Conditioned Simulation of Ground-Motion Time Series at Uninstrumented Sites Using Gaussian Process Regression

Aidin Tamhidi[*1], Nicolas Kuehn[2], S. Farid Ghahari[1], Arthur J. Rodgers[3], Monica D. Kohler[4], Ertugrul Taciroglu[1], and Yousef Bozorgnia[1]

ABSTRACT

Ground-motion time series are essential input data in seismic analysis and performance assessment of the built environment. Because instruments to record free-field ground motions are generally sparse, methods are needed to estimate motions at locations with no available ground-motion recording instrumentation. In this study, given a set of observed motions, ground-motion time series at target sites are constructed using a Gaussian process regression (GPR) approach, which treats the real and imaginary parts of the Fourier spectrum as random Gaussian variables. Model training, verification, and applicability studies are carried out using the physics-based simulated ground motions of the 1906 M_w 7.9 San Francisco earthquake and M_w 7.0 Hayward fault scenario earthquake in northern California. The method's performance is further evaluated using the 2019 M_w 7.1 Ridgecrest earthquake ground motions recorded by the Community Seismic Network stations located in southern California. These evaluations indicate that the trained GPR model is able to adequately estimate the ground-motion time series for frequency ranges that are pertinent for most earthquake engineering applications. The trained GPR model exhibits proper performance in predicting the long-period content of the ground motions as well as directivity pulses.

KEY POINTS

- Ground-motion time series are not available for uninstrumented sites.
- We estimate the time series at a site based on observed ground motions at surrounding sites.
- The method can be used for estimating and understanding causes of earthquake damage at uninstrumented sites.

INTRODUCTION

Although the number of available recorded earthquake ground motions has increased in the last few decades, current sensor networks are still sparse for various site-specific earthquake applications. Thus, an estimation of either ground-motion intensity measures (GMIM), for example, peak ground acceleration, peak ground velocity (PGV), and spectral response ordinates, or the entire ground-motion time series is required to evaluate the damage state or performance level of a specific structure for post-event assessment. Because there are only approximately 2000 ground-level stations to record the free-field ground motions in California (Southern California Seismic Network, Northern California Seismic Network, California Strong Motion Instrumentation Program) (Southern California Earthquake Data Center, 2021), site-specific structural assessments invariably require estimations using interpolation methods.

Presently, "ShakeCast," "ShakeMap," and the U.S. Geological Survey (USGS) "Did You Feel It?" platforms offer estimates of the shaking level and GMIM after an event using various techniques (Fraser *et al.*, 2008; Wald *et al.*, 2008, 2012; Lin *et al.*, 2018; Worden *et al.*, 2018). Some of these techniques involve estimating the GMIMs at target sites using the

1. Civil and Environmental Engineering Department, University of California, Los Angeles, California, U.S.A., https://orcid.org/0000-0002-3254-1720 (AT); https://orcid.org/0000-0002-3847-5277 (SFG); https://orcid.org/0000-0001-9618-1210 (ET); https://orcid.org/0000-0003-1773-2489 (YB); 2. University of California, Los Angeles, J. Garrick Institute for the Risk Science, Los Angeles, California, U.S.A., https://orcid.org/0000-0002-3512-5300 (NK); 3. Lawrence Livermore National Laboratory, Livermore, California, U.S.A., https://orcid.org/0000-0002-6784-5695 (AJR); 4. Department of Mechanical and Civil Engineering, California Institute of Technology, Pasadena, California, U.S.A., https://orcid.org/0000-0002-4703-190X (MDK)

*Corresponding author: aidintamhidi@ucla.edu

Cite this article as Tamhidi, A., N. Kuehn, S. F. Ghahari, A. J. Rodgers, M. D. Kohler, E. Taciroglu, and Y. Bozorgnia (2021). Conditioned Simulation of Ground-Motion Time Series at Uninstrumented Sites Using Gaussian Process Regression, *Bull. Seismol. Soc. Am.* **112**, 331–347, doi: 10.1785/0120210054

surrounding observations (Worden *et al.*, 2018; Baker and Chen, 2020; Otake *et al.*, 2020). However, for nonlinear response-history analyses of structural systems and analysis of the degree and distribution of damage in a structure, the entire ground-motion time series is needed. Therefore, the generation of realistic time series is needed at sites where recorded motions are not available (Petrone *et al.*, 2020). The generated motions should be able to capture reasonable variations in the amplitude, phase, and frequency content over an area (Zerva and Zervas, 2002; Zerva, 2009; Chen and Baker, 2019) because such spatial variations can have considerable effects, especially on distributed lifeline structures (Jayaram and Baker, 2009; Adanur *et al.*, 2016; Tian *et al.*, 2016; Todorovska *et al.*, 2017; Zerva *et al.*, 2018).

There has been extensive work on "conditioned ground-motion simulations" wherein the time series at target sites are constructed using surrounding measurements (Kameda and Morikawa, 1992; Konakli and Der Kiureghian, 2012; Zentner, 2013; Alimoradi and Beck, 2015; Wu *et al.*, 2016; Huang and Wang, 2017; Rodda and Basu, 2018, 2019; Lu *et al.*, 2021). The majority of conditioned ground-motion simulations are based on the use of cross-spectral density (CSD) and auto-spectral density (ASD) functions to determine the covariance between the Fourier series coefficients for neighboring stations (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 CSD is determined using coherency functions, and the coefficients of these functions are assigned empirically using data-driven methods (Abrahamson *et al.*, 1991). Moreover, a detailed description of the site properties and wave propagation characteristics is sometimes needed for generating the simulated motions, which can be computationally expensive and thus time-consuming, especially when an ensemble of ultradense sites is needed.

In this study, the Gaussian process regression (GPR) method, also known as Kriging (Rasmussen and Williams, 2006), is employed to generate the ground-motion time series at target sites where there are no available recording instruments. This method is able to construct the entire ground-motion time series properly at the target site using limited input information such as geographical coordinates and the average shear-wave velocity in the uppermost 30 m, V_{S30}, from each site. Therefore, it is able to estimate the motion time series with lower computational costs in comparison with the aforementioned methods. The GPR method spatially interpolates the real and imaginary parts of the observed frequency content of the neighboring motions using an assumed covariance function to establish the ground-motion time series at the target site. The spatial correlation of the ground motions is computed to estimate the entire time series at a target site using the observed dataset.

THEORETICAL BACKGROUND

Suppose the ground-motion acceleration time series, $a_s(t)$, at location s is constructed of N discrete data points, $a_s(t_i)$, $i = 1, \ldots, N$, at equal time intervals, Δt. The accelerations, $a_s(t_i)$, are then expressed using their discrete Fourier transform (DFT) coefficients A_k (e.g., Oppenheim *et al.*, 1997) as

$$a_s(t_i) = \sum_{k=0}^{N-1} A_k e^{j\omega_k t_i}, \tag{1}$$

in which

$$A_k = \frac{1}{N}\sum_{i=0}^{N-1} a_s(t_i)[\cos(\omega_k t_i) + j\sin(\omega_k t_i)] = \mathcal{R}e_k + j\mathcal{I}m_k. \tag{2}$$

In equations (1) and (2), ω_k is the k^{th} natural frequency (at equal frequency intervals) of the DFT, $j = \sqrt{-1}$ and $\mathcal{R}e_k$ and $\mathcal{I}m_k$ are the real and imaginary parts of the DFT coefficient, A_k, respectively, at the k^{th} frequency.

Here, we assume that $\mathcal{R}e_k$ and $\mathcal{I}m_k$ (at k^{th} frequency, $k = 0,\ldots, N-1$) are random Gaussian variables for any location, s, within a region. We also consider that $\mathcal{R}e_k$ at location s, is spatially correlated to $\mathcal{R}e_k^{'}$ at location $s^{'}$ where s and $s^{'}$ are neighbors. A similar assumption for $\mathcal{I}m_k$ is taken. In this study, we aim to implement GPR as a method to estimate the values of $\mathcal{R}e_k$ (and $\mathcal{I}m_k$) at the k^{th} frequency (k = 0, …, N − 1) using the corresponding $\mathcal{R}e_k^{'}$ (and $\mathcal{I}m_k^{'}$) from the surrounding station observations. We then reconstruct the entire acceleration time series at the target location with all estimated $\mathcal{R}e_k$ and $\mathcal{I}m_k$ using equation (1).

It is assumed that there is a statistically insignificant correlation between $\mathcal{R}e_k$ (or similarly $\mathcal{I}m_k$) and $\mathcal{R}e_j$ (or similarly $\mathcal{I}m_j$) at the same location, s, for different frequencies k and j, in which $k \neq j$, to construct the mean estimated ground-motion time series. It is worth noting that the mean estimated values for multivariate Gaussian variables (here $\mathcal{R}e$ and $\mathcal{I}m$) are independent of the interfrequency correlation between amplitudes at various frequencies; yet, the interfrequency correlations of the DFT coefficients need to be accounted for generating random ground-motion realizations (see the Realizations of Ground Motion section).

Gaussian process regression

GPR is a supervised learning method that has numerous applications in earthquake engineering and seismology, such as ground-motion time-series estimation, post-earthquake damage assessment, development of performance models of engineering materials, and seismic fragility assessment (Landwehr *et al.*, 2016; Sun *et al.*, 2018; Tamhidi *et al.*, 2019, 2020; Gentile and Galasso, 2020; Ghaderi *et al.*, 2020; Sajedi and Liang, 2020; Sheibani and Ou, 2020). A Gaussian process (GP) (Rasmussen and Williams, 2006) is a collection of indexed random variables such that every finite subset is distributed according to a multivariate normal distribution. In general terms, GP can be

understood as a multivariate normal distribution for an infinite number of random variables. More specifically, GP is a distribution over function $f(\boldsymbol{x}) \in \mathrm{R}$

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}^{'})), \tag{3}$$

which reads as "the function value, $f(\boldsymbol{x})$, at input location $\boldsymbol{x}$ is drawn from a GP with the mean function, $m(\boldsymbol{x})$, and the covariance function $k(\boldsymbol{x}, \boldsymbol{x}^{'})$." As equation (3) indicates, a GP is entirely defined by its mean, $m(\boldsymbol{x})$, and covariance, $k(\boldsymbol{x}, \boldsymbol{x}^{'})$ functions, which are

$$m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})], \tag{4}$$

$$k(\boldsymbol{x}, \boldsymbol{x}^{'}) = \mathbb{E}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x}^{'}) - m(\boldsymbol{x}^{'}))]. \tag{5}$$

In equations (4) and (5), $\mathbb{E}$ stands for mathematical expectation. The covariance function $k(\boldsymbol{x}, \boldsymbol{x}^{'})$ indicates the degree of similarity between function values at data points, $\boldsymbol{x}$ and $\boldsymbol{x}^{'}$. In Bayesian nonparametric statistics (Hjort *et al.*, 2010), a GP is often used to specify a prior distribution over possible functions. Here, we assume that the real and imaginary parts of the DFT coefficients (compare with equation 2) are functions of the location and possibly other geotechnical or seismological parameters such as the local site condition. Because the functional form is unknown and complicated, we replace it with a GP. To estimate the DFT coefficients, we carry out a GPR over the observed values, $\boldsymbol{f}$, which are either the real or imaginary parts of the DFT coefficients at each frequency.

It is worth noting that other regression methods such as Nadaraya–Watson kernel regressions (Nadaraya, 1964; Watson, 1964) or the Savitzky–Golay filter (Savitzky and Golay, 1964) are possible alternatives for interpolation purposes. Both GPR (see the Realizations of Ground Motion section) and Kernel regression methods (Rubin, 1981) can estimate the uncertainty of the predicted values. The Kernel regression methods can implement adaptive kernels that change with data (Huang *et al.*, 2014), and the GPR is able to implement a combination of multiple kernel functions (through multiplication and summation) to estimate the covariance among observations from a complex function. In this study, we use GPR as it simultaneously optimizes the covariance function parameters based on observations and estimates GP values at target locations without imposing a high computational cost.

One can compute the predictive distribution for function values $\boldsymbol{f}_*$ at new (target) locations (without recorded ground motions) by conditioning on the observed data. The joint distribution of observed data and new simulated data is

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu}_* \end{bmatrix}, \begin{bmatrix} \boldsymbol{K}_{\boldsymbol{xx}} + \sigma_y^2 \boldsymbol{I} & \boldsymbol{K}_{\boldsymbol{xx}*} \\ \boldsymbol{K}_{\boldsymbol{x}*\boldsymbol{x}} & \boldsymbol{K}_{\boldsymbol{x}_*\boldsymbol{x}_*} \end{bmatrix} \right), \tag{6}$$

in which $\boldsymbol{K}_{\boldsymbol{xx}}$ denotes the covariance matrix of the DFT coefficients at the observed locations. The entries of $\boldsymbol{K}_{\boldsymbol{xx}}$ are calculated from the covariance function via $\boldsymbol{K}_{\boldsymbol{xx}_{ij}} = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$, in which k denotes the covariance function between two locations (compare with equation 5). Correspondingly, $\boldsymbol{K}_{\boldsymbol{xx}_*}$ describes the covariance between the observed DFT coefficients and the estimated ones at the target locations, and $\boldsymbol{K}_{\boldsymbol{x}_*\boldsymbol{x}_*}$ are the covariances of the DFT coefficients at the target locations. The term σ_y denotes the observation noise; $\boldsymbol{I}$ is the identity matrix; and $\boldsymbol{\mu}$ and $\boldsymbol{\mu}_*$ are the prior mean vectors at the observed and target locations, respectively. Here, the observed ground motions, and subsequently their DFT coefficients, are considered noise-free ($\sigma_y = 0$). The predictive distribution for the function values $\boldsymbol{f}_*$ at the target locations is then (Rasmussen and Williams, 2006)

$$\boldsymbol{f}_* | \boldsymbol{X}_*, \boldsymbol{X}, \boldsymbol{f} \sim \mathcal{N}(\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_{**}), \tag{7}$$

in which

$$\boldsymbol{\mu}_* = \boldsymbol{\mu} + \boldsymbol{K}_{\boldsymbol{x}_*\boldsymbol{x}} \boldsymbol{K}_{\boldsymbol{xx}}^{-1} (\boldsymbol{f} - \boldsymbol{\mu}), \tag{8}$$

$$\boldsymbol{\Sigma}_{**} = \boldsymbol{K}_{\boldsymbol{x}_*\boldsymbol{x}_*} - \boldsymbol{K}_{\boldsymbol{x}_*\boldsymbol{x}} \boldsymbol{K}_{\boldsymbol{xx}}^{-1} \boldsymbol{K}_{\boldsymbol{xx}_*}, \tag{9}$$

and $\boldsymbol{X}$ denotes the input matrix of the observations, each row of which is one observed location's input vector including its geographical coordinates and possibly other features. Similarly, $\boldsymbol{X}_*$ is the input matrix of all new target locations.

The GPR's output and smoothness depend on the computed covariance function, which is defined based on a kernel, $k(r)$, in which r is the distance between the input vectors $\boldsymbol{x}$ and $\boldsymbol{x}^{'}$ given by the following equation:

$$r = \theta \sqrt{\sum_{i=1}^{d} (x_i - x_i^{'})^2}. \tag{10}$$

In equation (10), θ is a positive normalizing factor (also known as the inverse of length-scale, l, where $\theta = 1/l$) and d is the size of the input vector (number of attributes). There are several established covariance functions such as exponential and Matérn, which are given by

$$k_{\exp}(r) = \sigma_f^2 \exp(-r), \tag{11}$$

and

$$k_{\text{Matérn}}(r) = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu} r)^{\nu} K_{\nu}(\sqrt{2\nu} r), \tag{12}$$

respectively. In equation (12), Γ is the Gamma function $\Gamma(\mathrm{n}) = (n-1)!$; K_{ν} is the modified Bessel function (Abramowitz and Stegun, 1972); and ν is a positive parameter

that controls the smoothness of the output function. In equations (11) and (12), σ_f is the variance that governs how uncertain the GPR's estimate is for a given input location.

In this study, a single θ value is used to normalize all attributes within an input vector (compare with equation 10). Such a covariance function is called isotropic. As an alternative, an anisotropic covariance function in which each attribute has its own specific length-scale also can be used (Rasmussen and Williams, 2006). The θ value specifies the rate of decay for the covariance function. Higher values for θ (smaller length-scale) result in a faster decay of covariance, and subsequently correlation, by increasing the distance. More detailed descriptions of GPR can be found in, for example, Li and Sudjianto (2005) and chapters 2 and 4 of Rasmussen and Williams (2006).

PROPOSED MODELS

A proper input vector for the observed and target sites needs to be defined to start fitting the GPR. It is possible to consider the homogeneity assumption for regions with fairly uniform site conditions. In this case, all of the GP's stochastic descriptors depend only on the geographical separation distance between the stations (Zerva and Zervas, 2002). In this study, we consider two types of input vectors (corresponding to two GPR models) for the stations. These are namely, type 1, the 3D Cartesian components of each station (after converting the geographical coordinates longitude and latitude into 3D Cartesian coordinates), $\boldsymbol{x} = \{x_1, x_2, x_3\}$, in which the homogeneity assumption is valid. In fact, $\{x_1, x_2, x_3\}$ are the Cartesian coordinates of the station on the Earth's surface. In type 2, the 3D Cartesian components are stacked up with $\log(V_{S30})$ as the fourth component, $\boldsymbol{x} = \{x_1, x_2, x_3, \log(V_{S30})\}$, in which the homogeneity assumption is invalid.

More precisely, the GPR model type 1 is a specific case of the more inclusive GPR model type 2. The GPR model type 1 is used here to investigate the applicability of a simpler attribute vector (using only the 3D Cartesian coordinates), imposing lower computational cost for regions with fairly uniform soil conditions (i.e., the variation of V_{S30} is negligible).

The GPR model type 2 input vector can also be extended to include more 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), and R_{JB} (closest distance to the surface projection of coseismic rupture). In the GPR model type 2, input vector attributes are normalized, such that the mean and standard deviation of each distribution are zero and one, respectively. This normalization is required to convert all of the attributes to a similar range of values.

Model parameters and optimization

The parameters of the GPR model are the distance normalizing factor θ and the GP mean μ and variance σ_f, which need to be predefined to implement the GPR. Denoting the model parameters as $\gamma = (\theta, \mu, \sigma_f)$, a commonly used method to find the optimum γ is to maximize the log-marginal likelihood of the n observations given γ, using

$$\log p(\boldsymbol{f}|\boldsymbol{X}, \boldsymbol{\gamma}) = -\frac{1}{2}(\boldsymbol{f}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{K}_{\boldsymbol{xx}}^{-1}(\boldsymbol{f}-\boldsymbol{\mu}) - \frac{1}{2}\log|\boldsymbol{K}_{\boldsymbol{xx}}| - \frac{n}{2}\log 2\pi. \tag{13}$$

In equation (13), the superscript T indicates the transpose operator, and $|\boldsymbol{K}_{\boldsymbol{xx}}|$ denotes the determinant of the matrix $\boldsymbol{K}_{\boldsymbol{xx}}$. In equation (13), $\boldsymbol{\mu}$ and $\boldsymbol{K}_{\boldsymbol{xx}}$ are functions of θ (Li and Sudjianto, 2005). Parameter estimates found by maximizing equation (13) are the maximum-likelihood estimates (MLEs). The MLEs have considerable variance near their optimum solution because the likelihood function is almost flat close to its extremum, especially when observations are sparse (Li and Sudjianto, 2005). To tackle this issue, one can maximize the penalized log likelihood (log posterior) rather than log-marginal likelihood. Equation (14) shows the penalized log-likelihood, $Q(\gamma)$, formulation:

$$Q(\boldsymbol{\gamma}) = -\frac{1}{2}(\boldsymbol{f}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{K}_{\boldsymbol{xx}}^{-1}(\boldsymbol{f}-\boldsymbol{\mu}) - \frac{1}{2}\log|\boldsymbol{K}_{\boldsymbol{xx}}| - \frac{n}{2}\log 2\pi - ndp_{\lambda}(\theta). \tag{14}$$

In equation (14), $p_\lambda(\theta)$ is a nonnegative penalty function for normalizing factor θ. The λ is a nonnegative regularization factor that needs to be tuned using data-driven methods, as elaborated in the Hyperparameter Optimization section. There are several choices for the penalty function in equation (14), such as the least absolute shrinkage and selection operator (e.g., Tibshirani, 1996) and smoothly clipped absolute deviation Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001). In this study, the SCAD penalty function is used and is given by

$$p_\lambda(\theta) = \begin{cases} \lambda\theta & \theta \le \lambda \\ -\frac{\lambda^2+\theta^2-2a\lambda\theta}{2(a-1)} & \lambda < \theta \le a\lambda \\ \frac{\lambda^2(a+1)}{2} & a\lambda < \theta \end{cases}, \tag{15}$$

in which a is a constant, which is assumed to be 3.7 based on Fan and Li (2001) who illustrated that a model's performance is not considerably improved choosing a through a data-driven method. The penalized log likelihood, $Q(\boldsymbol{\gamma})$, in equation (14), is the log posterior distribution of $\boldsymbol{\gamma}$, given the observations. In other words, the maximum a posteriori estimates of parameters, $\hat{\boldsymbol{\gamma}} = (\hat{\theta}, \hat{\mu}, \widehat{\sigma_f})$, are employed as an alternative to the commonly used MLEs by maximizing equation (14). The GP mean, μ, in equation (8), and the variance, σ_f, are updated into $\hat{\mu}$ and $\widehat{\sigma_f}$, respectively, given $\hat{\theta}$ (e.g., Li and Sudjianto, 2005).

The GPR is completely defined by its optimized parameters $\hat{\theta}$, $\hat{\mu}$, and $\widehat{\sigma_f}$. The regularization factor, λ, needs to be defined before optimizing these parameters through maximizing $Q(\boldsymbol{\gamma})$. More specifically, λ governs the derivation of optimized parameters $\hat{\theta}$, $\hat{\mu}$, and $\widehat{\sigma_f}$. As a hierarchical view, one can recognize θ, μ, and σ_f as the parameters of the GPR model, whereas λ is its hyperparameter. The process of optimization of this hyperparameter is elaborated next.

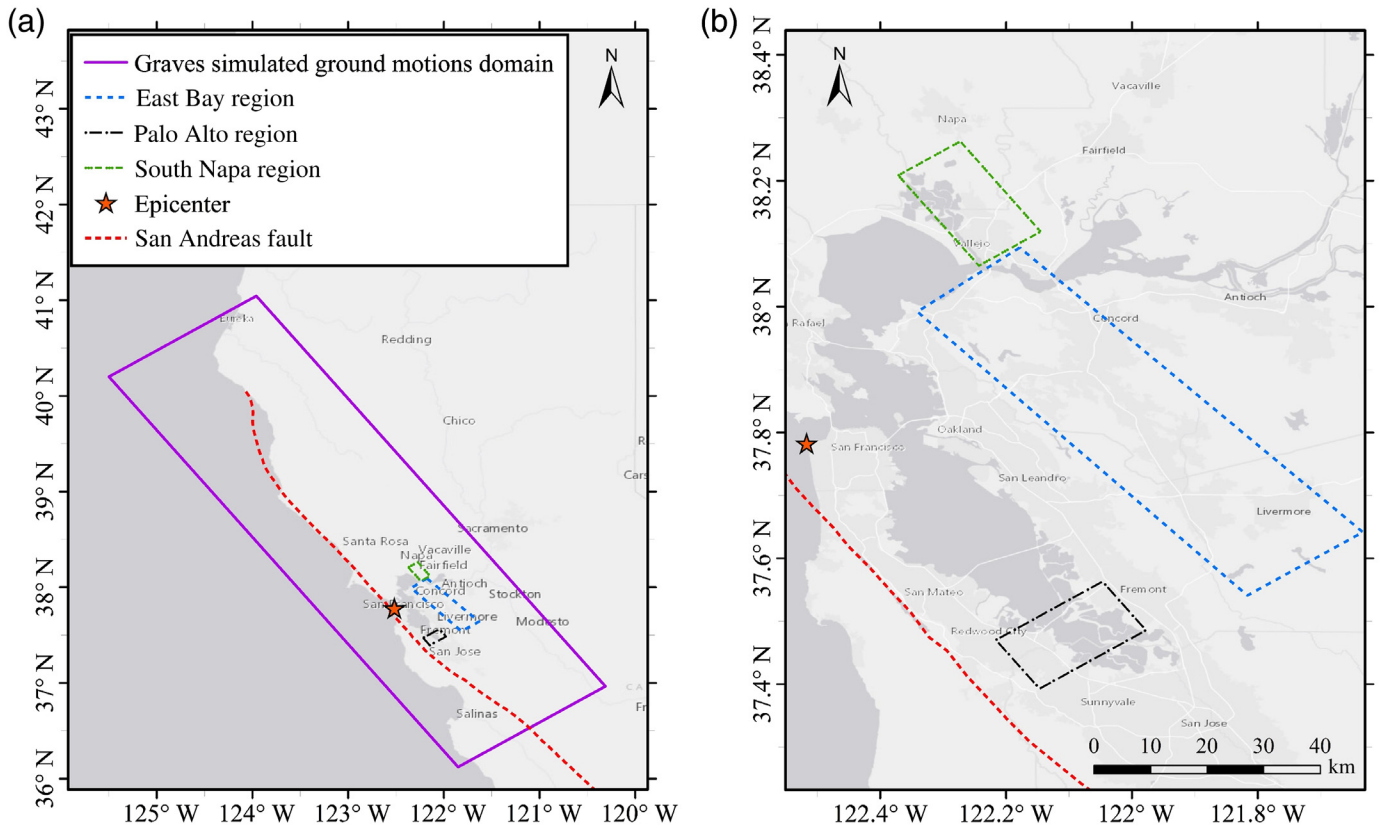

Figure 1. (a) Aagaard *et al.* (2008) 1906 M_w 7.9 San Francisco earthquake simulated ground motions domain and (b) the study regions corresponding to the type 1 (East Bay) and type 2 Gaussian process regression (GPR) models (Palo Alto and South Napa). The color version of this figure is available only in the electronic edition.

Hyperparameter optimization

It is common to use data-driven methods such as cross validation (CV) to find the optimum hyperparameter values, here the regularization factor, $\hat{\lambda}$. In our case, the "data" to be used in the "data-driven" methodology is a set of "observed" ground motions, which is a subset of physics-based simulated ground motions for the 1906 M_w 7.9 San Francisco earthquake. Here, we used broadband ground motions generated using Graves's hybrid simulation wave propagation code (Aagaard *et al.*, 2008). These ground-motion time series were generated at 40,700 locations on a 1.5 km × 1.5 km uniform grid along three orthogonal directions. Table 1 displays various features of the physics-based simulated ground motions for the 1906 M_w 7.9 San Francisco earthquake. A minimum V_{S30} value of 760 m/s was used for these simulations. Correction factors were applied for site effects at locations with V_{S30} lower than 760 m/s.

Two different optimum $\hat{\lambda}$ values must be obtained for the two GPR models introduced in the Proposed Models section. For GPR model type 1, 396 locations with the same V_{S30} (560 m/s) within a 20 km × 75 km rectangular region are chosen with the homogeneity assumption. We refer to this region hereafter as the "East Bay" region (Fig. 1). For GPR model type 2, two regions where the homogeneity assumption is invalid are chosen. These two regions are hereafter referred to as the "Palo Alto" and "South Napa" regions (Fig. 1), with 104 and 111 chosen sites, respectively. The sites within each of

TABLE 1
1906 M 7.9 San Francisco Physics-Based Simulated Wave Propagation Parameters from Aagaard *et al.* (2008)

Domain			Resolution		Features			
Length (km)	**Width (km)**	**Maximum depth**	**Bandwidth**	**Minimum V_S**	**Topography**	**Water**	**Material Properties**	**Attenuation**
555	162	45	T > 1.0 s	760 m/s	Bulldozed	Sediment filled	USGS 05.1.0	Graves (Aagaard *et al.*, 2008)

USGS, U.S. Geological Survey.

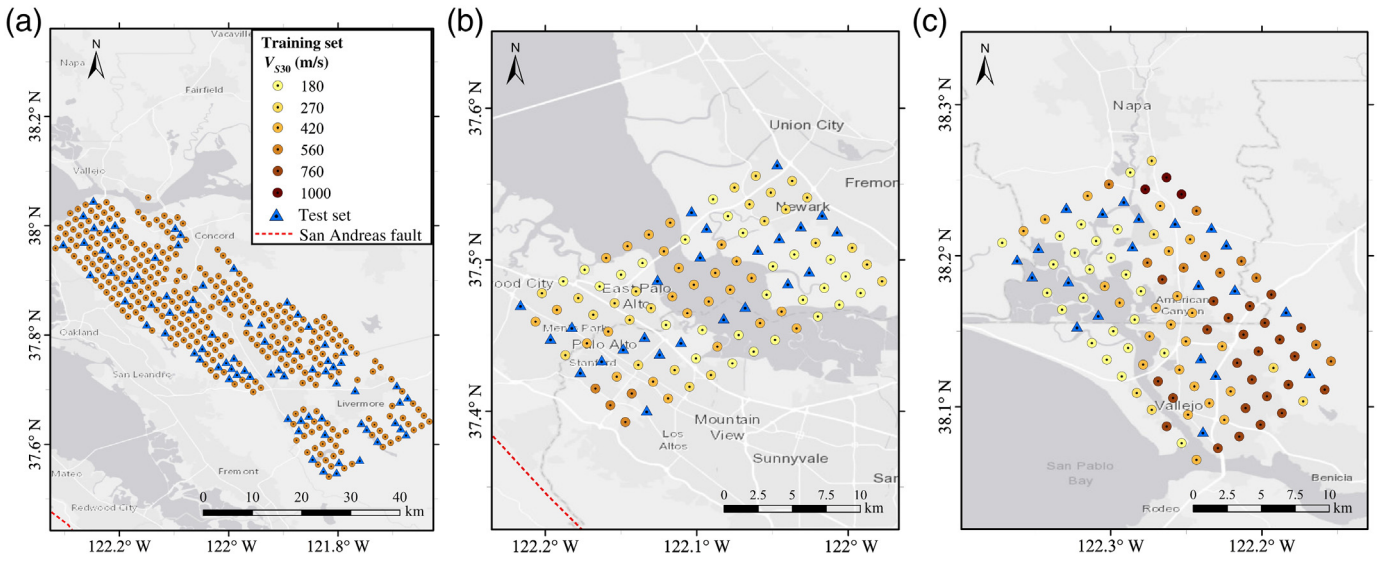

Figure 2. Distribution of the training and test sets for the (a) East Bay, (b) Palo Alto, and (c) South Napa study regions within the 1906 San Francisco simulated motions domain. The color version of this figure is available only in the electronic edition.

the East Bay, Palo Alto, and South Napa regions are randomly split into a training set (80% of the total number of sites), which makes up the "observed" ground motions, while the remaining 20% are considered the test set (target sites) (Fig. 2).

A five-fold CV procedure is implemented over the training set (observed ground motions) within each region to select the best regularization factor, $\hat{\lambda}$, for the corresponding GPR model. The accuracy criterion for this selection is the normalized root mean square error (NRMSE) between the exact (physics-based simulated) and the estimated (conditioned simulated) ground motions' 5% damped pseudo-spectral acceleration (PSA) at the target site. The NRMSE is computed as

$$\mathrm{NRMSE} = \sqrt{\frac{1}{\tau}\sum_{i=1}^{\tau}\frac{(\mathrm{PSA}_i - \widehat{\mathrm{PSA}}_i)^2}{\widehat{\mathrm{PSA}}_i^2}}, \tag{16}$$

in which, τ equals 85, which is the number of periods included in the PSA ranging from 0.1 to 20 s, and PSA_i and $\widehat{\mathrm{PSA}}_i$ are the predicted and exact ground motions' PSA values at the i^{th} period, respectively. A lower NRMSE value indicates a greater similarity between the estimated and exact response spectra. One can use Fourier amplitude spectrum (FAS) NRMSE; however, the response spectrum is smoother than the FAS, which makes the NRMSE criterion better suited to measure the degree of similarity between the estimated and exact ground motions. The NRMSE value computed across all frequencies for the FAS is highly sensitive to the rapid variations of amplitude from one frequency to another, whereas such changes are far smaller for the PSA. Moreover, the PSA spectrum is representative of the GMIM, which is commonly used for engineering applications.

The following steps are taken to select $\hat{\lambda}$. First, we randomly split the training (observed) dataset into five separate folds. For each λ_{test} to be evaluated, we carry out the following procedure:

1. For each fold i = 1, …, 5:
   1.1 Find the optimum parameters $\hat{\theta}$, $\hat{\mu}$, and $\hat{\sigma_f}$ for the observed motions within all folds except the i^{th}-fold using λ_{test} and maximizing $Q(\boldsymbol{\gamma})$ in equation (14). These parameters need to be found for each frequency and for both real and imaginary parts of the DFT coefficients.
   1.2 Estimate the ground-motion time series at each site within the i^{th}-fold using the posterior mean (equation 8) for the DFT coefficients, using $\hat{\theta}$, $\hat{\mu}$, and $\hat{\sigma}_f$ determined in step 1.1.
   1.3 Compute the NRMSE between the estimated (step 1.2) and exact ground-motion response spectra (equation 16) at each site within the i^{th}-fold, and store their averages as Error_i.
2. Take the average of Error_i (i = 1, …, 5), that is, $\mathrm{Error}_{\mathrm{avg}}$ and record it as being associated with λ_{test}.

Eventually, we choose the λ_{test} with the lowest $\mathrm{Error}_{\mathrm{avg}}$ computed in step 2 as the optimized regularization factor, $\hat{\lambda}$.

Covariance function selection

We investigate the performance of the model using three different covariance functions, exponential (compare with equation 11), Matérn with $\nu = 1.5$, and $\nu = 2.5$ (shown in equations 17 and 18), to find the optimized covariance function for the GPR model. The values $\nu = 1.5$ and $\nu = 2.5$ are widely used for Matérn covariance functions in GPR applications (Rasmussen and Williams, 2006). Exponential covariance

TABLE 2
Optimized Regularization Factor, $\hat{\lambda}$, and Obtained Average Normalized Root Mean Square Error (NRMSE) over the South Napa Training Set

Covariance Kernel	$\hat{\lambda}$ (FN)	$\hat{\lambda}$ (FP)	$\text{Error}_{\text{avg}}$ (FN)	$\text{Error}_{\text{avg}}$ (FP)
Exponential	1.3	1.3	0.36	0.36
Matérn ($\nu = 1.5$)	0.7	0.7	0.28	0.28
Matérn ($\nu = 2.5$)	0.7	0.7	0.30	0.31

FN, fault normal; FP, fault parallel.

functions have been used in the estimation of spatially distributed GMIMs (Jayaram and Baker, 2009). Matérn covariance functions have also been used to model the spatial correlation of ground motions for developing nonergodic ground-motion models (Kuehn and Abrahamson, 2020):

$$k_{\nu=1.5}(r) = \sigma_f^2(1 + \sqrt{3}r)\exp(-\sqrt{3}r), \tag{17}$$

$$k_{\nu=2.5}(r) = \sigma_f^2\left(1 + \sqrt{5}r + \frac{5}{3}r^2\right)\exp(-\sqrt{5}r). \tag{18}$$

We use the South Napa training set (Fig. 2c) to conduct the five-fold CV procedure for GPR models type 2 constructed with each of the three aforementioned covariance functions. First, the optimized regularization factor, $\hat{\lambda}$, is obtained for each of the GPR models; then, the average NRMSE ($\text{Error}_{\text{avg}}$ described in the Hyperparameter Optimization section) for the corresponding obtained $\hat{\lambda}$ is determined for each GPR model. Table 2 illustrates derived $\hat{\lambda}$ values for each covariance function as well as the average NRMSE obtained for the corresponding model over the South Napa training set in both the fault-normal (FN) and fault-parallel (FP) directions. As is illustrated in Table 2, the Matérn with $\nu = 1.5$ covariance function outperformed the other two covariance functions. Thus, in this study, we used the Matérn covariance function with $\nu = 1.5$ (compare with equation 17) to establish the GPR models type 1 and 2.

Table 3 displays the $\hat{\lambda}$ values for the FN and FP directions for each model using the Matérn ($\nu = 1.5$) covariance function. The CV procedure yielded the same $\hat{\lambda}$ values within the South Napa and Palo Alto study regions for the GPR model type 2. It is worth noting that the optimized regularization factor, $\hat{\lambda}$, is dependent on the density of the observations (recall that this is the number of observed sites divided by the area of the network). We observe that a smaller number of available observations (lower density of observed sites) leads to higher required regularization factor values, consistent with Li and Sudjianto (2005). The observation densities for the East Bay and Palo Alto (or similarly South Napa) regions are 0.26 and 0.29 sites/km^2, respectively.

TABLE 3
Optimized Regularization Factor, $\hat{\lambda}$, for the Models Type 1 and Type 2

GPR Model Type 1		GPR Model Type 2	
Fault Normal	**Fault Parallel**	**Fault Normal**	**Fault Parallel**
1.2	1.2	0.7	0.7

GPR, Gaussian process regression.

MODEL EVALUATION

The performance of the proposed GPR models is evaluated next. To do so, we compare the estimated ground motions generated by the GPR model with the exact ground motions at the same sites. The following procedure is used to estimate the ground-motion time series at any target site:

1. Given the observed ground motions (training set), the model parameters $\hat{\theta}$, $\hat{\mu}$, and $\widehat{\sigma_f}$ are obtained at each frequency for the real and imaginary parts of the DFT coefficients using the corresponding $\hat{\lambda}$ given in Table 3.
2. The posterior means (equation 8) at the desired sites for the DFT coefficients are calculated for each frequency using the values of $\hat{\theta}$, $\hat{\mu}$, and $\widehat{\sigma_f}$ from step 1.
3. The entire ground-motion time series is constructed using equation (1).

The tuned $\hat{\lambda}$ based on the 1906 M_w 7.9 San Francisco earthquake subset (as elaborated in the Covariance Function Selection section) is used to validate the GPR models' estimation of ground motions in simulated datasets of the 1906 M_w 7.9 San Francisco (Aagaard *et al.*, 2008) and the M_w 7.0 Hayward fault scenario earthquakes (Rodgers *et al.*, 2019), as well as the 2019 M_w 7.1 Ridgecrest earthquake recorded by the Community Seismic Network (CSN) (Clayton *et al.*, 2020).

The 1906 M_w 7.9 San Francisco simulated motions

The training set for each study region (East Bay, Palo Alto, and South Napa) for the 1906 M_w 7.9 San Francisco earthquake simulated motions are shown in Figure 2. The corresponding GPR model is implemented for each region to estimate the ground-motion time series at each test site within the test set (colorful circular points in Fig. 3). Figure 3 illustrates the distribution of NRMSE between the estimated and exact motions' linear response spectra 5% damped RotD50 (Boore, 2010) values. In Figure 3, there are three chosen test sites for each region. The prediction results for the RotD50 spectrum, velocity time series, and FAS are shown for selected sites in the East Bay, Palo Alto, and South Napa regions in Figures 4–6, respectively. Table 4 summarizes the 1906 San Francisco test set's NRMSE for FN, FP, and RotD50 linear response spectra.

Regarding the RotD50 spectrum NRMSE, Figure 3 demonstrates that the GPR model is able to estimate the ground-motion

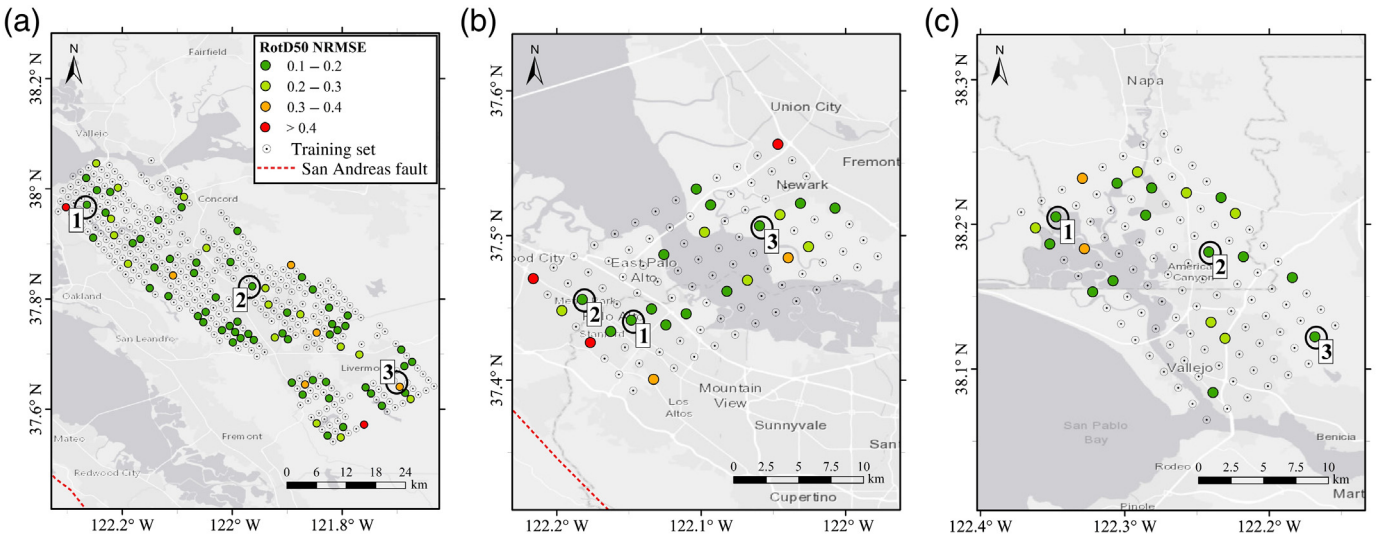

Figure 3. The distribution of the test set's normalized root mean square error (NRMSE) for the 5% damped RotD50 spectrum for the GPR model (a) type 1 in East Bay, (b) type 2 in Palo Alto, and (c) type 2 in South Napa study regions. The color version of this figure is available only in the electronic edition.

time series at most of the target sites reasonably well. The estimation can be less accurate for sites at the boundaries of the network (as shown in Fig. 3), where there is a less uniform distribution of observations. By comparing the results of Figure 3b,c, it is apparent that the estimation accuracy for the sites far away from the causative fault might be higher than for those close to the fault (also shown in Table 4). This could be due to the use of an isotropic covariance function, which allocates a uniform correlation to the surrounding locations based on the separation distance. The GPR model predictions can be improved by employing an anisotropic covariance structure, which uses different normalizing factors for each attribute to compute the separation distance (Rasmussen and Williams, 2006) for regions closer to the fault. Figure 3b indicates that the test sites with less accurate ground-motion estimation (higher RotD50 NRMSE) within the Palo Alto region are mainly restricted to the edge stations, yet the trained GPR model is able to predict ground motions for sites close to the fault appropriately. As evidenced in Figure 4, the GPR model type 1 is capable of estimating the entire ground-motion time series properly for the structural period ranges pertinent to most earthquake engineering applications.

Figure 5a,b show that the GPR model is able to estimate the long-period pulses along the FN direction due to the directivity effect for the sites far away from the epicenter but close to the fault (sites 1 and 2 in Fig. 3b).

Figures 4–6 show that the RotD50 response spectrum and FAS errors are lower for longer periods, whereas the difference between the estimated and the exact spectra increases for the shorter periods. This might be due to two reasons: first, the short-period motions of the 1906 San Francisco earthquake are constructed stochastically (Aagaard *et al.*, 2008), which results in lower correlations for the short-period content of the neighboring motions. Thus, the GPR estimation for the short-period motion could be less accurate than it is for longer-period motion. Second, the motions are less well-correlated to each other at higher frequencies and longer geographical separation distances because of the smaller wavelengths associated with those frequencies. Therefore, short-period waves of the ground motions may be less accurately synthesized, especially when the neighboring stations are not sufficiently close to each other. This phenomenon is observed in existing "lagged coherency" models in which the lagged coherency between two stations, as a representative of the correlation between the frequency content, drops with increasing frequency and separation distance (e.g., Abrahamson *et al.*, 1991; Liao and Zerva, 2006; Rodda and Basu, 2018).

Figure 7 displays the DFT coefficients' real part $\hat{\theta}$ values for the GPR model type 2 implemented within the Palo Alto and South Napa study regions along the FN and FP directions. As indicated, we incorporated the effects of variations in soil conditions in these regions. Figure 7 demonstrates the $\hat{\theta}$ growth as a function of increasing frequency. A similar observation exists for the imaginary part $\hat{\theta}$ values. It is recognizable from equations (10) and (17) that covariance (and subsequently correlation) among the observed values decreases with increasing $\hat{\theta}$ (equivalently decrease of length-scale). In other words, there is a lower correlation between the higher-frequency content of the ground motions, which is consistent with the established lagged coherency models.

M_w 7.0 Hayward fault scenario earthquake simulated motions

We evaluate the performance of the trained GPR on another simulated earthquake dataset, which was not used during the

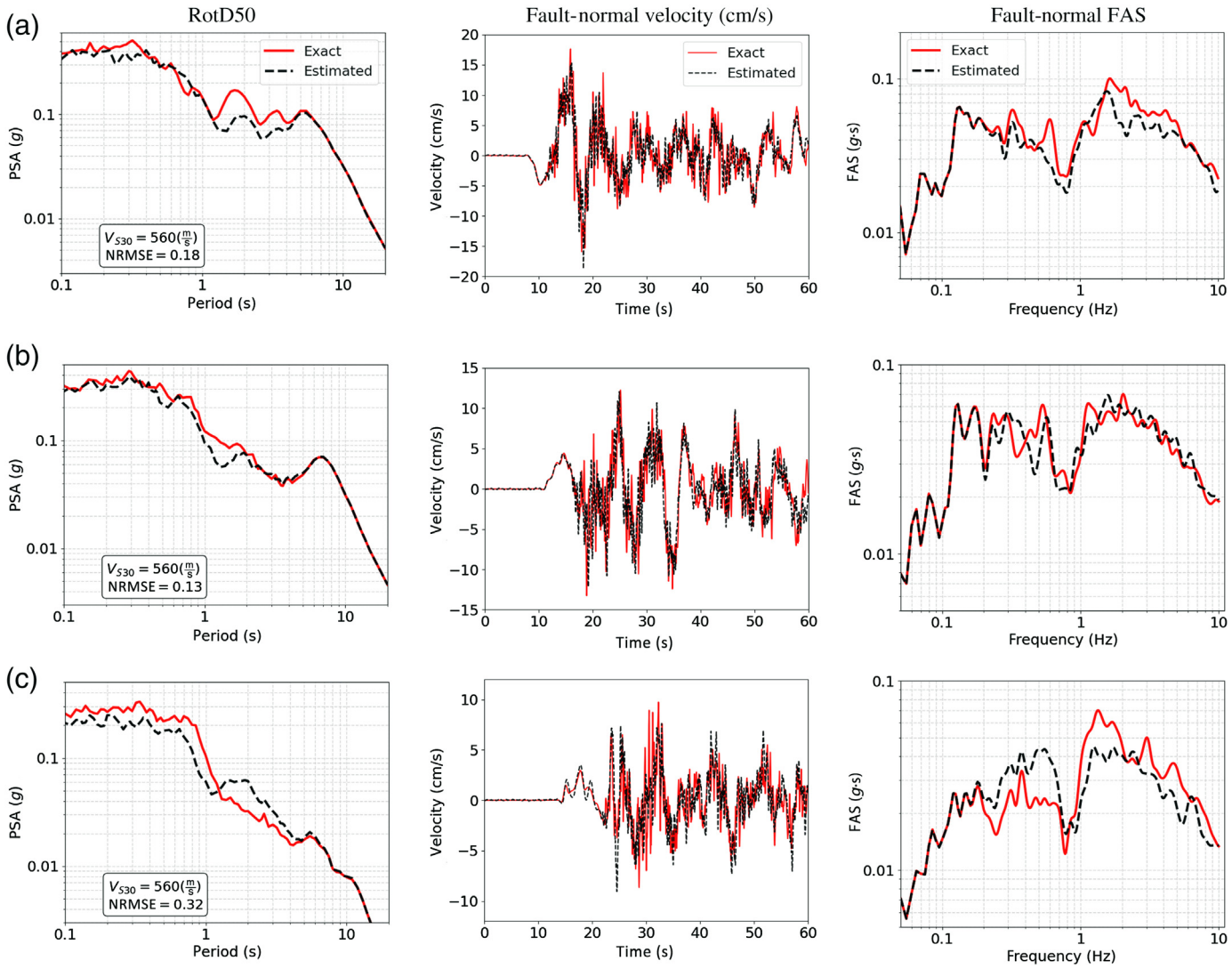

Figure 4. The RotD50, velocity time series, and Fourier amplitude spectrum (FAS) of the predicted and the exact motions along fault-normal direction for the chosen test sites: numbers (a) 1, (b) 2, and (c) 3 within the East Bay study region. The color version of this figure is available only in the electronic edition.

hyperparameter optimization procedure. To do this, the M_w 7.0 Hayward fault scenario earthquake-simulated ground motions (Rodgers *et al.*, 2019) are employed. In the present study, motions for the 3D model ("3DTOPO") are used to evaluate the accuracy of the trained GPR's estimation. The 3DTOPO Earth model has a $V_{S_{\min}} = 500$ m/s; therefore, the simulation results for the sites with $V_S > 500$ m/s are more reliable. The 3D subsurface material properties and the topography (3DTOPO) simulations are obtained based on the USGS model (USGS, 2018). There is a total of 2301 locations within a 120 km × 80 km rectangular domain on a uniform 2 km × 2 km grid for which velocity time series are generated along the FN, FP, and vertical directions.

In this study, 326 locations with fairly uniform site conditions (520 m/s > V_{S30} > 500 m/s) are chosen within the East Bay study region (Fig. 8). Because all sites are located on a fairly uniform site condition, the GPR model type 1 is implemented to estimate the test sites' ground-motion time series.

About 80% of the 326 locations are randomly selected as the training set, whereas the remaining 20% are used as the test set. The observation density of the training set is about 0.25 stations/km^2, which makes the $\hat{\lambda}$ obtained for the GPR model type 1 (with approximately the same observation density in Table 3) usable to estimate the ground-motion time series at target sites. The distribution of the training set and test sites is shown in Figure 8. In addition, the distribution of the NRMSE between the estimated and "exact" (physics-based simulated) motions' RotD50 spectra are shown in Figure 8. Two test sites are shown in Figure 8, for which the prediction results are shown in Figure 9.

Table 5 summarizes the NRMSE of the linear response spectrum in both the FN and FP directions and for RotD50. Table 5 shows that the M_w 7.0 Hayward fault scenario earthquake test

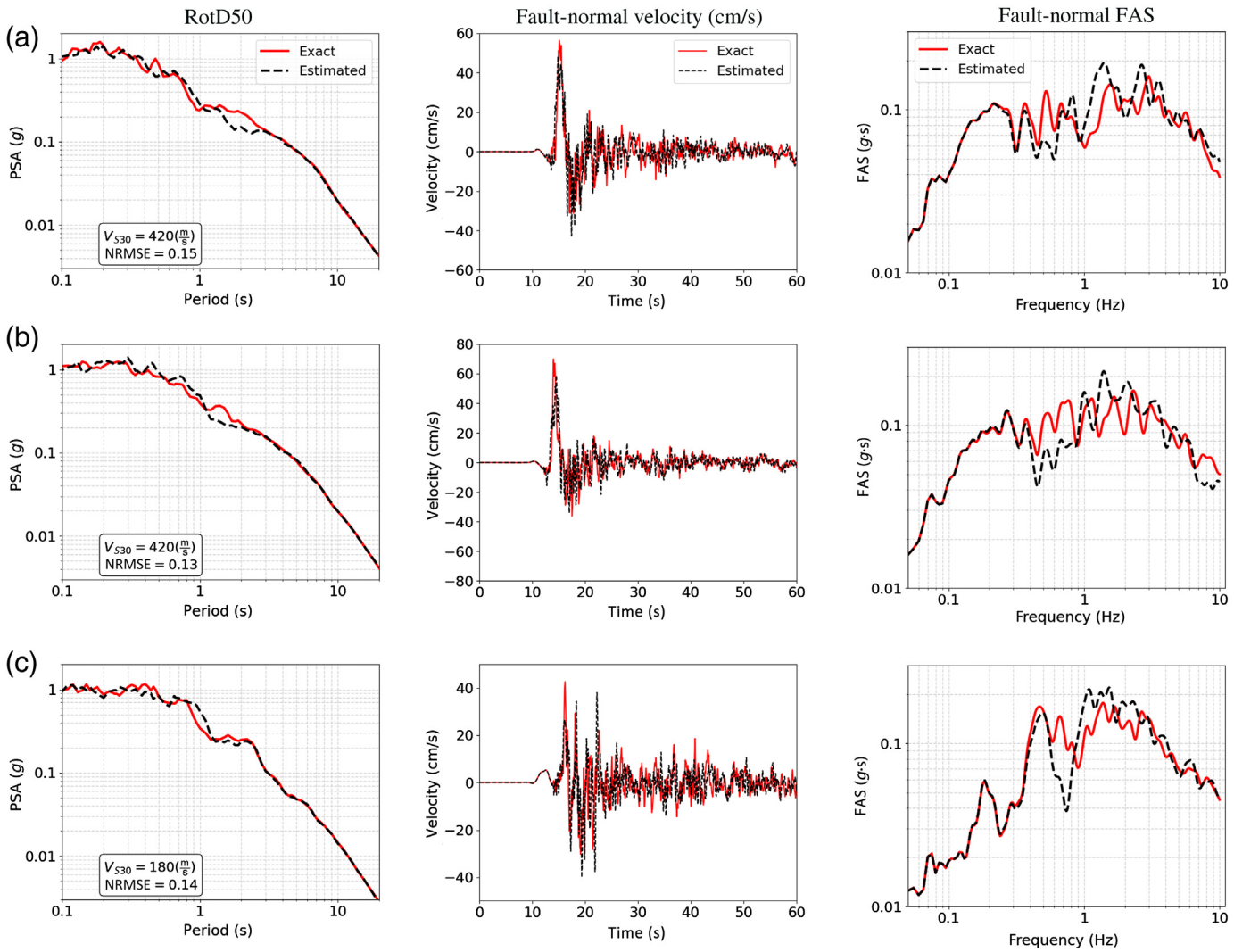

Figure 5. The RotD50, velocity time series, and FAS of the predicted and the exact motions along fault-normal direction for the chosen test sites: numbers (a) 1, (b) 2, and (c) 3 within the Palo Alto study region. The color version of this figure is available only in the electronic edition.

set, which was not used during the hyperparameter optimization, resulted in a higher average NRMSE for the RotD50 spectrum in comparison with the 1906 M_w 7.9 San Francisco simulated motion dataset. In addition, the Hayward fault simulations consider the topography of the region, whereas the 1906 San Francisco simulations were carried out for a horizontal surface. Moreover, the M_w 7.0 Hayward fault simulations used wave propagation for all frequencies, whereas the 1906 San Francisco simulations implemented stochastic noise for frequencies above 1 Hz. Therefore, the Hayward fault dataset includes more complexity, which might lead to a higher estimation error than for the 1906 San Francisco dataset.

Figure 8 demonstrates that the RotD50 spectra of the estimated ground motions have acceptable NRMSE for the majority of the test locations for the M_w 7.0 Hayward fault-simulated motions. Moreover, Figure 8 illustrates the applicability of the trained GPR model in the prediction of ground-motion time series for the sites close to the fault (less than 4 km away) and close to the epicenter (less than 9 km away). For a few locations mainly located at the boundary edges of the simulation network, the estimated ground-motion time series are less accurate. Figure 9a demonstrates that the GPR model appropriately predicted the long-period pulses due to the directivity effect for site 1. Figure 9 illustrates that the ground-motion time-series estimation is more accurate for the long-period motions, whereas the shorter-period shear waves might be less accurately predicted. This feature of our results is consistent with those generated using the 1906 San Francisco earthquake.

2019 M_w 7.1 Ridgecrest earthquake

It is useful to examine the trained GPR model's prediction with an actual recorded earthquake dataset. To do so, we used motions of the 2019 M_w 7.1 Ridgecrest earthquake that were recorded by the CSN within the northern Los Angeles basin as another test set for the GPR model type 2 (Clayton *et al.*, 2020; Kohler *et al.*, 2020; Filippitzis *et al.*, 2021). The M_w 7.1 Ridgecrest earthquake

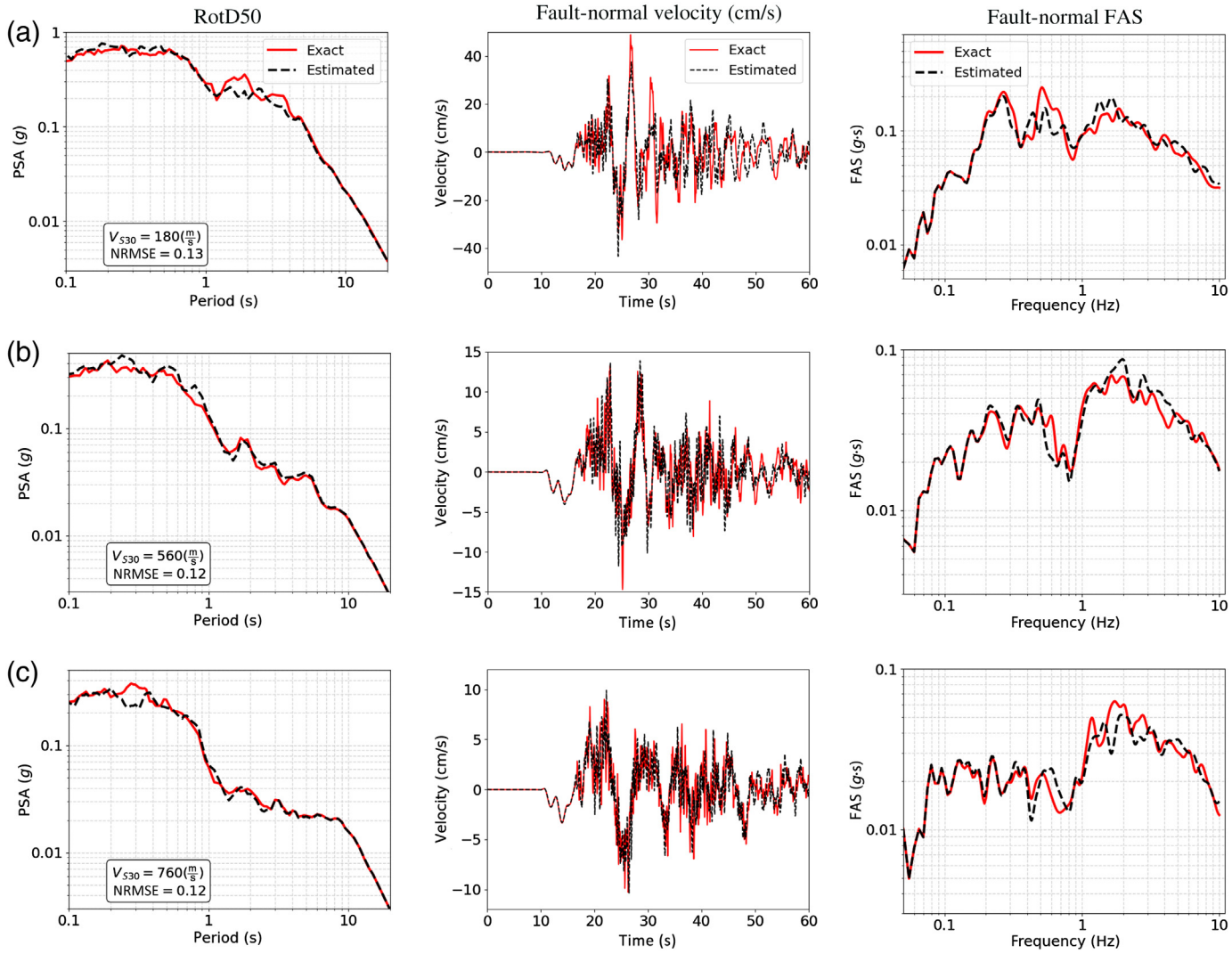

Figure 6. The RotD50, velocity time series, and FAS of the predicted and the exact motions along fault-normal direction for the chosen test sites: numbers (a) 1, (b) 2, and (c) 3 within the South Napa study region. The color version of this figure is available only in the electronic edition.

ground motions were poorly recorded in the epicentral region because there are only about three seismic stations within a 20 km × 20 km area surrounding the epicenter (USGS ShakeMap for 2019 M_w 7.1 Ridgecrest earthquake) (U.S. Geological Survey, 2019). Therefore, an area with an adequate number of recording stations is chosen to evaluate the trained GPR model's performance. In this study, we chose 151 CSN ground-level stations that recorded the 2019 M_w 7.1 Ridgecrest earthquake. The site condition, V_{S30}, of the recording stations is estimated using a proxy-based model as described in Ahdi *et al.* (2020).

One hundred forty nine recording stations are considered as the training set, whereas the remaining two stations are used as the test stations. The observation density for 149 observed sites distributed over a 492 km^2 region is about 0.30 stations/km^2,

TABLE 4
1906 M_w 7.9 San Francisco Test Set's NRMSE for Model Type 1 and Type 2

		FN		FP		RotD50	
Model Type	**Study Region**	**Average**	**Standard Deviation**	**Average**	**Standard Deviation**	**Average**	**Standard Deviation**
Type 1	East Bay	0.23	0.08	0.23	0.08	0.19	0.07
Type 2	Palo Alto	0.34	0.29	0.38	0.38	0.31	0.36
Type 2	South Napa	0.23	0.06	0.26	0.1	0.19	0.05

FN, fault normal; FP, fault parallel.

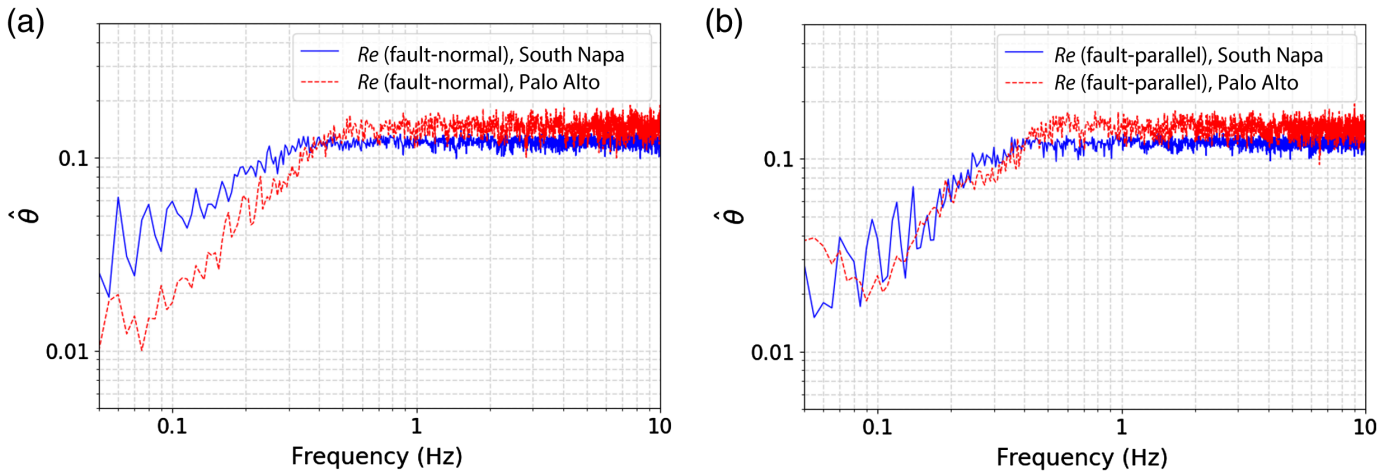

Figure 7. $\hat{\theta}$ for real part (Re) covariance functions along (a) fault-normal and (b) fault-parallel directions within Palo Alto and South Napa study regions. The color version of this figure is available only in the electronic edition.

which makes the $\hat{\lambda}$ obtained for the GPR model type 2 (for 0.29 stations/km^2 density) applicable. Figure 10 shows the distribution of the observed stations (training set) and their site conditions, V_{S30}, as well as the test stations for which the ground-motion time series are estimated. The trained GPR model type 2 is implemented for the conditioned simulation procedure. As indicated, this GPR model is capable of incorporating variations of local soil conditions. Figure 11 displays the prediction results for the two test stations shown in Figure 10. It is noted that the predicted time series are reliable only within the mutually usable frequency bandwidth (Ancheta *et al.*, 2014) among all observed motions, which is the reliable frequency range after the noise removal of the recorded motions. Figure 11 displays the estimated and the exact (recorded) ground motions' RotD50 and FAS within the overlapping usable frequency bandwidth of the observed motions.

Figure 11 shows that the results for ground-motion time-series estimation for the 2019 M_w 7.1 Ridgecrest earthquake sequence are auspicious. The PGV and the long-period pulses of the recorded motions (Filippitzis *et al.*, 2021) are captured fairly accurately at both test stations. This implies that the GPR model type 2 can generate ground motions with acceptable accuracy. We plan to investigate further the applicability of the GPR model type 2 using a broader set of recorded earthquake datasets.

Realizations of ground motion

The trained GPR model provides the posterior mean vector and posterior covariance matrix for the DFT coefficients at each frequency for the target sites based on equations (8) and (9). In this study, there is only one target site to estimate the DFT coefficients at each prediction step. Thus, equations (8) and (9) provide the DFT coefficients' real and imaginary parts' posterior mean and posterior standard deviation at each frequency. In addition, we estimate the correlation between $\mathcal{R}e_k$ and $\mathcal{I}m_k$ at the k^{th} frequency, $k = 0, \ldots, N - 1$, at the target site using the observed ground motions DFT coefficients at the same frequency. Therefore, we can generate

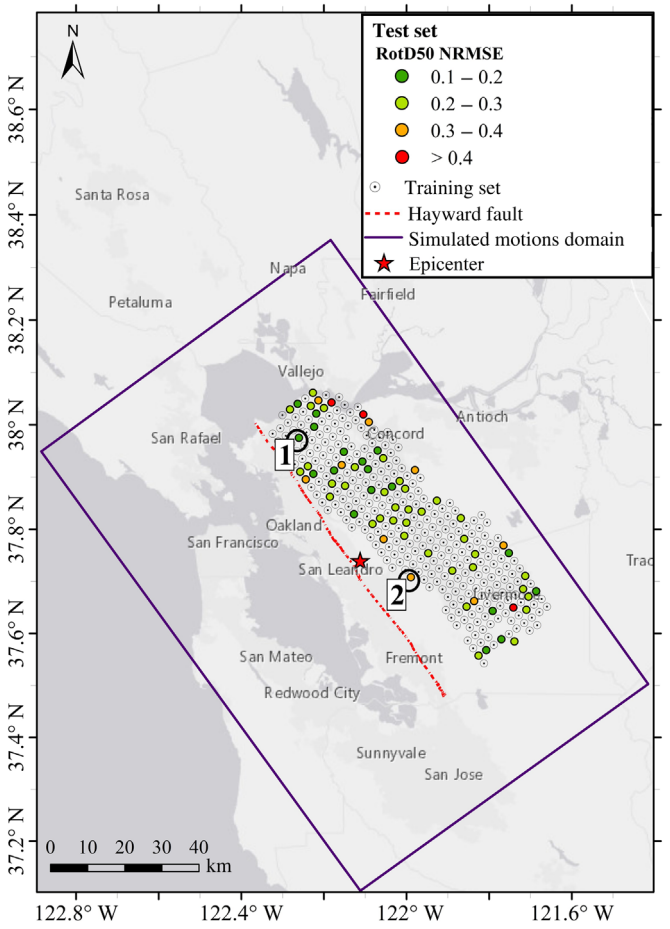

Figure 8. Distribution of the training and test set and the test set's NRMSE for the 5% damped RotD50 spectrum for the M_w 7.0 Hayward fault scenario earthquake simulated motions study region. The color version of this figure is available only in the electronic edition.

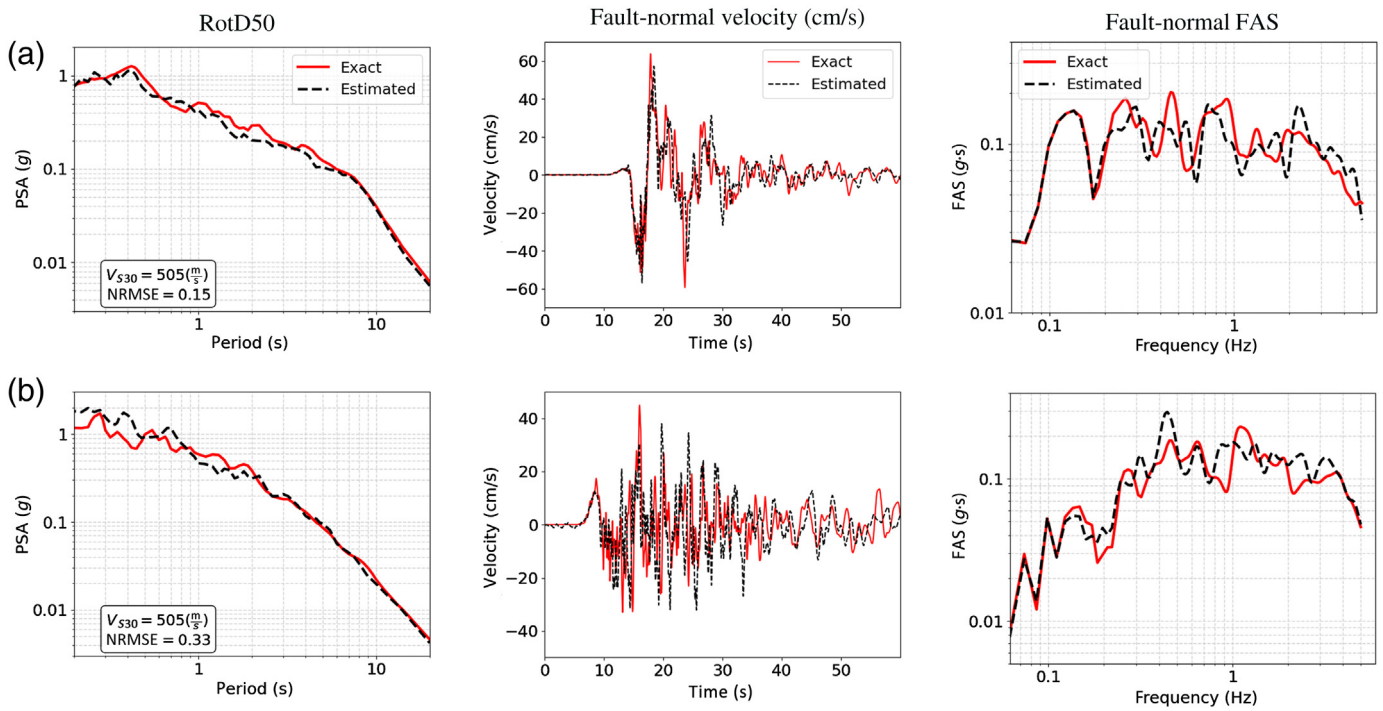

Figure 9. The RotD50, velocity time series, and FAS of the predicted and the exact motions along fault-normal direction for the chosen test sites: numbers (a) 1 and (b) 2 within the M_w 7.0 Hayward fault scenario earthquake simulated motions study region. The color version of this figure is available only in the electronic edition.

pairs of $(\mathcal{R}e_k, \mathcal{I}m_k)$ 2×1 random sample vectors having the 2×1 mean vector and 2×2 covariance matrix of the real and imaginary parts at each frequency. These generated samples can then be converted to samples of amplitude, $|A_k|$ (compare with equation 2). We estimate the logarithmic mean and standard deviation of amplitudes at each frequency using the $|A_k|$ samples. Eventually, we implement the interfrequency correlation model established by Bayless and Abrahamson (2019) to develop the covariance matrix of $\log(|A_k|)$ for all frequencies, $k = 0, \ldots, N-1$. We generate 150 multivariate Gaussian random samples of FAS using the established $N \times 1$ mean vector and $N \times N$ covariance matrix. These FAS samples are then combined with the phase spectrum constructed with mean estimated real and imaginary parts at the target site to generate 150 random ground-motion realizations.

Figure 12a depicts 150 random generated ground-motion realizations' 5% damped response spectra as well as the logarithmic mean of those samples' response spectra along the east–west direction for the test station 1 shown in Figure 10. In addition, Figure 12a demonstrates that the estimated ground-motion time series using mean DFT coefficients has a similar response spectrum (solid black line in Fig. 12a) to the logarithmic mean of the realizations' response spectra. Figure 12b shows the 68% confidence interval (mean ± standard deviation) of the ground motions' realizations on a logarithmic scale. It is observable in Figure 12b that the estimated ground motion's response spectrum has higher uncertainty at shorter periods, whereas this uncertainty decreases at longer periods. As evidenced in Figure 12b, the recorded motion's response spectrum is located within the 68% confidence interval at most periods within the usable frequency bandwidth.

CONCLUSION AND DISCUSSION

A novel approach to estimate the entire ground-motion time series at a target location using the observed surrounding motions was developed. The generated motions at target

TABLE 5
M_w 7.0 Hayward Fault Scenario Earthquake Simulated Motions Test Set's NRMSE for Model Type 1 and Type 2

		FN		FP		RotD50	
Model Type	**Study Region**	**Average**	**Standard Deviation**	**Average**	**Standard Deviation**	**Average**	**Standard Deviation**
Type 1	M_w 7.0 Hayward Fault	0.28	0.07	0.31	0.11	0.25	0.07

FN, fault normal; FP, fault parallel.

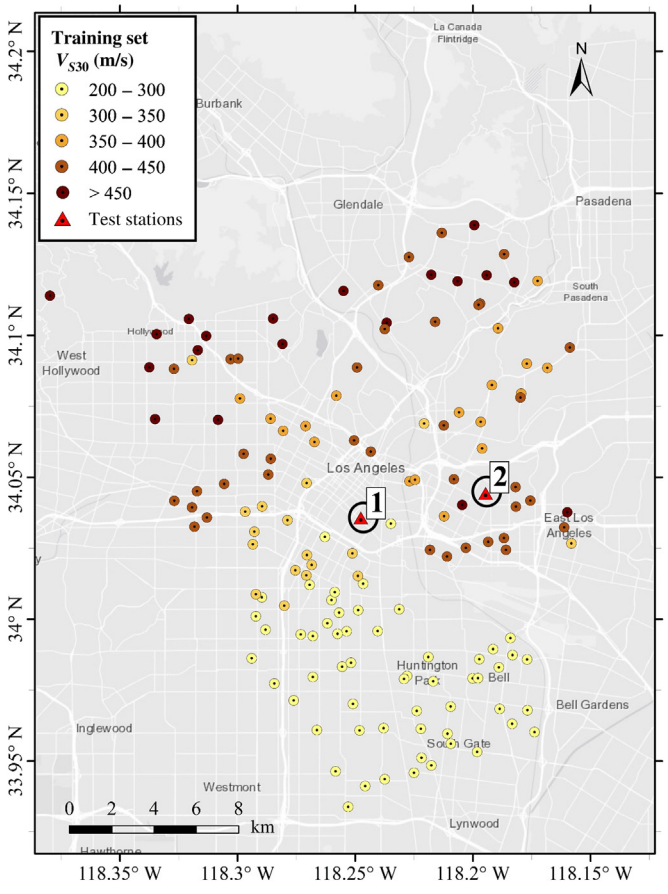

Figure 10. Distribution of the training set and test stations for the Community Seismic Network network. The color version of this figure is available only in the electronic edition.

(uninstrumented) sites can be used for site-specific nonlinear structural analysis as well as quantification of spatial damage distribution.

In this study, GPR was employed to estimate time series at the target sites through interpolating the real and imaginary parts of the DFT coefficients. To do so, the GPR model's hyperparameter, λ, was tuned using ground motions of the physics-based simulated 1906 San Francisco earthquake. Two GPR models were developed: one applicable to the homogeneous regions (relatively uniform site condition) and the other usable in regions with considerable local site condition variation. The optimized $\hat{\lambda}$ for these models are applicable for the regions with an approximately similar observation density. Both models demonstrated acceptable performance for estimation of the ground motion, as well as the response spectra for the 1906 M_w 7.9 San Francisco and M_w 7.0 Hayward fault-simulated ground motions. In addition, our investigation demonstrated the applicability of the trained GPR model for estimation of the 2019 M_w 7.1 Ridgecrest earthquake recorded ground motions.

The trained GPR models estimated the long-period pulses properly. The estimation of the motions for locations at the edges of the network, where there is a nonuniform distribution of observations or regions with fewer observations, may not be as accurate as those at other locations. In addition, the length-scale parameter of the covariance functions demonstrated that there is a higher correlation for the long-period content of the ground motions compared with the short-period content within a region. Therefore, the conditioned simulated ground motions are generally more reliable in the long-period range than those at short periods. In addition, we incorporated the posterior mean and standard deviation of the DFT coefficients as well as the interfrequency correlations among neighboring frequencies to generate random realizations of ground motions at the target site. The ground-motion realizations depicted that the uncertainty of the estimated ground motions is higher for the short periods.

The GPR models can be expanded by considering other site attributes such as $Z_{1.0}$, $Z_{2.5}$, and R_{JB} as well as combining the covariance functions. In addition, using an anisotropic covariance function, especially for regions closer to the fault, may improve the estimation.

DATA AND RESOURCES

The 1906 M_w 7.9 San Francisco earthquake simulated ground motions were provided by Robert W. Graves (Aagaard *et al.*, 2008). The RotD50 and orthogonal directions linear response spectra of the ground motions were constructed using the R package for computation of earthquake ground-motion response spectra (Wang *et al.*, 2017), which is accessible through https://peer.berkeley.edu/peer-reports. The M_w 7.0 Hayward fault scenario earthquake simulated motions (Rodgers *et al.*, 2019) were provided by Arthur J. Rodgers. The M_w 7.1 2019 Ridgecrest earthquake data recorded by the Community Seismic Network (CSN) were obtained from http://csn.caltech.edu/data/. The processed recorded motions for the 2019 M_w 7.1 Ridgecrest earthquake can be retrieved from https://www.risksciences.ucla.edu/nhr3/gmdata. The average shear-wave velocity values, V_{S30}, at each CSN station were provided by Pengfei Wang using the proxy-based model (Ahdi *et al.*, 2020). All websites were last accessed in February 2021.

DECLARATION OF COMPETING INTERESTS

The authors acknowledge that there are no conflicts of interest recorded.

ACKNOWLEDGMENTS

This study was partially supported by the University of California, Los Angeles (UCLA) Graduate Fellowship to the first author, which is gratefully acknowledged. Partial supports of the National Science Foundation (Award Number 2025310), California Department of Transportation and Pacific Gas & Electric Company are also fully appreciated. Arthur Rodgers' work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the authors and do not necessarily reflect those of the

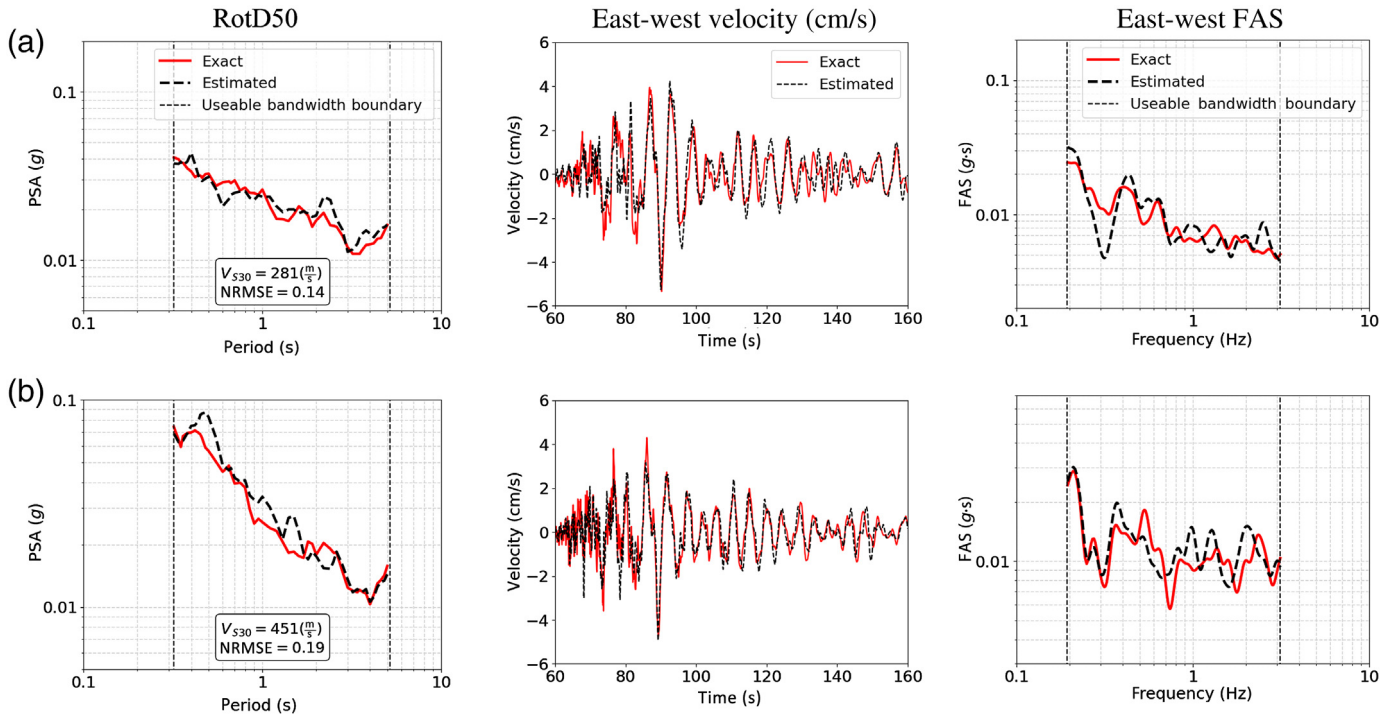

Figure 11. The RotD50, velocity time series, and FAS of the predicted and the exact motions along east–west direction for the chosen test stations: numbers (a) 1 and (b) 2 within the CSN network that recorded the 2019 M_w 7.1 Ridgecrest earthquake. The color version of this figure is available only in the electronic edition.

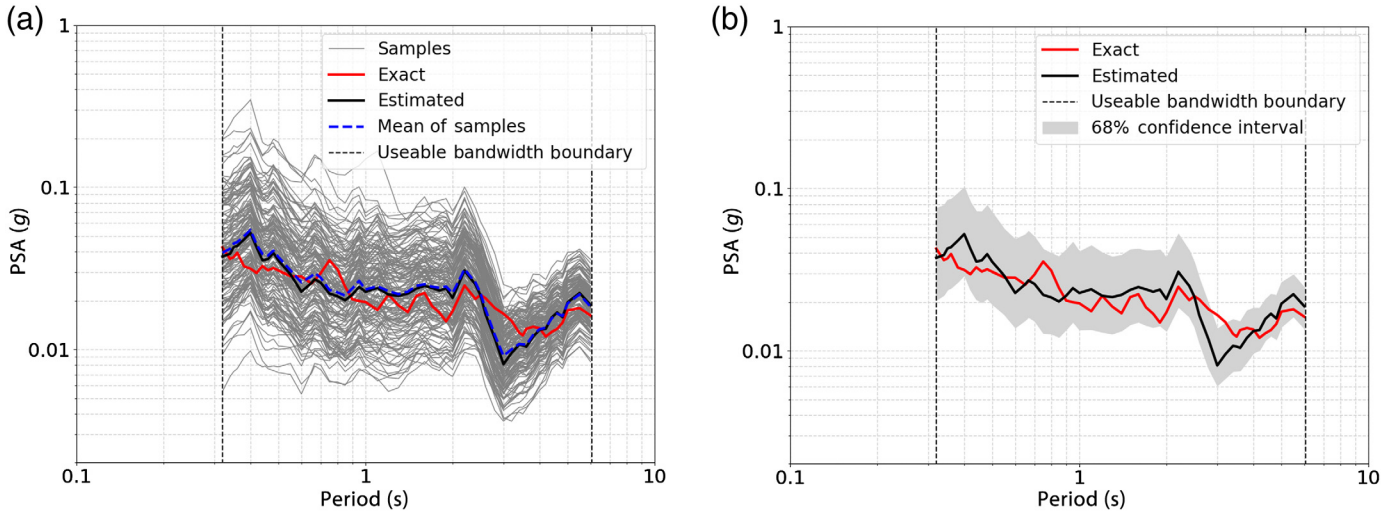

Figure 12. The 5% damped pseudospectral acceleration (PSA) along east–west direction at test station 1 within the CSN network that recorded the 2019 M_w 7.1 Ridgecrest earthquake for (a) 150 random ground-motion realizations and (b) 68% confidence interval. The color version of this figure is available only in the electronic edition.

supporting agencies. The authors would like to also thank Robert Graves for providing the simulated ground-motion dataset of the 1906 event and Tadahiro Kishida for his efforts in organizing and processing it. Sean Ahdi and Pengfei Wang have kindly assisted in the estimation of V_{S30} values at the recording stations of the 2019 Ridgecrest earthquake. The authors also benefitted from constructive discussions with Silvia Mazzoni. The comments from two *BSSA* anonymous reviewers are greatly appreciated.

REFERENCES

Aagaard, B. T., T. M. Brocher, D. Dolenc, D. Dreger, R. W. Graves, S. Harmsen, S. Hartzell, S. Larsen, K. McCandless, S. Nilsson, *et al.* (2008). Ground-motion modeling of the 1906 San Francisco earthquake, part II: Ground-motion estimates for the 1906 earthquake and scenario events, *Bull. Seismol. Soc. Am.* **98,** no. 2, 1012–1046.

Abrahamson, N. A., J. F. Schneider, and J. C. Stepp (1991). Empirical spatial coherency functions for application to soil–structure interaction analyses, *Earthq. Spectra* **7,** no. 1, 1–27.

Abramowitz, M., and I. A. Stegun (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards, Applied Mathematics Series, Vol. 55, U.S. Government Printing Office and Tenth Printing, Washington, D.C., XIII pp.

Adanur, S., A. C. Altunisik, K. Soyluk, A. A. Dumanoglu, and A. Bayraktar (2016). Contribution of local site-effect on the seismic response of suspension bridges to spatially varying ground motions, *Earthq. Struct.* **10,** no. 5, 1233–1251.

Ahdi, S. K., S. Mazzoni, T. Kishida, P. Wang, C. C. Nweke, N. M. Kuehn, V. Contreras, B. Rowshandel, J. P. Stewart, and Y. Bozorgnia (2020). Engineering characteristics of ground motions recorded in the 2019 Ridgecrest earthquake sequence, *Bull. Seismol. Soc. Am.* **110,** no. 4, 1474–1494.

Alimoradi, A., and J. L. Beck (2015). Machine-learning methods for earthquake ground motion analysis and simulation, *J. Eng. Mech.* **141,** no. 4, 04014147, doi: 10.1061/(ASCE)EM.1943-7889.0000869.

Ancheta, T. D., R. B. Darragh, J. P. Stewart, E. Seyhan, W. J. Silva, B. S.-J. Chiou, K. E. Wooddell, R. W. Graves, A. R. Kottke, D. M. Boore, *et al.* (2014). NGA-West2 database, *Earthq. Spectra* **30,** no. 3, 989–1005.

Baker, J. W., and Y. Chen (2020). Ground motion spatial correlation fitting methods and estimation uncertainty, *Earthq. Eng. Struct. Dynam.* **49,** no. 15, 1662–1681.

Bayless, J., and N. A. Abrahamson (2019). An empirical model for the interfrequency correlation of epsilon for Fourier amplitude spectra, *Bull. Seismol. Soc. Am.* **109,** no. 3, 1058–1070.

Boore, D. M. (2010). Orientation-independent, nongeometric-mean measures of seismic intensity from two horizontal components of motion, *Bull. Seismol. Soc. Am.* **100,** no. 4, 1830–1835.

Chen, Y., and J. W. Baker (2019). Spatial correlations in CyberShake physics-based ground-motion simulations, *Bull. Seismol. Soc. Am.* **109,** no. 6, 2447–2458.

Clayton, R. W., M. Kohler, R. Guy, J. Bunn, T. Heaton, and M. Chandy (2020). CSN-LAUSD network: A dense accelerometer network in Los Angeles Schools, *Seismol. Res. Lett.* **91,** no. 2A, 622–630.

Der Kiureghian, A. (1996). A coherency model for spatially varying ground motions, *Earthq. Eng. Struct. Dynam.* **25,** no. 1, 99–111.

Fan, J., and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Am. Stat. Assoc.* **96,** no. 456, 1348–1360.

Filippitzis, F., M. D. Kohler, T. H. Heaton, R. W. Graves, R. W. Clayton, R. G. Guy, J. J. Bunn, and K. M. Chandy (2021). Ground motions in urban Los Angeles from the 2019 Ridgecrest earthquake sequence, *Earthq. Spectra* doi: 10.1177/87552930211003916.

Fraser, W. A., D. J. Wald, and K.-W. Lin (2008). Using ShakeMap and ShakeCast to prioritize post-earthquake dam inspections, in *Geotechnical Earthquake Engineering and Soil Dynamics IV*, 1–10.

Gentile, R., and C. Galasso (2020). Gaussian process regression for seismic fragility assessment of building portfolios, *Struct. Saf.* **87,** 101980, doi: 10.1016/j.strusafe.2020.101980.

Ghaderi, A., V. Morovati, and R. Dargazany (2020). A Bayesian surrogate constitutive model to estimate failure probability of rubber-like materials, available at https://arxiv.org/pdf/2010.13241.pdf (last accessed February 2021).

Hjort, N. L., C. Holmes, P. Müller, and S. G. Walker (Editors) (2010). *Bayesian Nonparametrics*, Vol. 28, Cambridge University Press, Cambridge, United Kingdom, 22–25.

Huang, C., Y. Liang, X. Ding, and C. Fang (2014). Generalized joint kernel regression and adaptive dictionary learning for single-image super-resolution, *Sig. Process.* **103,** 142–154.

Huang, D., and G. Wang (2017). Energy-compatible and spectrum-compatible (ECSC) ground motion simulation using wavelet packets, *Earthq. Eng. Struct. Dynam.* **46,** no. 11, 1855–1873.

Jayaram, N., and J. W. Baker (2009). Correlation model for spatially distributed ground-motion intensities, *Earthq. Eng. Struct. Dynam.* **38,** no. 15, 1687–1708.

Kameda, H., and H. Morikawa (1992). An interpolating stochastic process for simulation of conditional random fields, *Probab. Eng. Mech.* **7,** no. 4, 243–254.

Kohler, M. D., F. Filippitzis, T. Heaton, R. W. Clayton, R. Guy, J. Bunn, and K. M. Chandy (2020). 2019 Ridgecrest earthquake reveals areas of Los Angeles that amplify shaking of high-rises, *Seismol. Res. Lett.* **91,** no. 6, 3370–3380.

Konakli, K., and A. Der Kiureghian (2012). Simulation of spatially varying ground motions including incoherence, wave-passage and differential site-response effects, *Earthq. Eng. Struct. Dynam.* **41,** no. 3, 495–513.

Kuehn, N. M., and N. A. Abrahamson (2020). Spatial correlations of ground motion for non-ergodic seismic hazard analysis, *Earthq. Eng. Struct. Dynam.* **49,** no. 1, 4–23.

Landwehr, N., N. M. Kuehn, T. Scheffer, and N. Abrahamson (2016). A nonergodic ground-motion model for California with spatially varying coefficients, *Bull. Seismol. Soc. Am.* **106,** no. 6, 2574–2583.

Li, R., and A. Sudjianto (2005). Analysis of computer experiments using penalized likelihood in Gaussian Kriging models, *Technometrics* **47,** no. 2, 111–120.

Liao, S., and A. Zerva (2006). Physically compliant, conditionally simulated spatially variable seismic ground motions for performance-based design, *Earthq. Eng. Struct. Dynam.* **35,** no. 7, 891–919.

Lin, K., D. J. Wald, C. A. Kircher, D. Slosky, K. Jaiswal, and N. Luco (2018). USGS shakecast system advancements, *11th National Conf. on Earthquake Engineering*, 3458–3468.

Lu, X., Q. Cheng, Y. Tian, and Y. Huang (2021). Regional ground-motion simulation using recorded ground motions, *Bull. Seismol. Soc. Am.* **111,** no. 2, 825–838.

Nadaraya, E. A. (1964). On estimating regression, *Theor. Probab. Appl.* **9,** no. 1, 141–142.

Oppenheim, A. V., A. S. Willsky, and S. H. Nawab (1997). *Signals and Systems*, Prentice Hall Inc., Upper Saddle River, New Jersey, 7458.

Otake, R., J. Kurima, H. Goto, and S. Sawada (2020). Deep learning model for spatial interpolation of real-time seismic intensity, *Seismol. Soc. Am.* **91,** no. 6, 3433–3443.

Petrone, F., N. Abrahamson, D. McCallen, and M. Miah (2020). Validation of (not-historical) large-event near-fault ground-motion

simulations for use in civil engineering applications, *Earthq. Eng. Struct. Dynam.* **50,** no. 1, 116–134.

Rasmussen, C. E., and C. K. I. Williams (2006). *Gaussian Processes for Machine Learning*, The MIT Press, Cambridge, Massachusetts.

Rodda, G. K., and D. Basu (2018). Spatial variation and conditional simulation of seismic ground motion, *Bull. Earthq. Eng.* **16,** no. 10, 4399–4426.

Rodda, G. K., and D. Basu (2019). On conditional simulation of spatially varying rotational ground motion, *J. Earthq. Eng.* **25,** no. 6, 1191–1226.

Rodgers, A. J., N. A. Petersson, A. Pitarka, D. B. McCallen, B. Sjogreen, and N. Abrahamson (2019). 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, *Seismol. Res. Lett.* **90,** no. 3, 1268–1284.

Rubin, D. B. (1981). The Bayesian bootstrap, in *The Annals of Statistics*, 130–134.

Sajedi, S. O., and X. Liang (2020). A data-driven framework for near real-time and robust damage diagnosis of building structures, *Struct. Contr. Health Monit.* **27,** no. 3, e2488, doi: 10.1002/stc.2488.

Savitzky, A., and M. J. E. Golay (1964). Smoothing and differentiation of data by simplified least squares procedures, *Anal. Chem.* **36,** no. 8, 1627–1639.

Sheibani, M., and G. Ou (2020). The development of Gaussian process regression for effective regional post-earthquake building damage inference, *Comput. Aided Civ. Infrastruct. Eng.* **36,** no. 3, 264–288.

Southern California Earthquake Data Center (2021). [Online], available at https://service.scedc.caltech.edu/SCSNStationMap/station.html (last accessed February 2021).

Sun, H., H. Burton, Y. Zhang, and J. Wallace (2018). Interbuilding interpolation of peak seismic response using spatially correlated demand parameters, *Earthq. Eng. Struct. Dynam.* **47,** no. 5, 1148–1168.

Tamhidi, A., N. Kuehn, Y. Bozorgnia, E. Taciroglu, and T. Kishida (2019). Prediction of ground-motion time-series at an arbitrary location using Gaussian process interpolation: Application to the Ridgecrest earthquake, *Poster Presentation at the 2019 SCEC Annual Meeting*, Palm Springs, California, 7–11 September.

Tamhidi, A., N. M. Kuehn, M. D. Kohler, F. Ghahari, E. Taciroglu, and Y. Bozorgnia (2020). Ground-motion time-series interpolation within the community seismic network using Gaussian process regression: Application to the 2019 Ridgecrest earthquake, *Poster Presentation at the 2020 SCEC Annual Meeting*, 14–17 September.

Tian, L., X. Gai, B. Qu, H. Li, and P. Zhang (2016). Influence of spatial variation of ground motions on dynamic responses of supporting towers of overhead electricity transmission systems: An experimental study, *Eng. Struct.* **128,** 67–81.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso, *J. Roy. Stat. Soc. B* **58,** no. 1, 267–288.

Todorovska, M. I., H. Ding, and M. D. Trifunac (2017). Coherency of synthetic earthquake ground motion for the design of long structures: Effect of site conditions, in *International Collaboration in Lifeline Earthquake Engineering 2016*, American Society of Civil Engineers, Reston, Virginia, 427–434.

U.S. Geological Survey (2018). 3-D geologic and seismic velocity models of the San Francisco Bay region, available at https://www.usgs.gov/natural-hazards/earthquake-hazards/science/3-d-geologic-and-seismic-velocity-models-san-francisco (last accessed February 2021).

U.S. Geological Survey (2019). M7.1-2019 Ridgecrest earthquake sequence, available at https://earthquake.usgs.gov/earthquakes/eventpage/ci38457511/executive (last accessed May 2021).

Wald, D., K.-W. Lin, K. Porter, and L. Turner (2008). ShakeCast: Automating and improving the use of ShakeMap for post-earthquake decision-making and response, *Earthq. Spectra* **24,** no. 2, 533–553.

Wald, D. J., V. Quitoriano, C. B. Worden, M. Hopper, and J. W. Dewey (2012). USGS "Did You Feel It?" internet-based macroseismic intensity maps, *Ann. Geophys.* **54,** no. 6, doi: 10.4401/ag-5354.

Wang, P., J. P. Stewart, Y. Bozorgnia, D. M. Boore, and T. Kishida (2017). *R package for computation of earthquake ground motion response spectra*, No. 2017/09. Report.

Watson, G. S. (1964). Smooth regression analysis, *Sankhyā: Indian J. Stat. Series A* **26,** no. 4, 359–372.

Worden, C. B., E. M. Thompson, J. W. Baker, B. A. Bradley, N. Luco, and D. J. Wald (2018). Spatial and spectral interpolation of ground-motion intensity measure observations, *Bull. Seismol. Soc. Am.* **108,** no. 2, 866–875.

Wu, Y., Y. Gao, N. Zhang, and D. Li (2016). Simulation of spatially varying ground motions in V-shaped symmetric canyons, *J. Earthq. Eng.* **20,** no. 6, 992–1010.

Zentner, I. (2013). Simulation of non-stationary conditional ground motion fields in the time domain, *Georisk* **7,** no. 1, 37–48.

Zerva, A. (2009). *Spatial Variation of Seismic Ground Motions: Modeling and Engineering Applications*, CRC Press, Boca Raton, Florida, 9–64.

Zerva, A., and V. Zervas (2002). Spatial variation of seismic ground motions: An overview, *Appl. Mech. Rev.* **55,** no. 3, 271–297.

Zerva, A., M. R. Falamarz-Sheikhabadi, and M. K. Poul (2018). Issues with the use of spatially variable seismic ground motions in engineering applications, *European Conf. on Earthquake Engineering*, Thessaloniki, Greece, 225–252.

Manuscript received 21 February 2021
Published online 14 September 2021