Armen Der Kiureghian. "Simulation of spatially varying ground motions including incoherence, wave-passage and differential site-response effects." *Earthquake Engineering & Structural Dynamics* 41, no. 3 (2012): 495-513.

Landwehr, Niels, Nicolas M. Kuehn, Tobias Scheffer, and Norman Abrahamson. "A nonergodic ground-motion model for California with spatially varying coefficients." *Bulletin of the Seismological Society of America* 106, no. 6 (2016): 2574-2583.

Li, Runze, and Agus Sudjianto. "Analysis of computer experiments using penalized likelihood in Gaussian Kriging models." *Technometrics* 47, no. 2 (2005): 111-120.

Liao, S., and A. Zerva. "Physically compliant, conditionally simulated spatially variable seismic ground motions for performance-based design." Earthquake engineering & structural dynamics 35, no. 7 (2006): 891-919.

Lin K., Wald D., Kircher C., Slosky D., Jaiswal K., and Luco. N. "USGS SHAKECAST SYSTEM ADVANCEMENTS." (2018).

Oppenheim, Alan V., Alan S. Willsky, and S. Hamid Nawab. "Signals and systems Prentice Hall." Inc., Upper Saddle River, New Jersey 7458 (1997).

Otake, Ryota, Jun Kurima, Hiroyuki Goto, and Sumio Sawada. "Deep Learning Model for Spatial Interpolation of Real-Time Seismic Intensity." Seismological Research Letters (2020).

Rasmussen, Carl Edward, and Christopher KI Williams. "Gaussian Processes for Machine Learning", the MIT Press. Cambridge, Mass (2006).

Rodda, Gopala Krishna, and Dhiman Basu. "Spatial variation and conditional simulation of seismic ground motion." *Bulletin of Earthquake Engineering* 16, no. 10 (2018): 4399-4426.

Rodda, Gopala Krishna, and Dhiman Basu. "On Conditional Simulation of Spatially Varying Rotational Ground Motion." *Journal of Earthquake Engineering* (2019): 1-36.

Rodgers, Arthur J., N. Anders Petersson, Arben Pitarka, David B. McCallen, Bjorn Sjogreen, and Norman Abrahamson. "Broadband (0–5 Hz) fully deterministic 3D ground-motion simulations of a magnitude 7.0 Hayward fault earthquake: Comparison with empirical ground-motion models and 3D path and site effects from source normalized intensities." Seismological Research Letters 90, no. 3 (2019): 1268-1284.

Sajedi, Seyed Omid, and Xiao Liang. "A data-driven framework for near real-time and robust damage diagnosis of building structures." *Structural Control and Health Monitoring* 27, no. 3 (2020): e2488.

Sheibani, Mohamadreza, Ge Ou, and Shandian Zhe. "Rapid Seismic Risk Assessment of Structures with Gaussian Process Regression." Dynamic Substructures, Volume 4. Springer, Cham, 2020. 159-165.

Southern California Earthquake Data Center. [Online]. Available: https://service.scedc.caltech.edu/SCSNStationMap/station.html

Sun, Han, Henry Burton, Yu Zhang, and John Wallace. "Interbuilding interpolation of peak seismic response using spatially correlated demand parameters." Earthquake Engineering & Structural Dynamics 47, no. 5 (2018): 1148-1168.

Tamhidi, A., Kuehn, N., Bozorgnia, Y., Taciroglu, E., & Kishida, T. "Prediction of Ground-Motion Time-Series at an arbitrary location using Gaussian Process Interpolation: Application to the Ridgecrest Earthquake". Poster Presentation at 2019 SCEC Annual Meeting, (2019).

Tamhidi, A., Kuehn, N. M., Kohler, M. D., Ghahari, F., Taciroglu, E., & Bozorgnia, Y., "Ground-Motion Time-Series Interpolation within the Community Seismic Network using Gaussian Process Regression: Application to the 2019 Ridgecrest Earthquake". Poster Presentation at 2020 SCEC Annual Meeting, (2020).

Tian, Li, Xia Gai, Bing Qu, Hongnan Li, and Peng Zhang. "Influence of spatial variation of ground motions on dynamic responses of supporting towers of overhead electricity transmission systems: An experimental study." *Engineering Structures* 128 (2016): 67-81.

Tibshirani, Robert. "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society: Series B (Methodological) 58, no. 1 (1996): 267-288.

Todorovska, Maria I., Haiping Ding, and Mihailo D. Trifunac. "Coherency of Synthetic Earthquake Ground Motion for the

Design of Long Structures: Effect of Site Conditions." In International Collaboration in Lifeline Earthquake Engineering 2016, pp. 427-434. Reston, VA: American Society of Civil Engineers, (2017).

Wald, David Jay, Vincent Quitoriano, Charles Bruce Worden, Margaret Hopper, and James W. Dewey. "USGS “Did You Feel It?” internet-based macroseismic intensity maps." Annals of geophysics 54, no. 6 (2012).

Wald, David, Kuo-Wan Lin, Keith Porter, and Loren Turner. "ShakeCast: Automating and improving the use of ShakeMap for post-earthquake decision-making and response." Earthquake Spectra 24, no. 2 (2008): 533-553.

Worden, C. Bruce, Eric M. Thompson, Jack W. Baker, Brendon A. Bradley, Nicolas Luco, and David J. Wald. "Spatial and spectral interpolation of ground-motion intensity measure observations." Bulletin of the Seismological Society of America 108, no. 2 (2018): 866-875.

Wu, Yongxin, Yufeng Gao, Ning Zhang, and Dayong Li. "Simulation of spatially varying ground motions in V-shaped symmetric canyons." Journal of Earthquake Engineering 20, no. 6 (2016): 992-1010.

Zerva, Aspasia. "Spatial variation of seismic ground motions: modeling and engineering applications.", Crc Press, (2009).

Zerva, Aspasia, Mohammad Reza Falamarz-Sheikhabadi, and Masoud Khazaei Poul. "Issues with the use of spatially variable seismic ground motions in engineering applications." In European Conference on Earthquake Engineering Thessaloniki, Greece, pp. 225-252. Springer, Cham, 2018.

Zerva, Aspasia and Zervas Vassilios. "Spatial variation of seismic ground motions: an overview." Applied Mechanics Reviews 55, no. 3 (2002): 271-297.