SELECTING EARTHQUAKE SCENARIOS FOR RISK ANALYSIS THAT PRESERVE JOINT DISTRIBUTIONS OF REGIONAL GROUND-MOTION INTENSITIES

N. Sharma[1] & J.W. Baker[2]

[1] Stanford University, Stanford, USA, neetesh@stanford.edu

[2] Stanford University, Stanford, USA

Abstract: *Regional seismic risk analysis involves assessing the probability and evaluating the socioeconomic impact of various seismic events. Accounting for the uncertainty in future earthquake characteristics requires evaluating a large number of events. However, assessments of infrastructure functionality loss and the ensuing socioeconomic impact are computationally intensive, making brute-force simulation approaches impractical. Previous research has tackled this problem by selecting a limited number of earthquake scenarios to represent the complete seismic hazard. However, these approaches only consider the marginal distributions of ground-motion intensities across a region and fail to account for correlations between different intensity measures at different locations. Neglecting these spatial and inter-intensity measure correlations may lead to underestimating the seismic risk in a given area. This work proposes a formulation that incorporates spatial and inter-intensity measure correlations when selecting earthquake scenarios. Additionally, we develop an open-source Python package allowing users to implement the proposed methodology. Furthermore, we provide an illustrative case study showcasing the application of our proposed method. By facilitating expanded and risk-consistent studies of the impact of infrastructure networks on regional seismic risk and resilience, our work can potentially enhance the understanding of seismic risks and inform effective risk management strategies.*

1. Introduction

Regional seismic risk analysis involves assessing the probability and evaluating the socioeconomic impact of various seismic events. However, earthquakes are low-probability and high-consequence events, and there is high uncertainty associated with their occurrence, as well as the consequences on society given the occurrence. Governments, communities, and industrial entities such as catastrophe insurers often need to make decisions about mitigating and managing earthquake risk. Making such decisions while accounting for the uncertainty in future earthquake characteristics and the associated societal impacts requires evaluating a large number of earthquake scenarios. Many types of earthquake impacts are of interest to different stakeholders, for example, the damage to structures, loss of functionality in critical infrastructure, societal impacts such as casualties, health, and shelter loss, as well as the economic losses to individuals, businesses, and economies (Rose and Lim 2002; Chang 2003). However, evaluating these consequences is computationally intensive, and in the lack of closed-form formulations, simulations remain the widely used solution (Ebel and Kafka 1999; Crowley and Bommer 2006). However, running many simulations is often infeasible, and probabilistic results are difficult to communicate to the public (Corotis et al. 2012). So, most studies tend not to account for seismic hazard uncertainty and focus on a single or handful of earthquake scenarios (Chang and Nojima 2001; Elnashai et al. 2008; Tabandeh et al. 2022).

Uncertainty quantification literature indicates that the hazard models contribute a majority of uncertainty in seismic risk (Ellingwood and Kinali 2009). So, the challenge in seismic risk analyses is to account for hazard uncertainty while minimizing the required impact assessment simulations. Previous research has tackled this problem by selecting a limited number of earthquake scenarios to represent the complete seismic hazard. Importance sampling, clustering, and optimization-based methodologies have been applied in the literature for selecting a limited number of scenarios to represent the probability hazard in the region (for example, Jayaram and Baker 2010; Han and Davidson, 2012; Vaziri et al. 2012; Miller and Baker 2015; Manzour et al. 2016; Soleimani et al. 2021; Rosero-Velásquez and Straub 2022). The typical objectives for selecting the representative scenarios have satisfied the occurrence probabilities of intensity measures for the various sites in the region of interest. However, there has been a limited focus on maintaining consistency with the joint distribution of various intensity measures. Maintaining the joint distribution of hazard intensity measures involves satisfying the spatial and inter-intensity measure correlations while selecting the hazard scenarios. Neglecting these spatial and inter-intensity measure correlations may lead to underestimating the seismic risk for the region. The limited literature attempting to address this gap reports challenges in dealing with the high dimensionality of the problem and the high computational cost (Kavvada et al. 2022). Furthermore, the performance of the current methods has often been worse when selecting small sets of scenarios (less than a hundred) and satisfying the consistency at high return periods, which is required for most planning and preparedness applications.

This work proposes a formulation that incorporates spatial and inter-intensity measure correlations when selecting earthquake scenarios. The large number of sites controls the high dimensionality of the joint intensity measures in the region. However, the spatial correlation models indicate that it is possible to represent the regional seismic hazard with smaller dimensions (Loth and Baker 2013). Furthermore, Markhvida et al. (2018) show that the inter-intensity measure correlations can also be represented by a reduced set of dimensions using Principal Component Analysis (PCA) (Hotelling 1933). Exploiting these insights, in this paper, we use Principal Component Analysis to find a low-dimensional representation of the regional seismic intensity measures. We then use optimization to maintain consistency of the probabilities of occurrence for the principal components that implicitly improve performance in maintaining the correlations over space and among intensity measures. Then, to improve the performance of the optimization in selecting smaller sets of scenarios we use parameter tuning similar to LASSO (least absolute shrinkage and selection operator) in regression (Tibshirani 1996). We also identify improvements in weights and target return periods. We illustrate the proposed formulation for selecting earthquake scenarios for a region in Sonoma County, California. Preliminary results indicate that PCA-based optimization formulation performs better at selecting a smaller number of scenarios while improving computational efficiency and solution stability.

The rest of the paper is organized into four sections. Following this introduction, Section 2 discusses some essential background and formulations required to explain the paper's contributions. Section 3 presents the main methodological contribution of the paper; Section 4 then discusses the case study and presents some preliminary results. Finally, Section 5 summarises the paper and draws conclusions from the case study results.

2. Background on seismic hazard mapping and scenario selection

This section describes how the ground-motion intensity realizations are produced for a region. We also discuss the current methods for optimal scenario selection, which are conceptually similar to the proposed formulation and serve as a benchmark for comparison of improvements.

2.1. Generating ground motion intensity maps

For the selected region of interest, the extensive (close to accurate) representation of regional seismic hazard is first captured by a large number of selected scenarios that discretize the infinite-dimensional hazard. This set of scenarios is commonly known as the rupture forecast, which provides the probability of occurrence of various magnitudes, locations, and faulting types in the vicinity of the region. For example, the rupture forecast may provide Q number of scenarios with corresponding reoccurrence rates of w_q, where $q \in \{1,2,\ldots,Q\}$. This information is based on available seismological data (e.g., Uniform California Earthquake Rupture Forecast, Version 3 (UCERF3), 2013).

Now, for each earthquake scenario in the extensive set of scenarios, the shaking intensity at each location of interest is characterized probabilistically using a ground motion model (GMM) (for example, Campbell and Bozorgnia 2014). Various types of ground motion models (for example, physics-based simulations or empirical equations) can be used to find the intensity measures at various locations. However, for regional risk assessment applications, the following type of empirical model is the most common, which we also use in this paper:

$$\ln Y_{isj} = \ln \bar{Y}_{isj}\left[M_j, R_{i,j}, \dots\right] + \phi_{isj}\epsilon_{isj} + \tau_{isj}\eta_{sj} \tag{1}$$

where Y_{isj} is the realized ground motion intensity measure of type s (for example, spectral acceleration, S_a at a particular period), i is the index for a location in the region, and j is the ground motion intensity map index. M_j is the moment magnitude of the j^{th} scenario, R_{ij} is a distance measure from the rupture location to the location i. If the initial earthquake rupture forecast provides Q earthquake scenarios and we sample b spatially and inter-intensity measure correlated maps, then the index $j \in \{1,2,\dots,m\}$, where $m = Q \times b$. The variable, ϕ_{isj} is the intra-event standard deviation for the joint normally distributed residual ϵ_{isj}. Similarly, τ_{isj} is the inter-event standard deviation for the joint normally distributed residual η_{sj}. The ground motion model (GMM) thus provides the median $\bar{Y}_{isj}$, and the standard deviation terms ϕ_{isj} and τ_{isj} for each of the intensity measures of interest, s, for each location, i. The residuals ϵ_{isj} and η_{sj} are sampled using the spatial and inter-intensity measure correlation models such as Baker and Jayaram (2008), Loth and Baker (2013) and Markhvida et al. (2018) to obtain the intensity realizations Y_{isj}.

Following the above explanation, if the number of uniformly sampled realizations is b from each of the Q scenarios, the recurrence rate associated with each of the realized maps is $w_j = w_q/b, j \in \{1,2,\dots,m\}$ where w_q is the recurrence rate of the scenario q as obtained from the earthquake rupture forecast. Combining Y_{isj} and w_j, we can estimate for each site i and intensity measure t, the marginal exceedance rate for an intensity, also known as the hazard curve $\lambda(Y_{is} \geq y)$ as

$$\lambda_{is}(Y_{is} \geq y) = \sum_{j=1}^{m} w_j \,.\, \mathbf{1}_{\{Y_{isj} \geq y\}} \tag{2}$$

where $\mathbf{1}_{\{.\}}$ is an indicator function, which takes the value 1 if and only if the Boolean $Y_{isj} \geq y$, in the subscript is True, and zero otherwise. Hazard curves are traditionally used for site-specific probabilistic seismic hazard analysis for individual structures. As apparent from Equation 2, hazard curves only capture the marginal probability of exceedance of various intensity measures at any site and do not consider the joint distribution of intensities over several sites.

2.2. Optimal selection of ground motion maps

The goal of this analysis is to select a set of k ground-motion intensity maps and assign to them an adjusted rate of occurrence w_j', so that the subset represents well the regional seismic hazard. Most current methodologies attempt to maintain the similarity in terms of the hazard curves defined in Equation 2. Here, we discuss a specific contribution by Miller and Baker (2015), which formulates the problem using convex optimization. Mathematically, we can write the problem as

minimize

$$\sum_{s=1}^{S}\sum_{i=1}^{n} \|\mathrm{diag}(\boldsymbol{\lambda})^{-1}(\boldsymbol{\lambda} - \boldsymbol{\Theta}_{is}\mathbf{w})\|_1 \tag{3a}$$

subject to

$$\|\mathbf{w}\|_1 \leq \sum_{j=1}^{m} w_j \tag{3b}$$

$$\mathbf{0} \leq \mathbf{w} \tag{3c}$$

where the decision variable is the vector $\mathbf{w} \in \mathbb{R}^{m\times 1}$, each element of which is the adjusted recurrence rate w_j', $j \in \{1,2,\ldots,m\}$, $\|\cdot\|_1$is the L_1 norm, i.e., the sum of absolute values of the elements in the vector. The vector $\boldsymbol{\lambda}$ is the exceedance rates over which we are trying to minimize the deviation from the hazard curve. The elements of this vector $\lambda_r, r \in \{1,2,\ldots,R\}$ are constants where R is the total number of return periods of interest; for example, if the first return period of interest is 100, then the corresponding λ_1 would be 0.01. The matrix $\mathrm{diag}(\boldsymbol{\lambda})^{-1}$ is a matrix with the principal diagonal having the only non-zero values, which are $(\boldsymbol{\lambda})^{-1}$. Here, $\boldsymbol{\Theta}_{is} \in \mathbb{R}^{R\times m}$ is a binary matrix where each element is

$$\theta_{is;r,j} = \mathbf{1}_{\{Y_{isj} \geq \tilde{y}_{isr}\}} \tag{4}$$

where i is the index for a location in the region, $i \in \{1,2,\ldots,n\}$ over which the objective function will be minimized, s is the intensity measure. The constant $\tilde{y}_{isr}$ is the y obtained from Equation 2, when we set the left-hand side to be one of the exceedance rates of interest, λ_r, i.e., $\tilde{y}_{isr}$ corresponds to the intensity that has λ_r as the exceedance rate for site i and intensity measure s from the extensively sampled set. Finally, the constraint $\|\mathbf{w}\|_1 \leq \sum_{j=1}^{m} w_j$, limits the total recurrence rate to be not more than the extensively sampled set, the constraint, $\mathbf{0} \leq \mathbf{w}$, ensures that the adjusted rates are non-negative. Miller and Baker then suggest using the k maps with the largest w_j' as the representative set for the regional seismic hazard.

The objective function in Equation (3a) minimizes the gaps between the exceedance rate curves from the extensively sampled set (notated as "fullset" in Figure 1) and from the subset; each site i and intensity measure s in the objective function contributes a ground-motion intensity exceedance rate curve. The difference between the hazard curves for the intensity measures corresponding to λ_r are further multiplied by $\mathrm{diag}(\boldsymbol{\lambda})^{-1}$, which effectively performs a log transformation on the exceedance rate axis, i.e., highly weighting the errors at lower exceedance rates, which are typically of greater importance. The optimization objective function thus minimizes the weighted sum of vertical distances between the hazard curves at the R number of exceedance rates that discretize the range of importance shown by the grey region Figure 1.

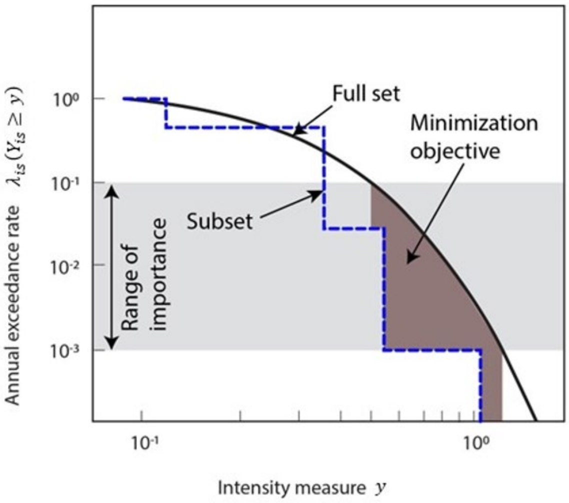

Figure 1. Hazard curves from extensive set and subset of scenarios, with the area related to the minimization objective. The range of importance for exceedance rates may exclude events that do not cause significant damage and events that are too unlikely to occur or infeasible to plan for.

3. Proposed formulation for optimal selection of ground motion maps

The spatial variability in hazard intensities plays a significant role in estimating the regional seismic risk. Typical regional seismic risk analysis is performed on a large number of sites, which may correspond to the locations of thousands of buildings or infrastructure components within a region of interest. The number of locations, n, thus controls the dimensions of the joint distributions of intensity. This section discusses a weighted Principal Component analysis-based dimension reduction algorithm to reduce the computational cost. We then discuss the optimization formulation that translates the marginal intensity measure-based objective into the corresponding preference in terms of the principal components.

3.1. Dimension reduction and correlation embedding

As explained in section 2.1, the residuals ϵ_{isj} and η_{sj} follow normal distributions. Hence the $\ln Y_{isj}$ is a good candidate for dimension reduction using PCA. However, the samples of $\ln Y_{isj}$ over j are not equally likely samples, instead, they are occurrences from random processes with occurrence rates w_j. So, we first obtain the weight of j^{th} map in the PCA. The weight for the sample should be the relative likelihood of observing the sample given a hazard occurrence. For the case of multiple independent Poisson processes, say $N_j(t)$ with recurrence rates w_j, the encompassing Poisson process, $N_T(t)$ will have a reoccurrence rate $w_T = \sum_{j=1}^{m} w_j$. The relative likelihood of observing a sample from $N_j(t)$ can then be written as $\mathbb{P}\big(N_j(t) = 1 | N_T(t) = 1\big)$, which can be shown to follow

$$\mathbb{P}\big(N_j(t) = 1 | N_T(t) = 1\big) = \frac{w_j}{\sum_{j=1}^{m} w_j} \tag{5}$$

In other words, the weights for the various sample ground motion intensity maps are the recurrence rates normalized by their total.

We can then use weighted PCA to reduce the distribution dimensions and embed the correlation into independent dimensions. We use the algorithm in Delchambre (2015) that uses a weighted covariance eigen decomposition approach to fit the weighted PCA (VanderPlas 2016). The weighted PCA finds a linear transformation of the data Y_{isj} represented by a matrix $\mathbf{Y} \in \mathbb{R}^{m \times nS}$, where m is the number of rows and the number of intensity maps, n is the number of sites, and S is the number of intensity measures of interest. The row weights of these data are $w_j / \sum_{j=1}^{m} w_j$. We obtain, using weighted PCA, orthogonal principal components that we denote as $\mathbf{X}$. Mathematically PCA finds parameters $\mathbf{B}$, such that

$$\mathbf{X} = \mathbf{YB} \tag{6}$$

We denote the number of principal components we extract as ν, i.e., from the complete $\mathbf{X} \in \mathbb{R}^{m \times nS}$**,** we only keep $\mathbf{X}' \in \mathbb{R}^{m \times \nu}$. The number of components required to describe the data can be chosen based on the proportion of variance explained by the first ν.components. In general, the number of components required to represent the data is much smaller than the original dimensions, i.e., $\nu \ll nS$, which leads to high computational savings.

3.2. Optimal selection of ground motion maps in reduced dimensions

Once we have an orthogonal and lower dimensional representation of the ground motion intensity maps, various optimization and clustering techniques become applicable to select the representative set of maps. The problem can be formulated quite similarly to the problem in the original space of sites, i, and intensity measure types, s. However, in this paper, we wanted to study the improvement we can expect based on the conceptual contribution of reduced dimensionality using weighted PCA. Hence, we modify a simple optimization algorithm described in Section 2.2 to work in the transformed space. The optimization formulation in the transformed space with some additional improvements is the following:

minimize

$$\sum_{l=1}^{\nu} \|\mathrm{diag}(\boldsymbol{\vartheta}_l)(\boldsymbol{\lambda}_l - \boldsymbol{\Theta}_l \mathbf{w})\ \|_1 \tag{7a}$$

subject to

$$\|\mathbf{w}\|_1 \leq \omega \tag{7b}$$

$$\mathbf{0} \leq \mathbf{w} \tag{7c}$$

where the decision variable remains the same, i.e., the vector $\mathbf{w} \in \mathbb{R}^{m \times 1}$, each element of which is the adjusted recurrence rate $w_j', j \in \{1,2,\dots,m\}$. The first difference, however, occurs in vectors $\boldsymbol{\lambda}_l$, which are now different for each component $\mathbf{X}_l, l \in \{1,2,\dots,\nu\}$. Furthermore, the weights for the difference between the exceedance curves at various values of λ_l captured by $\boldsymbol{\vartheta}_l$ are not the same as $\boldsymbol{\lambda}_l$, as in the original formulation. This is because the rate of exceedance for principal component values need not correspond to the preferences expressed by the exceedance rates in the original space. So, we need to ascertain what should be the $\boldsymbol{\lambda}_l$ and

$\boldsymbol{\vartheta}_l$ for a better selection of ground motion maps. Ascertaining these values rigorously is ongoing work, which can be based on the transformation obtained in Equation 6, and further modified by tuning for better performance on achieving consistency in hazard curves and correlation in the original space. However, as part of this paper we present a simple empirical approach that individually sets $\boldsymbol{\lambda}_l$ for each principal component. We first find for each map indexed j the geometric mean, $\bar{\lambda}_{is}(Y_{is} \geq y_j)$, over each site i and intensity measure s. We then find the $\boldsymbol{\vartheta}_l$, for each principal component based on a linear fit between $\bar{\lambda}_{is}(Y_{is} \geq y_j)$ as the response and corresponding $\bar{\lambda}_l(X_l \geq x_j)$ as the predictor. The matrix, $\boldsymbol{\Theta}_l \in \mathbb{R}^{R\times m}$ is a binary matrix, where each element is

$$\theta_{l;r,j} = \mathbf{1}_{\{X_{lj} \geq \tilde{x}_{lr}\}} \tag{8}$$

where l is the principal component index, $l \in \{1,2,\dots,v\}$ over which the objective function will be minimized, t is the intensity measure. The constant $\tilde{x}_{lr}$ corresponds to the principal component value that has λ_{lr} as the exceedance rate for component l from the extensively sampled set. Finally, the constraint $\|\mathbf{w}\|_1 \leq \omega$, creates a LASSO-like penalty for selecting a larger number of scenarios. Using ω, we can tune for improving the performance while selecting small number of maps.

3.3. Evaluation metrics

Two types of evaluation metrics have been used in the literature for the selection of scenarios and corresponding ground motion maps. These metrics capture the performance in terms of error in capturing the hazard curves at all the sites in the region as well as the errors in correlations over pairs of sites and intensity measures.

Mean Hazard Curve Error ($MHCE$)

Han and Davidson (2012) define the Hazard Curve Error, HCE as the error in the hazard curve in terms of an intensity measure as the percentage of the 'true' value. The horizontal distance from the reduced set hazard curve to the 'true' hazard curve for site i and exceedance rate λ_r by the true ground motion at site i with an exceedance rate λ_r.

$$HCE_{isr} = \frac{\hat{y}_{isr} - y_{isr}}{y_{isr}} \times 100 \tag{9}$$

$$MHCE = \sum_{s=1}^{S}\sum_{i=1}^{n}\sum_{r=1}^{R} |HCE_{isr}| \tag{10}$$

where $\hat{y}_{isr}$ and y_{isr} denote the intensity measures at site i and intensity type s for the exceedance rate λ_r for the hazard curves based on the subset and the extensively sampled set, respectively.

Mean Absolute Correlation Error ($MACE$)

Mean absolute correlation error is defined by Kavvada et al. (2022) by extending the definition of Mean Spatial Correlation Error ($MSCE$) from Han and Davidson (2012) for multiple intensity measures. First, the weighted correlation coefficient between a pair of intensity measures, s_1 and s_2 over a pair of sites i_1 and i_2 is defined as

$$\rho_{i_1 s_1 i_2 s_2} = \frac{\operatorname{cov}(Y_{i_1 s_1}, Y_{i_2 s_2})}{\sqrt{\operatorname{cov}(Y_{i_1 s_1}, Y_{i_1 s_1})\operatorname{cov}(Y_{i_2 s_2}, Y_{i_2 s_2})}} \tag{11a}$$

where

$$\operatorname{cov}(Y_{i_1 s_1}, Y_{i_2 s_2}) = \frac{\sum_{j=1}^{m} w_j \left(Y_{i_1 s_1 j} - \mathbb{E}_j[Y_{i_1 s_1}]\right)\left(Y_{i_2 s_2 j} - \mathbb{E}_j[Y_{i_2 s_2}]\right)}{\sum_{j=1}^{m} w_j} \tag{11b}$$

$$\mathbb{E}_j[Y_{is}] = \frac{\sum_{j=1}^{m} w_j Y_{isj}}{\sum_{j=1}^{m} w_j} \tag{11c}$$

where $Y_{is}$is the intensity measure s at site i and w_j is the recurrence rate of the ground motion map j.The correlation error CE between a pair of intensity measures, s_1 and s_2 over a pair of sites i_1 and $i_2$is defined as

$$CE_{i_1 s_1 i_2 s_2} = \rho_{i_1 s_1 i_2 s_2} - \hat{\rho}_{i_1 s_1 i_2 s_2} \tag{12}$$

where $\hat{\rho}_{i_1 s_1 i_2 s_2}$ and $\rho_{i_1 s_1 i_2 s_2}$ denote the correlation coefficients for site pairs i_1 and i_2 and intensity measures, s_1 and s_2 based on the subset and the extensively sampled set of ground motion maps, respectively.

$$MACE = \frac{\sum_{i_1}^{n} \sum_{s_1}^{S} \sum_{i_2}^{n} \sum_{s_2}^{S} \left| CE_{i_1 s_1 i_2 s_2} \right|}{(nS)^2} \tag{13}$$

4. Case Study and Preliminary Results

We use a region (Figure 2) in Sonoma County, California to illustrate the proposed formulation. We divide the region into $2\ \mathrm{km} \times 2\ \mathrm{km}$ grid sites, which gives us $n = 380$ sites. For intensity measures, we use Peak Ground Acceleration (PGA), i.e., $S = 1$. For the optimization, we fix the relevant range of exceedance rates to be $10^{-1.5}$ to 10^{-4}. To get the initial earthquake rupture forecast, we use the OpenSHA event set generator application (Field et al. 2003) based on the Uniform California Earthquake Rupture Forecast (UCERF3 2013) with a wrapper code written in Python by the authors (Sharma 2023). For simplicity, we use a constant V_{s30} of $760\ \mathrm{m/s}$. As the ground motion model we use Campbell and Bozorgnia (2014). We model the spatial correlation using Baker and Jayaram (2008) and Loth and Baker (2013). The initial ERF returns $Q = 2091$ scenarios, and we then uniformly sample $b = 2$ maps from each of the 2091 scenarios.

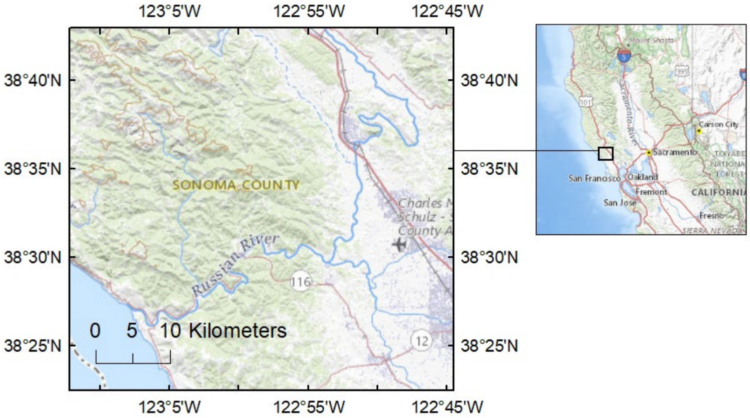

Figure 2. Region of interest.

4.1. Weighted PCA

The PGA data for $S = 380$ sites and $Q = 2091$ scenarios with $b = 2$ sample maps each, result in $\mathbf{Y} \in \mathbb{R}^{4182 \times 380}$. We then use weighted PCA and extract 8 principal components. Figure 3 shows the explained variance ratio for the principal components.

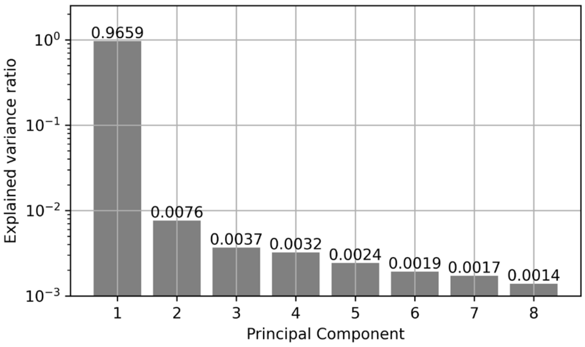

Figure 3. Explained variance ratio for principal components.

We see that the first principal component explains more than 95% of the total variance, which means that the 380 sites have a large amount of redundant information regarding the spatial variability of PGA. We note, however, that the study region is only $\sim 40 \times 40$ km, so most locations have highly correlated PGA values. A larger study region with lower spatial correlation, or a study with multiple intensity measures, would have greater contributions from more principal components.

4.2. PCA-based convex optimization

Now, for the optimal selection of the ground motion maps, we need to specify the values of $\boldsymbol{\vartheta}_l$ and $\boldsymbol{\lambda}_l$. For the same, we plot the geometric mean for exceedance rates $\bar{\lambda}_{is}(Y_{is} \geq y_j)$ as the response and corresponding exceedance rate of the principal component $\lambda_l(X_l \geq x_j)$ in Figure 4. The geometric mean $\bar{\lambda}_{is}(Y_{is} \geq y_j)$ captures the weight pattern for each map considering all the sites, comparing that with $\lambda_l(X_l \geq x_j)$ we can mimic the same weight pattern based on the principal component values. We observe that for the first principal component, the exceedance rate of the principal component closely follows the $1{:}1$ line with the geometric mean of the exceedance rate for the PGA for each site. In comparison, all the rest of the principal components are almost uncorrelated.

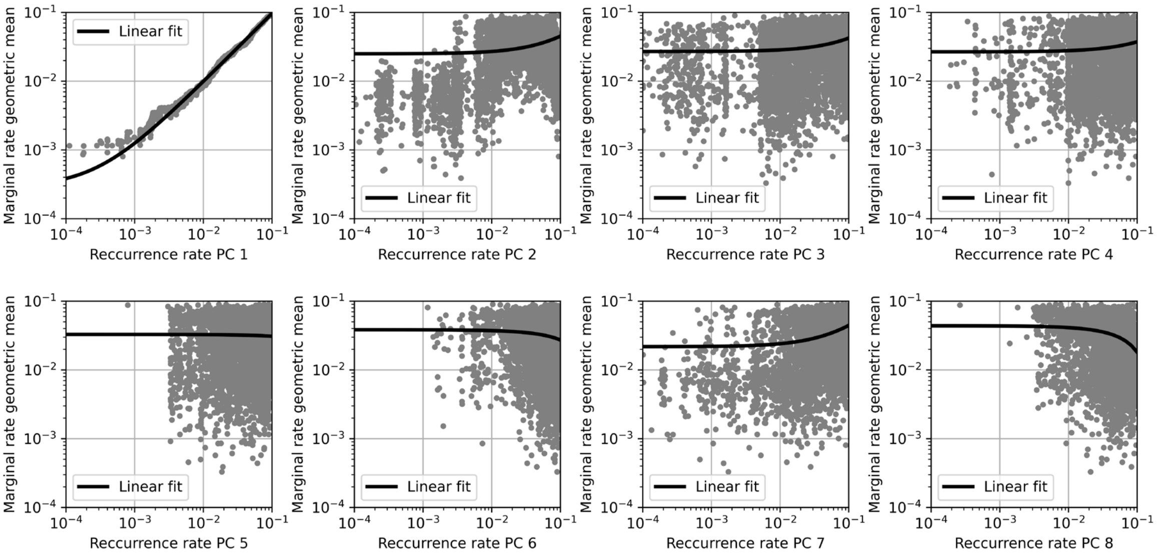

Figure 4. Mean exceedance rates in original space versus exceedance rates of principal components.

Hence for the current case study, we use for $l = 1$, $\boldsymbol{\vartheta}_l = (\boldsymbol{\lambda}_l)^{-1}$ and we use $\boldsymbol{\lambda}_l$ to follow 50 log scale distributed points over the range 10^{-4} to $10^{-1.5}$. Whereas for $l \in \{2, \dots, 8\}$, we use constant $\boldsymbol{\vartheta}_l$ as obtained from the intercepts of the linear fits in Figure 4, and we symmetrically distribute $\boldsymbol{\lambda}_l$ to linearly cover the quantiles of the principal component distributions at 50 points.

4.3. Evaluation metrics

We perform the original convex optimization from Section 2.2 and the updated weighted PCA-based optimization and compare their performance.

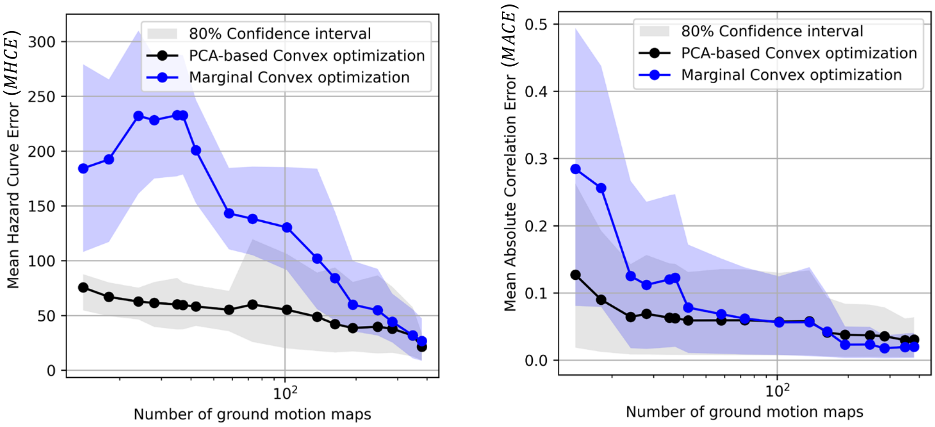

Figure 5. Mean hazard curve and absolute correlation error comparison for the two optimization formulations. The Marginal Convex optimization results are obtained from scenario selection using Equation 3, and PCA-based Convex optimization results are obtained from scenario selection using Equation 7.

From the preliminary results, we observe that PCA-based optimization shows substantial improvements compared to the original convex optimization formulation. The proposed formulation provides better and more stable performance in terms of the $MHCE$ for low number of ground motion maps. For $MACE$, again the PCA-based optimization performs better for a low number of ground motion maps, with nearly no difference when the number of selected maps are more than 100. We also note that we studied the impact of PCA-based improvements in a convex optimization formulation, and hence the performance for $MHCE$ in absolute terms is not desirable. However, the proposed dimension reduction and correlation embedding novelties in the paper are equally applicable for other non-convex optimization and clustering based algorithms. Finally, because we used convex optimization the formulations discussed in the paper is highly efficient with requiring few seconds to obtain PCA based results for a selected set of hyperparameters.

5. Conclusion

This paper developed a computationally efficient formulation for selecting a limited number of earthquake scenarios that can represent the complete seismic hazard in a region. The proposed formulation can incorporate spatial and inter-intensity measure correlations while selecting earthquake scenarios to help maintain probabilistic consistency in scenario-based assessment of regional seismic risk. Specifically, the paper proposed a weighted PCA-based dimension reduction approach combined with convex optimization to select earthquake ground motion maps. Compared with a marginal convex optimization formulation that directly minimizes the sites' hazard curve errors, the weighted PCA step reduces the problem's dimensionality while embedding the correlations into orthogonal principal components. The paper also made improvements to the convex optimization constraints in the form of hyperparameter tuning similar to LASSO regression to improve the optimization performance for a lower number of maps. The paper then provided a case study to illustrate the proposed formulation and provide preliminary results for a comparative analysis of the formulation's performance in benchmark metrics from the literature. The proposed formulation shows substantial performance improvement for both the marginal distribution fit in terms of the Mean Hazard Curve Error and capturing correlation in terms of the Mean Absolute Correlation Error. The proposed formulation is computationally efficient, requiring a few seconds to select the ground motion maps. Furthermore, the dimension reduction and correlation embedding novelties in the paper are equally applicable for other non-convex optimization and clustering-based algorithms, which can be used to improve the efficiency and accuracy of other algorithms better than convex optimization. The code and case study data presented in the paper are available at the authors' GitHub accounts (Sharma 2023).

6. References

Baker, J.W. and Jayaram, N., 2008. Correlation of spectral acceleration values from NGA ground motion models. *Earthquake Spectra*, *24*(1), pp.299-317.

Campbell, K.W. and Bozorgnia, Y., 2014. NGA-West2 ground motion model for the average horizontal components of PGA, PGV, and 5% damped linear acceleration response spectra. *Earthquake Spectra*, *30*(3), pp.1087-1115.

Chang, S.E. and Nojima, N., 2001. Measuring post-disaster transportation system performance: the 1995 Kobe earthquake in comparative perspective. *Transportation research part A: policy and practice*, *35*(6), pp.475-494.

Chang, S.E., 2003. Evaluating disaster mitigations: Methodology for urban infrastructure systems. *Natural Hazards Review*, *4*(4), pp.186-196.

Corotis, R. B., Bonstrom, H. and Porter K. 2012. Overcoming Public and Political Challenges for Natural Hazard Risk Investment Decisions. *IDRiM Journal*, 2(1), pp. 26-48.

Crowley, H. and Bommer, J.J., 2006. Modelling seismic hazard in earthquake loss models with spatially distributed exposure. *Bulletin of Earthquake Engineering*, *4*, pp.249-273.

Delchambre, L., 2015. Weighted principal component analysis: a weighted covariance eigen decomposition approach. *Monthly Notices of the Royal Astronomical Society*, *446*(4), pp.3545-3555.

Ebel, J.E. and Kafka, A.L., 1999. A Monte Carlo approach to seismic hazard analysis. *Bulletin of the Seismological Society of America*, *89*(4), pp.854-866.

Ellingwood, B.R. and Kinali, K., 2009. Quantifying and communicating uncertainty in seismic risk assessment. *Structural safety*, *31*(2), pp.179-187.

Elnashai, A.S., Cleveland, L.J., Jefferson, T. and Harrald, J., 2008. Impact of Earthquakes on the Central USA. *MAE Center Report 08-02*.

Field, E.H., Jordan, T.H. and Cornell, C.A., 2003. OpenSHA: A developing community-modeling environment for seismic hazard analysis. *Seismological Research Letters*, *74*(4), pp.406-419.

Han, Y. and Davidson, R.A., 2012. Probabilistic seismic hazard analysis for spatially distributed infrastructure. *Earthquake Engineering & Structural Dynamics*, *41*(15), pp.2141-2158.

Hotelling, H., 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, *24*(6), p.417.

Jayaram, N. and Baker, J.W., 2010. Efficient sampling and data reduction techniques for probabilistic seismic lifeline risk assessment. *Earthquake Engineering & Structural Dynamics*, *39*(10), pp.1109-1131.

Kavvada, I., Moura, S., Horvath, A. and Abrahamson, N., 2022. Probabilistic seismic hazard analysis for spatially distributed infrastructure considering the correlation of spectral acceleration across spectral periods. *Earthquake Spectra*, *38*(2), pp.1148-1175.

Loth, C. and Baker, J.W., 2013. A spatial cross-correlation model of spectral accelerations at multiple periods. *Earthquake Engineering & Structural Dynamics*, *42*(3), pp.397-417.

Manzour, H., Davidson, R.A., Horspool, N. and Nozick, L.K., 2016. Seismic hazard and loss analysis for spatially distributed infrastructure in Christchurch, New Zealand. *Earthquake Spectra*, *32*(2), pp.697-712.

Markhvida, M., Ceferino, L. and Baker, J.W., 2018. Modeling spatially correlated spectral accelerations at multiple periods using principal component analysis and geostatistics. *Earthquake Engineering & Structural Dynamics*, *47*(5), pp.1107-1123.

Miller, M. and Baker, J., 2015. Ground-motion intensity and damage map selection for probabilistic infrastructure network risk assessment using optimization. *Earthquake Engineering & Structural Dynamics*, *44*(7), pp.1139-1156.

Rose, A. and Lim, D., 2002. Business interruption losses from natural hazards: conceptual and methodological issues in the case of the Northridge earthquake. *Global Environmental Change Part B: Environmental Hazards*, *4*(1), pp.1-14.

Rosero-Velásquez, H. and Straub, D., 2022. Selection of representative natural hazard scenarios for engineering systems. *Earthquake Engineering & Structural Dynamics*, *51*(15), pp.3680-3700.

Sharma, N., 2023. Package for regional probabilistic seismic hazard analysis, pypsha 0.0.3. Available at: https://github.com/neetesh-nks/pypsha.

Soleimani, N., Davidson, R.A., Davis, C., O'Rourke, T.D. and Nozick, L.K., 2021. Multihazard scenarios for regional seismic risk assessment of spatially distributed infrastructure. *Journal of Infrastructure Systems*, *27*(1), p.04021001.

Tabandeh, A., Sharma, N. and Gardoni, P., 2022. Uncertainty propagation in risk and resilience analysis of hierarchical systems. *Reliability Engineering & System Safety*, *219*, p.108208.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), pp.267-288.

VanderPlas, J., 2016. Weighted Principal Component Analysis, wpca 0.1. Available at: http://github.com/jakevdp/wpca/.

Uniform California Earthquake Rupture ForecastVersion 3 (UCERF3) 2013. Cumulative fault participation MFDs. Available at: https://pubs.usgs.gov/of/2013/1165/data/UCERF3_SupplementalFiles/UCERF3.3/Model/FaultMFDs/ParticipationCumulative/index.html.

Vaziri, P., Davidson, R., Apivatanagul, P. and Nozick, L., 2012. Identification of optimization-based probabilistic earthquake scenarios for regional loss estimation. *Journal of Earthquake Engineering*, *16*(2), pp.296-315.