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

DOI: xxx/xxxx

ARTICLE TYPE

Scalable Scenario-based Earthquake Risk Modeling via Linearized Ground-Motion–Fragility Coupling and PPCA

Soung Eil Houng | **Luis Ceferino**

[1]Civil and Environmental Engineering, University of California, Berkeley, CA

Correspondence
Corresponding author Soung Eil Houng
Email: shoung@berkeley.edu

Present address
Davis Hall, Berkeley, CA 94706

Abstract

In scenario-based regional risk modeling, the traditional workflow simulates spatially correlated ground motions, and subsequently samples building damage states from lognormal fragility functions. In this procedure, the dimensionality grows with the number of assets and quickly becomes computationally prohibitive for large cities. To overcome this limitation, we introduce a scalable computational framework that (i) reformulates the ground-motion–fragility coupling through an exact linearization and (ii) employs probabilistic principal component analysis (PPCA) to identify low-dimensional latent variables for efficient simulation. We validate the proposed approach on San Francisco's downtown portfolio of 1,000 buildings, benchmarking against SimCenter R2D's computational testbed. The modal damage states of > 95% buildings match exactly, with a mean difference below 0.04 (on a 0–4 ordinal scale representing none to complete damage), confirming the framework's accuracy. The achievable dimensionality reduction depends primarily on the portfolio's spatial extent rather than building density, implying that—given a fixed spatial region—the computational burden remains nearly constant with portfolio size, unlike current approaches. In tests on downtown San Francisco (15,836 buildings) and the broader Bay Area, a single latent dimension and 20 dimensions, respectively, reproduce the benchmark loss distributions within $<\sim$ 2.5%. The method reduces pre-processing complexity from $O(N^3)$ to $O(N^2)$ and simulation complexity from $O(N^2)$ to $O(N)$, where N denotes the number of buildings, yielding roughly 3$\times$ faster pre-processing and 110$\times$ faster simulation for a 30,000-building subset, with speedups growing linearly with portfolio size. The framework substantially lowers the computational barrier for high-resolution regional seismic risk assessment.

KEYWORDS
regional seismic risk; scenario-based risk modeling, ground-motion–fragility coupling, dimensionality reduction, probabilistic principal component analysis

1 | INTRODUCTION

Scenario-based seismic risk analysis (SRA) estimates losses for a building portfolio subjected to a specified earthquake scenario (Figure 1). By focusing on a single representative event rather than aggregating outcomes across all possible earthquakes, this approach provides an actionable view of risk, supporting risk mitigation decisions on where and how much to invest, which assets to prioritize for retrofitting, and how to plan evacuations [1,2,3,4,5]. The type of loss considered depends on the intended application and may include financial losses, casualties and injuries, or infrastructure downtime [6,7,8,9,10,11]. However, most of these analyses rely on computationally intensive frameworks originally developed for individual structures, limiting their scalability to large regional portfolios.

The performance-based earthquake engineering (PBEE) framework [12] represents one of the most widely adopted methodologies for individual structures. This framework sequentially quantifies loss by first generating ground motion intensities, simulating structural responses, estimating damage states based on those responses, and finally calculating the resulting losses. Regional-scale risk analysis fundamentally follows this same logic, albeit by utilizing spatially distributed quantities such as ground motion intensity maps, damage maps, and loss maps (Figure 1). At this scale, explicit response analysis is frequently bypassed

due to its high computational cost; instead, fragility curves are employed to directly evaluate building damage from ground shaking intensity. The total regional loss is then determined by aggregating the losses of all constituent buildings.

While this workflow is conceptually straightforward, its implementation at scale faces computational hurdles. Specifically, the generation of the ground motion intensity maps is recognized as the most computationally demanding phase of regional SRA (Figure 1 (b) and (c)). Regional shaking is typically modeled using empirical ground-motion models integrated within a Gaussian process framework to account for spatial ground motions' correlation (Figure 1 (b))[13,14,15,16]. As a result, this step presents a major bottleneck as it necessitates simulating multivariate normally distributed random variables (one per building) through an expensive Cholesky decomposition of a large, dense correlation matrix and repeated sampling of within-event residuals[17,18]. Consequently, the computational burden scales poorly ($O(N^3)$) with the size of the building inventory[19].

Once the ground motion intensity map is established, building damage is modeled using fragility curves , which at a given ground shaking intensity defines the probability that a building will exceed a specific damage state; thus, this process is inherently stochastic (Figure 1 (d) and (e)). In this stage, because damage must be simulated across all portfolio locations, this phase remains high-dimensional and computationally intensive. However, since damage is typically modeled as independent between buildings, it is generally less burdensome than ground shaking modeling.

Following the damage assessment, loss metrics—such as repair costs or recovery times—are modeled based on the predicted damage states (Figure 1 (f) and (g)). The complexity of the loss model significantly influences the overall computational demand of the risk analysis[20]. The simplest approach utilizes look-up tables or deterministic ratios for each damage state[21]. Nevertheless, again, because this process must be executed for every building in the inventory, the total computational time scales linearly with the number of assets. As regional portfolios grow in size, even these relatively simple operations can add computational burden.

To mitigate the computational burden of regional risk analysis, two primary strategies have been developed, both centered on the simulation of ground-motion intensity maps. The first approach seeks to reduce the per-simulation cost, while the second focuses on minimizing the total number of simulations required. While both strategies alleviate computational demands, the latter offers the additional benefit of reducing the time required for loss calculations— a critical advantage when employing complex loss models[22].

In the first approach, researchers often employ coarse "ground-shaking grids" to alleviate computational demands, assigning each asset the intensity from the nearest grid point or via interpolation[17]. Although these techniques substantially reduce the overhead of generating correlated ground motions, they can introduce bias by assigning identical intensities to distinct sites that, in reality, experience different shaking[23,24,25]. Recent efforts have also utilized principal component analysis (PCA) to accelerate the simulation of spatially correlated intensities[26,27]; however, these methods primarily benefit the modeling of multiple intensity measures across various periods and do not fundamentally resolve the challenges of spatial scalability.

The second approach aims to decrease the number of ground-shaking simulations required for accurate loss estimation. One such method involves Importance Sampling (IS)[28,29], which samples from "most contributing" shaking fields with appropriate weights to yield unbiased loss estimates. More recently, adaptive importance sampling (AIS) has been applied to accelerate probabilistic seismic hazard analysis by factors of up to $\sim 10^3$ [30,31]. Although AIS provides a principled framework for optimizing computational effort, applying existing AIS algorithms to high-dimensional problems—such as regional risk analysis—remains a significant challenge[32,33].

Another category of methods in the second approach seeks to minimize errors associated with the reduced sets of selected ground-motion scenarios that match target metrics, such as hazard curve[34,35,36,37]. This approach formulates an optimization problem that yields a unique set of deterministic ground-motion maps, which may conflict with the inherently stochastic nature of seismic risk. Furthermore, unbiasedness of the target performance is not guaranteed without proper correction factors. For instance, a set of ground-motion simulations that best fits a hazard curve does not necessarily ensure an unbiased risk estimate, as the sample distribution may deviate from the underlying stochastic model. Additionally, these approaches do not reduce the dimensionality of the underlying problem; consequently, the per-simulation computational cost remains high.

In this study, we introduce a scalable computational framework for scenario-based regional risk modeling that overcomes the limitations of prior methods. First, we derive an exact linearized formulation that merges ground-motion and damage simulations into a single-variable operation within a Gaussian space, demonstrating how this formulation simplifies the interpretation of ground motion–damage interaction. Second, we integrate Probabilistic Principal Component Analysis (PPCA)[38] into the risk modeling framework to identify low-dimensional latent variables, thereby enabling high-performance computations for large-scale building portfolios. The proposed framework is mathematically formulated, interpreted using simple examples, and its performance is evaluated through a case study of the San Francisco Bay Area—extending from the urban downtown to the entire regional scale—under a major earthquake scenario on the San Andreas Fault.

2 | TRADITIONAL REGIONAL SCENARIO-BASED SEISMIC RISK ANALYSIS

Following the principles of PBEE, the resulting loss distribution is generally mathematically intractable in closed form. Thus, Monte Carlo (MC) simulation is utilized to achieve numerical convergence. The required number of MC realizations is typically dictated by the specific exceedance probabilities or tail risks of interest. The following sections detail the traditional implementation of each phase, including the simulation of ground-shaking (Figure 1(b) and (c)), damage and subsequent loss (Figure 1(d)–(g)).

2.1 | Ground Shaking

Given an earthquake scenario, the ground motion $y \in \mathbb{R}_+$ at a single site is lognormally distributed. Thus, we can simulate it using the logarithmic mean of the ground motion (μ) and samples of between-event and within-event residuals, expressed as a linear combination of two standard normal random variables. For a single site:

$$\ln y = \mu + \tau\,\zeta + \phi\,\xi, \tag{1}$$

where τ and ϕ are between- and within-event standard deviations, respectively; and ζ and ξ are independent standard normal random variables, $\mathcal{N}(0, 1)$. The terms $\tau\zeta$ and $\phi\xi$ provide stochastic realizations of the between-event and within-event residuals, respectively. The logarithmic mean μ is modeled as a function of magnitude, source-to-site distance, and site conditions, with additional adjustments to account for effects such as rupture directivity and stress-drop variability[39,40]. The parameter τ, between-event standard deviation, quantifies the variability in ground-shaking intensity arising from random rupture scenarios—the range of rupture characteristics that are not captured in the calculation of μ. Similarly, the parameter ϕ, within-event standard deviation,

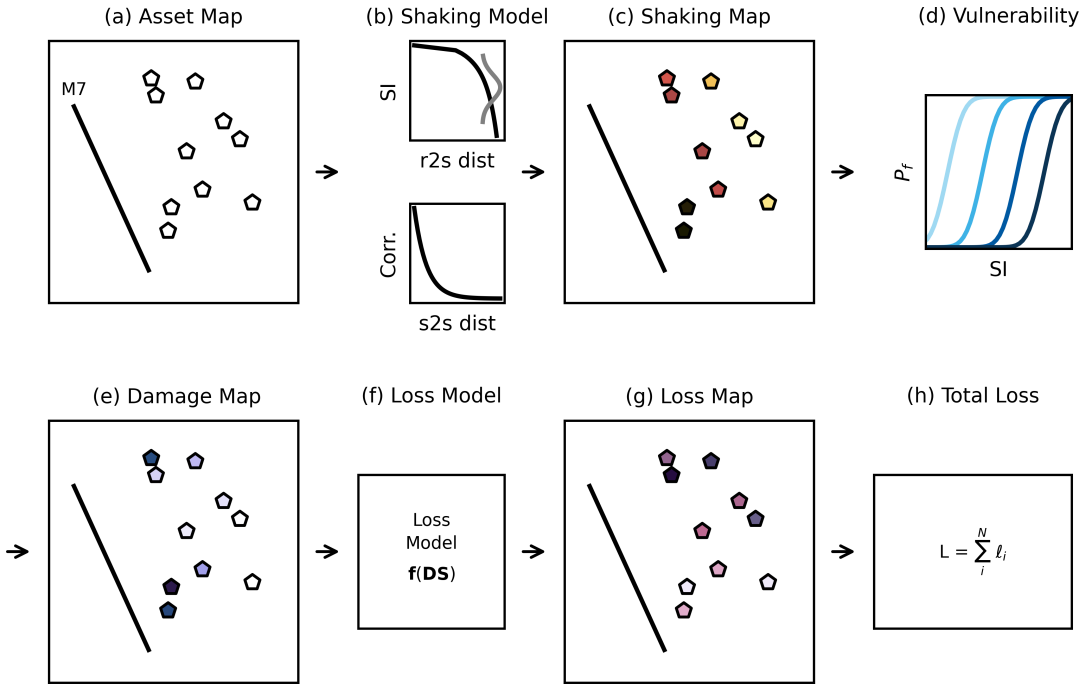

FIGURE 1 Schematic diagram of the traditional scenario-based risk modeling framework. (a) Asset distribution map featuring individual assets (pentagons) and a scenario earthquake rupture of M 7.0 (solid black line). (b) Models for generating shaking maps: (top) Ground Motion Model (GMM) and (bottom) correlation model. r2s dist: rupture-to-site distance; SI: shaking intensity; s2s distance: inter-site distance; and Corr.: within-event spatial correlation. (c) A realization of a shaking intensity map, where increasing color intensity indicates higher ground shaking. (d) Fragility curves representing vulnerability for multiple damage limit states (P_f). (e) A realization of a damage map. (f) Implementation of a user-defined loss model. (g) Resultant loss map illustrating the distribution of consequences. (h) Total loss calculated as the sum of all individual losses.

represents the variability in ground-shaking intensity due to random wave propagation and site effects, capturing variations in propagation paths and site amplification that are likewise not incorporated into μ.

For a region, we model ground shaking $\mathbf{y}$ at multiple sites simultaneously, explicitly incorporating the spatial correlation of ground motions. This is necessary because the within-event residuals ($\phi\,\xi$ in Equation 1) at nearby sites are not independent, as these sites share most of the seismic wave propagation path as well as averaged local site responses. The within-event correlation matrix $\mathbf{C}$ is typically modeled as a function of inter-site distance, with higher correlation assigned to adjacent sites and lower correlation to more distant sites[15,14,13]:

$$\mathbf{C} = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{N1} & \rho_{N2} & \cdots & \rho_{NN} \end{bmatrix}$$

where $\rho_{ii} = 1$ because each site has a unique ground-shaking intensity, and $\rho_{ij} = \rho_{ji}$ since the correlation between sites i and j depends solely on their inter-site distance, which is symmetric.

To model the within-event residuals with correlation matrix $\mathbf{C}$, we seek any matrix $\mathbf{L}$ such that

$$\mathbf{L}\mathbf{L}^{\mathsf{T}} = \mathbf{C}.$$

The matrix $\mathbf{L}$ is typically obtained via the Cholesky decomposition of $\mathbf{C}$, for its simplicity and numerical stability[41]. Accordingly, $\mathbf{L}$ is the lower triangular matrix.

Therefore, for N buildings, the ground motions $\mathbf{y} \in \mathbb{R}_+^N$ across the sites are given by

$$\ln \mathbf{y} = \boldsymbol{\mu} + \boldsymbol{\tau}\,\zeta + \boldsymbol{\phi} \circ (\mathbf{L}\,\boldsymbol{\xi}) \tag{2}$$

, where $\boldsymbol{\mu} \in \mathbb{R}^N$ is the vector of logarithm of median shaking intensities; $\boldsymbol{\tau} \in \mathbb{R}^N$ and $\boldsymbol{\phi} \in \mathbb{R}^N$ are the vectors of between- and within-event standard deviations; $\mathbf{L} \in \mathbb{R}^{N\times N}$ is the lower-triangular Cholesky factor of the within-event spatial correlation matrix $\mathbf{C} \in \mathbb{R}^{N\times N}$; $\zeta \in \mathbb{R} \sim \mathcal{N}(0,1)$ is the scalar standard normal random variable; and $\boldsymbol{\xi} \in \mathbb{R}^N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the standard normal random vector. The operator $\circ$ denotes element-wise multiplication. Note that ζ, which generates the between-event residual, is a scalar for a given earthquake scenario and affects all sites the same. By contrast, the within-event residuals are represented by a correlated N-dimensional random vector and therefore require $\boldsymbol{\xi} \in \mathbb{R}^N$.

The total covariance matrix of the ground shaking intensity given earthquake scenario ($\ln \mathbf{y}$ in Equation 2) is:

$$\mathbf{Cov}(\ln \mathbf{y}) = \boldsymbol{\tau}\boldsymbol{\tau}^{\mathsf{T}} + \mathbf{FCF} \tag{3}$$

where $\mathbf{F} = \mathrm{diag}(\boldsymbol{\phi})$. Also, the total correlation of the ground shaking intensity is the normalized version of Equation 3, which is $\mathbf{C}_{\ln \mathbf{y}} = \mathbf{Cov}(\ln \mathbf{y})/\sum_{i=1}^{N} \mathbf{Cov}(\ln \mathbf{y})_{ii}$.

In most recent ground-motion models, $\boldsymbol{\phi}$ and $\boldsymbol{\tau}$ are site dependent and vary with local soil conditions, particularly when the shaking intensity amplification term is nonlinear with respect to ground-shaking intensity[42,43,44,40]. If nonlinearity is neglected, some models can be modeled as magnitude-dependent only. Thus, for a fixed scenario, we could treat these standard deviations across sites as constant[40]. However, for generality, we retain the vector notation. For peak ground acceleration (PGA), $\boldsymbol{\tau}$ typical ranges from 0.30 to 0.50, and $\boldsymbol{\phi}$ from 0.50 to 0.75, with both tending to decrease for larger magnitudes; exact ranges depend on the model[39].

2.2 | Damage States

We define damage as an ordinal variable with n_{ds} states, indexed by $k = 0, 1, \ldots, n_{\mathrm{ds}} - 1$, where larger k indicates more severe damage. A damage state of $k = 0$ denotes no damage, and $k = n_{\mathrm{ds}} - 1$ denotes the most severe state. In typical applications, we take $n_{\mathrm{ds}} = 5$, corresponding to the following categories: No damage ($k = 0$), Slight ($k = 1$), Moderate ($k = 2$), Extensive ($k = 3$), and Complete ($k = 4 = n_{\mathrm{ds}} - 1$)[21].

The damage probabilities are typically modeled via lognormal fragility functions, which describe the likelihood of exceeding a given damage threshold as a function of ground-motion intensity. For damage state k, the exceedance probability is expressed as

$$P(DS \geq k \mid y) = \Phi\left(\frac{\ln y - \ln \theta_k}{\beta}\right) \tag{4}$$

, where $P(DS \geq k \mid y)$ is the probability that the building experiences a damage state k or greater at shaking intensity y (e.g., PGA of 0.3g), θ_k is the logarithm of the shaking intensity with 50% probability that $DS \geq k$, β is the logarithmic standard deviation, which is often assumed constant across damage states to avoid crossing fragility curves[21], and $\Phi(\cdot)$ is the CDF of $\mathcal{N}(0, 1)$.

Damage states can be correlated when structure–soil–structure interaction (SSSI) is significant for adjacent large structures or one building falls on another; however, these effects are often negligible[45,46], thus, the damage is modeled independently conditioned on the ground shaking intensity in general.

In practice, to simulate damage, we draw $u \sim \mathcal{U}(0, 1)$. If $u > P(DS \geq k \mid y)$ or equivalently, $u - P(DS \geq k \mid y) > 0$, then $DS < k$; otherwise, $DS \geq k$. In order to derive our approach, we define the limit-state function G_k:

$$G_k = \begin{cases} -\infty, & \text{if } k = 0, \\ u - \Phi\left(\frac{\ln y - \ln \theta_k}{\beta}\right), & \text{if } k = 1, \ldots, n_{\text{ds}} - 1. \end{cases} \tag{5}$$

If $u > P(DS \geq k \mid y)$, then $G_k > 0$, implying that the damage state is lower than k ($DS < k$). Conversely, if $u < P(DS \geq k \mid y)$, then $G_k < 0$, implying that the damage state is at least k ($DS \geq k$). The simulated damage state is determined by selecting the maximum k such that $G_k < 0$:

$$DS = \max\{k \in \{0, \ldots, n_{\text{ds}} - 1\} \mid G_k < 0\} \tag{6}$$

That is, DS is the largest k for which $G_k < 0$. Note that such a k always exist because we set $G_0 = -\infty$ (Equation 5), ensuring $G_0 < 0$. Thus, the minimum possible value is $DS = 0$ (No damage). Importantly, as long as the sign of G_k is preserved, the simulated damage state does not change. Given DS, the loss associated with DS, l, is computed by multiplying the building's value V (e.g., replacement cost, number of occupants) by the loss ratio function, $r(\cdot)$:

$$l = V \times r(DS). \tag{7}$$

Also, the total loss is then obtained by summing over all buildings, $l_{\text{t}} = \sum_{i=1}^{N} l_i$.

3 | EXACT LINEARIZATION OF GROUND-MOTION–FRAGILITY COUPLING

As discussed earlier, the risk is typically computed in two separate stages: ground shaking and damage simulations. Integrating these stages is challenging because the damage simulation involves a non-linear relationship between the Gaussian shaking intensity, $\ln y$ (Equation 1), and the uniform damage-state random variable, u (Equation 5). This non-linearity precludes the use of simplified computational implementations available for linear systems.

In this section, we reformulate this traditional two-step procedure into a single linear matrix operation. Notably, this is an exact linearization performed without approximation. We first derive the formulation for a single site and then extend it to the multi-site case.

3.1 | Single Site Linearization

We begin by transforming u via the inverse transform sampling method: $u = \Phi(\varepsilon)$, where ε is a standard normal random variable and $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF). Then, Equation 5 for $k \geq 1$ becomes:

$$G_k = \Phi(\varepsilon) - \Phi\left(\frac{\ln y - \ln \theta_k}{\beta}\right)$$

Since $\Phi(\cdot)$ is monotonically increasing, it always preserves the order of inputs; thus, G_k has the same sign as the difference of its inputs. Thus, we can model damage exactly via the sign of the following variable, g_k:

$$g_k = \varepsilon + \frac{1}{\beta}(\ln \theta_k - \ln y) \tag{8}$$

This yields the linearized formulation for a single-building damage simulation by substituting Equation 1 for $\ln y$:

$$g_k = -\frac{1}{\beta}(\tau\zeta + \phi\xi - \beta\varepsilon) + \frac{1}{\beta}(\ln \theta_k - \mu) \tag{9}$$

The first term of Equation 9, $-\frac{1}{\beta}(\tau\zeta + \phi\xi - \beta\varepsilon)$, is a linear combination of three independent standard normal variables (ε, ζ, and ξ), which still is a normally distributed and its variance is $1 + (\tau^2 + \phi^2)/\beta^2$ or $1 + \sigma^2/\beta^2$, where $\sigma = \sqrt{\tau^2 + \phi^2}$ is the total standard deviation of the ground shaking intensity. The second term, $\frac{1}{\beta}(\ln \theta_k - \mu)$, contains no random variables and thus acts as a deterministic bias, i.e., a shift in the mean of the distribution. Therefore, g_k is normally distributed as:

$$g_k \sim \mathcal{N}\left(\frac{\ln \theta_k - \mu}{\beta}, 1 + \frac{\sigma^2}{\beta^2}\right) \tag{10}$$

Damage simulation can be performed similarly to the traditional approach (Equation 6), with the substitution of G_k by g_k:

$$DS = \max\{k \in \{0, \ldots, n_{\text{ds}} - 1\} \mid g_k < 0\} \tag{11}$$

Recall that in the traditional framework, damage simulation involves a two-step procedure: i) simulating ground motion (Equation 1), and ii) evaluating Equations 4 to compute G_k. The procedure incorporates three random variables—ζ, ξ, and u (Equations 1 and 5). In contrast, the linearized damage simulation using g_k based on Equations 10 streamlines the traditional two-step modeling process into a single-step formulation. Furthermore, this approach eliminates redundant random variables, enabling the damage to be modeled using a single Gaussian random variable.

Furthermore, the linearization provides a transparent perspective on the coupling between ground-motion models and fragility curve parameters, elucidating their joint influence on the resulting risk (Equation 10). For instance, an increase in θ_k or a decrease in μ shifts the distribution toward the positive domain, thereby reducing the probability of damage. An increase in σ flattens the distribution, which leads to an increase in risk when $\ln \theta_k > \mu$ and a decrease when $\ln \theta_k < \mu$. Finally, as β increases, the distribution converges toward $\mathcal{N}(0, 1)$; in this limiting case, $P[DS > k]$ approaches 0.5, and the specific values of μ and θ_k become increasingly irrelevant to the risk assessment. Further details regarding these parametric sensitivities are provided in the Section S1 of Supplementary Material.

3.2 | Geometric Interpretation of the Linearized Formulation

We demonstrate geometrically how the linearized formulation reduces the traditional framework's three-dimensional system (ζ, ξ, and u) to a single dimension. We begin by analyzing the distribution of G_k (Equation 5) in the traditional framework within the space spanned by the ground shaking intensity, $\ln y$, and the uniform damage sample, u. For a given k, we map the regions where G_k is negative or positive within the $(\ln y, u)$ space (Figure 2(a)). As shown in Figure 2(a), the regions corresponding to $G_k > 0$ (or $DS < k$) and $G_k \leq 0$ (or $DS \geq k$) are well separated, with $G_k = 0$ acting as the boundary (red and blue shaded regions separated by a black curve). This boundary naturally traces the normal CDF fragility function, forming a non-linear curve, as the limit state $G_k = 0$ is defined by $u = P(DS \geq k)$ (Equation 5). Due to this non-linearity, the gradient vector of G_k—the direction of maximum change—varies depending on the specific values of $\ln y$ and u. This is evident from the varying orientations of the contour lines (gray dotted lines in Figure 2(a)).

In contrast, the linearized formulation employing g_k (Equation 9) transforms this non-linear boundary into a linear one characterized by a unique gradient direction of g_k. The limit-state boundary transforms into a straight line in the $(\ln y, \varepsilon)$ space (Figure 2(b); green and purple shaded regions separated by a black line). Thus, to maintain $g_k = 0$, a change in $\ln y$ requires a constant proportional change in ε. This linearity applies not only to $g_k = 0$ but to any constant value of g_k; thus, the contours (isolines) of g_k are parallel to the limit-state line. Therefore, the variation of g_k occurs exclusively in the direction perpendicular to these contours, corresponding to a slope of $1/\beta$ (Figure 2(b)).

We extend this concept to three-dimensional space by linearly decomposing $\ln y$ into its components, ξ and ζ (Equation 1). In the traditional framework, a curved surface (black surface in Figure 2(c)) separates the failure ($G_k < 0$) and non-failure ($G_k \geq 0$)

regions (red and blue shaded regions). As discussed with Figure 2(a), the curvature of this surface prevents the identification of a single, constant direction responsible for the variation in G_k. In contrast, our linearized formulation yields a linear planar boundary in the three-dimensional $(\xi, \zeta, \varepsilon)$ space (Figure 2(d); green and purple shaded regions separated by a plane), analogous to the linear boundary in Figure 2(b). Consequently, the variation in g_k occurs entirely along the unique direction defined by the normal vector of this planar boundary.

This normal vector (**n**) is derived from Equation 9:

$$\mathbf{n} = \begin{bmatrix} \tau/\beta & \phi/\beta & -1 \end{bmatrix}.$$

This property allows us to model g_k effectively as a one-dimensional problem along this direction using a single Gaussian random variable, as any vector component orthogonal to **n** does not cause variation in g_k (Equation 10).

We can extend this visualization to include all possible k values, drawing the limit state boundaries for multiple damage state categories (Figure 3). Figure 3(a) illustrates these boundaries in the traditional framework using uniform damage samples u, while Figure 3(b) shows them in the linearized framework with standard normal damage samples ε. In the traditional framework, the limit state boundaries appear as non-linear curved surfaces with slopes that vary by location and k. Conversely, the linearized formulation exhibits perfectly linear planar boundaries, defined by the normal vector **n** (Figure 3(b)). Notably, if the same β is used for all k, these linear boundaries are parallel as seen in Figure 3(b). If different β values are used, they do not share **n**, thus the planes are not parallel and may intersect.

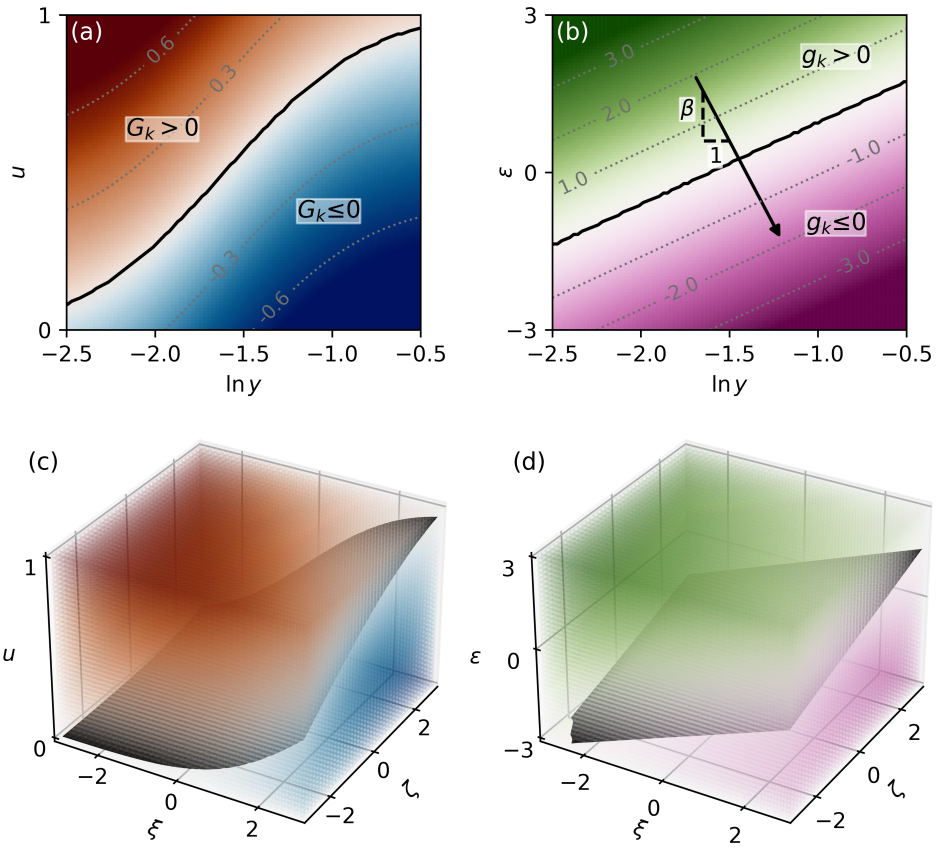

FIGURE 2 Comparison of the damage function behavior in the traditional and linearized frameworks. (a) Variation of G_k in the traditional $(\ln y, u)$ space. (b) Variation of g_k in the linearized $(\ln y, \varepsilon)$ space. Red and green shaded zones indicate regions with positive values ($DS < k$), while blue and purple zones indicate negative values ($DS \geq k$). The thick black line indicates the limit state ($G_k = 0$ or $g_k = 0$), and gray dotted lines represent contours. Darker colors indicate larger absolute values. The black arrow in (b) denotes the direction of variation for g_k. (c) Variation of G_k in the 3D (ξ, ζ, u) space. (d) Variation of g_k in the 3D $(\xi, \zeta, \varepsilon)$ space. The color code matches (a) and (b): (c) uses red/blue and (d) uses green/purple. The black curved surface in (c) and the planar surface in (d) represent the limit states.

3.3 | Multiple-Site Linearization

For a region, the linearized damage-state formulation for ith building can be expressed similar to single building case by incorporating the standard normal damage variable ε_i instead of u_i (Equation 9), but incorporating the within-event spatial correlation of ground shaking (Equation 2):

$$g_{ik} = -\frac{1}{\beta_i}\left(\tau_i\zeta + \phi_i(\mathbf{L}_i \cdot \boldsymbol{\xi})\right) + \varepsilon_i + \frac{1}{\beta_i}(\ln\theta_i^k - \mu_i)$$

, where $\mathbf{L}_i$ is the ith row of $\mathbf{L}$ from Equation 2, and $\boldsymbol{\xi} \in \mathbb{R}^N \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This can be expressed in vector form as:

$$\mathbf{g}_k = \mathbf{A}\mathbf{z} + \mathbf{b}_k \tag{12}$$

, where the damage for the i-th building is simulated using the sign of the i-th entry of $\mathbf{g}_k \in \mathbb{R}^N$ (Equation 11). The components are defined as $\mathbf{A} \in \mathbb{R}^{N\times(2N+1)}$, $\mathbf{b}_k \in \mathbb{R}^N$, and $\mathbf{z} \in \mathbb{R}^{2N+1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$A = \left[-\mathbf{B}\boldsymbol{\tau}\ -\mathbf{B}\mathbf{F}\mathbf{L}\ \mathbf{I}\right], \qquad \mathbf{b}_k = \mathbf{B}(\ln\boldsymbol{\theta}_k - \boldsymbol{\mu}), \qquad \mathbf{z} = \begin{bmatrix}\zeta\\ \boldsymbol{\xi}\\ \boldsymbol{\varepsilon}\end{bmatrix}. \tag{13}$$

Here, $\mathbf{B} = \mathrm{diag}(1/\beta_1, \ldots, 1/\beta_N)$ and $\mathbf{F} = \mathrm{diag}(\phi_1, \ldots, \phi_N)$. The matrix $\mathbf{A}$ contains the uncertainty-related components: the ground-motion standard deviations ($\boldsymbol{\tau}$ and $\boldsymbol{\phi}$), the Cholesky factor $\mathbf{L}$, and the dispersions of the fragility curves $\boldsymbol{\beta}$. Similar to Equation 9, the vector $\mathbf{b}_k$ contains the deterministic, mean-shifting components: the median ground shaking of the fragility curve $\boldsymbol{\theta}_k$ minus logarithmic median of ground shakings $\boldsymbol{\mu}$, divided by dispersion parameter of fragility curve $\boldsymbol{\beta}$.

Equation 12 defines a transformation of a $(2N + 1)$-dimensional standard normal vector $\mathbf{z}$ into an N-dimensional normally distributed vector, $\mathbf{g}_k$, with mean $\mathbf{b}_k$ and covariance matrix $\mathbf{A}\mathbf{A}^\mathsf{T}$:

$$\mathbf{g}_k \sim \mathcal{N}(\mathbf{b}_k, \mathbf{A}\mathbf{A}^\mathsf{T}) \tag{14}$$

, where the covariance matrix $\mathbf{A}\mathbf{A}^\mathsf{T} \in \mathbb{R}^{N\times N}$ is:

$$\mathbf{Cov}(\mathbf{g_k}) = \mathbf{A}\mathbf{A}^\mathsf{T} = \mathbf{I} + (\mathbf{B}\boldsymbol{\tau})(\mathbf{B}\boldsymbol{\tau})^\mathsf{T} + \mathbf{B}\mathbf{F}\mathbf{C}\mathbf{F}\mathbf{B}. \tag{15}$$

This indicates that the damage to N buildings can be modeled using a set of correlated N-dimensional Gaussian random variables, where the damage of the ith building corresponds to the ith Gaussian random variable, and the categorical damage states are

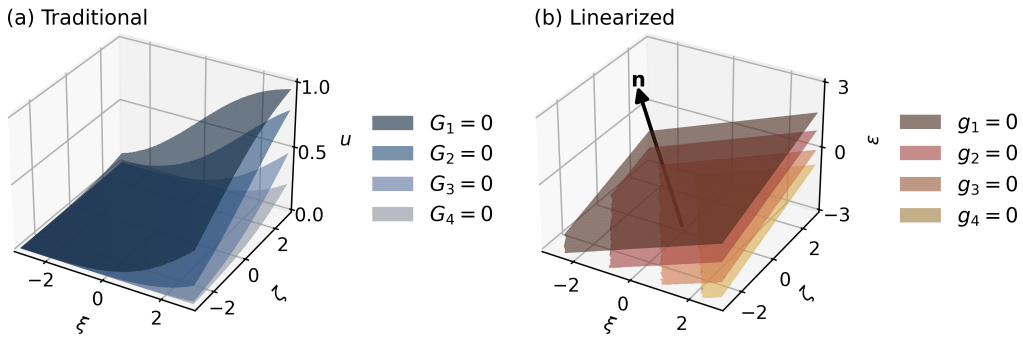

FIGURE 3 Boundaries separating damage state categories (k = 1 to 4) in: (a) the (ξ, ζ, u) space using the traditional framework, and (b) the $(\xi, \zeta, \varepsilon)$ space using the linearized framework. Darker planes indicate lower damage state thresholds, while lighter planes indicate higher damage state thresholds. The black arrow indicates the normal vector ($\mathbf{n}$) applicable to all damage states.

encoded by the signs of these random samples. Negative entries of $\mathbf{g}_k$ correspond to damage states $DS \geq k$, whereas non-negative entries correspond to $DS < k$. In practice, the zero mean random samples are generated once for all ks, and mean shifting term is applied to decide if $DS \geq k$ or not.

Similar to the single-building case, damage modeling of N buildings in the traditional framework requires a two-step procedure involving $2N + 1$ random variables: $N + 1$ for ground-motion simulation (Equation 2) and N for damage modeling (Equation 6). By contrast, the proposed formulation enables a one-shot procedure using Equation 14 and reduces the dimensionality by approximately half for large N ($2N + 1$ to N), without introducing additional mathematical assumptions.

3.4 | Statistical Characteristics of Linearized Damage Function

For any single entry in $\mathbf{g}_k$, the behavior of an individual building remains the same as the single-building case, as the marginal distributions of an N-dimensional Gaussian are still Gaussian. The difference in the multi-building context lies in the dependency (correlation) between buildings. The distribution of $\mathbf{g}_k$ is defined by two components: the deterministic mean vector $\mathbf{b}_k$ and the covariance matrix $\mathbf{Cov}(\mathbf{g}_k)$. Thus, the dependency effects arise solely from the covariance matrix, $\mathbf{Cov}(\mathbf{g}_k)$.

We illustrate this using a simple toy example and interpret this in comparison with the ground-motion correlation matrix $\mathbf{Cov}(\ln \mathbf{y})$ (Equation 3), whose structure has been well established and understood [13,14]. We assume three sites, where Sites 1 and 2 are adjacent and have a within-event ground-motion correlation coefficient of 0.9, whereas Site 3 is sufficiently distant that its within-event correlations with the other two sites are zero (Figure 4(a)). The resulting total ground-shaking correlation matrix obtained for $\phi = 0.7$, and $\tau = 0.4$, which reflects the combined effects of within-event ($\boldsymbol{\phi}$ and $\mathbf{C}$) and between-event variability ($\boldsymbol{\tau}$), is presented in Figure 4(b). The diagonal entries are equal to 1, as expected. A high correlation is observed between Site 1 and Site 2 (0.97), represented by a bright tone, which is result of the combined effect of fully correlated between-event residual and highly correlated within-event residual (ρ of 0.9). In contrast, correlations involving Site 3 are moderate (0.49), attributed to the fully correlated between-event component but the absence of correlation in the within-event component.

Using Equation 15, $\mathbf{Cov}(\mathbf{g_k})$ is obtained for $\beta = 0.6$ and the same ground motion standard deviations for Figure 4(b), i.e., $\phi = 0.7$, and $\tau = 0.4$ (Figure 4(c)). The strong spatial ground motion correlation ($\mathbf{Cov}(\mathbf{ln\,y})$ between Sites 1 and 2 is substantially reduced in $\mathbf{Cov}(\mathbf{g_k})$, 0.97 to 0.60 (Figure 4 (b) and (c)). In the same way, the correlation terms regarding Site 3 are reduced from 0.49 to 0.16. These reductions reflect the structure of $\mathbf{Cov}(\mathbf{g_k})$ (Equation 15): the addition of $\mathbf{I}$, which comes from independent damage state modeling and affects only the diagonal entries, decreases the magnitude of the off-diagonal correlations, so that even highly correlated ground motions produce only moderately correlated damage outcomes, making $\mathbf{g}_k$ have lower inter-site correlation than $\ln \mathbf{y}$. i.e., $\mathbf{Cov}(\mathbf{g_k})$ is diagonal dominant with small-to-modest inter-site dependencies.

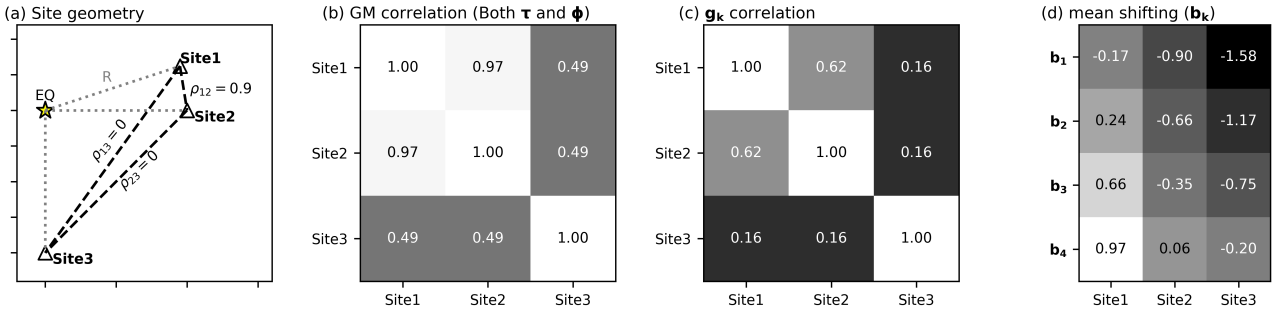

FIGURE 4 (a) Assumed three-site geometry. Sites 1 and 2 are close, with a within-event ground-motion correlation of 0.9, while Site 3 is distant and has zero correlation with the others. (b) Ground-motion correlation matrix illustrating the combined effects of within-event and between-event variability. Brighter colors indicate higher correlation coefficients. (c) Covariance matrix $\mathbf{A}\mathbf{A}^{\mathsf{T}}$ of $\mathbf{g}_k$, obtained using the coupled linearized ground-motion and damage model. (d) Mean-shifting vectors $\vec{b}_k$.

4 | DIMENSIONALITY REDUCTION IN REGIONAL SEISMIC RISK

Our linearized formulation simplifies damage simulation and facilitates modeling in a reduced-dimensional space from $2N + 1$ to N because the traditional damage modeling framework involving $N + 1$ ground shaking and N damage state random variables (Equations 2 and 5) is now reduced to N-dimensional Gaussian process (Equation 14). However, simulating N-dimensional correlated Gaussian random variables remains computationally demanding for large N. To address this challenge, we investigate the potential for further dimensionality reduction within the linearized ground-motion–damage coupling formulation (Equations 12 and 14).

We first examine two different strategies for dimensionality reduction in simulating $\mathbf{g}_k$: the standard Principal Component Analysis (PCA)[47] and PPCA[38]. We discuss their effectiveness from an eigenvalue-eigenvector perspective, assessing their applicability to our model $\mathbf{g}_k$. Subsequently, we reformulate the simulation of $\mathbf{g}_k$ using PPCA, demonstrating that it is better suited for dimensionality reduction in regional risk analysis.

4.1 | PCA vs. PPCA

PCA is a widely used dimensionality reduction technique for multivariate Gaussian random variables[47,26]. The method approximates these variables as linear combinations of lower-dimensional principal components by exploiting their underlying linear correlations. These dependencies are encoded within the covariance matrix—specifically $\mathbf{Cov}(\mathbf{g}_k)$ in this study—whose eigenvalues are real and non-negative due to the matrix's positive semi-definite property. The eigenvector corresponding to the largest eigenvalue of $\mathbf{Cov}(\mathbf{g}_k)$ identifies the primary axis of linear dependency, with the eigenvalue quantifying the variance captured along this vector. Subsequent eigenvectors, ordered by their corresponding eigenvalues, identify directions of decreasing variance under the constraint of orthogonality to preceding components. Since the sum of the eigenvalues represents the variance within the space spanned by the corresponding eigenvectors, the number of components m required to capture a significant proportion (e.g., 90%) of the total variance in $\mathbf{g}_k$ is determined by:

$$F_m = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{j=1}^{N_b} \lambda_j} \times 100 \tag{16}$$

where F_m is the cumulative variance explained, expressed as a percentage. A reduction in the value of m necessary to reach a specific threshold indicates stronger correlations between the g_k values across various sites, thereby enhancing the efficiency of PCA representation.

PPCA extends the standard PCA framework[38]. While standard PCA is most effective when strong correlations exist, PPCA enables dimensionality reduction even in cases with low-to-moderate correlation. This approach decomposes the covariance matrix $\mathbf{Cov}(\mathbf{g}_k)$ into a highly correlated component and an independent noise component, such that $\mathbf{Cov}(\mathbf{g}_k) = \mathbf{Cov}(\mathbf{g}_k)^{(cor)} + \mathbf{Cov}(\mathbf{g}_k)^{(ind)}$. Standard PCA is then applied to the correlated portion, $\mathbf{g}_k^{(cor)}$, to obtain the dimensionality reduced approximation of $\tilde{\mathbf{g}}_k^{(cor)}$. Consequently, $\mathbf{g}_k$ is represented by the low-dimensional $\tilde{\mathbf{g}}_k^{(cor)}$ and the full-dimensional $\mathbf{g}_k^{(ind)}$. Since standard PCA is still applied to $\tilde{\mathbf{g}}_k^{(cor)}$, the efficiency of PPCA heavily lies in how to decompose it among infinite possible combinations to maximize the linear correlations of variables in $\mathbf{Cov}(\mathbf{g}_k)^{(cor)}$. The more $\mathbf{g}_k^{(cor)}$ is highly correlated, the more efficient.

As demonstrated in the preceding section, the linearized damage function $\mathbf{g}_k$ typically exhibits low-to-moderate correlations, suggesting that PPCA would be more effective than standard PCA. However, a significant challenge in applying PPCA to $\mathbf{g}_k$ is that the optimal decomposition is not known a priori. The optimal decomposition generally necessitates a complete eigendecomposition of the covariance matrix $\mathbf{Cov}(\mathbf{g}_k)$ to extract the shared independent noise component of $\mathbf{Cov}(\mathbf{g}_k)$, based on its low-order eigenvalues. However, for large N, this approach is computationally prohibitive due to its $O(N^3)$ complexity[48]. To fully leverage the efficiency of PPCA, a decomposition that avoids full numerical eigendecomposition is preferable. In the following sections, we develop an efficient PPCA decomposition of $\mathbf{g}_k$ through an analytical derivation and numerical experiments that circumvent the need for explicit full eigendecomposition.

For more detailed comparison between standard PCA and PPCA using illustrative examples, see Section S2 of the Supplementary Material.

4.2 | Analytical Derivation of Eigenvalues of $\mathbf{Cov}(\mathbf{g}_k)$

To determine the optimal decomposition of the covariance matrix $\mathbf{Cov}(\mathbf{g}_k)$ for PPCA, we examine its eigenvalue properties. We focus primarily on the lower-order eigenvalues of $\mathbf{Cov}(\mathbf{g}_k)$, which are instrumental in identifying independent noise components and isolating the highly correlated structure. We first present an analytical derivation of the lower bound eigenvalues for $\mathbf{Cov}(\mathbf{g}_k)$ using an idealized spatial distribution that represents a limiting case in regional risk analysis. Subsequently, we investigate more realistic scenarios through illustrative examples across various spatial configurations. These cases serve to validate the analytical derivation and demonstrate how the eigenvalues evolve in response to diverse building distributions.

The theoretical analysis focuses on a limiting geometric configuration designed to facilitate a formal derivation: t building clusters separated by large spatial distances, with each cluster containing densely packed buildings. A building cluster represents a group of buildings in close spatial proximity. Each cluster i ($i = 1, 2, \ldots, t$) contains n_i buildings, satisfying $\sum_{i=1}^{t} n_i = N$. The correlation between buildings belonging to distinct clusters is assumed to be zero. Within each cluster, all the inter-building within-event ground shaking correlation is assumed to be a constant $\rho_{\max} < 1$. This value represents the maximum achievable inter-site correlation, accounting for the nugget effect[14].

Under these idealized conditions, the lower bound eigenvalues of $\mathbf{Cov}(\mathbf{g}_k)$ are derived as $1 + \min_i \left(\frac{\phi_i^2}{\beta_i^2}\right)(1 - \rho_{\max})$ (see Section S3 of Supplementary Material). In the specific case where the within-event standard deviations (ϕ) and fragility curve dispersion parameter (β) are constant across all sites, there are exactly $(N - t)$ minimum eigenvalues equal to:

$$1 + \frac{\phi^2}{\beta^2}(1 - \rho_{\max}). \tag{17}$$

Therefore, there exist t, the assumed number of building clusters, eigenvalues that strictly exceed this minimum value.

4.3 | Impact of Building Distribution on the Eigenvalues of $\mathbf{Cov}(\mathbf{g}_k)$

Building upon this analytical derivation, we investigate how the eigenvalues evolve under more realistic spatial configurations. More realistic scenarios are constructed by relaxing the idealized constraints in two primary ways. First, the inter-cluster correlation constraint is loosened; by adjusting the spatial separation between building clusters, we generate geometries with various inter-cluster correlations. (Figure 5 (a)). These configurations are categorized as close, intermediate, or distant, corresponding to within-event ground shaking intensity correlations between cluster centers of 0.8, 0.38, and 0, respectively. Second, the intra-cluster correlation constraint is relaxed; by increasing the spatial extent of a cluster, the within-event correlations between buildings inside that cluster may fall below $\rho_{\max}$ (Figure 5 (b)). These cases are categorized by building density as high, moderate, or low, corresponding to average within-event correlations between buildings within the cluster of 0.8, 0.23, and 0.002, respectively.

Together, these modifications represent a broad spectrum of spatial distributions encountered in practice. We demonstrate these effects using a system of 15 buildings organized into three clusters ($t = 3$), subjected to varying inter-cluster distances and within-cluster densities.

Figure 5 (c) shows the eigenvalue spectrum change when the inter-cluster distance is changed. First, the minimum eigenvalues of 1.29, which precisely match with the theoretical prediction of Equation 17, remain invariant with respect to the change in inter-cluster distances. Furthermore, the number of eigenvalues exceeding this minimum is exactly $t = 3$, the number of building clusters. The primary variation occurs in the relative magnitudes; as inter-cluster distances decrease, the second and third eigenvalues converge toward the theoretically derived minimum value. This behavior is consistent with the physical interpretation that as clusters move closer, they begin to behave as a single aggregate cluster rather than three distinct entities, effectively reducing the number of eigenvalues greater than the theoretically derived minimum toward one.

Figure 5 (d) illustrates the change in the eigenvalue spectrum as the building density within each cluster varies. These results demonstrate that within-cluster building density influences the minimum eigenvalues of the covariance matrix. The high-density case shown in Figure 5(b) exhibits distinctive three (t) non-minimum eigenvalues and 12 ($N - t$) minimum eigenvalues. However, as the spatial extent of each cluster increases—thereby reducing within-event correlations below $\rho_{\max}$—the eigenvalues no longer maintain this constant minimum value. This transition results in a gradual decay as observed in the "moderate" case. It decreases the first three eigenvalues, while the trailing eigenvalues increase. We also found that the minimum eigenvalue in this case closely approximates the value predicted by Equation 17 when $\rho_{\max}$ is replaced by the observed maximum correlation $\hat{\rho}_{\max}$

for the specific building geometry:

$$\lambda_{\min} \approx 1 + \frac{\phi^2}{\beta^2}(1 - \hat{\rho}_{\max}). \tag{18}$$

When the cluster size increases to the point that intra-cluster correlations approach zero (the "low density" case in Figure 5(b) and (d)), the spectrum reveals 14 nearly homogeneous minimum eigenvalues, with only a single eigenvalue exceeding this threshold. This behavior mirrors the "Close cluster" case shown in Figures 5(a) and (c), which also functions as a single cluster. In this state, the system effectively behaves as a single large cluster characterized by negligible spatial correlation across all buildings. Consequently, the number of building clusters t within our linearized formulation is defined in a relative sense, rather than by absolute physical distances. The primary distinction lies in the converged eigenvalue levels and the relative magnitude between the dominant and secondary eigenvalues. The converged eigenvalue in this case also follows Equation 18 with $\hat{\rho}_{\max} = 0$.

Across all examined cases, it is evident that a substantial number of principal components are required to explain a significant fraction of the total variance in $\mathbf{g}_k$. In the scenarios presented in Figure 5 (a), all three configurations require 12 out of 15 components to capture 90% of the variance (Figure 5 (e)). Similarly, in Figure 5 (b), all three cases require 13 out of 15 components to reach the same 90% threshold (Figure 5 (f)). Consequently, standard PCA yields a dimensionality reduction of only 13% to 20%, consistent with previous findings. However, having established the convergence behavior of these eigenvalues and close correlation between the first few dominant eigenvalue to the number of building clusters t, we can now leverage these insights to formulate a PPCA framework for the linearized damage function $\mathbf{g}_k$. This approach allows us to circumvent the computationally expensive full eigendecomposition of $\mathbf{Cov}(\mathbf{g}_k)$.

within-event ground motion correlations between cluster centers of 0.8, 0.38, and 0, respectively.

4.4 | Decomposition of $\mathbf{Cov}(\mathbf{g}_k)$ for PPCA

Since standard PCA proves inefficient for dimensionality reduction in our linearized damage modeling formulation, we adopt PPCA. This strategy reformulates the generative model of $\mathbf{g}_k$ by extracting the highly correlated component and isolating the independent component. This separation allows us to apply dimensionality reduction to the correlated component efficiently and facilitates fast simulation of the independent component.

4.4.1 | Decomposition of $\mathbf{g}_k$

Instead of Equation 12, we rewrite the generative model of $\mathbf{g}_k$ as:

$$\mathbf{g}_k = \mathbf{W}\mathbf{x} + c\mathbf{z} + \mathbf{b}_k \tag{19}$$

where $\mathbf{W} \in \mathbb{R}^{N\times N}$, c is a scalar constant, and $\mathbf{x}, \mathbf{z} \in \mathbb{R}^N$ are independent standard normal random vectors. To ensure this formulation reproduces the statistics of the original model (Equation 12), the covariance of the right-hand side of Equation 19 must match $\mathbf{Cov}(\mathbf{g}_k)$ (Equation 15). This condition requires (see Section S4 of the Supplementary Material for derivation):

$$\mathbf{W} = \mathbf{U}\mathbf{\Sigma} \tag{20}$$

where

$$\mathbf{\Sigma} = (\mathbf{\Lambda} - c^2\mathbf{I})^{1/2} \tag{21}$$

Here, $\mathbf{\Lambda} \in \mathbb{R}^{N\times N}$ is a diagonal matrix containing the eigenvalues of $\mathbf{Cov}(\mathbf{g}_k)$ ordered in descending magnitude, $\mathbf{I} \in \mathbb{R}^{N\times N}$ is the identity matrix, and $\mathbf{U} \in \mathbb{R}^{N\times N}$ is the matrix of eigenvectors corresponding to $\mathbf{\Lambda}$. The $\mathbf{U}$ and $\mathbf{\Sigma}$ in Equation 20 can be visualized in matrix form as:

$$\mathbf{U} = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_N\end{bmatrix}, \qquad \mathbf{\Sigma} = \begin{bmatrix}\sqrt{s_1} & 0 & \cdots & 0 \\ 0 & \sqrt{s_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sqrt{s_N}\end{bmatrix} \tag{22}$$

where

$$s_i = \lambda_i - c^2 \tag{23}$$

and λ_i are the eigenvalues of $\mathbf{Cov}(\mathbf{g}_k)$ and $\mathbf{u}_i$ are the corresponding eigenvectors.

4.4.2 | Selection of Σ and c

While infinite pairs of $\mathbf{W}$ and c can satisfy Equation 19, our goal is to minimize computational cost of $\mathbf{g}_k$ while maintaining its accuracy. The primary burden is the calculation of the first term:

$$\boldsymbol{\omega} = \mathbf{W}\mathbf{x} \tag{24}$$

which scales with $O(N^2)$, while the remaining terms ($c\mathbf{z}$ and $\mathbf{b}_k$) scale with $O(N)$. Therefore, we aim to select an optimal constant c and corresponding $\boldsymbol{\Sigma}$ that allows reducing the effective number of columns of matrix $\mathbf{W}$ from N to a smaller dimension.

To do that, we desire the covariance of $\boldsymbol{\omega}$ to have many zero eigenvalues and only a few dominant ones to effectively reduce dimensionality. The covariance of $\boldsymbol{\omega}$ is given by:

$$\mathbf{Cov}(\boldsymbol{\omega}) = \mathbf{W}\mathbf{W}^{\mathsf{T}} = \mathbf{U}\boldsymbol{\Sigma}^2\mathbf{U}^{\mathsf{T}} \tag{25}$$

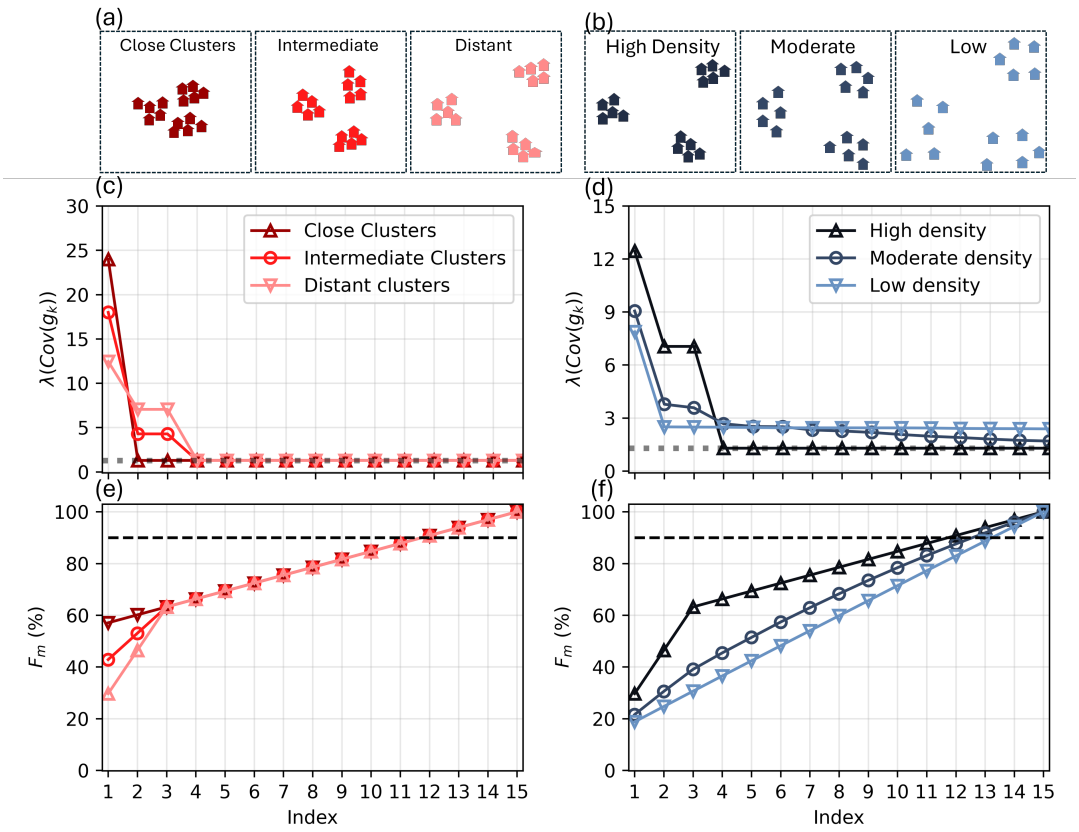

FIGURE 5 Eigenvalue analysis of building configurations across varying spatial distributions. (a) Configurations with varying inter-cluster distances, categorized as Close (dark red), Intermediate (red), and Distant (light red) clusters, corresponding to within-event ground shaking intensity correlations between cluster centers of 0.8, 0.38, and 0, respectively. (b) Configurations with varying cluster sizes, categorized as High (dark blue), Moderate (blue), and Low (light blue) density, corresponding to average within-event correlations between buildings within the cluster of 0.8, 0.23, and 0.002, respectively. (c) and (d): the eigenvalues for the configurations in (a) and (b), respectively; the theoretically driven lower bound is indicated by the gray dotted horizontal lines. (e) and (f): the cumulative variance contribution (F_m) for the configurations in (a) and (b), respectively, with the dashed black line denoting the 90% threshold.

where

$$\mathbf{U} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_N \end{bmatrix}, \quad \mathbf{\Sigma}^2 = \begin{bmatrix} s_1 & 0 & \cdots & 0 \\ 0 & s_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & s_N \end{bmatrix}, \quad \mathbf{U}^\mathsf{T} = \begin{bmatrix} \mathbf{u}_1^\mathsf{T} \\ \mathbf{u}_2^\mathsf{T} \\ \vdots \\ \mathbf{u}_N^\mathsf{T} \end{bmatrix} \tag{26}$$

Here, $\mathbf{u}_i \in \mathbb{R}^N$ represents the eigenvectors of $\mathbf{Cov}(\mathbf{g}_k)$ (Equation 22). Note that $\mathbf{u}_i$ also represents the eigenvectors of $\mathbf{Cov}(\boldsymbol{\omega})$, and the constant c is selected as the square root of the lower bound eigenvalue of $\mathbf{Cov}(\mathbf{g}_k)$ (Equation 18). This choice ensures that s_i values for $i > t$ remain close to zero (See Section S5 of the Supplementary Material for derivation):

$$c^* = \sqrt{1 + (1 - \hat{\rho}_{\max})\frac{\phi^2}{\beta^2}} \tag{27}$$

Substituting this value into Equation 20, then we define $\mathbf{W}^*$ as:

$$\mathbf{W}^* = \mathbf{U}\mathbf{\Sigma}^* \tag{28}$$

where $\mathbf{\Sigma}^* = (\mathbf{\Lambda} - c^{*2}\mathbf{I})^{1/2}$ and $s_i^* = \lambda_i - (c^*)^2$ Then, Equation 19 can be also re-written as:

$$\mathbf{g}_k = \mathbf{W}^*\mathbf{x} + c^*\mathbf{z} + \mathbf{b}_k \tag{29}$$

and

$$\mathbf{Cov}(\boldsymbol{\omega}^*) = \mathbf{W}^*\mathbf{W}^{*\mathsf{T}} \tag{30}$$

where $\boldsymbol{\omega}^* = \mathbf{W}^*\mathbf{x}$.

The total variance of $\mathbf{g}_k$ explained by the first m largest principal components of $\mathbf{Cov}(\boldsymbol{\omega}^*)$ is defined as:

$$F_m = \frac{Nc^* + \sum_{i=1}^{m} \lambda_i}{Nc^* + \sum_{j=1}^{N_b} \lambda_j} \times 100 \tag{31}$$

where Nc^* represents the variance contribution from the independent noise component. Because Nc^* acts as a non-zero constant contributor to the total variance regardless of the chosen m, and since the high correlation within $\boldsymbol{\omega}^*$ allows for a minimal m to explain the majority of the variance in $\mathbf{g}_k$, this formulation is expected to reach the variance contribution threshold more efficiently than standard PCA.

Figure 6 illustrates the impact of applying the optimal c^* to the eigenvalues of $\mathbf{Cov}(\boldsymbol{\omega})(i.e., \mathbf{Cov}(\boldsymbol{\omega}^*))$ for the building configurations previously analyzed in Figure 5, along with the resulting cumulative variance contribution to $\mathbf{g}_k$ (Equation 31). Figure 6(a) displays the eigenvalue spectra of $\mathbf{Cov}(\boldsymbol{\omega}^*)$ for the building configurations shown in Figure 5(a). In contrast to the original covariance $\mathbf{Cov}(\mathbf{g}_k)$, $\mathbf{Cov}(\boldsymbol{\omega}^*)$ possesses zero eigenvalues, as seen by comparing Figure 6(a) to Figure 5(c). Furthermore, the number of non-zero eigenvalues is identical to the number of building clusters, which is $t = 3$ in this case. If the density of buildings within each cluster decreases (e.g., the moderate density case in Figure 5(b) and Figure 6(b)), several trailing components remain non-zero, although they converge toward zero. In cases where the building density within a cluster is low (Figure 6(b)), the spectra contain many zeros with only one non-zero eigenvalue because the framework treats them as a single cluster with no ground-shaking correlation.

The cumulative variance contribution shown in Figures 6(c) and (d) reaches the threshold far more rapidly than the case where independent noise components are not extracted from $\mathbf{Cov}(\mathbf{g}_k)$. In all instances, the variance contribution exceeds 90% within only three principal components, whereas 12 or 13 components were previously required (Figures 5(e) and (f)). Interestingly, the least favorable scenario for this PPCA framework occurs when the building distribution is neither highly dense nor highly sparse. A "moderate density" distribution results in the slowest convergence of the total variance contribution (intermediate blue curve of Figure 6(d)). This indicates that the framework is highly efficient for dense metropolitan areas. For cases with moderate ground-motion correlations, the approach remains efficient, though the efficiency gains are less pronounced than in dense urban settings. For independently located buildings, the framework is also efficient as the system can be modeled with a single principal component. However, the ground-shaking and damage modeling for independent buildings is already computationally manageable, as the complexity scales with $O(N)$ when correlations are neglected.

Since the diagonal entries of $\mathbf{\Sigma}^*$ become zero for all indices greater than t, we can effectively truncate $\mathbf{W}^*$ without loss of information. The reduced matrix $\mathbf{W}_t^*$ is defined as:

$$\mathbf{W}_t^* = \mathbf{U}_t\mathbf{\Sigma}_t \tag{32}$$

where $\mathbf{W}_t^* \in \mathbb{R}^{N\times t}$ and $\mathbf{\Sigma}_t \in \mathbb{R}^{t\times t}$ are defined as:

$$\mathbf{U}_t = \begin{bmatrix}\mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_t\end{bmatrix} \tag{33}$$

and

$$\mathbf{\Sigma}_t = \begin{bmatrix}\sqrt{s_1^*} & 0 & \cdots & 0\\ 0 & \sqrt{s_2^*} & \cdots & 0\\ \vdots & \vdots & \ddots & 0\\ 0 & 0 & \cdots & \sqrt{s_t^*}\end{bmatrix} \tag{34}$$

.

Using this reduced representation, Equation 19 is approximated as:

$$\mathbf{g}_k \approx \mathbf{W}_t^*\mathbf{x}_t + c^*\mathbf{z} + \mathbf{b}_k \tag{35}$$

where $\mathbf{x}_t \in \mathbb{R}^t$ and $\mathbf{z} \in \mathbb{R}^N$ are independent standard normal random vectors. Equation 35 constitutes our proposed computational framework for regional damage modeling. It integrates the linearization of the ground motion–damage coupling formulation and the dimensionality reduction using PPCA, reducing the computational complexity of the matrix-vector product from $O(N^2)$ to $O(tN)$—a significant efficiency gain when $t \ll N$.

5 | PROPOSED COMPUTATIONAL FRAMEWORK

A schematic comparison of our proposed framework in comparison with the traditional framework is presented in Figure 7. The procedures are decomposed into two: 1) pre-processing step and 2) simulation step. Pre-processing is executed once for a given portfolio and does not repeat over MC realizations. In contrast, the simulation step includes operations repeated for each MC realization (M times).

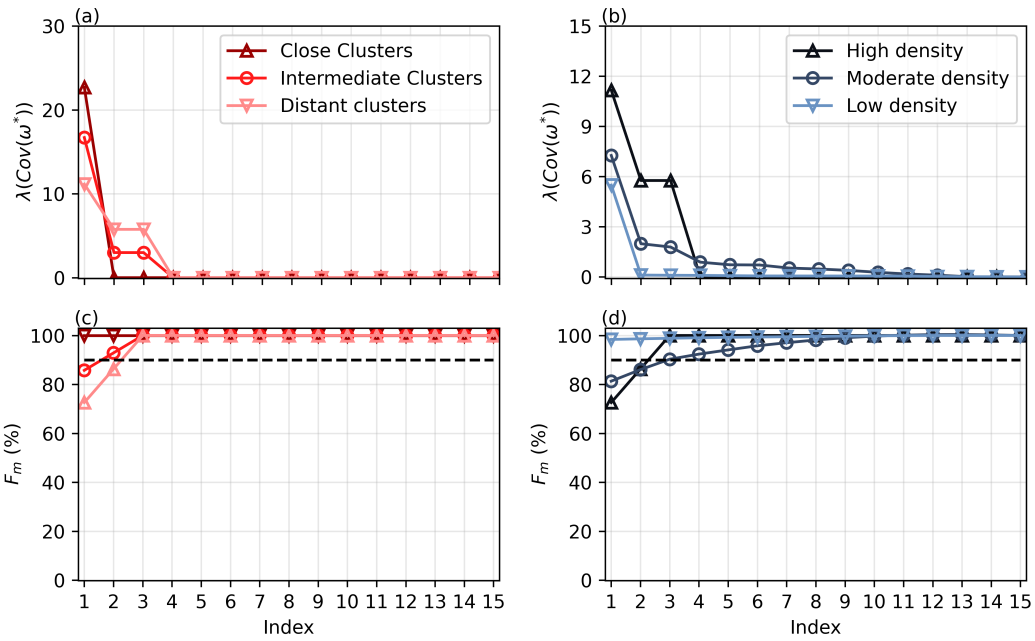

FIGURE 6 Eigenvalue analysis of $\mathbf{Cov}(\boldsymbol{\omega}^*)$ for the spatial configurations depicted in Figure 5(a) and (b): (a) Eigenvalue spectra corresponding to the cases in Figure 5(a); (b) Eigenvalue spectra corresponding to the cases in Figure 5(b); (c) and (d) Cumulative variance explained (F_m) as a function of the number of retained principal components. The horizontal dashed line denotes the 90% threshold.

In the pre-processing step, for both traditional and proposed frameworks, we need to specify the parameters of the buildings' fragility curves ($\boldsymbol{\beta}_k$ and $\boldsymbol{\theta}_k$), the ground motion between- and within-event standard deviation models ($\boldsymbol{\tau}$ and $\boldsymbol{\phi}$), and the within-event ground shaking correlation matrix $\mathbf{C}$. Afterward, in the traditional framework, one must perform the Cholesky decomposition of $\mathbf{C}$ to create $\mathbf{L}$, which is used to simulate the spatially correlated ground shaking. However, in the proposed framework, first we construct the covariance matrix $\mathbf{g}_k$, or $\mathbf{AA}^\mathsf{T}$, using Equation 15. Note that the Cholesky decomposition is not required in our framework. Then, $\mathbf{W}_t^*$ and c^* are computed using Equations 27 and 32.

In the simulation step, in both frameworks, mean ground shaking intensities ($\boldsymbol{\mu}$) at the buildings are calculated using empirical ground motion models for the earthquake scenario. This step is performed once for a scenario-based risk model. For an event-based model (random rupture magnitude and location), this process is repeated inside the MC loop (red dotted box in Figure 7). Next, in the traditional framework, the calculated $\boldsymbol{\mu}$ is combined with the standard deviation models and $\mathbf{L}$ to simulate the random ground shaking vector ($\ln \mathbf{y}$). In the proposed framework, however, $\boldsymbol{\mu}$ is not used at this stage. Instead, the matrix $\mathbf{W}_t^*$

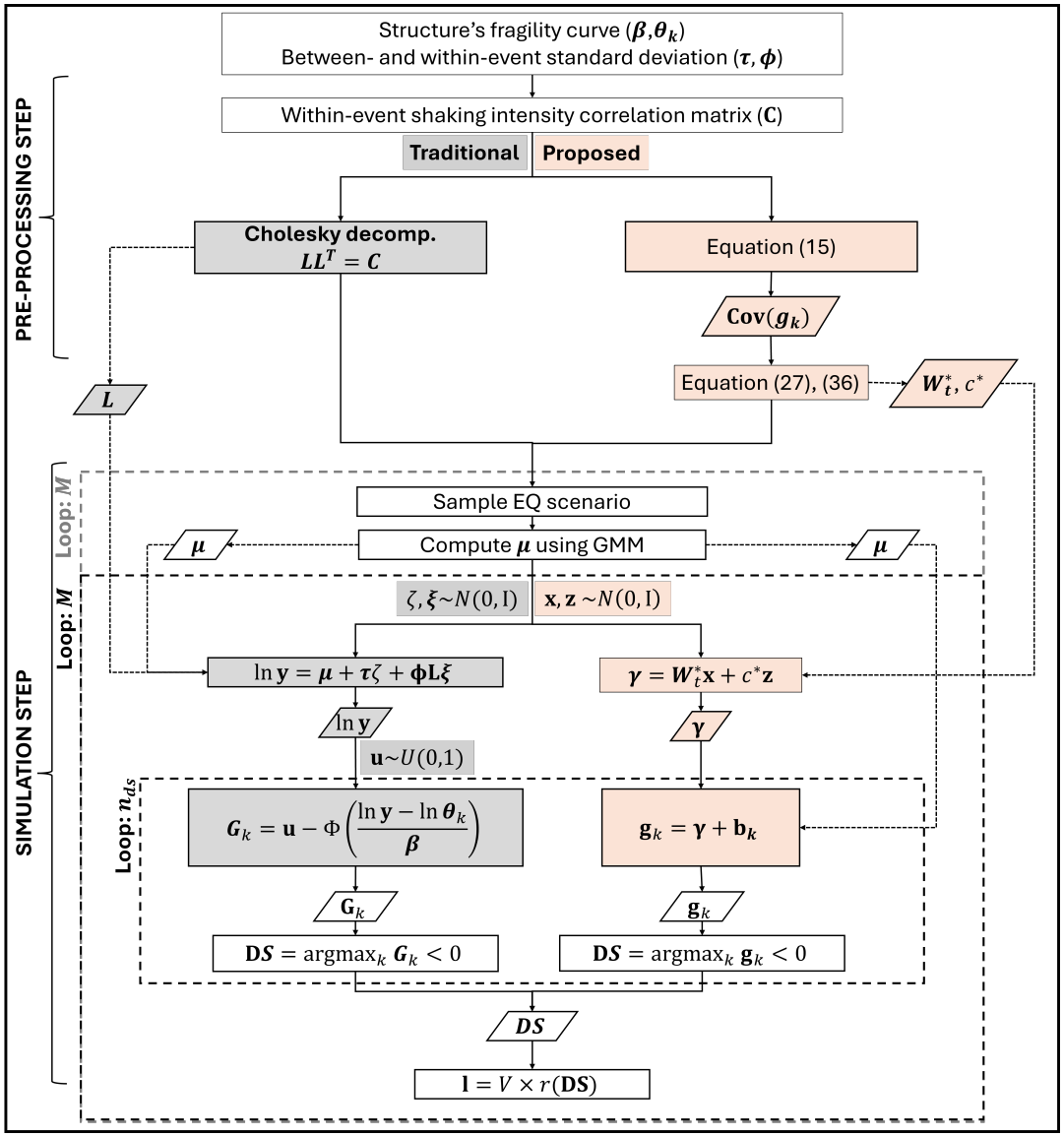

FIGURE 7 Schematic diagram comparing the traditional and the proposed frameworks for regional-scale seismic risk analysis. The steps for both traditional and proposed frameworks are white boxes; unique processes in the traditional framework are filled with light gray, while unique steps in the proposed framework are filled with light orange. The loops are presented as dotted line boxes. The red dotted loop over the *M* box corresponds to event-based (random rupture magnitude and location) risk modeling, while the black loop corresponds to scenario-based (specified rupture magnitude and location) risk modeling.

and c^* are used to calculate the zero-mean component of $\mathbf{g}_k$ (denoted as $\boldsymbol{\gamma}$ in Figure 7), which is the part of Equation 35 before adding the mean $\mathbf{b}_k$. For the damage simulation, the traditional framework requires N uniform samples ($\mathbf{u}$) per MC realization and the calculation of the standard normal CDF ($\Phi(\cdot)$). The proposed framework, at this point, adds the mean-shifting term $\mathbf{b}_k$ (which depends on $\boldsymbol{\mu}$) to $\boldsymbol{\gamma}$ to get the final $\mathbf{g}_k$. Both frameworks require a repetitive loop over the number of categorical damage states (n_{ds}). Given $\mathbf{G}_k$ or $\mathbf{g}_k$, the damage state is determined using Equation 6 or 11, and the corresponding loss is also calculated using Equation 7.

Table 1 summarizes the computational complexities of both the traditional and proposed frameworks for regional scenario-based seismic risk models. In the traditional framework, the pre-processing bottleneck lies in the Cholesky decomposition of $\mathbf{C}$ to construct the lower triangular matrix $\mathbf{L}$, which entails a computational complexity of $O(N^3)$. Conversely, the proposed framework bypasses the Cholesky decomposition during pre-processing. Instead, it requires the construction of $\mathbf{W}_t^*$ via the eigen-decomposition of $\mathbf{Cov}(\mathbf{g}_k)$. While a full decomposition scales as $O(N^3)$, computing only the top t eigenpairs reduces the complexity to $O(N^2t)$, which is effectively $O(N^2)$ when $t \ll N$. Regarding the simulation step, the primary computational bottleneck in the traditional framework is ground-motion generation, which scales with $O(N^2M)$ for large values of N and M. In contrast, the dominant procedure in the proposed framework is the simulation of $\mathbf{g}_k$, which scales with $O(tNM)$, or effectively $O(NM)$ when $t \ll N$. Overall, the proposed framework reduces the computational complexity by one order of magnitude: from $O(N^3)$ to effectively $O(N^2)$ in the pre-processing stage, and from $O(N^2M)$ to effectively $O(NM)$ in the simulation stage.

TABLE 1 Computational complexity of our proposed framework in comparison with the traditional method for regional scenario-based seismic risk model. N denotes the number of buildings, and M denotes the number of MC simulations

Traditional		**Proposed**	
Pre-processing		**Pre-processing**	
Calc. $\mathbf{C}$	$O(N^2)$	Calc. $\mathbf{C}$	$O(N^2)$
Calc. $\mathbf{L}$	$O(N^3)$	Construct $\mathbf{Cov}(\mathbf{g}_k)$	$O(N^2)$
		Construct $\mathbf{W}_t^*$	$O(tN^2)$
Simulation		**Simulation**	
Calc. $\boldsymbol{\mu}$	$O(N)$	Calc. $\boldsymbol{\mu}$	$O(N)$
Sampling ζ, $\boldsymbol{\xi}$, u	$O(NM)$	Sampling $\mathbf{x}$, $\mathbf{z}$	$O((N+t)M)$
GM simulation (Equation 2)	$O(N^2M)$	$\mathbf{g}_k$ simulation	$O(tNM)$
DS simulation (Equation 5)	$O(NM)$		

For a detailed implementation of the algorithm and an illustrative case study, see Section S6 of the Supplementary Material. A more comprehensive discussion of the computational complexity is also provided in Section S7 of the Supplementary Material. In the subsequent sections, we provide a numerical demonstration of the proposed framework for a building portfolio in the San Francisco Bay Area.

6 | TESTBEDS: FROM SAN FRANCISCO DOWNTOWN TO BAY AREA

We conduct empirical analyses on our proposed framework using San Francisco (SF) Downtown and Bay Area building portfolios. In Example 1, we demonstrate the accuracy of the framework using 1,000 buildings in SF downtown. In Example 2, we examine t, the building cluster parameter, to keep high accuracy for different region extents and building portfolio sizes, using cases for SF downtown and the whole Bay Area comprising 15,836 buildings (Figure 9). In Example 3, we study our framework's computational efficiency in the northern SF peninsula region (Figure 11(a)).

The earthquake rupture scenario for all the examples is a magnitude 7.2 event along the northern segment of the vertically dipping San Andreas Fault (Figure 9 (b)). The rupture is assumed to span 50 km in length and 10 km in width, extending from the west coast of San Francisco to the southern Peninsula. The mean, between-event, and within-event standard deviations are computed using the ground-motion model developed for shallow crustal earthquakes in California[49]. The within-event spatial correlation is computed using the model developed based on extensive crustal earthquake records[15]. The building inventory was

obtained from the SimCenter Earthquake Testbed[50], and the structural-type–dependent fragility functions were sourced from Hazus 6.1[21]. These are lognormal fragility functions comprising four limit states and five damage states (No Damage, Slight, Moderate, Extensive, and Complete) that take Peak ground acceleration (PGA) as input.

6.1 | Example 1: Accuracy for San Francisco Downtown

We compared the results of our framework against a benchmark generated using the R2D software[51,52]. The benchmark utilizes the building portfolio from R2D Example 1 – Basic HAZUS, which performs analysis via HAZUS loss assessment (the traditional framework). From the dataset containing 15,836 buildings in downtown San Francisco, a subset of 1,000 buildings was selected to avoid excessive computational overhead when running the R2D software on a desktop-level machine (Figure 8(a)). This subset captures the full spectrum of characteristics found in the original dataset, including spatial locations, design levels, and structural types. The selected portfolio consists of 544 Pre-Code, 35 Low-Code, 322 Moderate-Code, and 99 High-Code buildings. In terms of structure, the set includes 267 wood-frame, 217 steel, 321 concrete, 194 masonry, and one unreinforced masonry building. Damage for each building was simulated over 10,000 MC realizations to ensure accuracy.

We executed both the traditional and the proposed approaches, demonstrating that $t = 1$ is sufficient to identify the damage-state distribution with high accuracy. Figure 8(b) illustrates a representative damage-state histogram for Building X (identified in Figure 8(a)). While the R2D simulation identifies the modal damage as "Extensive" and the proposed method yields "Complete," the overall distribution are nearly equal. The modal damage states for all 1,000 buildings were compared with the R2D simulation

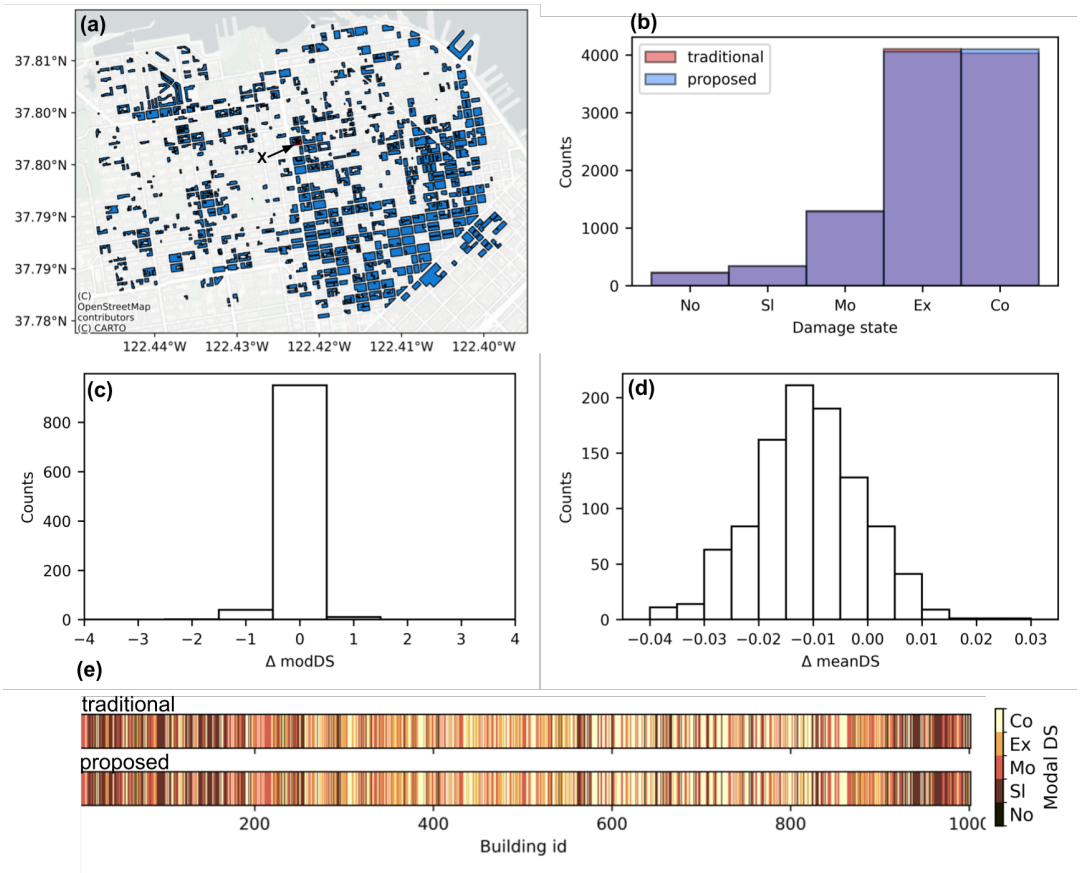

FIGURE 8 (a) Spatial distribution of the building portfolio for Example 1. (b) Comparative damage-state histogram for Building X from panel (a), illustrating a case with a one-step discrepancy in the modal damage state between the two methods. Histogram of (c) modal and (d) mean damage state differences between the R2D baseline and the proposed framework. (e) Comparison of modal damage states simulated using the R2D software (traditional) and the proposed framework ($t = 1$) based on 10,000 MC simulations.

results in Figure 8(c) and (e). The two methods show strong agreement across whole buildings: 952 buildings match exactly, while the remaining 48 buildings exhibit a difference of only one damage-state level. For the buildings exhibiting a one-level difference, their distributions still closely resemble each other, as shown in Figure 8(b). The difference in the mean damage state across all 1,000 buildings—which reflects the overall damage distribution more accurately than the mode—ranges from -0.036 to 0.039 (Figure 8(d)). This further confirms that the proposed framework closely approximates the traditional distribution and the discrepancies between the two methods may predominantly originate from the inherent stochasticity of MC simulations rather than inaccuracies in our framework.

6.2 | Example 2: Loss Estimation by Spatial Extent

The proposed framework is applied to loss estimation for two building portfolios with distinct spatial extents to investigate how the building cluster parameter, t, scales with geographic spread: 1) San Francisco Downtown, comprising 15,836 buildings (representing a dense urban area); and 2) the entire Bay Area, comprising 15,836 buildings randomly sampled from a total inventory of approximately 1.3 million structures (representing a metropolitan-scale simulation; Figure 9). By limiting the sample size, the computation of an accurate benchmark loss curve via the traditional framework—utilizing 15,836 distinct building-specific ground motions—remains feasible on a standard workstation without compromising accuracy, e.g. though coarse grid[25,17,53].

The loss curves for the SF Downtown and Bay Area cases are presented in Figures 10(a) and (b), respectively. For both cases, the traditional and proposed framework curves are generated using 100,000 MC realizations. This sample size results in a coefficient of variation (COV) for the MC loss estimation of approximately 1%. Consequently, if the loss curve produced by the proposed framework remains within $\sim 2.5\%$ error margin—corresponding to an approximate 99% confidence interval—the framework is considered to represent the benchmark accurately, as the deviation falls within the range of MC estimation error.

The error is calculated as:

$$e_t(p) = \frac{l_t(p) - l(p)}{l(p)} \times 100 \tag{36}$$

where $l(p)$ and $l_t(p)$ are the losses corresponding to exceedance probability p from the traditional and proposed frameworks, respectively.

In the San Francisco Downtown case (Figure 10(a)), the proposed framework yields a close approximation of the benchmark when $t = 1$. Within the exceedance probability range down to 10^{-3}, the error fluctuates between -0.0% and 1.4% (blue solid

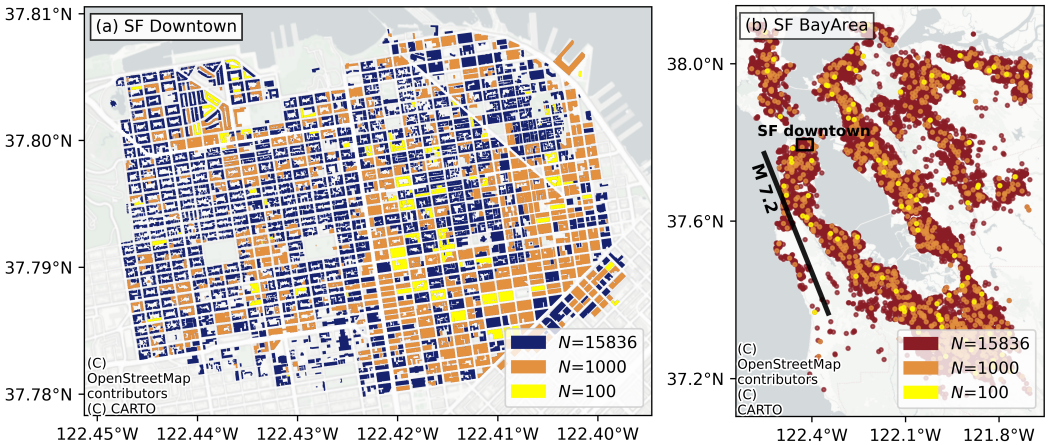

FIGURE 9 Spatial distribution of building portfolios for Example 2: (a) Downtown San Francisco and (b) the broader Bay Area. An assumed M 7.2 rupture is indicated by a thick solid black line, and SF downtown area in panel (a) is denoted by a small black square. The combined yellow, orange, and blue (or red) markers represent the full portfolio of 15,836 buildings; the orange markers denote a selected subset of 1,000 buildings; and the yellow markers specifically highlight a further subset of 100 selected buildings.

curve, Figure 10(c)). Notably, increasing t beyond 1 does not yield a significant improvement in model fit, as evidenced by the blue solid and dashed curves in Figure 10(c).

Conversely, the Bay Area case exhibits larger discrepancies at $t = 1$, with errors ranging from 2.3% to 9.9% (lightest red dotted curve of Figure 10 (d)). The error decreases as t increases; specifically, at $t = 20$, the error falls within the target 2.5% threshold in general, ranging from 1.2% to 2.9% (red solid curve). Further increases in t lead to asymptotic convergence toward zero, for example for $t = 200$, the error ranges from 0.1% to 1.5%. However, as t increases, the rate of improvement diminishes (Figure 10(d)).

While numerical simulations demonstrate that the proposed framework significantly reduces dimensionality, selecting an optimal value of t is key. We found that the optimal t is invariance regardless of the total building count, provided the spatial extent of the portfolio remains consistent (Section S8 of the Supplementary Material). Thus, we suggest determining t from small, randomly selected subset of buildings that spans the entire region: 1) Conduct a traditional regional risk analysis using a small, randomly selected subset of buildings that spans the entire region. This is computationally inexpensive due to the small N; 2) Perform the risk analysis with the proposed framework on the same subset for a sufficiently large t; 3) Determine an optimal t for the region based on the agreement between the resulting loss curves; 4) Apply the determined t to the full-scale analysis of the entire building portfolio.

In addition, we observed that the eigenvectors of $\mathbf{Cov}(\boldsymbol{\omega}^*)$ emerge as spatial harmonics (Figure S6). As the eigenvalues decrease, the spatial frequency of these modes increases, with the order of their magnitudes being directly related to the spatial extent of the building distribution, higher order in the direction. The dominant component corresponds to the direction along which the building distribution is most elongated. Furthermore, the product of each eigenvalue and its corresponding eigenvector

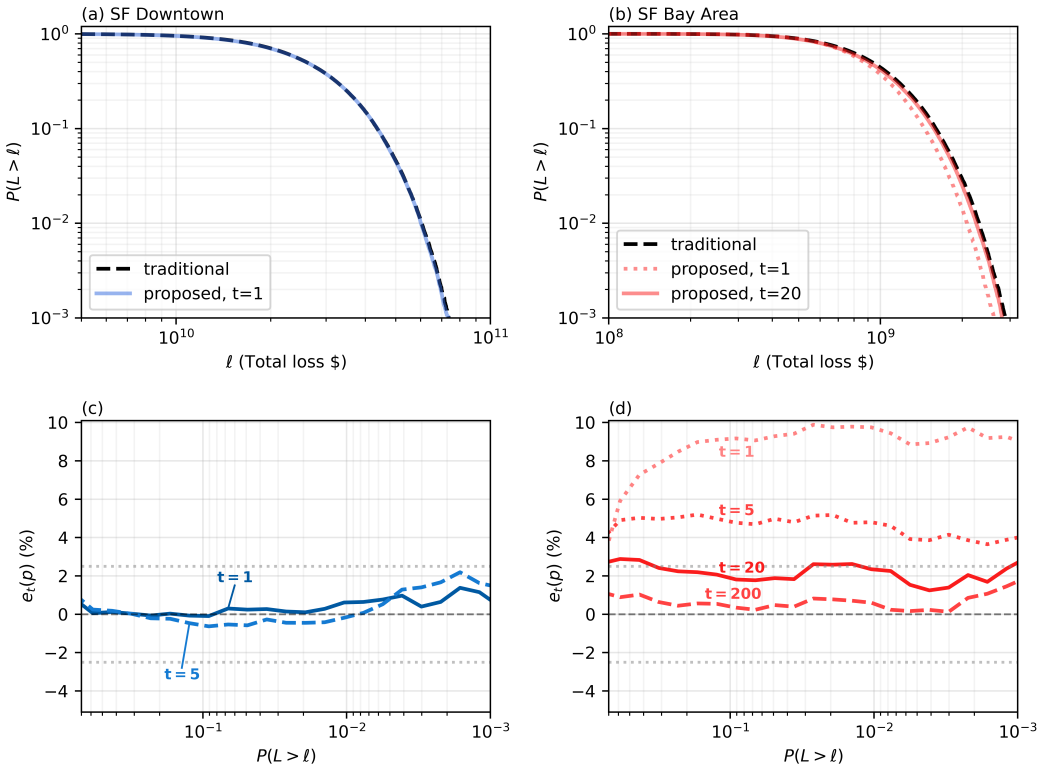

FIGURE 10 (a) Exceedance probability of total loss for the SF Downtown portfolio; the black dashed line represents the traditional framework benchmark, while the blue solid line represents the proposed framework with $t = 1$. (b) Equivalent plot for the Bay Area; the benchmark (black dashed) is compared against the proposed framework with $t = 20$ (red solid line) and $t = 1$ (red dotted line). (c) Relative error of the proposed framework across various exceedance probabilities for SF Downtown ($t = 1$ and $t = 5$, indicated by blue solid and dashed lines, respectively) and (d) the Bay Area ($t = 1$ and 5 indicated by red dotted lines, $t = 20$ by red solid line, and $t = 200$ by red dashed line).

illustrates how the magnitude of each eigenvector component diminishes as one moves toward the minor components. For a more detailed analysis of these spatial modes, please refer to Section S9 of Supplementary Material.

Finally, we note that the relationship between t and spatial extent is conceptually analogous to the selection of a ground-motion grid in traditional risk assessment. Just as a traditional approach requires a large number of ground motion grids to resolve ground motion across larger spatial extents, our framework necessitates a larger t as well. However, a fundamental distinction exists: while conventional gridding suppresses small-scale ground-shaking anomalies by assuming uniform intensities within each cell, our framework identifies a latent space—defined by eigenvectors—that captures the majority of the correlated variance without requiring spatial discretization. This approach facilitates the representation of macro-scale variations without the inherent loss of resolution typical of spatial gridding.

6.3 | Example 3: Gains in Computational Efficiency

To compare computational efficiency between the traditional framework and ours, we selected the Northern San Francisco Peninsula, which includes more buildings than the downtown area (Figure 11(a)). Following the procedure described above, we measured runtime for both frameworks by increasing the number of randomly selected buildings from 5,000 to 30,000. The analysis was performed in Python 3.11 using NumPy vectorization and Numba for numerical computation[54,55]. All simulations were executed on an Intel Core i7-13700 2.1 GHz processor with 64 GB RAM, without parallelization. We distinguish between pre-processing and simulation runtimes, since simulation scales with both the number of buildings (N) and MC realizations (M), while pre-processing only with the number of buildings (Table 1).

In pre-processing, the proposed framework is faster than the traditional framework by up to $\times$ 3.4 depending on the number of buildings (squares in Figure 11(b)). At $N = 5,000$, runtimes are nearly equal, but as N increases, the runtime ratio reaches 3.4 at $N = 30,000$ (Figure 11(b)). This growth in efficiency is due to the traditional framework's pre-processing scales approximately as $N^{2.5}$, while the proposed framework scales as N^2 (Figures 11(c)). Although Cholesky decomposition theoretically scales as $O(N^3)$ and should dominate for large N (Table 1), this dominance is not clearly observed for $N \in [5,000, 30,000]$. Instead, the overall runtime lies between the $O(N^2)$ cost of constructing $\mathbf{C}$ and the $O(N^3)$ cost of Cholesky, i.e., roughly $O(N^{2.5})$. This suggests a larger constant factor for correlation matrix construction, delaying cubic scaling. For larger N, we expect $O(N^3)$ behavior to dominate, further widening the gap.

In the simulation step, the performance difference is even more pronounced (circles in Figure 11(b)). As expected (Table 1), the traditional and proposed methods scale as $O(N^2)$ and $O(N)$, respectively (Figures 11(d)). For 30,000 buildings, the proposed framework is approximately 110 times faster than the traditional method. With more buildings, the runtime ratio will be even larger and is theoretically expected to reach $\times$3500 for an N of one million.

7 | SUMMARY

We developed a scalable computational framework for regional scenario-based seismic risk modeling based on an exact linearized reformulation of the mathematical problem and dimensionality reduction via PPCA. First, we demonstrate that canonical coupling between ground motions and lognormal fragility functions can be reformulated as a linear problem. This transformation reveals that the traditional two-step process to simulate ground-motion–fragility coupling, which is $2N + 1$ dimensional problem for ground shaking simulation and corresponding damage modeling for N building simulation, can be simplified into a one-shot N-dimensional correlated Gaussian process, each dimension representing each building's damage function.

The theoretical and empirical covariance analysis of the model shows that 1) its smallest eigenvalues converge to a non-zero constant as a closed form solution as a function of the maximum within-event ground motion correlation ($\rho_{\max}$), within-event ground motion standard deviation (ϕ), and the dispersion parameter of the lognormal fragility curve (β) and 2) the number of eigenvalues greater than this lower bound is representing the number of spatial building clusters in the portfolio (t). In addition, it is also found that the further dimensionality reduction of the model using standard PCA is not efficient because the smallest eigenvalues are not zero.

Alternatively, we applied PPCA. We first isolate the strongly correlated structure of the model's covariance by extracting the independent noise component. Using both theoretical and empirical eigenvalue lower bounds, we identified the optimal decomposition, which facilitates easy decomposition and enhances computational efficiency by reducing the dimensionality of the correlated structure to the number of building clusters, t. This value t is significantly smaller than the total number of

buildings in the portfolio (N). This formulation reduces the computational complexity of the risk modeling from cubic complexity (e.g., $O(N^3)$ or $O(N^2M)$) to quadratic complexity ($O(N^2)$ or $O(NM)$), where M represents the number of MC simulations.

Verification using the R2D example—an established software for regional seismic risk modeling—demonstrates that the resulting damage-state distributions are nearly identical to those obtained from the traditional framework. The mean damage state differs by at most ~ 0.04. We also found that for the San Francisco downtown portfolio containing 15,836 buildings, a value of $t = 1$ suffices for accurate modeling. In contrast, for the entire Bay Area with the same number of buildings, $t = 20$ is required, yielding a loss estimation error within approximately 2.5% of the benchmark. Importantly, in this realistic portfolio, the building cluster parameter t depends primarily on the spatial extent of the portfolio rather than the building density. This suggests that the building cluster parameter t can be determined through computationally efficient small-sample simulations, e.g., using 100 subsamples for a full inventory of $\sim$ 15,000 buildings. Additionally, we found that the eigenvectors exhibit spatial harmonics, where the order of these harmonics is closely related to the spatial elongation of the portfolio. The computational gain of the proposed framework was also experimentally verified; for a 30,000-building portfolio of the San Francisco Northern Peninsula, this translates to 3.4$\times$ faster pre-processing and 110$\times$ faster simulation. This advantage is expected to grow with the portfolio size N, given the lower asymptotic complexity in both steps.

Overall, the proposed framework enables large-scale regional seismic risk analysis that remains computationally tractable without compromising accuracy. It substantially reduces computation time while mitigating the uncertainties associated with ground-shaking gridding and exposure resolution in regional seismic risk assessments. This improvement is particularly significant for metropolitan areas characterized by densely distributed building portfolios.

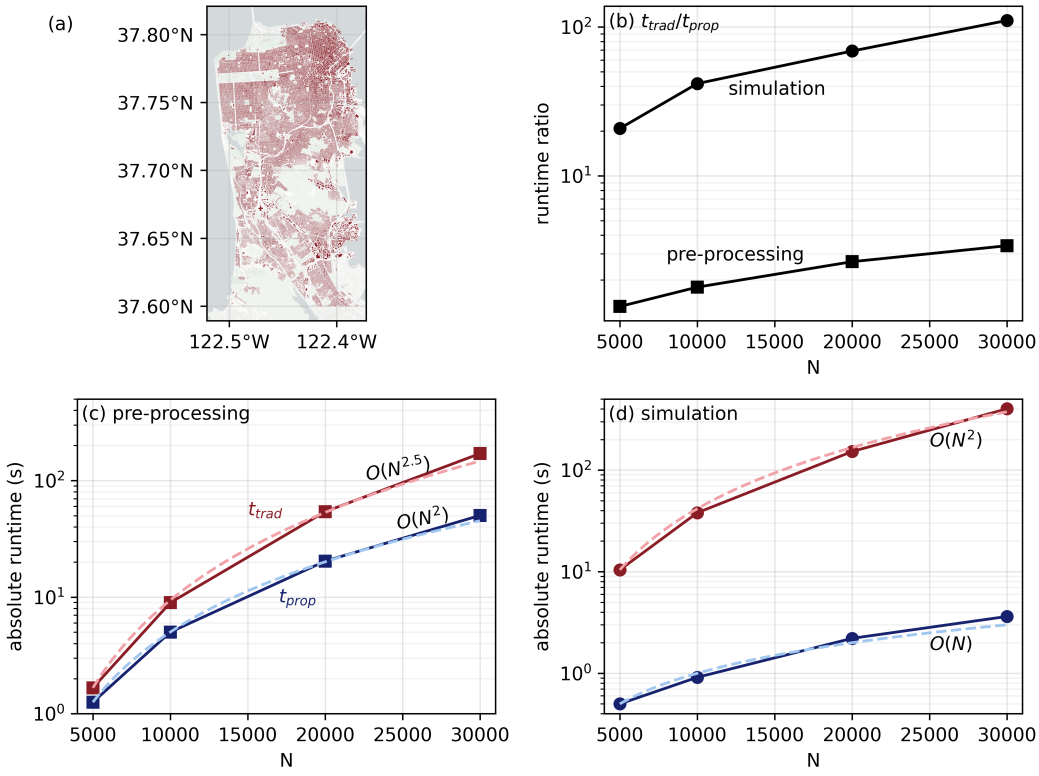

FIGURE 11 (a) San Francisco Northern Peninsula Building portfolio ($N = 30,000$) used to measure the computation time of the proposed framework. (b) Runtime ratio between the traditional and proposed methods for pre-processing (squares) and simulation (circles). t_{trad} denotes the computation time of the traditional framework, while t_{prop} denotes that of the proposed framework. The MC simulation uses $M = 10^4$. (c, d) Absolute computation times as a function of the number of buildings in (c) pre-processing and (d) the simulation steps for traditional (red) and proposed (blue) methods.

ACKNOWLEDGEMENTS

The authors acknowledge the financial support provided by Department of Civil and Environmental Engineering at the University of California, Berkeley.

DATA AVAILABILITY STATEMENT

The source code for computing regional scenario-based damage modeling using the proposed framework explained in this paper is available at https://github.com/sehoung/Linearizaiton_PPCA

References

1. Ceferino L, Merino Y, Pizarro S, Moya L, Ozturk B. Placing engineering in the earthquake response and the survival chain. *Nature Communications.* 2024;15(1):4298.

2. Ceferino L, Kukunoor C, Zhao J, et al. Accessing acute care hospitals in the San Francisco Bay Area after a major hayward earthquake. *Nature Communications.* 2025;16(1):9328.

3. Kiremidjian A, Moore J, Fan YY, Yazlali O, Basoz N, Williams M. Seismic risk assessment of transportation network systems. *Journal of Earthquake Engineering.* 2007;11(3):371–382.

4. Silva-Lopez R, Baker JW. Optimal Bridge retrofitting selection for seismic risk management using genetic algorithms and neural Network–Based surrogate models. *Journal of Infrastructure Systems.* 2023;29(4):04023030.

5. Chen R, Jaiswal KS, Bausch D, Seligson H, Wills C. Annualized earthquake loss estimates for California and their sensitivity to site amplification. *Seismological Research Letters.* 2016;87(6):1363–1372.

6. Ceferino L, Kiremidjian A, Deierlein G. Regional multiseverity casualty estimation due to building damage following a Mw 8.8 earthquake scenario in Lima, Peru. *Earthquake Spectra.* 2018;34(4):1739–1761.

7. Ceferino L, Mitrani-Reiser J, Kiremidjian A, Deierlein G, Bambarén C. Effective plans for hospital system response to earthquake emergencies. *Nature communications.* 2020;11(1):4325.

8. Liu Y, Wotherspoon L, Nair NKC, Blake D. Quantifying the seismic risk for electric power distribution systems. *Structure and Infrastructure Engineering.* 2021;17(2):217–232.

9. Yoshikawa H, Goda K. Financial seismic risk analysis of building portfolios. *Natural Hazards Review.* 2014;15(2):112–120.

10. Goda K, Hong H. Estimation of seismic loss for spatially distributed buildings. *Earthquake Spectra.* 2008;24(4):889–910.

11. Crowley H, Bommer JJ. Modelling seismic hazard in earthquake loss models with spatially distributed exposure. *Bulletin of Earthquake Engineering.* 2006;4:249–273.

12. Moehle J, Deierlein GG. A framework methodology for performance-based earthquake engineering. In: . 679. WCEE Vancouver. 2004:12.

13. Goda K, Hong HP. Spatial correlation of peak ground motions and response spectra. *Bulletin of the Seismological Society of America.* 2008;98(1):354–365.

14. Jayaram N, Baker JW. Correlation model for spatially distributed ground-motion intensities. *Earthquake Engineering & Structural Dynamics.* 2009;38(15):1687–1708.

15. Loth C, Baker JW. A spatial cross-correlation model of spectral accelerations at multiple periods. *Earthquake Engineering & Structural Dynamics.* 2013;42(3):397–417. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/eqe.2212doi: 10.1002/eqe.2212

16. Weatherill G, Silva V, Crowley H, Bazzurro P. Exploring the impact of spatial correlations and uncertainties for portfolio analysis in probabilistic seismic loss estimation. *Bulletin of Earthquake Engineering.* 2015;13(4):957–981.

17. Dabbeek J, Crowley H, Silva V, Weatherill G, Paul N, Nievas CI. Impact of exposure spatial resolution on seismic loss estimates in regional portfolios. *Bulletin of Earthquake Engineering.* 2021;19(14):5819–5841.

18. Bhattacharya A, Chakraborty A, Mallick BK. Fast sampling with Gaussian scale mixture priors in high-dimensional regression. *Biometrika.* 2016:asw042.

19. Trefethen LN, Bau D. *Numerical linear algebra*. SIAM, 2022.

20. Li P, Wang Z, Zhao B, Becker T, Soga K. Surrogate modeling for identifying critical bridges in traffic networks under earthquake conditions. *Transportation Research Part D: Transport and Environment.* 2025;138:104512.

21. Federal Emergency Management Agency (FEMA) . *HAZUS Earthquake Model Technical Manual: HAZUS 6.1*. FEMA; : 2024. Version 6.1.

22. Handy SL, Niemeier DA. Measuring accessibility: an exploration of issues and alternatives. *Environment and planning A.* 1997;29(7):1175–1194.

23. Bazzurro P, Park J. The effects of portfolio manipulation on earthquake portfolio loss estimates. In: . 31. 2007:8.

24. Scheingraber C, Käser MA. The impact of portfolio location uncertainty on probabilistic seismic risk analysis. *Risk Analysis.* 2019;39(3):695–712.

25. Bal IE, Bommer JJ, Stafford PJ, Crowley H, Pinho R. The influence of geographical resolution of urban exposure data in an earthquake loss model for Istanbul. *Earthquake Spectra.* 2010;26(3):619–634.

26. Markhvida M, Ceferino L, Baker JW. Modeling spatially correlated spectral accelerations at multiple periods using principal component analysis and geostatistics. *Earthquake Engineering & Structural Dynamics.* 2018;47(5):1107–1123.

27. Du W, Ning CL. Modeling spatial cross-correlation of multiple ground motion intensity measures (SAs, PGA, PGV, Ia, CAV, and significant durations) based on principal component and geostatistical analyses. *Earthquake Spectra.* 2021;37(1):486–504.

28. Kiremidjian AS, Stergiou E, Lee R. Issues in seismic risk assessment of transportation networks. *Geotechnical, Geological and Earthquake Engineering.* 2007;6:461–480. doi: 10.1007/978-1-4020-5893-6_19

29. Jayaram N, Baker JW. Efficient sampling and data reduction techniques for probabilistic seismic lifeline risk assessment. *Earthquake Engineering & Structural Dynamics.* 2010;39(10):1109–1131.

30. Houng SE, Ceferino L. Fast probabilistic seismic hazard analysis through adaptive importance sampling. *Bulletin of the Seismological Society of America.* 2025;115(2):646–663.

31. Houng SE, Ceferino L, Abrahamson N. Fast Propagation of Epistemic Uncertainty in Seismic Hazard via Adaptive Importance Sampling.

32. Wang Z, Song J. Cross-entropy-based adaptive importance sampling using von Mises-Fisher mixture for high dimensional reliability analysis. *Structural Safety.* 2016;59:42–52.

33. Katafygiotis LS, Zuev KM. Geometric insight into the challenges of solving high-dimensional reliability problems. *Probabilistic Engineering Mechanics.* 2008;23(2-3):208–218.

34. Han Y, Davidson RA. Probabilistic seismic hazard analysis for spatially distributed infrastructure. *Earthquake Engineering & Structural Dynamics.* 2012;41(15):2141–2158.

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

36. Manzour H, Davidson RA, Horspool N, Nozick LK. Seismic hazard and loss analysis for spatially distributed infrastructure in Christchurch, New Zealand. *Earthquake Spectra.* 2016;32(2):697–712.

37. Christou V, Bocchini P, Miranda MJ, Karamlou A. Effective sampling of spatially correlated intensity maps using hazard quantization: Application to seismic events. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering.* 2018;4(1):04017035.

38. Tipping ME, Bishop CM. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 1999;61(3):611–622.

39. Bozorgnia Y, Abrahamson NA, Atik LA, et al. NGA-West2 research project. *Earthquake Spectra.* 2014;30(3):973–987.

40. Abrahamson NA, Silva WJ, Kamai R. Summary of the ASK14 ground motion relation for active crustal regions. *Earthquake Spectra.* 2014;30(3):1025–1055.

41. Golub GH, Van Loan CF. *Matrix computations*. JHU press, 2013.

42. Boore DM, Stewart JP, Seyhan E, Atkinson GM. NGA-West2 equations for predicting PGA, PGV, and 5% damped PSA for shallow crustal earthquakes. *Earthquake Spectra.* 2014;30(3):1057–1085.

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

44. Chiou BSJ, Youngs RR. Update of the Chiou and Youngs NGA model for the average horizontal component of peak ground motion and response spectra. *Earthquake Spectra.* 2014;30(3):1117–1153.

45. Alexander NA, Ibraim E, Aldaikh H. A simple discrete model for interaction of adjacent buildings during earthquakes. *Computers & Structures.* 2013;124:1–10.

46. Luco JE, Contesse L. Dynamic structure-soil-structure interaction. *Bulletin of the Seismological Society of America.* 1973;63(4):1289–1303.

47. Abdi H, Williams LJ. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics.* 2010;2(4):433–459.

48. Dongarra JJ, Moler CB, Bunch JR, Stewart GW. *LINPACK users' guide*. SIAM, 1979.

49. Sadigh K, Chang CY, Egan J, Makdisi F, Youngs RR. Attenuation relationships for shallow crustal earthquakes based on California strong motion data. *Seismological research letters.* 1997;68(1):180–189.

50. Zsarnóczay A, Elhaddad W, Cetiner B, Zhong K, McKenna F, Deierlein G. SimCenter Earthquake Testbed: San Francisco, CA.; 2023

51. Deierlein GG, McKenna F, Zsarnóczay A, et al. A Cloud-Enabled Application Framework for Simulating Regional-Scale Impacts of Natural Hazards on the Built Environment. *Frontiers in Built Environment.* 2020;6. Publisher: Frontiersdoi: 10.3389/fbuil.2020.558706

52. McKenna F, Gavrilovic S, Zhao J, et al. NHERI-SimCenter/R2DTool: Version 5.2.0. https://zenodo.org/records/14219019; 2024

53. Goda K, Zhang L, Tesfamariam S. Portfolio Seismic Loss Estimation and Risk-based Critical Scenarios for Residential Wooden Houses in Victoria, British Columbia, and Canada. *Risk analysis.* 2021;41(6):1019–1037.

54. Lam SK, Pitrou A, Seibert S. Numba: a LLVM-based Python JIT compiler. In: LLVM '15. Association for Computing Machinery 2015; New York, NY, USA:1–6

55. Harris CR, Millman KJ, Van Der Walt SJ, et al. Array programming with NumPy. *nature.* 2020;585(7825):357–362.

56. Marsaglia G, Tsang WW. The ziggurat method for generating random variables. *Journal of statistical software.* 2000;5:1–7.

57. Joyner WB, Boore DM. Peak horizontal acceleration and velocity from strong-motion records including records from the 1979 Imperial Valley, California, earthquake. *Bulletin of the seismological Society of America.* 1981;71(6):2011–2038.

58. Horn RA, Johnson CR. *Matrix Analysis*. Cambridge: Cambridge University Press. 2nd ed., 2013.

59. Sadigh K, Chang CY, Egan J, Makdisi F, Youngs R. Attenuation Relationships for Shallow Crustal Earthquakes Based on California Strong Motion Data. *Seismological Research Letters.* 1997;68(1):180–189. doi: 10.1785/gssrl.68.1.180

60. Park J, Bazzurro P, Baker JW, others . Modeling spatial correlation of ground motion intensity measures for regional seismic hazard and portfolio loss estimation. *Applications of statistics and probability in civil engineering.* 2007;2:1–8.

61. DeBock DJ, Liel AB. A comparative evaluation of probabilistic regional seismic loss assessment methods using scenario case studies. *Journal of Earthquake Engineering.* 2015;19(6):905–937.

62. Antonietti PF, Mazzieri I, Melas L, et al. Three-dimensional physics-based earthquake ground motion simulations for seismic risk assessment in densely populated urban areas. *Mathematics in Engineering.* 2021;3(2):1–31.

63. Graves R, Jordan TH, Callaghan S, et al. CyberShake: A physics-based seismic hazard model for southern California. *Pure and Applied Geophysics.* 2011;168(3):367–381.

64. Smerzini C, Pitilakis K. Seismic risk assessment at urban scale from 3D physics-based numerical modeling: the case of Thessaloniki. *Bulletin of Earthquake Engineering.* 2018;16(7):2609–2631.

65. Boore DM. Simulation of ground motion using the stochastic method. *Pure and applied geophysics.* 2003;160(3):635–676.

66. Irikura K. Prediction of strong acceleration motion using empirical Green's function. In: . 151. 1986:151–156.

67. Lee R, Kiremidjian AS. Uncertainty and correlation for loss assessment of spatially distributed systems. *Earthquake Spectra.* 2007;23(4):753–770.

68. Sokolov V, Wenzel F. Influence of spatial correlation of strong ground motion on uncertainty in earthquake loss estimation. *Earthquake Engineering & Structural Dynamics.* 2011;40(9):993–1009.

69. Heresi P, Miranda E. Structure-to-structure damage correlation for scenario-based regional seismic risk assessment. *Structural Safety.* 2022;95:102155.

70. Mueller CS. Source pulse enhancement by deconvolution of an empirical Green's function. *Geophysical Research Letters.* 1985;12(1):33–36.

71. Field EH, Biasi GP, Bird P, et al. Long-term time-dependent probabilities for the third Uniform California Earthquake Rupture Forecast (UCERF3). *Bulletin of the Seismological Society of America.* 2015;105(2A):511–543.

72. Horn RA, Johnson CR. *Matrix analysis*. Cambridge university press, 2012.