Received: 2 February 2024 | Revised: 3 May 2024 | Accepted: 8 June 2024

DOI: 10.1002/eqe.4184

RESEARCH ARTICLE

WILEY

# Damage-informed ground motion and semi-empirical fragility assessment

**Iunio Iervolino**[1,2] | **Annalisa Rosti**[3] | **Andrea Penna**[3] | **Massimiliano Giorgio**[1]

[1]Università degli Studi di Napoli Federico II, Naples, Italy

[2]IUSS – Scuola Superiore Universitaria di Pavia, Pavia, Italy

[3]Università degli Studi di Pavia, Pavia, Italy

**Correspondence**
Annalisa Rosti
Email: annalisa.rosti@unipv.it

Iunio Iervolino
Email: iunio.iervolino@unina.it

**Abstract**

Calibrating parametric fragility curves via empirical damage data is one of the standard approaches to derive seismic structural vulnerability models. Fragilities based on empirical data require the characterization of the ground motion (GM) intensity at the building sites in the area affected by the earthquake producing the observed damages. This is commonly conducted via ShakeMap, that is, a map of the expected values of a Gaussian random field (GRF) of the logarithms of a GM intensity measure conditional to magnitude, location, and possibly a set of recordings of the earthquake. Once that intensity and damage data at the same sites are available, the typical approach calibrates a two-parameter fragility model. However, ShakeMap estimates are affected by uncertainty deriving from that of the GM model used to characterize it. Furthermore, such an uncertainty can be reduced by building damage data, which provide information on the shaking intensity at the sites where damage is observed. It is shown herein that if this uncertainty is not addressed, also considering the shaking information provided by damage, the estimates of the fragility parameters obtained using a median ShakeMap only can be biased, and a recommended maximum likelihood estimation procedure – which exploits the *expectation maximization algorithm* – is provided. These arguments are illustrated via an application considering damage data from the 2009 L'Aquila earthquake in central Italy.

**KEYWORDS**

earthquake damage data, generalized linear models, ShakeMap, uncertainty, vulnerability

## 1 | INTRODUCTION AND MOTIVATION

One of the approaches to derive seismic fragility curves for structures is based on damage data observed in earthquakes. Although this approach inherently derives a vulnerability model describing the average behavior of a population of structures, typically referred to as a *typology*, it is seen as an alternative to the simulation-based approach that entails solving the equation of motion for nonlinear structural models representing specific constructions.[1] The observed-damage-based approach to derive fragility curves is sometimes considered especially appealing, because of its semi-empirical nature. Herein, *semi-empirical* refers to observed damage calibrating a parametric fragility model, to distinguish from *empirical* fragilities, which are non-parametric models.[2]

  

The derivation of fragility curves boils down to obtain the probability of observing some structural performance of interest (i.e., exceedance of a structural damage level) conditional to the values of a ground motion (GM) intensity measure (IM). In the probabilistic context, *conditional* means that the IM is assumed known with certainty. However, one of the distinctive features of this approach is that the GM intensity, at the sites where damage is observed, is – in most cases – not directly available. (This is not an issue in numerical fragility curves, which are developed subjecting the structural model to a sample of assigned GMs with known features.[3]) In the context of semi-empirical fragility, the IMs at the damage sites where seismic stations are missing are often obtained via the ShakeMap[4] developed for the earthquake which produced these damages.[4] The information employed to develop ShakeMap is classically made of earthquake magnitude and location, as well as recordings available at a few sites. In recent ShakeMap implementations,[4] the logarithms of IMs in the area affected by the earthquake are assumed to form a Gaussian random field (GRF). Therefore, the joint distribution of IMs is specified by the mean vector and the covariance matrix, which are obtained by a GM model, for example a ground motion prediction equation (GMPE) and a spatial correlation model for its intra-event residuals. Despite rules of risk analysis[5] requiring that no information is neglected in evaluating the fragility parameters, in most cases only the ShakeMap collecting the mean of the logarithms of the IMs, hereafter referred to as *median ShakeMap*, is considered in literature.[6–12]

Once it is recognized that intensities from ShakeMap are affected by uncertainty that should be accounted for in semi-empirical (also referred to as *observational*) fragility curve derivation, it must be acknowledged that damage data also provide information about the shaking intensity in the area where such damages are observed. Given earthquake magnitude and source-to-site distance, observing damage [no damage] on a building suggests that intensity at that site was more likely *large* [*low*] compared with the case in which the damage information at that site was not available. In other words, each building in an earthquake works as a seismic station in which the ground shaking record at the site is measured in terms of observed damage levels that – in turn – can be used to inform the ShakeMap for the event.

In general terms, the problem at hand consists in determining the fragility curve from a set of data where the shaking intensity at the damage sites are replaced by estimates. This is a problem known in statistics as *missing-data*,[13] that can be effectively addressed in a number of ways. One of them, particularly efficient on the computational side, is the maximum likelihood estimation via the expectation maximization (EM) algorithm. In this context, the scope of the study presented in the following was threefold: (i) to provide a EM-based procedure to account for shaking uncertainty in semi-empirical fragility calibration, also considering the information damage data provide to GM in the area of interest; (ii) to demonstrate that the parameters of the fragility curve obtained considering only the median ShakeMap – that is the typical approach in literature – can be different from those obtained considering shaking estimation uncertainty; and (iii) to show that GRF modelling of IMs, which includes the uncertainty in the GM model as informed only by the recordings available, as done in some studies,[14] without considering the information provided by damage, still leads to biased estimation of the fragility curves parameters.

The remainder of this paper is structured briefly recalling first the modelling of fragility based on logistic regression, which is considered especially suitable in the case of observed damages. Second, the conditional GRF approach to model GM intensity is recapped. Third, the question of why damage data informs the shaking estimates is addressed, as well as how to account for such information, in the context of maximum likelihood estimation of fragility parameters. Then, the procedure is applied to the case-study of fragility assessment based on a few thousand observations of damage data of reinforced concrete buildings affected by the 2009 L'Aquila earthquake in central Italy. The discussion of the results and the simulation-based validation of the approach precede some final remarks.

## 2 | LOGISTIC REGRESSION FOR FRAGILITY ASSESSMENT

When determining a semi-empirical fragility curve, data about a set of buildings hit by an earthquake are available. The performance of each building can be seen as the realization of a *Bernoulli* random variable (RV), where the outcome is one if the damage exceeds the threshold related to the damage state (DS) of interest, and zero otherwise; the exceedance of the damage threshold will be conventionally indicated hereafter as *failure*.

If the constructions all match a taxonomy, the observations of the performances of buildings experiencing the same IM can be seen as the outcomes of the repetition of the same random trial. If damage data are indicated as $\{y_1, y_2, \dots, y_k\}$, where $y_j$ is equal to 1 if there is damage in the $j$-th structure (i.e., building) and 0 otherwise, and the shaking intensities $\{im_1, im_2, \dots, im_k\}$ experienced by the $k$ buildings are also available, it is possible to link the probability of failure to the

GM, $\pi\,(im) = P[failure|IM = im]$, via, for example, a *generalized linear model* (GLM):[15]

$$\log\left[\frac{\pi\,(im)}{1-\pi\,(im)}\right] = \beta_0 + \beta_1 \cdot \log\,(im) \tag{1}$$

where $\{\beta_0, \beta_1\}$ are the parameters to be determined. The application of the model in Equation (1) is defined as *logistic regression*, and the fragility function obtained has the shape of the cumulative distribution function of a logistic RV:

$$\pi\,(im) = P\,[failure\,|IM = im\,] = \frac{1}{1+e^{-[\beta_0+\beta_1\cdot\log(im)]}}. \tag{2}$$

The logistic model will be considered in the following to represent the fragility, although the same reasoning applies to any semi-empirical model analogously. It is useful to recall the likelihood, $L(\beta_0, \beta_1)$, associated to this model, which is function of the sample $\{y_j, im_j\},\ j = 1,\ 2, \dots, k$:

$$L(\beta_0, \beta_1) = \prod_{j=1}^{k} [\pi(im_j)]^{y_j} \cdot [1-\pi(im_j)]^{1-y_j}, \tag{3}$$

where $\pi(im_j)$ is computed according to Equation (2). The corresponding log-likelihood is:

$$\log[L(\beta_0, \beta_1)] = \sum_{j=1}^{k} \{y_j \cdot \log[\pi(im_j)] + (1-y_j) \cdot \log[1-\pi(im_j)]\}. \tag{4}$$

In fact, in the context of semi-empirical fragility, the $\{y_1, y_2, \dots, y_k\}$ damage data are available, yet the $\{im_1, im_2, \dots, im_k\}$ intensity values at the building sites are not (usually for none of them), while estimates (e.g., via ShakeMap) are usually used instead. However, treating the estimates as the true intensity values leads to a biased assessment of the fragility parameters, that is, $\{\beta_0, \beta_1\}$, while the uncertainty on $\{im_1, im_2, \dots, im_k\}$ must be characterized based on all information available.

## 3 | CONDITIONAL GRF GROUND MOTION MODELLING AND SHAKEMAP

For seismic vulnerability assessment, the effect of earthquake $i$ at the sites, $s$ in number, where buildings are located, can be defined as the vector collecting the (logarithms of) the IMs it produces, $\{\log(IM_{1,i}), \log(IM_{2,i}), \dots, \log(IM_{s,i})\}$ at those sites. Most GMPEs model the logarithm of IM at the generic site $j = 1, \dots, s$ due to earthquake $i$, as:

$$\log(IM_{j,i}) = \mathrm{E}[\log(IM_{j,i})|m_i, r_{j,i}, \theta_j] + \eta_i + \varepsilon_{j,i} \tag{5}$$

where $E[\log(IM_{j,i})|m_i, r_{j,i}, \theta_j]$ is the mean of $\log(IM_{j,i})$ conditional to earthquake magnitude, $m_i$, source-to-site distance, $r_{j,i}$, and one or more other factors, such as the local soil site condition, indicated as $\theta_j$ (other factors, such as the faulting style, can also be represented by similar terms). The term $\eta_i$ is the inter-event residual, which is common to all sites in the $i$-th event, while $\varepsilon_{j,i}$ is the intra-event residual of the logarithms of IM at site $j$ in earthquake $i$. At each site, inter-event and intra-event residuals are assumed to be normally distributed stochastically independent RVs. Both have zero mean and variance denoted as $\sigma^2_{inter}$ and $\sigma^2_{intra}$, respectively. Thus, $\log(IM_{j,i})$, at site $j$ and conditional to $\{M = m_i\,, R = r_{j,i}\,\}$, is a Gaussian RV, with mean $E[\log(IM_{j,i})|m_i, r_{j,i}, \theta_j]$ and variance $\sigma^2 = \sigma^2_{inter} + \sigma^2_{intra}$.

By extension, state-of-the-art of multi-site probabilistic seismic hazard analysis[16] assumes that $\{\log(IM_{1,i}), \log(IM_{2,i}), \dots, \log(IM_{s,i})\}$ form a GRF; that is, the logarithms of the IMs jointly have a multivariate normal distribution conditional to the features of the $i$-th earthquake. The mean vector, $\{\mu\}$ and the covariance matrix, $[\Sigma]$

are:

$$\begin{cases} \{\mu\} = \{E[\log(IM_{1,i})\,|m_i, r_{1,i}, \theta_1\,], E[\log(IM_{2,i})\,|m_i, r_{2,i}, \theta_2\,], \dots, E[\log(IM_{s,i})\,|m_i, r_{s,i}, \theta_s\,]\}^T \\ [\Sigma] = \sigma^2_{inter} \cdot \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} + \sigma^2_{intra} \cdot \begin{bmatrix} 1 & \rho_{1,2} & \cdots & \rho_{1,s} \\ \rho_{2,1} & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{s,1} & \rho_{s,2} & \cdots & 1 \end{bmatrix} \end{cases} \tag{6}$$

where, $\rho_{j,h}$, is the correlation coefficient between intra-event residuals at two generic sites $\{j, h\}$, which is typically function of the sites' distance.[17] (In Equation (6), subscript $i$ was dropped because, according to most GMPEs, variances and covariances of the residuals do not depend on magnitude and location of the earthquake.)

The main advantage of the GRF assumption is that Equation (6) completely specifies the stochastic model of GM intensity in one earthquake and that the IMs at a subset of the sites, and/or conditional to the intensities at other sites, still follow a multivariate Gaussian model. If the IMs are known for $s-k$ of the sites ($k < s$), for example because some seismic monitoring stations recording the earthquake of interest were operating at the time of the event, then the IMs at the remaining $k$ sites, where damages are observed, still form a GRF. Renaming the IMs such as to split the vector of the logarithms to obtain two sub-vectors collecting the $k$ sites first and the $s-k$ sites then, Equation (6) reads:

$$\begin{cases} \{\mu\} = \left\{\{\mu\}_k, \{\mu\}_{s-k}\right\}^T \\ [\Sigma] = \begin{bmatrix} [\Sigma]_{k,k} & [\Sigma]_{k,s-k} \\ [\Sigma]_{s-k,k} & [\Sigma]_{s-k,s-k} \end{bmatrix} \end{cases}, \tag{7}$$

where, $\{\mu\}_k = \{E[\log(IM_{1,i})|m_i, r_{1,i}, \theta_1], \dots, E[\log(IM_{k,i})|m_i, r_{k,i}, \theta_k]\}^T$, while $\{\mu\}_{s-k} = \{E[\log(IM_{k+1,i})|m_i, r_{k+1,i}, \theta_{k+1}], \dots, E[\log(IM_{s,i})|m_i, r_{s,i}, \theta_s]\}^T$; $[\Sigma]_{k,k}$ and $[\Sigma]_{s-k,s-k}$ are the covariance submatrices referring to the former and latter subset of sites, respectively, while $[\Sigma]_{s-k,k} = [\Sigma]^T_{k,s-k}$ are the submatrices relating the two subsets. Then, the mean vector, $\{\mu\}'_k$, and covariance matrix, $[\Sigma]'_k$, for the $k$ sites, conditional on earthquake magnitude, location, and also to the measurements at the $s-k$ sites,* can be readily obtained by those in Equation (6):

$$\begin{cases} \{\mu\}'_k = \{\mu\}_k - [\Sigma]_{k,s-k} \cdot [\Sigma]^{-1}_{s-k,s-k} \cdot [\{\mu\}_{s-k} - \{\log(im)\}_{s-k}] \\ [\Sigma]'_k = [\Sigma]_{k,k} - [\Sigma]_{k,s-k} \cdot [\Sigma]^{-1}_{s-k,s-k} \cdot [\Sigma]_{s-k,k} \end{cases}, \tag{8}$$

where $\{\log(im)\}_{s-k} = \{\log(im_{k+1,i}), \log(im_{k+2,i}), \dots, \log(im_{s,i})\}^T$ is the vector collecting the IMs values available for the *s-k* sites providing the records for the *i*-th earthquake. This approach is exploited in the recent versions of ShakeMap,[4] providing $\{\mu\}'_k$ in the area affected by the earthquake. The covariance matrix in Equation (8) describes the uncertainty in the ShakeMap estimates. Such an uncertainty depends on that of the considered GM model (i.e., GMPE), although modified in virtue of the information provided by the $s-k$ sites where IM measurements are available.

## 4 | DAMAGE-INFORMED SEMI-EMPIRICAL FRAGILITY CALIBRATION VIA EXPECTATION MAXIMIZATION

The uncertainty around ShakeMap intensity values should be accounted for in the fragility calibration. This is because the median estimates (i.e., mean of the logarithms), from the first of Equation (8), is only one of the possible realizations of the random field, and to use the median IM values in Equation (1) neglects the fact that any other realization of the GRF would lead to a different estimate of $\{\beta_0, \beta_1\}$, which therefore, are factually subjected to the propagation, through the logistic regression, of the uncertainty in the (unobserved) IM values at the damage sites.

A second, equally relevant, issue is that the damages $\{y_1, y_2, \dots, y_k\}$ provide information about the $\{im_1, im_2, \dots, im_k\}$ values, thus affecting the uncertainty, in a very analogous manner as the earthquake magnitude, location, and the $\{im_{k+1}, im_{k+2}, \dots, im_s\}$ values at the recording stations. To understand this issue, the reader has to just consider the IM

occurred during a given earthquake at a given site. If no further information is provided, and if no recording stations were present at the site, such IM is a Gaussian RV specified as per Equation (8). However, if now one adds the information that at the site there was a building for which damage exceeded [not exceeded] a given threshold, then one must gain belief that the *im* occurred must have been relatively *strong* [*weak*] to cause [not to cause] such damage level.[†] In other words, the building experiencing or not experiencing the damage at the site has worked as a recording station measuring the IM in terms of observed structural response; that is, the $\{im_1, im_2, \dots, im_k\}$ values are actually *damage-informed*. In fact, due to the spatial correlation of the IMs in one given earthquake, also the damages at the other sites provide information about the IM at the site of interest. Such information, must be thus considered in the fragility calibration to get a probabilistically consistent evaluation of the fragility parameters $\{\beta_0, \beta_1\}$ of Equation (1). A straightforward way to consistently account for these issues in the fragility calibration is to maximize the following likelihood function:

$$L(\beta_0, \beta_1) \propto \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \left\{ \prod_{j=1}^{k} [\pi(im_j)]^{y_j} \cdot [1 - \pi(im_j)]^{1-y_j} \right\} \cdot f_{IM_1,\dots,IM_k | IM_{k+1},\dots,IM_s}(im_1, \dots, im_k \,|\, im_{k+1}, \dots, im_s) \cdot d(im_1) \dots \cdot d(im_k). \tag{9}$$

In the equation, which reduces to Equation (3) when the IMs at the building sites are known, $f_{IM_1,\dots,IM_k | IM_{k+1},\dots,IM_s}$ is the joint distribution at the $k$ sites, conditional on the magnitude and location of the earthquake, as well as the data from the available recording stations. In the considered context, this latter distribution, when passing to the logarithms of IMs, becomes Gaussian with mean vector and covariance matrix by Equation (8). However, as illustrated in the application, the sites of interest can be some thousands and therefore to determine $\{\beta_0, \beta_1\}$, directly maximizing Equation (9), can be computational challenging, if not unfeasible. A computationally efficient strategy that literature has already proven to provide the same result as the direct maximization of the likelihood function is by the EM algorithm,[13] the application of which requires to distinguish available and missing data: the available data are $\{y_1, y_2, \dots, y_k, im_{k+1}, \dots, im_s\}$ and those missing are $\{im_1, \dots, im_k\}$.[‡] The EM algorithm implemented here is an iterative procedure based on sequential sampling. Its steps are listed in the following.

1. Assign tentative fragility parameters $\{\beta_0^{(z)}, \beta_1^{(z)}\}$, where $z = 0$ indicating the first iteration of the procedure.
2. Select a number $h \geq 1$ (the lower $h$ is, the better) of sites among the $k$ damage sites (for which intensity is not available), and sample $l$ times the GM distribution $f_{IM_1,\dots,IM_h | IM_{k+1},\dots,IM_s}$, which is the joint IM distribution conditional to magnitude, location, and the available earthquake measurements (that is corresponding to a GRF model characterized by equations analogous to (8)). This enables obtaining $l$ realizations of the IMs at the $h$ selected sites:

$$\{im_1, \dots, im_h\}_i, \; i = 1, 2 \dots, l. \tag{10}$$

For each of these samples, compute the following weight (where $r$ is not to confuse with source-to-site distance):

$$r_i\left(\beta_0^{(z)}, \beta_1^{(z)}\right) = \frac{f_{Y_1,\cdots,Y_h | IM_1,\cdots,IM_h}\left(y_1, \cdots, y_h | \{im_1, \cdots, im_h\}_i\right)}{\sum_{i=1}^{l} f_{Y_1,\cdots,Y_h | IM_1,\cdots,IM_h}\left(y_1, \cdots, y_h | \{im_1, \cdots, im_h\}_i\right)}, \; i = 1, 2, \cdots, l, \tag{11}$$

where $f_{Y_1,\dots,Y_h | IM_1,\dots,IM_h}(y_1, \dots, y_h | \{im_1, \dots, im_h\}_i) = \prod_{j=1}^{h} [\pi(\{im_j\}_i)]^{y_j} \cdot [1 - \pi(\{im_j\}_i)]^{1-y_j}$; that is, the fragility likelihood for the selected sites, when $\{\beta_0^{(z)}, \beta_1^{(z)}\}$ are the parameters. Use the weights $r_i(\beta_0^{(z)}, \beta_1^{(z)})$, $i = 1, 2, \dots l$ to resample with replacement $\{im_1, \dots, im_h\}_i, i = 1, 2 \dots l$, to obtain a population where each of them is represented proportionally to its weight.
3. Select the next $h$ damage sites (different to those in step #2), and generate $l$ realizations of the IM at these sites sampling from the distribution $f_{IM_{h+1},\dots,IM_{2\cdot h} | IM_1,\dots,IM_h,IM_{k+1},\dots,IM_s}(im_{h+1}, \dots, im_{2\cdot h} | \{im_1, \dots, im_h\}_i, im_{k+1}, \dots, im_s)$. When passing to the logarithms, this is a GRF conditional to magnitude, location, and measurements of the earthquake, as well as the simulated (and resampled) data at the first $h$ sites from step #2, which account for the effect of damage. This enables obtaining:

$$\{im_{h+1}, \dots, im_{2\cdot h}\}_i, \; i = 1, 2 \dots, l. \tag{12}$$

For these realizations, compute the weights in analogy with Equation (11):

$$r_i\left(\beta_0^{(z)},\beta_1^{(z)}\right)=\frac{f_{Y_{h+1},\cdots,Y_{2\cdot h}|IM_{h+1},\cdots,IM_{2\cdot h}}\left(y_{h+1},\cdots,y_{2\cdot h}|\{im_{h+1},\cdots,im_{2\cdot h}\}_i\right)}{\sum_{i=1}^{l}f_{Y_{h+1},\cdots,Y_{2\cdot h}|IM_{h+1},\cdots,IM_{2\cdot h}}\left(y_{h+1},\cdots,y_{2\cdot h}|\{im_{h+1},\cdots,im_{2\cdot h}\}_i\right)},\ i=1,2,\cdots,l. \tag{13}$$

Append the vectors $\{im_{h+1},\dots,im_{2\cdot h}\}_i$ to the vectors $\{im_1,\dots,im_h\}_i$, so as to obtain $\{im_1,\dots,im_{2\cdot h}\}_i$, $i=1,2,\dots l$, and use the weights just computed to resample them with replacement. This will lead to a population of $l$ IMs realizations at the $2\cdot h$ sites where each of them is represented proportionally to the just computed weights.

4. In analogy with steps #2 and #3, generate $l$ realizations of the IMs at the next $h$ sites sampling $f_{IM_{2\cdot h+1},\dots,IM_{3\cdot h}|\{IM_1,\dots,IM_{2\cdot h}\}_i,IM_{k+1},\dots,IM_s}(im_{2\cdot h+1},\dots,im_{3\cdot h}|\{im_1,\dots,im_{2\cdot h}\}_i,im_{k+1},\dots,im_s)$. Then compute the weights for these new realizations and resample with replacement the vectors $\{im_1,\dots,im_{3\cdot h}\}_i$, $i=1,2,\dots l$. This sequential sampling procedure proceeds until all $k$ damage sites are considered and $l$ realizations of the GM fields at the damage sites, $\{im_1,\dots,im_k\}_i$, are obtained.
5. Formulate the likelihood of all (available and missing) data, $L_{C,i}(\beta_0,\beta_1)$, $i=1,2,\dots l$, as:

$$L_{C,i}(\beta_0,\beta_1)\propto f_{IM_1,\dots,IM_k|IM_{k+1},\dots,IM_s}\left(\{im_1,\dots,im_k\}_i\,|im_{k+1},\dots,im_s\right)\cdot\prod_{j=1}^{k}\left[\pi\left(\{im_j\}_i\right)\right]^{y_j}\cdot\left[1-\pi\left(\{im_j\}_i\right)\right]^{1-y_j}. \tag{14}$$

6. Compute the conditional expectation, $Q(\beta_0,\beta_1|\beta_0^{(z)},\beta_1^{(z)})$, of the log-likelihood with respect to missing data, given the available data:

$$Q\left(\beta_0,\beta_1\left|\beta_0^{(z)},\beta_1^{(z)}\right.\right)\approx\frac{1}{l}\cdot\sum_{i=1}^{l}\log\left(L_{C,i}\right). \tag{15}$$

7. Maximize $Q(\beta_0,\beta_1|\beta_0^{(z)},\beta_1^{(z)})$, with respect to $\{\beta_0,\beta_1\}$, to obtain a new set of tentative estimates $\{\beta_0^{(z+1)},\beta_1^{(z+1)}\}$:

$$\left\{\beta_0^{(z+1)},\beta_1^{(z+1)}\right\}=\underset{\{\beta_0,\beta_1\}}{\operatorname{argmax}}\left[Q\left(\beta_0,\beta_1\left|\beta_0^{(z)},\beta_1^{(z)}\right.\right)\right]. \tag{16}$$

8. Once $\{\beta_0^{(z+1)},\beta_1^{(z+1)}\}$ are found, one can return to step #2 and continue until the absolute value of the relative difference between the fragility parameters (or likelihood estimates) in two subsequent iterations is less than a tolerance value. Figure 1 provides an indicative flowchart of the EM algorithm as implemented herein.

It is to note that steps from #2 to #4 represent sequential sampling so as to generate realizations of GM compatible with the damage as represented by the fragility function with the candidate parameters. This is not a strictly necessary approach, and the EM algorithm can be applied generating $l$ samples at all $k$ damage sites at once, directly by Equation (8), at the beginning of each iteration of the procedure. Then the weight of each realization is computed based on the fragility function adopted and the GRF. These weights provide a measure of the compatibility of each realization with respect to damage. Nevertheless, it is easy to prove that the procedure based on sequential sampling is computationally more efficient, which may be necessary in the case at hand where buildings/sites amount to thousands.

## 5 | ILLUSTRATIVE APPLICATION

### 5.1 | Damage data and earthquake information

The case study of the building damage in the 2009 L'Aquila earthquake (moment magnitude 6.1), in central Italy, is considered. The damage surveyed in this earthquake has been used in literature to derive semi-empirical fragility curves already.[8] The damage data are available from the *Observed Damage Database* or Da.D.O.[18] The L'Aquila damage database includes about forty thousand residential buildings, from municipalities with *survey completeness ratio* (i.e., number of surveyed buildings over the total number of residential buildings from national building census) larger than ninety percent,

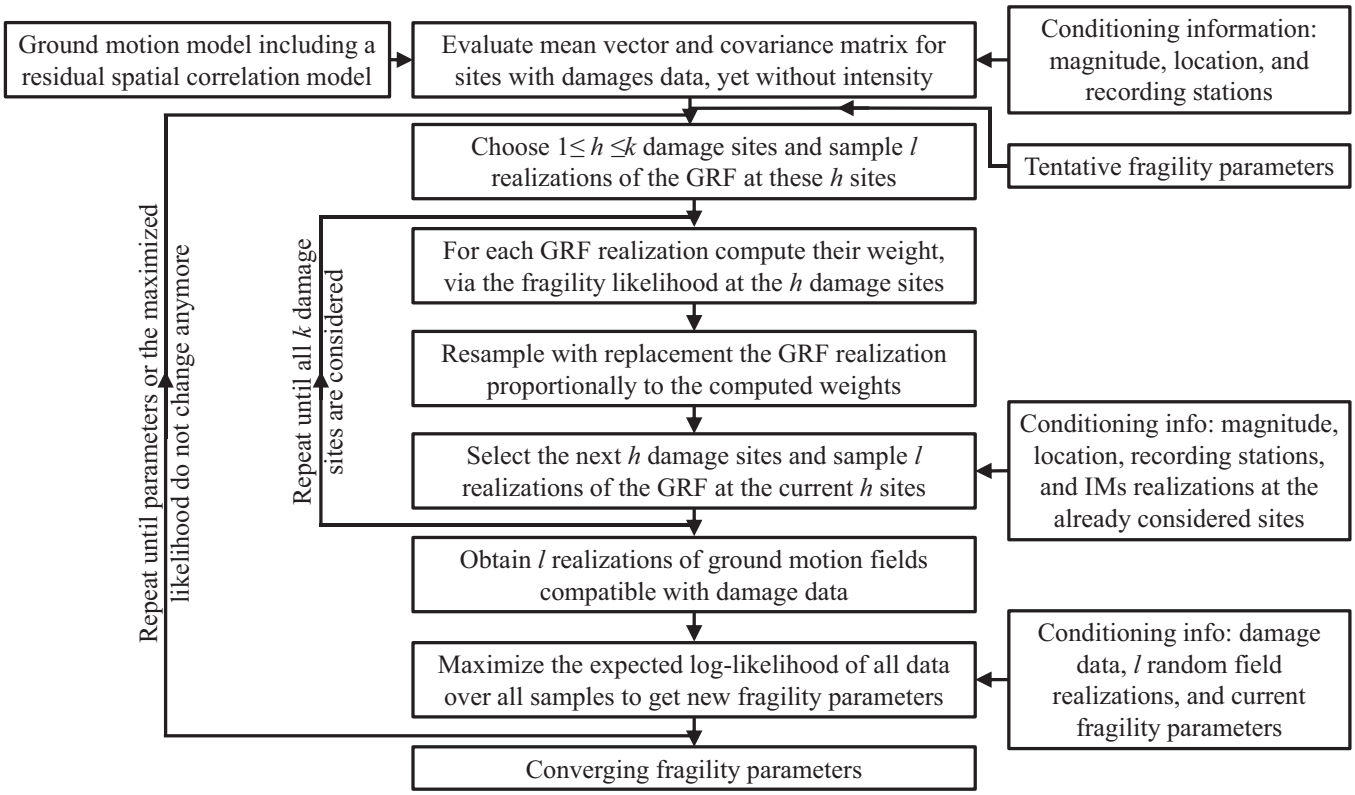


**FIGURE 1** Flowchart sketch for the determination of the distribution of the fragility parameters considering ShakeMap uncertainty and the informative effect of damage on GM estimates.

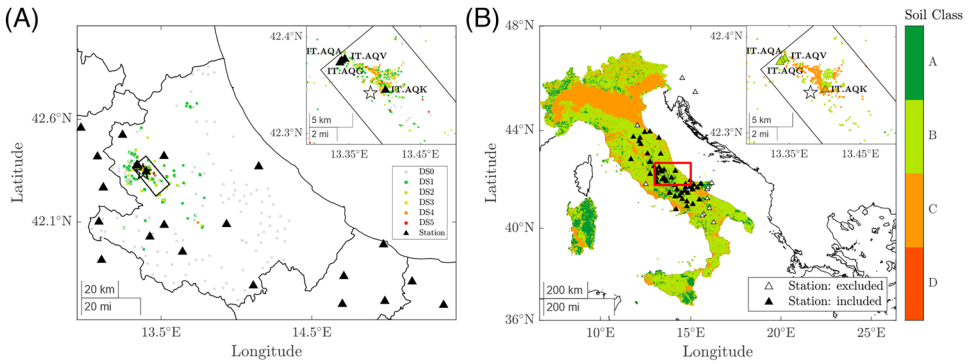


**FIGURE 2** (A) Considered damage data in the Abruzzo region, along with regional borders, source (rectangle) and epicenter (star) of the 2009 L'Aquila earthquake. (B) Stations recording the earthquake used to inform the ShakeMap and local site conditions according to a large-scale geological model (ID from https://itaca.mi.ingv.it/, last accessed Jan. 2024).

complemented by undamaged buildings from non-surveyed and partially surveyed municipalities in the Abruzzo region, the most affected by the earthquake.

This application considers data referring to a single residential building typology, that is, mid-rise (three and four storeys) reinforced concrete (RC) buildings, designed for seismic actions before 1981; that is, an obsolete code with respect to the one currently enforced in the country. Such damage data are 5247 and distributed among 1058 sites, as shown in Figure 2(A). Each building is assigned a level of damage compatible with the *European Macroseismic Scale*),[19] accounting for what observed on both the vertical structure and masonry infills/partitions.[8] (Different DS' at the same site are represented by overlapping markers in the figure.)

Figure 2(B) shows the recording stations conditioning the ShakeMap for the earthquake, they are sixty-four in number. In the figure, the local site conditions – according to Eurocode 8[20] – from a geology-based model used by the INGV implementation of ShakeMap are also given; they are used to assign local conditions to recording and damage sites.

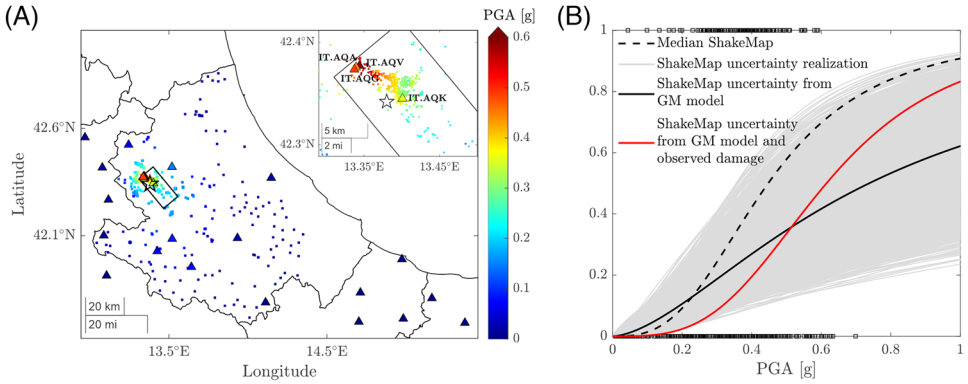


**FIGURE 3** (A) Median PGA, at damage sites, conditional to magnitude, location and IMs recorded close to the source of the earthquake. (B) Fragilities obtained using different levels of information.

## 5.2 | Ground motion model

Peak ground acceleration (PGA) is the selected IM to describe fragility. The GMPE to describe the GRF in the area is that of Bindi et al.,[21] complemented by the spatial correlation model of intra-event residual of Esposito and Iervolino.[17] PGA measurements from fifty-one, out of sixty-four stations mentioned in the previous section, were considered herein. Two stations were excluded as it was not possible to retrieve recorded data, while eleven stations with *Joyner & Boore distance*[22] exceeding 200 km were also further ruled out, as beyond the applicability limit of the selected GMPE. Figure 3(A) shows the median PGAs, according to the chosen GMPE (considering normal faulting style for the event), at the sites with damage data used to compute the fragility function (to follow);[§] the figure also shows the PGAs at the stations in the epicentral region (i.e., some of those in Figure 2B). In other words, it is the geographical representation of the conditional mean vector $\{\mu\}'_k$ in Equation (8), which is equivalent to a (median) ShakeMap.

Figure 3(B) shows the calibrated fragility function of Equation (2) obtained from applying the logistic regression in Equation (1) to the failure data, in terms of exceedance of damage level DS3 (black dashed line) from Figure 2(A) and the median ShakeMap estimates of Figure 3(A), that is, the most common approach in literature. The figure also shows the dots representing the realizations of failure data (zeros and ones) determining the logistic regression using median ShakeMap. The other curves in the figure will be discussed later.

## 5.3 | Propagation of ShakeMap uncertainty in fragility calibration

Given the damage data for the buildings at the sites, and the conditioning GM information that is, earthquake magnitude, location and recorded shaking at the stations, it is possible to appreciate how the ShakeMap uncertainty, represented by the GRF of Equation (8) alone, reflects in the fragility assessment, which is also found in literature.[14] In fact, one may fit a fragility curve for any realization of the GRF according to the following procedure.

a. Evaluate the conditional mean vector of the IMs at the damage sites via the GMPE according to Equation (8).
b. With the GMPE and the spatial correlation model for intra-event residuals, evaluate the conditional covariance matrix according to Equation (8). At this point the conditional GRF for the earthquake is fully characterized.
c. Generate a realization of the random field of the (logarithms) of PGA at the sites with damages sampling a multivariate Gaussian function featuring mean and covariance matrix evaluated at steps *a* and *b*.
d. Use the damage data (i.e., a given building performance of interest) and the realization of the GRF from step *c* to fit a fragility model according to a parametric model, for example, the GLM in Equation (1). This yields a realization of the parameters, $\{\beta_0, \beta_1\}_i$, corresponding to the realization of the GRF used for the fit.
e. Repeat the last two steps as many times as deemed sufficient to characterize the variability of the parameters $\{\beta_0, \beta_1\}_i$ reflecting ShakeMap uncertainty.

As an example, the results of the simulations are given for the L'Aquila data, referring to DS3, in Figure 3(B), where the gray curves are the results of the fragility calibration per each simulated GRF (ten thousand simulations are represented

**TABLE 1** Number of buildings exceeding each damage level in the data and fragility parameters by different levels of information: (i) median ShakeMap only, (ii) uncertainty in ShakeMap, (iii) uncertainty in ShakeMap, and observed damages.

| | No. of buildings exceeding DS | Median ShakeMap | | ShakeMap uncertainty | | Damage-informed GM | |
|---|---|---|---|---|---|---|---|
| | | $\beta_0$ | $\beta_1$ | $\beta_0$ | $\beta_1$ | $\beta_0$ | $\beta_1$ |
| **DS1** | 1345 | 5.223 | 2.932 | 3.668 | 2.124 | 5.183 | 3.045 |
| **DS2** | 696 | 3.112 | 2.72 | 1.497 | 1.713 | 3.071 | 2.967 |
| **DS3** | 390 | 2.289 | 2.846 | 0.497 | 1.633 | 1.605 | 3.302 |
| **DS4** | 164 | 0.747 | 2.492 | −0.772 | 1.464 | 0.651 | 3.746 |
| **DS5** | 30 | −1.811 | 1.858 | −2.813 | 1.241 | −1.252 | 3.102 |

in the figure). In other words, the gray curves represent the population of fragility curves obtained considering ShakeMap uncertainty. Because each curve corresponds to a pair of parameters $\{\beta_0, \beta_1\}_i$, Figure 3(B) also provides the fragility curve (i.e., the continuous black line) featuring the mean parameters of the curves from the simulation. The difference between the dashed and solid line curves is somewhat evident. Nevertheless, the solid line still does not contemplate the effect of the damage information on the GRF.

## 5.4 | EM-based fragility

To also include the information provided by damage to the IM random field in the fragility parameters calibration, the procedure outlined in Figure 1 was implemented in MATHWORK-MATLAB® (ver. 2022b).[¶] The number of realizations of IMs' random fields was set equal to $l = 10^4$. The stopping rule for the EM algorithm was the absolute value of the relative differences in the fragility parameters, in two successive iterations, lower than 0.01. The resulting fragility is given as the solid red line in Figure 3(B). It is apparent the difference with both the dashed black line, which is calibrated considering only the median ShakeMap, and the solid black curve, which considers the ShakeMap uncertainty in terms of the GRF determined by the GM model and the recording stations only, that is what discussed in the previous section, yet neglecting damage information. The differences in the fragility curves show that neglecting the information on GM provided by damages determines a bias in both $\beta_0$ and $\beta_1$, with the latter reciprocally related to the standard deviation of the logistic model, that is the slope of the fragility curve, as also discussed in the following.

Figure 4 shows the results of the EM algorithm for all damage states observed in the data (see Figure 2A). (The starting tentative parameters for each DS were obtained based on the EM algorithm without the sequential sampling considering $10^4$ random field realizations of the model in Equation (8)). Each row in the figure refers to one DS (including DS3, already shown), with the leftmost panel being similar to Figure 3(B) and providing: (i) the damage-informed fragility curve (red), (ii) the fragility curve obtained considering ShakeMap uncertainty only (continuous black line) obtained averaging the fragility parameters from the GRF samples (i.e., the gray curves in Figure 3(B), which are omitted for readability), and (iii) the fragility based on median ShakeMap estimates (dashed black line). The same issues discussed for DS3 in Figure 3(B), are observed for all damage states. From Figure 4 it is apparent that the red curves (i) are different than the corresponding case (iii) and are steeper than case (ii). Both issues, which are more evident for the most severe DS', are because of the influence of damage information. The fragility parameters for all discussed cases are collected in Table 1, where the number of buildings exceeding each DS are also reported.

To investigate how the damage informs the GM, Figure 4 features two additional panels for each DS. More specifically, the average IMs at each of the damage sites, are obtained averaging the $l = 10^4$ GM field realizations at convergence of the EM algorithm. These average PGAs, which are compatible with the observed DS according to the algorithm, are divided by the PGA from the median ShakeMap at the same site. Then these ratios are separated considering the sites where only buildings with damage lower than the considered DS are present, from those where at least one building with damage equal, or larger than the considered DS, is present. This criterion is to distinguish *no damage sites* from *damage sites*. The resulting two PGA ratios distributions are shown in the central and right panels of Figure 4 (vertical lines are the distribution medians). It is apparent that the PGAs at the damage sites are generally larger than those from the median ShakeMap, indicating that the damage suggests intensity values larger than those estimated without such an information. For the sites with no damage, the computed ratios tend to be lower than those at the sites with damages. The separation between the two distributions increases with the severity of the DS.

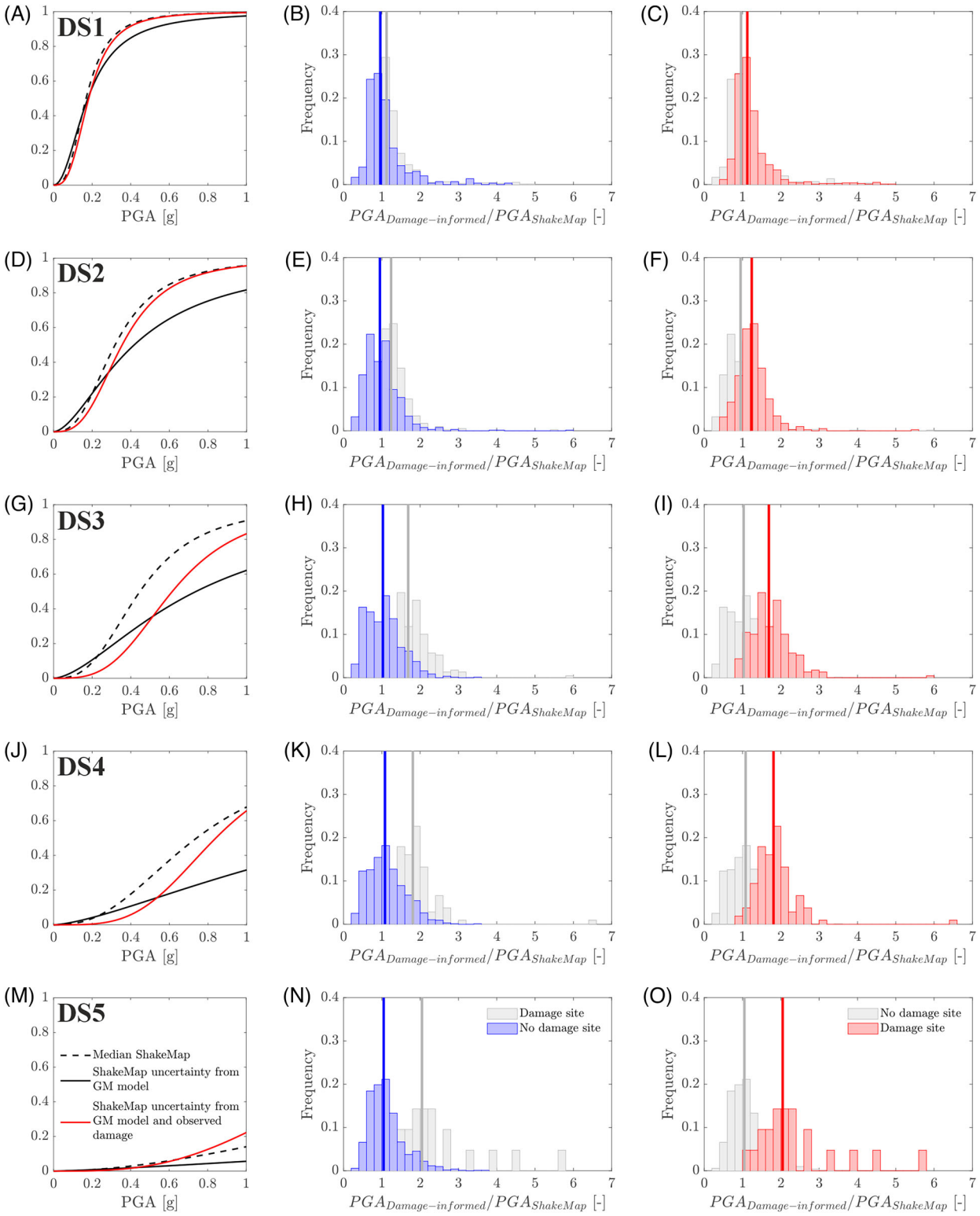


**FIGURE 4** (Left panels) Fragilities obtained considering: (i) the median ShakeMap (dashed black curve), (ii) uncertainty in ShakeMap only, averaging the parameters from ten thousand GRF simulations (continuous black line), (iii) effect of the information provided by damage on GM (red curve). (Center panels) Distribution of the expected PGAs from the EM algorithm divided by the median ShakeMap estimates at sites without damage. (Right panels) Distribution of the ratios of the expected PGAs from the EM algorithm and the median ShakeMap estimates at the sites with at least one damage. Grey histograms in the no damage [damage] sites panels represent the damage [no damage] sites distribution, for comparison.

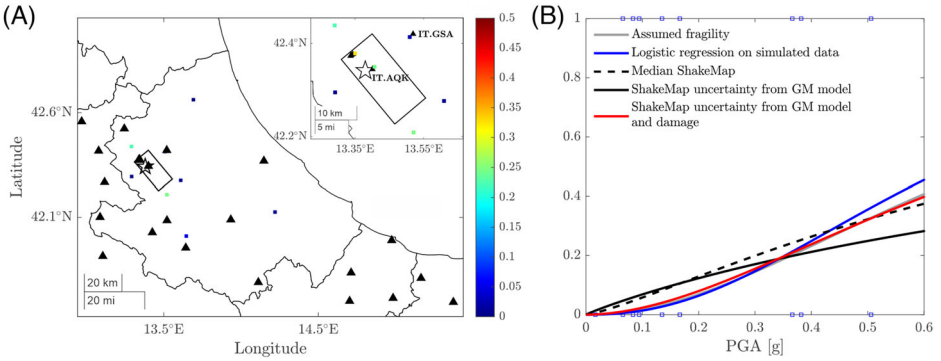


**FIGURE 5** (A) Selected ten sites and simulated damage data. (b) Fragility curves obtained for this case of simulated data.

## 6 | VALIDATION

The algorithm validation can be carried out assuming a known fragility model. To reduce the computational effort, a fictitious example was set up, where only ten sites are considered (shown in Figure 5A) among those in the illustrative application of Section 5, and at each site five-hundred buildings are located. Then, the validation example refers to DS3 and proceeds as follows:

1. the fragility parameters $\{\beta_0 = 0.659,\ \beta_1 = 2.024\}$ are assumed for DS3, and five thousand PGA values for the fictitious buildings are sampled from this model;
2. for the selected sites, a GRF realization of PGA values is generated via the model of Equation (8) adjusted for this case, representing the *true* IMs the considered sites have experienced in the earthquakes;
3. the damage data are simulated comparing, for each building at each site, the PGA values from step #1 and those from step #2: damage occurs if the PGA from the GRF at each site is larger than the one from the fragility, and does not occur otherwise; the simulation results are shown in Figure 5(A), where the color of the markers of the sites represents the fraction, over five-hundred, of the number of buildings at that site with damage equal to DS3 or worse;
4. the logistic regression in Equation (1) is applied to the simulated damage data and the fragility parameters are obtained $\{\beta_0 = 0.983,\ \beta_1 = 2.278\}$; such a curve, shown as the blue line in Figure 5(B), does not coincide with the assumed fragility because of estimation uncertainty (damages are shown as squares)[2];
5. then the EM algorithm is applied to the simulated damage data via $10^5$ random field realizations, as in Section 5, and the damage-informed fragility curve is obtained;
6. the fragilities considering the median ShakeMap, and only the uncertainty in the GM model, are also obtained for the same simulated damage data.

The resulting curves are given in Figure 5(B). The EM algorithm, considering a relative difference between parameters in successive iterations lower than 0.001 as the convergence criterion, returns a fragility $\{\beta_0 = 0.531,\ \beta_1 = 1.849\}$. The obtained fragility is similar to the logistic regression of the simulated data, which is the reference one for this exercise, as it is obtained from the damage data if the IMs would be known at the sites, and performs apparently better than considering the median ShakeMap or the uncertainty in the GM model only. Nevertheless, the algorithm does not exactly returns the fragility curve fitted on the simulated data as there is always residual uncertainty in the IMs that caused the damages the EM can reduce, yet not eliminate. (Repetitions of this damage simulation, not shown herein for brevity, show consistent results.)

## 7 | FINAL REMARKS

Semi-empirical fragility curves, obtained calibrating a parametric model via damage observations and GM intensity data from earthquakes, are a common tool used to describe the seismic vulnerability of the built environment. Because, in most of cases, the intensity recorded in the earthquake at the sites where damage data are collected is not available, such an information is surrogated by ShakeMap estimates. In the typical approach, the uncertainty affecting these estimates is neglected. The study presented herein aimed at investigating the effect of such an uncertainty on the semi-empirical

fragility curves. It benefitted from the fact that the current ShakeMap approach describes the random field of (the logarithms of) shaking intensity as a multivariate Gaussian distribution conditional to earthquake magnitude, location, and possibly recordings by some seismic monitoring stations. The mean vector and the covariance matrix of such distribution can be characterized via a GM prediction equation and a spatial correlation model of its intra-event residuals. Most importantly, the information the damage provides to the GM is also considered for the fragility parameters' estimation, via the expectation-maximization (EM) algorithm based on sequential sampling.

Using the data from the 2009 L'Aquila earthquake (central Italy), already used to derive semi-empirical fragility curves, fragility curves for different damage states were derived considering increasing level of information on the GM estimates: (i) median ShakeMap only; (ii) uncertainty in ShakeMap from the GM model and available recordings in the area; and (iii) uncertainty in ShakeMap from the GM model, available recordings, and observed damages. It was found that:

- to consider only the median ShakeMap, that is (i), the most common approach in the literature, leads to bias in the estimation of fragility parameters with respect to case (iii), that is, when modelling of GM intensity considers all the information available;
- considering only the information provided by the GM model and shaking recordings (ii), which is also found in literature, also causes bias with respect to (iii);
- the effect of damage information on GM is such that the ratio of the expected IMs at the sites with damages and what predicted from the median ShakeMap tends to be larger than one and larger than the same ratio computed at the sites without damage, an effect that increases with the severity of the damage state.

To validate the procedure, damage data at a few sites were simulated from an assumed fragility model showing that the algorithm returns a curve similar to the fragility fitted on the simulated data, and performs better than the fragility curve assuming only the median ShakeMap or only the uncertainty in the GM model, even if the EM-based fragility remains affected by some – inevitable – residual uncertainty on the IMs that caused the damages.

Although left out from the scope of this study for simplicity, it is to remark that specific damage states (rather than exceedance of damage levels as per common fragility curves), and damages to different typologies in the same earthquake, can inform the fragility calibration of a given typology (e.g., damages to masonry buildings can inform RC fragility in the same area). Moreover, in this work, fragilities for different damage states were treated separately, whereas an approach which would further exploit the available data would be to estimate the parameters of all fragilities jointly, as this would account for the exact damage state of each building rather than just using damage state exceedance information. Finally, following the logic of this study, it can be argued that also ShakeMap for a given earthquake can be revised after the damages are surveyed; this can be addressed maximizing a likelihood function that also includes, as arguments, the GM model parameters.

## ORCID

*Iunio Iervolino* https://orcid.org/0000-0002-4076-2718
*Annalisa Rosti* https://orcid.org/0009-0009-6896-8417
*Andrea Penna* https://orcid.org/0000-0001-6457-7827
*Massimiliano Giorgio* https://orcid.org/0000-0002-5348-5289

## ENDNOTES

*The uncertainty in the measurement of the recording stations, is neglected, yet could be also accounted for. This also applies to the magnitude attributed to an earthquake, which is also characterized by uncertainty.

†Building damage attribution is considered unaffected by uncertainty, which is the typical approach in semi-empirical seismic fragility literature. Nevertheless, uncertain damage attribution could also be considered within the approach pursued herein.

‡The procedure is written considering one building per site for simplicity of notation. Nevertheless, according to equation (8), if multiple buildings share the same site, the IMs correlation of the two is perfect and their IM is the same, and the algorithm perfectly contemplates the (actual) case of multiple buildings at each site.

§The map is in terms of maximum PGA of the horizontal components. Because the considered GMPE defines PGA in terms of geometric mean of the horizontal components, the empirical correction by Beyer & Bommer[23] is employed for the conversion.

¶In fact, even in the context of the EM algorithm, the weights, $r_i$, can be hard to evaluate numerically, due to the considerable size of the building dataset, a common issue in computing likelihoods for large data samples. The artifact of introducing a scale factor, suitably applied not to alter the results, enabled to overcome the issue.

## ACKNOWLEDGMENTS

The presented study was developed within the framework of the *Rete dei Laboratori Unversitari di Ingegneria Sismica* (ReLUIS) within the 2022–2024 ReLUIS-DPC research program (WP4).

## DATA AVAILABILITY STATEMENT

All data used and the software developed (by Annalisa Rosti) in this study are available to the reader upon request.

## REFERENCES

1. Baraschino R, Baltzopoulos G, Iervolino I. R2R-EU: software for fragility fitting and evaluation of estimation uncertainty in seismic risk analysis. *Soil Dyn Earthquake Eng*. 2020;132:106093. doi:10.1016/j.soildyn.2020.106093
2. Iervolino I, Assessing uncertainty in estimation of seismic response for PBEE. 2017;46(10):1711-1723. doi:10.1002/eqe.2883
3. Iervolino I, Manfredi G. A review of ground motion record selection strategies for dynamic structural analysis. In: Bursi O, Wagg DJ, eds. CISM International Centre for Mechanical Sciences, Courses and Lectures. *Vol 502*. Springer; 2008:131-163. doi:10.1007/978-3-211-09445-7_3
4. Worden BC, Thompson EM, Baker JW, Bradley BA, Luco N, Wald DJ. Spatial and spectral interpolation of ground-motion intensity measure observations. *Bull Seismol Soc Am*. 2018;108(2):866-875. doi:10.1785/0120170201
5. Erto P, Giorgio M, Iervolino I. About knowledge and responsibility in probabilistic seismic risk management. *Seismol Res Lett*. 2016;87(5):1161-1166. doi:10.1785/0220160001
6. Rosti A, Rota M, Penna A. An empirical seismic vulnerability model. *Bull Earthquake Eng*. 2022;20(8):4147-4173. doi:10.1007/S10518-022-01374-3/FIGURES/13
7. Giordano N, De Luca F, Sextos A, Ramirez Cortes F, Fonseca Ferreira C, Wu J, Empirical seismic fragility models for Nepalese school buildings. *Natural Hazards*. 2021;105(1):339-362. doi:10.1007/S11069-020-04312-1/TABLES/4
8. Rosti A, Del Gaudio C, Rota M, et al. Empirical fragility curves for Italian residential RC buildings. *Bull Earthquake Eng*. 2021;19(8):3165-3183. doi:10.1007/S10518-020-00971-4/TABLES/3
9. Del Gaudio C, De Martino G, Di Ludovico M, et al. Empirical fragility curves for masonry buildings after the 2009 L'Aquila, Italy, earthquake. *Bull Earthquake Eng*;17(11):6301-6330. doi:10.1007/S10518-019-00683-4/TABLES/8
10. Del Gaudio C, De Martino G, Di Ludovico M, et al. Empirical fragility curves from damage data on RC buildings after the 2009 L'Aquila earthquake. *Bull Earthquake Eng*. 2017;15(4):1425-1450. doi:10.1007/S10518-016-0026-1/FIGURES/16
11. Ceroni F, Casapulla C, Cescatti E, Follador V, Prota A, da Porto F. Damage assessment in single-nave churches and analysis of the most recurring mechanisms after the 2016–2017 central Italy earthquakes. *Bull Earthquake Eng*. 2022;20(15):8031-8059. doi:10.1007/S10518-022-01507-8/FIGURES/21
12. Di Ludovico M, Cattari S, Verderame G, et al. Fragility curves of Italian school buildings: derivation from L'Aquila 2009 earthquake damage via observational and heuristic approaches. *Bull Earthquake Eng*. 2023;21(1):397-432. doi:10.1007/S10518-022-01535-4/FIGURES/25
13. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Royal Statist Soc B (Methodol)*. 1977;39(1):1-22. doi:10.1111/J.2517-6161.1977.TB01600.X
14. Miano A, Jalayer F, Forte G, Santo A. Empirical fragility assessment using conditional GMPE-based ground shaking fields: application to damage data for 2016 Amatrice Earthquake. *Bull Earthquake Eng*. 2020;18(15):6629-6659. doi:10.1007/S10518-020-00945-6/TABLES/5
15. Agresti A. *Categorical Data Analysis*. 3rd ed. Wiley; 2012.
16. Giorgio M, Iervolino I. On multisite probabilistic seismic hazard analysis. *Bull Seismol Soc Am*. 2016;106(3):1223-1234. doi:10.1785/0120150369
17. Esposito S, Iervolino I. PGA and PGV spatial correlation models based on European multievent datasets. *Bull Seismol Soc Am*. 2011;101(5):2532-2541.
18. Dolce M, Giordano F, Borzi B, Bocchi F, Di Meo A, Faravelli M. Observed damage database of past Italian earthquakes: the Da.D.O. *WebGIS*. 2019;60(2):141-164. doi:10.4430/bgta0254
19. Grünthal G, European Macroseismic Scale1998. Cahiers du Centre Europèen de Gèodynamiqueet de Seismologie. Published online; 1998.
20. CEN. EN 1998-1 - Eurocode 8: Design of Structures for Earthquake Resistance—Part 1: General Rules, Seismic Actions and Rules for Buildings; 2004.
21. Bindi D, Pacor F, Luzi L, et al. Ground motion prediction equations derived from the Italian strong motion database. *Bull Earthquake Eng*. 2011;9(6):1899-1920. doi:10.1007/s10518-011-9313-z
22. Joyner WB, Boore DM. Peak horizontal acceleration and velocity from strong-motion records including records from the 1979 Imperial Valley, California, earthquake. *Bull Seismol Soc Am*. 1981;71(6):2011-2038. Accessed January 10, 2019. https://pubs.geoscienceworld.org/ssa/bssa/article-abstract/71/6/2011/102109
23. Beyer K, Bommer JJ. Relationships between median values and between aleatory variabilities for different definitions of the horizontal component of motion. *Bull Seismol Soc Am*. 2006;96(4 A):1512-1522. doi:10.1785/0120050210

**How to cite this article:** Iervolino I, Rosti A, Penna A, Giorgio M. Damage-informed ground motion and semi-empirical fragility assessment. *Earthquake Engng Struct Dyn*. 2024;53:3514-3526. https://doi.org/10.1002/eqe.4184