Received: 1 March 2022 | Revised: 18 July 2022 | Accepted: 11 September 2022

DOI: 10.1002/eqe.3743

RESEARCH ARTICLE

WILEY

Selection of representative natural hazard scenarios for engineering systems

Hugo Rosero-Velásquez | **Daniel Straub**

Engineering Risk Analysis Group, Technical University of Munich, Munich, Germany

Correspondence
Hugo Rosero-Velásquez, Engineering Risk Analysis Group, Technical University of Munich, 80333 Munich, Germany.
Email: hugo.rosero@tum.de

Funding information
Federal Ministry of Education and Research; RIESGOS, Grant/Award Number: 03G0876A-J; RIESGOS 2.0, Grant/Award Number: 03G0905A-H

Abstract

Representative hazard scenarios are essential for many tasks in risk management, such as preparedness and emergency response planning. However, criteria and methods for systematically selecting such scenarios for natural hazards are lacking. From a risk perspective, such scenarios should be selected considering the losses they incur. Hence, we propose to define a scenario that is representative for a certain degree of loss, for example, the 100-year loss, as the most likely one among all possible scenarios leading to this loss. Taking basis in a generic model of natural hazards and their impact on engineering systems, we formally introduce the representative scenarios. We then develop algorithms that enable an efficient evaluation of these scenarios. The method and algorithms are demonstrated on a hypothetical example considering a spatially distributed infrastructure system subjected to earthquakes.

KEYWORDS
engineering system, loss, natural hazard, representative scenario, return period, risk analysis

1 | INTRODUCTION

Natural hazards pose significant risks to engineering systems in general and infrastructure systems in particular. Past infrastructure failures caused by storms, floods and seismic events, or combinations thereof, attribute to this fact, for example during hurricane Kathrina in 2005[1,2] or the 2010 Chile earthquake and tsunami.[3–5]

To limit the impact of these failures, authorities and utility operators aim at an effective risk management.[6] In this context, risk managers commonly work with representative scenarios to assess the risk[7–9] and resilience,[10] to test and validate risk management procedures,[11–13] and for effective risk communication.[6,14] Such scenarios are usually selected based on expert knowledge[15] and past events.[16]

Surprisingly, there has been little research on how to systematically identify representative scenarios for risk management in general, and for natural hazard events in particular. Miller and Baker[8] propose to identify a limited set of hazard and damage scenarios that best approximate the loss exceedance curve, with application to seismic hazards. Romero et al.[17] develop an optimization approach for selecting seismic hazard and damage scenarios. Seismic hazards scenarios are selected as those that best approximate the annual exceedance probability (AEP) curve of the PGA, based on Vaziri et al.[18] Damage scenarios are selected as those that best approximate the component vulnerabilities, following Brown et al.[19] All these approaches aim at an accurate risk estimation with only a limited set of (computationally expensive) evaluations of the full hazard and damage models.

Salgado-Gálvez et al.,[9] in the context of probabilistic seismic hazard analysis (PSHA), select a single scenario 'based on choosing the loss corresponding to a return period of 1000 years from the fully probabilistic analysis as the target loss, and then, a single event from the complete stochastic set which caused a similar value of loss'.

Berk et al.[20] propose a method to identify representative rainfall events for flood analysis. Their approach is based on inverse FORM (first-order reliability method),[21] which identifies the hazard event that is the most likely one among all possible events leading to a demand of a certain return period.

In this contribution, we propose a risk-oriented definition of representative hazard scenarios, which takes up ideas from these previous approaches. In line with Miller and Baker[8] and Salgado-Gálvez et al.,[9] we identify hazard scenarios that are associated with a given loss return period; for example, the seismic event that is representative for the loss with a 100yr return period. In contrast to these approaches, we define the representative scenario as the most likely one to lead to such a loss. This definition is inspired by the inverse FORM approach, which, however, does not consider losses but failure events and assumes that the system performance is deterministic for given parameters. Our proposed definition addresses the uncertainty in the system response for given hazard parameters. This leads to additional computational challenges in evaluating these scenarios. Therefore, we also develop efficient algorithms to estimate the scenarios.

To enable a systematic and generic definition of representative hazard scenarios, Section 2 presents a general framework for risk assessment of (spatially distributed) engineering systems subject to natural hazards, which is inspired by the PEER framework.[22,23] On this basis, we introduce and discuss the definition of the representative hazard scenarios in Section 3. In Section 4, we propose a workflow for identifying such scenarios based on a combination of surrogate models with active learning (AL) techniques. In Section 2.5, we investigate and demonstrate the methodology on an idealized example of a power network subject to earthquakes. We end this contribution in Section 6 with a discussion on the limitations and possible extensions of the proposed definitions and methods.

2 | PROBABILISTIC HAZARD AND RISK ANALYSIS

2.1 | General framework

Risk analysis of infrastructure systems subject to natural hazards requires the combination of models from multiple domains, including models of hazard occurrence, hazard propagation, response of system components, overall system performance and losses. In order to identify hazard scenarios that are representative for losses with a certain return period, it is necessary to combine the different models. In this Section 2, we present a framework for combining these models as a basis for identifying representative hazard scenarios.

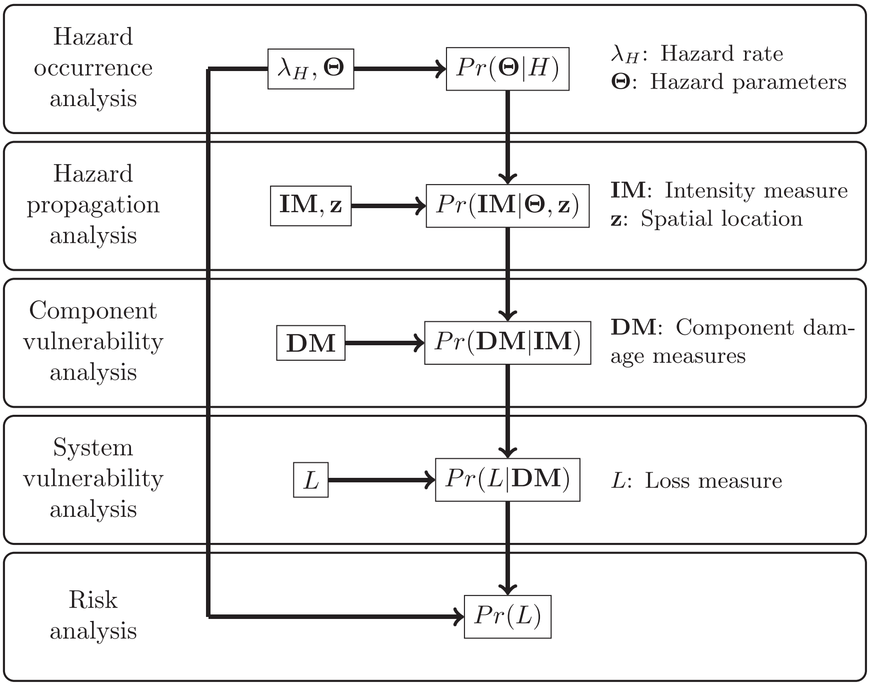
In the context of earthquake engineering, PSHA is a methodology for identifying ground motions with a specified exceedance probability. First formulated by Cornell,[24] it is a widely used framework for hazard analysis of engineering structures under seismic hazards.[25–28] On this basis, the Pacific Earthquake Engineering Research Center (PEER) formulated a performance- or risk-based framework for seismic hazard, which systematically integrates the different model components to assess the seismic risk.[22,23,28,29]

For other types of hazards, risk analysis is also based on a similar combination of models. Examples found in the literature include wind storms,[30,31] floods[32,33] and volcanoes.[34,35]

In Figure 1, we present a generic framework for risk analysis of infrastructure systems under natural hazards. It is inspired by the PEER framework, yet is compatible with modelling approaches used for different natural hazards. It is applicable to different types of engineering systems and infrastructure; in particular, it allows consideration of the spatially distributed nature of most infrastructure systems. The framework establishes the connection between the probabilistic model of the hazard occurrence (the hazard rate λ_H and the hazard parameters $\mathbf{\Theta}$) and the target variable (the loss measure L).

The framework consists of a hazard occurrence analysis described in Section 2.2, a hazard propagation analysis (Section 2.3), the component fragility and damage analysis (Section 2.4) and the system response and loss analysis (Section 2.5). The models resulting from these analyses are combined in the risk analysis summarized in Section 2.6. Since the interest is ultimately in the relation between the hazard parameters $\mathbf{\Theta}$ and the loss L, Section 2.7 discusses the sensitivity of the losses to the uncertainty in $\mathbf{\Theta}$.

FIGURE 1 General framework for risk assessment of spatially distributed engineering systems under natural hazards. $Pr(\cdot|\cdot)$ indicates a conditional probability.


Hazard occurrence analysis
$\lambda_H, \mathbf{\Theta}$
$Pr(\mathbf{\Theta}|H)$
λ_H: Hazard rate
$\mathbf{\Theta}$: Hazard parameters
Hazard propagation analysis
$\mathbf{IM}, \mathbf{z}$
$Pr(\mathbf{IM}|\mathbf{\Theta}, \mathbf{z})$
$\mathbf{IM}$: Intensity measure
$\mathbf{z}$: Spatial location
Component vulnerability analysis
$\mathbf{DM}$
$Pr(\mathbf{DM}|\mathbf{IM})$
$\mathbf{DM}$: Component damage measures
System vulnerability analysis
L
$Pr(L|\mathbf{DM})$
L: Loss measure
Risk analysis
$Pr(L)$

2.2 | Hazard occurrence analysis

The occurrence of natural hazard events can be described by random processes. For most hazards, the occurrence of extreme events H is well-described by a stationary Poisson process, or more generally a point process.[36–38] We denote the mean rate of occurrence of hazard events by λ_H.

Each hazard event is described by a set of parameters $\theta = [\theta_1; \theta_2; \cdots; \theta_m]$. For instance, parameters describing a seismic hazard are earthquake magnitude, location and slip type (among others); similarly, for flood hazards, example parameters are rainfall intensity and duration. The selection of these parameters is application-specific and might depend on the model and data availability. For example, for floods, the hazard parameters could also be maximum discharge instead of rainfall parameters. The hazard parameters are modelled as random variables $\mathbf{\Theta}$ with conditional joint probability density function (PDF) $f_{\mathbf{\Theta}|H}(\mathbf{\Theta})$.

In some cases, the analysis includes only a single hazard parameter Θ, for example, in a fluvial flood risk assessment, the hazard might be characterized only by the maximum discharge.[39,40] In this case, one can find the hazard parameter corresponding to a return period t by

$$\theta_t = F_\Theta^{-1}\left(\frac{1}{\lambda_H t}\right) \tag{1}$$

where F_Θ^{-1} is the inverse cumulative distribution function (CDF) of Θ. If the relation between hazard parameter θ and resulting losses L is deterministic and monotonously increasing, then θ_t is also the hazard leading to the losses with return period t. This corresponds to an AEP neutrality between the losses and the hazard parameter.[41]

In general, however, the hazard parameters form a vector and the losses L are not deterministic for given $\boldsymbol{\theta}$. As a consequence, AEP neutrality does not hold and Equation (1) is not applicable. This motivates the criterion and methods for identifying representative hazard scenarios that we propose in Section 3.

2.3 | Hazard propagation analysis

To assess the performance of the engineering system, the impact of the hazard at the locations $\mathbf{z}_i = [x_i, y_i]$ at any of the n system components must be determined, for $i = 1, \dots, n$. The impact is quantified through one or more intensity measures

IM at each of the locations $\mathbf{z}_i$. Examples of intensity measures are the PGA in seismic hazard analysis or the inundation depth in flood risk analysis.

For given hazard parameters, the intensity measure IM_i at a location $\mathbf{z}_i$ is evaluated through a model. One example of such a model is the so called ground motion prediction equation (GMPE) that predicts the PGA for a given seismic event. We represent such models by a function ξ, which can be either deterministic or stochastic, depends on additional deterministic and stochastic parameters, such as shear velocity and roughness, and maps the hazard parameters θ to the intensity measure at $\mathbf{z}_i$:

$$IM_i = \xi(\theta, \mathbf{z}_i) \tag{2}$$

The intensity measure is evaluated at all n locations $\mathsf{Z} = [\mathbf{z}_1; \cdots; \mathbf{z}_n]$, resulting in a vector $\mathbf{IM} = [IM_1; \dots; IM_n]$. In the general case, ξ is a stochastic function and $\mathbf{IM}$ is a discrete representation of a random field, defined conditional on the hazard parameters θ. A commonly used random field model for a scalar intensity measure is based on multiplying a deterministic model μ_ξ with a lognormal random field:

$$\mathbf{IM} = \xi(\theta, \mathsf{Z}) = \mu_\xi(\theta, \mathsf{Z}) \exp\left[\sigma U(\mathsf{Z})\right] \tag{3}$$

where $U(\mathsf{Z})$ is a standard normal random field, and σ controls the variance of the random field. This model is commonly used in seismic risk analysis,[42,43] wherein μ_ξ is the deterministic GMPE.

More generally, the hazard intensity at a location $\mathbf{z}_i$ can be described by a vector of intensity measures $\mathbf{IM}_i = [IM_{i1}, \dots, IM_{ir}]$. For example, both PGA and PGV might be used as intensity measures in a seismic risk analysis. In this case, applying the model ξ to all locations Z, one obtains the matrix of intensity measures IM, whose ijth entry corresponds to the jth intensity measure evaluated at the location of the ith component.

Based on a model such as given in Equation (3), the intensity measures at all n components can be summarized by a conditional joint PDF $f_{\mathbf{IM}|\boldsymbol{\Theta}}(\mathbf{im}|\theta)$.

2.4 | Component vulnerability analysis

The performance of the system components is represented by fragility models in function of the intensity measures at the component locations. Fragility functions for different hazard types and system components are described extensively in the literature.[25,27,44–48]

In agreement with the PEER framework, we use the term damage measure DM_i to describe the performance of the ith system component. In the simplest case, one distinguishes only between a functioning $DM_i = 0$ and a failed $DM_i = 1$ component. However, consideration of multiple damage states is straightforward.[27]

For a scalar intensity measure IM_i at a location $\mathbf{z}_i$, and considering only a failure state $DM_i = 1$, the fragility function is

$$Fr(im_i) = \Pr(DM_i = 1 | IM_i = im_i) \tag{4}$$

In spatially distributed systems, the damage measure is a random vector $\mathbf{DM} = [DM_1; \dots; DM_n]$. Commonly, it is assumed that the damage states are conditionally independent among the components given the vector of intensity measures $\mathbf{IM}$. Consideration of such dependence among components is, however, possible[46] and can be included in the framework.

The fragility functions for all components define a conditional joint PMF of component states given the vector of intensity measures, $p_{\mathbf{DM}|\mathbf{IM}}(\mathbf{dm}_k|\mathbf{im})$.

2.5 | System vulnerability analysis

The performance of the overall system is represented by a loss measure L. It can be measured, for example, by connectivity loss,[49–52] efficiency loss,[53–55] or a factor describing the impact on the population.[51] For power networks, it can be measured by the energy not supplied (ENS),[56] and for road networks by the travel time delay.[57] Examples of further metrics can be found in the literature.[50,58] The L can also consider the resilience of the system, that is, time until the system recovers.[56,59]

Depending on the scope of the analysis, L can include only direct or additionally indirect consequences associated with system disruptions.

In the loss analysis, **DM** is mapped to L by means of a system model ν, which can be deterministic or stochastic:

$$L = \nu(\mathbf{DM}) \tag{5}$$

A large variety of system models ν can be found in the literature, ranging from simple connectivity-based models[51,52,54,55,60] to fully physics-based network models.[61,62] In the numerical investigations presented in Section 2.5, we utilize a generic network model that allows accounting for cascading effects.

In the general case of a stochastic loss model, ν defines the conditional PDF of L given the vector of component damage states, $f_{L|\mathbf{DM}}(l|\mathbf{dm}_k)$.

2.6 | Risk analysis

The components of the probabilistic hazard analysis described in Sections 2.2–2.5 can be combined to determine the CDF of L conditional on a hazard event H:

$$F_{L|H}(l|H) = \int_{\mathbb{R}^m} F_{L|\boldsymbol{\Theta}}(l|\boldsymbol{\Theta}) f_{\boldsymbol{\Theta}|H}\left(\theta|H\right) d\theta \tag{6}$$

wherein

$$F_{L|\boldsymbol{\Theta}}(l|\theta) = \int_{\mathbb{R}^n} \sum_{k=1}^{2^n} F_{L|\mathbf{DM}}(l|\mathbf{dm}_k) p_{\mathbf{DM}|\mathbf{IM}}\left(\mathbf{dm}_k|\mathbf{im}\right) f_{\mathbf{IM}|\boldsymbol{\Theta}}\left(\mathbf{im}|\theta\right) d\mathbf{im} \tag{7}$$

is the conditional CDF of the losses given $\boldsymbol{\Theta}$.

A common measure of risk is the loss exceedance rate.[8,26] Under the common (albeit not necessarily correct) assumption of independence between the number of hazard events and the losses in any given hazard event, the loss exceedance rate is

$$\lambda_L(l) = [1 - F_{L|H}(l)]\lambda_H \tag{8}$$

From the inverse of $\lambda_L(l)$, one can derive losses l_t with given return period t as

$$l_t = \lambda_L^{-1}\left(\frac{1}{t}\right) \tag{9}$$

2.7 | Sensitivity of the losses to the hazard scenarios

Equation (6) represents the uncertainty in the losses given a hazard event H, whereas Equation (7) corresponds to the uncertainty in the losses due to all factors different from $\boldsymbol{\Theta}$. In this section, we show how one can quantify the contribution of the uncertainty in $\boldsymbol{\Theta}$ to the overall uncertainty in the losses. For ease of notation, we drop the explicit conditioning on the hazard occurrence H in the following. One can decompose the conditional variance of L given a hazard occurrence by the law of total variance:

$$\sigma_L^2 = \mathbb{V}_{\boldsymbol{\Theta}}(\mu_L(\boldsymbol{\Theta})) + \mathbb{E}_{\boldsymbol{\Theta}}\left(\sigma_L^2(\boldsymbol{\Theta})\right) \tag{10}$$

wherein $\mathbb{V}_{\boldsymbol{\Theta}}$ and $\mathbb{E}_{\boldsymbol{\Theta}}$ are the variance and expected value with respect to $\boldsymbol{\Theta}$; whereas $\mu_L(\theta) = \mathbb{E}_{L|\boldsymbol{\Theta}}(L|\theta)$ and $\sigma_L^2(\theta) = \mathbb{V}_{L|\boldsymbol{\Theta}}(L|\theta)$ are, respectively, the conditional mean and variance of L given θ.

The first term in Equation (10) corresponds to the combined effect of $\boldsymbol{\Theta}$ on the variance of L. One can observe that if $\sigma_L^2 = \mathbb{V}_{\boldsymbol{\Theta}}(\mu_L(\boldsymbol{\Theta}))$, then $\sigma_L^2(\theta) = 0$ and the losses are a deterministic function of θ.

Normalizing Equation (10) by the total variance σ_L^2, one obtains the following equality:

$$1 = \frac{\mathbb{V}_{\boldsymbol{\Theta}}(\mu_L(\boldsymbol{\Theta}))}{\sigma_L^2} + \frac{\mathbb{E}_{\boldsymbol{\Theta}}\left(\sigma_L^2(\boldsymbol{\Theta})\right)}{\sigma_L^2} = S_{\boldsymbol{\Theta}} + S_{-\boldsymbol{\Theta}}^{tot} \tag{11}$$

The term $S_{\boldsymbol{\Theta}}$ is the closed Sobol' index of $\boldsymbol{\Theta}$, that is, the sum of the Sobol' sensitivity indices of all orders involving only the hazard parameters.[63] The term $S_{-\boldsymbol{\Theta}}^{tot}$ is the total effect sensitivity index of $-\boldsymbol{\Theta}$, that is, the sum of the remaining Sobol' sensitivity indices, which are those that consider at least one input random variable that is not a hazard parameter. One can observe that $0 \leq S_{\boldsymbol{\Theta}} \leq 1$, and that $S_{\boldsymbol{\Theta}} = 1$ corresponds to the case where the loss function is deterministic for given $\boldsymbol{\Theta}$, whereas $S_{\boldsymbol{\Theta}} = 0$ is the (hypothetical) case where the uncertainty in the hazard parameter has no effect on the expected losses.

3 | REPRESENTATIVE HAZARD SCENARIOS

3.1 | Definition

The aim of this section is to define hazard scenarios that are representative of specific loss values, typically losses l_t with a specified return period t following Equation (9). A hazard scenario is characterized by a vector θ, and we denote the representative scenario of a loss l_t by θ_t. We present and later investigate the definition of a representative hazard scenario. All distributions utilized in the following are conditional on the occurrence of a hazard event H. For readability, we do not write out this condition.

We define the representative scenario θ_t as the most likely parameter values leading to a loss l_t, that is,

$$\theta_t = \underset{\theta}{\operatorname{argmax}}\, f_{\boldsymbol{\Theta}|L}\left(\theta | l_t\right) \tag{12}$$

Here $f_{\boldsymbol{\Theta}|L}(\theta|l_t)$ is the conditional joint PDF of the hazard parameters, given that the loss equals l_t.

By Bayes' rule, $f_{\boldsymbol{\Theta}|L}(\theta|l_t)$ is equal to

$$f_{\boldsymbol{\Theta}|L}\left(\theta|l_t\right) = f_{L|\boldsymbol{\Theta}}\left(l_t|\theta\right)\frac{f_{\boldsymbol{\Theta}}(\theta)}{f_L(l_t)} \tag{13}$$

Combining Equations (12) and (13), and noticing that $f_L(l_t)$ is constant, θ_t equals

$$\theta_t = \underset{\theta}{\operatorname{argmax}}\, f_{L|\boldsymbol{\Theta}}\left(l_t|\theta\right) f_{\boldsymbol{\Theta}}(\theta) \tag{14}$$

$f_{\boldsymbol{\Theta}}(\theta)$ can be derived from historical records, literature and expert knowledge; $f_{L|\boldsymbol{\Theta}}(l_t|\theta)$ corresponds to the derivative of Equation (7) with respect to l, evaluated at l_t.

Equation (14) shows that θ_t depends on the uncertainty in $\boldsymbol{\Theta}$, as well as the uncertainty associated with the conditional density $f_{L|\boldsymbol{\Theta}}(l|\boldsymbol{\Theta})$. In fact, if $S_{\boldsymbol{\Theta}} = 1$, that is, if all uncertainty in L comes from $\boldsymbol{\Theta}$, then $f_{L|\boldsymbol{\Theta}}(l_t|\boldsymbol{\Theta}) = \delta(l_t - l(\boldsymbol{\Theta}))$ and Equation (14) becomes a constrained optimization problem:

$$\begin{gathered}\theta_t = \underset{\theta}{\operatorname{argmax}}\, f_{\boldsymbol{\Theta}}(\theta) \\ \text{s.t. } l(\theta) = l_t\end{gathered} \tag{15}$$

Equation (15) is analogous to the inverse FORM approach[20,21] for finding representative design parameters. Thereby, $l_t - l(\theta)$ would be equal to the limit state function.

3.2 | Illustration

We illustrate the definition of Section 3.1 through a simple example. The goal is to determine a representative seismic event. At a seismic fault, strong earthquakes occur with a rate of $\lambda_H = 1.0\,\text{yr}^{-1}$. The engineering system under consideration consists of a single component. The hazard parameters $\boldsymbol{\Theta} = [M, \ln R]^\intercal$ are the magnitude M and the log of the hypocentral distance R from the earthquake source to the component. The distribution of $\boldsymbol{\Theta}$ is normal with mean vector $\boldsymbol{\mu}_{\boldsymbol{\Theta}}$ and covariance matrix $\Sigma_{\boldsymbol{\Theta}}$ given as follows:

$$\boldsymbol{\mu}_{\boldsymbol{\Theta}} = \begin{bmatrix} 7.00 \\ 4.38 \end{bmatrix}, \quad \Sigma_{\boldsymbol{\Theta}} = \begin{bmatrix} 0.36 & -0.08 \\ -0.08 & 0.49 \end{bmatrix} \tag{16}$$

The intensity measure IM is the PGA in m/s^2. The GMPE from Esteva and Villaverde[64] is taken as the hazard propagation model of Equation (2):

$$\begin{aligned} PGA(\theta) &= 56\frac{\exp(0.8m + Q_{PGA})}{r^2} \\ &= 56\exp\left(\mathbf{a}^\intercal\theta + Q_{PGA}\right) \end{aligned} \tag{17}$$

wherein $\theta = [m, \ln r]^\intercal$ are realizations of the hazard parameters $\boldsymbol{\Theta}$, $\mathbf{a} = [0.8, -2]^\intercal$, and Q_{PGA} is a normal random variable with zero mean and standard deviation $\sigma_{Q,PGA}$.

For illustrative purposes and to obtain analytical solutions, we assume here that the damage measure DM is continuous and proportional to the intensity measure multiplied with a lognormal random variable. We furthermore take the losses to be proportional to the damage measure, also multiplied with a lognormal random variable. Omitting the proportionality constant, it follows that the losses are

$$L(\theta) = \exp\left(\mathbf{a}^\intercal\theta + Q_{PGA} + Q_{DM} + Q_L\right) \tag{18}$$

wherein Q_{DM} and Q_L are normal random variables with zero mean and standard deviations $\sigma_{Q,DM}$, $\sigma_{Q,L}$, respectively. They represent the uncertainty in the damage state given the PGA and the losses given the damage state. Equation (18) can be re-written in terms of a single standard normal random variable U and a single standard deviation σ:

$$\begin{aligned} L(\theta) &= \exp\left(\mathbf{a}^\intercal\theta + \sigma U\right) \\ \sigma &= \sqrt{\sigma^2_{Q,PGA} + \sigma^2_{Q,DM} + \sigma^2_{Q,L}} \end{aligned} \tag{19}$$

For given values of the hazard parameters, $L(\theta)$ is a lognormal random variable with parameters:

$$\mu_{\ln L|\boldsymbol{\Theta}}(\theta) = \mathbf{a}^\intercal\boldsymbol{\Theta}, \quad \sigma_{\ln L|\boldsymbol{\Theta}}(\theta) = \sigma \tag{20}$$

The unconditional loss L is also lognormal with parameters $\mu_{\ln L} = -3.16$, $\sigma_{\ln L} = \sqrt{2.46 + \sigma^2}$.

Unless otherwise noted, we set $\sigma = 1$ in the following and we consider a scenario with return period $t = 100$yr. The loss associated with this return period is $l_t = 3.21$.

Because the parameters $\boldsymbol{\Theta}$ and the log-loss $\ln L$ are jointly normal, the conditional $f_{\boldsymbol{\Theta}|L}(\boldsymbol{\Theta}|l_t)$ can be evaluated analytically. It is normal with mean $\boldsymbol{\mu}_{\boldsymbol{\Theta}|L=l_t}$ and covariance matrix $\Sigma_{\boldsymbol{\Theta}|L=l_t}$ equal to

$$\boldsymbol{\mu}_{\boldsymbol{\Theta}|L=l_t} = \begin{bmatrix} 7.57 \\ 3.07 \end{bmatrix}, \quad \Sigma_{\boldsymbol{\Theta}|L=l_t} = \begin{bmatrix} 0.30 & 0.05 \\ 0.05 & 0.17 \end{bmatrix} \tag{21}$$

Figure 2 compares this conditional distribution to the unconditional distribution of $\boldsymbol{\Theta}$. Following Equation (12), the representative hazard scenario is $\theta_t = [7.57; 3.07]$. In this model, σ reflects the uncertainty from all sources except $\boldsymbol{\Theta}$. For $\sigma = 0$, the only uncertainty in the losses given an earthquake event comes from $\boldsymbol{\Theta}$. In contrast, as σ becomes large, the contribution of the uncertainty in $\boldsymbol{\Theta}$ to the total uncertainty in L becomes small.

Following Section 2.7, the relative contribution of the uncertainty in $\boldsymbol{\Theta}$ to the uncertainty in the losses can be expressed by the closed Sobol' sensitivity index of $\boldsymbol{\Theta}$. This index is

$$S_{\boldsymbol{\Theta}} = \frac{10.71}{11.71\exp\left(\sigma^2\right) - 1} \tag{22}$$

$S_{\boldsymbol{\Theta}}$ is shown in Figure 3 as a function of σ. For $\sigma = 1$, it is $S_{\boldsymbol{\Theta}} = 0.347$.

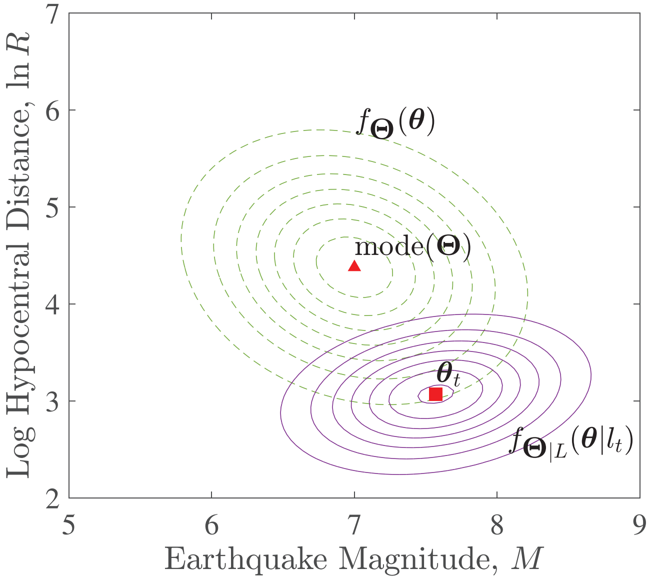


FIGURE 2 The conditional distribution of $\boldsymbol{\Theta}$ given $L = l_t$ and the representative hazard scenario θ_t, with $\sigma = 1$

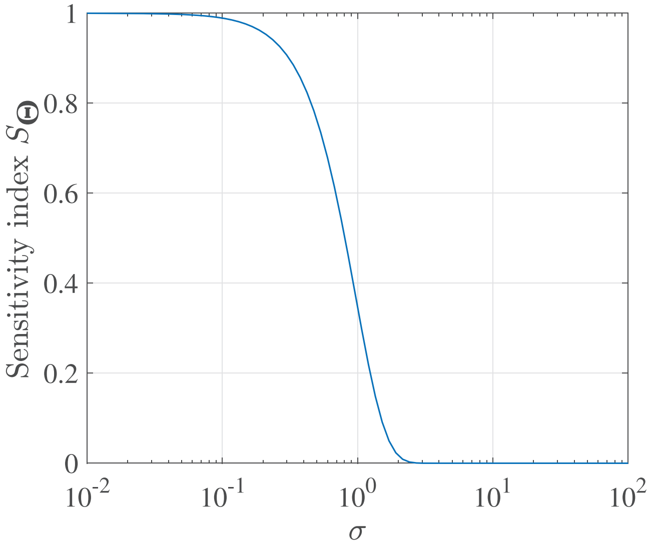


FIGURE 3 Closed Sobol' sensitivity index S_{Θ} for different values of σ

In the following, we investigate the effect of σ on θ_t. Figure 4 shows θ_t for different values of σ. One can observe that for increasing values of σ, θ_t approaches the mode of the distribution of $\boldsymbol{\Theta}$. When $\sigma = 0$, $L(\theta)$ is a deterministic function $l(\theta)$, and l_t can be evaluated following Equation (15), that is, as the value θ with the highest density along the black line shown in Figure 4. For values $\sigma \to 0$, that is, as the combined effect sensitivity index of $\boldsymbol{\Theta}$ tends to 1, θ_t approaches the solution on this line.

Figure 4 illustrate how the relative share of the uncertainty that is associated with the hazard parameters affects the representative scenario. If σ is large, and hence S_{Θ} is small, then the representative hazard scenario for any loss return period is close to the mode of $f_{\Theta}(\theta)$. In this case, the extent of the loss is not determined by the intensity of the hazard, but by other factors. In contrast, if σ is small, and hence S_{Θ} is large, then extreme losses will always be associated with extreme hazard scenarios.

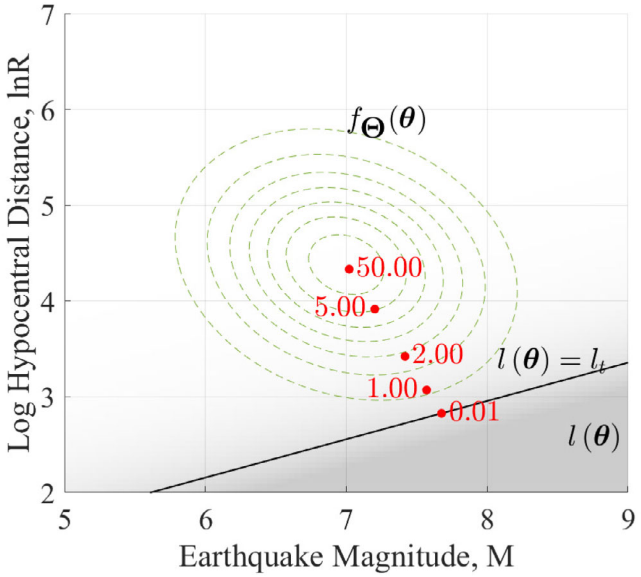


FIGURE 4 Representative hazard scenarios θ_t shown as red dots for different values of σ. The black line corresponds to the values of $\boldsymbol{\Theta}$ for which $\bar{l}(\theta) = l_t$ when $\sigma = 0$. The shade indicates the value of $\bar{l}(\theta)$ for $\sigma = 0$.

4 | NUMERICAL EVALUATION OF REPRESENTATIVE HAZARD SCENARIOS

Generally, analytical solutions to Equation (12) are not available and the representative scenarios must be found numerically. Furthermore, the loss exceedance rate λ_L is also not available in analytical form; it must be evaluated numerically to determine l_t. In this section, we propose a procedure for doing so efficiently. The procedure assumes that the hazard rate and the probability distribution of the hazard parameters are available.

The procedure starts with an initial Monte Carlo sampling. The resulting sample evaluations of the losses are used for estimating l_t and to set up an initial surrogate model of the losses using Gaussian process (GP) regression. Depending on the variance in L given $\boldsymbol{\Theta}$, the procedure chooses between two approximation methods. When this conditional variance of the losses is small, that is, for large $S_{\boldsymbol{\Theta}}$, the procedure employs a GP surrogate of the losses for given θ. If the conditional variance is large, the procedure approximates $f_{L|\boldsymbol{\Theta}}(l_t|\theta)$ with a kernel-density estimation (KDE), which requires additional model evaluations, and learns a GP surrogate of $f_{L|\boldsymbol{\Theta}}(l_t|\theta)$ in function of θ.

To efficiently learn the GP surrogates, an AL method is utilized, which optimally chooses additional samples of $\boldsymbol{\Theta}$ for which the system should be evaluated. AL is a commonly used strategy for improving the predictions of GP regression in optimization[65,66] and reliability analysis.[67,68] At the core of AL is the learning or acquisition function, which determines the best next point to be evaluated for reducing the prediction error of the GP regression. The implemented AL methods are detailed in Sections 4.2 and 4.3. The final surrogate model learned from all sampled scenarios is then employed to solve the optimization problem defined by Equation (14), resp. Equation (15).

Figure 5 shows the different steps of the procedure, which are detailed in Sections 4.1–4.3.

4.1 | Initial sampling

The first step is to generate n_0 random samples of $\boldsymbol{\Theta}$, denoted by $\theta^{(1)}, \ldots, \theta^{(n_0)}$. For each sample, one evaluates the losses with the model, resulting in loss samples $l^{(1)}, \ldots, l^{(n_0)}$.

The choice of n_0 depends on the computational cost of one model evaluation. If the computational budget allows it, we recommend $n_0 \geq 10\lambda_H t$ and generate the samples independently from the distribution $f_{\boldsymbol{\Theta}}(\theta)$. Then one can obtain an estimate of l_t from the ordered samples $l_{(1)} \leq \ldots \leq l_{(n_0)}$:

$$\hat{l}_t = l_{(q-1)}(1-\omega) + \omega l_{(q)}, \tag{23}$$

where q is the smallest integer larger or equal than $n_0(1 - \frac{1}{\lambda_H t})$, and ω is their difference, that is, $\omega = n_0(1 - \frac{1}{\lambda_H t}) - q$.

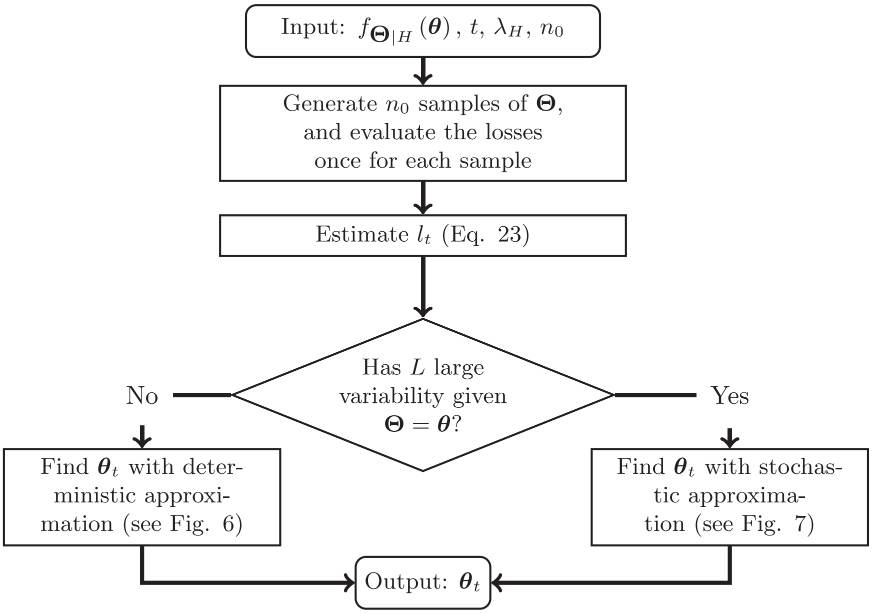

Input: $f_{\boldsymbol{\Theta}|H}(\boldsymbol{\theta})$, t, λ_H, n_0
Generate n_0 samples of $\boldsymbol{\Theta}$, and evaluate the losses once for each sample
Estimate l_t (Eq. 23)
Has L large variability given $\boldsymbol{\Theta}=\boldsymbol{\theta}$?
No
Yes
Find $\boldsymbol{\theta}_t$ with deterministic approximation (see Fig. 6)
Find $\boldsymbol{\theta}_t$ with stochastic approximation (see Fig. 7)
Output: $\boldsymbol{\theta}_t$

FIGURE 5 Scheme of numerical approximation of the representative hazard scenario associated to return period t

Following the initial sampling, it must be decided if the conditional variance of the losses given the hazard parameters is small or large. Quantitatively, this is expressed by $S^{tot}_{-\boldsymbol{\Theta}}$, see Section 2.7. In many cases, it will be clear to the analyst prior to the analysis whether $S^{tot}_{-\boldsymbol{\Theta}}$ is small or large. If not, $S^{tot}_{-\boldsymbol{\Theta}}$ can be estimated with a sampling approximation based on the generated samples of $\boldsymbol{\Theta}$ and L.

If the conditional variance is deemed to be small, the procedure proceeds with learning a surrogate of the losses given $\boldsymbol{\theta}$ following Section 4.2 (deterministic approximation). Otherwise, a surrogate modelling of the objective function is employed as described in Section 4.3 (stochastic approximation). If one is unsure about whether the conditional variance is 'large' or 'small', one should choose the stochastic approximation, since it works in all cases (but at a higher computational cost than the deterministic approximation).

4.2 | Deterministic approximation of the conditional losses

If the conditional variance of the losses given $\boldsymbol{\theta}$ is small, it is justified to replace $f_{L|\boldsymbol{\Theta}}(l|\boldsymbol{\theta})$ with a deterministic model of the losses in function of $\boldsymbol{\theta}$, that is, $l(\boldsymbol{\theta}) \approx \mathbb{E}[L|\boldsymbol{\theta}]$. With this approximation, the representative scenario can be found by solving Equation (15).

To model $l(\boldsymbol{\theta})$, we employ GP regression[69] and obtain the mean $\mu_{\mathcal{G}}(\boldsymbol{\theta})$ and standard deviation function $\sigma_{\mathcal{G}}(\boldsymbol{\theta})$. The loss function is approximated by the mean function, and Equation (15) is replaced with the approximation:

$$\begin{aligned} \boldsymbol{\theta}_t \approx \underset{\boldsymbol{\theta}}{\operatorname{argmax}} f_{\boldsymbol{\Theta}}(\boldsymbol{\theta}), \\ \text{s.t. } \mu_{\mathcal{G}}(\boldsymbol{\theta}) = l_t \end{aligned} \tag{24}$$

The approximation of the loss function with a GP introduces a prediction error. However, one can reduce it near the equality constraint of Equation (24) by adding more scenarios in this region. One can achieve this with an AL method whose acquisition function looks for new scenarios close to the constraint. For that purpose, we employ the Expected Feasibility Function EFF as acquisition function,[67] defined as

$$EFF(\boldsymbol{\theta}) = \int_{l_t-\epsilon}^{l_t+\epsilon} \frac{\epsilon - |l_t - u|}{\sigma_{\mathcal{G}}(\boldsymbol{\theta})} \phi\left(\frac{u-\mu_{\mathcal{G}}(\boldsymbol{\theta})}{\sigma_{\mathcal{G}}(\boldsymbol{\theta})}\right) du \tag{25}$$

wherein $\epsilon \propto \sigma_{\mathcal{G}}(\boldsymbol{\theta})$.

An optimization algorithm is employed to find the scenario $\boldsymbol{\theta}$ with the maximum $EFF(\boldsymbol{\theta})$. This scenario is added to the training set and the GP is re-trained. This procedure is repeated until the computational budget n_3 is exceeded. After

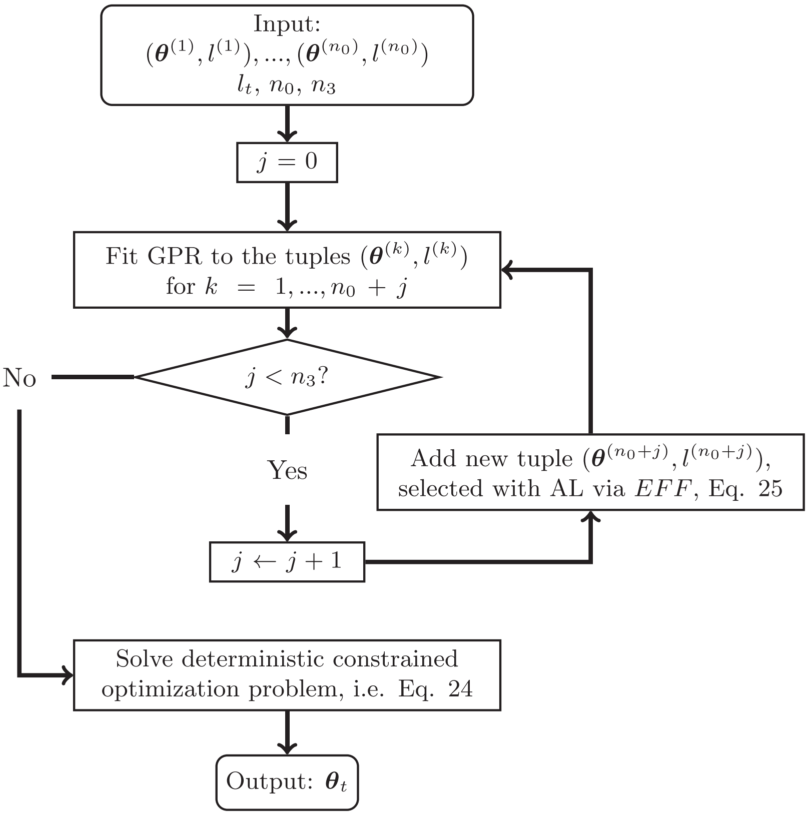

Input:
$(\boldsymbol{\theta}^{(1)}, l^{(1)}), ..., (\boldsymbol{\theta}^{(n_0)}, l^{(n_0)})$
l_t, n_0, n_3
$j = 0$
Fit GPR to the tuples $(\boldsymbol{\theta}^{(k)}, l^{(k)})$ for $k = 1, ..., n_0 + j$
$j < n_3?$
No
Yes
Add new tuple $(\boldsymbol{\theta}^{(n_0+j)}, l^{(n_0+j)})$, selected with AL via EFF, Eq. 25
$j \leftarrow j + 1$
Solve deterministic constrained optimization problem, i.e. Eq. 24
Output: $\boldsymbol{\theta}_t$

FIGURE 6 Scheme for finding the representative hazard scenario with the deterministic approximation of the losses

adding sequentially n_3 new scenarios with AL, one solves Equation (24) and finds the numerical approximation of θ_t. Figure 6 summarizes the approach.

4.3 | Stochastic approximation of the conditional distribution of the losses

If the conditional variance of the losses given θ is large, then one cannot utilize the deterministic approximation $l(\theta)$, and a normal approximation of the conditional density of the losses is inappropriate in most cases. Therefore, we propose to approximate $f_{L|\Theta}(l_t|\theta)$ with KDE for selected scenarios $\theta^{(k)}$, $k = 1, \dots, n_1$. At these scenarios, the objective function in Equation (14) is approximated as

$$\begin{aligned} f_{L|\Theta}\left(l_t|\theta^{(k)}\right) f_{\Theta|H}\left(\theta^{(k)}\right) &\approx \hat{f}_{L|\Theta}\left(l_t|\theta^{(k)}\right) f_{\Theta|H}\left(\theta^{(k)}\right) \\ &= \frac{f_{\Theta|H}\left(\theta^{(k)}\right)}{n_2\gamma} \sum_{i=1}^{n_2} \kappa\left(\frac{l_t - l^{(k,i)}}{\gamma}\right) = g\left(\theta^{(k)}\right) \end{aligned} \tag{26}$$

wherein κ is a kernel function and γ the bandwidth, which can depend on $\theta^{(k)}$. $l^{(k,i)}$ are loss samples conditional on $\theta^{(k)}$, which are obtained from system model evaluations. n_2 is the number of model evaluations at the scenario $\theta^{(k)}$. Note that one can use the same n_2 model evaluations for estimating the conditional density at multiple loss values, corresponding to different return periods.

To limit the number of evaluations of Equation (26), a GP surrogate of the objective function $f_{L|\Theta}(l_t|\Theta) f_\Theta(\theta)$ is constructed. One needs to select the training set for this GP regression efficiently, because each evaluation of $g(\theta)$ (Equation 26) requires n_2 model evaluations. To this end, we propose to choose from the n_0 scenarios of the initial sampling the n_a scenarios whose sampled losses are closest to l_t. One then clusters these n_a scenarios into n_1 clusters with k-medoids,[70] and takes the medoids of the clusters $\theta^{(c_1)}, \dots, \theta^{(c_{n_1})}$ as the initial training set for learning the GP. We consider $n_a = \min(n_0, 5n_1)$.

Density estimation introduces uncertainty. We assess this uncertainty with bootstrapping.[71] That is, at scenario $\theta^{(c_k)}$ and after estimating $g^{(k)}$, one computes the standard deviation $s_g^{(k)}$ of n_b KDEs from re-sampled loss values. The bootstrap standard deviations of the KDEs $s_g^{(1)}, \dots, s_g^{(n_1)}$ evaluated at the points $\theta^{(c_1)}, \dots, \theta^{(c_{n_1})}$ are the training set for a surrogate GP S to learn the noise variance over the hazard parameters.

We employ AL to select informative scenarios θ at which to evaluate the objective function. A suitable acquisition function in the context of an optimization problem is the Augmented Expected Improvement AEI,[65,66] which we employ here. The AEI is

$$AEI(\theta) = \mathbb{E}\left(\max\left(\mu_G(\theta) - \mu_q^*, 0\right)\right)\left(1 - \frac{\tau(\theta)}{\sqrt{\sigma_G^2(\theta) + \tau^2(\theta)}}\right) \tag{27}$$

wherein $\tau^2(\theta)$ is the noise variance, which we approximate with the mean function of S, and μ_q^* is a representative value of the maximum observed value at AL iteration q. Based on a similar idea of Picheny et al.,[72] with $0 \le \alpha \le 1$, we consider the following expression for μ_q^*:

$$\mu_q^* = \max_{j=1,\dots,n_1+q}\left(g^{(j)} + \Phi^{-1}(1-\alpha/2)s_g^{(j)}\right) \tag{28}$$

We set $\alpha = 0.01$.

The scenario with the largest AEI is found via a standard optimization algorithm. The scenario is added to the training set and the GP is re-trained. This procedure is repeated until the computational budget n_3 is exceeded.

After adding sequentially n_3 new scenarios with AL, we solve the optimization problem with the GP surrogate of the objective function, that is,

$$\theta_t \approx \hat{\theta}_t = \underset{\theta}{\operatorname{argmax}}\, \mu_G(\theta) \tag{29}$$

Figure 7 summarizes the method.

The total number of model evaluations n_{tot} is $n_0 + n_2(n_1 + n_3) - n_1$. The last term is related to the re-use of loss evaluations from the initial sampling when evaluating the KDE at the n_1 scenarios $\theta^{(c_1)}, \dots, \theta^{(c_{n_1})}$.

The choice of n_1 and n_2 has a significant impact on the performance of the methodology. n_1 is the number of scenarios of the initial training size of the GP regression, hence it should be chosen in function of the dimensionality of $\boldsymbol{\Theta}$. Based on numerical investigations, we recommend to choose n_1 in the range $20 \le n_1 \le \max(20, 0.1n_0)$. The reason for the upper bound on n_1 is to avoid including too many scenarios in the training set that are far from the solution. If one wants a large n_1, then one should increase n_0 as well.

n_2 is the number of loss evaluations necessary for a KDE of L given $\boldsymbol{\Theta} = \theta$. For n_2, the recommendation is to use a value $n_2 \ge 20$. n_3 is the number of AL steps, which implies additional n_3 KDEs. For that reason, we suggest a value $5 \le n_3 \le n_1$.

We observed that the uncertainty in the density estimation dominates the overall uncertainty in the GP predictions, together with the uncertainty in estimating l_t. Therefore, if the computational budget is large enough, additional model evaluations should focus on n_0 and n_2, rather than on n_1 and n_3. Exemplary choices of the number of samples for different computational budgets are shown in the next section.

4.4 Illustration

We illustrate the performance of the proposed method by approximating the solution of the basic example presented in Section 3. The analytical solution of Section 3 is utilized to assess the quality of the approximation method. The return period of interest is $t = 100$yr. To investigate the deterministic approximation (Section 4.2), we use the example with the setting $\sigma = 0.01$, in which case the uncertainty of the losses given θ is small. To investigate the stochastic approximation (Section 4.3), we set $\sigma = 1$.

In the deterministic case, we set a computational budget of $n_{tot} = 10^3$ model evaluations. From that budget, we generate $n_0 = 980$ random scenarios with Monte Carlo Simulation (MCS), and the remaining $n_3 = 20$ with AL. Figure 8 shows the

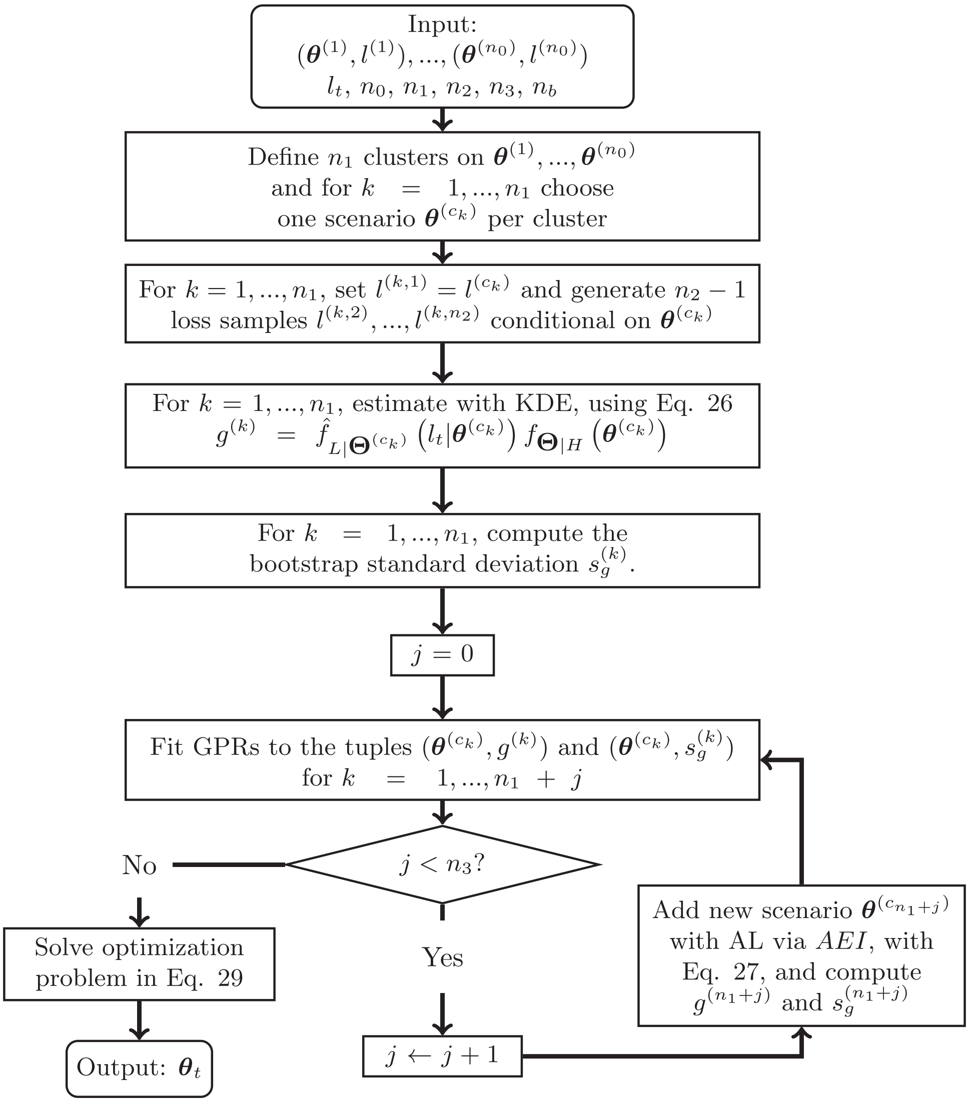

Input:
$(\boldsymbol{\theta}^{(1)}, l^{(1)}), ..., (\boldsymbol{\theta}^{(n_0)}, l^{(n_0)})$
$l_t,\ n_0,\ n_1,\ n_2,\ n_3,\ n_b$
Define n_1 clusters on $\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(n_0)}$ and for $k = 1, ..., n_1$ choose one scenario $\boldsymbol{\theta}^{(c_k)}$ per cluster
For $k = 1, ..., n_1$, set $l^{(k,1)} = l^{(c_k)}$ and generate $n_2 - 1$ loss samples $l^{(k,2)}, ..., l^{(k,n_2)}$ conditional on $\boldsymbol{\theta}^{(c_k)}$
For $k = 1, ..., n_1$, estimate with KDE, using Eq. 26
$g^{(k)} = \hat{f}_{L|\boldsymbol{\Theta}^{(c_k)}}\left(l_t|\boldsymbol{\theta}^{(c_k)}\right) f_{\boldsymbol{\Theta}|H}\left(\boldsymbol{\theta}^{(c_k)}\right)$
For $k = 1, ..., n_1$, compute the bootstrap standard deviation $s_g^{(k)}$.
$j = 0$
Fit GPRs to the tuples $(\boldsymbol{\theta}^{(c_k)}, g^{(k)})$ and $(\boldsymbol{\theta}^{(c_k)}, s_g^{(k)})$ for $k = 1, ..., n_1 + j$
$j < n_3$?
No
Yes
Solve optimization problem in Eq. 29
Output: $\boldsymbol{\theta}_t$
$j \leftarrow j + 1$
Add new scenario $\boldsymbol{\theta}^{(c_{n_1+j})}$ with AL via AEI, with Eq. 27, and compute $g^{(n_1+j)}$ and $s_g^{(n_1+j)}$

FIGURE 7 Scheme of stochastic approximation of the representative hazard scenario associated to return period t

variability of the estimation of $\boldsymbol{\theta}_t$ with 20 experiments. One can observe in Figure 8 that 10^3 model evaluations are sufficient for obtaining an estimation near to the true solution in the deterministic case. The estimation lies close to the loss contour line corresponding to l_t.

For the stochastic case, we evaluate the solution with different computational budgets n_{tot}, as detailed in Figure 9. In all cases, we compute the bootstrap standard deviation of the KDEs with a re-sampling size of 400.

The stochastic approximation requires 10 times more model evaluations than the deterministic approximation for obtaining an acceptably accurate estimate of the representative hazard scenario, as seen from Figure 9. Most of the evaluations are invested in conditional KDEs close to the solution. In terms of training set sizes, the stochastic approximation utilizes between 20 and 60 scenarios for learning the initial GP surrogate, all of them selected close to the solution.

5 | NUMERICAL EXAMPLE

We apply the methodology to the IEEE39 bus system, whose topology is displayed in Figure 10. It consists of 39 nodes and 43 edges, which represent substations and transmission lines, respectively. The system is assumed to be exposed to earthquake events, whose hazard parameters are the hypocenter location $\mathbf{Z}_0 = [X; Y; H]$ and the magnitude M. We consider only events with magnitudes between 6 and 9.5, with a hazard rate of $\lambda_H = 1yr^{-1}$. The representative hazard scenario is estimated for a loss return period of $t = 100$ yr.

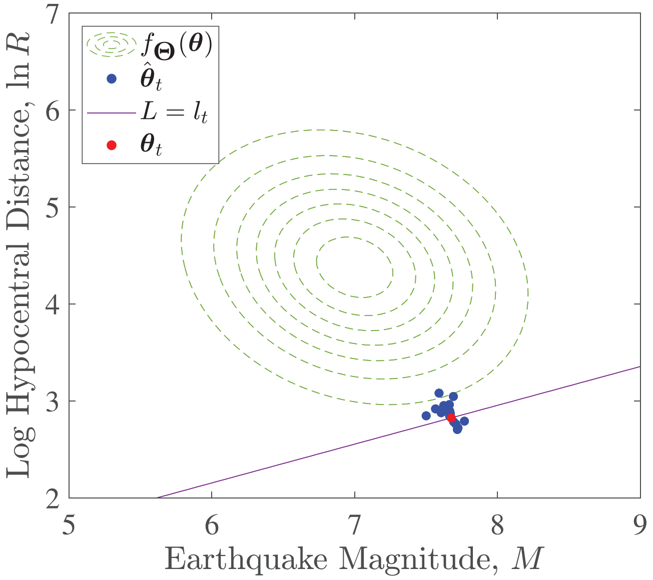


FIGURE 8 Variability of estimation of θ_t, based on 20 runs, for the case of deterministic approximation ($\sigma = 0.01$). $n_{tot} = 10^3$: $n_0 = 980$, $n_3 = 20$

TABLE 1 Shape parameters α and β, mean, standard deviation and bounds of the beta distributed hazard parameters

Parameter	Distr.	α	β	Min	Max	μ	σ
x-coordinate, X [km]	Beta	2	5	−400	400	−171.43	127.77
y-coordinate, Y [km]	Beta	2	5	−400	400	−171.43	127.77
Depth, H [km]	Beta	2	6	−80	0	−60.00	11.55
Magnitude, M [-]	Beta	1	3.5	6.0	9.5	6.78	0.62

The hazard model is presented in Section 5.1, and the system model in Section 5.2. The numerical results are shown in Section 5.3.

5.1 | Hazard model

The hazard parameters $\boldsymbol{\Theta} = [X; Y; H; M]$ are summarized in Table 1. They all follow a beta distribution with parameters α and β. The hazard parameters are assumed to be statistically independent.

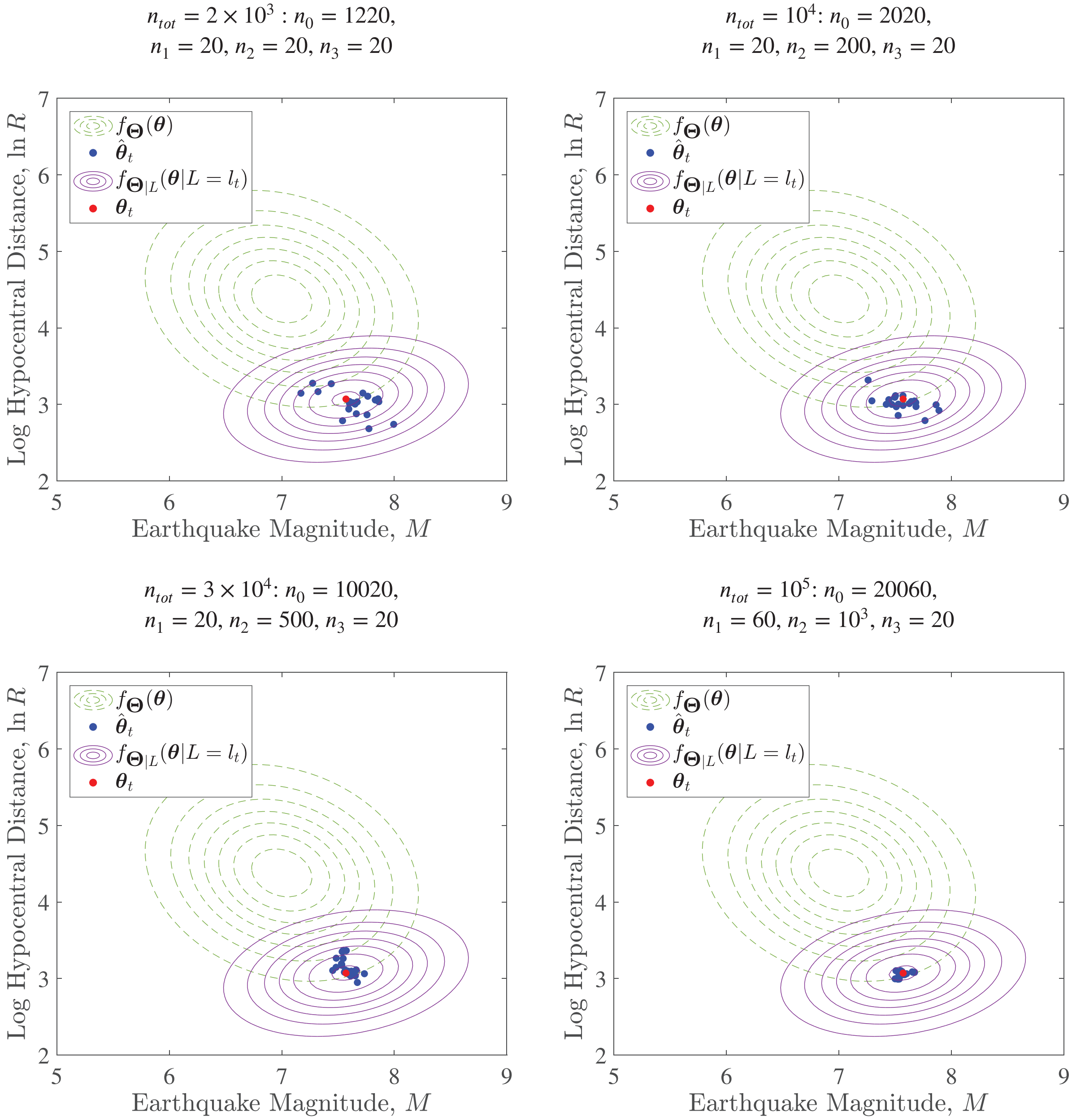
The output of the hazard model is the PGA in m/s^2. A GMPE of the form described by Esteva and Villaverde[64] is taken as the hazard propagation model of Equation (2) at the ith system component with location $\mathbf{z}_i$:

$$\begin{aligned} PGA_i(\theta) &= \xi(\theta, \mathbf{z}_i, U_{PGA}(\mathbf{z}_i))) \\ &= 56\frac{\exp(0.8m + 0.04 + 0.64U_{PGA}(\mathbf{z}_i))}{(\|\mathbf{z}_0 - \mathbf{z}_i\|_2 + 40)^2} \end{aligned} \tag{30}$$

wherein $\mathbf{z}_0 = [x, y, h]$ is the hypocenter location and $\|\cdot\|_2$ evaluates the Euclidian distance. $U_{PGA}(\mathbf{z})$ is a Gaussian random field with zero mean, unit variance, and a squared exponential auto-correlation coefficient function with a correlation length of 50 km.

5.2 | System model

All substations have the same fragility function, represented by a lognormal CDF with parameters $\mu_{Fr} = -1.77$ and $\sigma_{Fr} = 0.35$. The parameters correspond to the fragility function of the extensive damage state for high voltage substations with unanchored elements.[73] No direct failures of transmission lines due to an earthquake are considered.



$n_{tot} = 2 \times 10^3 : n_0 = 1220,$
$n_1 = 20, n_2 = 20, n_3 = 20$

$n_{tot} = 10^4 : n_0 = 2020,$
$n_1 = 20, n_2 = 200, n_3 = 20$

$n_{tot} = 3 \times 10^4 : n_0 = 10020,$
$n_1 = 20, n_2 = 500, n_3 = 20$

$n_{tot} = 10^5 : n_0 = 20060,$
$n_1 = 60, n_2 = 10^3, n_3 = 20$

FIGURE 9 Variability of estimation of θ_t with stochastic approximation, based on 20 runs. $n_{tot} = n_0 + n_2(n_1 + n_3) - n_1$

We simulate the system response with the generic model of cascading failures proposed by Crucitti et al.,[54] using the line reactances as edge weights. The loss measure L depends on the network efficiency $\mathcal{E}$:

$$L = 1 - \frac{\mathcal{E}(\mathbf{dm})}{\mathcal{E}(\mathbf{0})} \tag{31}$$

$$\mathcal{E}(\mathbf{dm}) = \frac{1}{|\mathcal{S}||\mathcal{T}|} \sum_{\substack{s \in \mathcal{S} \\ t \in \mathcal{T} \\ s \neq t}} \varepsilon_{st}(\mathbf{dm}) \tag{32}$$

where $\varepsilon_{st}(\mathbf{dm})$ is the efficiency of the most efficient path from source node s to terminal node t, $\mathcal{S}$ is the set of source nodes, $\mathcal{T}$ the set of terminal nodes. Equation (32) is the network efficiency associated with the component damage vector $\mathbf{dm}$ and $\mathcal{E}(\mathbf{0})$ is the efficiency of the intact system. In power networks, the efficiency of a path is equal to the inverse of the sum of the reactance values along that path. Equation (31) is a modified version of the network efficiency.[54]

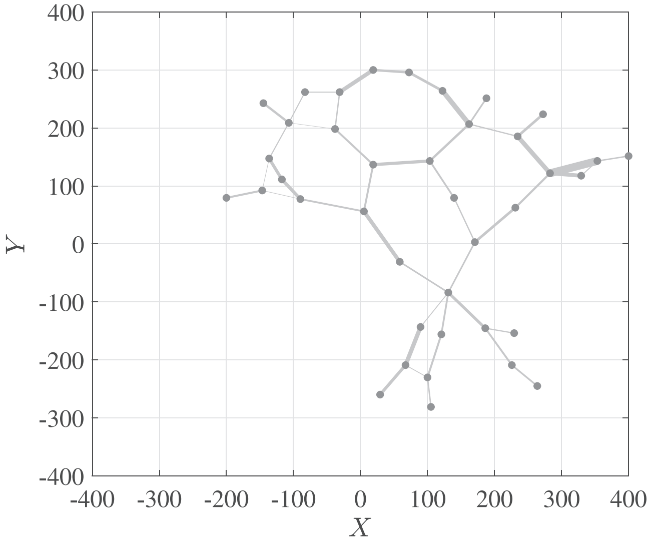


FIGURE 10 Geolocation of the IEEE39 power system, with hypothetical coordinates in km. The edge thicknesses are proportional to their reactance.

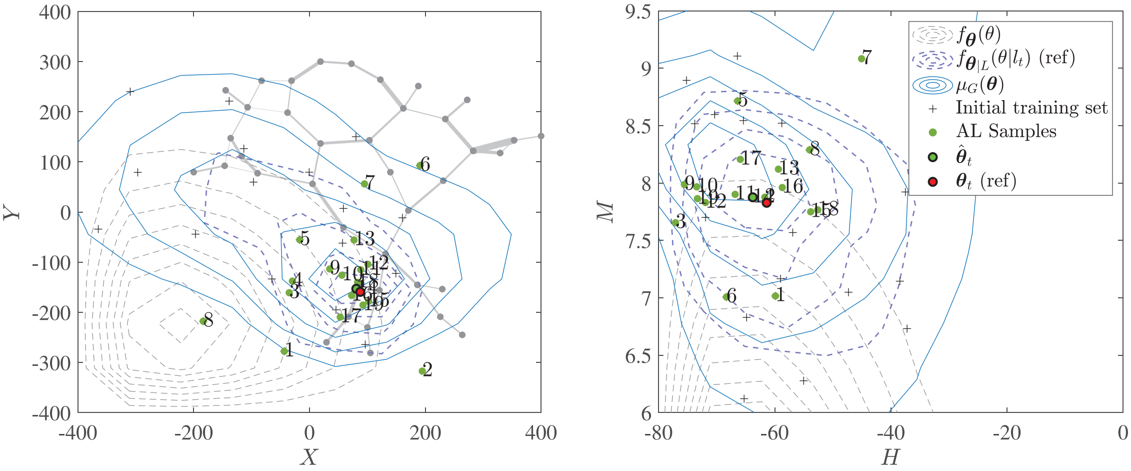


FIGURE 11 Estimation of θ_t for one experiment with 10^4 model evaluations, showing contour lines of the GP mean function, the initial training set, and the additional AL scenarios. The reference PDF contours correspond to sections at the reference solution

5.3 | Estimation of θ_t

A preliminary analysis of the system model clearly shows that the conditional variance of the losses given the hazard parameters is considerable and the stochastic approximation should be employed. (While not necessary in practice, we also evaluate the contribution of the uncertainty in the system model and the loss evaluation on the overall uncertainty as $S^{tot}_{-\boldsymbol{\Theta}} = 0.45$.)

We consider a computational budget of $n_{tot} = 10^4$ model evaluations, which we distribute as $n_0 = 2020$, $n_1 = 20$, $n_2 = 200$, $n_3 = 20$.

We compare the performance of the numerical approximation with a reference solution. This solution utilizes first 10^6 randomly sampled scenarios for estimating l_t. Thereafter, an additional 10^4 scenarios are evaluated in a full factorial experimental design whereby each hazard parameter is discretized to 10 values. At each scenario, the losses are evaluated 220 times. Finally, linear interpolation is employed over the four-dimensional grid for finding θ_t. In total, the reference solution is based on 3.2×10^6 model evaluations.

Figure 11 shows the results of one computation of θ_t. One can observe that the initial training set, which was selected from the initial MCS, is spread, but not located around the prior mode. The scenarios added with AL show that the

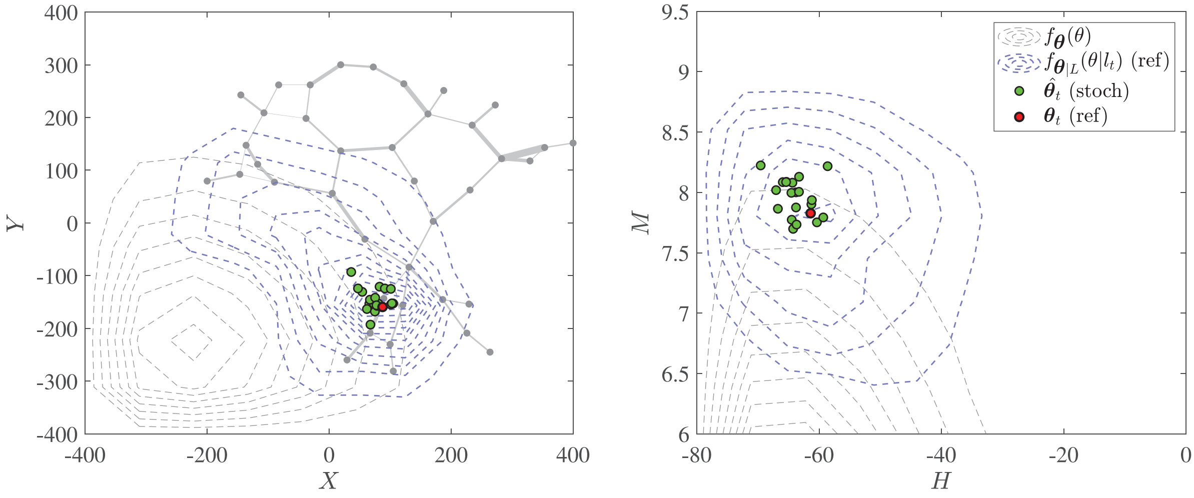


FIGURE 12 Variability of estimation of θ_t, based on 20 runs. PDF contours correspond to sections at the reference solution.

chosen acquisition function balances well exploration and exploitation. The resulting GP estimate $\mu_G(\theta)$ matches well the conditional distribution $f_{\Theta|L}(\theta|l_t)$ obtained with the reference solution.

Figure 12 shows the resulting $\hat{\theta}_t$ from 20 repeated evaluations. One can observe that the variability is close to the reference solution

6 | CONCLUDING REMARKS

Representative hazard scenarios are of relevance for many tasks in risk management of engineering systems. In cases where the hazard is characterized by one dominant parameter, such as the annual maximum discharge in fluvial flood events, such scenarios are typically defined as those with a specific return period, for example, the 100-year flood event. For hazards that are characterized by multiple stochastic parameters, such as the earthquake hazard for a given region, it was unclear how a scenario can be associated to a given return period. In this contribution, we close this gap by proposing a definition of representative hazard scenarios for engineering systems. The return period of a scenario is thereby evaluated with respect to the losses, that is, a 100-year hazard scenario is one that leads to losses with an AEP of 100 years. The representative scenario is then defined as the one with the highest probability among all scenarios leading to these losses. The proposed definition is applicable to systems with a single location, such as buildings, as well as spatially distributed systems, such as power networks, or even systems of systems. The definition also considers the uncertainty in the system response and the losses.

It is important to realize that the proposed representative hazard scenario is a function not only of the hazard but also of the considered system. One should not refer to a representative scenario without specifying the system for which it is representative. A consequence of this definition is that any action that modifies the system (e.g., adding redundancy or reinforcing component structures) can lead to a change of the representative hazard scenarios. In most cases, in particular for small changes in the system, this change will not be large, and one can keep working with the previously determined representative scenario. In some cases, when larger changes to the system are made, the representative scenario can indeed change, but we see this as a positive feature of our definition. Assume that in a spatially distributed network, a weak section is strengthened, then it can be that the representative earthquake scenario changes from one that it is in the vicinity of this section to another one that is closer to another weak section. Such a change in the system should indeed be reflected in the representative hazard scenario.

We developed a methodology for evaluating such representative hazard scenarios numerically. The focus of the proposed algorithms is on an efficient evaluation; nevertheless, the necessary number of system performance evaluations is still in the order of 10^3–10^4 when evaluating scenarios with a return period of 100 years.

This large number restricts the complexity of the system models that can be used within the framework. However, using simplified system models for identifying the representative hazard scenarios seems justified. Once the scenarios are determined, more advanced models can be applied to these scenarios to assess the appropriateness of the simpler models.

In this paper, we show how a single representative hazard scenario can be evaluated. The method can be extended easily to obtain representative hazard scenarios for multiple return periods. Essentially, all model evaluations utilized for determining the representative scenario for one return period can be re-used when determining the one for another return period. Only the AL steps need to be performed separately for each return period.

In some cases, one might also wish to obtain multiple hazard scenarios that are representative of the same return period. Our definition does not cover this case. When the distribution of the hazard parameters given the losses l_t is multi-modal, our definition, which currently only identifies the highest mode, can be extended to cover these multiple modes. If $f_{\Theta|L}(\theta|l_t)$ has only a single mode, then one might identify additional 'representative' scenarios at some distance around this mode. We leave this question for future research.

While the examples in this paper consider single-hazard scenarios, the scenario definition and the proposed methodology for evaluation are applicable to multi-hazard scenarios, for example, earthquake followed by tsunami, floods and landslides. In this case, the hazard model corresponds to the combination of the models of the individual hazards.

At present, the methodology is applicable to hazards that are characterized by continuous parameters, whose joint probability distribution is available. Future work should extend the methodology to discrete and categorical hazard parameters, for example, if an earthquake can originate in multiple faults. Furthermore, in some cases, available information on hazards is not in the form of models with continuous random input parameters, but rather in the form of selected scenarios, for example, from an earthquake catalogue. In this case, the scenarios must be parameterized before the methodology is applicable.

The methodology is also restricted to hazard scenarios described by a limited number of parameters, in the order of up to 10 parameters. This is due to the use of GP regression, whose performance deteriorates with increasing dimensions. For most hazards, this restriction is not crucial. However, the method might have difficulties when one wants to include information about a spatial distribution of a hazard, for example, if one would want to describe earthquake hazard in terms of a ground motion map. In such cases, the methodology might be extended with dimensionality reduction techniques, such as principal component analysis (PCA) or partial least squares regression (PLS).[74,75]

The general idea behind the definition of representative scenarios could also be extended beyond the hazard parameters to system parameters, for example, to identify scenarios of network failures associated with a specific return period. In these cases, the number of parameters describing these scenarios would be even higher. Therefore, it seems worthwhile to further investigate the extension of the methodology to scenarios defined by a larger number of input parameters.

ACKNOWLEDGEMENTS

This work has been sponsored by the research and development project RIESGOS (Grant No. 03G0876A-J) and RIESGOS 2.0 (Grant No. 03G0905A-H), both funded by the German Federal Ministry of Education and Research (BMBF), also known in German as Bundesministerium für Bildung und Forschung as part of the funding programme 'BMBF CLIENT II – International Partnerships for Sustainable Innovations'.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Hugo Rosero-Velásquez https://orcid.org/0000-0003-0615-9293

REFERENCES

1. DesRoches R, ed. Hurricane Katrina: Performance of Transportation Systems. *Number 29 in Technical Council on Lifeline Earthquake Engineering Monographs*. American Society of Civil Engineers; 2006.
2. Kwasinski A, Weaver WW, Chapman PL, Krein PT. Telecommunications power plant damage assessment for Hurricane Katrina – site survey and follow-up results. *IEEE Sys J*. 2009;3(3):277-287.
3. Tang A, Eidinger JM, eds. Chile Earthquake of 2010: Lifeline Performance. *Number 36 in Technical Council on Lifeline Earthquake Engineering Monographs*. American Society of Civil Engineers; 2006.
4. Wu J, Dueñas-Osorio L. Calibration and validation of a seismic damage propagation model for interdependent infrastructure systems. *Earthq Spectra*. 2013;29(3):1021-1041.
5. Krishnamurthy V, Kwasinski A, Dueñas-Osorio L. Comparison of power and telecommunications dependencies and interdependencies in the 2011 Tohoku and 2010 Maule earthquakes. *J Infrastruct Sys*. 2016;22(3):04016013.
6. Bründl M, Romang HE, Bischof N, Rheinberger CM. The risk concept and its application in natural hazard risk management in Switzerland. *Nat Hazards Earth Syst Sci*. 2009;9(3):801-813.

7. Jayaram N, Baker JW. Efficient sampling and data reduction techniques for probabilistic seismic lifeline risk assessment. *Earthq Eng Struct Dyn*. 2010;39(10):1109-1131.
8. Miller M, Baker J. Ground-motion intensity and damage map selection for probabilistic infrastructure network risk assessment using optimization. *Earthq Eng Struct Dyn*. 2015;44(7):1139-1156.
9. Salgado-Gálvez MA, Zuloaga D, Henao S, Bernal GA, Cardona OD. Probabilistic assessment of annual repair rates in pipelines and of direct economic losses in water and sewage networks: application to Manizales, Colombia. *Nat Hazards*. 2018;93(1):5-24.
10. Ouyang M, Wang Z. Resilience assessment of interdependent infrastructure systems: with a focus on joint restoration modeling and analysis. *Reliab Eng Syst Saf*. 2015;141:74-82.
11. Rogers MD. Scientific and technological uncertainty, the precautionary principle, scenarios and risk management. *J Risk Res*. 2001;4(1):1-15.
12. Duinker PN, Greig LA. Scenario analysis in environmental impact assessment: improving explorations of the future. *Environ Impact Assess*. 2007;27(3):206-219.
13. Hassani BK. *Scenario analysis in risk management – theory and practice in finance*. Springer; 2016.
14. IPCC. Summary for policymakers. In: *Climate Change 2014: Mitigation of Climate Change: Working Group III Contribution to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press; 2014: 151-206.
15. Aguirre P, Vásquez J, de la Llera JC, González J, González G. Earthquake damage assessment for deterministic scenarios in Iquique, Chile. *Nat Hazards*. 2018;92(3):1433-1461.
16. Riga E, Karatzetzou A, Apostolaki S, Crowley H, Pitilakis K. Verification of seismic risk models using observed damages from past earthquake events. *Bull Earthq Eng*. 2021;19(2):713-744.
17. Romero N, Nozick LK, Dobson I, Xu N, Jones DA. Seismic retrofit for electric power systems. *Earthq Spectra*. 2015;31(2):1157-1176.
18. Vaziri P, Davidson R, Apivatanagul P, Nozick L. Identification of optimization-based probabilistic earthquake scenarios for regional loss estimation. *J Earthq Eng*. 2012;16(2):296-315.
19. Brown NJ, Gearhart JL, Jones DA, Nozick LK, Romero N, Xu N.: *Optimizing the Selection of Scenarios for Loss Estimation in Transportation Networks*. Sandia National Lab. (SNL-NM); 2011.
20. Berk M, Špačková O, Straub D. Probabilistic design storm method for improved flood estimation in ungauged catchments. *Water Resour Res*. 2017;53(12):10701-10722.
21. Winterstein S, Ude TC, Cornell CA, Bjerager P, Haver S. Environmental parameters for extreme response: inverse FORM with omission factors. In Proceedings of the International Conference on Structural Safety and Reliability ICOSSAR93. 1993:551-557.
22. Moehle J, Deierlein G. A framework methodology for performance-based earthquake engineering. In: Proceedings of the 13th World Conference on Earthquake Engineering (13WCEE). 2004:679.
23. Kiureghian AD. Non-ergodicity and PEER's framework formula. *Earthq Eng Struct Dyn*. 2005;34(13):1643-1652.
24. Cornell CA. Engineering seismic risk analysis. *Bull Seismol Soc Am*. 1968;58(5):1583-1606.
25. FEMA. *Seismic Performance Assessment of Buildings*. Report P-58-1. 2012.
26. Baker JW. *An Introduction to Probabilistic Seismic Hazard Analysis*. White Paper, Version 2.0.1. 2013.
27. Pitilakis K, Crowley H, Kaynia A, eds. *SYNER-G: Typology Definition and Fragility Functions for Physical Elements at Seismic Risk*. Springer; 2014.
28. Beer M, Kougioumtzoglou IA, Patelli E, Au IS-K, eds. *Encyclopedia of Earthquake Engineering*. Springer; 2015.
29. Yang TY, Moehle J, Stojadinovic B, Kiureghian AD. Seismic performance evaluation of facilities: methodology and implementation. *J Struct Eng*. 2009;135(10):1146-1154.
30. Scherb A, Garrè L, Straub D. Probabilistic risk assessment of infrastructure networks subjected to hurricanes. In: Proceedings of the 12th International Conference on Applications of Statistics and Probability in Civil Engineering ICASP12. 2015:388.
31. Khajwal AB, Noshadravan A. Probabilistic hurricane wind-induced loss model for risk assessment on a regional scale. *ASCE-ASME J Risk Uncertain Eng Syst A: Civ Eng*. 2020;6(2):04020020.
32. Pant R, Thacker S, Hall JW, Alderson D, Barr S. Critical infrastructure impact assessment due to flood exposure. *J Flood Risk Manag*. 2018;11(1):22-33.
33. Alipour A, Ahmadalipour A, Moradkhani H. Assessing flash flood hazard and damages in the southeast United States. *J Flood Risk Manag*. 2020;13(2):e12605.
34. Alberico I, Petrosino P, Lirer L. Volcanic hazard and risk assessment in a multi-source volcanic area: the example of Napoli city (Southern Italy). *Nat Hazards Earth Syst Sci*. 2011;11(4):1057-1070.
35. Retnowati DA, Meilano I, Riqqi A. Modeling of volcano eruption risk toward building damage and affected population in Guntur, Indonesia. In: Proceedings of the 2018 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology AGERS. 2018:60-66.
36. Papale P. Global time-size distribution of volcanic eruptions on Earth. *Sci Rep*. 2018;8:6838.
37. Smid M, Russo S, Costa AC, Granell C, Pebesma E. Ranking European capitals by exposure to heat waves and cold waves. *Urban Clim*. 2019;27:388-402.
38. Rao VD, Choudhury D. Probabilistic modelling for earthquake forecasting in the northwestern part of Haryana State, India. *Pure Appl Geophys*. 2020;177(7):3073-3087.
39. Forzieri G, Feyen L, Russo S, et al. Multi-hazard assessment in Europe under climate change. *Climatic Change*. 2016;137(1):105-119.

40. Ourloglou O, Stefanidis K, Dimitriou E. Assessing nature-based and cassical engineering solutions for flood-risk reduction in urban streams. *J Ecol Eng*. 2020;21(2):46-56.
41. Stuart J, Keogh R, Hughes L. 100 or 10000 year flood, who knows? Implications for dam, floodplain and emergency management. In: Floodplain Management Association National Conference. 2016.
42. Wang M, Takada T. Macrospatial correlation model of seismic ground motions. *Earthq. Spectra*. 2005;21(4):1137-1156.
43. Park J, Bazzurro P, Baker J. Modeling spatial correlation of ground motion intensity measures for regional seismic hazard and portfolio loss estimation. In: Proceedings of the 10th International Conference on Applications of Statistics and Probability in Civil Engineering ICASP10. 2007:579.
44. Pires JA, Ang AH-S, Villaverde R. Seismic reliability of electrical power transmission systems. *Nucl Eng Des*. 1996;160(3):427-439.
45. Hwang HHM, Huo J-R. Seismic fragility analysis of electric substation equipment and structures. *Probab Eng Mech*. 1998;13(2):107-116.
46. Straub D, Kiureghian AD. Improved seismic fragility modeling from empirical data. *Struct Saf*. 2008;30(4):320-336.
47. Zuccaro G, Gregorio D. Time and space dependency in impact damage evaluation of a sub-Plinian eruption at Mount Vesuvius. *Nat Hazards*. 2013;68(3):1399-1423.
48. Gokon H, Koshimura S, Imai K, Matsuoka M, Namegaya Y, Nishimura Y. Developing fragility functions for the areas affected by the 2009 Samoa earthquake and tsunami. *Nat Hazards Earth Syst Sci*. 2014;14(12):3231-3241.
49. Dueñas-Osorio L, Craig JI, Goodno BJ. Seismic response of critical interdependent networks. *Earthq Eng Struct Dyn*. 2007;36(2):285-306.
50. Ghosn M, Dueñas-Osorio L, Frangopol DM, et al. Performance indicators for structural systems and infrastructure networks. *J Struct Eng*. 2016;142(9):F4016003.
51. Poljanšek K, Bono F, Gutiérrez E. Seismic risk assessment of interdependent critical infrastructure systems: the case of European gas and electricity networks. *Earthq Eng Struct Dyn*. 2011;41(1):61-79.
52. Hernández-Fajardo I, Dueñas-Osorio L. Probabilistic study of cascading failures in complex interdependent lifeline systems. *Reliab Eng Syst Saf*. 2013;111:260-272.
53. Latora V, Marchiori M. Efficient behavior of small-world networks. *Phys Rev Lett*. 2001;87:198701.
54. Crucitti P, Latora V, Marchiori M. Model for cascading failures in complex networks. *Phys Rev E*. 2004;69:045104.
55. Scherb A, Garrè L, Straub D. Reliability and component importance in networks subject to spatially distributed hazards followed by cascading failures. *ASCE-ASME J Risk Uncertain Eng Syst B: Mech Eng*. 2017;3(2):021007.
56. Navarro-Espinosa A, Moreno R, Lagos T, et al. Improving distribution network resilience against earthquakes. In: Proceedings of the IET International Conference on Resilience of Transmission and Distribution Networks (RTDN). 2017:1-6.
57. Jayaram N, Baker JW. Deaggregation of lifeline risk: insights for choosing deterministic scenario earthquakes. In: Proceedings of the Technical Council on Lifeline Earthquake Engineering Conference. 2009:1051-1060.
58. Ouyang M. Comparisons of purely topological model, betweenness based model and direct current power flow model to analyze power grid vulnerability. *Chaos*. 2013;23(2):023114.
59. Cimellaro GP, Solari D, Bruneau M. Physical infrastructure interdependency and regional resilience index after the 2011 Tohoku Earthquake in Japan. *Earthq Eng Struct Dyn*. 2014;43(12):1763-1784.
60. Albert R, Albert I, Nakarado GL. Structural vulnerability of the North American power grid. *Phys Rev E*. 2004;69(2):025103(R).
61. Rosato V, Issacharoff L, Tiriticco F, Meloni S, Porcellinis S, Setola R. Modelling interdependent infrastructures using interacting dynamical models. *Int J Crit Infrastruct*. 2008;4(01):63-79.
62. Ferrario E, Poulos A, de la Llera JC, Lorca A, Oneto A, Magnere C. Representation and modeling of the Chilean electric power network for seismic resilience analysis. In: Proceedings of the 29th International European Safety and Reliability Conference ESREL 2019. 2019:3374-3381.
63. Prieur C, Tarantola S. Variance-based sensitivity analysis: theory and estimation algorithms. In: Ghanem R, Higdon D, Owhadi H, eds. *Handbook of Uncertainty Quantification*. Springer; 2017:1217-1240.
64. Esteva L, Villaverde R. Seismic risk, design spectra and structural reliability. In: Proceedings of the 5th World Conference on Earthquake Engineering (5WCEE). 1973;2:2586-2596.
65. Huang D, Allen TT, Notz WI, Miller RA. Sequential Kriging optimization using multiple-fidelity evaluations. *Struct Multidiscip Optim*. 2006;32(5):369-382.
66. Jalali H, Van Nieuwenhuyse I, Picheny V. Comparison of Kriging-based algorithms for simulation optimization with heterogeneous noise. *Eur J Oper Res*. 2017;261(1):279-301.
67. Bichon B, Mahadevan S, Eldred M. Reliability-based design optimization using efficient global reliability analysis. In: Proceedings of the 50th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference. 2009:2261.
68. Echard B, Gayton N, Lemaire M. AK-MCS: An active learning reliability method combining Kriging and Monte Carlo simulation. *Struct Saf*. 2011;33(2):145-154.
69. Rasmussen CE, Williams CKI. *Gaussian Processes for Machine Learning*. The MIT Press; 2006.
70. Bishop CM. *Pattern recognition and Machine Learning*. Springer; 2006.
71. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. *Number 57 in Monographs on Statistics and Applied Probability*. Chapman and Hall/CRC; 1993.
72. Picheny V, Wagner T, Ginsbourger D. A benchmark of Kriging-based infill criteria for noisy optimization. *Struct Multidiscip Optim*. 2013;48:607-626.
73. FEMA. *HAZUS MH MR4 Multi-hazard Loss Estimation Methodology – Earthquake Model*. 2003. Technical manual.

74. Yoon S, Mukherjee S, Hastak M. A Framework to Assess Natural Hazard Induced Service Inoperability in the Electricity Sector. In: Proceedings of the Canadian Society of Civil Engineering (CSCE) Conference. 2019.

75. Yu J, Kim JE, Lee JH, Kim TW. Development of a PCA-based vulnerability and copula-based hazard analysis for assessing regional drought risk. *KSCE J Civ Eng*. 2021;25:1901-1908.

How to cite this article: Rosero-Velásquez H, Straub D. Selection of representative natural hazard scenarios for engineering systems. *Earthquake Engng Struct Dyn*. 2022;1-21. https://doi.org/10.1002/eqe.3743