THE SEISMOLOGICAL SOCIETY OF AMERICA
400 Evelyn Ave., Suite 201
Albany, CA 94706-1375
(510) 525-5474; FAX (510) 525-7204
www.seismosoc.org

A Non-Ergodic Ground-Motion Model for California with Spatially Varying Coefficients

Niels Landwehr*[1], Nicolas M. Kuehn†[2], Tobias Scheffer‡[1], and Norman Abrahamson§[2]

[1] *Department of Computer Science, University of Potsdam*

[2] *Pacific Earthquake Engineering Research Center, University of California, Berkeley*

submitted to *Bulletin of the Seismological Society of America*

Abstract

Traditional probabilistic seismic hazard analysis (PSHA), as well es the estimation of ground-motion models (GMMs), is based on the ergodic assumption, which means that the distribution of ground motions over time at given site is the same as their spatial distribution over all sites for the same magnitude, distance, and site condition. With a large increase in the number of recorded ground-motion data, there are now repeated observations at given sites and from multiple earthquakes in small regions, so that assumption can be relaxed. We use a novel approach to develop a non-ergodic GMM, which is cast as a varying coefficients model (VCM). In this model, the coefficients are allowed to vary by geographical location, which makes it possible to incorporate effects of spatially varying

*landwehr@cs.uni-potsdam.de
†kuehn@berkeley.edu
‡scheffer@cs.uni-potsdam.de
§abrahamson@berkeley.edu

source, path and site conditions. Hence, a separate set of coefficients is estimated for each source and site coordinate in the data set. The coefficients are constrained to be similar for spatially nearby locations. This is achieved by placing a Gaussian process prior on the coefficients. The amount of correlation is determined by the data. The spatial correlation structure of the model allows one to extrapolate the varying coefficients to a new location and trace the corresponding uncertainties. The approach is illustrated with the NGA West2 data set, using only Californian records. The VCM outperforms a traditionally estimated GMM in terms of generalization error, and leads to a reduction in the aleatory standard deviation by about 40%, which has important implications for seismic hazard calculations. The scaling of the model with respect to its predictor variables, such as magnitude and distance, is physically plausible. The epistemic uncertainty associated with the predicted ground motions is small in places where events or stations are close and large where data are sparse.

Introduction

Probabilistic seismic hazard analysis (PSHA) estimates the expected future distribution of a ground-motion parameter of interest at a specific site. To this end, it is important to have a model that accurately predicts ground motion given source, path, and site related parameters such as magnitude and distance. This is usually done by a ground-motion model (GMM), which is an estimate of the conditional distribution of the ground-motion parameter of interest given magnitude, distance, and other parameters.

On the other hand, there are also GMMs that are developed on smaller, regional data sets (e.g., for Greece (Danciu and Tselentis, 2007), Italy (Bindi et al., 2011), the Eastern Alps (Bragato and Slejko, 2005), and Turkey (Akkar and Cagnan, 2010)), though these suffer from smaller numbers of data points. Remedies have been proposed that attempt to directly estimate regionally-varying GMMs from a larger data set by constraining the coefficients to be similar across regions (Gianniotis et al., 2014; Stafford, 2014).

A GMM with different coefficients for different regions is a first step in removing the assumption of ergodicity from PSHA (Stafford, 2014), which states that the conditional distribution of the ground-motion parameter of interest at a given site is identical to the conditional distribution at any other site, given the same magnitude, distance, and site conditions (Anderson and Brune, 1999). For a fully non-ergodic PSHA, regional adjustments are broken down into smaller and smaller geographical units, assuming that there are repeatable source, path, and site effects for different locations, which, in principle, can be known and estimated. This reduces the overall value of aleatory variability (Al-Atik et al., 2010; Lin et al., 2011). Basically, one trades apparent aleatory variability against epistemic uncertainty, which has large effects on the resulting hazard curve distribution (Kuehn and Abrahamson, 2015). Typically, these repeatable effects are estimated from residuals and then added to the median GMM prediction as adjustment terms (Anderson and Uchiyama, 2011; Douglas and Aochi, 2016; Villani and Abrahamson, 2015). The repeatable effects are spatially correlated (Jayaram and Baker, 2009; Lin et al., 2011; Walling, 2009), which makes it possible to estimate them from a limited amount of data.

In this paper, we develop a GMM in which the coefficients of the model can vary smoothly with geographical location. Hence, an individual model for each location is estimated, but the coefficients are enforced to be similar for nearby locations. The model is cast as a varying-coefficient model (Bussas et al., 2015; Gelfand et al., 2003), which places a Gaussian-process (GP) prior (Rasmussen and Williams, 2006) over the coefficients. Since there is not enough data to estimate independent models for each location, the GP prior is used to constrain the coefficients based on their spatial correlation. In the varying-coefficient model, the spatial correlation structure of the coefficients also introduces spatial random effects for both events and stations; the amount of smoothing and variability in the coefficients is determined by the data.

The model is developed and evaluated on a subset of the NGA West 2 dataset (Ancheta et al., 2014), based on Californian data used by Abrahamson et al. (Abrahamson et al., 2014). In California, regional differences between Northern California and Southern California have been found previously (Atkinson and Morrison, 2009; Chiou et al., 2010), though the recent NGA West 2 models treat California as a whole.

Non-Ergodic Seismic Hazard

Traditionally, GMMs are developed under the assumption of ergodicity, which means that the conditional distribution of the ground-motion parameter of interest at any site, given the predictor variables, is the same as the conditional ground-motion distribution at any other site (Anderson and Brune, 1999). Typically, the distribution is assumed to be log-normal, and functions for both the median and the variance of this distribution are estimated, depending on predictor variables such as magnitude, distance, and other parameters. The variance is typically partitioned into a between-event and within-event term. Hence, the typical form of a GMM is

$$y = f(\boldsymbol{x}) + \eta_e \tau + \epsilon_{es} \phi \tag{1}$$

where y is the ground-motion parameter of interest from event e recorded at station s, $\boldsymbol{x}$ is a vector of predictor variables, τ and ϕ are the between-event and within-event standard deviation, respectively, and η and ϵ are normally distributed with mean zero and standard deviation one. It has been recognized that the variance can be further partitioned, to account for repeatable source, path, and site effects (Al-Atik et al., 2010; Anderson and Brune, 1999). In particular, the use of single-station sigma (Atkinson, 2006; Rodriguez-Marek et al., 2011) has been used in hazard studies for critical facilities in recent years.

Hence, a GMM becomes

$$y = f(\boldsymbol{x}) + \eta_e \tau_0 + \epsilon_{es} \phi_0 + \lambda_s \phi_{S2S} + \delta_e \tau_S + \xi_{es} \phi_P \tag{2}$$

where τ_S, ϕ_{S2S} and ϕ_P are the standard deviations of repeatable source, site and path effects, respectively (e.g., Al-Atik et al. (2010); Villani and Abrahamson (2015)). The adjustment terms, $\lambda_s \phi_{S2S}$, $\delta_e \tau_S$, and $\xi_{es} \phi_P$, can be estimated from observed residuals or simulations (Anderson and Uchiyama, 2011; Douglas and Aochi, 2016; Villani and Abrahamson, 2015). In Equation (2), only τ_0 and ϕ_0 describe aleatory variability, while the other variance components are part of epistemic uncertainty. In the absence of information about the repeatable effects, their uncertainty needs to be taken into account. The resulting mean hazard curve is the same as when calculating hazard with the ergodic assumption; however, the fractiles of the hazard curve distribution are different. If some of the repeatable effects are known, this reduces their respective uncertainty but changes the median prediction. Removing the ergodic assumption from calculating seismic hazard may have a large impact on hazard results because the reduction in the aleatory variability is about 50% (see Table 5 of Lin et al. (2011)), which has large effects on the resulting hazard calculations (Bommer and Abrahamson, 2006). See Villani and Abrahamson (2015) for an example calculation of non-ergodic seismic hazard.

The repeatable source, path, and site effects $\delta_e \tau_S$, $\xi_{es} \phi_P$, and $\lambda_s \phi_{S2S}$ are different for different locations. Hence, the median predictions of a non-ergodic GMM vary spatially, because they are obtained by incorporating the epistemic terms according to Equation (2). The repeatable effects are spatially correlated (Jayaram and Baker, 2009; Lin et al., 2011; Walling, 2009).

In the present work, we take a slightly different approach on the estimation of a GMM suitable for non-ergodic seismic hazard. We estimate a model where the coefficients vary smoothly with geographical location. Thus, we put the repeatable source, path, and site effects into the coefficients, not into an adjustment term to a "global" model (here, global means a model that has constant coefficients over the underlying data set). The model is cast as a varying coefficient model (VCM) (Gelfand et al., 2003). To be able to constrain the coefficients, these are assumed to be spatially correlated, which is achieved by placing a Gaussian process prior over them (Bussas et al., 2015). This basically smoothes the coefficients spatially, where the amount of smoothing is determined by the data. The model is explained in more detail later.

Data

We use the same subset of the NGA West 2 data set as Abrahamson et al. (2014), who also describe the selection process in more detail. We use only the data from California and Nevada, since data from other regions will be spatially uncorrelated to this region. Figure 1 shows a map of California, together with the event and station locations in the data set. In total, there are 10,692 records from 221 earthquakes, recorded at 1425 stations. The magnitude/distance distribution is shown in Figure 2. The ground motion parameter of interest is logarithmic horizontal peak ground acceleration (PGA, equal to spectral acceleration at a period T = 0.01s) and logarithmic spectral acceleration at periods T = 0.02, 0.05, 0.1, 0.2, 0.5, 1, and 4s

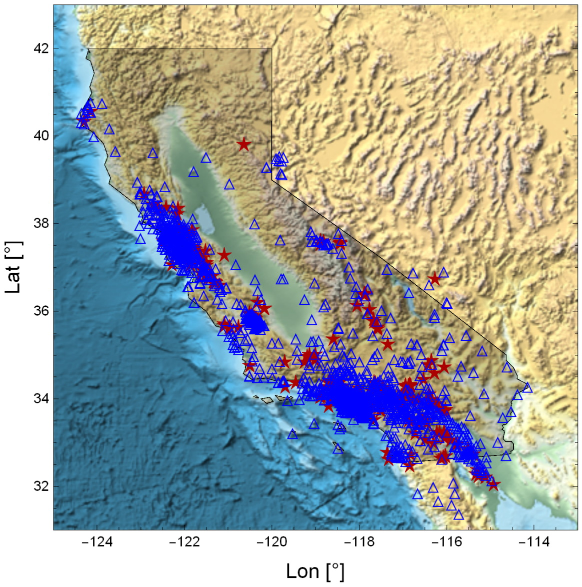

Figure 1: Map of earthquakes and seismograph stations. Stars show event locations, and triangles the station locations. The color version of this figure is available only in the electronic edition.

in units of g.

Model

As discussed in Sections *Introduction* and *Non-Ergodic Seismic Hazard*, it can be useful to view a GMM as a model that varies continuously on a spatial scale. Building on techniques presented by Bussas et al. (2015), this section presents a model in which the coefficients are spatially dependent, but also spatially correlated, corresponding to the assumption that coefficient values change smoothly in space.

We have a data set of N pairs $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)$ of inputs $\boldsymbol{x}_i$ and outputs y_i, where the outputs y_i are measurements of the ground-motion parameter of interest; in our case, response spectral ordinates at different periods. Each input $\boldsymbol{x}_i$ is a vector $\boldsymbol{x}_i = [\boldsymbol{M}, R_{JB}, V_{S30}, F]$, comprising the magnitude, Joyner-Boore distance, time-averaged shear wave velocity in the upper 30ms, and the style of faulting for the i-th recording. We model the dependency of the

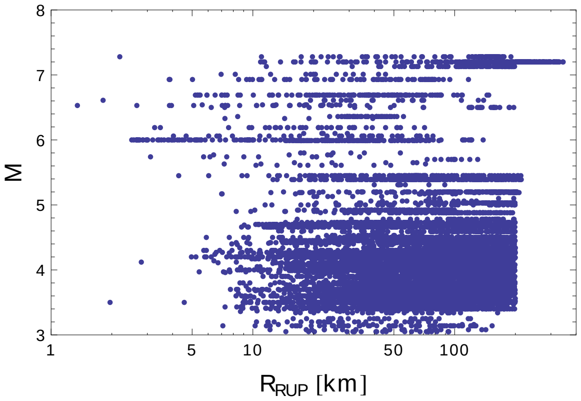

Figure 2: Magnitude and R_{JB}-distance scatterplot of the data used in this study.

outputs y_i on the inputs as

$$
\begin{aligned}
y &= f(\boldsymbol{\beta}; \boldsymbol{x}) + \epsilon && (3)\\
&= \beta_0 + \beta_1 \boldsymbol{M} + \beta_2 \boldsymbol{M}^2 + (\beta_3 + \beta_4 \boldsymbol{M}) \ln \sqrt{R_{JB}^2 + h^2} + \beta_5 R_{JB} \\
&\quad + \beta_6 \ln V_{S30} + \beta_7 F_R + \beta_8 F_{NM} + \epsilon && (4)
\end{aligned}
$$

where F_{NM} and F_R are indicator variables that take the value of one for normal and reverse faulting, respectively, and zero otherwise; ϵ denotes the overall residual. Equation (4) is a simple functional form, but can nevertheless capture the main characteristics of ground-motion scaling with the predictor variables.

In Section *Non-Ergodic Seismic Hazard*, we have seen that by means of the adjustment terms that account for repeatable source, path, and site effects (cf. Equation (2)), ground-motion predictions used in PSHA become essentially spatially dependent. In our model, the spatial dependency is expressed by assuming that the coefficients $\boldsymbol{\beta}$ of the model in Equation (4) are a function of spatial location. For each ground-motion record, there are two coordinates available: the event latitude and longitude, and the station latitude and longitude. We denote these by $\boldsymbol{t}_e \in \mathbb{R}^2$ and $\boldsymbol{t}_s \in \mathbb{R}^2$, respectively. For event coordinates $\boldsymbol{t}_e$, we use the horizontal projection of the (geographical) center of the rupture, estimated from the NGA West 2 source flatfile (Ancheta et al., 2014). The coefficients of Equation (4) are now modeled as (partially) depending on the coordinates $\boldsymbol{t}_e$ and $\boldsymbol{t}_s$. Specifically, we revise Equation (4) to

$$\begin{aligned} y = {} & \beta_{-1}(\boldsymbol{t}_e) + \beta_0(\boldsymbol{t}_s) + \beta_1 \boldsymbol{M} + \beta_2 \boldsymbol{M}^2 + (\beta_3(\boldsymbol{t}_e) + \beta_4 \boldsymbol{M}) \ln \sqrt{R_{JB}^2 + h^2} + \beta_5(\boldsymbol{t}_e) R_{JB} \\ & + \beta_6(\boldsymbol{t}_s) \ln V_{S30} + \beta_7 F_R + \beta_8 F_{NM} + \epsilon, \end{aligned} \tag{5}$$

where $\beta_j(\boldsymbol{t}_e)$ denotes a coefficient varying with event location, $\beta_j(\boldsymbol{t}_s)$ denotes a coefficient varying with station location, and β_j denotes a coefficient which does not vary spatially, that is, which is modeled as identical for all locations. The term ϵ is the remaining residual, after all spatial correlations are taken into account, and is distributed normally with mean zero and standard deviation σ_0. Random effects corresponding to event and site terms are modeled via the constant coefficients $\beta_{-1}(\boldsymbol{t}_e)$ and $\beta_0(\boldsymbol{t}_s)$.

The rationale for the chosen dependencies on $\boldsymbol{t}_e$ and $\boldsymbol{t}_s$ are as follows. The coefficients that control the scaling with distance, β_3 and β_5, vary with event location. This corresponds to an average path attenuation for each event. This does not correspond to a single path attenuation, as expressed by the adjustment term $\xi_{es}\phi_P$ in Equation (2). However, most of the records in the data set have a Joyner-Boore distance that is smaller than 200km, so the variation in the distance scaling coefficients corresponds to a small-scale variation in distance attenuation, averaged over all directions. The near-source saturation term is fixed to a value $h = 6$. This ensures that the model is linear in the coefficients. Since there are very few normal faulting events, we fix the coefficient $\beta_8 = -0.1$, as do Abrahamson et al. (2014). For periods $T \geq 1$, the coefficient corresponding to anelastic attenuation is set to zero, $\beta_5 = 0$. Coefficient β_6, which controls the scaling with V_{S30}, varies with station location because V_{S30} as a proxy for site scaling is primarily correlated with local geology under the station. To avoid problems of extrapolating the predictions to large magnitudes, and due to under-representation of different styles of faulting in the data set, the coefficients associated with magnitude and style-of-faulting (SOF) do not vary spatially.

Finally, the constant coefficient β_0 in Equation (4) is split up into two constant coefficients $\beta_{-1}(\boldsymbol{t}_e)$ and $\beta_0(\boldsymbol{t}_s)$ that vary with event coordinates $\boldsymbol{t}_e$ and station coordinates $\boldsymbol{t}_s$, respectively. Because these constant terms are different for different stations and events, they introduce random effects for events and stations into the model. Due to the spatial correlation of coefficients, the random effects are also spatially correlated such that, for example, the random effects of two stations that are very close to each other tend to be similar. Thus, β_{-1} and β_0 capture the repeatable, epistemic part of between-event and within-event variability.

The noise term of the VCM, as described in Equation (5) by the residual ϵ, contains all parts of ground-motion generation that cannot be explained by systematic effects under the parameterization of the model. Thus, it corresponds to the aleatory part, which is quantified by the standard deviation σ_0. Because we implicitly assume that the coefficients vary smoothly between different location, abrupt changes in local geology are not modeled, but smoothed through. Hence, the residual ϵ can still contain some potentially systematic parts.

In the following, we provide a brief, more technical discussion of how to mathematically specify and estimate the varying-coefficient model. The model follows the formulation of a varying-coefficient model by Bussas et al. (2015); here, we briefly restate their main results in the context of the particular model that is given by Equation (5). More details and mathematical background are included in Bussas et al. (2015). According to Equation (5), the model is given by the values of the coefficients $\beta_{-1}, ..., \beta_8$ at all combinations of event and station coordinates $\boldsymbol{t}_e$ and $\boldsymbol{t}_s$; however, in practice, we are only concerned with the values of the coefficients for all coordinates that appear in the data. We denote by $\boldsymbol{t} = (\boldsymbol{t}_e, \boldsymbol{t}_s)$ a combined station and event coordinate, comprising two latitudes and two longitudes. Let $\boldsymbol{t}_1, ..., \boldsymbol{t}_N \in \mathbb{R}^4$ denote the combined station and event coordinates of the N observed data points, and let $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_N \in \mathbb{R}^d$ denote the complete vectors of coefficients at these coordinates as needed for Equation (5), where $d = 10$ is the number of coefficients in the model. The model as described in Equation (5) defines the conditional distribution of an observed response spectral value y_i given the inputs $\boldsymbol{x}_i$ and coefficients $\boldsymbol{\beta}_i$, by

$$y_i \sim p(y|\boldsymbol{x}_i, \boldsymbol{\beta}_i). \tag{6}$$

The spatially varying coefficients are modeled as evaluations of a function $\boldsymbol{\omega} : \mathbb{R}^4 \to \mathbb{R}^d$ that specifies a coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^d$ for any combined event and station coordinate $\boldsymbol{t} \in \mathbb{R}^4$. The function $\boldsymbol{\omega}$ associates any coordinate $\boldsymbol{t}$ with the corresponding values of the model coefficients $\boldsymbol{\beta}$; the model coefficients at the data points are obtained by evaluating this function at the respective coordinates: that is, $\boldsymbol{\beta}_i = \boldsymbol{\omega}(\boldsymbol{t}_i)$. To enforce a spatial correlation between the $\boldsymbol{\beta}_i$, we require the function $\boldsymbol{\omega}$ to be smooth; that is, $\boldsymbol{\omega}(\boldsymbol{t}_i)$ should be close to $\boldsymbol{\omega}(\boldsymbol{t}_l)$ if coordinates $\boldsymbol{t}_i$ are close to coordinates $\boldsymbol{t}_l$. Smoothness over $\boldsymbol{\omega}$ is imposed by modeling $\boldsymbol{\omega}$ as being drawn from a Gaussian-process prior (Rasmussen and Williams, 2006),

$$\boldsymbol{\omega} \sim GP(\mathbf{0}, \boldsymbol{\kappa}). \tag{7}$$

In Equation (7), $\boldsymbol{\kappa} = \boldsymbol{\kappa}(\boldsymbol{t}, \boldsymbol{t'})$ is a kernel function that specifies the correlations between coefficients at the locations $\boldsymbol{t} = (\boldsymbol{t}_e, \boldsymbol{t}_s)$ and $\boldsymbol{t'} = (\boldsymbol{t}'_e, \boldsymbol{t}'_s)$. Because $\boldsymbol{\omega}$ is a vector-valued function, the covariance function is matrix-valued, that is, $\boldsymbol{\kappa}(\boldsymbol{t}, \boldsymbol{t}') \in \mathbb{R}^{d \times d}$; this matrix represents the prior covariances between elements of the vectors $\boldsymbol{\omega}(\boldsymbol{t})$ and $\boldsymbol{\omega}(\boldsymbol{t}')$. Values $\boldsymbol{\kappa}(\mathbf{t}, \mathbf{t'})$ are assumed to be diagonal matrices with diagonal entries $\kappa_1(\boldsymbol{t}, \boldsymbol{t}'), ..., \kappa_d(\boldsymbol{t}, \boldsymbol{t}')$, where $\kappa_j(\boldsymbol{t}, \boldsymbol{t'}) \in \mathbb{R}$ is a scalar-valued kernel function. This means that each dimension of $\boldsymbol{\omega}(\boldsymbol{t})$—corresponding to a particular coefficient—is generated by an independent scalar-valued Gaussian process whose covariance is given by $\kappa_j(\boldsymbol{t}, \boldsymbol{t'}) \in \mathbb{R}$, where j indexes the coefficients. The kernel functions $\kappa_j(\boldsymbol{t}, \boldsymbol{t'})$ are given by

$$\kappa_j(\boldsymbol{t}, \boldsymbol{t'}) = \begin{cases} \theta_j & \text{if } j \in \{1, 2, 4, 7, 8\} \\ \theta_j \exp\left(-\frac{\|\boldsymbol{t}_e - \boldsymbol{t}'_e\|}{\rho_j}\right) + \pi_j & \text{if } j \in \{-1, 3, 5\} \\ \theta_j \exp\left(-\frac{\|\boldsymbol{t}_s - \boldsymbol{t}'_s\|}{\rho_j}\right) + \pi_j & \text{if } j \in \{0, 6\} \end{cases} \tag{8}$$

where θ_j, ρ_j, and π_j are kernel parameters—different for each coefficient—that are determined from data. For coefficients that do not depend on either event or station coordinates ($j \in \{1, 2, 4, 7, 8\}$), the kernel function $\kappa_j(\boldsymbol{t}, \boldsymbol{t'})$ is constant, which implies that any function $\boldsymbol{\omega}$ drawn from the GP prior (Equation (7)) is constant in its j-th dimension. For coefficients that depend on event or station coordinates ($j \in \{-1, 3, 5\}$ or $j \in \{0, 6\}$, respectively), kernel function $\kappa_j(\boldsymbol{t}, \boldsymbol{t'})$ is a Matérn kernel function of degree $\nu = 1/2$ based on the Euclidian distance between event or station coordinates, implying that the j-th dimension of $\boldsymbol{\omega}$ varies with $\boldsymbol{t}_s$ or $\boldsymbol{t}_e$. Parameter π_j is a constant offset to the Matérn kernel. Parameter ρ_j can be understood as a coefficient-specific length scale that determines how rapidly the coefficient changes with location, and θ_j determines how much the coefficient can change. Note that the definition of the VCM given above slightly extends the formulation by Bussas et al. (2015), in which the kernel functions κ_j for the different dimensions are assumed to be identical: that is, $\kappa_j(\boldsymbol{t}, \boldsymbol{t}') = \kappa(\boldsymbol{t}, \boldsymbol{t}')$ for some fixed scalar-valued kernel function $\kappa(\boldsymbol{t}, \boldsymbol{t}') \in \mathbb{R}$. We will discuss how this affects the calculations when applying the model to data.

Given a data set of ground-motion records at different locations with different predictor variables, it is possible to estimate all parameters of the model. These comprise the parameters of the GP (cf. Equation (8)) and the noise term σ_0. The parameters are estimated by optimizing the marginal likelihood (cf. Section 5.2 of Rasmussen and Williams (2006)). We present tables of the parameters of the GP for the different response spectral periods in the electronic supplement.

When applying the model, a prediction of the ground-motion distribution for a new location $\boldsymbol{t}_\star$ with predictor variables $\boldsymbol{x}_\star$ needs to be made. The Gaussian process prior over $\boldsymbol{\omega}$ enables us to derive fully Bayesian predictions from the model. Informally speaking, this works by first computing the conditional distribution of the coefficients at the new location given the data observed at the known locations, and afterwards calculating the median ground-motion prediction using Equation (4) from the new coefficients. The uncertainty associated with the coefficients is translated into additional epistemic uncertainty of the ground-motion predictions.

The full derivation of the predictive distribution at a new location $\boldsymbol{t}_\star$ is presented by Bussas et al. (2015). The main insight from this work is that the VCM model is identical to a standard Gaussian process that uses concatenated inputs $(\boldsymbol{x}_i, \boldsymbol{t}_i)$ and a product kernel function $\bar{\kappa}((\boldsymbol{x}_i, \boldsymbol{t}_i), (\boldsymbol{x}_l, \boldsymbol{t}_l)) = \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_l \kappa(\boldsymbol{t}_i, \boldsymbol{t}_l) = \sum_{j=1}^{d} x_{ij} x_{lj} \kappa(\boldsymbol{t}_i, \boldsymbol{t}_l)$, as stated in Theorem 1 in Bussas et al. (2015). Here, x_{ij} and x_{lj} denote the j-th component of the vectors $\boldsymbol{x}_i$ and $\boldsymbol{x}_l$. When using different kernel functions κ_j for the dimension of $\boldsymbol{\omega}$ instead of a single kernel function κ, the VCM is identical to a standard Gaussian process with kernel function $\bar{\kappa}((\boldsymbol{x}_i, \boldsymbol{t}_i), (\boldsymbol{x}_l, \boldsymbol{t}_l)) = \sum_{j=1}^{d} x_{ij} x_{lj} \kappa_j(\boldsymbol{t}_i, \boldsymbol{t}_l)$, as is easily seen by retracing the derivation of the product kernel in Theorem 1 of Bussas et al. (2015).

We now reproduce this main result of Bussas et al. (2015), taking into account the modification of using dimension-specific kernel functions κ_j. The full ground-motion distribution at a new location $\boldsymbol{t}_\star$, with predictor variables $\boldsymbol{x}_\star$, given the observed data set $\boldsymbol{y} = (y_1, \ldots, y_N)$ and $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ at locations $\boldsymbol{T} = (\boldsymbol{t}_1, \ldots, \boldsymbol{t}_N)$, is calculated by

$$p(y_\star | \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{T}, \boldsymbol{x}_\star, \boldsymbol{t}_\star) = \mathcal{N}(\mu, \psi^2 + \sigma_0^2) \tag{9}$$

where $\psi^2 + \sigma_0^2$ is the full predictive variance. It is the sum of two terms, the noise variance σ_0^2 that represents aleatory variability, and ψ^2 which is a measure of epistemic uncertainty due to uncertainty in the coefficients at the new coordinates. In PSHA calculations, σ_0^2 and ψ^2 are generally treated differently—the former is integrated out, while the latter is sampled using logic trees and leads to a distribution of hazard curves.

The mean μ and the epistemic variance ψ^2 can be calculated as

$$\mu = \boldsymbol{k}^{\mathrm{T}}(\boldsymbol{K} + \tau^2 \boldsymbol{I}_{x \times n})^{-1} \boldsymbol{y} \tag{10}$$

$$\psi^2 = k_\star - \boldsymbol{k}^{\mathrm{T}}(\boldsymbol{K} + \sigma_0^2 \boldsymbol{I}_{x \times n})^{-1} \boldsymbol{k}. \tag{11}$$

Here, $\boldsymbol{K}$ is a matrix with components $K_{i,l} = \sum_{j=1}^{d} x_{ij} x_{lj} \kappa_j(\boldsymbol{t}_i, \boldsymbol{t}_l) \in \mathbb{R}^{N \times N}$. The variable $\boldsymbol{k}$ is a vector with components $k_i = \sum_{j=1}^{d} x_{ij} x_{\star j} \kappa_j(\boldsymbol{t}_i, \boldsymbol{t}_\star)$, where $x_{\star j}$ denotes the j-th component of $\boldsymbol{x}_\star$. The variable $k_\star$ is given by $k_\star = \sum_{j=1}^{d} x_{\star j} x_{\star j} \kappa_j(\boldsymbol{t}_\star, \boldsymbol{t}_\star)$. One interesting point to note about Equations (10) and (11) is that coefficients $\boldsymbol{\beta}$ are integrated out in the predictions, and therefore do not have to be explicitly computed in order to calculate μ and ψ.

As one can see in Equation (11), the epistemic variance is the difference between two terms. The first term is the "prior" variance, which represents the epistemic uncertainty of a new prediction at a new location, for a new set of predictor variables. It is reduced by a positive term which depends on the distance of the new location, $\boldsymbol{t}_\star$, to the existing locations, $\boldsymbol{T}$, in the data set. Hence, the epistemic variance becomes small for a prediction at a new location which is close to an observed station or earthquake location, whereas it remains large in regions where no data are available.

As the model is equivalent to a standard Gaussian process with an appropriate kernel function, it is straightforward to implement using standard toolboxes for Gaussian processes. We have used the *GPML (Gaussian Processes for Machine Learning)* toolbox (Rasmussen and Nickisch, 2010).

Results

In this section, we evaluate the presented VCM regarding its capability to predict logarithmic spectral acceleration at periods T = 0.01, 0.02, 0.05, 0.1, 0.2, 0.5, 1, and 4s. For reference, we compare the model against an ergodic model with (spatially) fixed coefficients, which we call a "global" model—global in the sense that it is estimated for the entire data set. Figure 3 shows the root-mean-squared prediction error (RMSE), estimated by 10-fold cross-validation.

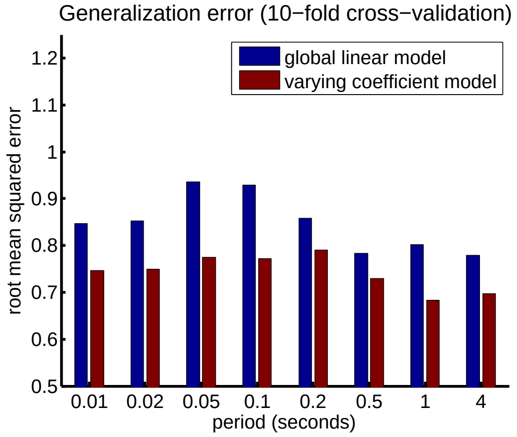

Figure 3: Prediction of logarithmic spectral acceleration: RMSE for the VCM and the ergodic global model, estimated by 10-fold cross validation.

Table 1: Non-spatially varying coefficients for spectral accelerations at different periods.

coefficient	$T = 0.01s$	$T = 0.02s$	$T = 0.05s$	$T = 0.10s$	$T = 0.20s$	$T = 0.50s$	$T = 1.00s$	$T = 4.00s$
β_1	2.4228	2.3162	1.9596	2.5858	3.3492	4.3298	4.9558	3.4599
β_2	-0.17267	-0.16278	-0.14328	-0.18871	-0.22996	-0.28533	-0.31213	-0.13699
β_4	0.1983	0.1991	0.22429	0.16462	0.11759	0.098985	0.089561	0.1238
β_7	0.074761	0.073866	0.11079	0.11489	0.058449	0.056953	0.10464	0.057233
β_8	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1
σ_0	0.5219	0.5255	0.5387	0.5292	0.5392	0.5034	0.4690	0.4588
σ_T	0.8127	0.8176	0.8741	0.8733	0.8433	0.7534	0.7056	0.6848

In 10-fold cross-validation, the data set is split into 10 subsets. In each of ten iterations, the model is estimated based on 9 subsets while one subset is set aside for evaluation. All data associated to any single earthquake event are joined into the same subset; therefore, the model is always evaluated on events that have not been used for parameter estimation. On this tenth subset, the RMSE between the predicted ground motion and the actual ground motion documented in the data is calculated. After 10 iterations, the mean RMSE over is determined by averaging the 10 measurements. The RMSE, as determined by 10-fold cross validation, is an indicator of whether the model is able to generalize well: that is, predict ground motion for new, previously unobserved values of the predictor variables. Figure 3 shows that the VCM has consistently lower RMSE than the global model. This indicates that incorporating spatial differences improves ground-motion prediction, even for a relatively small region such as California.

The spatially varying coefficients for PGA (that is, T = 0.01s) are shown as a color-coded plot in Figure 4. As stated in Equation (5), the constant β_0 and the coefficient that controls

Spatial variation of coefficient β_{-1}

Spatial variation of coefficient β_0

Spatial variation of coefficient β_3

Spatial variation of coefficient β_5

Spatial variation of coefficient β_6

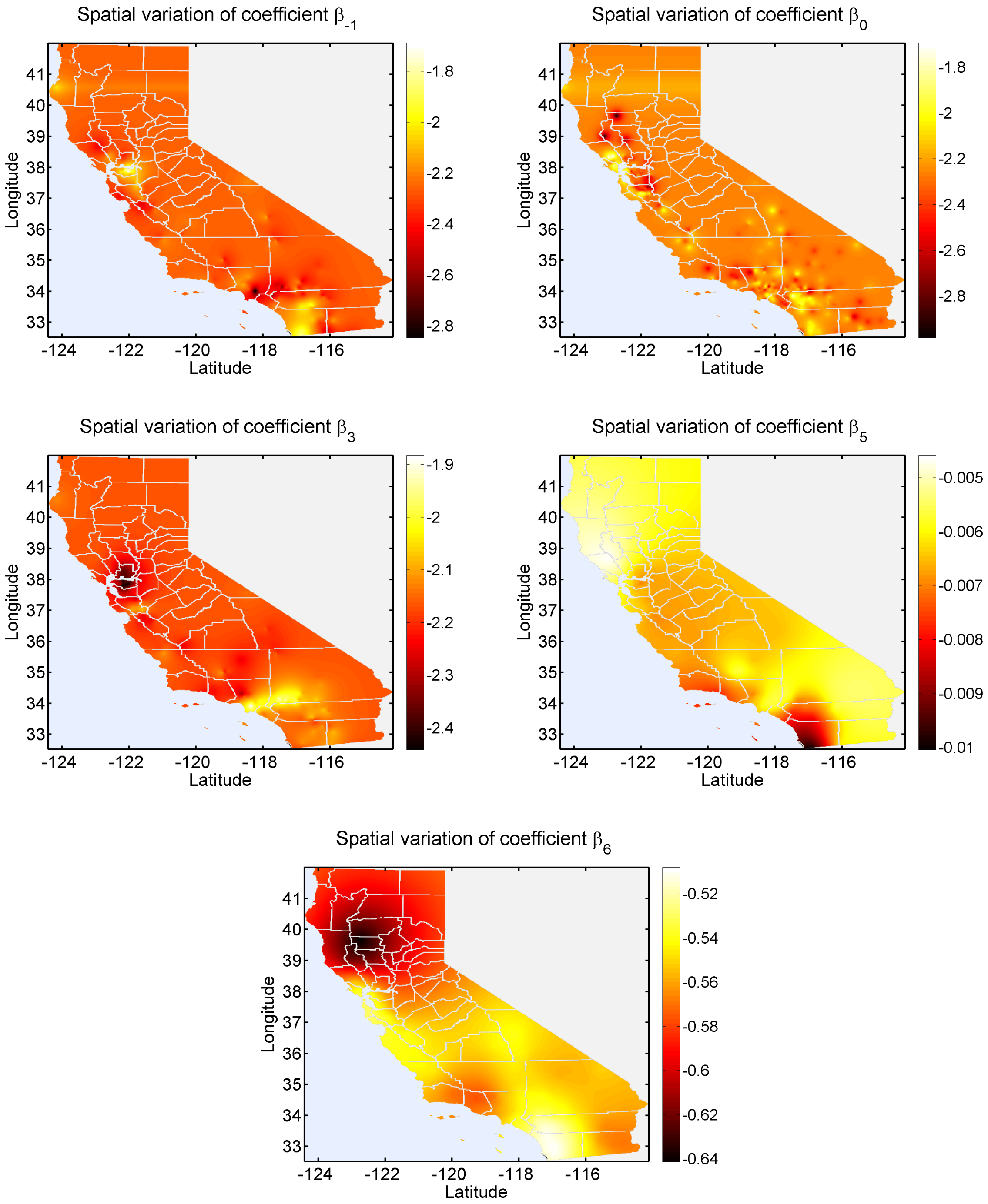

Figure 4: Spatial variation of coefficients for PGA. The color version of this figure is available only in the electronic edition.

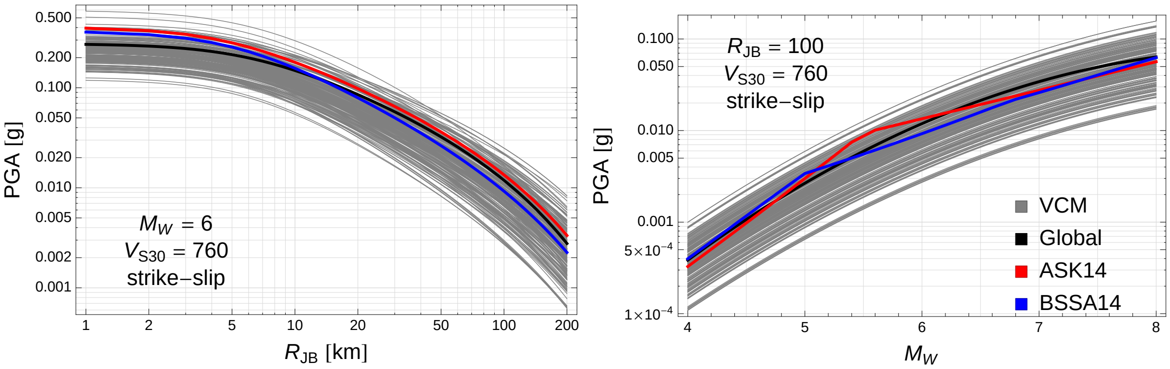

Figure 5: Scaling of the VCM with magnitude and distance. The VCM is evaluated at 221 event locations. For comparison, the scaling of the ergodic global linear model and the ASK14 (Abrahamson et al., 2014) and BSSA14 (Boore et al., 2014) models are shown. The color version of this figure is available only in the electronic edition.

scaling with V_{S30} (β_6) vary with the station location, while the coefficients that control geometrical spreading (β_3) and anelastic attenuation (β_5) as well as β_{-1} vary with source location. The other coefficients, which do not vary by location, are displayed in Table 1. This table also shows the estimated noise term (non-ergodic standard deviation σ_0). Its value is drastically reduced compared to the global, ergodic standard deviation ($\sigma_0 = 0.52$ versus $\sigma_T = 0.81$ for PGA). As stated in the previous section, the values of the median and epistemic uncertainty, μ and ψ, are calculated using only the parameters of the correlation functions and the data. These are presented in the electronic supplement.

Figure 5 shows the scaling of the model with magnitude and distance. Because the model is location dependent, we show the scaling for each event location (thin gray lines in Figure 5). For simplicity, when evaluating the coefficients, we set the station location to the be identical to the event location in Equation (10); this does not affect the distance scaling, whose coefficients depend on event location. The predictor variables $\boldsymbol{x}_\star$ are set independently to the values in Figure 5. In a real application, one needs to use coordinates that correspond to the problem at hand: in particular, the station/event locations need to be consistent with the distance metrics. For comparison, Figure 5 also shows the California models BSSA14 (Boore et al., 2014) and ASK14 (Abrahamson et al., 2014). For the ASK14 model, the values of the depth to the top of the rupture are calculated using the model presented in Chiou and Youngs (2014). As one can see, the overall behavior of the global ergodic model and the California GMMs is similar: differences at small distances are due to the VCM having a constant, magnitude-independent

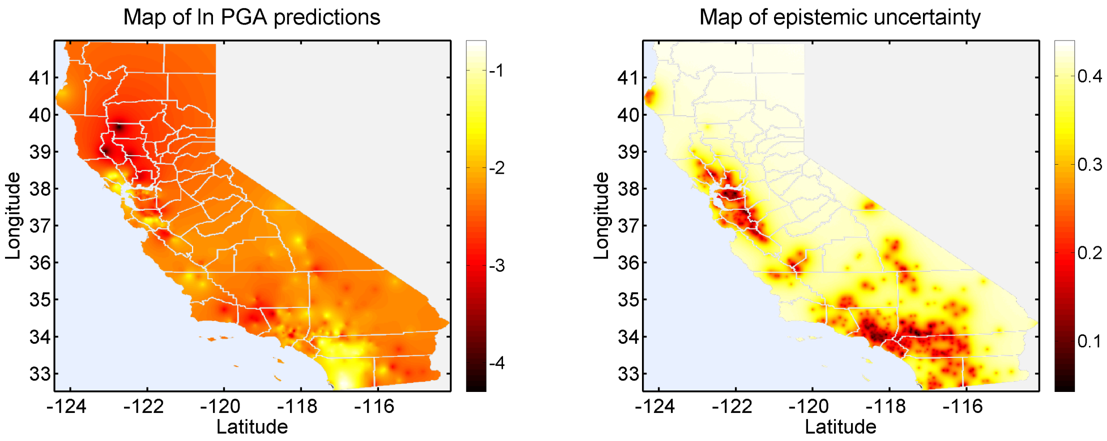

Figure 6: **Left**: Map of ln PGA predictions, coded by ground-motion value. **Right**: Epistemic predictive uncertainty ψ associated with ln PGA predictions. For simplicity, in both plots the same event/station coordinate is used for the coefficients. Predictor variables are set to $M = 6$, $R_{JB} = 10$ km, $SOF = 0$, $V_{S30} = 760$ m/s. The color version of this figure is available only in the electronic edition.

h-term. The individual VCM models scatter around the global model, with some variation in scaling. They all exhibit a reasonable physical scaling with both magnitude and distance. The variation seen in the gray lines in Figure 5 illustrates the trade-off between apparent aleatory variability and epistemic uncertainty. The decrease in the standard deviation leads to an increase in variation of median predictions between different locations.

The left part of Figure 6 shows the spatial variation of median ground-motion predictions for PGA. Specifically, we calculate the median prediction for a set of predictor variables $M = 6$, $R_{JB} = 10$km, $V_{S30} = 760$m/s and $F_N = F_R = 0$, for a location grid across California. As for Figure 5, for simplicity we use same coefficients at the same coordinates for station and event. Similar maps can be found for other periods in the electronic supplement. As one can see, there is some spatial variation in the median predictions over California, but this variation is constrained to locations where data are available: that is, close to stations or observed events (cf. Figure 1).

The right part of Figure 6 shows the epistemic uncertainty ψ associated with the predictions across California, as calculated by Equation (11). The values of ψ are calculated using the same settings as for the calculation of the median predictions. We can see that the predictive uncertainty increases for regions where data are sparse. It is important to include this uncertainty in hazard calculations. Currently, this uncertainty is included as part of aleatory variability, and epistemic uncertainty of median predictions using alternative GMMs is modeled for example

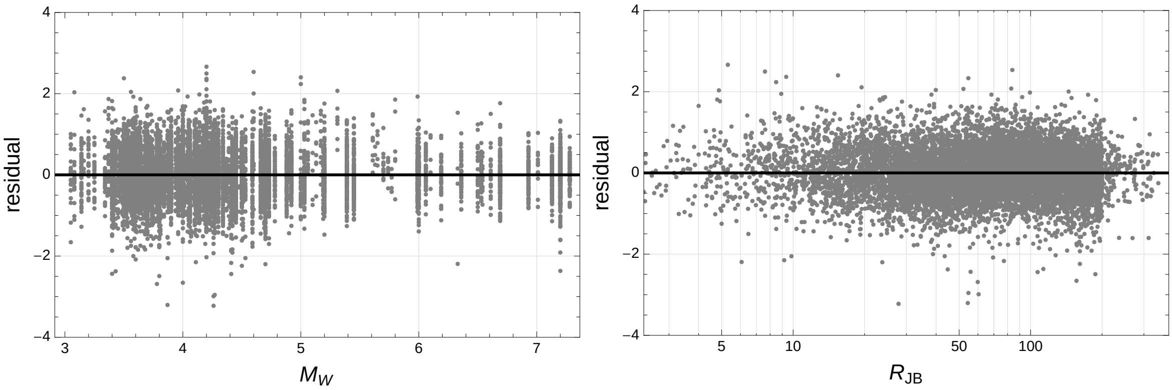

Figure 7: Residuals of the VCM, calculated as observed minus predicted PGA value.

by Al-Atik and Youngs (2014), which gives a value of the epistemic standard deviation of 0.083 for $\boldsymbol{M} = 6, R_{JB} = 100$ and PGA. By contrast, the standard deviation of median predictions of the VCM at the event coordinates has a value of 0.45 for $\boldsymbol{M} = 6, R_{JB} = 100$ and PGA.

The spatial variation of median predictions seen in Figures 5 and 6 is an example of trading epistemic uncertainty and apparent aleatory variability. In the global model, the spatial variation is accounted for by the estimated (ergodic) variance, which includes the repeatable source, path, and site terms, whereas these are translated into the median predictions in the VCM framework. For PGA, we see a reduction of about 35% in the value of the standard deviation (from 0.81 to 0.52). This is comparable with results seen for Taiwan (cf. Table 5 of Lin et al. (2011)), where approximately a 50% reduction going from full-ergodic to non-ergodic standard deviation is estimated.

Figure 7 shows the residuals of the VCM, plotted against magnitude and distance. There are no obvious trends with the predictor variables.

Discussion and Conclusions

We have presented a (relatively) simple GMM that can take into account regional differences in ground-motion scaling. In contrast to other approaches to regionalizing GMMs (Gianniotis et al., 2014; Stafford, 2014), which estimate separate GMMs for distinct regions, the presented model works spatially on a continuous scale. Hence, the predictions of the model vary smoothly across California. This is a next step to go from a partially non-ergodic PSHA (having separate GMMs for small regions, as well as using single-station sigma) to an almost fully non-ergodic

PSHA. The presented model does not take true 3D-path effects into account, because the coefficients that account for scaling with distance describe an average distance attenuation for each event in all directions. The next logical step in the development of a fully non-ergodic GMM is to take such 3D-path effects into account.

The underlying assumption of the VCM is that some of its coefficients vary continuously with station or event location, as shown in Figure 4. From Equations (9) to (11) we can see that the predictive distribution for a new set of predictor variables at a new location can be computed without explicitly computing the spatially varying coefficients, and is completely determined by the parameters of the covariance function $\boldsymbol{\kappa}$.

We have seen in Figure 3 that the presented VCM has a low average prediction RMSE on unseen test data (that is, for events and data points that have not been used in estimating the model parameters), which provides strong evidence that the VCM is a viable, even superior alternative to a non-ergodic GMM. For the application of the VCM in seismic hazard analysis, the following two points need to be taken into account. First, there are different coefficients for different locations, and different subsets for station and earthquake locations. Hence, to estimate seismic hazard at a particular site, the appropriate sets of coefficients need to be applied for the site and the relevant sources. Second, it is important to keep track of the (epistemic) predictive uncertainty of the model. The VCM has a much smaller value of the aleatory variability than the global ergodic model. The reduction in aleatory variability is due to including spatial variation of the coefficients, which results in the variation of median predictions as seen in Figure 5 and in the left part of Figure 6. However, Figure 6 also shows that there is less variation in the median predictions for locations that are far from observed data (stations or earthquakes); here, the VCM predictions resort to mean predictions. Accordingly, this is accommodated by a larger epistemic predictive uncertainty ψ. This increased uncertainty must be incorporated into PSHA calculations, because just using a smaller value for the aleatory variability, without any adjustment for median predictions, will lead to an underestimation of the seismic hazard.

Data and Resources

The data used in this study come from the PEER NGA West 2 data base (Ancheta et al., 2014) (http://peer.berkeley.edu/ngawest2/databases/, last accessed 07/10/2015) and comprise the Californian/Nevada data used for the model of Abrahamson et al. (2014).

Acknowledgements

We would like to thank Frank Scherbaum for comments on an early draft of the manuscript. Support of the Pacific Earthquake Engineering Research Center (PEER) and the German Research Foundation (DFG), grant LA 3270/1-1, is gratefully acknowledged. We would like to thank Ivan Wong and two anonymous reviewers for comments that helped to improve the manuscript.

References

Abrahamson, N. A., W. J. Silva, and R. Kamai (2014). Summary of the ASK14 Ground Motion Relation for Active Crustal Regions. *Earthquake Spectra* **30**, no. 3, 1025–1055. doi: 10.1193/070913EQS198M.

Akkar, S., and Z. Cagnan (2010). A Local Ground-Motion Predictive Model for Turkey, and Its Comparison with Other Regional and Global Ground-Motion Models. *Bulletin of the Seismological Society of America* **100**, no. 6, 2978–2995. doi: 10.1785/0120090367.

Al-Atik, L., N. Abrahamson, J. J. Bommer, F. Scherbaum, F. Cotton, and N. Kuehn (2010) The Variability of Ground-Motion Prediction Models and Its Components. *Seismological Research Letters* **81**, no. 5, 794–801. doi: 10.1785/gssrl.81.5.794.

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, T. Kishida, and J. L. Donahue (2014) NGA-West2 Database. *Earthquake Spectra* **30**, no. 3, 989–1005. doi: 10.1193/070913EQS197M.

Anderson, J. D., and J. N. Brune (1999). Probabilistic Seismic Hazard Analysis without the Ergodic Assumption. *Seismological Research Letters* **70**, no. 1, 19–28. doi: 10.1785/gssrl.70.1.19.

Anderson, J. G., and Y. Uchiyama (2011). A Methodology to Improve Ground-Motion Prediction Equations by Including Path Corrections. *Bulletin of the Seismological Society of America* **101**, no. 4, 1822–1846. doi: 10.1785/0120090359.

Al-Atik, L., and R. R. Youngs (2014). Epistemic Uncertainty for NGA-West2 Models. *Earthquake Spectra* **30**, no. 3, 1301–1318. doi: 10.1193/062813EQS173M.

Atkinson, G. M (2006). Single-Station Sigma. *Bulletin of the Seismological Society of America* **96**, no. 2, 446–455. doi: 10.1785/0120050137.

Atkinson, G. M., and M. Morrison (2009). Observations on Regional Variability in Ground-Motion Amplitudes for Small-to-Moderate Earthquakes in North America. *Bulletin of the Seismological Society of America* **99**, no. 4, 2393–2409. doi: 10.1785/0120080223.

Bindi, D., F. Pacor, L. Luzi, R. Puglia, M. Massa, G. Ameri, and R. Paolucci (2011). Ground motion prediction equations derived from the Italian strong motion database. *Bulletin of Earthquake Engineering* **9**, no. 6, 1899–1920. doi: 10.1007/s10518-011-9313-z.

Bommer, J. J., and N. A. Abrahamson (2006). Why Do Modern Probabilistic Seismic-Hazard Analyses Often Lead to Increased Hazard Estimates? *Bulletin of the Seismological Society of America* **96**, no. 6, 1967–1977. doi: 10.1785/0120060043.

Boore, D. M., J. P. Stewart, E. Seyhan, and G. M. Atkinson (2014). NGA-West2 Equations for Predicting PGA, PGV, and 5% Damped PSA for Shallow Crustal Earthquakes. *Earthquake Spectra* **30**, no. 3, 1057–1085. doi: 10.1193/070113EQS184M.

Bozorgnia, Y., N. A. Abrahamson, L. Al-Atik, T. D. Ancheta, G. M. Atkinson, J. W. Baker, A. Baltay, D. M. Boore, K. W. Campbell, B. S.-J. Chiou, R. Darragh, S. Day, J. Donahue, Robert W. Graves, Nick Gregor, Thomas Hanks, I. M. Idriss, Ronnie Kamai, T. Kishida, A. Kottke, S. A. Mahin, S. Rezaeian, B. Rowshandel, E. Seyhan, S. Shahi, T. Shantz, W. Silva, P. Spudich, J. P. Stewart, Jennie Watson-Lamprey, K. Wooddell, and R. Youngs

(2014). NGA-West2 Research Project. *Earthquake Spectra* **30**, no. 3 973–987. doi: 10.1193/072113EQS209M.

Bragato, P. L., and D. Slejko (2005). Empirical Ground-Motion Attenuation Relations for the Eastern Alps in the Magnitude Range 2.5-6.3. *Bulletin of the Seismological Society of America* **95**, no. 1, 252276. doi: 10.1785/0120030231.

Bussas, M., C. Sawade, T. Scheffer, and N. Landwehr (2015). Varying-coefficient models with isotropic Gaussian process priors. ArXiv:1508.07192.

Campbell, K. W., and Y. Bozorgnia (2014). NGA-West2 Ground Motion Model for the Average Horizontal Components of PGA, PGV, and 5% Damped Linear Acceleration Response Spectra. *Earthquake Spectra* **30**, no. 3, 1087–1115. doi: 10.1193/062913EQS175M.

Chiou, B. S.-J., and R. R. Young (2014). Update of the Chiou and Youngs NGA Model for the Average Horizontal Component of Peak Ground Motion and Response Spectra. *Earthquake Spectra* **30**, no. 3, 1117–1153. doi: 10.1193/072813EQS219M.

Chiou, B., R. Youngs, N. Abrahamson, and K. Addo (2010). Ground-Motion Attenuation Model for Small-To-Moderate Shallow Crustal Earthquakes in California and Its Implications on Regionalization of Ground-Motion Prediction Models. *Earthquake Spectra* **26**, no. 4, 907–926. doi: 10.1193/1.3479930.

Danciu, L., and G. A. Tselentis (2007). Engineering Ground-Motion Parameters Attenuation Relationships for Greece. *Bulletin of the Seismological Society of America* **97**, no. 1, 162183. doi: 10.1785/0120050087.

Douglas, J., S. Akkar, G. Ameri, P.-Y. Bard, D. Bindi, J. J. Bommer, S. S. Bora, F. Cotton, B. Derras, M. Hermkes, N. M. Kuehn, L. Luzi, M. Massa, F. Pacor, C. Riggelsen, M. A. Sandkkaya, F. Scherbaum, P. J. Stafford, and P. Traversa (2014). Comparisons among the five ground-motion models developed using RESORCE for the prediction of response spectral accelerations due to earthquakes in Europe and the Middle East. *Bulletin of Earthquake Engineering* **12**, no. 1, 341–358. doi: 10.1007/s10518-013-9522-8.

Douglas, J., and H. Aochi (2016). Assessing Components of GroundMotion Variability from

Simulations for the Marmara Sea Region (Turkey). *Bulletin of the Seismological Society of America* **106**, no. 1, 300–306. doi: 10.1785/0120150177.

Gelfand, A. E., H.-J. Kim, C. F. Sirmans, and S. Banerjee (2003). Spatial Modeling With Spatially Varying Coefficient Processes. *Journal of the American Statistical Association* **98**, no. 462, 387–396. doi: 10.1198/016214503000170.

Gianniotis, N., N. Kuehn, and F. Scherbaum (2014). Manifold aligned ground motion prediction equations for regional datasets. *Computers and Geosciences* **69**, 72–77. doi: 10.1016/j.cageo.2014.04.014.

Jayaram, N., and J. W. Baker (2009). Correlation model for spatially distributed ground-motion intensities. *Earthquake Engineering & Structural Dynamics* **38**, no. 15 1687–1708. doi: 10.1002/eqe.922.

Kuehn, N. M., and N. A. Abrahamson (2015). Non-Ergodic Seismic Hazard : Using Bayesian Updating for Site-Specific and Path-Specific Effects for Ground-Motion Models In *CSNI Workshop on Testing PSHA Results and Benefit of Bayesian Techniques for Seismic Hazard Assessment*, 1–15.

Lin, P.-S., B. Chiou, N. Abrahamson, M. Walling, C.-T. Lee, and C.-T. Cheng (2011). Repeatable Source, Site, and Path Effects on the Standard Deviation for Empirical Ground-Motion Prediction Models. *Bulletin of the Seismological Society of America* **101**, no. 5, 2281–2295. doi: 10.1785/0120090312.

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

Rasmussen, C. E., and H. Nickisch (2010). Gaussian Processes for Machine Learning (GPML) Toolbox. *Journal of Machine Learning Research* **11**, 3011–3015.

Rodriguez-Marek, A., G. A. Montalva, F. Cotton, and F. Bonilla (2011). Analysis of Single-Station Standard Deviation Using the KiK-net Data *Bulletin of the Seismological Society of America* **101**, no. 3, 1242–1258. doi: 10.1785/0120100252.

Stafford, P. J. (2014). Crossed and Nested Mixed-Effects Approaches for Enhanced Model Development and Removal of the Ergodic Assumption in Empirical Ground-Motion Models *Bulletin of the Seismological Society of America* **104**, no. 2, 702–719. doi: 10.1785/0120130145.

Villani, M., and N. A. Abrahamson (2015). Repeatable Site and Path Effects on the Ground-Motion Sigma Based on Empirical Data from Southern California and Simulated Waveforms from the CyberShake Platform *Bulletin of the Seismological Society of America* **105**, no. 5, 2681–2695. doi: 10.1785/0120140359.

Walling, M. A. (2009). *Non-Ergodic Probabilistic Seismic Hazard Analysis and Spatial Simulation of Variation in Ground Motion.* PhD thesis, UC Berkeley.