Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Ground motion spatial correlation fitting methods and estimation uncertainty

Jack W. Baker* | Yilin Chen

[1]Department of Civil & Environmental Engineering, Stanford University, California, USA

Correspondence
*Jack W. Baker Email: bakerjw@stanford.edu

Summary

Ground shaking intensity varies spatially in earthquakes, and many studies have estimated correlations of intensity from past earthquake data. This paper presents a framework for quantifying uncertainty in the estimation of correlations, and true variability in correlations from earthquake to earthquake. A procedure for evaluating estimation uncertainty is proposed and used to evaluate several methods that have been used in past studies to estimate correlations. The results indicate that a weighted-least-squares algorithm is most effective in estimating spatial correlation models, and that earthquakes with at least 100 recordings are needed to produce informative earthquake-specific estimates of spatial correlations. The proposed procedure is also used to distinguish between estimation uncertainty and the true variability in model parameters that exist in a given data set. The estimation uncertainty is seen to vary between well-recorded and poorly recorded earthquakes, while the true variability is more stable.

KEYWORDS:
spatial correlation, spectral accelerations, infrastructure systems, event-to-event variability

1 | INTRODUCTION

Spatial correlations in ground motion intensities have been studied for nearly two decades [e.g., 1,2,3,4,5,6,7]. These studies calibrate models using statistical analysis of recorded data from past earthquakes, in a manner similar to ground motion prediction models. The resulting models are important for estimating risks to distributed systems such as portfolios of insured properties and distributed infrastructure systems[e.g., 8,9,10,11,12].

As our library of recorded ground motions grows over time, spatial correlation studies have grown in refinement, with several studies exploring factors that may cause spatial correlation to vary from one earthquake to another. Goda and Hong[13] report differences in correlations between California and Taiwan ground motions, but no effect of earthquake magnitude. Jayaram and Baker[14] and Sokolov et al.[15] speculate that soil condition heterogeneity may influence spatial correlation. Goda[5] and Heresi and Miranda[16] report that spatial correlations vary from individual earthquake to earthquake, but find no earthquake characteristic that clearly predicts this variation. Schiappapietra and Douglas[17] report high variability in correlations amongst a sequence of earthquakes in Central Italy, and list local site effects or path and azimuthal effects as possible causes, noting Sokolov et al.'s[15] similar speculation. Other studies note a possible trend with earthquake magnitude[18,19] or variation regionally[20]. Other studies group data from multiple earthquakes into a single data set, making the assumption that correlations from the earthquakes are equivalent[4,7,6,21,22].

Several issues make it difficult to definitively identify important factors. First, because the spatial correlation estimation is empirical and requires many observed ground motions from an earthquake, it is difficult to obtain sufficient data under the conditions of interest. Second, there are no closed-form results from spatial correlation estimation that allow for a quantitative assessment of the uncertainty in a given estimate. Bootstrap estimation is another popular technique to quantify estimation uncertainty. However, it is difficult to apply to spatial data because of challenges with maintaining spatial dependence structure in the replicates while avoiding resampling the same location within a replicate (which provides no information about spatial structure). Some studies have proposed bootstrap techniques that resample from an estimated nonparametric distribution[23] or resampling transformed data[24], but the methods are somewhat complex and have not been adopted widely. For the above reasons, no results have been presented in previous studies to quantify the estimation uncertainty in spatial correlations computed from individual earthquakes.

Another issue that has not been evaluated in this literature is the role of the method used to fit parameters for the models, and relative performance of alternative methods. Fitting methods used in prior ground motion studies include manual visual fitting[14], least squares regression on transformed data[16,7], and least squares regression with weighting according to distance or number of data[25]. One general study evaluated several fitting methods using Monte Carlo simulation of spatial data with a known correlation structure[26], and found that the above fitting methods produce systematic differences in results. But that study considered a small number of observed data (16 or 36 stations) on a regular grid–a situation very different than ground motion data coming from greater numbers of irregularly spaced stations. A recent study examined this issue for realistic numbers of stations in earthquake ground motion studies, but the locations in that study were randomly simulated[20]. There are no general statistical results for estimators under arbitrary station configurations[27].

Exact estimation of a correlation model from data (i.e., consistency in statistical estimation language) requires a large number of observations at closely spaced distances, with the dense locations not too concentrated at a single location[28]. The above-cited studies of ground motions consider well-recorded earthquakes, but none systematically study the impact of well-recorded versus more-poorly-recorded earthquakes on resulting spatial correlation estimates. Further, the above-cited studies use different methods to fit models to data, and it is unknown what impact those fitting methods have on estimation uncertainty. The baseline estimation uncertainty is important as it establishes a threshold at which variability in observed correlations can be credibly linked to some causal source rather than being due to estimation variability.

Given the potential variation in correlations between earthquakes and a lack of consensus predictive physical mechanism, several authors have suggested that correlation model parameters should be considered uncertain in risk analysis calculations forecasting the impacts of future events[15,16,17]. While uncertainties in correlations may be relevant in some cases, care is needed to distinguish *true variation* in correlations among earthquakes from our *measurement errors* caused by having small samples of recordings; only the former is relevant to risk analysis.

To address the above issues, this paper proposes a framework to quantify uncertainty in spatial correlation models, and uses the framework to evaluate estimation uncertainty associated with individual earthquakes and model fitting methods. Section 2 introduces the basic framework for characterizing ground motion amplitudes using ground motion models, and introduces the semivariogram as a tool for quantifying spatial correlations. Methods for fitting semivariogram models are also introduced. Section3 then introduces a model to describe the various components of apparent uncertainty in semivariogram parameters. A method is proposed for quantifying estimation uncertainty, by synthetically simulating ground motion amplitudes with a known spatial correlation model but observed only at locations corresponding to those of past earthquakes. This method is then applied to the considered earthquakes and fitting methods. The results are then discussed, along with the limitations and broader implications of the work.

2 | MODELS FOR GROUND MOTION AMPLITUDE AND SPATIAL CORRELATION

Models for ground motion amplitude correlation utilize the typical ground motion model (GMM) formulation. This formulation is written in the following equations, and written for two sites to illustrate how differences in amplitude at the two sites are considered. A GMM predicts a ground motion intensity measure (IM) from earthquake rupture i at site j (here $j = 1$ and 2, indicated by subscripts) as a function of rupture and site properties.

$$\ln IM_{i,1} = \mu_{\ln IM}(rup_i, site_1) + \delta B_i + \delta W_{i,1} \quad (1)$$

$$\ln IM_{i,2} = \mu_{\ln IM}(rup_i, site_2) + \delta B_i + \delta W_{i,2} \quad (2)$$

where $\mu_{\ln IM}()$ is the mean predicted $\ln IM$ value, as a function of rupture (*rup*) parameters and site (*site*) parameters. The predictor parameters depend upon the particular GMM, but rupture parameters typically include earthquake magnitude and rupture mechanism. Site parameters typically include source-to-site distance, and a metric for near-surface geology, among others. Rupture parameters depend only upon the rupture, and are fixed for all locations given that rupture. The site parameters vary by location, and so are indexed by the site number in the above equations.

The δB_i and $\delta W_{i,j}$ terms are the between- and within-event residuals, representing deviations between observed and mean predicted $lnIM$ values. These are normally distributed random variables with means of zero and standard deviations denoted τ and ϕ, respectively. These standard deviations are also sometimes a function of rupture and site parameters, but this dependence is omitted here for brevity. The between-event residual, δB_i, is common for all sites because it depends upon the rupture and not the specific site.

Standard GMMs provide the function $\mu_{\ln IM}()$, and the standard deviations τ and ϕ. So, for this spatial correlation application, the only uncharacterized portion of the model is the dependence of $\delta W_{i,1}$ and $\delta W_{i,2}$–the model for how within-event residuals vary in space. Since these two parameters are each normally distributed, we make the additional assumption that they are jointly normal[29], so we can fully characterize their dependence with a correlation coefficient (or a semivariance, which will be introduced in Section 2.2).

2.1 | Ground motion data

We consider ground motion IM data from the NGA-West2 database[30], to illustrate the proposed calculations below. The IMs of interest are spectral accelerations at a range of periods ($SA(T)$). We consider the SA_{RotD50} definition of spectral acceleration; this is the median spectral amplitude over all horizontal orientations and is the metric used by the adopted GMM.

We restrict the database to consider only ground motions from earthquakes with moment magnitude (M_w) $\geq$ 4, closest distance to rupture $<$ 300 km, 180 $\leq V_{S30} \leq$ 760 m/s, and a maximum usable period within the period range of interest. We then consider all earthquakes with greater than 40 stations that satisfy the above criteria. Table A1 provides a summary of the resulting data.

For these ground motion data, we use the Chiou and Youngs GMM[31] to compute residuals, and mixed-effects regression to estimate within-event residuals[32]. Because the GMM already provides the standard deviation of the residuals, we divided the residuals from each earthquake by their sample standard deviation and worked with these standardized residuals for the calculations below. This standardization simplifies the model fitting below because it ensures a known sample standard deviation.

Figure 1 shows within-event residuals from two example earthquakes. The two events have differing numbers of recordings (290 for El Mayor-Cucapah and 118 for Yorba Linda), which will be important later. The areas of similarly-colored symbols in both maps (typically with separation distances of less than 40 km) indicate that similarly-located stations had similar ground shaking intensity residuals. We next use these data to build a quantitative model of spatial correlations.

2.2 | The semivariogram

A popular tool to estimate spatial correlations is the semivariogram, which measures dissimilarity of two values $\delta W_{i,j}$ and $\delta W_{i,k}$

$$\gamma_{j,k} = \frac{1}{2} E\left[\left(\delta W_{i,j} - \delta W_{i,k}\right)^2\right] \quad (3)$$

where j and k are two locations of interest, γ denotes the semivariogram, and $E[\,]$ denotes expectation[33].

The semivariogram can be empirically estimated from observed data. Because we rarely have data to make estimates for specific locations j and k, we typically make an assumption of stationarity and isotropy: all locations separated by a distance h have the same semivariance. In this case, we can estimate the semivariogram by pooling all observations with a given separation distance h and using them to estimate the semivariance:

$$\hat{\gamma}(h) = \frac{1}{2n(h)} \sum_{d(j,k)=h} \left(\delta W_{i,j} - \delta W_{i,k}\right)^2 \quad (4)$$

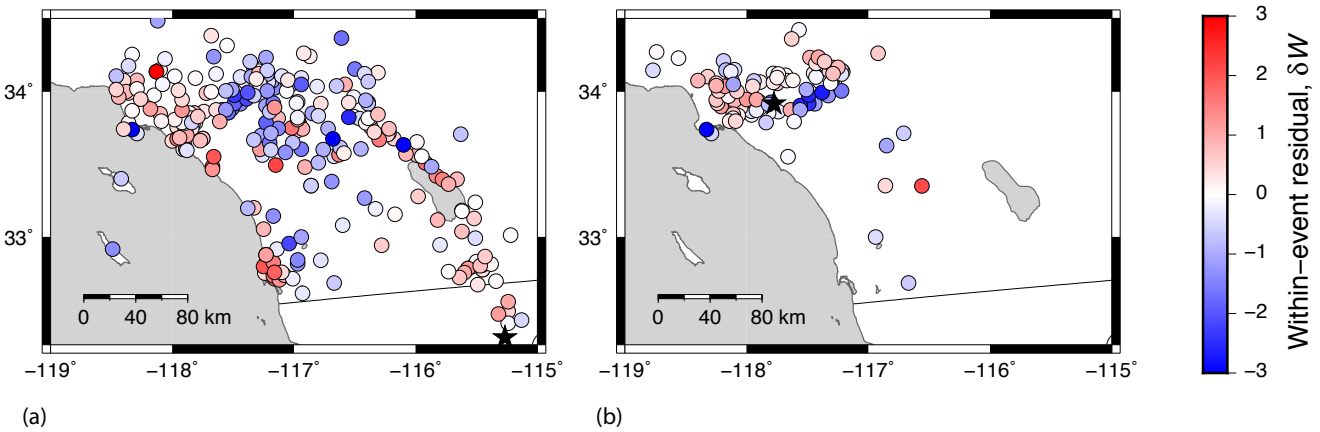

FIGURE 1 Observed $SA(1s)$ within-event residuals from (a) the M_w 7.2 2010 El Mayor-Cucapah, earthquake, and (b) the M_w 4.3 2002 Yorba Linda, earthquake. The earthquake epicenter is shown with a black star, and residuals are shown with colored circles.

where $d(j,k)$ is the distance between sites j and k, the summation is over all j and k with separation distance h (within a user-specified tolerance), and $n(h)$ is the number of observed station pairs with separation distance h (i.e., the counts shown in Figure 3). Note that while Equation 4 is the most common semivariogram estimator, some studies[3,21,17] have used an alternate estimator developed to be less sensitive to outlier data[34].

Empirical semivariograms obtained from the ground motion data shown in Figure 1 are computed using the Equation 4 estimator and plotted in circles in Figure 2. The semivariogram values start near 0 at small separation distances h, and increase to values of approximately one at large distances. The circles in the figure represent estimates at increments of 3km, and each estimate uses a tolerance of $\pm$1.5 km in the distance criterion.

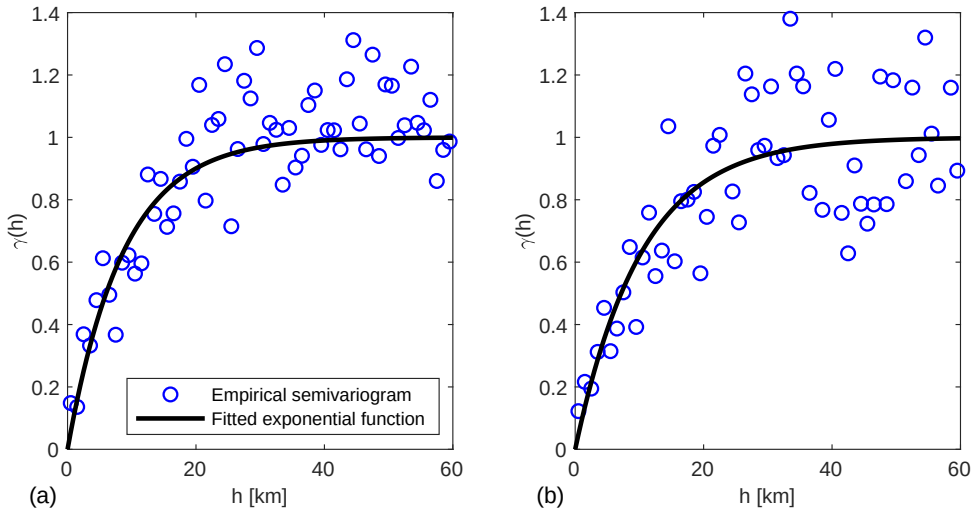

FIGURE 2 Example empirical semivariograms (Equation 4) and fitted semivariograms (Equation 5) for $SA(1s)$ within-event residuals from the (a) El Mayor-Cucapah, and (b) Yorba Linda, earthquakes.

In Figure 2, the large variation in the circles' vertical axis values with small changes in h is an indication of estimation uncertainty (because the true semivariance varies smoothly with small changes in h). This is expected, as the estimate is made from a finite number of data pairs, leading to estimation uncertainty. Further, given the maps of Figure 1, we know that the

El Mayor-Cucapah earthquake has more than double the number of recordings than the Yorba Linda earthquake (290 versus 118), so we expect the Figure 2b semivariogram values to be more uncertain than the Figure 2a values. To further quantify this issue, Figure 3 shows the number of data pairs at each separation distance for each earthquake. Depending upon the specific distance, the El Mayor-Cucapah earthquake has 1.6 to 6.5 times the number of data pairs. The numbers of pairs for all other earthquakes are also shown in the figure, to illustrate that these two earthquakes are typical of the broader data set (although El Mayor-Cucapah is one of the better-recorded events).

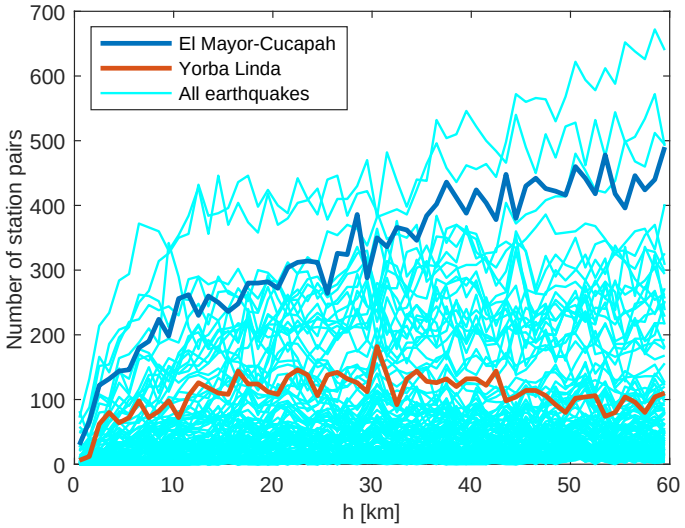

FIGURE 3 Number of paired observations of within-event residuals ($\delta W_{i,j}$) for $SA(1s)$ values observed in all considered earthquakes, at each considered separation distance. The El Mayor-Cucapah and Yorba Linda earthquakes are highlighted with distinct line styles.

2.3 | Semivariogram model fitting

The data from the previous section is next used to fit a semivariogram model that can be used for forward predictions. The model is a continuous function of distance (and possibly other parameters such as azimuthal angle). It must be positive definite to satisfy the requirements of the semivariance definition.

Here we will consider the common 'exponential' semivariogram model (but note that other models can be considered, as will be discussed later):

$$\tilde{\gamma}(h) = s\left[1 - \exp\left(\frac{-3h}{r}\right)\right] \tag{5}$$

Where h is the separation distance and s and r are parameters to be fitted (referred to as the sill and range, respectively). Figure 2 illustrates fitted exponential semivariograms.

The semivariogram is directly related to a correlation function, by the following relationship:

$$\gamma(h) = Var[\delta W_{i,j}]\,(1 - \rho(h)) \tag{6}$$

where $\rho(h)$ is the correlation at separation distance h, $Var[\delta W_{i,j}]$ is variance of the residuals, and it can shown that $Var[\delta W_{i,j}] = s$. We will assume throughout this study that $s = 1$, due to the previous standardization of the residual data. Nonetheless, for generality, we provide general formulas for estimating s in the equations below. We could study correlation instead of semivariance, but the semivariogram is often preferred in geostatistical practice as it does not require a prior estimation of the mean or standard deviation of the considered parameter.

Determining the parameters of the fitted semivariogram model, such that it is a 'good' fit with the empirical semivariogram data, requires a metric to evaluate misfit. We can then choose the parameter values that minimize this misfit. Generally, we can write this:

$$\hat{s}, \hat{r} = \arg\min_{r,s} \sum_{i=1}^{n} f\left(\hat{\gamma}(h_i), \tilde{\gamma}(h_i)\right) \tag{7}$$

where i is the i^{th} separation distance of interest, n is the number of separation distances, $\hat{\gamma}(h_i)$ and $\tilde{\gamma}(h_i)$ are defined in Equations 4 and 5, and $f(\,)$ is some function to evaluate misfit between $\hat{\gamma}(h_i)$ and $\tilde{\gamma}(h_i)$. Several fitting methods have been proposed in the earthquake engineering and general geostatistical literature, and almost all follow the general form of Equation 7. Several specific examples are given below, along with citations to studies that proposed or adopted each approach. Note that some of the cited studies considered correlation coefficients rather than semivariograms, but the formulas have been converted using Equation 6 so that each approach is presented in a consistent format.

An ***Ordinary Least Squares*** approach simply minimizes the sum of squared differences between the empirical data and fitted model:

$$\hat{s}, \hat{r} = \arg\min_{r,s} \sum_{i=1}^{n} \left(\hat{\gamma}(h_i) - \tilde{\gamma}(h_i)\right)^2 \tag{8}$$

This is conceptually simple, is implemented in several numerical algorithms, and one numerical package names it as 'the best general approach'[35]. It has been used in a number of ground motion studies[13,36,21,37]. But this approach does not account for the fact that the empirical $\hat{\gamma}(h_i)$ values differ in variance depending upon the $\hat{\gamma}(h_i)$ and number of data points, and ignores that the model at small-h values is of much greater practical importance than at large-h values.

A ***Weighted Least Squares*** approach incorporates weights on the squared errors to refine the fitting:

$$\hat{s}, \hat{r} = \arg\min_{r,s} \sum_{i=1}^{n} w_i \left(\hat{\gamma}(h_i) - \tilde{\gamma}(h_i)\right)^2 \tag{9}$$

where w_i is the weight for the i^{th} term in the summation. Weights are typically used to increase the relative importance of small-distance values (because they are of more practical importance and are lower-variance by nature of their smaller $\hat{\gamma}$ value) and to increase the relative importance of $\hat{\gamma}(h_i)$ values estimated from more data (because they are thus known with more precision). Several ground motion studies have used weighted least squares functions[4,15]. The following weighting function, which will refer to below simply as ***Weighted Least Squares***, is proposed in this study:

$$w_i = n(h_i)\, e^{-h_i/c} \tag{10}$$

where c is a coefficient that controls how quickly the weights decrease with increasing distance. The value of c can be adjusted depending on the anticipated extent of spatial correlation. Here we will use $c = 5$ km, indicating that an error at a distance of 5 km receives approximately 1/3 the weight of an error at a distance of 0 km. This choice of coefficient value is discussed further in the Appendix.

The R software's gstat package[38] provides a number of weighting functions for use in weighted least squares semivariogram fitting, and by default uses the following weighting function, which we will refer to as ***Weighted Least Squares,*** $\mathbf{w = n/h^2}$ below:

$$w_i = \frac{n(h_i)}{h_i^2} \tag{11}$$

This function places greater weights on values estimated with high numbers of data points (i.e., large $n(h_i)$) and at small distances.

Other fitting approaches perform a transformation on the semivariogram values, as part of the process of evaluating misfit. ***Cressie***[25] proposed the following function based on a theoretical derivation that minimized estimation errors:

$$\hat{s}, \hat{r} = \arg\min_{r,s} \sum_{i=1}^{n} n(h_i) \left(\frac{\hat{\gamma}(h_i)}{\tilde{\gamma}(h_i)} - 1\right)^2 \tag{12}$$

Note that the $n(h_i)$ term in the summation is a weighting term based on the number of data pairs, so this approach also incorporates weighting, and the theory behind Cressie's derivation led to the inclusion of this term. Two recent ground motion studies have utilized this approach for fitting semivariograms[39,40].

Two ground motion studies used a ***Fisher Transformation*** during fitting to account for differing estimation variances for highly-correlated versus less correlated conditions[16,41]

$$\hat{s}, \hat{r} = \arg\min_{r,s} \sum_{i=1}^{n} \left(\ln \frac{2 - \hat{\gamma}(h_i)}{\hat{\gamma}(h_i)} - \ln \frac{2 - \tilde{\gamma}(h_i)}{\tilde{\gamma}(h_i)} \right)^2 \tag{13}$$

This transformation compensates for the fact that estimated semivariances (and correlation coefficients) have estimation uncertainty that depends on the actual semivariance value. The Fisher-transformed estimated parameter has a normal distribution with mean zero and variance of $1/(n(h_i) - 3)$. For this approach to have errors with uniform variances, the misfit metric should thus also include a weight factor of $n(h_i) - 3$, though neither of the cited studies include this factor.

One ground motion study used a transformation that then allows for fitting of the semivariogram parameter r using ***Linear Regression***[7] (the s parameter was estimated separately):

$$\hat{r} = \arg\min_{r} \sum_{i=1}^{n} \frac{1}{h_i} \left[\ln\left(1 - \min\left\{\hat{\gamma}(h_i), 0.99\right\}\right) - \ln\left(1 - \tilde{\gamma}(h_i)\right) \right]^2 \tag{14}$$

This transformation requires truncating the empirical semivariogram at 0.99 to avoid taking logarithms of negative numbers. Note the $1/h_i$ weighting to more heavily weight small-distance values, reflecting one goal of some weighting schemes above.

A few other approaches have been used with ground motion data. One common approach is to perform a visual fit[14,42,22,43]. This allows an informed user to tailor the fit to distances of interest, down-weight outliers, etc., and thus is considered a reasonable approach[44]. It is not considered further here, however, as it is not replicable and thus not compatible with the calculations below. Others have proposed simultaneously fitting the ground motion model and a spatial correlation model[45,46], but these approaches are more algorithmically complex and less compatible with the calculations below, so are also not considered further here.

Reviewing Equations 8-14, it is clear that there are multiple possible semivariogram fitting methods. The earthquake engineering and geostatistics communities have not reached consensus on a preferred approach. This is because there is no theoretically optimal approach (unlike classic statistical problems such as least-squares linear regression) and because the effectiveness of the algorithms depends upon the number and spatial configuration of the data being studied. Several of the proposers of the above approaches explicitly note that the formulations are pragmatic rather than being based on theory[7,38]; in other studies this issue is implicit, but it remains. For the same reason, most geostatistical software packages provide multiple fitting methods in their libraries and leave it to the user to select an appropriate one[35,38].

To illustrate the practical implications of these fitting options, Figure 4 shows the empirical semivariogram data from Figure 2, with exponential models fitted using four of the above approaches. Because the residuals are standardized, the sill is set equal to 1 in each case, and only the range is estimated. The methods produce similar, but not identical, fitted functions. For each specific event, the estimated r values vary by approximately 6 km among the methods. We also see that the El-Major Cucapah ranges (Figure 2b) are approximately 5km smaller than the Yorba Linda ranges (Figure 2a). These anecdotal results suggest that estimation uncertainty and event-to-event uncertainty in ranges are both non-trivial. These issues will be explored more quantitatively in the following sections.

3 | MODEL FOR OBSERVED SEMIVARIOGRAM UNCERTAINTY

We propose the following random effects model to characterize real and apparent uncertainty in semivariograms. We assume that the semivariogram for a particular earthquake can be represented by a parametric model with an uncertain parameter or parameters. In this discussion, we will consider the range parameter r from Equation 5, but the same process can be applied to other model functions or parameters.

We then assume that a population of earthquakes has a distribution of r values, with earthquake i having a true (but unknown to us) value of r_i. We consider the population of earthquakes to have a mean value of μ_{r_i} and a standard deviation of σ_{r_i}. For earthquake i and its associated recorded ground motions, we do not know the true value of r_i, but we can estimate it using one of the methods from Section 2.3. We denote this estimate $\hat{r}_i$. Because of our limited observational data, the estimate will have uncertainty. We denote the estimator's standard deviation, conditional on the true value r_i, as $\sigma_{\hat{r}_i|r_i}$ (where '$A|B$' denotes that A is conditional on B). We also denote the estimator's mean, conditional on the true value r_i, as $\mu_{\hat{r}_i|r_i}$. We would like an estimator that is unbiased (i.e., $\mu_{\hat{r}_i|r_i} = r_i$) and that has small variance (i.e., $\sigma_{\hat{r}_i|r_i}$ is small). We will explore these estimator properties in Section 3.1.

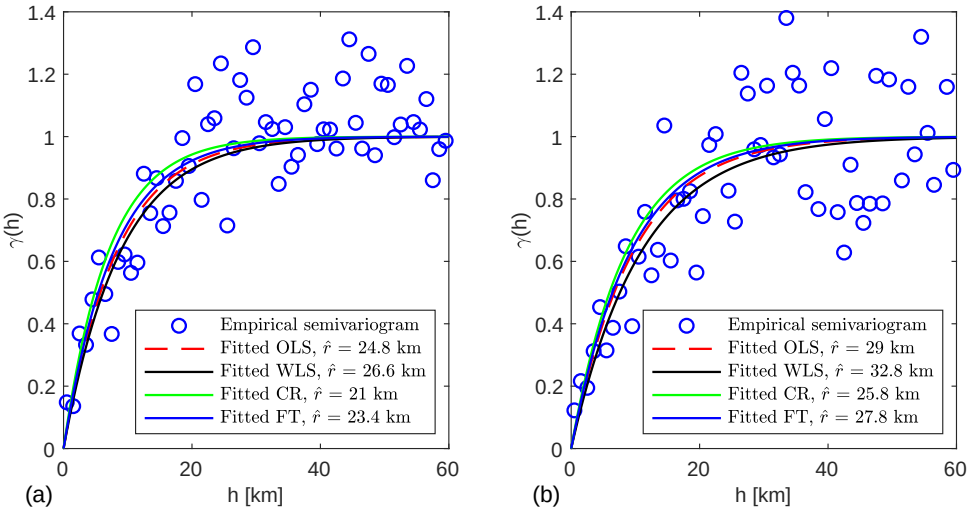

FIGURE 4 Example empirical semivariograms and fitted semivariograms using four methods, for $SA(1s)$ within-event residuals from the (a) El Mayor-Cucapah, and (b) Yorba Linda, earthquakes. Legend abbreviations: OLS = Ordinary Least Squares, WLS = Weighted Least Squares, CR = Cressie, FT = Fisher Transform.

When we estimate a semivariogram model from observational data, we obtain an $\hat{r}_i$ that reflects both the variability in r_i, and the estimation variability in $\hat{r}_i|r_i$. According to the variance decomposition formula, the total variance of $\hat{r}_i$ is:

$$\sigma^2_{\hat{r}_i} = E\left[Var\left(\hat{r}_i|r_i\right)\right] + Var\left(E\left[\hat{r}_i|r_i\right]\right) \tag{15}$$

where $Var(\,)$ denotes variance. Note that, given r_i, $\hat{r}_i$ is conditionally independent of μ_{r_i} and σ_{r_i}. If $\hat{r}_i$ is an unbiased estimator of r_i, Equation 15 can be simplified to

$$\sigma^2_{\hat{r}_i} = \sigma^2_{\hat{r}_i|r_i} + \sigma^2_{r_i} \tag{16}$$

We will quantify and study this total uncertainty in Section 3.2.

3.1 | Quantification of estimation uncertainty

We next consider the estimation uncertainty associated with a particular fitting method and a particular set of ground motion data, to quantify $\mu_{\hat{r}_i|r_i}$ and $\sigma_{\hat{r}_i|r_i}$. We propose the following five-step procedure to estimate these properties:

1. Specify station locations, an assumed semivariogram model, and a semivariogram estimation method. In this study, we use locations of recordings from past earthquakes, an exponential semivariogram with $r = 30$ km (a typical range seen in real ground motion data) and $s = 1$, and consider the six fitting methods described in the previous section.
2. Generate many Monte Carlo simulations of ground motion data (i.e., $\delta W_{i,j}$ values at each station location), using a multivariate normal distribution with mean values of zero, and a covariance matrix specified by the semivariogram.
3. For each Monte Carlo simulation of residuals, compute an empirical semivariogram using Equation 4, and estimate a semivariogram model using the chosen fitting method.
4. Take the set of range estimates from step 3, and compare them to the (known) range in order to study the estimation error associated with that set of station locations and fitting method.
5. Repeat steps 1-4 for all station location configurations and all fitting methods of interest.

Figure 5 graphically illustrates the procedure, with steps 1-4 numbered on the figure.

This calculation process addresses several issues needed to decompose the total apparent variance in Equation 15. First, by specifying the semivariogram model and range in Step 1, and sampling data from that model, we can generate data with a known

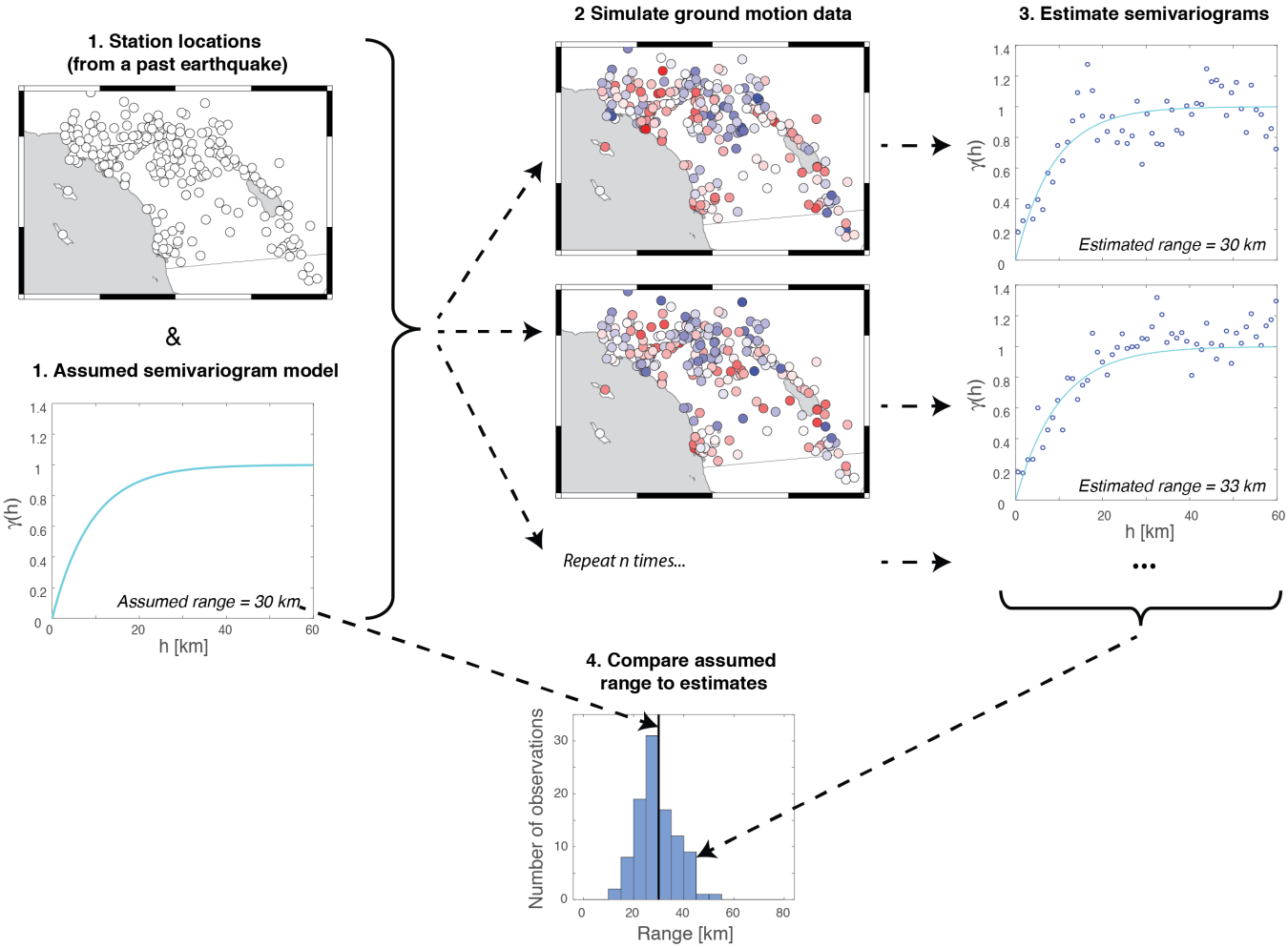

FIGURE 5 Illustration of the process used to quantify semivariogram model parameter estimation uncertainty.

r_i (unlike with real data). Second, by Monte Carlo sampling synthetic data, we can obtain multiple parameter estimates for a given earthquake (unlike with real data). Third, by using station locations from real earthquakes, we can study the performance of fitting algorithms in exactly the conditions we use them for, and we are also able to obtain event-specific estimates of estimation uncertainty.

Figure 6 shows histograms of range estimates obtained from this procedure, using the station locations shown in Figure 1 and the Weighted Least Squares fitting method. As anticipated, the El Mayor-Cucapah range estimates are less variable than the Yorba Linda estimates, because of the greater data: the standard deviations of the data are 7.7 and 9.3 km, respectively. In both cases, the average range estimate is fairly close to the true value of 30km (the sample means are 29.1 and 29.5 km, respectively).

To study estimation variability and bias more comprehensively, Figure 7 shows scatter plots of estimated ranges for all considered earthquakes and fitting methods. Each subfigure was produced with a single fitting method. Within each subfigure, the data for each earthquake is plotted versus the number of recording stations from that earthquake (so the replicates from a single earthquake are plotted as a vertical stripe of data). The correct range value of 30km is shown with a dotted line for reference. To aid interpretation, a moving average and moving standard deviation (σ) are estimated from the data and also plotted on the figures.

We see in Figure 7 that for all fitting methods, more reliable estimates are obtained with earthquakes with a greater number of stations, as expected (i.e., the moving average tends towards the correct range and the $\pm\sigma$ interval gets smaller). For earthquakes with small numbers of stations, the estimates are biased and have large $\pm\sigma$ intervals. Results of this type can guide investigators in choosing appropriate data sets and can quantify the fitting uncertainty and bias associated with a particular earthquake or set

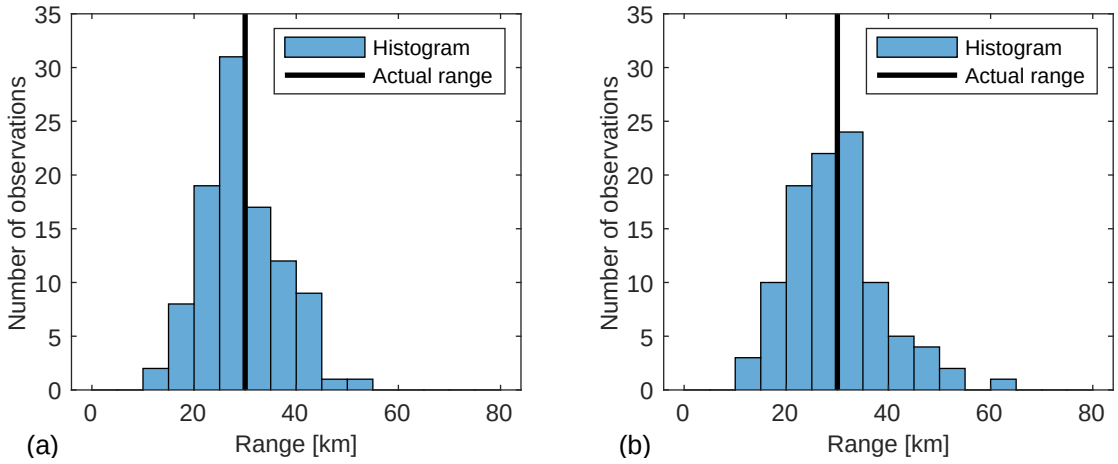

FIGURE 6 Histograms of ranges estimated when sampling residuals from an exponential semivariogram model and using the Weighted Least Squares fitting method. (a) Using station locations from the El Mayor-Cucapah, earthquake (b) Using station locations from the Yorba Linda, earthquake.

of earthquakes. Note that several previous studies have used only earthquakes with $n \geq 100$ stations to perform earthquake-specific semivariogram fitting[5,16,19], and others have used earthquakes with as few as 90 stations[47,7]. The results here suggest that those less-well-recorded events will be subject to substantial estimation uncertainty.

With the above general observations, we then group the data into three groups: the 48 earthquakes with $n \leq 65$ stations (when estimation uncertainty is large), the 23 earthquakes with $n > 130$ stations (when estimates are less biased and have smaller estimation uncertainty), and the 58 earthquakes with $65 < n \leq 130$ stations (in the middle). Estimation biases are reported in Table 1, and estimation sample standard deviations are reported in Table 2 for these pools of data. From Figure 7 and these Tables, we see that the Weighted Least Squares and Fisher Transformation methods have lower bias than the other four methods for the $n > 130$ earthquakes of greatest interest. And between these two, the Weighted Least Squares approach has the smaller estimation standard deviation.

To evaluate the robustness of these results to the assumed range value, we repeated the calculations in this section for r values between 15 and 45 km. The bias results were quite stable for all assumed ranges. The estimation standard deviations did vary somewhat with range (with larger ranges producing larger standard deviations), but the increase in standard deviation was not proportional to the increase in r. The effect of the number of recording stations and the relative performance of the six fitting methods were consistent across all cases.

Event-specific estimation standard deviations obtained from the Weighted Least Squares approach and using $r = 30$ are reported in Table A1.

TABLE 1 Fitting bias for the six considered fitting methods. The bias is the mean of the sample ranges from earthquakes with the given numbers of recording stations, minus the true range of 30.

Fitting method	**Bias when** $n \leq 65$	**Bias when** $65 < n \leq 130$	**Bias when** $n > 130$
Ordinary Least Squares	10.3	7.3	3.7
Weighted Least Squares	4.5	1.2	-1.2
Weighted Least Squares, $w = n/h^2$	4.2	0.5	-3.0
Cressie	-17.3	-10.2	-4.9
Fisher Transform	24.2	9.1	0.3
Linear Regression	10.6	4.8	3.0

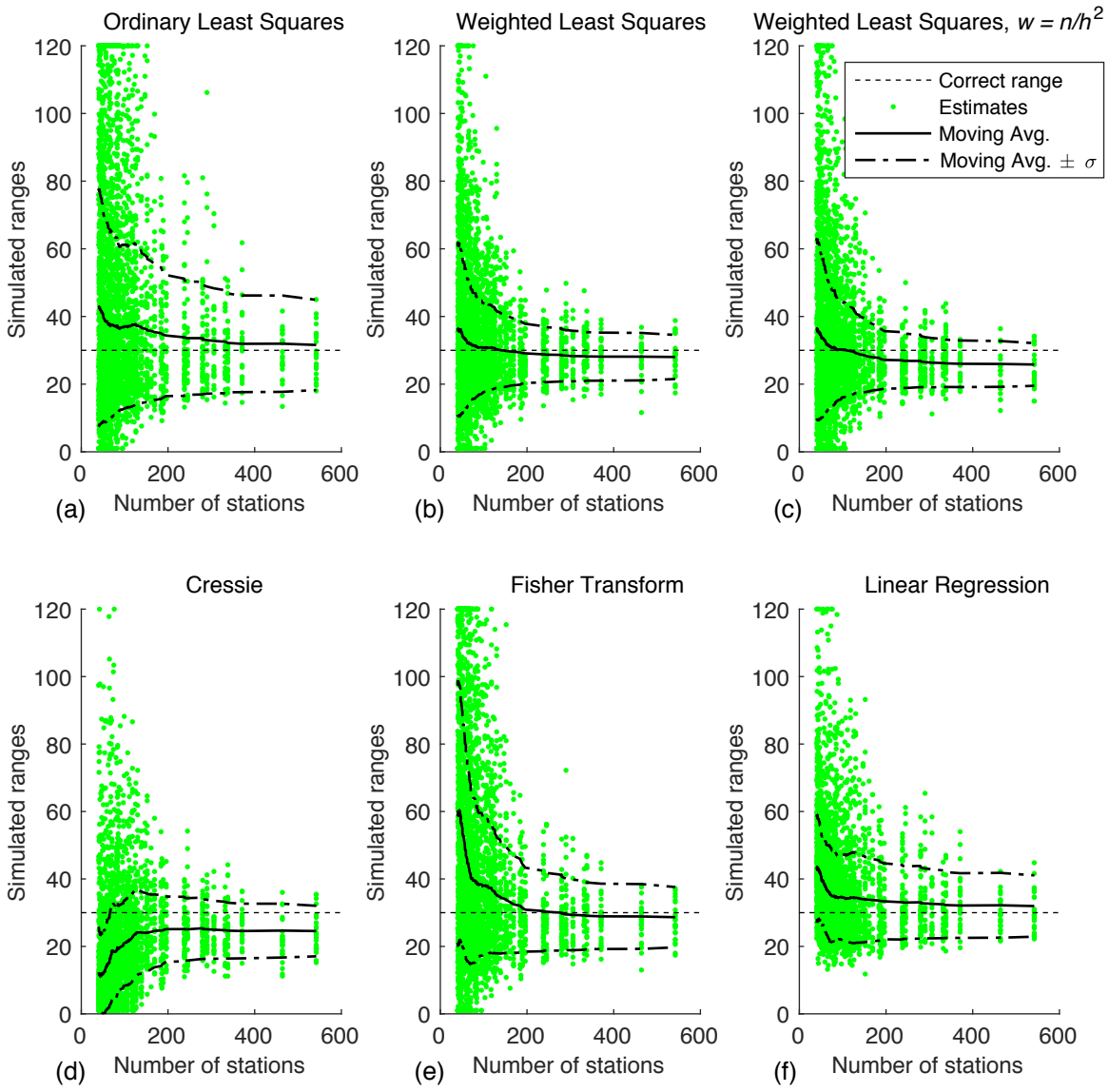

FIGURE 7 Replicates of range estimates using each considered earthquake's station locations and an assumed exponential semivariogram model with $r = 30$ km, plotted versus the number of stations. Estimates using (a) Ordinary Least Squares, (b) Weighted Least Squares, (c) Weighted Least Squares, n/h^2, (d) Cressie, (e) Fisher Transform, (e) Linear Regression. The correct range, the moving average of the estimates, and the moving average ± the moving standard deviation are plotted for reference.

3.2 | Total uncertainty

We next consider the total uncertainty that is apparent in $\hat{r}_i$ from real earthquake data, using the preferred Weighted Least Squares fitting method. We again consider the sets of earthquakes with $n \leq 65$, $65 < n \leq 130$, and $n > 130$ stations, and compute the sample standard deviations of the ranges from the *real* $SA(1s)$ ground motion data. For the Weighted Least Squares method, the sample standard deviations ($\sigma_{\hat{r}_i}$) are 32.0, 25.7, and 20.5 km, respectively. From Table 2, we see that the corresponding estimation standard deviations ($\sigma_{\hat{r}_i|r_i}$) are 23.0, 14.7, and 8.1 km, respectively. Substituting these values into Equation 16 gives an estimate of the true underlying standard deviation of r: 22.2, 21.1, and 18.8 km, respectively. These same calculations are repeated for spectral acceleration data at other periods, and results are plotted in Figure 8. While the results vary somewhat

TABLE 2 Estimation standard deviation for the six considered fitting methods. Sample standard deviations of the ranges are computed for the data from earthquakes with the given numbers of recording stations.

Fitting method	$\sigma_{\hat{r}_i \mid r_i}$ **when** $n \leq 65$	$\sigma_{\hat{r}_i \mid r_i}$ **when** $65 < n \leq 130$	$\sigma_{\hat{r}_i \mid r_i}$ **when** $n > 130$
Ordinary Least Squares	31.8	25.4	16.9
Weighted Least Squares	23.0	14.7	8.1
Weighted Least Squares, $w = n/h^2$	24.0	15.7	8.0
Cressie	12.5	13.4	9.3
Fisher Transform	34.8	22.6	11.6
Linear Regression	14.1	13.3	10.8

by period, there is a general trend of the total standard deviation and the estimation standard deviation being larger for poorly recorded earthquakes. In contrast, the underlying standard deviation (σ_{r_i}) is fairly constant near 20 to 25 km for most periods and groups of earthquakes. The consistency of σ_{r_i} among groups of earthquakes is reassuring, as the proposed model assumed that this was a property of the earthquake itself (and not dependent upon the number of stations that recorded it).

These results, and those of Figure 7, suggest that at least 100 (and preferably 200) stations should be available in order to obtain an event-specific range estimate with reasonably low estimation uncertainty. More precisely, the estimation uncertainty depends in a more complex way on the number of station pairs with small separation distances, but the results vary depending upon the fitting method and are omitted here for brevity. The following section will present a technique for quantifying uncertainty in event-specific range estimates using these results.

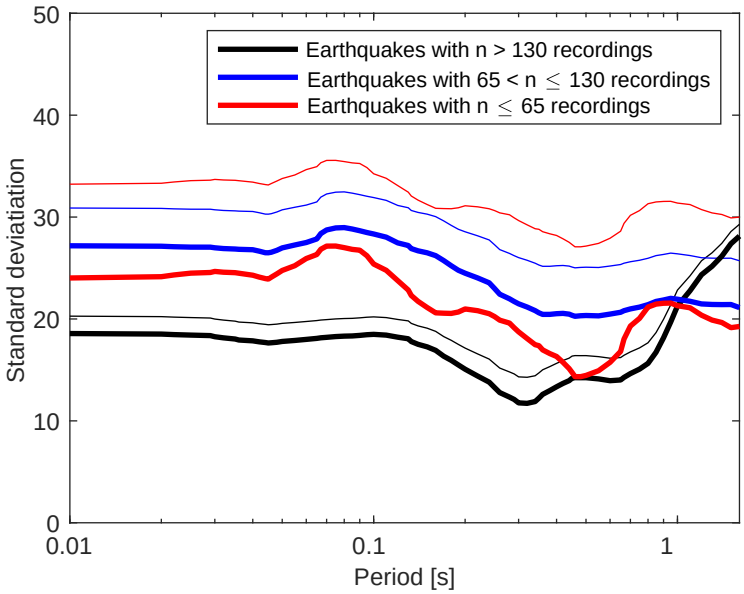

FIGURE 8 Apparent and actual standard deviations of range estimates obtained from Weighted Least Squares fitting to spectral acceleration values at a range of periods. Separate results are shown separately for data from earthquakes with a small, medium, or large number of recordings. Estimated $\sigma_{\hat{r}_i}$ values are shown in light lines and estimated σ_{r_i} values are shown in heavy lines. The lines have been smoothed with a moving average function for ease of reading.

3.3 | Earthquake-specific posterior parameter estimates

The decomposition of the earthquake-specific parameter estimates into two types of uncertainty allows for estimation of a posterior distribution of r_i, conditional upon an estimated prior distribution for r_i and the event-specific estimate $\hat{r}_i$. If we assume that r_i and $\hat{r}_i$ are jointly normally distributed (which is true if r_i is marginally normal and $\hat{r}_i|r_i$ is marginally normal with mean r_i and standard deviation independent of r_i), the posterior distribution of $r_i|\hat{r}_i$ can be shown to be normally distributed with the following mean and variance[48]

$$\mu_{r_i|\hat{r}_i} = \frac{\sigma^2_{r_i}}{\sigma^2_{\hat{r}_i|r_i} + \sigma^2_{r_i}}\hat{r}_i + \frac{\sigma^2_{\hat{r}_i|r_i}}{\sigma^2_{\hat{r}_i|r_i} + \sigma^2_{r_i}}\mu_{r_i} \tag{17}$$

$$\sigma^2_{r_i|\hat{r}_i} = \left(\frac{1}{\sigma^2_{\hat{r}_i|r_i}} + \frac{1}{\sigma^2_{r_i}}\right)^{-1} \tag{18}$$

Equation 17 shows that for small estimation uncertainty ($\sigma^2_{\hat{r}_i|r_i}$) the posterior mean tends towards the estimated value ($\hat{r}_i$) because the small estimation uncertainty makes the estimate highly informative. On the other hand, for large estimation uncertainty, the posterior tends toward the overall population mean (μ_{r_i}). Similarly, Equation 18 shows that the posterior variance decreases as $\sigma^2_{\hat{r}_i|r_i}$ decreases, because the estimate is more informative.

To use these equations on our $SA(1s)$ earthquake data, we assume the range estimates are unbiased (a reasonable assumption for at least the well-recorded earthquakes, per Table 1), obtain event-specific estimates of $\sigma^2_{\hat{r}_i|r_i}$ from the data of Figure 6 and 7, estimate $\mu_{r_i} = 30$ km from the sample mean of the real earthquake data range estimates (29.6 km), and estimate $\sigma_{r_i} = 20$ km from Figure 8. This approach uses an Empirical Bayesian formulation, in which the population's prior parameters are estimated from the data. With this approach, we are also using a $\sigma^2_{\hat{r}_i|r_i}$ estimate obtained assuming $r_i = 30$ km, while it was noted above that this standard deviation varies somewhat with r_i; this assumption adds some degree of approximation to the results.

Figure 9 shows the posterior distributions estimated for the two example earthquakes considered throughout this paper. The point estimates of the ranges (26.6 and 32.8 km for El Mayor-Cucapah and Yorba Linda, respectively, from Figure 4) are shown in dotted lines. The posterior mean values from Equation 17 are shown in dashed lines; they both move slightly towards the global mean value of 30km, relative to the point-estimate values. The posterior probability density functions are shown in solid lines. These probability density functions have mean values from Equation 17 and standard deviations from Equation 18. The posterior standard deviations (7.2 and 8.4 km) are substantial relative to the difference in posterior means between these two earthquakes (5.2 km), indicating that there is no strong evidence that the ranges between these two earthquakes differ. The posterior distribution for El Major-Cucapah is narrower than the distribution for Yorba Linda, due to El Major-Cucapah's smaller estimation uncertainty. These results are consistent with qualitative evaluation of the data presented earlier in Figures 1 and 4.

Posterior distributions are computed for all 129 considered earthquakes and shown in Figure 10. Each earthquake's range estimate is shown with a single point, and its posterior mean and ± one standard deviation is shown with an error bar. The horizontal axis of the figure shows the number of recordings for each earthquake (in log scale to slightly ease viewing of results from the poorly-recorded earthquakes). We see that the error bars are wider at the left side of the figure, as these are the events with larger estimation uncertainty. Further, 83% of the earthquakes with $n \leq 65$ recordings include the prior mean (30 km) in the ± one standard deviation interval, but only 35% of the earthquakes with $n > 130$ recordings include it. This implies that only the well-recorded earthquakes can provide clear information regarding the potential deviations of ranges from a general distribution. Note also that for the poorly recorded earthquakes with point estimates of $\hat{r}_i > 80$ km, the posterior distributions do not even include that point estimate in the ± one posterior standard deviation interval. This is because the poorly recorded earthquakes have significant estimation uncertainty, so the extreme values of range estimates are likely resulting from poor estimation rather than clear evidence of a truly extreme range. Mean posterior range estimates for each considered earthquake are reported in Table A1.

4 | DISCUSSION

The above results have several limitations due to assumptions in the formulation and several broader implications for the field of spatial correlation estimation. While the above framework for estimating estimation uncertainty and decomposing it into underlying range variation versus estimation error is general, it was applied under a specific set of assumptions here, and those assumptions influence the presented results.

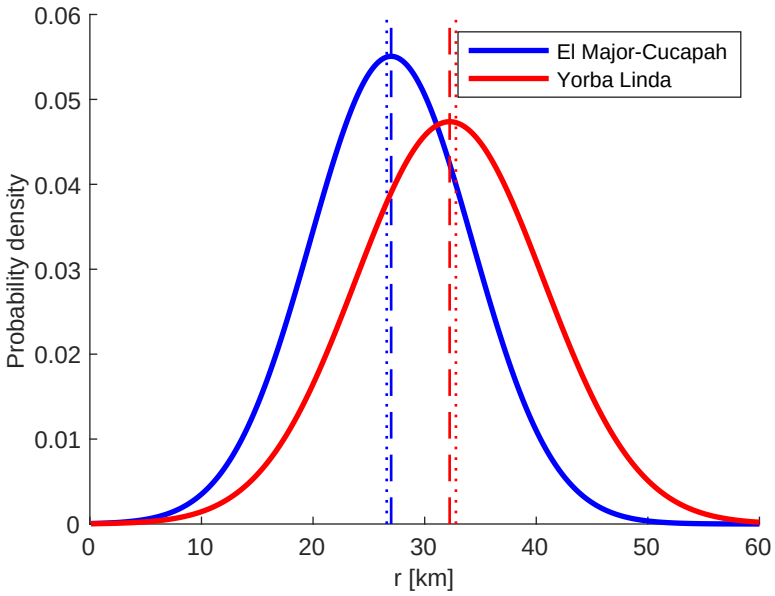

FIGURE 9 Posterior distributions of r_i for the example El Mayor-Cucapah and Yorba Linda earthquakes, estimated using the Weighted Least Squares method. Posterior probability density functions are shown in solid lines, point estimates of $\hat{r}_i$ are shown in dotted lines, and posterior mean values of $\hat{r}_i$ are shown in dashed lines.

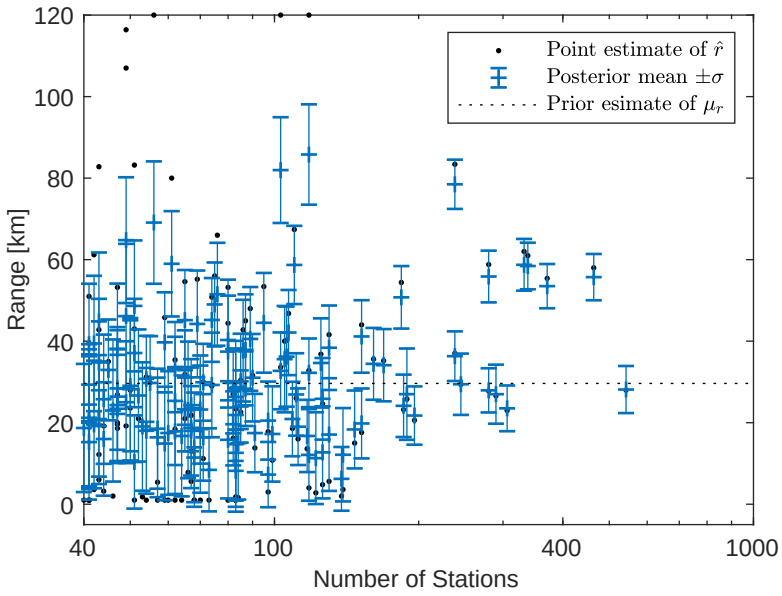

FIGURE 10 Point estimates and posterior distributions of r for all earthquakes, plotted versus the earthquakes' number of recordings.

First, we have assumed that an exponential semivariogram model correctly captures the semivariance. This is reasonable for some earthquakes (e.g., Figure 2), but it is less reasonable for others, and a number of other studies have considered alternate semivariogram functional forms[1,13,7,16]. If the assumed functional form is a poor representation of the data's actual spatial structure, the functional-form error will partially manifest as parameter error, complicating the statistical analysis above. Further,

the estimation-error calculations of Section 3.1 generate data from a specified semivariogram function, and so the results will not reflect the effects of a semivariogram functional form that is incorrectly assumed or is varying from earthquake to earthquake.

Second, we have considered a semivariogram functional form (exponential) with only a single unknown parameter (r). Many other semivariogram functions have more than one parameter. The procedures above can be employed to estimate uncertainty in multiple parameters, with only minor adjustments. From our limited experience with other functional forms, we have seen that when there is more than one model parameter, the uncertainty in each individual parameter can increase substantially. This is because there are often multiple combinations of parameter values that lead to comparably good fits, leading individual parameters to be poorly constrained. A multi-parameter estimation technique could also be used to estimate the standard deviation (s) parameter in Equation 5, and it may produce evidence of varying standard deviations from earthquake to earthquake. However, this quantification would be more appropriate within the initial GMM development stage, rather than during spatial correlation estimation based on the GMM.

Third, we have used the classical estimator of the semivariance (Equation 4). Some ground motion studies[21,17] have used an alternate robust estimator proposed by Cressie and Hawkins[34]. Schiappapietra and Douglas[20] explore the performance of the classical versus robust empirical semivariogram estimators, and find that the performance is generally comparable between the two. So we expect that the results measured above would be similar if the robust estimator were used in the procedure instead.

Fourth, the conditional event-specific parameter distributions of Section 3.3 requires some additional assumptions. We have assumed that the parameter, and its estimation error, are both normally distributed, and that the estimation error is independent of the true parameter value. Further, we have estimated the true parameter value's distribution from population data. (A similar alternative formulation, not presented here, would assume that the parameter and its estimation uncertainty are both lognormally distributed.) These assumptions are almost certainly not strictly true, and the parameter distribution estimate is approximate, so the conditional distributions of Equations 17-18 are approximate. They nonetheless are informative in illustrating the factors influencing parameter estimates and their uncertainty.

There are two notable broader implications of these results. First, these results indicate our ability to detect earthquake-specific variation in semivariances. Specifically, they show that subtle variations will be challenging to detect, given the inherent estimation uncertainty associated with the ground motion data we are currently able to obtain. The best-recorded earthquakes are the most useful for this purpose, but there are only 12 earthquakes in this catalog with more than 200 usable recordings. It is unlikely that this number will grow rapidly in the near future. Numerical simulations of ground motions are likely to provide the next generation of insights on causes of variations in correlations, as there is no data limitation and causality is easier to infer, though the simulation algorithms will require validation for use in this way[43,49].

A second implication is that there is a tradeoff between single-earthquake semivariogram estimates (which let us study the effect of a particular earthquake or region on ranges) and estimates obtained from pooling data from multiple earthquakes (which have reduced estimation uncertainty). Some ground motion correlation studies pool all data from all available earthquakes and produce one semivariogram model. In contrast, others split the data into specific earthquakes or specific regions, as discussed in the Introduction. Here there is a clear analogy with ground motion prediction models, where some models are general to a range of conditions, while others are tuned to specific regions or locations [*e.g.*, 50,51].

5 | CONCLUSIONS

This study has investigated the role of the fitting method and of estimation uncertainty on models for spatial correlation in ground motion amplitudes. This work was motivated by the variations in approaches used to estimate spatial correlations from ground motion data, and by the varying conclusions in the literature regarding the role of earthquake and site properties on resulting correlations. Spatial correlations of ground motion residuals were studied and were quantified using a semivariogram. The semivariogram is a measure of dissimilarity in values at two sites having a given separation distance, and can easily be converted into a correlation model. Here we have assumed that the semivariogram can be represented by a parametric model, and that the model parameter uncertainty is of interest.

The model parameter variability we observe when estimating semivariograms from a set of earthquakes comes from two sources: true variability from earthquake to earthquake, and estimation uncertainty resulting from limited observational data. The true variability is of primary interest and is what should be considered in calculations of possible future ground motions. The estimation uncertainty is an artifact of our analysis rather than a true phenomenon in nature; further, it is larger for poorly recorded earthquakes and smaller for well-recorded earthquakes.

We proposed a technique for quantifying the estimation uncertainty for a particular earthquake. With this technique, we assume a semivariogram model, take the station locations that recorded the earthquake, and then simulate data from the assumed correlation model at those locations. We then re-estimate the semivariogram model using a particular fitting method. By repeatedly simulating data and re-estimating the model, we can measure uncertainty and potential bias associated with a particular fitting method, and for a specific earthquake's recordings.

Applying this technique to the NGA West-2 database of crustal earthquakes, we evaluated a number of semivariogram fitting methods that have been used in the literature. The results suggest that a Weighted Least Squares fitting method is most effective in estimating the semivariogram, due to its limited bias and small estimation uncertainty. This technique identified model parameters by minimizing the squared differences between the observed semivariogram data and fitted function, where the squared differences are weighted by the number of data pairs and the inverse exponential of the distance value. That fitting method was then used to study earthquake-specific semivariogram models.

When considering earthquake-specific semivariogram models, we found that estimation uncertainty decreased as a function of the number of stations that recorded the earthquake. At least 100, and preferably 200, stations are needed to limit estimation uncertainty. For earthquakes with fewer stations, earthquake-specific estimation of spatial correlations should be performed with caution. When this estimation uncertainty was removed from the apparent total variability in semivariogram estimates, the implied underlying true variability was seen to be relatively stable among well-recorded and poorly recorded earthquakes, as expected. To rigorously incorporate estimation uncertainty, we also proposed an empirical Bayes approach to obtain a posterior estimate of semivariogram parameters, conditional on an estimated parameter, and an estimated prior distribution for the parameter from the data. These posterior estimates for the NGA West-2 data showed that extreme parameter values estimated for a few poorly recorded earthquakes are not well constrained and that the posterior distributions lie much closer to typical values than to the earthquake-specific point estimate.

Even for well-recorded earthquakes, fitted semivariograms have substantial estimation uncertainty. This fundamentally limits our ability to detect subtle earthquake-specific or region-specific features of spatial correlations from empirical data. That is not to say that no such features exist, but rather that these features will likely be difficult to discern from empirical data sets available at present. Researchers evaluating whether earthquake-specific patterns in spatial correlations are present in their data can use the approach proposed here to assess whether any potential trends rise above the level of estimation noise inherent to the data.

The results above assumed that ground motions are well described by an exponential semivariogram function with an average range of 30km (a typical value observed in past studies of real data) and utilized the Weighted Least Squares fitting method that was observed to perform better than alternatives. The calculations have been repeated for other range values, for other fitting methods, and for other semivariogram functional forms. Results in these other cases are broadly consistent those presented here, but have not yet been studied as extensively. Software to perform these calculations is freely available (see Data and Resources section), and is set up to evaluate alternative assumptions for analysts interested in other cases. While the numerical results presented here should be useful, the insights from the application of this framework to other data sets and assumed models are likely to have lasting value as researchers continue to investigate causes of spatial correlations in ground motions.

ACKNOWLEDGMENTS

We thank Pablo Heresi, Iunio Iervolino, Sabine Loos, Eduardo Miranda, Erika Schiappapietra, Peter Stafford, and an anonymous reviewer for helpful feedback and advice that improved the quality of this manuscript. This work was supported by the U.S. Geological Survey (USGS) via External Research Program award G20AP00019. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the USGS.

6 | DATA AND RESOURCES

The ground motion data in this study came from the NGA-West2 Database Flatfile (https://apps.peer.berkeley.edu/ngawest2/databases/). Maps were produced using Generic Mapping Tools [52]. Source code used to perform the above analysis and produce the figures in this paper is available at https://github.com/bakerjw/spatialCorrelationEstimation. The authors hope that the source code will help interested readers to perform further investigations of the type proposed here, and will also provide documented functions to perform basic semivariogram calculation and fitting.

How to cite this article: Baker J.W. and Y. Chen (2020), Uncertainty in estimating ground motion intensity spatial correlations, *Earthquake Engng Struct Dyn*, *2020* (in press).

APPENDIX

A WEIGHTED LEAST SQUARES COEFFICIENT

Given the good performance of the weighted least squares method in the sections above, that method is further scrutinized here. Specifically, the c coefficient in Equation 10 is a tunable parameter, and the fitting performance will depend upon the choice of that parameter. To explore the performance of the fitting approach as a function of the fitting parameter and the properties of the ground motion, the above assessment scheme was performed for several approaches. Two figures are provided to document the findings. Figure A1 provides results for estimated ranges from sythetic data (using the same approach as used to produce Figure 7). The four panels show results when using Weighted Least Squares fitting with four values of c in the weight function. Results like this were produced for a range of c values, and for each one, the average bias was computed:

$$\text{Average Bias} = \frac{1}{n}\sum_{i=1}^{n} |\bar{r}_i - r| \tag{A1}$$

where $\bar{r}_i$ is the mean estimated range for earthquake i, and r is the true range used to simulate the data (i.e., 30 in the case of Figure A1). Additionally, the average coefficient of variation was computed

$$\text{Average coefficient of variation} = \frac{1}{n}\sum_{i=1}^{n} \frac{s_{r,i}}{r} \tag{A2}$$

where $s_{r,i}$ is the sample standard deviation of the r estimates for earthquake i.

The above calculations were repeated for several c values, and for several r values used to simulate the data. The variation in r is important to consider, to ensure that the fitting function is not overly tuned to one particular semivariogram model, recognizing that real-world data will not conform to one specific semivariogram and so the estimator needs to be robust. The results are shown in Figure A2, and a few observations can be made. The estimation bias is minimized for $c \approx 5$ km. The coefficient of variation is minimized for $c \approx 4$ km. In general, $3 < c < 9$ km all perform reasonably well; this is also seen in Figure A1. The results are remarkably stable when the true underlying range used to simulate the ground motion data is varied. This was somewhat unexpected, but appears to result in part from the fact that typical earthquake data have relatively few paired observations with $h < 10km$ (Figure 3). While those pairs are important, a very small c (which would emphasize those pairs) would ignore most of the available data. Conversely, a very large c would emphasize distant station pairs that do not strongly constrain the model regardless of the true range. So the optimal weighting function seems to be dependent upon both the data configuration and the true underlying semivariogram.

Noting that the distance taper in this weighting scheme has the same functional form as the semivariogram of Equation 5, we see that an effective weighting scheme is one that tapers weights roughly as $1 - \gamma(h)$. This finding raises the question of whether an iteratively reweighted least squares approach would be effective. Such an approach would estimate a semivariogram and use that semivariogram to produce an updated weighting function, which is then used to re-estimate the semivariogram. The process would then be iterated until convergence. The authors found, however, that in application this approach is not always stable and so it is not recommended.

TABLE A1 Summary of earthquake ground motion data considered in this study. The event name and event index are from the NGA-West2 database[30]. The number of usable stations and posterior mean r values are for $SA(1s)$ data.

Event name	Year	Event index	Magnitude	# of usable stations	Estimated $\sigma_{\hat{r}_i \mid r_i}$	Posterior mean r_i [km]
Chuetsu-oki	2007	278	6.8	542	6.0	28.1
Niigata, Japan	2004	180	6.6	464	5.9	55.7
Chi-Chi, Taiwan	1999	137	7.6	371	5.6	53.5
Tottori, Japan	2000	176	6.6	338	6.0	58.4

TABLE A1 Summary of earthquake ground motion data considered in this study. The event name and event index are from the NGA-West2 database[30]. The number of usable stations and posterior mean r values are for $SA(1s)$ data.

Event name	Year	Event index	Magnitude	# of usable stations	Estimated $\sigma_{\hat{r}_i \mid r_i}$	Posterior mean r_i [km]
Iwate	2008	279	6.9	332	6.7	58.7
Chi-Chi, Taiwan-05	1999	174	6.2	306	5.8	23.5
El Mayor-Cucapah	2010	280	7.2	290	7.8	27.0
Chi-Chi, Taiwan-02	1999	171	5.9	280	6.7	55.9
Chi-Chi, Taiwan-06	1999	175	6.3	280	5.6	27.9
10370141	2009	1018	4.5	245	8.1	29.4
Chi-Chi, Taiwan-03	1999	172	6.2	238	6.3	78.5
Chi-Chi, Taiwan-04	1999	173	6.2	238	6.4	36.3
10275733	2007	1028	4.7	196	7.7	21.8
14312160	2007	1019	4.7	189	13.5	27.0
Anza-02	2001	163	4.9	186	8.3	24.1
14383980	2008	1002	5.4	184	8.3	50.8
10410337	2009	1011	4.7	169	9.9	34.1
40199209	2007	1045	4.2	161	9.8	34.4
40204628	2007	1001	5.5	152	9.9	41.2
71336726	2010	1221	4.0	152	9.5	19.8
Northridge-01	1994	127	6.7	147	10.9	18.3
14138080	2005	1014	4.6	139	14.0	12.1
21522424	2006	1021	4.3	138	8.5	6.2
Parkfield-02, CA	2004	179	6.0	130	19.1	17.1
51207740	2008	1051	4.1	130	12.1	38.4
14151344	2005	1003	5.2	126	11.6	25.9
14186612	2005	1012	4.7	126	13.7	12.7
Hector Mine	1999	158	7.1	125	13.1	34.7
9753485	2002	1015	4.2	122	13.6	11.3
Yorba Linda	2002	167	4.3	118	9.3	32.2
14095628	2004	1006	5.0	118	15.6	85.8
9983429	2004	1016	4.3	118	13.6	12.1
51177644	2007	1101	3.7	117	9.9	16.7
21305648	2003	1060	4.0	112	10.6	19.0
Whittier Narrows-01	1987	113	6.0	111	11.7	26.9
21530368	2006	1023	4.5	110	10.9	58.7
21437727	2005	1050	4.2	109	10.3	20.9
14155260	2005	1007	4.9	107	12.2	42.1
9941081	2003	1066	3.9	106	15.6	36.2
9173365	2001	1035	4.3	105	17.2	35.6
Wenchuan, China	2008	277	7.9	103	17.0	82.0
30226086	2003	1059	4.0	103	10.7	32.7
40194055	2007	1046	4.2	99	14.4	17.2
21465580	2005	1008	4.8	97	8.7	7.3
21510121	2006	1079	3.7	97	10.9	20.5
10059745	2004	1064	4.2	95	15.5	44.5
21339029	2004	1096	3.6	91	11.0	17.5
21502994	2006	1125	3.6	90	12.7	31.0
14077668	2004	1038	4.3	89	16.5	40.6
Big Bear City	2003	170	4.9	87	18.6	37.9

TABLE A1 Summary of earthquake ground motion data considered in this study. The event name and event index are from the NGA-West2 database[30]. The number of usable stations and posterior mean r values are for $SA(1s)$ data.

Event name	Year	Event index	Magnitude	# of usable stations	Estimated $\sigma_{\hat{r}_i \mid r_i}$	Posterior mean r_i [km]
51203888	2008	1138	3.5	87	10.3	28.3
14201764	2005	1075	4.2	86	16.0	37.7
10403777	2009	1025	4.4	85	14.4	30.1
21414391	2004	1098	3.7	85	9.9	24.0
14295640	2007	1020	4.3	84	19.1	14.9
21422178	2004	1022	4.3	83	11.6	8.2
51203773	2008	1070	4.0	83	11.5	28.3
14376612	2008	1113	4.0	83	18.9	26.3
9970349	2003	1152	3.5	83	13.6	10.6
14519780	2009	1186	5.2	83	19.2	14.7
21350824	2004	1031	4.2	82	11.7	19.6
30225187	2002	1069	3.9	82	12.1	27.7
Darfield, New Zealand	2010	281	7.0	80	22.0	40.3
Christchurch, New Zealand	2011	346	6.2	80	18.9	37.4
14239184	2006	1108	3.9	80	20.7	15.8
Landers	1992	125	7.3	76	16.3	51.5
51183708	2007	1032	4.2	75	12.0	49.0
21455182	2005	1044	4.1	74	12.0	45.2
9644101	2001	1128	3.6	74	16.7	29.3
30225889	2003	1043	4.1	73	11.8	8.4
Loma Prieta	1989	118	6.9	71	10.7	29.9
21262721	2003	1030	4.3	71	12.5	16.4
14118096	2005	1054	4.3	70	17.7	23.2
9652545	2001	1095	3.8	70	24.9	18.4
14242516	2006	1140	3.7	69	17.3	44.2
10321561	2008	1086	4.2	68	18.2	13.9
30226452	2003	1137	3.5	68	15.2	11.5
13692644	2002	1103	3.7	67	22.0	18.7
21335949	2004	1130	3.7	67	13.7	24.3
40193843	2007	1170	3.4	67	13.1	18.0
10299017	2008	1094	3.9	66	30.7	23.1
Whittier Narrows-02	1987	114	5.3	65	12.1	23.3
San Simeon, CA	2003	177	6.5	65	15.6	45.2
21266207	2003	1049	4.0	65	14.2	31.1
10972299	2001	1076	3.8	64	22.1	16.8
14146956	2005	1062	4.1	62	24.0	17.9
51177794	2007	1160	3.4	62	13.3	33.6
10477949	2009	1178	4.0	62	25.5	25.3
10067405	2004	1132	3.6	61	16.9	59.0
10249565	2007	1110	3.9	60	23.3	17.5
51182810	2007	1013	4.6	59	15.5	39.7
14403732	2008	1052	4.1	59	19.4	14.9
14366244	2008	1134	3.6	57	21.6	16.4
14295984	2007	1135	3.7	57	21.7	18.5
10216101	2006	1146	3.6	56	22.7	69.1
21397674	2004	1097	3.7	55	14.0	29.7

TABLE A1 Summary of earthquake ground motion data considered in this study. The event name and event index are from the NGA-West2 database[30]. The number of usable stations and posterior mean r values are for $SA(1s)$ data.

Event name	Year	Event index	Magnitude	# of usable stations	Estimated $\sigma_{\hat{r}_i \mid r_i}$	Posterior mean r_i [km]
14137160	2005	1085	3.9	54	24.2	18.0
21261124	2002	1123	3.6	54	12.4	30.8
10276197	2007	1068	4.1	53	25.6	19.1
14204720	2005	1139	3.6	52	27.7	26.7
40234037	2009	1033	4.3	51	13.9	10.4
40219463	2008	1081	3.8	51	17.8	37.1
14216544	2006	1154	3.5	51	27.0	48.7
9069997	1998	1034	4.5	50	27.0	29.1
14330056	2007	1053	4.3	50	35.9	28.2
10207681	2006	1089	4.0	49	24.2	64.8
14520900	2009	1190	4.2	49	38.4	46.1
14355256	2008	1264	3.4	49	22.5	25.0
L'Aquila, Italy	2009	274	6.3	47	28.1	26.4
9064093	1998	1027	4.8	47	26.6	38.1
21526081	2006	1099	3.7	47	18.8	24.3
21549979	2006	1100	3.7	47	17.6	23.4
10205997	2006	1163	3.6	47	22.9	28.4
14169456	2005	1048	4.3	46	35.5	23.0
Big Bear-01	1992	126	6.5	45	15.7	32.9
Coalinga-01	1983	76	6.4	44	19.4	16.0
14295380	2007	1118	4.1	44	26.0	25.8
Big Bear-02	2001	161	4.5	43	19.9	20.9
9171679	2000	1042	4.4	43	25.0	34.8
30192424	1998	1061	4.1	43	31.6	44.9
14079184	2004	1107	3.8	43	25.8	20.8
10285533	2007	1065	4.2	42	25.2	19.6
9950169	2003	1151	3.5	42	29.8	39.4
San Fernando	1971	30	6.6	41	19.9	15.2
9655209	2001	1091	3.8	41	28.9	20.4
9986489	2004	1131	3.6	41	31.0	21.2
14039128	2004	1145	3.5	41	28.1	20.0
10025757	2004	1166	3.4	41	22.1	39.3
14285852	2007	1156	3.6	40	25.4	18.7

References

1. Boore DM, Gibbs JF, Joyner WB, Tinsley JC, Ponti DJ. Estimated Ground Motion From the 1994 Northridge, California, Earthquake at the Site of the Interstate 10 and La Cienega Boulevard Bridge Collapse, West Los Angeles, California. *Bulletin of the Seismological Society of America* 2003; 93(6): 2737–2751.

2. Wang M, Takada T. Macrospatial Correlation Model of Seismic Ground Motions. *Earthquake Spectra* 2005; 21(4): 1137–1156.

3. Foulser-Piggott R, Stafford PJ. A Predictive Model for Arias Intensity at Multiple Sites and Consideration of Spatial Correlations. *Earthquake Engineering & Structural Dynamics* 2012; 41(3): 431–451. doi: 10.1002/eqe.1137

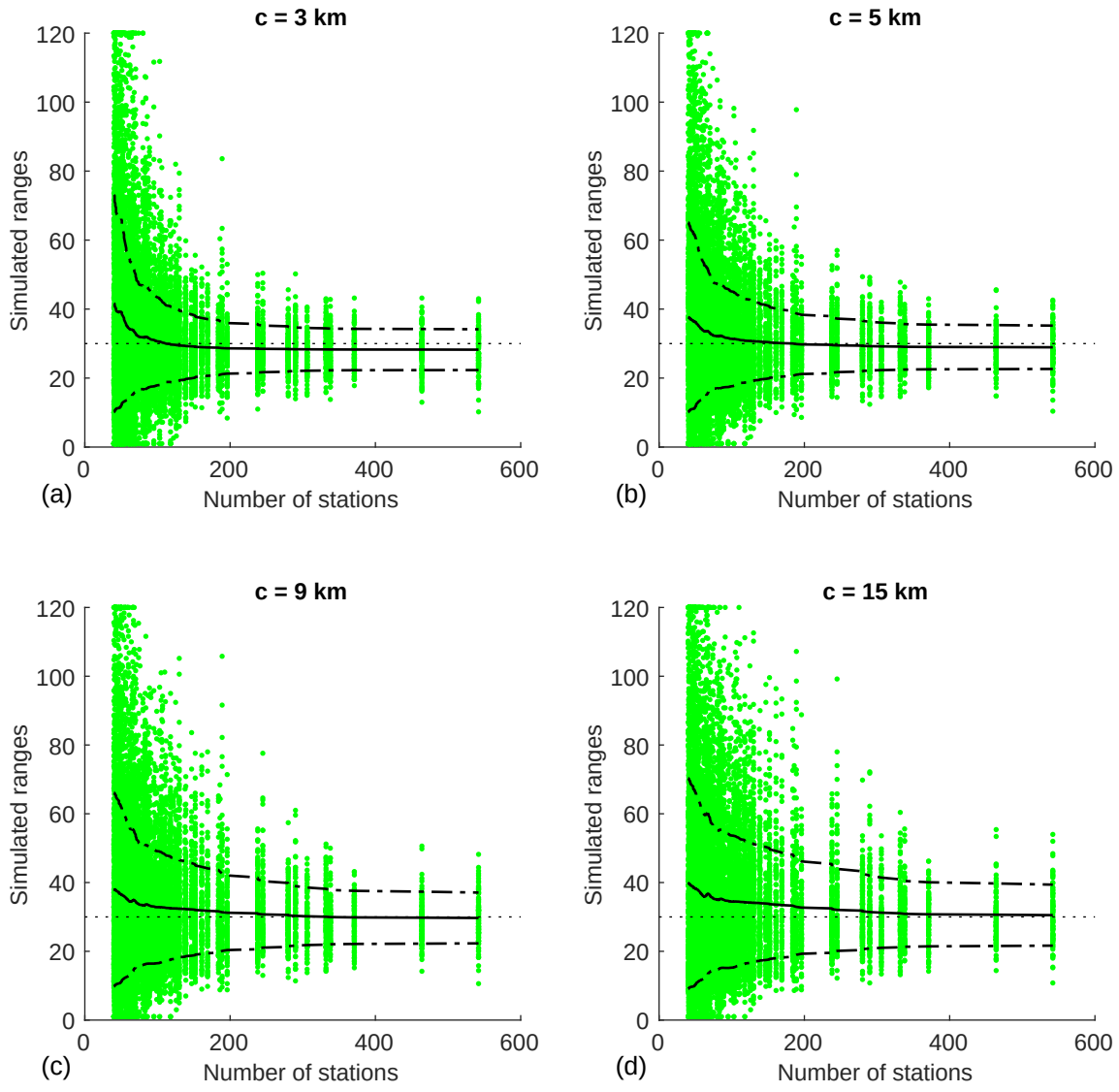

FIGURE A1 Plots of fitting performance for several choices of c, in the case where $r = 30km$.

4. Goda K, Atkinson GM. Intraevent Spatial Correlation of Ground-Motion Parameters Using SK-Net Data. *Bulletin of the Seismological Society of America* 2010; 100(6): 3055–3067. doi: 10.1785/0120100031

5. Goda K. Interevent Variability of Spatial Correlation of Peak Ground Motions and Response Spectra. *Bulletin of the Seismological Society of America* 2011; 101(5): 2522–2531. doi: 10.1785/0120110092

6. Esposito S, Iervolino I. PGA and PGV Spatial Correlation Models Based on European Multievent Datasets. *Bulletin of the Seismological Society of America* 2011; 101(5): 2532–2541. doi: 10.1785/0120110117

7. Loth C, Baker JW. A Spatial Cross-Correlation Model for Ground Motion Spectral Accelerations at Multiple Periods. *Earthquake Engineering & Structural Dynamics* 2013; 42(3): 397–417. doi: DOI: 10.1002/eqe.2212

8. Wesson RL, Perkins DM. Spatial Correlation of Probabilistic Earthquake Ground Motion and Loss. *Bulletin of the Seismological Society of America* 2001; 91(6): 1498–1515.

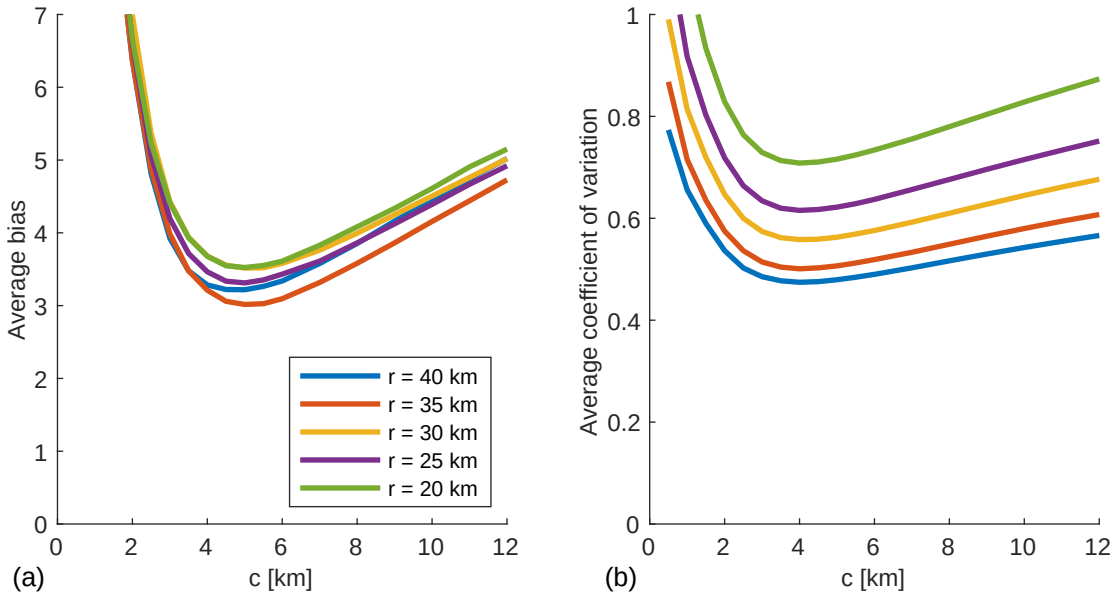

FIGURE A2 (a) Average bias and (b) average coefficient of variation for range estimates using weighted least squares fitting, as a function of the c value used in the fitting function. Each line represents errors when simulated data was produced with an assumed range (r) value.

9. Lee R, Kiremidjian AS. Uncertainty and Correlation for Loss Assessment of Spatially Distributed Systems. *Earthquake Spectra* 2007; 23(4): 753–770.

10. Adachi T, Ellingwood B. Impact of Infrastructure Interdependency and Spatial Correlation of Seismic Intensities on Performance Assessment of a Water Distribution System. In: ; 2007; Tokyo, Japan.

11. Park J, Bazzurro P, Baker JW. Modeling Spatial Correlation of Ground Motion Intensity Measures for Regional Seismic Hazard and Portfolio Loss Estimation. In: ; 2007; Tokyo, Japan: 8p.

12. Shiraki N, Shinozuka M, Moore JE, Chang SE, Kameda H, Tanaka S. System Risk Curves: Probabilistic Performance Scenarios for Highway Networks Subject to Earthquake Damage. *Journal of Infrastructure Systems* 2007; 13(1): 43–54.

13. Goda K, Hong HP. Spatial Correlation of Peak Ground Motions and Response Spectra. *Bulletin of the Seismological Society of America* 2008; 98(1): 354–365. doi: 10.1785/0120070078

14. Jayaram N, Baker JW. Correlation Model for Spatially Distributed Ground-Motion Intensities. *Earthquake Engineering & Structural Dynamics* 2009; 38(15): 1687–1708. doi: 10.1002/eqe.922

15. Sokolov V, Wenzel F, Wen KL, Jean WY. On the Influence of Site Conditions and Earthquake Magnitude on Ground-Motion within-Earthquake Correlation: Analysis of PGA Data from TSMIP (Taiwan) Network. *Bulletin of Earthquake Engineering* 2012; 10(5): 1401–1429. doi: 10.1007/s10518-012-9368-5

16. Heresi P, Miranda E. Uncertainty in Intraevent Spatial Correlation of Elastic Pseudo-Acceleration Spectral Ordinates. *Bulletin of Earthquake Engineering* 2019; 17(3): 1099–1115. doi: 10.1007/s10518-018-0506-6

17. Schiappapietra E, Douglas J. Spatial Correlation of Ground Motions in the 2016-2017 Central Italy Seismic Sequence. In: ; 2019; Greenwich, London, UK: 10p.

18. Sokolov V, Wenzel F. Further Analysis of the Influence of Site Conditions and Earthquake Magnitude on Ground-Motion within-Earthquake Correlation: Analysis of PGA and PGV Data from the K-NET and the KiK-Net (Japan) Networks. *Bulletin of Earthquake Engineering* 2013: 1–18. doi: 10.1007/s10518-013-9493-9

19. Foulser-Piggott R, Goda K. Ground-Motion Prediction Models for Arias Intensity and Cumulative Absolute Velocity for Japanese Earthquakes Considering Single-Station Sigma and within-Event Spatial Correlation. *Bulletin of the Seismological Society of America* 2015; 105(4): 1903–1918.

20. Schiappapietra E, Douglas J. Modelling the Spatial Correlation of Earthquake Ground Motion: Insights from the Literature, Data from the 2016–2017 Central Italy Earthquake Sequence and Ground-Motion Simulations. *Earth-Science Reviews* 2020: 103139. doi: 10.1016/j.earscirev.2020.103139

21. Esposito S, Iervolino I. Spatial Correlation of Spectral Acceleration in European Data. *Bulletin of the Seismological Society of America* 2012; 102(6): 2781–2788. doi: 10.1785/0120120068

22. Markhvida M, Ceferino L, Baker JW. Modeling Spatially Correlated Spectral Accelerations at Multiple Periods Using Principal Component Analysis and Geostatistics. *Earthquake Engineering & Structural Dynamics* 2018; 47(5): 1107–1123. doi: 10.1002/eqe.3007

23. García-Soidán P, Menezes R, Rubiños Ó. Bootstrap Approaches for Spatial Data. *Stochastic Environmental Research and Risk Assessment* 2014; 28(5): 1207–1219. doi: 10.1007/s00477-013-0808-9

24. Clark RG, Allingham S. Robust Resampling Confidence Intervals for Empirical Variograms. *Mathematical Geosciences* 2011; 43(2): 243–259. doi: 10.1007/s11004-010-9314-5

25. Cressie N. Fitting Variogram Models by Weighted Least Squares. *Journal of the International Association for Mathematical Geology* 1985; 17(5): 563–586.

26. Zimmerman DL, Zimmerman MB. A Comparison of Spatial Semivariogram Estimators and Corresponding Ordinary Kriging Predictors. *Technometrics* 1991; 33(1): 77–91.

27. Cressie NA. *Statistics for Spatial Data.* New York: John Willy and Sons . 1993.

28. Kerby B. *Semivariogram Estimation: Asymptotic Theory and Applications*. The University of Utah . 2016.

29. Jayaram N, Baker JW. Statistical Tests of the Joint Distribution of Spectral Acceleration Values. *Bulletin of the Seismological Society of America* 2008; 98(5): 2231–2243. doi: 10.1785/0120070208

30. Ancheta TD, Darragh RB, Stewart JP, et al. NGA-West2 Database. *Earthquake Spectra* 2014; 30(3): 989–1005. doi: 10.1193/070913EQS197M

31. Chiou BSJ, Youngs RR. Update of the Chiou and Youngs NGA Model for the Average Horizontal Component of Peak Ground Motion and Response Spectra. *Earthquake Spectra* 2014; 30(3): 1117–1153. doi: 10.1193/072813EQS219M

32. Jayaram N, Baker JW. Considering Spatial Correlation in Mixed-Effects Regression, and Impact on Ground-Motion Models. *Bulletin of the Seismological Society of America* 2010; 100(6): 3295–3303. doi: 10.1785/0120090366

33. Journel AG, Huijbregts CJ. *Mining Geostatistics*. Academic press . 1978.

34. Cressie N, Hawkins DM. Robust Estimation of the Variogram: I. *Journal of the International Association for Mathematical Geology* 1980; 12(2): 115–125.

35. Pardo-Iguzquiza E, Chica-Olmo M. Geostatistics with the Matern Semivariogram Model: A Library of Computer Programs for Inference, Kriging and Simulation. *Computers & Geosciences* 2008; 34(9): 1073–1079. doi: 10.1016/j.cageo.2007.09.020

36. Goda K, Atkinson GM. Probabilistic Characterization of Spatially Correlated Response Spectra for Earthquakes in Japan. *Bulletin of the Seismological Society of America* 2009; 99(5): 3003–3020. doi: 10.1785/0120090007

37. Wagener T, Goda K, Erdik M, Daniell J, Wenzel F. A Spatial Correlation Model of Peak Ground Acceleration and Response Spectra Based on Data of the Istanbul Earthquake Rapid Response and Early Warning System. *Soil Dynamics and Earthquake Engineering* 2016; 85: 166–178. doi: 10.1016/j.soildyn.2016.03.016

38. Pebesma E, Heuvelink G. Spatio-Temporal Interpolation Using Gstat. *The R Journal* 2016; 8(1): 204–218.

39. Stafford PJ, Zurek BD, Ntinalexis M, Bommer JJ. Extensions to the Groningen Ground-Motion Model for Seismic Risk Calculations: Component-to-Component Variability and Spatial Correlation. *Bulletin of Earthquake Engineering* 2018. doi: 10.1007/s10518-018-0425-6

40. Sgobba S, Lanzano G, Pacor F, et al. Spatial Correlation Model of Systematic Site and Path Effects for Ground-Motion Fields in Northern Italy. *Bulletin of the Seismological Society of America* 2019; 109(4): 1419–1434.

41. Huang C, Galasso C. Ground-Motion Intensity Measure Correlations Observed in Italian Strong-Motion Records. *Earthquake Engineering & Structural Dynamics* 2019; 48(15): 1634–1660. doi: 10.1002/eqe.3216

42. Du W, Wang G. Intra-Event Spatial Correlations for Cumulative Absolute Velocity, Arias Intensity, and Spectral Accelerations Based on Regional Site ConditionsIntra-Event Spatial Correlations for Cumulative Absolute Velocity, Arias Intensity, and Spectral Accelerations. *Bulletin of the Seismological Society of America* 2013; 103(2A): 1117–1129. doi: 10.1785/0120120185

43. Infantino M, Paolucci R, Smerzini C, Stupazzini M. Study of the Spatial Correlation of Earthquake Ground Motion by Means of Physics-Based Numerical Scenarios. In: ; 2018: 12p.

44. Deutsch CV, Journel AG. *GSLIB Geostatistical Software Library and User's Guide*. Version 2.0. New York: Oxford University Press . 1997.

45. Hong HP, Zhang Y, Goda K. Effect of Spatial Correlation on Estimated Ground-Motion Prediction Equations. *Bulletin of the Seismological Society of America* 2009; 99(2A): 928–934. doi: 10.1785/0120080172

46. Ming D, Huang C, Peters GW, Galasso C. An Advanced Estimation Algorithm for Ground-Motion Models with Spatial Correlation. *Bulletin of the Seismological Society of America* 2019; 109(2): 541–566. doi: 10.1785/0120180215

47. Wang G, Du W. Spatial Cross-Correlation Models for Vector Intensity Measures (PGA, Ia, PGV, and SAs) Considering Regional Site Conditions. *Bulletin of the Seismological Society of America* 2013; 103(6): 3189–3204. doi: 10.1785/0120130061

48. Bromiley P. Products and Convolutions of Gaussian Probability Density Functions. Tech. Rep. Tina Memo No. 2003-003, University of Manchester; Manchester, United Kingdom: 2003.

49. Chen Y, Baker JW. Spatial Correlations in CyberShake Physics-Based Ground-Motion Simulations. *Bulletin of the Seismological Society of America* 2019; 109(6): 2447–2458. doi: 10.1785/0120190065

50. Landwehr N, Kuehn NM, Scheffer T, Abrahamson N. A Nonergodic Ground-Motion Model for California with Spatially Varying Coefficients. *Bulletin of the Seismological Society of America* 2016; 106(6): 2574–2583. doi: 10.1785/0120160118

51. Phung VB, Loh CH, Chao SH, Abrahamson NA. Ground Motion Prediction Equation for Taiwan Subduction Zone Earthquakes. *Earthquake Spectra* 2020(in press): 8755293020906829. doi: 10.1177/8755293020906829

52. Wessel P, Smith WHF, Scharroo R, Luis J, Wobbe F. Generic Mapping Tools: Improved Version Released. *Eos, Transactions American Geophysical Union* 2013; 94(45): 409–410. doi: 10.1002/2013EO450001