Received: 1 June 2018 | Revised: 25 February 2019 | Accepted: 1 March 2019

DOI: 10.1002/env.2566

RESEARCH ARTICLE

WILEY

Bayesian inference in natural hazard analysis for incomplete and uncertain data

A. Smit[1,2] | A. Stein[2,3] | A. Kijko[1]

[1]University of Pretoria Natural Hazard Centre, Department of Geology, University of Pretoria, Pretoria, South Africa

[2]Department of Statistics, University of Pretoria, Pretoria, South Africa

[3]Faculty of Geo-information Science & Earth Observation (ITC), University of Twente, Enschede, The Netherlands

Correspondence
A. Smit, University of Pretoria Natural Hazard Centre, Department of Geology, University of Pretoria, Pretoria 0028, South Africa.
Email: ansie.smit@up.ac.za

Funding information
National Research Foundation of South Africa, Grant/Award Number: IFR160120157106, TP14072278140 (96412), and 94808

Abstract

This study presents a method for estimating two area-characteristic natural hazard recurrence parameters. The mean activity rate and the frequency–size power law exponent are estimated using Bayesian inference on combined empirical datasets that consist of prehistoric, historic, and instrumental information. The method provides for incompleteness, uncertainty in the event size determination, uncertainty associated with the parameters in the applied occurrence models, and the validity of event occurrences. This aleatory and epistemic uncertainty is introduced in the models through mixture distributions and weighted likelihood functions. The proposed methodology is demonstrated using a synthetic earthquake dataset and an observed tsunami dataset for Japan. The contribution of the different types of data, prior information, and the uncertainty is quantified. For the synthetic dataset, the introduction of model and event size uncertainties provides estimates quite close to the assumed true values, whereas the tsunami dataset shows that the long series of historic data influences the estimates of the recurrence parameters much more than the recent instrumental data. The conclusion of the study is that the proposed methodology provides a useful and adaptable tool for the probabilistic assessment of various types of natural hazards.

KEYWORDS

Bayesian estimation, incomplete data, natural hazard, power law, uncertain data

1 | INTRODUCTION

The successful modeling of natural hazards and their associated risks are important for human health, safety, and economic growth. Underestimation of natural hazards could lead to fatalities and economic losses, whereas overestimation could result in overpriced and excessive safety measures. Various industries use information, such as the probabilities of exceedances, return periods, and the upper limit of the event size to generate products for safeguarding society. The event size of an event refers to the scale at which the event is measured, for example, the earthquake magnitude, tsunami intensity, or the affected area of a fire.

Power laws are probably the models used most often to describe event size and frequency. Such equations are capable of modeling natural systems where a large number of small events occur compared with a small number of large events. These equations include the Gutenberg–Richter relation for earthquake magnitude (Gutenberg & Richter, 1956), tsunami intensity (Geist & Parsons, 2006; Soloviev, 1970), landslide area (e.g., Malamud, Turcotte, Guzzetti, & Reichenbach, 2004), solar flare intensity and the burned area for wild fires (e.g., Newman, 2005), and air pollution (Shi & Liu, 2009). The application of power law to various types of natural systems is discussed in, for example, Newman (2005), Burroughs and Tebbens (2001), and Geist and Parsons (2014).

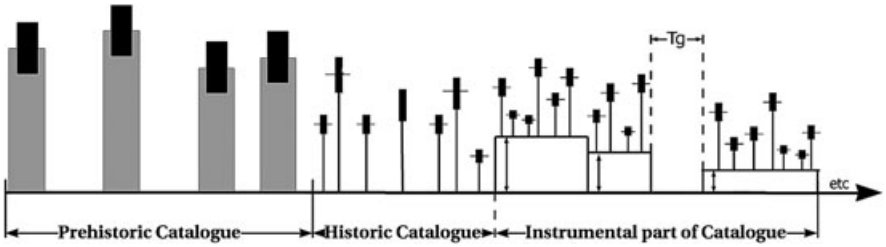

FIGURE 1 Illustration of typical data used for assessment of model recurrence parameters based on prehistoric, historic, and instrumental datasets (modified after Kijko, Smit, & Sellevoll, 2016)

Similar to any statistical distribution, the parameters of the power law are sensitive to the quality of the applied data. Data on natural systems can be highly incomplete and uncertain, thereby influencing the fitted power law (Burroughs & Tebbens, 2005). Accurate instrumental recordings of natural disasters represent only a very small part of the overall historical time line; therefore, there is a real possibility that the largest observed events are not included in the dataset. This omission could result in the underestimation of the hazard and the subsequent underestimation of the vulnerability or risk for the area under investigation.

In an effort to supplement instrumental datasets and reduce epistemic uncertainty, extensive research has been devoted to collecting reliable prehistoric and historic information. Prehistoric events are those recorded with palaeo-environmental studies. Such research is expensive as it requires the identification of areas where a prehistoric event could have taken place. Additional problems associated with prehistoric data are retrieving the exact date of the occurrence and the size of the event, which is often not possible. Historic events are those typically observed from the time of first human settlement. Their quality depends upon whether they were observed and described accurately.

Figure 1 provides a typical illustration of available data used in the assessment of recurrence parameters in virtually any natural system. Prehistoric data are subject to uncertainty relevant to the time of occurrence, the exact size of the event, and incompleteness in terms of the probability of detecting an event. Historic data, consisting of the largest observed events, and instrumental datasets are subject to incompleteness and uncertainty relevant to the observed event size, varying levels of certainty regarding the exact location of an event, and varying probabilities of all events above a certain minimum size being observed. Time gaps T_g represent missing event records.

Regardless the quality of the data, prehistoric and historic information can be combined with instrumental records using Bayesian statistics (e.g., Fernandes, Naghettini, & Loschi, 2010) or using the additive property of likelihood functions (Rao, 1973). Additional information from independent sources can be included to stabilize and constrain the estimates. Typical examples are found in seismology, such as recurrence parameters for geologically similar areas (Campbell, 1982, 1983). Silva, Portela, Naghettini, and Fernandes (2017) provide ideas for what could be used as prior information in extreme-frequency flood analyzes. Patskoski and Sankarasubramanian (2018) discuss various forms of potential prior information that are used in time-series-based hydrological studies, such as tree rings, sea surface temperature, probable maximum flood discharge (Fernandes et al., 2010), expert judgment (Viglione, Merz, Salinas, & Blöschl, 2013), and climate covariates (Sun, Thyer, Renard, & Lang, 2014). Other geologic hazards, such as tsunamis and landslides, can include prior information based on knowledge of seismotectonics (Geist & Uri, 2012), rainfall thresholds (Berti et al., 2012), and expert opinions (e.g., Yazdani & Kowsari, 2013). Uninformative priors occur commonly among different types of hazards (Cooley, Nychka, & Naveau, 2007; Lyubushin & Parvez, 2010).

The aim of this study is to present a generic methodology that is capable of utilizing different sources information and uncertainty in natural hazard modeling. The methodology is applicable to various types of natural hazards, where the frequency–size relation follows a power law. We show how prehistoric, historic, and instrumental data can be incorporated and can account for incomplete data, and uncertainty in event sizes and applied occurrence distributions. In addition, we present an argument for the inclusion of weighting information reflecting individual event validity. Bayesian inference is applied to obtain the recurrence parameter estimates.

2 | METHODOLOGY

The proposed methodology investigates natural systems by considering the frequency at which they occur and their respective sizes. First, we define the model before introducing its modifications and extensions to account for the different

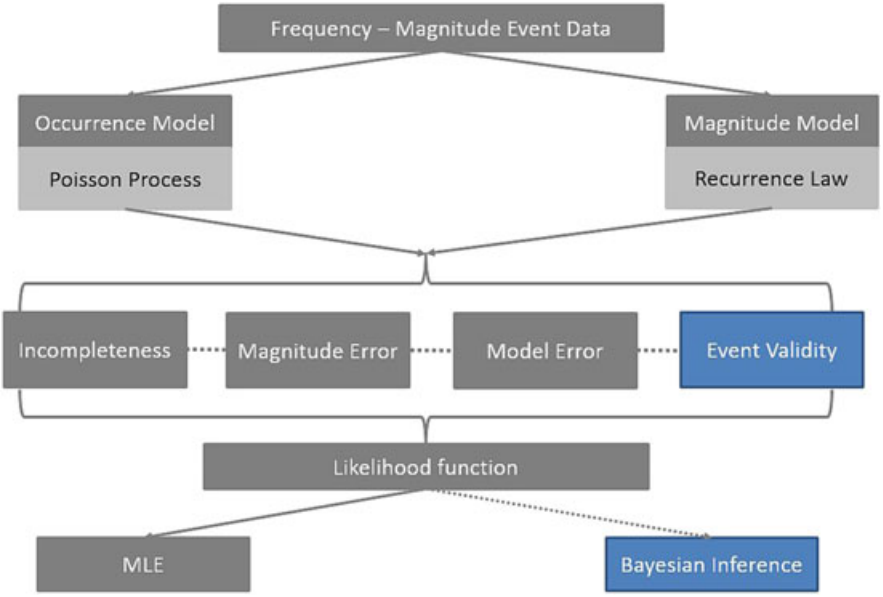

FIGURE 2 Schematic illustration of the proposed methodology, showing the types of data that can be used, the nature of the uncertainties that can be considered, and the estimation techniques for the model parameters

types of uncertainty. Figure 2 illustrates the applied methodology for the estimation of recurrence parameters. The methodology is similar to that described by Kijko et al. (2016) and Smit, Kijko, and Stein (2017). With reference to the applied power law, the proposed methodology accounts for aleatory and epistemic uncertainty by providing for incompleteness, uncertainty in the event size, and uncertainty in the applied occurrence models, with reference to the applied power law. As an extension of this methodology, our study introduces the validity of occurrence and Bayesian inference to constrain the estimated parameters.

2.1 | Size and frequency

As a first step, we assumed independence between the occurrence and the size of an event. As regards earthquakes, independence can be obtained by removing the events classified as swarms and/or foreshocks and aftershocks (e.g., Cornell, 1968). The random variable X refers to the size of the observed events. The sample space for event sizes is defined in terms of N possible independent and identically distributed (iid) variables $X = \{x_1, x_2, \dots x_N\}$. Following Figure 1, a dataset can be divided into subdatasets consisting of prehistoric (P), historic (H), and s complete instrumental datasets. Each of these subdatasets is considered complete for event sizes exceeding a certain level of completeness (LoC) $x_{\min}^{(i)}$ during a certain period of time t_i for $i = P, H, 1, \dots, s$. Incompleteness of the dataset is accounted for by defining the likelihood functions for each subdataset. Creating these subdatasets, each with different levels of completeness, allows more information to be included in the assessment of the recurrence parameters. Smaller event sizes are crucial, as they stabilize the slope of the power law, thereby providing superior estimates. Several authors have investigated the possibility of subsetting earthquake data for parameter estimation (Kijko & Smit, 2012, and the references therein). These authors subdivided the data according either to time or to earthquake magnitude levels over time. In this study, we subset the data according to the type of events (prehistoric, historic, and instrumental data), as well as the LoC over time, introduced by Kijko and Sellevoll (1989). By structuring the dataset in this fashion and building likelihood functions, data with different underlying assumptions could be included in the same assessment process. This approach permits the occurrence of "gaps" (T_g) to account for missing event records. Missing records are often attributable to event recording equipment or networks being nonoperational with other socioeconomic and social factors probably also playing a role.

The proposed methodology investigates natural systems by examining the frequency at which they occur in relation to their respective sizes in terms of power laws. For the investigated area, it is assumed that the probability to observe a number of n events in the time interval Δt is described by homogeneous Poisson distribution with parameter $\lambda \Delta t$:

$$P_N = P\left(N_{X \geq x_{\min}} = n\right) = \frac{e^{-\lambda \Delta t}(\lambda \Delta t)^n}{n!}, \quad n = 0, 1, 2 \ \dots, \tag{1}$$

where $N_{X \geq x_{\min}}$ is the number of events where $x_i \geq x_{\min}$. Parameter λ describes the mean, usually annual, area-characteristic activity rate of event occurrence, with $x_i \geq x_{\min}$. The parameter $x_{\min}$ is the size above, where it is certain that all the events were recorded and included in the analysis.

Furthermore, the number of observed events can be described in terms of event sizes in a specific time interval by using power laws. Power laws are used often when the frequency of the event occurrence with event size in the interval $[x, x + dx]$ tends to be linear on the log–log scale. Assuming the event size x is measured on a logarithmic scale, this linear relationship is expressed as $\ln n(x) = c - kx$ and transforms to $n(x) = Ce^{-kx}$, with k representing the power law exponent and $C = \exp(c)$ a constant. One disadvantage of using the frequency–size relation is the division of the size variable into bins, which is not always possible. The cumulative frequency–size relation (CFSR) overcomes this problem and allows the measurement sizes to fall on a continuous scale (Newman, 2005). The CFSR, indicated by $n_{X \geq x}$, is defined as

$$n_{X \geq x} = \frac{C}{k-1} 10^{-(k-1)x} \tag{2}$$

and represents the number of events observed, in a specified time interval, that are larger than or equal to x. By taking the logarithm, Equation (2) is transformed into a linear equation

$$\log n_{X \geq x} = a - bx, \tag{3}$$

with parameters $a = \log\left(\frac{C}{k-1}\right)$ and $b = (k-1)$; see Newman (2005). The cumulative-frequency power law probability distribution is also referred to as a Pareto distribution. Equation (3) can be transformed into an exponential distribution

$$n_{X \geq x} = \exp(\alpha - \beta x), \tag{4}$$

with parameters $\beta = b \ln 10$ and the $\alpha = a \ln 10$. Equation (3) is equivalent to the well-known frequency–magnitude Gutenberg–Richter relation in seismology, where β is a parameter expressing the relationship between large and small earthquakes. The size of a tsunami, expressed in terms of intensity, can be written in the same way (Smit et al., 2017). For purposes of convenience, $n_{X \geq x}$ will be referred to as n in the rest of this paper.

The characteristics of natural events are controlled by the forces of nature adhering to specific physical constraints. These constraints are unique to the various types of natural systems. For example, the potential earthquake magnitude of a specific geological fault is related to the length of the fault. This physical law indicates that an upper earthquake magnitude limit $x_{\max}$ has to exist. Similarly, the physical processes that govern events such as tsunamis, landslides, floods, and fires are subject to upper limits. The area-characteristic maximum possible event size represents the worst-case scenario and is an important parameter in hazard and risk modeling. Estimates of the upper limit of a distribution can be highly uncertain, particularly when the datasets are short. In an instance of an upper limit being assumed, the unknown true $x_{\max}$ forms part of the range over which the model parameters λ and β must be evaluated. Integrating the model parameter over an unknown value violates the required condition of regularity of likelihood functions (Cheng & Traylor, 1995). Methods designed to avoid this problem are discussed in Kijko and Singh (2011) and Vermeulen and Kijko (2017).

Based on the cumulative-frequency power law in Equation (3), a cumulative distribution function (CDF) for event sizes can be defined, taking into consideration the existence of the LoC of a dataset $x_{\min}$, as well as an upper limit $x_{\max}$. This yields a normalized, shifted-truncated CDF within the range $[x_{\min}, x_{\max}]$ (Cosentino, Ficara, & Luzio, 1977) as

$$F_X(x) = \begin{cases} 0 & x < x_{\min} \\ \frac{e^{-\beta x_{\min}} - e^{-\beta x}}{e^{-\beta x_{\min}} - e^{-\beta x_{\max}}} & x_{\min} \leq x \leq x_{\max} \\ 1 & x > x_{\max}. \end{cases} \tag{5}$$

Prehistoric and historic datasets contain only the largest events and are governed by extreme distributions. Utilizing the Poisson distribution (Equation (1)) and the shifted-truncated frequency–size distribution (Equation (5)), an extreme distribution is defined in terms of the recurrence parameters λ and β, thereby providing the link between extreme and instrumental datasets (Kijko & Sellevoll, 1989).

Let $x_{\min}$ represent the LoC across the entire dataset, x_H represent the smallest event in the historic subdataset, and x_P represent the smallest event in the prehistoric dataset such that $x_P \geq x_{\min}$ and $x_H \geq x_{\min}$. The derivation of the extreme distribution will be described only in terms of the prehistoric dataset, but it is translated easily into the historic data.

Let $X_P = \{x_1, x_2, \ldots x_{N_P}\}$ represent the sample space for prehistoric event sizes and N_P represent the number of prehistoric events where both the sizes and number of events are iid. The occurrence of observed events in time is assumed to

follow the homogeneous Poisson with parameter λ_P. The probability to observe a number of n_P events in the time interval Δt_P with sizes larger than $x_{\min}$ is defined as the Poisson distribution with parameter $\lambda_P \Delta t_P$.

Let x_0 be the largest event size in a specified time interval such that $F_X^{\mathrm{MAX}}(x_0) = [F_X(x_0)]^{n_P}$ (e.g., Bain & Engelhardt, 1992). Therefore, the distribution of the largest (extreme) events within the range $[x_{\min}, x_{\max}]$ takes the form

$$F_X^{\mathrm{MAX}}(x_0) = \begin{cases} 0 & x_0 < x_{\min} \\ \left[\frac{e^{-\beta x_{\min}} - e^{-\beta x_0}}{e^{-\beta x_{\min}} - e^{-\beta x_{\max}}}\right]^{n_P} & x_{\min} \le x_0 \le x_{\max} \\ 1 & x_0 > x_{\max}. \end{cases} \tag{6}$$

By the theorem of total probability (Cramér, 1961), the probability that sizes of all the events are less than x_0 in an arbitrary time interval Δt_P takes the form of the Gumbel I extreme distribution (e.g., Epstein & Lomnitz, 1966):

$$F_X^{\mathrm{MAX}}(x_0|\Delta t_P) = \exp\left(-\lambda_P \Delta t_P \left[1 - F_X(x_0)\right]\right) \quad x_{\min} \le x_0 \le x_{\max}. \tag{7}$$

2.2 | Accounting for uncertainties

Temporal, spatial, and/or spatial–temporal dependencies violate the independent and stationarity conditions in classic hazard assessment procedures. In the instance of earthquakes and tsunamis, this can be attributed to various factors, such as changes in the state of stress in rock (e.g., Ogata & Abe, 1991; Scholz, 2015). Rainfall, deforestation, and earthquakes affect the occurrence of landslides (Geist & Parsons, 2006; Tatard, Grasso, Helmstetter, & Garambois, 2010), whereas weather phenomena, such as the El Niño Southern Oscillation (ENSO) affect meteorological and hydrological occurrences (Egüen, Aguilar, Solari, & Losada, 2016; Khaliq, Ouarda, Ondo, Gachon, & Bobée, 2006). Accordingly, the assumption would fail that a homogeneous Poisson distribution describes the probability that the observed number of events in the time interval Δt equals n in a specific area of investigation. Similarly, the assumption of iid recorded event sizes might not hold. Therefore, the applied distributions should provide for this type of uncertainty in the data to ensure that the perceived hazard is not underestimated or over estimated.

The replacement of the classic distributions by their mixture counterparts is probably simplest way to accommodate the differences between the observed processes and the applied distributions (e.g., Vicini, Hotta, & Achar, 2013). By their very nature, mixture distributions can accommodate the temporal and spatial occurrence of natural events and can account for weak dependencies in the data. The assumption behind mixed distributions is that their parameters, in this study λ and β, are random variables and are subject to fluctuation (e.g., Cunningham, Herzog, & London, 2012; Daykin, Pentikainen, & Pesonen, 1993; Fernandes et al., 2010; Yadav, Tsapanos, Tripathi, & Chopra, 2013). In this approach, the Poisson occurrence distribution (Equation (1)) and the exponential event-size distributions (Equation (5)) are replaced by the Poisson mixture and the exponential mixture distributions.

The two-parameter gamma distribution is a frequent choice for a mixing distribution. It is used in many research fields, as it is flexible enough to adopt different forms by means of the shape (p) and scale (q) parameters of the two-parameter gamma distribution. Similar to the Poisson and exponential distributions, it is infinitely divisible, a condition for iid random variables. In addition, the gamma distribution is the maximum entropy distribution for the Pearson Type III distributions, committing only to the known information and limiting the prior information required. It adheres to the characteristic of systems governed by physical laws that move toward the maximum entropy. The resulting Poisson-gamma and exponential-gamma distributions are equivalent, respectively, to the negative binomial and Pareto distributions. Compared with classic distributions, mixture distributions have the additional benefit that the unconditional variance of the mixture distribution is larger because of the assumed uncertainty in the mixing distribution (Klugman, Panjer, & Wilmot, 2004). Examples of applying the gamma distribution can be seen in probabilistic models of failure times in engineering (Hamada, Wilson, Reese, & Martz, 2008), time to the kth event in seismology (Benjamin & Cornell, 2014), as well as in the insurance and risk industry for modeling claim numbers in risk theory (Daykin et al., 1993; Klugman et al., 2004).

In our work, the gamma distribution GAM(p_λ, q_λ) is used to model the uncertainty in the parameter λ. The parameters of the gamma function are defined through the mean and variance of λ as $p_\lambda = \bar{\lambda}/\sigma_\lambda^2$ and $q_\lambda \equiv v_\lambda^{-2} = \bar{\lambda}^2/\sigma_\lambda^2$, respectively, with v_λ set as the coefficient of variation for the activity rate λ:

$$P_N(n|\Delta t, v_\lambda) = \frac{\Gamma(n + q_\lambda)}{n!\Gamma(q_\lambda)} \left(\frac{p_\lambda}{\Delta t + p_\lambda}\right)^{q_\lambda} \left(\frac{\Delta t}{\Delta t + p_\lambda}\right)^{n}, \tag{8}$$

where $\bar{\lambda}$ denotes the mean parameter of λ and n is the value of $N_{X \ge x_{\min}}$.

In a similar way, the GAM(p_β, q_β) distribution is used to model the uncertainty in β parameter in Equations (5) and (7) with $p_\beta = \bar{\beta}/\sigma_\beta^2$ and $q_\beta \equiv v_\beta^{-2} = \bar{\beta}^2/\sigma_\beta^2$. The exponential-gamma distribution for the interval $x_{\min} \le x \le x_{\max}$ and the extreme exponential-gamma distribution for the interval $x_{\min} \le x_0 \le x_{\max}$, respectively, take the form

$$F_X\left(x|v_\beta\right) = \left[\frac{1-\left(\frac{q_\beta}{\bar{\beta}(x-x_{\min})+q_\beta}\right)^{q_\beta}}{1-\left(\frac{q_\beta}{\bar{\beta}(x_{\max}-x_{\min})+q_\beta}\right)^{q_\beta}}\right], \tag{9a}$$

$$F_X^{\mathrm{MAX}}\left(x_0|\Delta t_P, \boldsymbol{\nu_P}\right) = \exp\left(-\bar{\lambda}_P \Delta t_P\left[1-\frac{\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_0-x_{\min})+q_\beta}\right)^{q_\beta}\right]}{\left[1-\left(\frac{q_\beta}{\bar{\beta}(x_{\max}-x_{\min})+q_\beta}\right)^{q_\beta}\right]}\right]\right), \tag{9b}$$

where $\bar{\beta}$ denotes the mean parameter of β and $\boldsymbol{\nu_P} = (v_\lambda^P, v_\beta^P)$ is the vector of the coefficients of variation for λ and β. The respective probability distribution functions (PDFs) of the above CDFs are derived in Kijko et al. (2016).

In addition, event size uncertainty can be built into the formalism. Uncertainty in event size determination occurs because of uncalibrated instruments or when such determination is based upon the qualitative description of the effect of the event on the environment. One method of dealing with size uncertainty is to assume that the apparent (observed) size of an event $\check{x}$ consists of the true (unknown) size x and some error ε such that $\check{x} = x + \varepsilon$. The uncertainty in event size can be dealt with in different ways, with the hardbound and softbound models probably used most often.

The essence of the hardbound model is the assumption that the uncertainty of event size x is described by the uniform distribution between two values: the lower bound x_L and the upper bound x_U. If $\delta = \frac{1}{2}(x_U - x_L)$, then $x_L = \check{x} - \delta$ and $x_U = \check{x} + \delta$. Size uncertainty is implemented utilizing the convolution of size distribution with a uniform distribution. The PDF of event size, with an account of the hardbound size model using the shifted-truncated exponential-gamma distribution, is determined using the formula defined in Kijko and Sellevoll (1992).

$$f_X\left(\check{x}|\delta\right) = \frac{1}{2\delta}\begin{cases} F_X\left(\check{x}+\delta\right) - F_X\left(x_{\min}\right) & \check{x} < x_{\min} + \delta \\ \left[\frac{F_X(\check{x}+\delta)-F_X(\check{x}-\delta)}{1-F_X(\check{x}|\delta)}\right] & x_{\min} + \delta \le \check{x} < x_{\max} - \delta \\ 1 - F_X\left(\check{x}-\delta\right) & x_{\max} - \delta \le \check{x} < x_{\max} + \delta. \end{cases} \tag{10}$$

For this study, we assumed that event size uncertainty for prehistoric and historic data is described using hardbound models. The respective CDFs are obtained by substituting Equation (9b) into Equation (10) and solving the integral with numerical methods.

A softbound model of size error δ is a distribution with a continuous support, typically Gaussian, with mean equal to zero and a standard deviation of σ_x. The approximate CDF for the shifted-truncated exponential-gamma distribution (Equation (9a)) with an account of the softbound size model (Kijko et al., 2016):

$$F_X\left(x|v_\beta, \sigma_x\right) = \frac{C_\beta \bar{\beta} q_\beta^{q_\beta+1}}{2\sigma_x}\left\{A + B\right\}, \tag{11}$$

where

$$A = \frac{(r_1 + r_2\alpha)^{-q_\beta}}{r_2 q_\beta}\Bigg|_{\frac{x-x_{\min}}{\sigma_x}}^{\frac{x-x_{\max}}{\sigma_x}},$$

$$B = \left(\frac{2}{\pi}\right)^{\frac{1}{2}}\sum_{h=0}^{\infty}\frac{(-1)^h}{2^h h!\,(2h+1)}\frac{1}{b^{2h+2}} \times \sum_{j=0}^{2h+1}\frac{(2h+1)!(-r_1)^j(r_1+r_2\alpha)^{2h+1-q_\beta-j}}{(2h+1-j)!\,j!\,\left(2h+1-q_\beta-j\right)}\Bigg|_{\frac{x-x_{\min}}{\sigma_x}}^{\frac{x-x_{\max}}{\sigma_x}},$$

where $C_\beta = \left[1-\left(\frac{q_\beta}{q_\beta+\bar{\beta}(x_{\max}-x_{\min})}\right)^{q_\beta}\right]^{-1}$, $\alpha = q_\beta + \bar{\beta}\left(x - x_{\min}\right)$, $b = -\bar{\beta}\sigma_x$, $r_1 = q_\beta + \bar{\beta}\left(x - x_{\min}\right)$, and $r_2 = \bar{\beta}\sigma_x$.

3 | ESTIMATION OF PARAMETERS

3.1 | Likelihood functions

Using the additive property of likelihood functions, the individual likelihood functions are multiplied to provide a single likelihood function for the entire available dataset. Therefore, L_{Total}, the likelihood function of the unknown model parameters $(\bar{\lambda}, \bar{\beta})$, for given $x_{\max}$ and based on the entire dataset can be written as

$$L_{\text{Total}}\left(\bar{\lambda}, \bar{\beta} | \boldsymbol{I}_P, \boldsymbol{I}_H, \boldsymbol{I}_i, x_{\max}\right) = L_P\left(\bar{\lambda}, \bar{\beta} | \boldsymbol{I_P}, x_{\max}\right) \times L_H\left(\bar{\lambda}, \bar{\beta} | \boldsymbol{I_H}, x_{\max}\right) \times \prod_{i=1}^{s} L_i\left(\bar{\lambda}, \bar{\beta} | \boldsymbol{I}_i, x_{\min}^{(i)}, x_{\max}\right), \tag{12}$$

where L_P and L_H denote the likelihood functions based on the prehistoric and historic parts of the database; L_i is the likelihood function based on the ith subdataset ($i = 1, \ldots, s$); and $\boldsymbol{I}_P = (\boldsymbol{x_P}, \boldsymbol{t_P}, \boldsymbol{v})$, $\boldsymbol{I}_H = (\boldsymbol{x_H}, \boldsymbol{t_H}, \boldsymbol{v})$, and $\boldsymbol{I}_i = (n_i, t_i, \boldsymbol{x_i}, \boldsymbol{v})$ are the background information for the three types of data. For the prehistoric dataset, $\boldsymbol{x_P}$ and $\boldsymbol{t_P}$ are the $(n_P \times 1)$ vectors of prehistoric event sizes $x_{P,k}$ that occurred within time intervals $\Delta t_{P,k}$, where $k = 1, \ldots, n_P$. The vectors $\boldsymbol{x_H}$, $\boldsymbol{x_i}$, $\boldsymbol{t_H}$, and $\boldsymbol{t_i}$ are defined similarly for the historic and instrumental datasets. The number of events n_i is the number of events observed in the ith subdataset. The vector $\boldsymbol{v} = (v_\lambda, v_\beta)$ constitutes coefficients of variation for the unknown $\bar{\lambda}$ and $\bar{\beta}$.

The likelihood function for the complete, instrumental datasets $L_i(\bar{\lambda}, \bar{\beta} | \boldsymbol{I_i}, x_{\min}^{(i)}, x_{\max})$ is a combination of the likelihood functions following from the frequency and size event distributions. In the instance that the number of events is independent of their sizes, the respective likelihood functions for subdataset $i = 1, \ldots s$ are defined as

$$L_i\left(\bar{\lambda}, \bar{\beta} | \boldsymbol{I_i}, x_{\min}^{(i)}, x_{\max}\right) = L_{\lambda i}\left(\bar{\lambda} | n_i, \Delta t_i, v_\lambda\right) \times L_{\beta i}\left(\bar{\beta} | \boldsymbol{x_i}, \boldsymbol{v}, x_{\min}^{(i)}, x_{\max}\right), \tag{13}$$

where

$$L_{\lambda i}\left(\bar{\lambda} | n_i, \Delta t_i, v_\lambda\right) = \bar{\lambda}^{(i)}\left(\frac{1}{\bar{\lambda}_{\text{i}} \Delta t_i + q_\lambda}\right)^{q_\lambda}\left(\frac{\bar{\lambda}_{\text{i}} \Delta t_i}{\bar{\lambda}_{\text{i}} \Delta t_i + q_\lambda}\right)^{n_i} \tag{14}$$

and

$$L_{\beta i}\left(\bar{\beta} | \boldsymbol{x_i}, \boldsymbol{v}, x_{\min}^{(i)}, x_{\max}\right) = \left[C_\beta \bar{\beta}\right]^{n_i} \prod_{k=1}^{n_i}\left[1 + \frac{\bar{\beta}}{q_\beta}\left(x_{i,k} - x_{\min}^{(i)}\right)\right]^{-(q_\beta + 1)}, \tag{15}$$

with $x_{i,k}$ representing event k in the subdataset i, $\bar{\lambda}^{(i)} = \bar{\lambda}(x_{\min})\left[1 - F_X\left(x | v_\beta, \sigma_X\right)\right]$, where $\bar{\lambda}(x_{\min})$ is the mean activity rate for the LoC $x_{\min}$ (Kijko & Sellevoll, 1989) and $C_\beta = \left[1 - \left(\frac{q_\beta}{q_\beta + \bar{\beta}(x_{\max} - x_{\min})}\right)^{q_\beta}\right]^{-1}$, as defined in Equation (11).

Other examples of constructing a total likelihood function to combine different types of data are employed in earthquake, tsunami, and extreme flood analyzes (e.g., Fernandes et al., 2010; Kijko et al., 2016; Lam, Thompson, Croke, Sharma, & Macklin, 2017; Smit et al., 2017; Stedinger & Cohn, 1986).

The underlying mechanical trigger of natural catastrophes is sometimes questionable. This is particularly true for prehistoric and historic tsunami occurrences, as it could be difficult to distinguish between the effects of a tsunami, severe storm surges, and floods. In earthquake datasets, it is difficult to distinguish between triggered and induced events, and landslides can be caused by extreme rainfall or by earthquakes. Some datasets have an additional variable that expresses the validity of the event. For example, the GITEC catalogue criteria (e.g, Tinti, Maramai, & Graziani, 2001) are a validity index for each observation ranging from 0 for an event considered extremely improbable to 4 for a definite tsunami, with a probability close to one.

The presence of uncertain and questionable event data with respect to the applied model can have a serious effect on the estimated recurrence times for event sizes. This could lead to the an erroneous assessment of the recurrence parameters. To account for the validity of an event, the standard likelihood function is replaced with the weighted likelihood (WL) function. This procedure at least preserves the first-order asymptotic properties of the classic likelihood function, leading to estimators with the usual asymptotic behavior (Markatou, Basu, & Lindsay, 1998).

The WL function in the context of the above methodology equals

$$L_i\left(\bar{\lambda}, \bar{\beta} | \boldsymbol{I_i}, x_{\max}\right) \equiv \prod_{k=1}^{n_i} f_X^{\text{i}}\left(x_k | \bar{\lambda}, \bar{\beta}, \boldsymbol{I_i}, x_{\min}, x_{\max}\right)^{w_j}, \tag{16}$$

where $w_j \equiv w(x_j)$ is the known weight of observation j in subdataset i, and the PDF for the subdataset i, $f_X^i(x|\bar{\lambda}, \bar{\beta}, \boldsymbol{I_i}, x_{\min}, x_{\max})$, is then defined by the multiplication of the PDFs for the Poisson-gamma occurrence distribution and the exponential-gamma event size distribution, ie, $f_X^i(\bar{\lambda}|n_i, t_i, \boldsymbol{\nu}, x_{\max}) \times f_X^i(\bar{\beta}|\boldsymbol{x_i}, \boldsymbol{\nu}, x_{\min}, x_{\max})$. The weights range from 0 to 1, where 1 is equivalent to 100% certainty of the validity of an event. The effect of questionable events in the model is reduced in this way.

3.2 | Bayesian inference

The accuracy of the maximum likelihood (ML) estimates obtained by the maximization of likelihood (Equation (12)) depends upon the quality of the observed dataset. Small datasets often do not yield reliable estimates for natural hazards, as they provide only a limited view of the characteristics of the physical process. By including prior information in the estimation process, the hazard estimates are improved and stabilized.

Following the Bayesian rule, the posterior distribution of the parameters $z(\bar{\lambda}, \bar{\beta}, x_{\max}|I)$is constructed with the a priori probability defined as $\pi(\boldsymbol{\theta})$, in which $\boldsymbol{\theta} = (\bar{\lambda}, \bar{\beta}, x_{\max})$ and I is the background information that denotes all the assumptions related to the particular investigation. In seismology, Bayesian inference is performed typically directly on the shifted-truncated distribution. Common priors used in these cases are the gamma distribution for $\bar{\lambda}$ and the beta distribution for $\bar{\beta}$ (Campbell, 1982, 1983; Mortgat & Shah, 1979), thereby accounting for model uncertainty. The gamma and beta distributions are also conjugate priors, providing closed-form expressions. We explicitly accounted for the model uncertainty by using mixture distributions, as this affords the opportunity to use priors from additional but independent information. These priors could be taken as uninformative with $\pi(\boldsymbol{\theta}) = \text{const}$ when little or no information is available. Alternatively, following the law of large numbers and the central limit theorem, our choice of prior $\pi_\beta(\bar{\beta})$ for $\bar{\beta}$ is the Gaussian distribution with the moments μ and σ^2 (Aki, 1965; Kijko & Graham, 1999). The mean annual rate of occurrence is a parameter that is specific to the region; therefore, we assigned an uninformative prior $\pi_\lambda(\bar{\lambda})$ to $\bar{\lambda}$ in the form of a uniform distribution in the range $[\lambda_A, \lambda_B]$ (Dong, Shah, Bao, & Mortgat, 1984). Similarly, an uninformative prior is assumed for the estimation of the area-characteristic maximum possible event size of the range $[x^A_{\max}, x^B_{\max}]$. In Equation (17), we assume that the prior information for $\bar{\beta}$ and $\bar{\lambda}$ is independent. Alternatively, a dependent prior $\pi(\bar{\beta}, \bar{\lambda}, x_{\max})$ can be assumed. Pisarenko, Lyubushin, Lysenko, and Golubeva (1996) implemented uniform priors for λ, β, and $x_{\max}$ that are constant on a parallelepiped (Pisarenko & Lyubushin, 1997). Assuming that the prior information for the three recurrence parameters is independent, the joint prior distribution $\pi(\boldsymbol{\theta})$ is defined as $[\pi_\lambda(\bar{\lambda})\pi_\beta(\bar{\beta})\pi_{x_{\max}}(x_{\max})]$. The posterior distribution and its estimated mean $\widehat{\boldsymbol{\mu}}(\bar{\lambda}, \bar{\beta}, x_{\max}|I)$ and variance $\widehat{\mathbf{var}}(\bar{\lambda}, \bar{\beta}, x_{\max}|I)$ are evaluated numerically

$$z\left(\bar{\lambda}, \bar{\beta}, x_{\max}|I\right) = \frac{\left(\frac{1}{x^B_{\max} - x^A_{\max}}\right)\left(\frac{1}{\lambda_B - \lambda_A}\right)\exp\left(-\frac{(\bar{\beta}-\mu)^2}{\sigma^2}\right) L_{\text{Total}}\left(\bar{\lambda}, \bar{\beta}|\boldsymbol{I}\right)}{\int L_{\text{Total}}\left(\bar{\lambda}, \bar{\beta}|I\right)\left(\frac{1}{x^B_{\max} - x^A_{\max}}\right)\left(\frac{1}{\lambda_B - \lambda_A}\right)\exp\left(-\frac{(\bar{\beta}-\mu)^2}{\sigma^2}\right) d\bar{\beta}\,\bar{\lambda}}. \tag{17}$$

4 | APPLICATIONS

For purposes of demonstration, two datasets were analyzed. Section 4.1 shows the effect of introducing the different types of uncertainty, namely, combining different types of datasets, event size uncertainty, parameter uncertainty, and the uncertainty associated with occurrence. The assessment is based on a synthetic earthquake dataset. The analyses of the Japanese tsunami dataset are presented in Section 4.2. All the estimates were derived by applying the Bayesian inference.

4.1 | Simulated dataset

To evaluate the behavior of the proposed methodology, a typical earthquake dataset was generated using the Monte Carlo simulation to mimic typical prehistoric, historic, and instrumentally recorded earthquake data. The prehistoric data were generated using Equation (9b) substituted into Equation (10), as well as the historic data with the historic derivations of these equations. The individual instrumental datasets of different levels of completeness were generated using Equation (11). The apparent earthquake magnitudes in terms of moment magnitude M_W were generated for $\beta = 2.302$ ($b = 1.0$), and a mean annual activity rate for a LoC equal to 4.0, that is, $\lambda(x_{\min} = 4.0) = 10$. Variation of 25% was included in the data, as well as uniform event size errors for prehistoric and historic data and Gaussian errors for instrumental data.

TABLE 1 Input for the generation of a synthetic earthquake magnitude dataset for the shifted-truncated exponential-gamma distribution, with $b = 1$ and $\lambda(x_{\min} = 4) = 10$

Input	Prehistoric(*P*)	Historic(*H*)	Instrumental 1(*c*1)	Instrumental 2(*c*2)
Year begin	100,000 BC	1500-01-01	1970-01-01	2001-01-01
Year end	1 AD	1969-12-31	2000-12-31	2017-12-31
Time periods	5,000 years	[50–2.5] years	Annual	Annual
Level of completeness $x_{\min}^{(i)}$	7.0	6.0	5.0	4.0
Magnitude error	0.5	0.5	0.3	0.1
Number of events (n_i)	20	20	42	161
Maximum observed magnitude ($x_{\max}^{\text{obs}}$)	9.53	8.84	7.78	6.5

Table 1 shows the input for the generation of the data. The evaluated time periods were chosen to simulate the way earthquake data are usually observed. Each subdataset has different and decreasing with time LoC and earthquake magnitude errors.

The same equations used to generate the synthetic earthquake dataset were used to derive the total likelihood function L_{Total} utilized in the Bayesian estimation process defined by the posterior distribution of Equation (17). Two versions of L_{Total} were investigated, namely, the unweighted likelihood, as defined in Equation (12), and the WL function version of Equation (12) when applying Equation (16). To assess the effect of including the different epistemic and aleatory uncertainties on the recurrence parameters λ and β of the power law, eight different model scenarios were investigated using different variations of the equations defined above. The area-characteristic maximum possible event size $x_{\max}$, in this instance the earthquake magnitude, was set to 9.65 across all eight scenarios. The scenarios are defined as follows.

1. **LoC**: Earthquake event dataset contains only instrumental data, with different levels of completeness $x_{\min}^{(i)}$.
2. **LoC_MAG**: Earthquake event dataset contains only instrumental data, with different LoC $x_{\min}^{(i)}$. The earthquake event magnitudes are uncertain.
3. **LoC_MOD**: Earthquake event dataset contains only instrumental data, with different LoC $x_{\min}^{(i)}$. The parameters of the earthquake recurrence model are assumed uncertain.
4. **LoC_MAG_MOD**: Earthquake event dataset contains only instrumental data, with different LoC $x_{\min}^{(i)}$. The earthquake event magnitudes are uncertain, and the parameters of the earthquake recurrence model are assumed uncertain.
5. **LoC_OCC**: Earthquake event dataset contains only instrumental data, with different LoC $x_{\min}^{(i)}$. The validity of event occurrence is introduced.
6. **LoC_MAG_OCC**: Earthquake event dataset contains only instrumental data, with different LoC $x_{\min}^{(i)}$. The earthquake event magnitudes are uncertain. The validity of event occurrence is introduced.
7. **LoC_MOD_OCC**: Earthquake event dataset contains only instrumental data, with different LoC $x_{\min}^{(i)}$. The parameters of the earthquake recurrence model are assumed uncertain. The validity of event occurrence is introduced.
8. **LoC_MAG_MOD_OCC**: Earthquake event dataset contains only instrumental data, with different LoC $x_{\min}^{(i)}$. The earthquake event magnitudes are uncertain and the parameters of the earthquake recurrence model are assumed uncertain. The validity of event occurrence is introduced.

For all eight scenarios, the recurrence parameters $\hat{\bar{\lambda}}$ and $\hat{\bar{\beta}}$ are calculated by application of the Bayesian inference formalism. The results are summarized in Table 2. In all tests, two types of $\pi(\bar{\beta})$ and $\pi(\bar{\lambda})$ priors were considered: The Gaussian distribution for $b = 1.0 \pm 0.1$ and the uninformative uniform distribution for $\bar{\lambda}$. The choices for the priors follow the global *b*-value estimate for tectonically active areas (El-Isa & Eaton, 2014) for $\pi(\bar{\beta})$, and the assumption of a lack of knowledge of the activity rate in the region in question for $\pi(\bar{\lambda})$.

The estimates $\hat{\bar{\lambda}}$ and $\hat{\bar{b}}$ (where $\hat{\bar{\beta}} = \hat{\bar{b}} \ln(10)$) for each scenario are provided in Table 2. In addition, this table provides the percentage contribution of each type of dataset to each of the estimates. Figure 3 shows the output for the return periods, and Figures 4 and 5 show how much $\hat{\bar{\lambda}}$ and $\hat{\bar{b}}$ rely on the respective input information.

From Table 2, it is clear that the estimated *b*-value depends on the type of input information and the type of uncertainty considered. Figure 4 shows that historic and prehistoric data play an important role in the estimation process of the *b*-parameter. Prehistoric data contribute between 56% and 75% of the information, depending on which model is used. The instrumental data provide the second most information, closely followed by the historical data. Independent prior

TABLE 2 Output of the estimated earthquake recurrence parameters $\hat{\lambda}$ and $\hat{\beta}$ according to the mixture models for occurrence and the shifted-truncated magnitude distribution

Recurrence parameter	Scenario	Estimated parameter	Percentage contribution			
			Prehistoric	Historic	Instrumental	Prior
Mean annual rate of occurrence $\hat{\lambda}$ for $x_{\min}$= (true $\lambda = 10$)	LoC	9.7 ± 0.7	3.5	8.8	87.7	0
	LoC_MAG	9.3 ± 0.7	4.1	8.7	87.2	0
	LoC_MOD	7.4 ± 1.3	0.1	31.9	68.0	0
	LoC_MAG_MOD	7.5 ± 1.3	6.2	32.9	60.9	0
	LoC_OCC	4.9 ± 0.5	3.5	8.0	88.5	0
	LoC_MAG_OCC	4.6 ± 0.5	4.0	7.9	88.1	0
	LoC_MOD_OCC	4.6 ± 0.9	2.6	24.9	72.5	0
	LoC_MAG_MOD_OCC	4.1 ± 0.8	5.0	24.0	71.1	0
Gutenberg–Richter $\hat{b}$-value ($\hat{\beta} = b \ln(10)$)[a]	LoC	0.97 ± 0.02	7.36	10.6	13.3	2.8
	LoC_MAG	1.00 ± 0.02	68.9	13.7	14.1	3.5
	LoC_MOD	1.08 ± 0.03	68.3	11.6	14.9	5.2
	LoC_MAG_MOD	1.14 ± 0.04	65.0	14.3	14.6	6.10
	LoC_OCC	0.93 ± 0.03	65.2	11.4	17.7	5.7
	LoC_MAG_OCC	0.94 ± 0.03	62.0	13.5	18.0	6.4
	LoC_MOD_OCC	1.05 ± 0.04	58.6	13.5	18.9	9.0
	LoC_MAG_MOD_OCC	1.06 ± 0.04	56.6	15.1	18.7	9.6

[a]True Gutenberg–Richter b-value = 1.0.

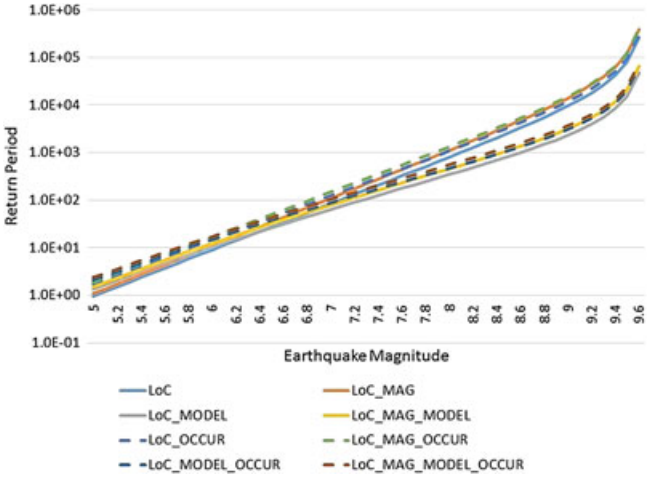

FIGURE 3 Comparison of return periods (on a log-scale) for the different model scenarios for earthquake magnitudes larger than or equal to 5.0

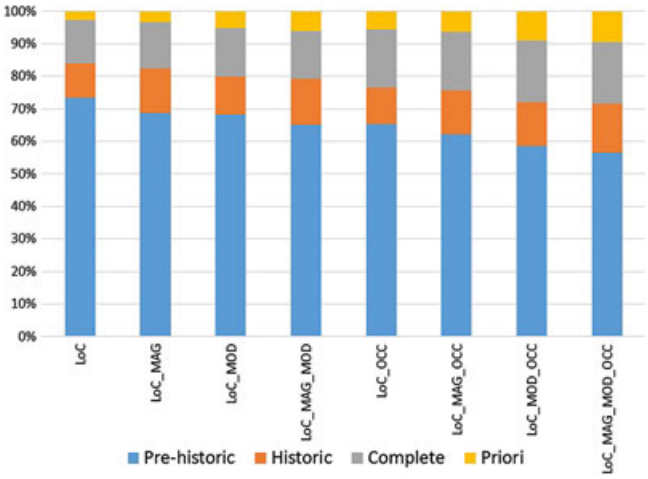

FIGURE 4 Percentage contribution of each subdataset to the Bayesian inference estimation of the Gutenberg–Richter b-value per scenario

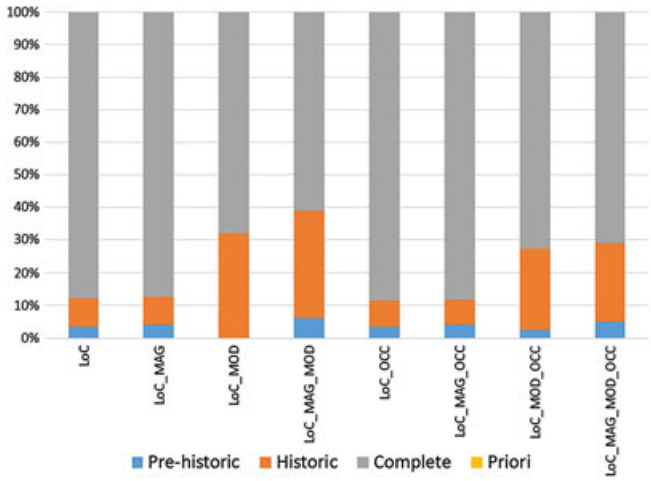

FIGURE 5 Percentage contribution of each subdataset to the Bayesian inference of the mean annual rate of occurrence per scenario. The uniform prior has no effect on the estimates

information contributes up to 10% to the estimation process. This relationship could change as the assumptions on the assumed uncertainties change.

For this example, Figure 5 shows that, unlike the estimates for the *b*-value, the estimate of the activity rate λ depends on the recent instrumental data. The contribution of the different types of data varies, depending upon the type and degree of uncertainty introduced. The presence of prehistoric and historic data with their respective uncertainties can have a notable effect on the estimated parameters. The effect of the validity of events is the smallest for the combined dataset. As an uninformative prior was used, it had no effect on the estimates.

In this example, it appears that the explicit introduction of model and earthquake magnitude uncertainties provides estimates quite close to the assumed true values of the recurrence parameters. However, the effect of data and model uncertainties manifests more clearly in the estimated return periods. The inclusion of validity of events generates higher return periods, which signifies a lower hazard. Uncertainty in size also leads to an increase of the return period, but the inclusion of model uncertainty reduces the return period (i.e., increased hazard). The introduction of the validity of events practically halved the mean annual rate of occurrence, which has a significant effect on hazard classifications.

Figure 4 shows that the inclusion of more uncertainty in the model increases the contribution of the prior information. An uninformative prior was used for the mean annual rate of occurrence λ, which has no effect on the parameter estimate. As the two recurrence parameters were solved simultaneously, the assumed prior for $\hat{\beta}$ had an indirect effect on $\hat{\lambda}$.

4.2 | The Japan tsunami dataset

The methodology was applied also to the Japan tsunami international dataset identified and analyzed in Smit et al. (2017). The dataset contains information about tsunami events occurring during the period 47 BC to 2015 that were measured according to the Soloviev-Imamura intensity scale. In addition, it contains a validity index according to the GITEC catalogue criteria. In Smit et al. (2017), the authors ignored events with a validity index less than 3. In this study, the dataset preparation followed the same process, with the exception that all events were included in the computation regardless of the validity index. Table 3 shows the input information used in this example, where the validity index is introduced into the likelihood function (Equation (16)). The recurrence parameter estimates, based on historic and instrumental data, are derived by applying the formalism of the Bayesian inference (Equation (17)). Equation (11) was used as the likelihood function for the instrumental data. Equations (9b) and (10) were used to derive the likelihood function for historic data. Table 4 provides a comparison between the estimated recurrence parameters, using the ML estimation (Smit et al., 2017) that considered model uncertainty and tsunami intensity uncertainty, with the ML and Bayesian estimates of the recurrence parameters when including the validity index. The maximum possible event size was assumed as $x_{\max} = 4.3 \pm 0.2$ as estimated in Smit et al. (2017).

Based on the above comparisons, it is clear that the introduction of the validity index and prior information can have a substantial effect on the hazard estimates, particularly on the estimation of the activity rate λ. The associated hazard increases when only the validity index is introduced. However, the introduction of prior information

TABLE 3 Probabilistic tsunami hazard assessment input parameters for Japan

	Smit et al. (2017) datasets	New datasets
Historical period	684 AD to 1960	684 AD to 1960
Number of events	79	118
LoC $\left(x_{\min}^{\text{Historic}}\right)$	1.0	1.0
Intensity SE	0.5	0.5
Complete period	1961–2011	1961–2011
Number of events	40	43
LoC $\left(x_{\min}^{\text{Complete}}\right)$	−2.0	−2.0
Intensity SE	0.1	0.1
Observed maximum intensity	4.2 ± 0.1	4.2 ± 0.1
Percentage variation	25	25
Prior information *b*-value	None	0.34
Prior information *b*-value SE	None	0.034
Coastline-characteristic intensity ($x_{\max}$)	4.3 ± 0.2	4.3 ± 0.2

TABLE 4 Return periods and probabilities of exceedance for tsunami intensities 1.5, 2.0, and 2.5 and percentage contribution of datasets to the estimates. Results are provided for the time periods 1, 10, and 25 years

Estimated recurrence parameters	Smit et al. (2017)	LoC_MOD_MAG_OCC_MLE	LoC_MOD_MAG_OCC_BI
Uncertainties accounted for	Model uncert. Intensity size	Model uncert. Intensity size Validity index	Model uncert. Intensity size Validity index
Intensity size	MLE	MLE	Bayesian inference
Mean annual rate of activity (λ)	1.5 ± 0.4	2.4 ± 0.8	1.8 ± 0.4
Frequency–magnitude *b*-value	0.4 ± 0.04	0.4 ± 0.05	0.4 ± 0.03
$x \geq 1.5$			
Return period	15	11	10
Prob. exceedance 1 year	7	0	0
Prob. exceedance 10 years	49	1	1
Prob. exceedance 25 years	80	1	1
$x \geq 2.0$			
Return period	23	17	16
Prob. exceedance 1 year	4	0	0
Prob. exceedance 10 years	35	0	1
Prob. exceedance 25 years	65	1	1
$x \geq 2.5$			
Return period	37	29	26
Prob. exceedance 1 year	3	0	0
Prob. exceedance 10 years	24	0	0
Prob. exceedance 25 years	48	1	1
Percentage contribution			
Historic *b*-value	92.9	94.3	73.1
Complete *b*-value	7.1	5.7	5.3
Prior Information *b*-value	0	0	21.6
Historic λ	98.5	94.9	98.5
Complete λ	1.5	5.1	1.5
Prior Information λ	0	0	0

(Orfanogiannaki & Papadopoulos, 2007) reduces this effect (Figure 6). The historic dataset spans more than 1,000 years compared with 50 years for the assumed instrumental dataset and, therefore, it has a larger influence on the estimates of the recurrence parameters. Both the validity index and the prior information lead to a decrease in the expected probabilities of exceedance for the tsunami intensity sizes compared with the ML estimates of Smit et al. (2017).

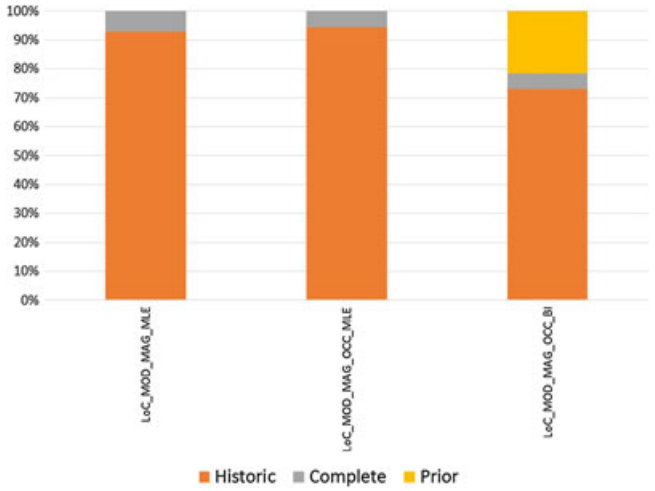

FIGURE 6 Percentage contribution of each subdataset to the maximum likelihood estimation and Bayesian inference (BI) of the *b*-value when taking into consideration the validity index associated with the Japan tsunami dataset

5 | DISCUSSION AND CONCLUSIONS

In this study, we introduced a generic methodology, based on empirical data and prior information, for modeling any type of natural hazard. The modeling process combines a nonhomogeneous Poisson distribution with the power law describing the relationship between the frequency and the sizes of events. Explicit provision is made for highly incomplete and uncertain data, the inclusion of prehistoric and historic information to constrain the results, uncertainties in event sizes, uncertainty in the applied occurrence models, and uncertainty in the occurrence of an event. The approach is applicable when event occurrence in time is nonhomogeneous or events are weakly dependent. The study shows how to different types of prior information can be combined for Bayesian inference.

The methodology was tested on a synthetic earthquake dataset, as well as a tsunami dataset for Japan. In testing the effects of the different types of uncertainty, eight model variations were applied to the synthetic dataset. The recurrence parameters, the mean activity rate of event occurrence λ, and the frequency-event size distribution parameter β were estimated using Bayesian inference while keeping the area-characteristic, maximum possible event size constant. In this example, the estimated parameters closely followed the original recurrence parameters used to generate the dataset. It was shown that the contribution of prehistoric, historic, and instrumental data, and prior information depended on the time span of the subdatasets and the combination of the different types of uncertainty introduced. Our synthetic example demonstrated that information on the rate of event occurrence λ derived mainly from the more-recent instrumental datasets, whereas the information on the power law parameter β, derived mainly from extreme events. The more uncertainty was included in the modeling process, the more the estimates depended on prior information. However, the opposite was observed in evaluating the tsunami dataset for Japan. Both model parameters λ and β relied heavily on the historic dataset, of which the time span was substantially longer than that of the instrumental dataset. The introduction of the tsunami event validity index and prior information not only increased the hazard estimates with short return periods but also decreased the probabilities of exceedance of high tsunami intensities. Prehistoric and historic information can be unreliable or unavailable. Therefore, to avoid placing too much emphasis on these datasets, the methodology allows reducing their contribution or continuing with the instrumental data only.

Three types of uncertainty were introduced to account for the lack of knowledge that often causes continuous problems in natural hazard datasets. The convolution theorem was used to introduce event size uncertainty and mixture distributions to allow for deviations from the strict Poisson and power law distributions. Employing the WL function to account for the validity of an event is an effective tool to ensure that the rates of occurrence and return periods are not overestimated.

The proposed methodology provides a useful and adaptable tool for the probabilistic assessment of various types of natural hazards by employing various modeling options to account for the different types of incompleteness and uncertainty commonly present in natural hazard datasets. The introduction of these types of uncertainty should be done prudently and with a thorough understanding of the data and the mechanics of the associated natural hazard.

ACKNOWLEDGEMENTS

This work is based on the research supported wholly/in part by the National Research Foundation of South Africa under Grants IFR160120157106, TP14072278140 (96412), and 94808. The authors thank the anonymous reviewer for the detailed input that significantly improved the manuscript.

DATA AND MATERIALS

MATLAB (https://www.mathworks.com/; last accessed 2018 03 04) was used in all computational analyzes. The synthetic data that support the findings of this study are available on request from the corresponding author. The tsunami dataset was provided by Dr. V. K. Gusiakov of the Novosibirsk Tsunami Laboratory of the Institute of Computational Mathematics and Mathematical Geophysics (NTL/ICMMG) SDRAS, Novosibirsk, Russia (HTDB/WLD, 2013).

ORCID

A. Smit https://orcid.org/0000-0003-0315-0875
A. Stein https://orcid.org/0000-0002-9456-1233
A. Kijko https://orcid.org/0000-0002-0949-0427

REFERENCES

Aki, K. (1965). Maximum likelihood estimate of *b* in the formula log $N = a\text{-}bM$ and its confidence limits. *Bulletin of the Earthquake Research Institute, University of Tokyo*, *43*, 237–239.

Bain, L. J., & Engelhardt, M. (1992). *Introduction to probability and mathematical statistics*. Pacific Grove, CA: Brooks/Cole.

Benjamin, J. R., & Cornell, C. A. (2014). *Probability, statistics, and decision for civil engineers*. Mineola, NY: Courier Corporation.

Berti, M., Martina, M. L. V., Franceschini, S., Pignone, S., Simoni, A., & Pizziolo, M. (2012). Probabilistic rainfall thresholds for landslide occurrence using a Bayesian approach. *Journal of Geophysical Research: Earth Surface*, *117*(F4), 1–20. https://doi.org/10.1029/2012JF002367

Burroughs, S. M., & Tebbens, S. F. (2001). Upper-truncated power laws in natural systems. *Pure and Applied Geophysics*, *158*(4), 741–757.

Burroughs, S. M., & Tebbens, S. F. (2005). Power-law scaling and probabilistic forecasting of tsunami runup heights. *Pure and Applied Geophysics*, *162*, 331–342.

Campbell, K. W. (1982). Bayesian analysis of extreme earthquake occurrences. Part I. Probabilistic hazard model. *Bulletin of the Seismological Society of America*, *72*, 1689–1705.

Campbell, K. W. (1983). Bayesian analysis of extreme earthquake occurrences. Part II. Application to the San Jacinto Fault Zone of Southern California. *Bulletin of the Seismological Society of America*, *73*, 1099–1115.

Cheng, R. C. H., & Traylor, L. (1995). Non-regular maximum likelihood problems. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*, 3–24.

Cooley, D., Nychka, D., & Naveau, P. (2007). Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, *102*(479), 824–840. https://doi.org/10.1198/016214506000000780

Cornell, C. A. (1968). Engineering seismic risk analysis. *Bulletin of the Seismological Society of America*, *58*, 1583–1606.

Cosentino, P., Ficara, V., & Luzio, D. (1977). Truncated exponential frequency – magnitude relationship in the earthquake statistics. *Bulletin of the Seismological Society of America*, *67*, 1615–1623.

Cramér, H. (1961). *Mathematical methods of statistics*. Princeton, NJ: Princeton University Press.

Cunningham, R. J., Herzog, T. N., & London, R. L. (2012). *Models for quantifying risk*. New Hartford, CT: ACTEX Publications.

Daykin, C. D., Pentikainen, T., & Pesonen, M. (1993). *Practical risk theory for actuaries*. Boca Raton, FL: Chapman and Hall/CRC.

Dong, W., Shah, H. C., Bao, A., & Mortgat, C. P. (1984). Utilization of geophysical information in Bayesian seismic hazard model. *International Journal of Soil Dynamics and Earthquake Engineering*, *3*(2), 103–111.

Egüen, M., Aguilar, C., Solari, S., & Losada, M. A. (2016). Non-stationary rainfall and natural flows modeling at the watershed scale. *Journal of Hydrology*, *538*, 767–782. https://doi.org/10.1016/j.jhydrol.2016.04.061

El-Isa, Z. H., & Eaton, D. W. (2014). Spatiotemporal variations in the *b*-value of earthquake magnitude–frequency distributions: Classification and causes. *Tectonophysics*, *615-616*, 1–11. https://doi.org/10.1016/j.tecto.2013.12.001

Epstein, B., & Lomnitz, C. (1966). A model for occurrence of large earthquakes. *Nature*, *211*, 954–956. https://doi.org/10.1038/211954b0

Fernandes, W., Naghettini, M., & Loschi, R. (2010). A Bayesian approach for estimating extreme flood probabilities with upper-bounded distribution functions. *Stochastic Environmental Research and Risk Assessment*, *24*(8), 1127–1143. https://doi.org/10.1007/s00477-010-0365-4

Geist, E. L., & Parsons, T. (2006). Probabilistic analysis of tsunami hazards. *Natural Hazards*, *37*, 277–314. https://doi.org/10.1007/s11069-005-4646-z

Geist, E. L., & Parsons, T. (2014). Undersampling power-law size distributions: Effect on the assessment of extreme natural hazards. *Natural Hazards*, *72*(2), 565–595. https://doi.org/10.1007/s11069-013-1024-0

Geist, E. L., & Uri, S. (2012). NRC/USGS workshop report: Landslide tsunami probability (USGS Administrative Report). Rockville, MD: US Nuclear Regulatory Commission. Retrieved from https://www.nrc.gov/docs/ML1227/ML12272A130.pdf

Gutenberg, B., & Richter, C. F. (1956). Earthquake magnitude, intensity, energy, and acceleration: (Second paper). *Bulletin of the Seismological Society of America*, *46*, 105–145.

Hamada, M. S., Wilson, A. G., Reese, C. S., & Martz, H. F. (2008). *Bayesian reliability*. New York, NY: Springer.

Khaliq, M. N., Ouarda, T. B. M. J., Ondo, J. C., Gachon, P., & Bobée, B. (2006). Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review. *Journal of Hydrology*, *329*(3–4), 534–552. https://doi.org/10.1016/j.jhydrol.2006.03.004

Kijko, A., & Graham, G. (1999). "Parametric-historic" procedure for probabilistic seismic hazard analysis. Part II: Assessment of seismic hazard at specified site. *Pure and Applied Geophysics*, *154*(1), 1–22.

Kijko, A., & Sellevoll, M. A. (1989). Estimation of earthquake hazard parameters from incomplete data files. Part I. Utilization of extreme and complete catalogues with different threshold magnitudes. *Bulletin of the Seismological Society of America*, *79*, 645–654.

Kijko, A., & Sellevoll, M. A. (1992). Estimation of earthquake hazard parameters from incomplete data files. Part II. Incorporation of magnitude heterogeneity. *Bulletin of the Seismological Society of America*, *82*(1), 120–134. https://pubs.geoscienceworld.org/ssa/bssa/article/82/1/120/119509/estimation-of-earthquake-hazard-parameters-from

Kijko, A., & Singh, M. (2011). Statistical tools for maximum possible earthquake magnitude estimation. *Acta Geophysica*, *59*, 674–700. https://doi.org/10.2478/s11600-011-0012-6

Kijko, A., & Smit, A. (2012). Extension of the Aki-Utsu *b*-value estimator for incomplete catalogs. *Bulletin of the Seismological Society of America*, *102*, 1283–1287. https://doi.org/10.1785/0120110226

Kijko, A., Smit, A., & Sellevoll, M. A. (2016). Estimation of earthquake hazard parameters from incomplete data files. Part III. Incorporation of uncertainty of earthquake-occurrence model. *Bulletin of the Seismological Society of America*, *106*(3), 1210–1222. https://doi.org/10.1785/0120150252

Klugman, S. A., Panjer, H. H., & Wilmot, G. E. (2004). *Loss models: From data to decisions* (2nd ed.). New York, NY: John Wiley & Sons.

Lam, D., Thompson, C., Croke, J., Sharma, A., & Macklin, M. (2017). Reducing uncertainty with flood frequency analysis: The contribution of paleoflood and historical flood information. *Water Resources Research*, *53*(3), 2312–2327. https://doi.org/10.1002/2016WR019959

Lyubushin, A. A., & Parvez, I. A. (2010). Map of seismic hazard of India using Bayesian approach. *Natural Hazards*, *55*(2), 543–556. https://doi.org/10.1007/s11069-010-9546-1

Malamud, B. D., Turcotte, D. L., Guzzetti, F., & Reichenbach, P. (2004). Landslide inventories and their statistical properties. *Earth Surface Processes and Landforms*, *29*(6), 687–711. https://doi.org/10.1002/esp.1064

Markatou, M., Basu, A., & Lindsay, B. G. (1998). Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, *93*, 740–750. https://doi.org/10.1080/01621459.1998.10473726

Mortgat, C. P., & Shah, H. C. (1979). A Bayesian model for seismic hazard mapping. *Bulletin of the Seismological Society of America*, *69*(4), 1237–1251.

Newman, M. E. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351. https://doi.org/10.1080/00107510500052444

Ogata, Y., & Abe, K. (1991). Some statistical features of the long-term variation of the global and regional seismic activity. *International Statistical Review/Revue Internationale de Statistique*, *59*, 139–161.

Orfanogiannaki, K., & Papadopoulos, G. A. (2007). Conditional probability approach of the assessment of tsunami potential: Application in three tsunamigenic regions of the Pacific Ocean. *Pure and Applied Geophysics*, *164*(2–3), 593–603. https://doi.org/10.1007/s00024-006-0170-7

Patskoski, J., & Sankarasubramanian, A. (2018). Reducing uncertainty in stochastic streamflow generation and reservoir sizing by combining observed, reconstructed and projected streamflow. *Stochastic Environmental Research and Risk Assessment*, *32*(4), 1065–1083. https://doi.org/10.1007/s00477-017-1456-2

Pisarenko, V. F., & Lyubushin, A. A. (1997). Statistical estimation of maximum peak ground acceleration at a given point of a seismic region. *Journal of Seismology*, *1*, 395–405.

Pisarenko, V. F., Lyubushin, A. A., Lysenko, V. B., & Golubeva, T. V. (1996). Statistical estimation of seismic hazard parameters: Maximum possible magnitude and related parameters. *Bulletin of the Seismological Society of America*, *86*(3), 691–700.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley & Sons.

Scholz, C. H. (2015). On the stress dependence of the earthquake *b* value. *Geophysical Research Letters*, *42*(5), 1399–1402. https://doi.org/10.1002/2014GL062863

Shi, K., & Liu, C.-Q. (2009). Self-organized criticality of air pollution. *Atmospheric Environment*, *43*(21), 3301–3304. https://doi.org/10.1016/j.atmosenv.2009.04.013

Silva, A. T., Portela, M. M., Naghettini, M., & Fernandes, W. (2017). A Bayesian peaks-over-threshold analysis of floods in the Itajaí-açu River under stationarity and nonstationarity. *Stochastic Environmental Research and Risk Assessment*, *31*(1), 185–204. https://doi.org/10.1007/s00477-015-1184-4

Smit, A., Kijko, A., & Stein, A. (2017). Probabilistic tsunami hazard assessment from incomplete and uncertain historical catalogues with application to tsunamigenic regions in the Pacific Ocean. *Pure and Applied Geophysics*, *174*(8), 3065–3081. https://doi.org/10.1007/s00024-017-1564-4

Soloviev, S. L. (1970). Recurrence of tsunamis in the Pacific. In W. M. Adams (Ed.), *Tsunamis in the Pacific Ocean* (pp. 149–163). Honolulu, HI: East-West Centre Press.

Stedinger, J. R., & Cohn, T. A. (1986). Flood frequency analysis with historical and paleoflood information. *Water Resources Research*, *22*(5), 785–793. https://doi.org/10.1029/WR022i005p00785

Sun, X., Thyer, M., Renard, B., & Lang, M. (2014). A general regional frequency analysis framework for quantifying local-scale climate effects: A case study of ENSO effects on Southeast Queensland rainfall. *Journal of Hydrology*, *512*, 53–68. https://doi.org/10.1016/j.jhydrol.2014.02.025

Tatard, L., Grasso, J. R., Helmstetter, A., & Garambois, S. (2010). Characterization and comparison of landslide triggering in different tectonic and climatic settings. *Journal of Geophysical Research: Earth Surface H*, *115*(F4), 1–18. https://doi.org/10.1029/2009JF001624

Tinti, S., Maramai, A., & Graziani, L. (2001). A new version of the European tsunami catalogue: Updating and revision. *Natural Hazards and Earth System Science*, *1*(4), 255–262.

Vermeulen, P. J., & Kijko, A. (2017). More statistical tools for maximum possible earthquake magnitude estimation. *Acta Geophysica*, *65*(4), 579–587. https://doi.org/10.1007/s11600-017-0048-3

Vicini, L., Hotta, L. K., & Achar, J. A. (2013). Non-homogeneous Poisson process in the presence of one or more change points: An application to air pollution data. *Journal of Environmental Statistics*, *5*(3). http://www.jenvstat.org/v05/i03

Viglione, A., Merz, R., Salinas, J. L., & Blöschl, G. (2013). Flood frequency hydrology: 3. A Bayesian analysis. *Water Resources Research*, *49*(2), 675–692. https://doi.org/10.1029/2011wr010782

Yadav, R. B. S., Tsapanos, T. M., Tripathi, J. N., & Chopra, S. (2013). An evaluation of tsunami hazard using Bayesian approach in the Indian Ocean. *Tectonophysics*, *593*, 172–182. https://doi.org/10.1016/j.tecto.2013.03.004

Yazdani, A., & Kowsari, M. (2013). Bayesian estimation of seismic hazards in Iran. *Scientia Iranica*, *20*(3), 422–430. https://doi.org/10.1016/j.scient.2012.12.032

How to cite this article: Smit A, Stein A, Kijko A. Bayesian inference in natural hazard analysis for incomplete and uncertain data. *Environmetrics*. 2019;e2566. https://doi.org/10.1002/env.2566