EARTHQUAKE ENGINEERING AND STRUCTURAL DYNAMICS
Earthquake Engng Struct. Dyn. 2010; **39**:1109–1131
Published online 27 January 2010 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/eqe.988

Efficient sampling and data reduction techniques for probabilistic seismic lifeline risk assessment

Nirmal Jayaram*,† and Jack W. Baker

Department of Civil and Environmental Engineering, Stanford University, Stanford, CA 94305-4020, U.S.A.

SUMMARY

Probabilistic seismic risk assessment for spatially distributed lifelines is less straightforward than for individual structures. While procedures such as the 'PEER framework' have been developed for risk assessment of individual structures, these are not easily applicable to distributed lifeline systems, due to difficulties in describing ground-motion intensity (e.g. spectral acceleration) over a region (in contrast to ground-motion intensity at a single site, which is easily quantified using Probabilistic Seismic Hazard Analysis), and since the link between the ground-motion intensities and lifeline performance is usually not available in closed form. As a result, Monte Carlo simulation (MCS) and its variants are well suited for characterizing ground motions and computing resulting losses to lifelines. This paper proposes a simulation-based framework for developing a small but stochastically representative catalog of earthquake ground-motion intensity maps that can be used for lifeline risk assessment. In this framework, Importance Sampling is used to preferentially sample 'important' ground-motion intensity maps, and K-Means Clustering is used to identify and combine redundant maps in order to obtain a small catalog. The effects of sampling and clustering are accounted for through a weighting on each remaining map, so that the resulting catalog is still a probabilistically correct representation. The feasibility of the proposed simulation framework is illustrated by using it to assess the seismic risk of a simplified model of the San Francisco Bay Area transportation network. A catalog of just 150 intensity maps is generated to represent hazard at 1038 sites from 10 regional fault segments causing earthquakes with magnitudes between five and eight. The risk estimates obtained using these maps are consistent with those obtained using conventional MCS utilizing many orders of magnitudes more ground-motion intensity maps. Therefore, the proposed technique can be used to drastically reduce the computational expense of a simulation-based risk assessment, without compromising the accuracy of the risk estimates. This will facilitate computationally intensive risk analysis of systems such as transportation networks. Finally, the study shows that the uncertainties in the ground-motion intensities and the spatial correlations between ground-motion intensities at various sites must be modeled in order to obtain unbiased estimates of lifeline risk.

Received 29 June 2009; Revised 18 November 2009; Accepted 25 November 2009

KEY WORDS: lifeline; seismic risk; efficient simulation; importance sampling; k-means clustering; transportation network

*Correspondence to: Nirmal Jayaram, Department of Civil and Environmental Engineering, Stanford University, Stanford, CA 94305-4020, U.S.A.
†E-mail: nirmalj@stanford.edu

Contract/grant sponsor: Stanford Graduate Fellowship

1. INTRODUCTION

Lifelines are large, geographically distributed systems that are essential support systems for any society. Owing to their known vulnerabilities, it is important to proactively assess and mitigate the seismic risk of lifelines. For instance, the Northridge earthquake caused over $1.5 billion in business interruption losses ascribed to transportation network damage [1]. The city of Los Angeles suffered a power blackout and $75 million of power-outage-related losses as a result of the earthquake, e.g. [2]. Recently, the analytical Pacific Earthquake Engineering Research Center (PEER) loss analysis framework has been used to perform risk assessment for a single structure at a given site, by estimating the site ground-motion hazard and assessing probable losses using the hazard information, e.g. [3]. Lifeline risk assessment, however, is based on a large vector of ground-motion intensities (e.g. spectral accelerations at all lifeline component locations). The intensities also show significant spatial correlation, which needs to be carefully modeled in order to accurately assess the seismic risk. Further, the link between the ground-motion intensities at the sites and the performance of the lifeline is usually not available in closed form. For instance, the travel time of vehicles in a transportation network, a commonly used performance measure, is only obtained using an optimization procedure rather than being a closed-form function of the ground-motion intensities. These additional complexities make it difficult to use the PEER framework for lifeline risk assessment. There are some analytical approaches that are sometimes used for lifeline risk assessment, e.g. [4, 5], but those are generally applicable to only specific classes of lifeline reliability problems. Hence, many past research works use simulation-based approaches instead of analytical approaches for lifeline risk assessment, e.g. [6–9]. One simple simulation-based approach involves studying the performance of lifelines under those earthquake scenarios that may dominate the hazard in the region of interest, e.g. [10]. While this approach is more tractable, it does not capture seismic hazard uncertainties in the way a Probabilistic Seismic Hazard Analysis (PSHA)-based framework would. Further, it is not easy to identify the earthquake scenario that dominates the hazard at the loss levels of interest [11]. A more comprehensive approach uses Monte Carlo simulation (MCS) to probabilistically generate ground-motion intensity maps (also referred to as intensity maps in this paper), considering all possible earthquake scenarios that could occur in the region, and then use these for the risk assessment. Ground-motion intensities are generated using an existing ground-motion model, which is described below.

We model the ground-motion intensity at a site as

$$\ln(S_{a_{ij}}) = \ln(\bar{S}_{a_{ij}}) + \sigma_{ij}\varepsilon_{ij} + \tau_{ij}\eta_{ij} \tag{1}$$

where $S_{a_{ij}}$ denotes the spectral acceleration (at the period of interest) at site i during earthquake j; $\bar{S}_{a_{ij}}$ denotes the predicted (by the ground-motion model) median spectral acceleration that depends on parameters such as magnitude, distance, period and local-site conditions; ε_{ij} denotes the normalized intra-event residual and η_{ij} denotes the normalized inter-event residual. Both ε_{ij} and η_{ij} are univariate normal random variables with zero mean and unit standard deviation. σ_{ij} and τ_{ij} are standard deviation terms that are estimated as part of the ground-motion model and are functions of the spectral period of interest, and in some models also functions of the earthquake magnitude and the distance of the site from the rupture. The term $\sigma_{ij}\varepsilon_{ij}$ is called the intra-event residual and the term $\tau_{ij}\eta_{ij}$ is called the inter-event residual. The inter-event residual is a constant across all the sites for a given earthquake.

DOI: 10.1002/eqe

Crowley and Bommer [7] describe the following MCS approach to simulate intensity maps using Equation (1):

Step 1: Use MCS to generate earthquakes of varying magnitudes on the active faults in the region, considering appropriate magnitude–recurrence relationships (e.g. the Gutenberg–Richter relationship).

Step 2: Using a ground-motion model (Equation (1)) obtain the median ground-motion intensities ($\bar{S}_{a_{ij}}$) and the standard deviations of the inter-event and the intra-event residuals (σ_{ij} and τ_{ij}) at all the sites.

Step 3: Generate the normalized inter-event residual term (η_{ij}) by sampling from the univariate normal distribution.

Step 4: Simulate the normalized intra-event residuals (ε_{ij}'s) using the parameters predicted by the ground-motion model. Jayaram and Baker [12] showed that a vector of spatially distributed normalized intra-event residuals $\boldsymbol{\varepsilon_j} = (\varepsilon_{1j}, \varepsilon_{2j}, \ldots, \varepsilon_{pj})$ follows a multivariate normal distribution. Hence, the distribution of $\boldsymbol{\varepsilon_j}$ can be completely defined using the mean (zero) and standard deviation (one) of ε_{ij}, and the correlation between all $\varepsilon_{i_1 j}$ and $\varepsilon_{i_2 j}$ pairs. The correlations between the residuals can be obtained from a predictive model calibrated using past ground-motion intensity observations [13, 14].

Step 5: Combine the median intensities, the normalized intra-event residuals and the normalized inter-event residual for each earthquake in accordance with Equation (1) to obtain ground-motion intensity maps (i.e. obtain $\mathbf{S}_{\mathbf{a}_j} = (S_{a_{1j}}, S_{a_{2j}}, \ldots, S_{a_{pj}})$).

Crowley and Bommer [7] used the above-mentioned approach to generate multiple earthquake scenarios that were then used for the loss assessment of a portfolio of buildings. They found that the results differed significantly from those obtained using other approximate approaches (e.g. using PSHA to obtain individual site hazard and loss exceedance curves, which are then heuristically combined to obtain the overall portfolio loss exceedance curve). Crowley and Bommer [7], however, ignored the spatial correlations of ε_{ij}'s when simulating intensity maps. Further, they used conventional MCS (i.e. brute-force MCS or random MCS), which is computationally inefficient because large magnitude events and above-average ground-motion intensities are considerably more important than small magnitude events and small ground-motion intensities while modeling lifeline risks, but these are infrequently sampled in conventional MCS. Kiremidjian *et al.* [8] improved the simulation process by preferentially simulating large magnitudes using importance sampling (IS). The normalized residuals (ε_{ij} and η_{ij}), however, were simulated using conventional MCS.

Shiraki *et al.* [9] also used an MCS-based approach to estimate earthquake-induced delays in a transportation network. They generated a catalog of 47 earthquakes and corresponding intensity maps for the Los Angeles area and assigned probabilities to these earthquakes such that the site hazard curves obtained using this catalog match with the known local-site hazard curves obtained from PSHA. In other words, the probabilities of the scenario earthquakes were made to be hazard consistent. Only median peak ground accelerations were used to produce the ground-motion intensity maps corresponding to the scenario earthquakes, however, and the known variability about these medians was ignored. While this approach is highly computationally efficient on account of the use of a small catalog of earthquakes, the selection of earthquakes is a somewhat subjective process, and the assignment of probabilities is based on hazard consistency rather than on actual event likelihoods. Moreover, the procedure does not capture the effect of the uncertainties in ground-motion intensities.

DOI: 10.1002/eqe

The current research work develops an importance sampling (IS)-based framework to efficiently sample important magnitudes and ground-motion residuals. It is seen that the number of IS simulations is about two orders of magnitude smaller than the number of MCSs required to obtain equally accurate lifeline loss estimates. Despite this improvement with respect to the performance of the conventional MCS approach, the number of IS intensity maps required for risk assessment is still likely to be an inconveniently large number. As a result, the K-means clustering technique is used to further reduce the number of intensity maps required for risk assessment by over an order of magnitude. The feasibility of the proposed framework is illustrated by assessing the seismic risk of an aggregated form of the San Francisco Bay Area transportation network using a sampled catalog of 150 intensity maps. The resulting risk estimates are shown to be in good agreement with those obtained using the conventional MCS approach (the benchmark method).

2. SIMULATION OF GROUND-MOTION INTENSITY MAPS USING IS

This section provides a description of the IS technique used in the current work to efficiently simulate ground-motion intensity maps. IS is a technique used to evaluate functions of random variables with a certain probability density function (PDF) using samples from an alternate density function [15]. This technique is explained in detail in Section 2.1. Sections 2.2–2.4 describe the application of IS to the simulation of ground-motion intensity maps, which involves probabilistically sampling a catalog of earthquake magnitudes and rupture locations (which are required for computing the median ground-motion intensities), the normalized inter-event residuals and the normalized intra-event residuals (Equation (1)).

2.1. Importance sampling procedure

Let $f(x)$ be a PDF defined over domain D for random variable X. Define an integral H as follows:

$$H = \int_D q(x) f(x)\,\mathrm{d}x \tag{2}$$

where $q(x)$ is an arbitrary function of x. The integral can be rewritten as follows:

$$H = \int_D q(x) \frac{f(x)}{g(x)} g(x)\,\mathrm{d}x \tag{3}$$

where $g(x)$ is any probability density assuming non-zero values over the same domain D. The term $f(x)/g(x)$ is called the IS weight.

Based on Equation (2), the integral H can be estimated using conventional MCS as follows:

$$\hat{H} = \frac{1}{n} \sum_{i=1}^{n} q(x_i) \tag{4}$$

where $\hat{H}$ is an estimate of H and $x_1, \ldots, x_n$ are n realizations of the random variable X obtained using $f(x)$. The IS procedure involves estimating the integral H using the alternate density $g(x)$ as follows (based on Equation (3)):

$$\hat{H} = \frac{1}{r} \sum_{i=1}^{r} q(y_i) \frac{f(y_i)}{g(y_i)} \tag{5}$$

DOI: 10.1002/eqe

where $y_1, \ldots, y_r$ are r realizations from $g(y)$, and $f(y_i)/g(y_i)$ is a weighting function (the IS weight) that accounts for the fact that the realizations are based on the alternate density $g(y)$ rather than the original density $f(y)$.

While Equations (4) and (5) provide two methods of estimating the same integral H, it can be shown that the variance of the estimate $\hat{H}$ obtained using Equation (5) can be made very small if an appropriate alternate density function $g(x)$ is chosen [15]. As a result of this variance reduction, the required number of IS realizations (r) is much smaller than the required number of conventional MCS realizations (n) for an equally reliable (i.e. same variance) estimate $\hat{H}$.

Intuitively, the density $g(x)$ should be such that the samples from $g(x)$ are concentrated in regions where the function $q(x)$ is 'rough'. This will ensure fine sampling in regions that ultimately determine the accuracy of the estimate and coarse sampling elsewhere. The challenge in implementing IS lies in choosing this alternate density $g(x)$. Useful alternate densities for this application are provided in the following subsections.

2.2. Simulation of earthquake catalogs

Let n_f denote the number of active faults in the region of interest and v_j denote the annual recurrence rate of earthquakes on fault j. Let $f_j(m)$ denote the density function for magnitudes of earthquakes on fault j. Let $f(m)$ denote the density function for the magnitude of an earthquake on any of the n_f faults (i.e. this density function models the distribution of earthquakes resulting from all the faults). Using the theorem of total probability, $f(m)$ can be computed as follows:

$$f(m) = \frac{\sum_{j=1}^{n_f} v_j f_j(m)}{\sum_{j=1}^{n_f} v_j} \tag{6}$$

In the event of an earthquake of magnitude m on a random fault, let $P_j(m)$ denote the probability that the earthquake rupture lies on fault j. The $P_j(m)$'s can be calculated using Bayes' theorem as follows:

$$P_j(m) = \frac{v_j f_j(m)}{\sum_{j=1}^{n_f} v_j f_j(m)} \tag{7}$$

A conventional MCS approach would use the density function $f(m)$ to simulate earthquake magnitudes, although this approach will result in a large number of small magnitude events since such events are considerably more probable than large magnitude events. This is not efficient since lifeline losses due to frequent small events are less important than those due to rare large events (although not negligible, so they cannot be ignored). It is desirable to improve the computational efficiency of the risk assessment process without compromising the accuracy of the estimates by using the IS technique described in Section 2.1 to preferentially sample large events while still ensuring that the simulated events are 'stochastically representative'. In other words, the magnitudes are simulated from a sampling distribution $g(m)$ (rather than $f(m)$), which is chosen to have a high probability of producing large magnitude events.

Let $m_{\min}$ and $m_{\max}$ denote the range of magnitudes of interest. This range $[m_{\min}, m_{\max}]$ can be stratified into n_m partitions as follows:

$$[m_{\min}, m_{\max}] = [m_{\min}, m_2) \cup [m_2, m_3) \cup \cdots \cup [m_{n_m}, m_{\max}] \tag{8}$$

DOI: 10.1002/eqe

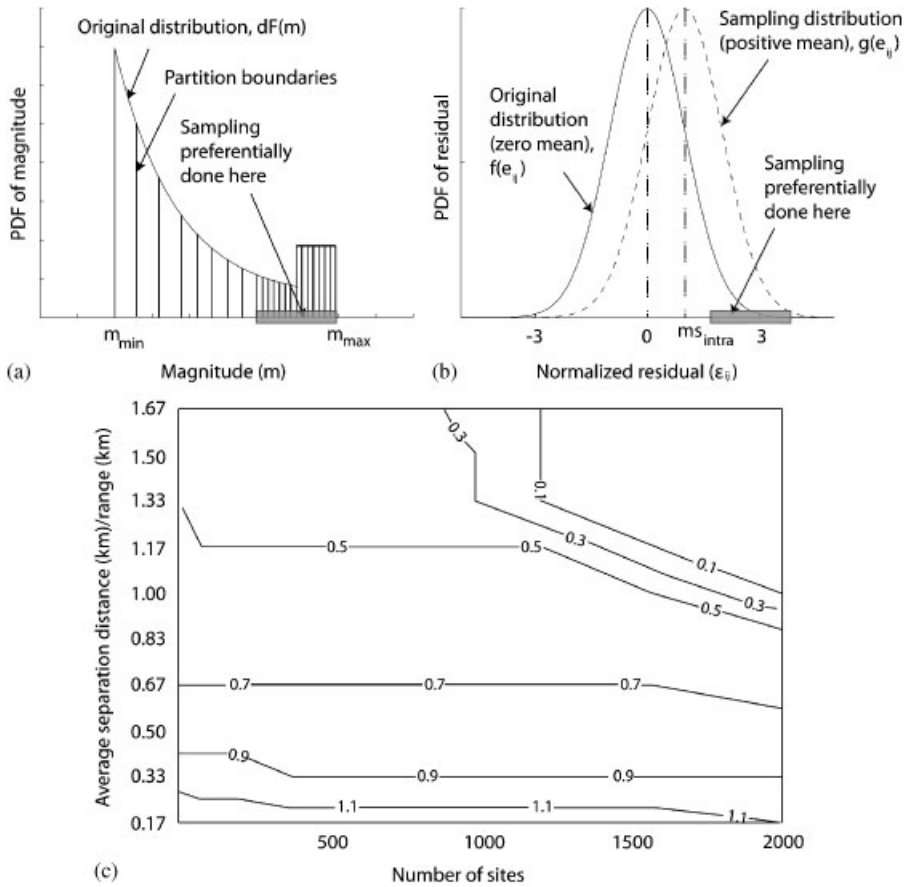

Figure 1. Importance sampling density functions for: (a) magnitude; (b) normalized intra-event residual; and (c) recommended mean-shift as a function of the average number of sites and the average site-to-site distance normalized by the range of the spatial correlation model.

In the current work, the partitions are chosen such that the width of the interval (i.e. $m_{k+1} - m_k$) is large at small magnitudes and small at large magnitudes (Figure 1(a)). A single magnitude is randomly sampled from each partition using the magnitude density function $f(m)$, thereby obtaining n_m realizations of the magnitudes. Since, the partitions are chosen to have small widths at large magnitudes, there is naturally a larger number of realizations of large magnitude events. In this case, the sampling distribution $g(m)$ is not explicit, but rather is implicitly defined by the magnitude selection partitioning. This procedure, sometimes called stratified sampling, has the advantage of forcing the inclusion of specified subsets of the random variable while maintaining the probabilistic character of random sampling [15].

The IS weight $f(m)/g(m)$ can be obtained by noting that the sampling distribution assigns equal weight to all the chosen partitions $(1/n_m)$, while the actual probability of a magnitude lying in a partition (m_k, m_{k+1}) is obtained by integrating the density function $f(m)$. Hence, the IS weight

DOI: 10.1002/eqe

for a magnitude m chosen from the kth partition is computed as follows:

$$\frac{f(m)}{g(m)} = \frac{\int_{m_k}^{m_{k+1}} f(m)\,\mathrm{d}m}{1/n_m} \tag{9}$$

Once the magnitudes are sampled using IS, the rupture locations can be obtained by sampling faults using fault probabilities $P_j(m)$ (Equation (7)). It is to be noted that $P_j(m)$ will be non-zero only if the maximum allowable magnitude on fault j exceeds m. Let $n_f(m)$ denote all such faults with non-zero values of $P_j(m)$. If $n_f(m)$ is small (around 10), a more efficient sampling approach will be to consider each of those $n_f(m)$ faults to be the source of the earthquake and consider $n_f(m)$ different earthquakes of the same simulated magnitude. It is to be noted that this fault sampling procedure is similar to the IS of magnitudes. The IS weight for fault j chosen by this procedure is computed as follows:

$$\frac{f(j|m)}{g(j|m)} = \frac{P_j(m)}{1/n_f(m)} \tag{10}$$

where $f(j|m)$ and $g(j|m)$ denote the original and the alternate (implicit) probability mass functions for fault j given an earthquake of magnitude m.

2.3. Simulation of normalized intra-event residuals

The set of normalized intra-event residuals at p sites of interest, $\boldsymbol{\varepsilon_j} = (\varepsilon_{1j}, \varepsilon_{2j}, \ldots, \varepsilon_{pj})$, follows a multivariate normal distribution $f(\boldsymbol{\varepsilon_j})$ [12]. The mean of $\boldsymbol{\varepsilon_j}$ is the zero vector of size p, while the variance of each ε_{ij} equals one. The correlation between the residuals at two sites is a function of the separation between the sites, and can be obtained from a spatial correlation model. In this work, the correlation coefficient between the residuals at two sites i_1 and i_2 separated by h km is computed using the following equation, which was calibrated using empirical observations [13]:

$$\rho_{\varepsilon_{i_1 j}, \varepsilon_{i_2 j}}(h) = \exp(-3h/R) \tag{11}$$

where R controls the rate of decay of spatial correlation and is called the 'range' of the correlation model. The range depends on the intensity measure being used. In this work, the intensity measure of interest is the spectral acceleration corresponding to a period of 1 s, and the corresponding value of R equals 26 km.

While a conventional MCS approach can be used to obtain realizations of $\boldsymbol{\varepsilon_j}$ using $f(\mathbf{e})$ [15], this will result in a large number of near-zero (i.e. near-mean) residuals and few realizations from the upper and the lower tails. This is inefficient since for the purposes of lifeline risk assessment it is often of interest to study the upper tail (i.e. the $\boldsymbol{\varepsilon_j}$ values that produce large intensities), which is not sampled adequately in the conventional MCS approach. An efficient alternate sampling density $g(\mathbf{e})$ is a multivariate normal density with the same variance and correlation structure as $f(\mathbf{e})$, but with positive means for all $\varepsilon'_{ij}s$ (i.e. a positive mean for the marginal distribution of each intra-event residual). In other words, the mean vector of $g(\mathbf{e})$ is the p-dimensional vector $\boldsymbol{ms}_{\text{intra}} = (ms_{\text{intra}}, ms_{\text{intra}}, \ldots, ms_{\text{intra}})$. Sampling normalized intra-event residuals from this distribution $g(\mathbf{e})$, which has a positive mean, will produce more realizations of large normalized intra-event residuals. Figure 1(b) shows the original and sampling marginal distributions for one particular ε_{ij}. It is to be

DOI: 10.1002/eqe

noted that this particular choice of the sampling distribution results in IS weights that are simple to estimate. The IS weights can be estimated as follows:

$$\frac{f(\mathbf{e})}{g(\mathbf{e})} = \exp\left(\frac{1}{2}(\mathbf{e}-\boldsymbol{ms}_{\text{intra}})'\Sigma^{-1}(\mathbf{e}-\boldsymbol{ms}_{\text{intra}}) - \frac{1}{2}\mathbf{e}'\Sigma^{-1}\mathbf{e}\right) \tag{12}$$

where Σ denotes the covariance matrix of $\boldsymbol{\varepsilon_j}$.

The positive mean of $g(\mathbf{e})$ will ensure that the realizations from $g(\mathbf{e})$ will tend to be larger than the realizations from $f(\mathbf{e})$. It is, however, important to choose a reasonable value of the mean-shift ms_{intra} to ensure adequate preferential sampling of large $\boldsymbol{\varepsilon_j}$'s, while avoiding sets of extremely large normalized intra-event residuals that will make the simulated intensity map so improbable as to be irrelevant. The process of selecting a reasonable value of ms_{intra} is described below.

The first step in fixing the value of ms_{intra} is to note that the preferred value depends predominantly on three factors, namely, the extent of spatial correlations (measured by the range parameter R in Equation (11)), the average site-to-site separation distance in the lifeline network being studied and the number of sites in the network. If sites are close to one another and if the spatial correlations are significant, the correlations between the residuals permit a larger mean-shift as it is reasonably likely to observe simultaneously large values of positively correlated random variables. Similarly, the presence of fewer sites permits larger mean-shifts since it is more likely to observe jointly large values of residuals over a few sites than over a large number of sites. Hence, it is intended to determine the preferred mean-shifts as a function of the number of sites and the average site-to-site separation distances normalized by the range parameter. This is done by simulating the normalized intra-event residuals in hypothetical analysis cases with varying numbers of sites and varying average site separation distances, considering several feasible mean-shifts in each case. The feasibility of the resulting residuals (i.e. whether the simulated set of residuals is reasonably probable) is then studied using the resulting IS weights. Based on extensive sensitivity analysis, the authors found that the best results are obtained when 30% of the IS weights fall below 0.1, if exceedance rates larger than 10^{-6} are of interest. The preferred mean-shifts are determined for each case based on this criterion, and are plotted in Figure 1(c). This figure will enable users to avoid an extremely computationally expensive search for an appropriate sampling distribution in a given analysis case. Incidentally the figure shows that the mean-shift increases with average site separation distance and decreases with the number of sites. This validates the above-mentioned statement that larger site separation distances and fewer sites permit larger mean-shifts.

2.4. Simulation of normalized inter-event residuals

Following standard conventions, since the inter-event residual is a constant across all the sites during a single earthquake, e.g. [16], the simulated normalized inter-event residuals should satisfy the following relation (which does not assume that the τ_{ij}'s are equal in order to be compatible with ground-motion models such as that of Abrahamson and Silva [17]):

$$\eta_{ij} = \frac{\tau_{1j}}{\tau_{ij}}\eta_{1j} \quad \forall j \tag{13}$$

Thus, the normalized inter-event residuals can be simulated by first simulating η_{1j} from a univariate normal distribution with zero mean and unit standard deviation, and by subsequently evaluating

DOI: 10.1002/eqe

other normalized inter-event residuals using Equation (13). The IS procedure for η_{1j} is similar to that for $\boldsymbol{\varepsilon_j}$, except that the alternate sampling distribution is univariate normal rather than multivariate normal, and has unit standard deviation and a positive mean ms_{inter}. The likelihood ratio in this case is

$$\frac{f(t)}{g(t)} = \exp\left(\frac{1}{2}(t - ms_{\text{inter}})^2 - \frac{1}{2}t^2\right) \tag{14}$$

where t denotes a realization of the normalized inter-event residual.

The authors have found that values of ms_{inter} between 0.5 and 1.0 produce an appropriate number of normalized inter-event residuals from the tail of the distribution.

3. LIFELINE RISK ASSESSMENT

In this paper, it is intended to obtain the exceedance curve for a lifeline loss measure denoted L (e.g. travel-time delay in a transportation network) considering seismic hazard. The exceedance curve, which provides the annual exceedance rates of various values of L, is the product of the exceedance probability curve and the total recurrence rate of earthquakes exceeding the minimum considered magnitude on all faults.

$$v_{L \geqslant u} = \left(\sum_{j=1}^{n_f} v_j\right) P(L \geqslant u) \tag{15}$$

A simple way to compute the annual exceedance rates, while treating each fault separately, would be to compute $\sum_{j=1}^{n_f} v_j P(L_j \geqslant u)$, where $P(L_j \geqslant u)$ denotes the exceedance probability for fault j, and the v_j values account for unequal recurrence rates across faults. That approach is not possible here because the IS of Equation (9) makes separation by faults difficult. In Equation (15), $P(L \geqslant u)$ is the probability that the loss due to any earthquake event of interest (irrespective of the fault of occurrence) exceeds u. It can be computed using the simulated maps, and in that form already accounts for the individual $P(L_j \geqslant u)$ values and the v_j values.

3.1. Risk assessment based on realizations from MCS

If a catalog of n intensity maps obtained using the conventional MCS approach is used for the risk assessment, the empirical estimate of the exceedance probabilities ($\hat{P}(L \geqslant u)$) can be obtained as follows (from Equation (4)):

$$\hat{P}(L \geqslant u) = \frac{1}{n}\sum_{i=1}^{n} I(l_i \geqslant u) \tag{16}$$

where l_i is the loss level corresponding to intensity map i, and $I(l_i \geqslant u)$ is an indicator function which equals 1 if $l_i \geqslant u$ and 0 otherwise.

DOI: 10.1002/eqe

3.2. Risk assessment based on realizations from IS

The summand in Equation (16) can be evaluated using the approach described in Section 2. Assuming that a catalog of r IS-based intensity maps is used for evaluating the risk, the estimate of the exceedance probability curve can be obtained as follows (from Equation (5)):

$$\hat{P}(L \geqslant u) = \frac{1}{r} \sum_{i=1}^{r} I(l_i \geqslant u) \frac{f_S(i)}{g_S(i)} \tag{17}$$

where $f_S(i)/g_S(i)$ is the IS weight corresponding to scenario intensity map i, which can be evaluated as follows:

$$\frac{f_S(i)}{g_S(i)} = \frac{f(m)}{g(m)} \frac{f(j|m)}{g(j|m)} \frac{f(\mathbf{e})}{g(\mathbf{e})} \frac{f(t)}{g(t)} = \Lambda_i \tag{18}$$

where m, j, $\mathbf{e}$, t denote the magnitude, fault, normalized intra-event residuals and normalized inter-event residual corresponding to map i, respectively. The terms in Equation (18) can be obtained from Equations (9), (10), (12) and (14).

Equation (17) shows that the exceedance probability curve is obtained by weighting the indicator functions by the IS weights for the maps. In the remainder of the paper, this weight is denoted Λ_i as shown in Equation (18). Using this notation for weight, Equation (17) can be rewritten as follows:

$$\hat{P}(L \geqslant u) = \frac{1}{r} \sum_{i=1}^{r} I(l_i \geqslant u) \Lambda_i = \frac{\sum_{i=1}^{r} I(l_i \geqslant u) \Lambda_i}{\sum_{i=1}^{r} \Lambda_i} \tag{19}$$

The second equality in the above equation comes from the fact that $\sum_{i=1}^{r} \Lambda_i = r$, as seen by substituting $u = 0$ in the equation and noting that $\hat{P}(L \geqslant 0) = 1$.

The variance (var) of this estimate can be shown to be

$$\mathrm{var}[\hat{P}(L \geqslant u)] = \frac{\sum_{i=1}^{r} [I(l_i \geqslant u) \Lambda_i - \hat{P}(L \geqslant u)]^2}{(\sum_{i=1}^{r} \Lambda_i)(\sum_{i=1}^{r} \Lambda_i - 1)} \tag{20}$$

4. DATA REDUCTION USING K-MEANS CLUSTERING

The use of IS causes a significant improvement in the computational efficiency of the simulation procedure, but the number of required IS intensity maps is still large and may pose a heavy computational burden. K-means clustering [18] is thus used as a data reduction technique in order to develop a smaller catalog of maps by 'clustering' simulated ground-motion intensity maps with similar properties (i.e. similar spectral acceleration values at the sites of interest). This data reduction procedure is also used in machine learning and signal processing, where it is called vector quantization [19].

K-means clustering groups a set of observations into K clusters such that the dissimilarity between the observations (typically measured by the Euclidean distance) within a cluster is minimized [18]. Let $\mathbf{S}_{\mathbf{a}_1}, \mathbf{S}_{\mathbf{a}_2}, \ldots, \mathbf{S}_{\mathbf{a}_r}$ denote r maps generated using IS to be clustered, where each

DOI: 10.1002/eqe

map $\mathbf{S_a}_j$ is a p-dimensional vector defined by $\mathbf{S_a}_j = [S_{a_1 j}, S_{a_2 j}, \ldots, S_{a_p j}]$. The K-means method groups these maps into clusters by minimizing V, which is defined as follows:

$$V = \sum_{i=1}^{K} \sum_{\mathbf{S_a}_j \in S_i} \| \mathbf{S_a}_j - \mathbf{C}_i \|^2 \tag{21}$$

where K denotes the number of clusters, S_i denotes the set of maps in cluster i, $\mathbf{C}_i = [C_{1i}, C_{2i}, \ldots, C_{pi}]$ is the cluster centroid obtained as the mean of all the maps in cluster i, and $\| \mathbf{S_a}_j - \mathbf{C}_i \|^2$ denotes the distance between the map $\mathbf{S_a}_j$ and the cluster centroid $\mathbf{C}_i$. If the Euclidean distance is adopted to measure dissimilarity, then the distance between $\mathbf{S_a}_j$ and $\mathbf{C}_i$ is computed as follows:

$$\| \mathbf{S_a}_j - \mathbf{C}_i \|^2 = \sum_{q=1}^{p} (S_{a_q j} - C_{qi})^2 \tag{22}$$

In its simplest version, the K-means algorithm is composed of the following four steps:

Step 1: Pick K maps to denote the initial cluster centroids. This selection can be done randomly.
Step 2: Assign each map to the cluster with the closest centroid.
Step 3: Recalculate the centroid of each cluster after the assignments.
Step 4: Repeat steps 2 and 3 until no more reassignments take place.

Once all the maps are clustered, the final catalog can be developed by selecting a single map from each cluster, which is used to represent all maps in that cluster on account of the similarity of the maps within a cluster. In other words, if the map selected from a cluster produces loss l, it is assumed that all other maps in the cluster produce the same loss l by virtue of similarity. The maps in this smaller catalog can be used in place of the maps generated using IS for the risk assessment (i.e. for evaluating $\hat{P}(L \geqslant u)$), which results in a dramatic improvement in the computational efficiency. This is particularly useful in applications where it is practically impossible to compute the loss measure L using more than K maps (where K equals a few hundreds). In such cases, the maps obtained using IS can be grouped using the K-means method into K clusters, and one map can be randomly selected from each cluster in order to obtain the catalog of intensity maps to be used for the risk assessment. This procedure allows us to select K strongly dissimilar intensity maps as part of the catalog (since the maps eliminated are similar to one of these K maps in the catalog), but will ensure that the catalog is 'stochastically representative'. Because only one map from each cluster is now used, the total weight associated with the map should be equal to the sum of the weights of all the maps in that cluster ($\sum_{i=1}^{r} \Lambda_i$). It is to be noted that even though the maps within a cluster are expected to be similar, for probabilistic consistency, a map must be chosen from a cluster with a probability proportional to its weight. Equation (19) can then be used with these sampled maps and the total weights to compute an exceedance probability curve using the catalog as follows:

$$\hat{P}(L \geqslant u) = \frac{\sum_{c=1}^{K} I(l_{(c)} \geqslant u)(\sum_{i \in c} \Lambda_i)}{\sum_{c=1}^{K} (\sum_{i \in c} \Lambda_i)} \tag{23}$$

where $l_{(c)}$ denotes the loss measure associated with the map selected from cluster c.

Appendix A shows that the exceedance probabilities obtained using Equation (23) will be unbiased. This and the fact that all the random variables are accounted for appropriately are the reason why the catalog selected is claimed to be stochastically representative. Incidentally,

DOI: 10.1002/eqe

the computational efficiency of this procedure can be improved with minor modifications to the clustering approach, as described in Appendix B.

5. APPLICATION: SEISMIC RISK ASSESSMENT OF THE SAN FRANCISCO BAY AREA TRANSPORTATION NETWORK

In this section, the San Francisco Bay Area transportation network is used to illustrate the feasibility of the proposed risk assessment framework. It is intended to show that the seismic risk estimated using the catalog of 150 intensity maps matches well with the seismic risk estimated using the conventional MCS framework and a much greater number of maps (which is the benchmark approach). The catalog size of 150 is chosen since it may be tractable to a real-life lifeline risk assessment problem. If reduced accuracy and reduced emphasis on very large losses is acceptable, the number of maps could be reduced even further.

5.1. Network data

The San Francisco Bay Area transportation network data are obtained from Stergiou and Kiremidjian [20]. Figure 2(a) shows the Metropolitan Transportation Commission (MTC) San Francisco Bay Area highway network, which includes 29 804 links (roads) and 10 647 nodes. The network also consists of 1125 bridges from the five counties of the Bay Area. Stergiou and Kiremidjian [20] classified these bridges based on their structural properties in accordance with the HAZUS [21] manual. (The HAZUS [21] fragility functions are used here only for illustrative purposes, and more realistic fragility functions can be used if applicable.) This classification is useful for estimating the structural damage to bridges due to various simulated intensity maps. The Bay Area network consists of a total of 1120 transportation analysis zones (TAZ), which are used to predict the trip demand in specific geographic areas. The origin-destination data provided by Stergiou and Kiremidjian [20] were obtained from the 1990 MTC household survey [22].

Analyzing the performance of a network as large and complex as the San Francisco Bay Area transportation network under a large number of scenarios is extremely computationally intensive. Therefore, an aggregated representation of the Bay Area network is used for this example application. The aggregated network consists predominantly of freeways and expressways, along with the ramps linking the freeways and expressways. The nodes are placed at locations where links intersect or change in characteristics (e.g. change in the number of lanes). The aggregated network comprises of 586 links and 310 nodes, and is shown in Figure 2(b). Of the 310 nodes, 46 are denoted centroidal nodes that act as origins and destinations for the traffic. These centroidal nodes are chosen from the centroidal nodes of the original network in such a way that they are spread out over the entire transportation network. The data from the 1990 MTC household survey are aggregated to obtain the traffic demands at each centroidal node. The aggregation involves assigning the traffic originating or culminating in any TAZ to its nearest centroidal node. Of the 1125 bridges in the original network, 1038 bridges lie on the links of the aggregated network and are considered in the risk assessment procedure.

While the performance of the aggregated network may or may not be similar to that of the full network, the aggregated network serves as a reasonably realistic and complex test case for the proposed framework, to demonstrate its feasibility. The goal is to demonstrate that the data reduction techniques proposed here produce the same exceedance curve as the more exhaustive

DOI: 10.1002/eqe

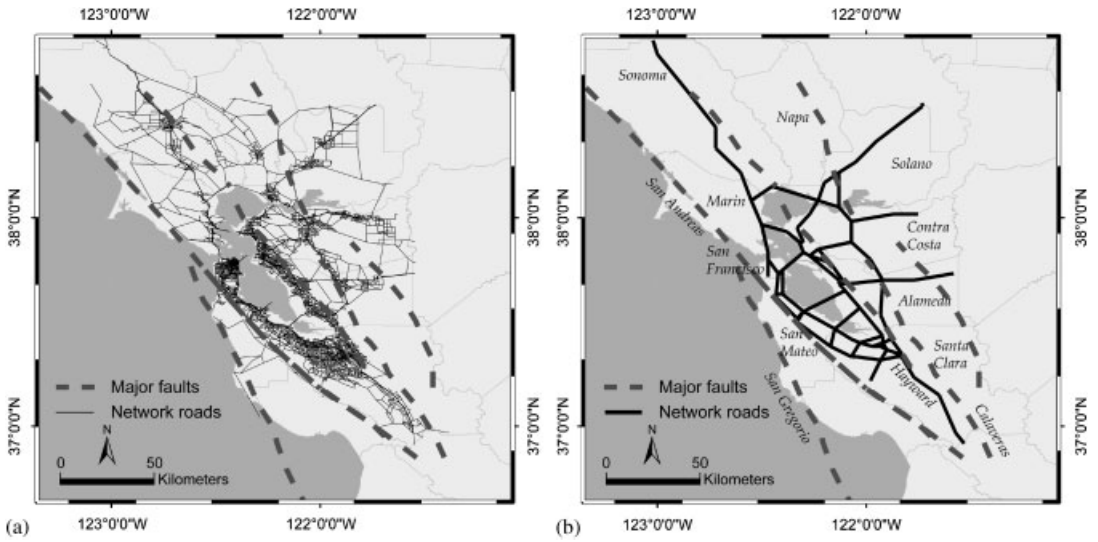

Figure 2. (a) San Francisco Bay Area transportation network and (b) aggregated network.

MCS. The simplified network is simple enough that MCS is feasible, but still retains the spatial distribution and network effects that are characteristic of more complex models. If the proposed techniques can be shown to be effective for this simplified model, then they can be used with more complex models where validation using MCS is not feasible.

5.2. *Transportation network loss measure*

A popular measure of network performance is the travel-time delay experienced by passengers in a network after an earthquake [9, 20]. The delay is computed as the difference between the total travel time in the network before and after an earthquake.

5.2.1. Estimating travel time in the network. The total travel time (T) in a network is estimated as follows:

$$T = \sum_{i \in \text{links}} x_i t_i(x_i) \tag{24}$$

where x_i denotes the traffic flow on link i and $t_i(x_i)$ denotes the travel time of an individual passenger on link i. The travel time on link i is obtained as follows [23]:

$$t_i(x_i) = t_i^f \left[1 + \alpha \left(\frac{x_i}{c_i} \right)^{\beta} \right] \tag{25}$$

where t_i^f denotes the free-flow link travel time (i.e. the travel time of a passenger if link i were to be empty), c_i is the capacity of link i, α and β are calibration parameters, taken as 0.15 and 4, respectively [9].

Travel times on transportation networks are usually computed using the user-equilibrium principle [24], which states that each individual user would follow the route that will minimize his or

DOI: 10.1002/eqe

her travel time. Based on the user-equilibrium principle, the link flows in the network are obtained by solving the following optimization problem:

$$\min \quad \sum_{i \in \{\text{links}\}} \int_0^{x_i} t_i(u)\,\mathrm{d}u \tag{26}$$

subject to the following constraints:

$$\sum_{j \in \text{paths}} f_j^{\text{od}} = Q^{\text{od}} \quad \forall o \in \{\text{org}\},\ \ \text{d} \in \{\text{dest}\} \tag{27}$$

$$x_i = \sum_{o \in \text{org}} \sum_{d \in \text{dest}} \sum_{j \in \text{paths}} f_j^{\text{od}} \delta_{ji}^{\text{od}} \quad \forall i \in \{\text{links}\} \tag{28}$$

$$f_j^{\text{od}} \geqslant 0 \quad \forall o \in \{\text{org}\},\ \ \text{d} \in \{\text{dest}\}\ \ , j \in \{\text{paths}\} \tag{29}$$

where f_j^{od} denotes the flow between origin o and destination d that passes through path j (here, a path denotes a set of links through which the flow between a specified origin and a specified destination occurs), Q^{od} denotes the desired flow between o and d, δ_{ji}^{od} is an indicator variable that equals 1 if the link i lies on path j and 0 otherwise, org denotes the set of all origins and dest denotes the set of all destinations. The current research work uses a popular solution technique for this optimization problem provided by Frank and Wolfe [25]. It is to be noted that there are also other travel time and traffic flow estimation techniques such as the dynamic user equilibrium formulation, e.g. [26] which could incorporate the non-equilibrium conditions which might exist after an earthquake.

5.2.2. Post-earthquake network performance. The current work assumes, for simplicity, that the post-earthquake demands equal the pre-earthquake demands even though this is known not to be true [27]. The changes in network performance after an earthquake are assumed to be due only to the delay and rerouting of traffic caused by structural damage to bridges. The damage states of the bridges are computed considering only the ground shaking, and other possible damage mechanisms such as liquefaction are not considered. The bridge fragility curves provided by HAZUS [21] are used to estimate the probability of a bridge being in a particular damage state (no damage, minor damage etc.) based on the simulated ground-motion intensity (spectral acceleration at 1 second) at the bridge site. These damage state probabilities are then used to simulate the damage state of the bridge following the earthquake. Damaged bridges cause reduced capacity in the link containing the bridge. The reduced capacities corresponding to the five different HAZUS damage states are 100% (no damage), 75% (slight damage/moderate damage) and 50% (extensive damage/collapse). The non-zero capacity corresponding to the bridge collapse damage state may seem surprising at first glance. This is based on the argument that there are alternate routes (apart from the freeways and highways considered in the model) that provide reduced access to transportation services in the event of a freeway or a highway closure [9]. Such redundancies are prevalent in most transportation networks.

A network can have several bridges in a single link, and in such cases, the link capacity is a function of the damage to all the bridges in the link. The current work assumes that the link capacity reduction equals the average of the capacity reductions attributable to each bridge in the link. This is a simplification, and further research is needed to handle the presence of multiple

DOI: 10.1002/eqe

bridges in a link. The post-earthquake network performance is then computed by solving the user-equilibrium problem using the new set of link capacities, and a new estimate of the total travel time in the network is obtained. It is to be noted that the current work estimates the performance of the network only immediately after an earthquake. The changes in the performance with network component restorations are not considered here for simplicity.

5.3. Ground-motion hazard

The San Francisco Bay Area seismicity information is obtained from USGS [28]. Ten active faults and fault segments are considered. The characteristic magnitude–recurrence relationship of Youngs and Coppersmith [29] is used to model $f(m)$ with the distribution parameters specified by the USGS, and 5.0 considered to be the lower bound magnitude of interest. The flattening of this magnitude distribution towards the maximum magnitude value (Figure 1) is to account for the higher probability of occurrence of the characteristic earthquake on the fault [29]. The ground-motion model of Boore and Atkinson [30] is used to obtain the median ground-motion intensities and the standard deviations of the residuals needed in Equation (1).

5.4. Results and discussion

5.4.1. Risk assessment using IS. The IS framework requires that the parameters of the sampling distribution for the magnitude and the residuals be chosen reasonably in order to obtain reliable results efficiently. The set of parameters includes the appropriate stratification for magnitudes, the mean-shift for normalized inter-event residuals (ms_{inter}) and the mean-shift for normalized intra-event residuals (ms_{intra}).

The stratification of the range of magnitudes is carried out so as to obtain a desired histogram of magnitudes. The partition width is chosen to be 0.3 between 5.0 and 6.5, 0.15 between 6.5 and 7.3 and 0.05 beyond 7.3. The results obtained using the simulations are not significantly affected by moderate variations in the partitions, suggesting that the stratification will be effective as long as it is chosen to preferentially sample large magnitudes. Normalized inter-event residuals are sampled using an ms_{inter} of 1.0. Using the procedure described earlier, the value of ms_{intra} is fixed at 0.3.

The loss measure of interest here is the travel-time delay (i.e. the variable L denoting loss measure in the previous section is the travel-time delay). Figure 3(a) shows the exceedance curve for travel-time delays obtained using the IS framework. This exceedance curve is obtained by sampling 25 magnitudes, each of which is then positioned on the active faults as described in Section 2.2, and 50 sets of inter and intra-event residuals for each magnitude-location pair (resulting in a total of 12 500 maps). To validate the IS, an exceedance curve is also estimated using the benchmark method (MCS). Strictly, the benchmark approach should use MCS to sample the magnitudes and the ground-motion residuals. This is computationally prohibitive, however, even for the aggregated network and hence the benchmark approach used in the current study uses IS for generating the magnitudes but MCS for the residuals. IS of a single random variable has been shown to be effective in a wide variety of applications including lifeline risk assessment [27], and so further validation is not needed. On the other hand, the simulation procedure for intra-event residuals involves the novel application of IS of a correlated vector of random variables, and hence, is the focus of the validation study described in this section.

Figure 3(a) shows the exceedance curve obtained using IS for generating 25 magnitudes and MCS for generating 500 sets of inter and intra-event residuals per magnitude-location pair, resulting in a total of 125 000 maps. As seen from the figure, the exceedance curve obtained using the IS

DOI: 10.1002/eqe

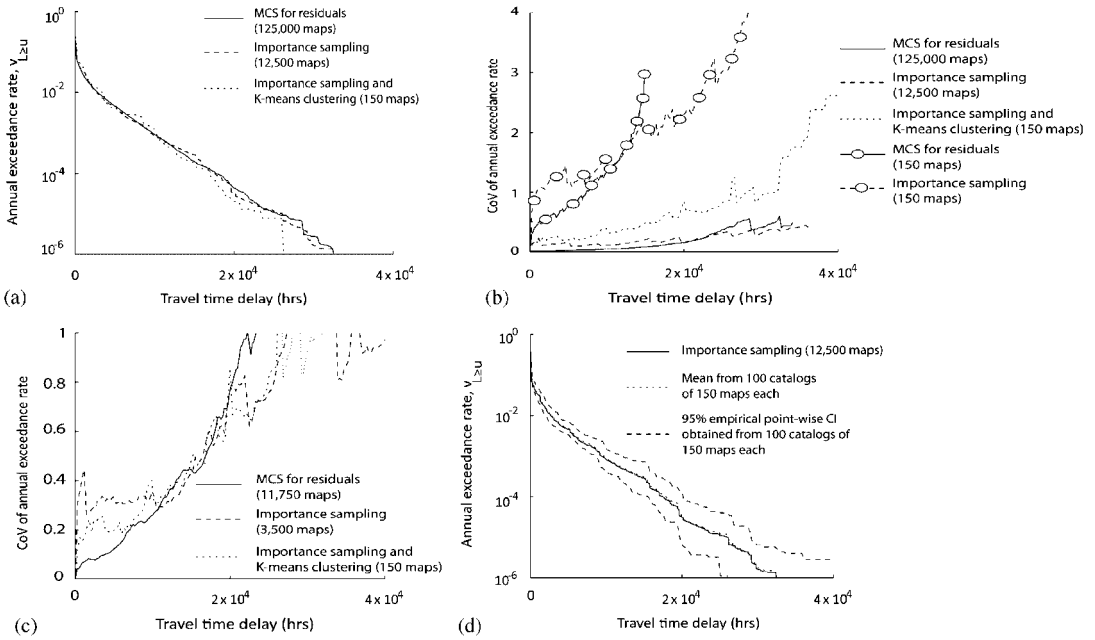

Figure 3. (a) Travel-time delay exceedance curves; (b) coefficient of variation of the annual exceedance rate; and (c) comparison of the efficiency of MCS, IS and the combination of K-means and IS; and (d) travel-time delay exceedance curve obtained using the K-means method.

framework closely matches that obtained using the benchmark method, indicating the accuracy of the results obtained using IS. This is further substantiated by Figure 3(b), which plots the estimated coefficient of variation (CoV) (computed using Equations (19) and (20)) of the exceedance rates obtained using the IS approach and the benchmark approach. It can be seen from the figure that the CoV values corresponding to travel-time delays obtained using IS are comparable to those obtained using MCS even though the IS uses one-tenth the number of simulations required by the MCS. Further, it is also seen that using IS in place of MCS for simulating magnitudes typically reduces the computational expense of the risk assessment by a factor of 10, and hence, the overall IS framework reduces the number of computations required for the risk assessment by a factor of nearly 100.

5.4.2. Risk assessment using IS and K-means clustering. The 12 500 maps obtained using IS are next grouped into 150 clusters using the K-means method. A catalog is then developed by randomly sampling one map from each cluster in accordance with the map weights as described in Section 4. This catalog is used to estimate the travel-time delay exceedance curve based on Equation (23), and the curve is seen to match reasonably well with the exceedance curve obtained using the IS technique (Figure 3(a)). Based on the authors' experience, the deviation of this curve from the IS curve at the large delay levels is a result of the variance of the exceedance rates rather than any systematic deviation. The variance in the exceedance curves is a consequence of the fact that the map sampled from each cluster is not identical to the other maps in the cluster (although they are similar).

DOI: 10.1002/eqe

To ascertain the variance of the exceedance rates, the clustering and the map selection processes are repeated several times in order to obtain multiple catalogs of 150 representative ground-motion intensities, which are then used for obtaining multiple exceedance curves. The CoV of the exceedance rates are then computed from these multiple exceedance curves and are plotted in Figure 3(b). It can be seen that the CoV values obtained using the 150 maps generated by the IS and K-means combination are about three times larger than those obtained using the 12 500 IS maps and the 125 000 MCS maps. This is to be expected, though, on account of the large reduction in the number of maps. The factor of three increase in the CoVs, however, is significantly smaller than what can be expected if IS and MCS are used to obtain the 150 maps directly. This can be seen from Figure 3(b), which shows the large CoV values of the exceedance rates obtained using 150 ground-motion maps selected directly using the IS and the MCS procedures. Alternately, the relative performances of the IS and K-means combination, the IS method and the MCS method can also be assessed by comparing the number of maps to be simulated using these methods in order to achieve the same CoVs. It is seen that 3500 IS maps and 11 750 MCS maps are necessary to produce similar CoVs (Figure 3(c)) achieved using the 150 IS and K-means combination maps.

Finally, Figure 3(d) shows the mean exceedance rates, along with the empirical 95 percentile (point-wise) confidence interval obtained using the K-means method. Also shown in this figure is the exceedance curve obtained using the IS technique. The mean K-means curve and the IS curve match very closely, indicating that the sampling and data reduction procedure suggested in this work results in unbiased exceedance rates (this is also theoretically established in Appendix A). This width of the confidence interval turns out to be reasonably small, especially considering that the exceedance rates have been obtained using only 150 intensity maps.

If the K-means clustering procedure is effective, intensity maps in a cluster will be similar to each other. Therefore, the travel-time delays associated with all the maps in a cluster should be similar to one another, and different from the travel-time delays associated with the maps in other clusters. In other words, the mean travel-time delays computed using all the maps in one cluster should be different from the mean from other clusters, while the standard deviation of the travel-time delays in a cluster should be small as a result of the similarity within a cluster. Conversely, 'random clustering' in which the maps obtained from the IS are randomly placed in clusters irrespective of their properties would be very inefficient. Figure 4 compares the mean and the standard deviation of cluster travel-time delays, obtained using K-means clustering and random clustering. The smoothly varying cluster means obtained using K-means as compared with the nearly uniform means obtained using random clustering shows that the K-means has been successful in separating dissimilar intensity maps. Similarly, the cluster standard deviations obtained using K-means are considerably smaller than the standard deviations obtained using random clustering for the most part (and are large for larger cluster numbers because all delays in these clusters are large). The occasional spikes in the standard deviations are a result of small sample sizes in some clusters.

In summary, the exceedance curves obtained and the results from the tests for the efficiency of K-means clustering indicate that the clustering method has been successful in identifying and grouping similar maps together. As a consequence, substantial computational savings can be achieved by eliminating redundant (similar) maps, without considerably affecting the accuracy of the exceedance rates.

5.4.3. Hazard consistency. The proposed framework not only produces reasonably accurate loss estimates, but also intensity maps that are hazard consistent. In other words, the site hazard

DOI: 10.1002/eqe

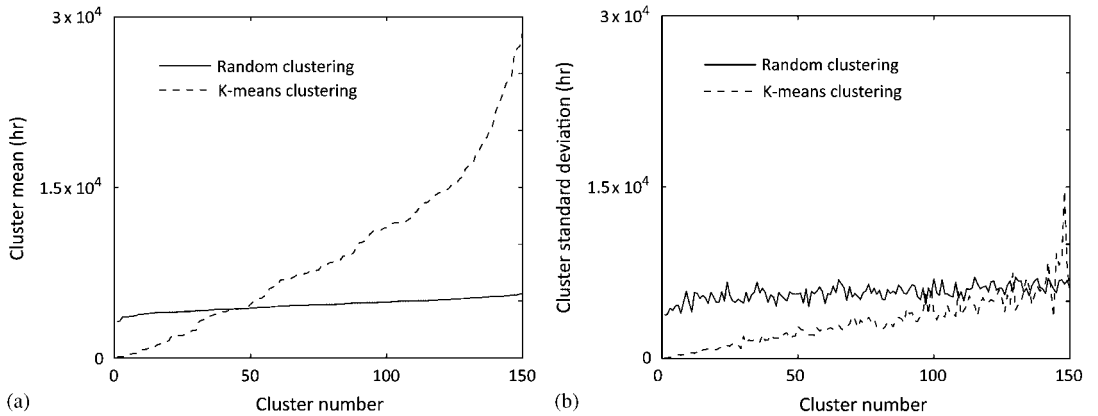

Figure 4. (a) Mean of travel-time delays within a cluster and (b) standard deviation of travel-time delays within a cluster. With both clustering methods, cluster numbers are assigned in order of increasing mean travel-time delay within the cluster for plotting purposes.

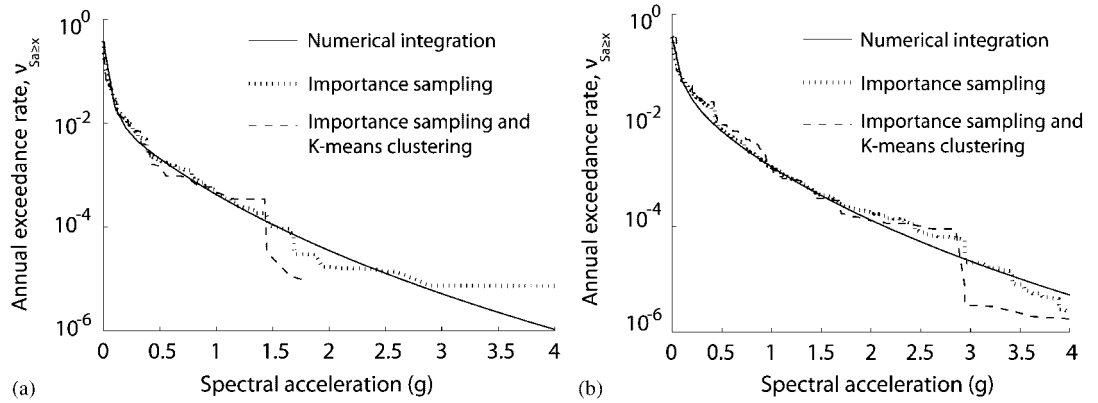

Figure 5. Comparison of site hazard curves obtained at two sample sites using the sampling framework with that obtained using numerical integration. (a) Sample site 1 and (b) sample site 2.

curves obtained based on the final catalog of intensity maps match the site ground-motion hazard curves obtained from the fault and the ground-motion model using numerical integration (i.e. traditional PSHA). Figures 5(a) and (b) show the site hazard curves at two different sites obtained using numerical integration, IS (for magnitudes and residuals) and the combination of importance sampling and K-means clustering. It can be seen that the sampling and clustering framework reasonably reproduces the site ground-motion hazard obtained through numerical integration.

5.5. Importance of modeling ground-motion uncertainties and spatial correlations

The transportation network risk assessment is repeated assuming uncorrelated intra-event residuals, and a new exceedance curve is obtained, and plotted in Figure 6. It can be seen that the risk is considerably underestimated when the spatial correlations are ignored. Further, some past risk

DOI: 10.1002/eqe

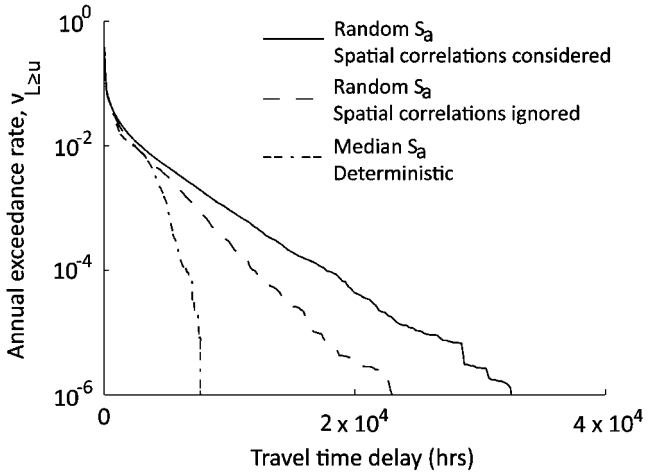

Figure 6. Exceedance curves obtained using simplifying assumptions.

assessments have completely ignored the uncertainty in the ground-motion intensities (i.e. median intensity maps are used, and inter- and intra-event residuals are ignored). A risk assessment carried out this way, and plotted in Figure 6 shows that the risk is even more substantially underestimated in this case. This happens because the possibility of observing above-median ground-motion intensities during a given earthquake is not considered. Such simplifications clearly introduce significant errors into the risk calculations, and should thus be avoided.

6. CONCLUSIONS

An efficient simulation-based framework based on IS and K-means clustering has been proposed, that can be used for the seismic risk assessment of lifelines. The framework can be used for developing a small, but stochastically representative catalog of ground-motion intensity maps that can be used for performing lifeline risk assessments. The IS technique is used to preferentially sample important ground-motion intensity maps, and the K-means clustering technique is used to identify and combine redundant maps. It is shown theoretically and empirically that the risk estimates obtained using these techniques are unbiased. The study proposes IS schemes that can be used for sampling earthquake magnitudes, rupture locations, inter-event residuals and spatially correlated maps of intra-event residuals. Magnitudes are sampled by first stratifying the magnitude range of interest in to smaller partitions and by selecting one magnitude from each partition. The partitions are made narrower at larger magnitudes to ensure that larger magnitudes are preferentially sampled. The normalized residuals are sampled from a normal distribution with a positive mean, rather than a zero mean, to sample more large positive residuals. Techniques are also suggested to estimate the optimal parameters of these alternate sampling density functions. The proposed framework was used to evaluate the exceedance rates of various travel-time delays on an aggregated form of the San Francisco Bay Area transportation network. Simplified transportation network analysis models were used to illustrate the feasibility of the proposed framework. The exceedance rates were obtained using a catalog of 150 maps generated using the combination

DOI: 10.1002/eqe

of importance sampling and K-means clustering, and were shown to be in good agreement with those obtained using the conventional MCS. Therefore, the proposed techniques can reduce the computational expense of a simulation-based risk assessment by several orders of magnitude, making it practically feasible. The efficiency of the proposed technique was compared with that of conventional techniques using the CoV of the exceedance rates. It was shown that the CoVs achieved using the 150 maps obtained from the combination of IS and K-means clustering can only be reproduced by 3500 IS maps and 11 750 MCS maps (conventional MCS for residuals and IS for magnitudes), thereby indicating the efficiency of the proposed technique. The study also showed that the proposed framework automatically produces intensity maps that are hazard consistent. Finally, the study showed that the uncertainties in ground-motion intensities and the spatial correlations between ground-motion intensities at multiple sites must be modeled in order to avoid introducing significant errors into the lifeline risk calculations. For the network considered in this work, ignoring spatial correlations results in about a 30% reduction in the estimated travel-time delays at small annual exceedance rates (10^{-6}/year), while ignoring uncertainties results in about a 70% reduction in the estimated travel-time delays at small exceedance rates.

APPENDIX A: PROOF THAT THE EXCEEDANCE RATES OBTAINED USING IS AND K-MEANS CLUSTERING ARE UNBIASED

This section illustrates that the loss (e.g. travel-time delay) exceedance rates obtained using a catalog of ground-motion intensities generated by the IS and K-means framework are unbiased. Since the IS procedure produces unbiased estimates [15], it will suffice to establish that the exceedance rates obtained using the K-means clustered catalog of maps are unbiased estimators of the exceedance rates obtained using the IS maps. This proof will further support the empirical observation that the example exceedance rates from the different procedures are equivalent.

Let $l_1, l_2, \ldots, l_r$ denote the loss measures (e.g. travel-time delay in a transportation network) corresponding to the r intensity maps obtained using IS. Let $\Lambda_1, \Lambda_2, \ldots, \Lambda_r$ denote the weights corresponding to the maps as defined in Equation (18). Let P_{IS} denote the exceedance probability curve obtained using the IS maps (Equation (19)). Assume that the r maps are grouped into K clusters. (This proof does not require knowledge about the clustering technique used.) Let $l_{(c)}$ be the travel-time delay in the network corresponding to the map selected from cluster c. The exceedance probability curve ($\hat{P}_{KM}(L \geqslant u)$) can be obtained from the catalog of $[l_{(1)}, l_{(2)}, \ldots, l_{(K)}]$ based on Equation (23).

Unbiasedness can be established by showing that the expected value of $\hat{P}_{KM}(L \geqslant u)$ equals $\hat{P}_{\mathrm{IS}}(L \geqslant u)$. The expected value of $\hat{P}_{KM}(L \geqslant u)$ is computed using the law of iterated expectations, by first conditioning it on a possible grouping G (i.e. a possible grouping of maps into clusters obtained using the clustering method), and then by computing the expectation over all possible groupings. The following equations describe this procedure:

$$E[\hat{P}_{KM}(L \geqslant u)] = E\left[\frac{\sum_{c=1}^{K} I(l_{(c)} \geqslant u) \sum_{i \in c} \Lambda_i}{\sum_{c=1}^{K} \sum_{i \in c} \Lambda_i}\right]$$

$$= E\left[\frac{\sum_{c=1}^{K} I(l_{(c)} \geqslant u) \sum_{i \in c} \Lambda_i}{\sum_{i=1}^{r} \Lambda_i}\right]$$

DOI: 10.1002/eqe

$$
\begin{aligned}
&= E_G\left\{E\left[\left.\frac{\sum_{c=1}^{K} I(l_{(c)}\geqslant u)\sum_{i\in c}\Lambda_i}{\sum_{i=1}^{r}\Lambda_i}\right| G\right]\right\}\\
&= E_G\left[\frac{1}{\sum_{i=1}^{r}\Lambda_i}\sum_{c=1}^{K} P(l_{(c)}\geqslant u \mid G)\sum_{i\in c}\Lambda_i\right]\\
&= E_G\left[\frac{1}{\sum_{i=1}^{r}\Lambda_i}\sum_{c=1}^{K}\frac{\sum_{j\in c} I(l_j\geqslant u)\Lambda_j}{\sum_{j\in c}\Lambda_j}\sum_{i\in c}\Lambda_i\right]\\
&= E_G\left[\frac{1}{\sum_{i=1}^{r}\Lambda_i}\sum_{c=1}^{K}\sum_{j\in c} I(l_j\geqslant u)\Lambda_j\right]\\
&= \frac{1}{\sum_{i=1}^{r}\Lambda_i}\sum_{c=1}^{K}\sum_{j\in c} I(l_j\geqslant u)\Lambda_j\\
&= \frac{\sum_{i=1}^{r} I(l_i\geqslant u)\Lambda_i}{\sum_{i=1}^{r}\Lambda_i}\\
&= \hat{P}_{\mathrm{IS}}(L\geqslant u)
\end{aligned}
\tag{A1}
$$

This shows that the exceedance rates obtained using the small catalog of ground-motion intensities are unbiased.

APPENDIX B: IMPROVING THE COMPUTATIONAL EFFICIENCY OF THE K-MEANS CLUSTERING METHOD

Clustering a large number of intensity maps (e.g. 12 500) in a single step may be computationally prohibitive on computers with limited memory and processing ability, because clustering involves repetitive computations of the distance between each map and the cluster centroids. In such cases, the authors propose the following two-step clustering technique in which the maps are preliminarily grouped into clusters using a simplified distance measure, followed by a rigorous final clustering step using the distance measure defined in Equation (22). This two-step process is described below.

In the preliminary clustering step, the intensity maps are grouped into a small number of preliminary clusters with the distance between map $\mathbf{S}_{\mathbf{a}_j}$ and centroid $\mathbf{C}_i$ computed as $(\sum_{q=1}^{p} S_{a_{qj}} - \sum_{q=1}^{p} C_{qi})^2$. In other words, the distance measure is based on the sum of the intensities corresponding to the intensity map. The sum of the intensities is chosen as the basis for clustering since it has been seen in past research [6] and in the current research work to be a reasonable indicator of the risk associated with an intensity map. Further, the K-means method is extremely fast when the distance is based on a single parameter.

The final clustering step is used to refine the preliminary clusters, and involves further clustering within each preliminary cluster using the distance measure defined in Equation (22). If 50 preliminary clusters are used, each of these could be subdivided into 3 clusters using the K-means

DOI: 10.1002/eqe

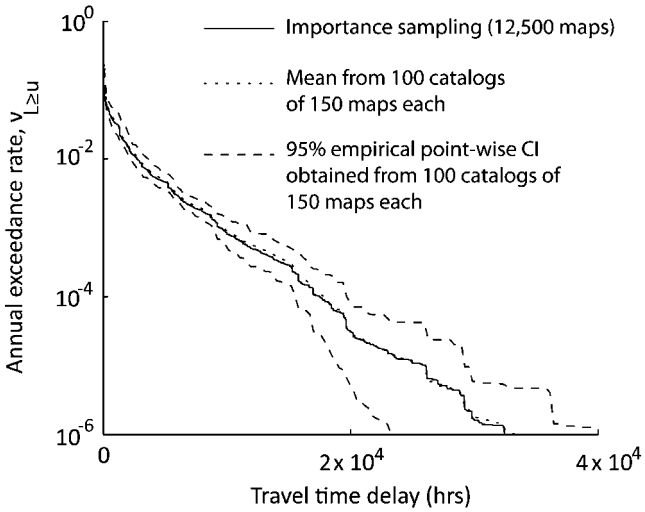

Figure B1. Travel-time delay exceedance curve obtained using the two-step clustering technique.

method. Although the more rigorous distance measure is used in this step, it is much faster because the final clustering is based on a far fewer number of maps stored within each preliminary cluster. Further, the memory demand in this case is much smaller than when clustering is carried out in a single step.

Figure B1 shows the (point-wise) confidence intervals of the travel-time delay exceedance curves obtained using the two-step clustering procedure, where 50 preliminary clusters are each subdivided into three final clusters. It can be seen from Figures 3(d) and (B1) that the results obtained using both the single-step and the two-step clustering approaches are essentially identical. For this application, the two-step clustering procedure is five times faster than the single-step clustering procedure.

ACKNOWLEDGEMENTS

The authors thank the two anonymous reviewers for their helpful reviews of the manuscript, and Prof. Anne Kiremidjian and Prof. Jerome Friedman from Stanford University for useful discussions on this research topic. The authors also thank Prof. Anne Kiremidjian and Evangelos Stergiou for providing data pertaining to the San Francisco Bay Area transportation network. The first author acknowledges the support of the Stanford Graduate Fellowship.

REFERENCES

1. Chang S. Evaluating disaster mitigations: methodology for urban infrastructure systems. *Natural Hazards Reviews* 2003; **4**:186.
2. Tanaka S, Shinozuka M, Schiff A, Kawata Y. Lifeline seismic performance of electric power systems during the Northridge earthquake. *Proceedings of the Northridge Earthquake Research Conference*, Los Angeles, CA, 1997.
3. McGuire RK. *Seismic Hazard and Risk Analysis*, Earthquake Engineering Research Institute, 2007.
4. Kang WH, Song J, Gardoni P. Matrix-based system reliability method and applications to bridge networks. *Reliability Engineering and System Safety* 2008; **93**(11):1584–1593.

DOI: 10.1002/eqe

5. Dueñas-Osorio L, Craig JI, Goodno BJ, Bostrom A. Interdependent response of networked systems. *Journal of Infrastructure Systems* 2005; **13**(3):185–194.
6. Campbell KW, Seligson HA. Quantitative method for developing hazard-consistent earthquake scenarios. *Proceedings of the 6th U.S. Conference and Workshop on Lifeline Earthquake Engineering*, Long Beach, CA, 2003.
7. Crowley H, Bommer JJ. Modelling seismic hazard in earthquake loss models with spatially distributed exposure. *Bulletin of Earthquake Engineering* 2006; **4**(3):249–273.
8. Kiremidjian AS, Stergiou E, Lee R. Issues in seismic risk assessment of transportation networks. *Earthquake Geotechnical Engineering*, Chapter 19. Springer: Berlin, 2007; 939–964.
9. Shiraki N, Shinozuka M, Moore II 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; **213**(1):43–54.
10. Adachi T, Ellingwood BR. Serviceability of earthquake-damaged water systems: effects of electrical power availability and power backup systems on system vulnerability. *Reliability Engineering and System Safety* 2008; **93**:78–88.
11. Jayaram N, Baker JW. Deaggregation of lifeline risk: insights for choosing deterministic scenario earthquakes. *Proceedings*, *TCLEE2009 Conference*: *Lifeline Earthquake Engineering in a Multihazard Environment*, Oakland, CA, 2009.
12. 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.
13. Jayaram N, Baker JW. Correlation model for spatially-distributed ground-motion intensities. *Earthquake Engineering and Structural Dynamics* 2009; **38**(15):1687–1708.
14. Wang M, Takada T. Macrospatial correlation model of seismic ground motions. *Earthquake Spectra* 2005; **21**(4):1137–1156.
15. Fishman GS. *A First Course in Monte Carlo*. Duxbury: Belmont, CA, 2006.
16. Abrahamson NA, Youngs RR. A stable algorithm for regression analyses using the random effects model. *Bulletin of the Seismological Society of America* 1992; **82**(1):505–510.
17. Abrahamson NA, Silva WJ. Summary of the Abrahamson & Silva NGA ground-motion relations. *Earthquake Spectra* 2008; **24**(1):99–138.
18. McQueen JB. Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, CA 1967.
19. Gersho A, Gray RM. *Vector Quantization and Signal Compression*. Springer: Berlin, 1991.
20. Stergiou E, Kiremidjian AS. Treatment of uncertainties in seismic risk analysis of transportation systems. *Technical Report No. 154*, Blume Earthquake Engineering Center, Stanford University, 2006.
21. HAZUS. Earthquake loss estimation technical manual. *Technical Report*, National Institute of Building Sciences, Washington, DC, 1999.
22. Purvis C. Peak spreading models: promises and limitations. *Seventh TRB Conference on the Application of Transportation Planning Models*, Boston, MA, 1999.
23. Bureau of Public Roads. *Traffic Assignment Manual*. U.S. Department of Commerce, Urban Planning Division, Washington, DC, 1964.
24. Beckman MJ, McGuire CB, Winsten CB. Studies in the economics of transportation. *Technical Report*, Cowles Comission Monograph, Yale University Press: New Haven, CT, 1956.
25. Frank M, Wolfe P. An algorithm for quadratic programming. *Naval Research Logistics Quarterly* 1956; **3**:95–110.
26. Friesz TL, Bernstein D, Smith TE, Tobin RL, Wie BW. A variational inequality formulation of the dynamic network user equilibrium problem. *Operations Research* 1993; **41**:179–191.
27. Kiremidjian AS, Moore J, Fan YY, Basiz N, Yazali O, Williams M. PEER highway demonstration project. *Sixth US Conference and Workshop on Lifeline Earthquake Engineering*, *TCLEE/ASCE*, *Monograph No.25*, Long Beach, CA, 2003.
28. USGS. Earthquake probabilities in the San Francisco bay region: 2002-2031. *Technical Report*, *Open File Report 03-214*, USGS, 2003.
29. Youngs RR, Coppersmith KJ. Implications of fault slip rates and earthquake recurrence models to probabilistic seismic hazard estimates. *Bulletin of the Seismological Society of America* 1985; **75**(4):939–964.
30. Boore DM, Atkinson GM. Ground-motion prediction equations for the average horizontal component of PGA, PGV and 5% damped SA at spectral periods between 0.01 s and 10.0 s. *Earthquake Spectra* 2008; **24**(1):99–138.

DOI: 10.1002/eqe