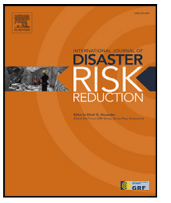
International Journal of Disaster Risk Reduction 132 (2026) 105947

Contents lists available at ScienceDirect

International Journal of Disaster Risk Reduction

journal homepage: www.elsevier.com/locate/ijdrr



A flexible simulation-based predictive approach to compare hazard and risk models: An example application to seismic hazard

Francesco Pauli [a],*, Stefano Parolai [b]

[a] Dipartimento di Scienze Economiche, Aziendali, Matematiche e Statistiche 'B. de Finetti', University of Trieste, Italy
[b] Dipartimento di Matematica, Informatica e Geoscienze, University of Trieste, Italy

ARTICLE INFO

Keywords:
Probabilistic Seismic Hazard Assessment (PSHA)
Peak ground acceleration
Tail-focused statistics
Simulation-based discrepancy profiling
Predictive *p*-value
Aleatory and epistemic uncertainty
Hazard curve discrepancy
Decision-support tools

ABSTRACT

We propose a simulation-based approach to compare probabilistic hazard and risk models, exploiting the Bayesian prior/posterior predictive p-values (PPP) framework. The comparison can utilize an arbitrary summary statistic and can be customized to the aspects of interest, particularly the right tail, which is crucial in risk assessment. The primary benefits of our approach in comparison to existing alternatives are twofold. Firstly, it incorporates both aleatory and epistemic variability in a natural probabilistic framework, secondly, it produces interpretable measures of discrepancy. The method is demonstrated on synthetic data and two state-of-the-art seismic hazard models for Italy (MPS19, Modello di Pericolosità Sismica 2019, and ESHM20, European Seismic Hazard Model 2020). The method is applicable in any domain involving probabilistic hazard or risk models, including flood, volcanic, or multi-layer single hazard or single risk assessments.

1. Introduction

Hazard or risk assessment of natural disasters must necessarily be carried out using probabilistic approaches to facilitate efficient land use planning or cost assessment over extended periods (e.g. [1]). There are two components of the uncertainty: the aleatory uncertainty, due to the randomness of the physical phenomenon, represented by exceedance probability curves, and the epistemic uncertainty of the curves themselves, due to the lack of knowledge in defining their main ingredients, which leads to a model being a set of exceedance probability curves (usually called ensemble or logic tree depending on the application context). Consequently, the interpretation of results frequently entails the consideration of mean or median curves, despite the fact that the distribution of the hazard/risk curves contains information of great importance. This information is essential for assessing models' differences and prioritizing choices based on risk estimation. In fact, an overview of the risks affecting a certain area and their correct ranking is of fundamental importance in enabling planners and decision-makers to make adequate decisions on risk reduction and loss prevention programs (e.g. [2–5]). In recent years, there has been an increased focus on the validation of probabilistic hazard models (e.g., [6–9]), the estimation of differences between different hazard models for the same natural hazard (e.g., [10]), and the comparison of models in the context of single-type hazard multi-risk assessment [5]. In particular, two approaches have been used in the latter two cases: one requiring an a priori assumption about the distribution (usually normal or log normal of the curves) and the other a nonparametric approach (independent distribution). However, these may not be optimal, especially when the comparison involves models comprising ensembles of exceedance probability curves of very different cardinality.

In this paper, we propose an extension of the approach proposed by Marzocchi and Jordan [8] and further developed by Marzocchi et al. [9] in the context of seismic hazard models to address these limitations and facilitate the comparison of the results

* Corresponding author.
E-mail address: francesco.pauli@deams.units.it (F. Pauli).

https://doi.org/10.1016/j.ijdrr.2025.105947
Received 28 August 2025; Received in revised form 29 November 2025; Accepted 1 December 2025
Available online 2 December 2025

of different models related to the same hazard. We note that the approach may be exploited also to facilitate the comparison of the annual exceedance probabilities of losses, although we do not pursue this objective in the present study. The proposal is predicated on a methodological approach utilized in Bayesian statistics for the evaluation of model adequacy. Specifically, observations are juxtaposed with the model based on the predictive distribution, and an indicator is computed. This methodology is then employed to facilitate model comparison by repeatedly simulating from one model and comparing each simulated sample to the alternative model. The proposed methodology is applied in an illustrative way to the results for different locations of two seismic hazard models developed for the Italian territory, the MPS19 (Modello di Pericolosità Sismica 2019, [11]) and the ESHM20 (European Seismic Hazard Model 2020, [12]). Note that we do not aim at performing a comparison of the two seismic models, which would entail a more thorough discussion, rather, we employ them to offer an illustrative example of an application of the methodology we propose. The ensuing discussion will address the results and the potential for application of the aforementioned method.

2. Hazard/risk model(s)

A model $\mathcal{H}_i = \{H_1^{(i)}, \dots, H_{M_i}^{(i)}\}$ is a collection of M_i hazard or risk curves that represent either the mean annual rate of exceedance or the probability of exceedance over a specified time window. These curves are weighted to reflect the epistemic uncertainty associated with the model parameters. Conditional on a specific hazard or risk curve $H_m^{(i)}$, let

$$\bar{F}(y|H_m^{(i)}) = 1 - F(y|H_m^{(i)}) = P\left(Y > y|H_m^{(i)}\right); \quad m = 1, \dots, M_i \tag{1}$$

be the hazard curves representing the conditional probability, and let

$$\pi_m^{(i)} = P(H_m^{(i)}) \tag{2}$$

be the weights, which can be interpreted as a probability distribution over the submodels $H_m^{(i)}$. The light gray curves in Fig. 1 represent the hazard or risk functions $\bar{F}^{(i)}(y|\lambda)$ for a toy model. More generally, a model is a collection of risk or hazard functions indexed by a parameter λ.

$$\bar{F}(y|\lambda) = 1 - F(y|\lambda) = P(Y > y|\lambda) \tag{3}$$

where

$$\lambda \sim G^{(i)}(\lambda) \tag{4}$$

is the distribution function of λ (representing the epistemic uncertainty in lieu of (2), and where $\sim$ stands for "distributed as"). The preceding scenario (illustrated by Eqs. (1) and (2)) corresponds to a discrete probability distribution on λ. While the models under consideration are specified as in (1) and (2), the more general notation (3) and (4) simplifies the description of the method.

The average hazard implied by the model can be calculated as follows:

$$F^{(i)}(y) = \int F_\lambda^{(i)}(y|\lambda) dG^{(i)}(\lambda) = \sum_m \pi_m^{(i)} F_m^{(i)}(y) \tag{5}$$

It should be noted that the mean hazard is sometimes obtained as $F^{(i)}(y|\bar{H}^{(i)}) = F^{(i)}(y|\bar{\lambda}_i)$ where $\bar{\lambda}_i$ is the mean or the median of $G^{(i)}(\lambda)$. The mean hazard for the model in Fig. 1 is represented by a thick black line.

Each hazard/distribution function, denoted $\bar{F}/F$, represents the aleatory uncertainty of the phenomenon y (the measure of ground motion intensity, in this application, the peak ground acceleration (PGA) with reference period of one year) under a specific set of hypotheses concerning the various model components (functional forms, parameter values, etc.). The multiplicity of hazard/distribution functions is due to the epistemic uncertainty affecting the different model components.

It is commonplace to summarize the model using either

- the distribution of the probability of exceeding a given threshold (blue curve in Fig. 1, which is an example for PGA). This distribution is defined as the distribution of the annual probability of exceedance (APE) $\bar{F}^{(i)}(y_0|\lambda)$, which is a transformation of the random variable λ. The distribution function of this quantity is denoted by $K^{(i)}_{\mathrm{APE},y_0}(\cdot)$;
- the distribution of the quantile corresponding to a certain APE, illustrated by the red curve in Fig. 1, which is an example for PGA. This distribution is defined as the distribution of the inverse function of $F^{(i)}$: $\overline{F}^{(i)}(1 - q_0|\lambda)$ that is also a transformation of the random variable λ. The distribution function of this quantity is denoted by $Q^{(i)}_{q_0}(\cdot)$.

It is evident that both probability distributions effectively capture the epistemic uncertainty on the quantity of interest. In the first instance, this uncertainty is represented by a probability, while in the second instance, it is denoted by a quantile.

In order to comprehend the comparative contributions of aleatory uncertainty and epistemic uncertainty one must consider a sample of PGAs: $y_1, \dots, y_n$ and potentially a statistic $T(y_1, \dots, y_n) : \mathbb{R}^n \to \mathbb{R}$, typically univariate, to summarize the sample. The purpose of the summary statistic $T()$ is to focus the analysis on a univariate random variable. In principle any real-valued function of $y_1, \dots, y_n$ could be used; however, for the discussion to be practically relevant, one should choose a summary statistic of substantial interest. Using the sample mean, for example, one focuses on the average hazard. If interest lies on the extreme behavior, one should prefer summary statistics such as the number of exceedances of a high threshold. Furthermore, in order to fully explore a model, multiple statistics ought to be considered, focusing on different model features.

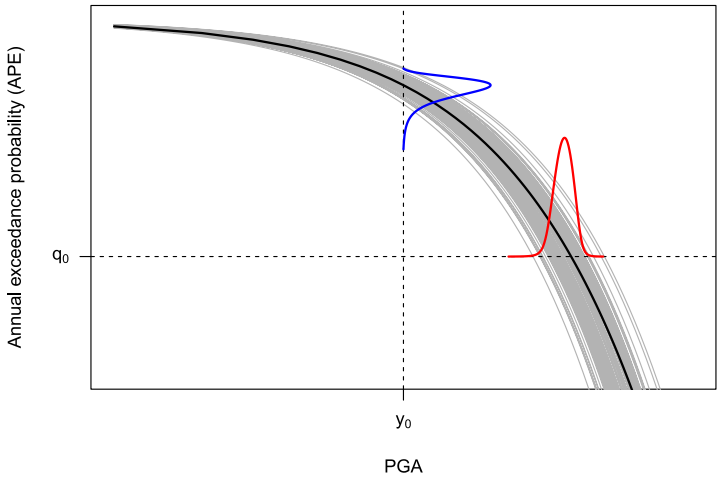


Fig. 1. Depiction of a single model, by means of the hazard functions in light gray. The black (thick) line is the mean hazard curve. The (probability density functions (PDFs) represent the conditional distribution of the Annual Frequency of Exceedance (AFE) or the annual probability of exceedance (APE) for a given parameter. In this case the parameter is exemplarily chosen as the peak ground acceleration, (PGA) in the context of earthquake hazards (PDFs of $K_{AFE,y_0}(\cdot)$, blue) and the conditional distribution of a PGA's quantile for a given probability (frequency), that is, the PDF of $Q_{q_0}(\cdot)$ (red). Both PDFs are transformations of the distribution of λ, representing epistemic uncertainty.

The distribution of T can be considered based on the mean hazard, $F_T^{(i)}(t|\bar{\lambda})$ or it can be considered as a mixture of the distributions conditional on λ,

$$F_T^{(i)}(t) = \int F_T^{(i)}(t|\lambda) dG^{(i)}(\lambda). \tag{6}$$

Quantity (6) accounts for both epistemic and aleatory uncertainty; quantity $F_T^{(i)}(t|\bar{\lambda})$ exclusively considers aleatory uncertainty conditional on an average hazard. Fig. 2 presents several examples from the toy model in Fig. 1 for some summary statistics $T()$. Regardless of the statistic under consideration, the conditional distribution $F_T^{(i)}(t|\bar{\lambda})$ exhibits less variability in comparison to the unconditional distribution $F_T^{(i)}(t)$. This discrepancy can be attributed to the fact that the latter considers both the aleatory uncertainty and the epistemic uncertainty, while the former neglects epistemic uncertainty.

Suppose now that two models $\mathcal{H}_1$ and $\mathcal{H}_2$ are available–$\mathcal{H}_2$ may be an update of model $\mathcal{H}_1$–as in Fig. 3. The objective of this study is to assess whether the new model $\mathcal{H}_2$ implies a relevant change in hazard estimates. In other words, the objective is to 'measure' the difference between the two models and possibly decide whether the difference is relevant in some sense.

The question's lack of an obvious answer is indicative of its complexity, stemming from the presence of both aleatory uncertainty, as represented by each hazard function, and epistemic uncertainty, which gives rise to the multiplicity of hazard functions. In essence, the task entails a comparison of two sets of distribution functions.

3. Model comparison in seismological literature

A substantial body of literature exists that compares hazard models. In this section, the focus is on a comparison of seismological models and a review of several methods.

Abrahamson [15] proposed two criteria. The first criterion entails a comparison of the mean hazard implied by model $\mathcal{H}_1$ for a specified q_0 value of the APE: $\bar{Y}_{q_0}^{(1)}$, with the distribution of the hazard for the same APE implied by model 2: $Q_{q_0}^{(2)}()$. The change is considered to be "robust" if $\bar{Y}_{q_0}^{(1)}$ lies outside the interquartile range of $Q_{q_0}^{(2)}() : [\overline{Q}_{q_0}^{(2)}(0.25); \overline{Q}_{q_0}^{(2)}(0.75)]$. The criterion is depicted in Fig. 4. Note that $\bar{Y}_{q_0}^{(1)}$ can be obtained either as $\bar{Y}_{q_0}^{(1)} = \overline{F}^{(1)}(1-q_0|\bar{\lambda})$ or $\bar{Y}_{q_0}^{(1)} = E(\overline{F}^{(1)}(1-q_0|\lambda)) = \int \overline{F}^{(1)}(1-q_0|\lambda) dG^{(1)}(\lambda)$ with slightly different results (see Eq. (5) in Section 2).

The second criterion proposed by Abrahamson [15] involves the mean hazards $\bar{Y}_{q_0}^{(1)}, \bar{Y}_{q_0}^{(2)}$. A change is considered robust if

$$\mathrm{ABR2}_{q_0} = \ln\left(\frac{\bar{Y}_{q_0}^{(2)}}{\bar{Y}_{q_0}^{(1)}}\right) - 0.5\sqrt{\sigma^2} > 0, \tag{7}$$

where σ^2 is the variance of $Q_{q_0}^{(2)}()$. To elaborate, the author is referring to a variance obtained from the logarithms of the fractiles which is used to compare the variation in the logarithm of the PGA to the uncertainty. The rationale behind the selection of the specific threshold remains ambiguous. However, it is customary to consider the difference in the quantity divided by the standard deviation. It is important to note that this second formula can detect as "robust" only changes implying an increased mean hazard.

Malhotra [13,16] poses the question whether a change in the hazard estimates is "significant enough to warrant a change in design ground motions". The author argues that the answer can be found in the effect size as measured by Cohen's d. Given two

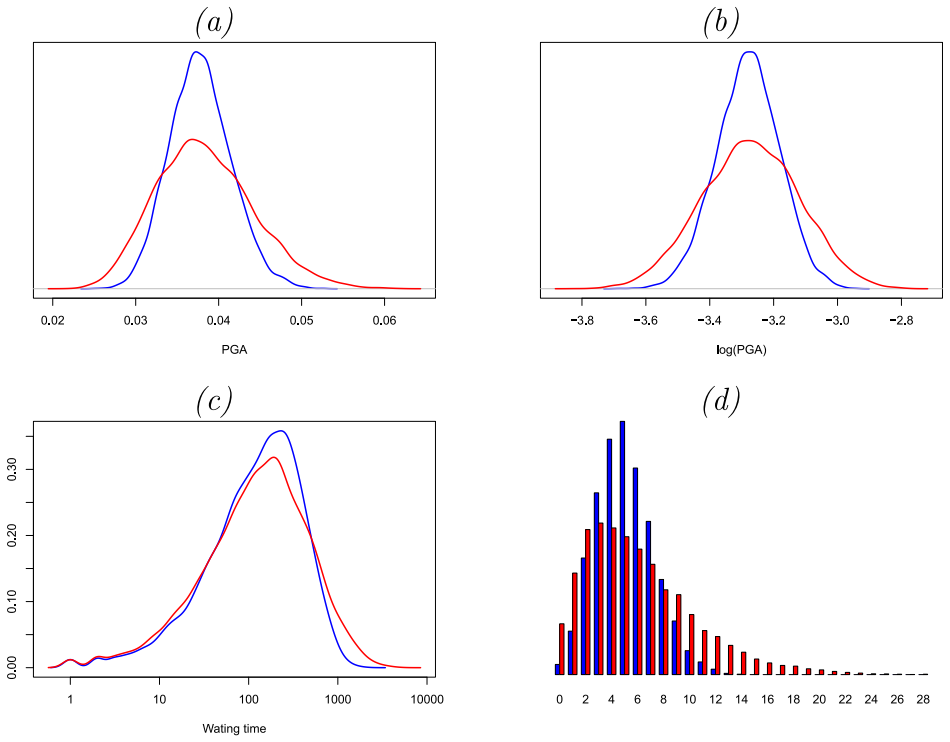


Fig. 2. Conditional ($F_T^{(i)}(t|\bar{\lambda})$, blue) and unconditional ($F_T^{(i)}(t)$, red) distributions of (a,b) the mean value of $y_1, \ldots, y_n$ for $n = 100$; (c) the waiting time for the (first) exceedance of the threshold y_0; (d) the number of exceedances of the threshold y_0 in a sample of size 1000. (All distributions are based on 10 000 simulations.)

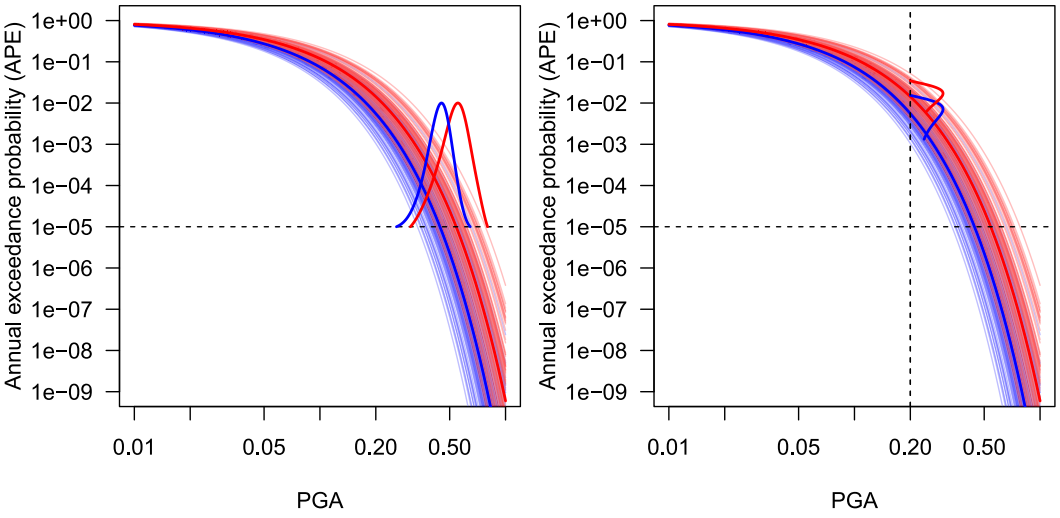


Fig. 3. Two models, that is, two sets of hazard functions. The two PDFs parallel to the x-axis in the left panel are the PDFs of $Q_{q_0}(\cdot)$ according to the two models (these are the distributions compared according to Malhotra [13] and Lund et al. [14]), the two PDFs parallel to the y-axis in the right panel represent the PDFs of $K_{AFE,0.2}(\cdot)$.

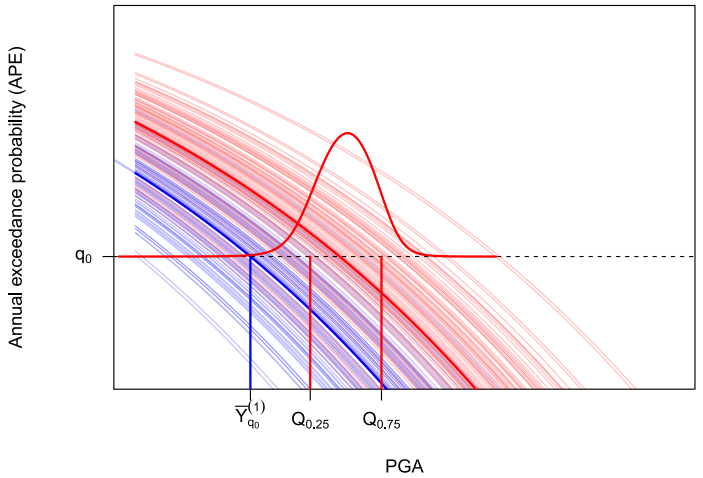


Fig. 4. The first criterion, as outlined in [15], entailing a comparison of the mean PGA for an APE equal to q_0 according to model 1: $\bar{Y}^{(1)}_{q_0}$ with the distribution of the PGA for an identical APE according to model 2 ($Q^{(2)}_{q_0}$, PDF in red). In particular the criterion stipulates that a change is deemed robust if $\bar{Y}^{(1)}_{q_0}$ is outside the interquartile range of $Q^{(2)}_{q_0}$: $[Q_{0.25}, Q_{0.75}]$, which is the case in this instance.

probability distributions, Cohen's d is the difference between the two means divided by the square root of the average of the variances

$$d = \frac{\mu_1 - \mu_2}{\sqrt{0.5(\sigma_1^2 + \sigma_2^2)}}. \tag{8}$$

Although there is no compelling reference value for Cohen d [17], a value of $|d|$ above 0.8 is widely considered to indicate a substantial difference. Malhotra [13] utilizes Cohen d to compare the distributions of the maximum PGA over a 50-year period in Los Angeles according to several updates of the USGS model. Lund [14] employed the same criterion to compare the distributions of the PGA at different return levels (in our notation $Q^{(1)}_q()$ and $Q^{(2)}_q()$ for various values of q, Fig. 3, left panel).

The aforementioned proposals [13–16] all rely on the comparison of $Q_q()$ distributions for a specific APE (return probability). In [18,19], the distributions of the APE (probability) are considered.

Douglas et al. [19] (Fig. 1b of the paper) employed the Cohen d to compare the distributions of the APE at different PGAs (corresponding to the means). To be more precise, the mean hazard curve is obtained for each model and, consequently, the mean PGAs are obtained

$$Y^{(i)}_{\bullet} = E(F^{(i)}(y)) = \sum_m \pi^{(i)}_m E(F^{(i)}_m(y)). \tag{9}$$

Subsequently, a comparison of the $K^{(i)}_{\mathrm{AFE},Y^{(i)}_{\bullet}}(q)$ distributions is carried out (Fig. 5). McGuire [18] proposed considering the mean APEs implied by the new model for ground motions corresponding to 10^{-4} and 10^{-6} annual rates in the old model. More precisely, the ground motion corresponding to the value 10^{-4} of the APE is computed according to the old model: $Y^{(1)}_{10^{-4}} = \overline{F}^{(1)}(1 - 10^{-4}|\bar{\lambda})$ (alternatively $Y^{(1)}_{10^{-4}} = \int \overline{F}^{(1)}(1 - 10^{-4}|\lambda) dG^{(1)}(\lambda)$). Subsequently, the APE corresponding to the ground motion $Y^{(1)}_{10^{-4}}$ is computed according to the new model: $\mathrm{AFE}^{new}_{10^{-4}} = F^{(2)}(\bar{\lambda}_2)$ (alternatively $\mathrm{AFE}^{new}_{10^{-4}} = \int F^{(2)}(\lambda) dG^{(2)}(\lambda)$). The same quantities are obtained for the value 10^{-6} of the APE. McGuire [18] proposal posits that the new and old model should not be deemed substantially divergent if the mean APE corresponding to a mean annual rate of 10^{-4} undergoes a changes of less than 25% and the mean APE corresponding to a mean annual rate of 10^{-6} exhibits a changes of less than 35%. The percentages are derived from an uncertainty analysis, the validity of which is contingent upon the specific type of study and the specified hypotheses regarding the model's parameters. This proposal is characterized by its specificity, which precludes the possibility of generalization. Also, it might refer to calculation of hazard for nuclear power plant, which is the reason why lower APEs than those typical of ordinary seismic hazard calculations are used. Consequently, it is not considered in the application.

As Douglas et al. [10] observe, the absence of a clear justification for any of these criteria is evident, with none pertaining to the notion of statistical significance due to the absence of an appropriate probabilistic framework. Instead, they function as heuristics, as also noted by Lund [14]. Cohen's d is the most standard measure among the ones reviewed; Cohen d is an index and not a statistical test, and thus, it does not rely on any assumption. However, it only considers second-order moments, which neglects possible skewness and kurtosis or higher-order characteristics (in a sense, it is most appropriate for Gaussian distributions).

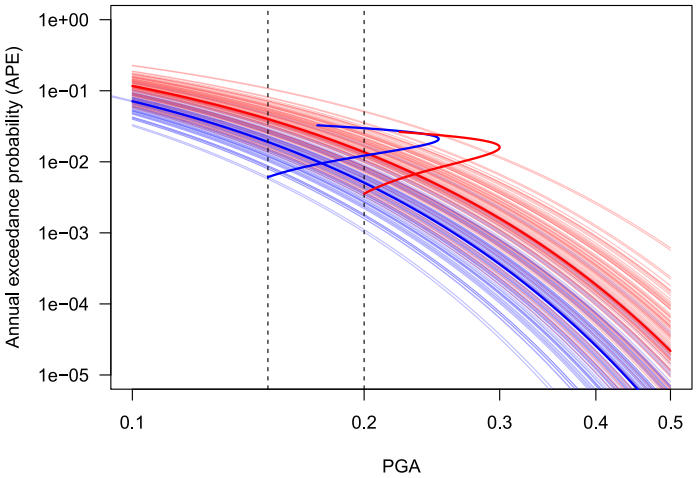


Fig. 5. Distributions of $F^{(i)}(y_0|\lambda)$ according to the two models, compared according to Douglas et al. [10].

4. Predictive *p*-value comparison

The proposed method is predicated on the framework developed by Marzocchi and Jordan [8,20] and Marzocchi et al. [9] to validate a single model by comparing it to actual observations. In this paper we extended to the comparison of two different models by employing simulated observations. Marzocchi and Jordan note that a comparison of a given sample $y = (y_1, \dots, y_n)$ (e.g. measures of ground motion intensity from actual earthquakes), possibly summarized by a statistic $T(\cdot)$, where we denote the observed value of the statistic as $t^* = T(y_1, \dots, y_n)$, with the mean hazard curve, by using a p-value calculated as follows

$$P(T(y) > T(y^*)|\bar{\lambda}_i) = 1 - F_T^{(i)}(t|\bar{\lambda}) \tag{10}$$

is erroneous because it fails to take into account epistemic uncertainty. Furthermore, they argue that the correct p-value for consideration would be instead

$$\int P(T(y) > T(y^*)|\lambda) dG^{(i)}(\lambda) = 1 - F_T^{(i)}(t^*). \tag{11}$$

where the calculation of $P(T(y) > T(y^*)|\lambda)$ is performed under the assumption that $y \sim F^{(i)}(\cdot|\lambda)$. The quantity in Eq. (11), which we call unconditional p-value, is ordinarily less extreme (approaching 0.5) than the quantity in Eq. (10), which we call conditional p-value), due to epistemic (ontological) uncertainty. In the context of Bayesian statistics, the quantity (11) is defined as a (prior/posterior) predictive p-value, as articulated by Gelman et al. [21,22]. This quantity serves as a reference to assess the adequacy of a model. The notion of the prior/posterior predictive p-value (PPP) is predicated on the assumption that if the model specified by Eqs. (3) and (4) adequately describes the phenomenon, then observations y^* should not be surprising according to the model (an observation being surprising if its discrepancy from expected behavior is very unlikely to occur). In order to assess this, one must simulate from the model and compare the simulated samples to the observed ones using a relevant summary statistic $T()$: the idea is that if the model consistently simulates samples that lead to a statistic $T()$ which is higher (lower) than the observed one then the model is not an adequate description of the phenomenon. Conversely, a value of the predictive p-value near 0.5 indicates a satisfactory fit. Different aspects of the model can be probed using different summary statistics $T()$, for instance using $T(y) = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ one checks that the model captures the mean behavior of the phenomenon, using $T(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ one checks that the model captures the variability of the phenomenon. Marzocchi and Jordan [20] employed the number of exceedances of a threshold, $T(y) = \#\{y > y_0\}$, as a metric to assess the fit in the tail of the distribution. This approach, utilizing a sufficiently high threshold, is particularly well-suited for earthquake modeling, as it emphasizes the fit in the tail of the distribution, which is of greater interest in this field than, for example, the mean.

Marzocchi and Jordan [20] consider (11) as an unconditional version of a unilateral p-value in that they rate low values as evidence against the model. This interpretation, although correct, is not fully in line with the standard way of using the predictive value p in (Bayesian) statistics, where values far from 0.5 are indications of the model discrepancy and no precise cut point is available. Following the latter approach, we think one can get more.

The PPP is a measure of discrepancy between a model and the observations. We propose an extension to the aforementioned methodology, which would entail the comparison of two models through the repeated measurement of the discrepancy between data simulated from one model (more precisely, from the hazard curves of one model) and the other. In essence, the methodology involves the repeated simulation from the first model and the subsequent assessment of the compatibility of each simulated dataset with the second model. This assessment is achieved by computing a PPP. The average of the PPPs over a large number of simulated samples provides a comprehensive measure of the compatibility between the two models

To be more precise, the strategy that is proposed here for the purpose of comparing models $\mathcal{H}_1$ and a model $\mathcal{H}_2$ is to simulate samples y^* according to model $\mathcal{H}_1$, that is, according to the mixture distribution

$$F^{(1)}(y) = \int F_{\lambda}^{(1)}(y|\lambda) dG^{(1)}(\lambda) = \sum_m \pi_m^{(1)} F_m^{(1)}(y). \tag{12}$$

Subsequently, after selecting a statistic $T(y)$ and defining $t^* = T(y)$ as the value of that statistic for the simulated sample, one computes the predictive p-value by means of the following procedure

$$\int P(T(y) > T(y^*)|\lambda, \mathcal{H}_2) dG^{(2)}(\lambda) = 1 - F_T^{(2)}(t^*). \tag{13}$$

This procedure is then repeated for a number B of simulated samples, and the values are averaged, thus obtaining an average measure of discrepancy between the two models. In summary, the following calculation is made:

$$\int \int \int P(T(y) > T(y^*)|\lambda, \mathcal{H}_2) dG^{(2)}(\lambda) dF_{\lambda}^{(1)}(y^*|\lambda) dG^{(1)}(\lambda). \tag{14}$$

A detailed description of the procedure is in Appendix B. As previously discussed, the utilization of distinct statistics, denoted T(), for the purpose of summarizing the sample enables the customization of the model comparison to diverse aspects of the model. In the context of earthquake models, it is appropriate to focus the comparison on the distribution's right tail. We subsequently propose the utilization of the maximum of a sample of size n, for varying values of n, and the waiting time for the initial exceedance of a high threshold, for different thresholds (the waiting time for the exceedance of the threshold y_0 being $T(y_1, \dots, y_n) = \min\{i | y_i > y_0\}$). An alternative that merits consideration is the empirical quantile.

Interpretation of the PPP for model comparison therefore follows the same logic as the interpretation of the PPP for model adequacy and is loosely based on two alternative points of view which have been discussed thoroughly in the literature [21–24]. In the following we adapt the guidelines and suggestions put forward in the literature to the present implementation.

In general, a PPP can be considered as an average of traditional p-values on the distribution of an unknown parameter (in this case, the distribution $G(\lambda)$). With this in mind, the value of the PPP can be judged to be similar to that of a p-value. Therefore, it is possible to use traditional significance thresholds (10%, 5%, 1%) and conclude that the two models diverge "significantly" when the PPP is less than 0.05 or greater than 0.95. This, however, creates a false dichotomy in which the two models are either different or they are not. A more promising approach is to use the PPP as a measure of discrepancy based on its interpretation as probability. In this sense, for example, a PPP of 60% means that if we could observe many replicas of the $T()$ statistic according to both model 1 and model 2, we would expect 60% of the replicas of model 1 to be greater than the replicas of model 2. Consider a quantity relevant for decision-making purposes, such as, in the seismic context, the 1/475 quantile (typically related to a 10% probability of exceedance in 50 years), which is relevant for infrastructure design. In this case, the two models will typically offer two different estimates of the same quantity.

To assess whether the difference in estimates is significant, standard statistical verification criteria are not easily applicable due to the presence of random and epistemic uncertainty. Indices such as Cohen's d are more appropriate, but beyond conventional thresholds they have no substantial meaning. Using the PPP approach, we can compare the empirical counterpart of the estimated quantities, such as the maximum in 475 years as the empirical counterpart of the 1/475 quantile, based on the frequency with which one model would imply a higher value than the other. A value close to 0 or 1 reveals that, according to one model, the experienced value would be almost systematically lower or higher than the other model. A value close to 0.5 indicates that the two models are very similar: the replicated values of $T()$ according to one model are higher than the replicated values according to the second model in 50% of cases. Intermediate values measure the discrepancy on a probability scale; values greater (less) than 0.5 indicate that the first model implies higher (lower) values for $T()$ on average, the higher (lower) the PPP value is. Although it is not possible to provide universally valid thresholds, it is reasonable to say that a value greater than 2/3 (less than 1/3), implying that according to one model the observed value should be higher than that of the other model 2 times out of 3, indicates a significant discrepancy.

5. Application

In the following section, the PPP approach delineated in Section 4 is employed to compare pairs of models. Both simulated models and the models resulting from MPS19 and ESHM20 for Udine, Napoli, and Reggio di Calabria are considered. An exhaustive description of the models and their derivation can be found in [11] for MPS19 and [12] for ESHM20. It is important to note that the model output was a table of APEs for a discrete set of PGAs. In order to apply the proposed method, it is necessary to fit a curve and take into account that the available data points cover a subset of the range of PGAs. The procedure that was followed is outlined in Appendix A.

It should be noted that weights are not applied in this context. Rather, the assumption is made that all sub-models $\bar{F}(y|H_m^{(i)})$, $m = 1, \dots, M_i$ have the same weight $\pi_m^{(i)} = P(H_m^{(i)}) = 1/M_i$.

The methodology can easily include weights (and the analytical presentation in Section 4 does include them as the probability distribution $G^{(i)}$), and we acknowledge that in order to make a relevant discussion of the seismic models some weights ought to be considered. The choice of the weights, however, is not trivial as there is not a unique established methodology to define them [25,26] (a meaningful comparison of the two models should consider different sets of weight to assess robustness of the conclusions). Since the purpose of the application in the present paper is to offer an illustrative example of the implementation of the PPP methodology

Table 1
Various measures of discrepancy between MPS19 and ESHM20 for six different locations, $q_0 = 1/475$, $q_1 = 1/2475$.

	Udine	Napoli	Reggio di Calabria
$\bar{Y}^{(1)}_{q_0}$	0.272	0.109	0.235
$\overline{Q}^{(2)}_{q_0}(0.25)$	0.170	0.113	0.234
$\overline{Q}^{(2)}_{q_0}(0.75)$	0.291	0.158	0.358
$\mathrm{ABR2}_{q_0}$	−0.226	0.266	0.221
Cohen d for $Q^{(i)}_{q_0}()$ comparison	0.569	1.030	0.881
PPP for max in 475 years	0.595	0.369	0.402
Cohen d comparing predictive distributions of maxima, $n = 475$	0.202	−0.448	−0.285
$\bar{Y}^{(1)}_{q_1}$	0.577	0.262	0.528
$\overline{Q}^{(2)}_{q_1}(0.25)$	0.390	0.446	0.455
$\overline{Q}^{(2)}_{q_1}(0.75)$	0.622	0.569	0.820
$\mathrm{ABR2}_{q_0}$	−0.246	0.626	0.089
Cohen d for $Q^{(i)}_{q_1}()$ comparison	0.489	2.840	0.633
PPP for max in 2475 years	0.600	0.238	0.439
Cohen d comparing predictive distributions of maxima, $n = 2475$	0.261	−0.836	−0.231
Cohen d to compare $K^{(i)}_{\mathrm{AFE},Y^{(i)}}(q_0)$	0.341	0.801	0.735
PPP for waiting time	0.575	0.417	0.429
Cohen d comparing predictive distributions of waiting time	−0.341	0.801	0.735
Cohen d to compare $K^{(i)}_{\mathrm{AFE},Y^{(i)}}(q_0)$	0.216	2.220	0.625
PPP for waiting time	0.601	0.290	0.442
Cohen d comparing predictive distributions of waiting time	−0.216	2.220	0.625

rather than comparing the two seismic models on substantive grounds, we feel that any choice would fit our purpose and, as such, we took the simplest solution.

The comparison is executed through the utilization of summary statistics, encompassing (a) the maximums of samples of varying dimensions, $T(y_1, \dots, y_n) = \max\{y_1, \dots, y_n\}$ ($n = 100, 476, 1000, 2476, 5000, 10\,000$) and (b) the waiting time for the first exceedance of 16 distinct thresholds ranging from 0.1 to 1.7, $T(y_1, \dots, y_n) = \min\{i : y_i > y_0\}$. Predictive p-values are computed for these statistics. Additionally, the Cohen d is employed to compare the predictive distributions of the statistics as well as the Cohen d to compare the conditional distributions of the expected PGA (as illustrated in the left panel of Fig. 3) and the conditional distributions of the probability of exceedance (as illustrated in the right panel of Fig. 3). Note that the latter two do not consider aleatory variability.

Fig. 6 depicts two simulated models (finite sets of exponential distribution functions) and the results of the comparisons based on the PPP approach as well as the more traditional Cohen d.

The subplot on the right of the main plot (second in lexicographic order) depicts the PPP for the maximum of a set of $n = 1/\mathrm{APE}$ replications (the orange dots correspond to $n = 475$ and $n = 2475$); the third subplot represents the Cohen d comparing the distributions of the sampled maxima (black) and the Cohen d comparing the distributions of the expected PGA (dark green). The subplots situated beneath the main plot (4th and 5th subplots) illustrate the PPP for the waiting time for the first exceedance of a threshold (x-axis), the Cohen d comparing the distributions of the waiting time (black) and the Cohen d comparing the distributions of the APEs for that threshold (dark green). Note that the Cohen d comparing the distributions of expected PGA and the distributions of the APE (dark green), which correspond to the relatively standard approaches discussed in Section 3, ignore aleatory variability, thus are expected to show higher values than the Cohen d comparing the distributions of waiting time and sample maxima.

By performing the comparison for multiple sample sizes (for the maxima) and thresholds (for the waiting time) a discrepancy profile is developed. This profile is more useful than a comparison specific to a single PGA threshold or APE since it enables to detect which parts of the model differ. For example the two models for Naples (Fig. 8) show a relevant difference for high thresholds/sample sizes while they are almost identical for lower thresholds/sample sizes. Conversely, the comparison for Udine (Fig. 7) suggests that the two models are very similar on the full range considered.

A comparison of the results from the proposed method with the existing comparison criteria described in the previous paragraph (reported in Table 1) reveals substantial agreement for Udine (Fig. 7), No indicator shows a relevant discrepancy, and the PPPs concur. For the Naples data set (see Fig. 8), there is a high degree of consensus among all results. The PPP-based discrepancy profile indicates that the models are nearly indistinguishable up to a certain point, but diverge significantly in the right tail. For Reggio di Calabria (Fig. 9), the conventional indicators offer equivocal results; indeed, the discrepancy discloses a divergence in the less extreme segment of the distributions (possibly due to different earthquake catalogs used to derive the models), which tends to dissipate as we progress toward the right tail. Consequently, the distinction between the models becomes moot when the objective is to evaluate risks with a low rate of occurrence (and a higher value).

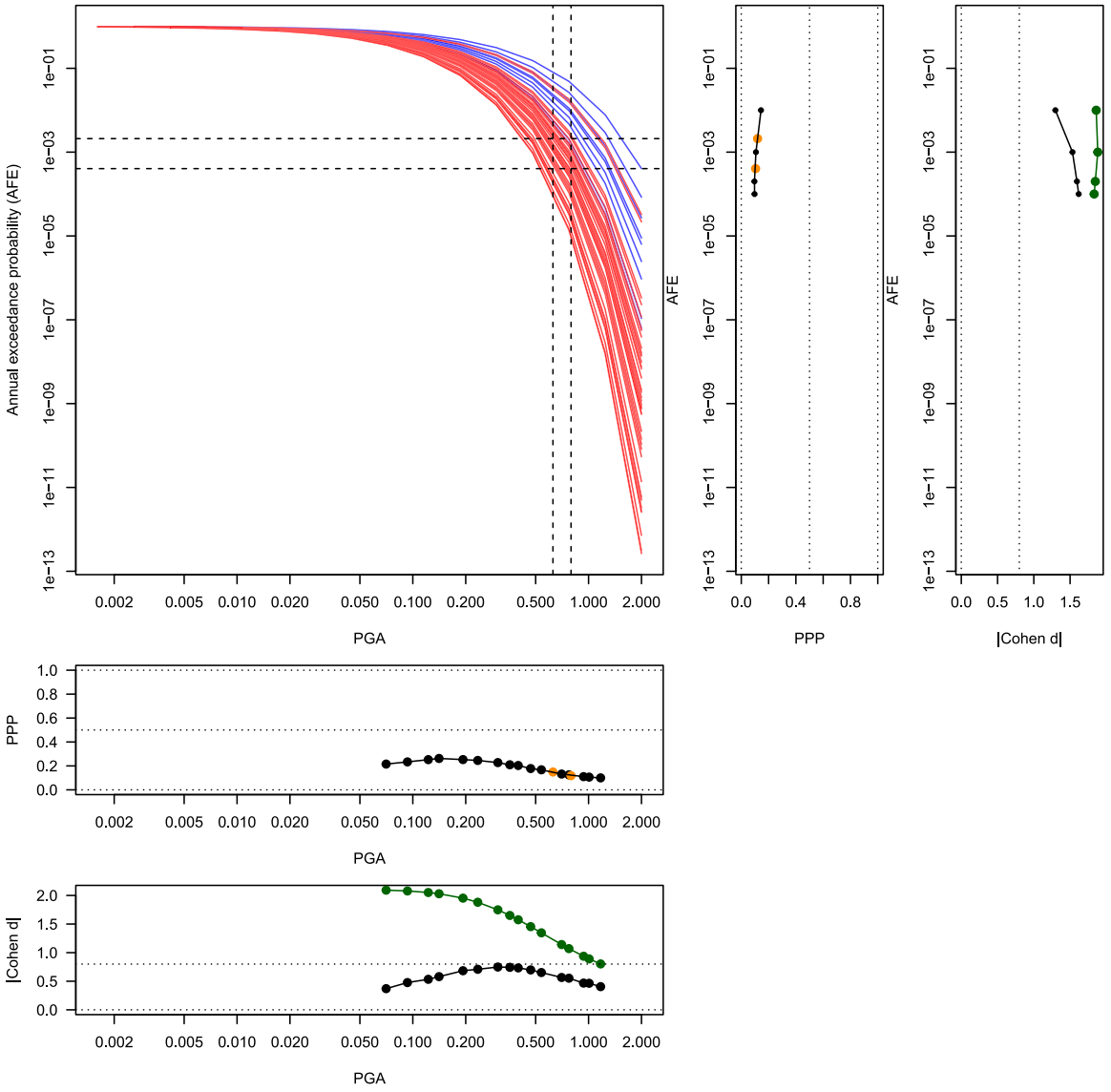


Fig. 6. Comparison of two simulated models. In lexicographic order: depiction of the two models; PPPs for the maximum of a set of $n = 1/\text{AFE}$ replications (the orange dots correspond to $n = 475$ and $n = 2475$); Cohen d comparing the distributions of the sampled maxima (black) and the Cohen d comparing the distributions of the expected PGA (dark green); PPPs for the waiting time for the first exceedance of a threshold (x-axis); Cohen d comparing the distributions of the waiting time (black) and the Cohen d comparing the distributions of the APEs for that threshold (dark green).

6. Epistemic and aleatory variability roles in the proposed approach

In order to emphasize the roles of epistemic and aleatory uncertainty, let us consider the examples in Fig. 10 in which two pairs of models are compared. These two pairs differ in particular for the role of epistemic uncertainty with respect to aleatory uncertainty.

Consider the comparison of the quantile for an APE equal to 0.1 for the two pairs of models. The distributions of the quantile according to the two models of the first pair (Fig. 11, left panel) differ slightly in absolute terms; however, since the epistemic variability is low, the Cohen d for the comparison is 1.034, relatively high, signaling a relevant difference. The distributions of the quantile according to the two models of the second pair (Fig. 12, left panel) exhibit a greater mean difference (in absolute terms). Nevertheless, the epistemic variability is elevated as well, resulting in a Cohen d of 1.067, which is analogous in value to the one obtained for the first pair.

Let us then consider the PPP comparison based on the maximums of samples of size $n = 10$. The predictive distributions for these statistics are shown in the right panels of Figs. 11 and 12 for the two pairs. The corresponding PPP values are 0.4717 for the first

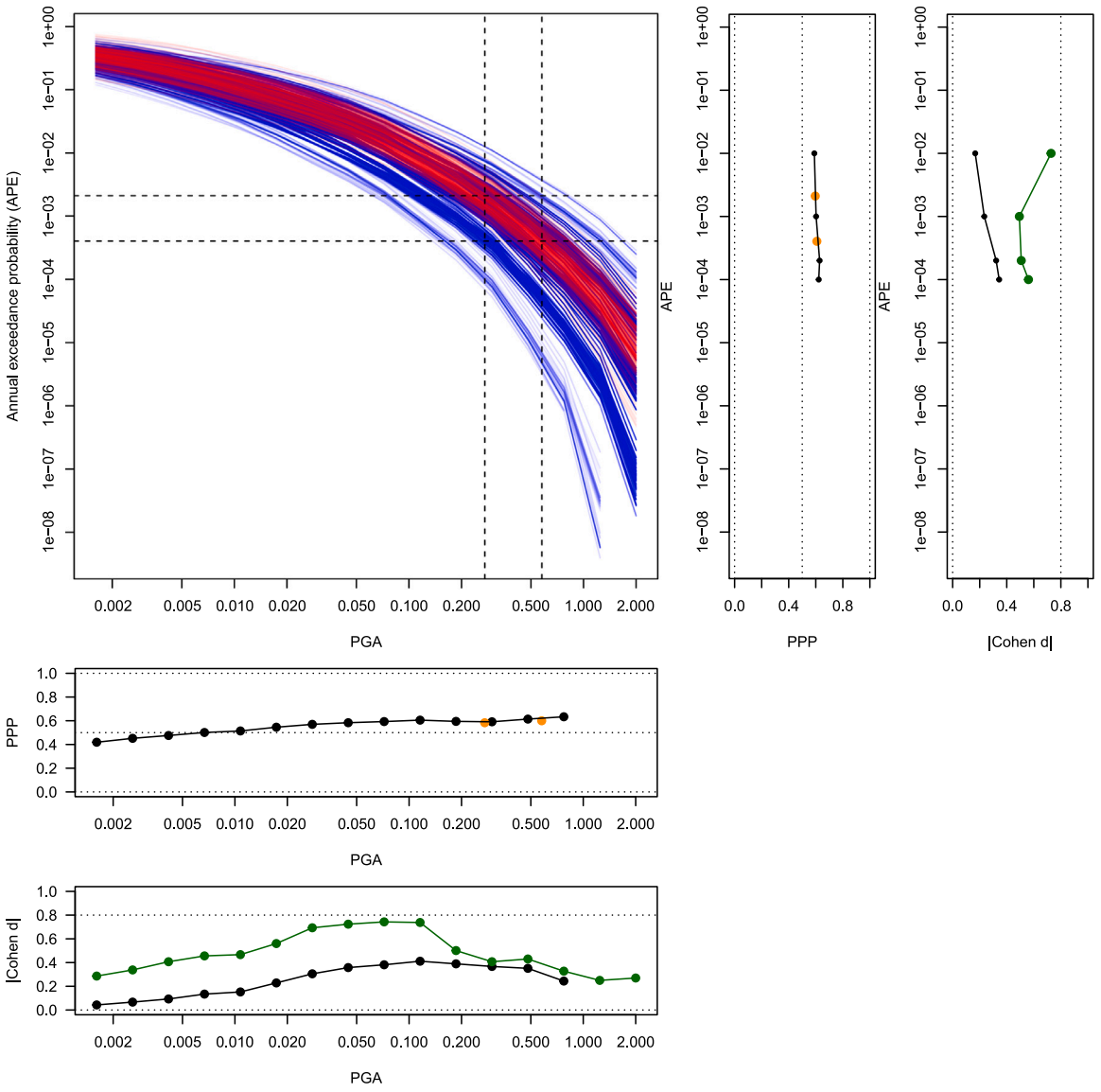


Fig. 7. Comparison of MPS19 (red) and ESHM20 (blue) models for Udine. In lexicographic order: depiction of the two models; PPPs for the maximum of a set of $n = 1/\text{AFE}$ replications (the orange dots correspond to $n = 475$ and $n = 2475$); Cohen d comparing the distributions of the sampled maxima (black) and the Cohen d comparing the distributions of the expected PGA (dark green); PPPs for the waiting time for the first exceedance of a threshold (x-axis); Cohen d comparing the distributions of the waiting time (black) and the Cohen d comparing the distributions of the APEs for that threshold (dark green).

pair and 0.3258 for the second pair. This discrepancy in the PPP values leads to different conclusions: the first pair of models are very similar while the second pair is different (Cohen d for the predictive distributions: the first pair yielded a result of −0.09382, while the second pair produced a result of −0.5801).

The distinguishing factor between the two cases under consideration is the relative importance of epistemic and aleatory uncertainties. In the first case, aleatory uncertainty plays a more significant role. Consequently, the discrepancy between the quantile distributions, which disregards aleatory uncertainty, indicates a notable difference. However, this discrepancy ceases to exist when aleatory uncertainty is taken into account. In the second case, the incorporation of aleatory uncertainty, which is less important relative to the epistemic one, does not alter the conclusion.

The same comparison made for a sample size $n = 10\,000$ is presented in Figs. 13 and 14. The Cohen d for the first pair is 1.034. while for the second pair, it is 1.067. The PPP values are 0.425 and 0.2305, respectively. The Cohen d for the comparison of the predictive distributions is −0.2564 and −0.9606.

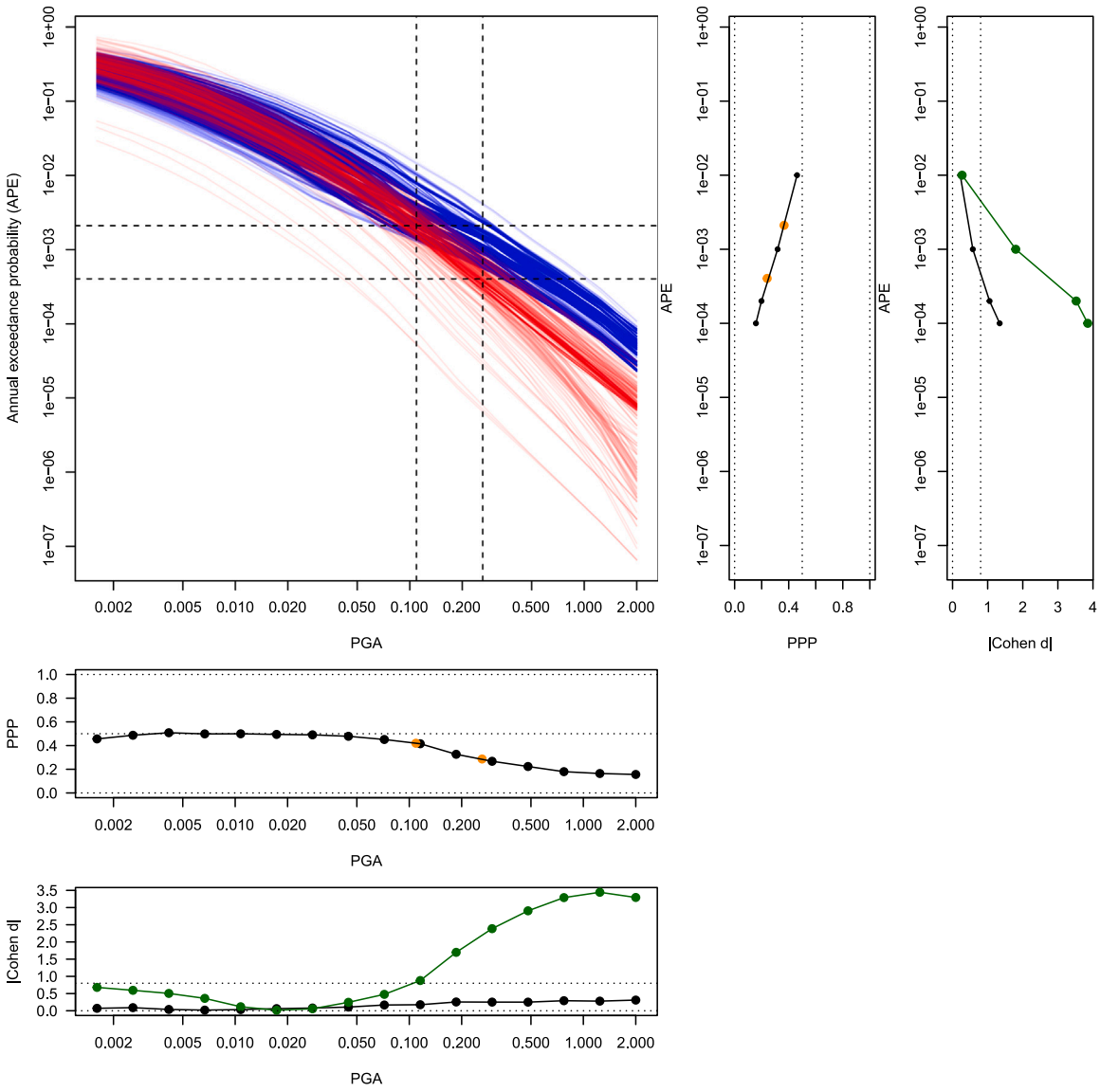


Fig. 8. Comparison of MPS19 (red) and ESHM20 (blue) models for Napoli. In lexicographic order: depiction of the two models; PPPs for the maximum of a set of $n = 1/\text{AFE}$ replications (the orange dots correspond to $n = 475$ and $n = 2475$); Cohen d comparing the distributions of the sampled maxima (black) and the Cohen d comparing the distributions of the expected PGA (dark green); PPPs for the waiting time for the first exceedance of a threshold (x-axis); Cohen d comparing the distributions of the waiting time (black) and the Cohen d comparing the distributions of the APEs for that threshold (dark green).

An alternative perspective on the role of aleatory variability can be obtained by examining a statistic such as *(a)* the number of exceedances of a specified threshold in a sample of size n, or *(b)* the mean of a sample of size n, for different sample sizes n. These statistics, in fact, are such that the aleatory variability reduces as the sample size increases. In essence, the impact of aleatory variability can be mitigated by increasing the sample size under consideration. In the limit, as the sample size becomes exceedingly large, aleatory uncertainty can be effectively disregarded.

Figs. 15 and 16 illustrate the predictive distributions for the number of exceedances and the mean, respectively, for samples of different sizes utilizing the first pair of models (where the epistemic variability is minimal), it is evident that the discrepancy between the predictive distributions is negligible for small samples. This prompts the reiteration of the aforementioned observations concerning the maximums, namely, that the distinction in the epistemic distributions becomes less consequential if we permit aleatory variability in the maximum. However, if a more substantial sample size is considered, the discrepancy becomes non-negligible.

When considering the second pair of models (where epistemic variability is larger, but the mean difference is also larger), the picture is different and, even for small sample sizes, the predictive distributions differ significantly (Figs. 17 and 18).

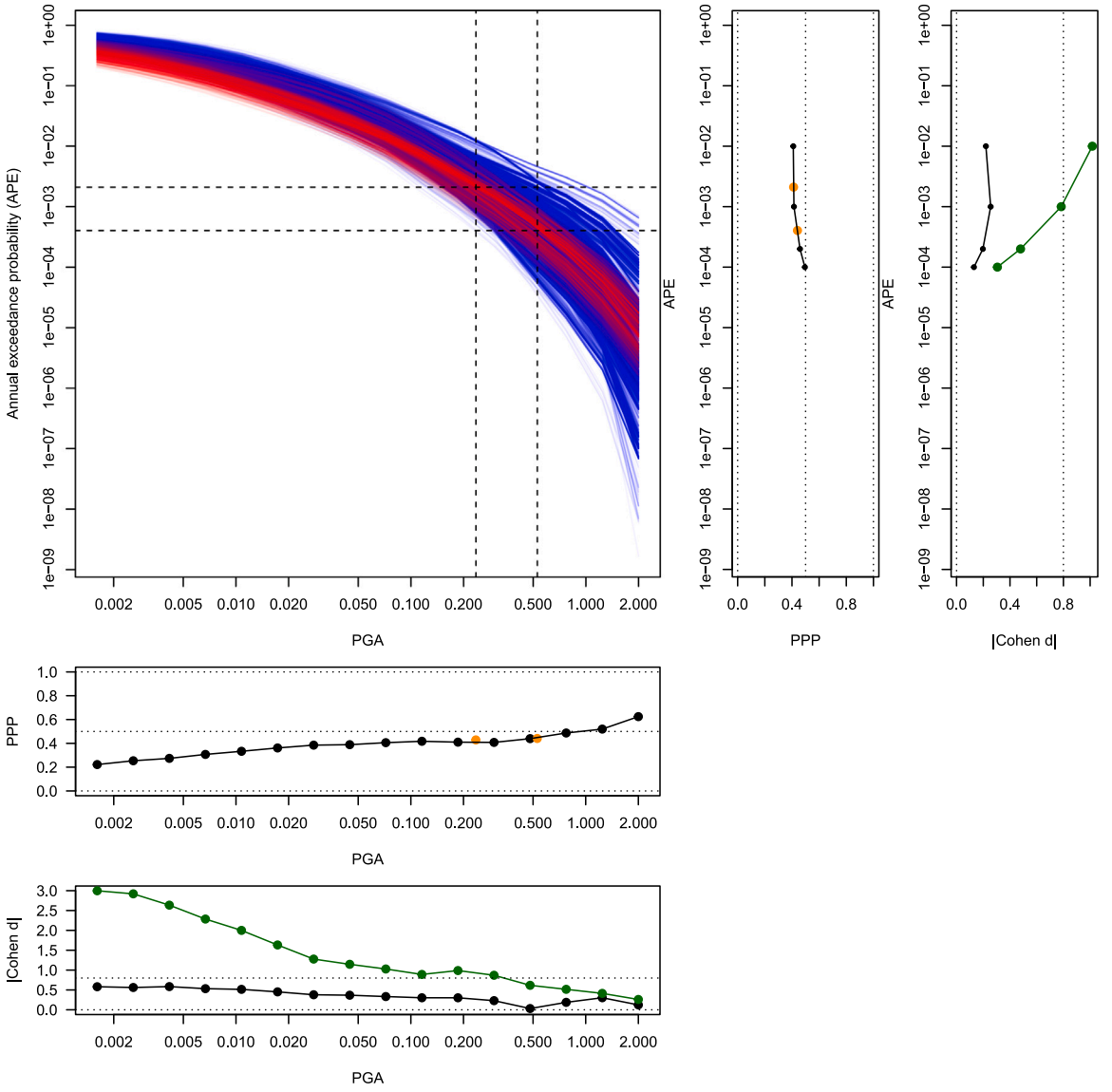


Fig. 9. Comparison of MPS19 (red) and ESHM20 (blue) models for Reggio di Calabria. In lexicographic order: depiction of the two models; PPPs for the maximum of a set of $n = 1/\text{AFE}$ replications (the orange dots correspond to $n = 475$ and $n = 2475$); Cohen d comparing the distributions of the sampled maxima (black) and the Cohen d comparing the distributions of the expected PGA (dark green); PPPs for the waiting time for the first exceedance of a threshold (x-axis); Cohen d comparing the distributions of the waiting time (black) and the Cohen d comparing the distributions of the APEs for that threshold (dark green).

This underscores the significance of selecting a statistic that can be regarded as substantially relevant also as far as the implied aleatory variability is concerned. For the purpose of illustration, let us consider the number of exceedances of a high threshold (e.g. a threshold that is exceeded with a probability less than 0.001 according to all sub-models involved) in a sample of size n. If a low sample size is employed, e.g. $n = 50$, most simulated values would be zero in both models, thereby suggesting no difference between them. Conversely, if a very high n, such as $n = 10^6$, is employed, the conditional (to each sub-model) variance of the statistic becomes negligible. Consequently, the comparison will consider only the epistemic variability.

It is important to note that the majority of the methods proposed in the existent literature for the comparison of seismological models, as reviewed in Section 4, ignore aleatory uncertainty by comparing the distributions representing the epistemic uncertainty on specific aspects of the hazard functions. These distributions may take the form of a probability of exceedance of a threshold or the quantile of the PGA for a given APE.

To the best of our knowledge, the only proposals that consider aleatory uncertainty are in [13,19]. The former approach utilizes Cohen's d to assess the distributions of the maximum PGA over a 50-year period, as previously referenced. In the latter, the authors

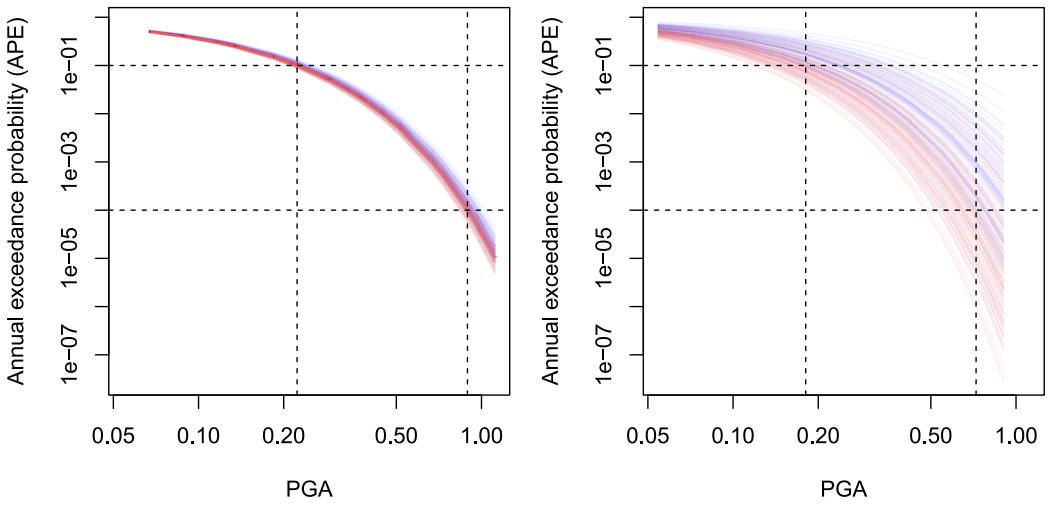


Fig. 10. Two Pairs of models: each plot comprises two sets of hazard curves (survival functions), constituting two models to be compared , the dashed horizontal lines correspond to APEs 10^{-1} and 10^{-4}, the vertical lines to the associated mean ground motions.

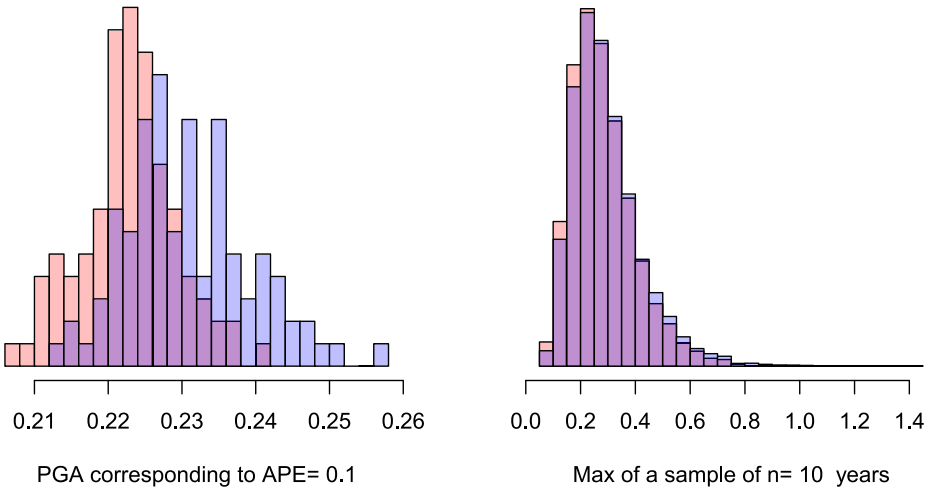


Fig. 11. Left: Distributions of PGA for a given APE; Right: Distribution of maxima of samples of size n. The colors correspond to the two models of Fig. 10-left.

employ Cohen d (improperly designated as a test statistic) to compare the "probability density functions (PDFs) derived from the mean hazard curves of the two studies being compared" (see Fig. 5), in our notation $F^{(1)}(y)$ and $F^{(2)}(y)$. It is important to note that the result of this analysis is contingent upon the full distribution of the PGA. The relevance of this distribution is not particularly pertinent to engineering applications; rather, the salient aspect is the right tail (i.e., the extreme quantiles) of the hazard distribution.

In the application, the relevance of the aleatory variability is evident in the evaluation for Udine. For thresholds ranging from 0.03 to 0.1 the Cohen d to compare the distributions of the corresponding APEs approaches 0.8. Conversely, the PPPs obtained for the same threshold using the waiting time for the first exceedance are approximately 0.5. This finding underscores the importance of incorporating aleatory variability into the PPP approach, thereby facilitating the attainment of a more authentic depiction of the discrepancy between the two models.

7. Conclusion

We propose a methodology for the comparison hazard and risk models. This methodology is based on a variant of the prior/posterior predictive checks, which are routinely employed in Bayesian statistics for model validation. In particular, we propose the utilization of a variant of predictive p-values as indices of discrepancy between the two models. While the proposed approach is generally valid, its use is illustrated in the context of seismic hazard. In that context, Marzocchi and Jordan [8,20] and Marzochi

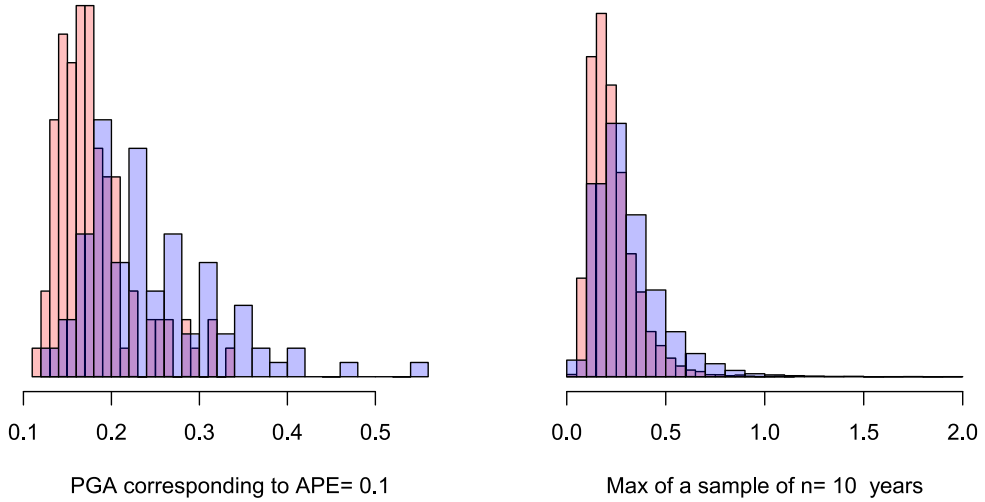


Fig. 12. Left: Distributions of PGA for a given APE; Right: Distribution of maxima of samples of size n. The colors correspond to the two models of Fig. 10-right.

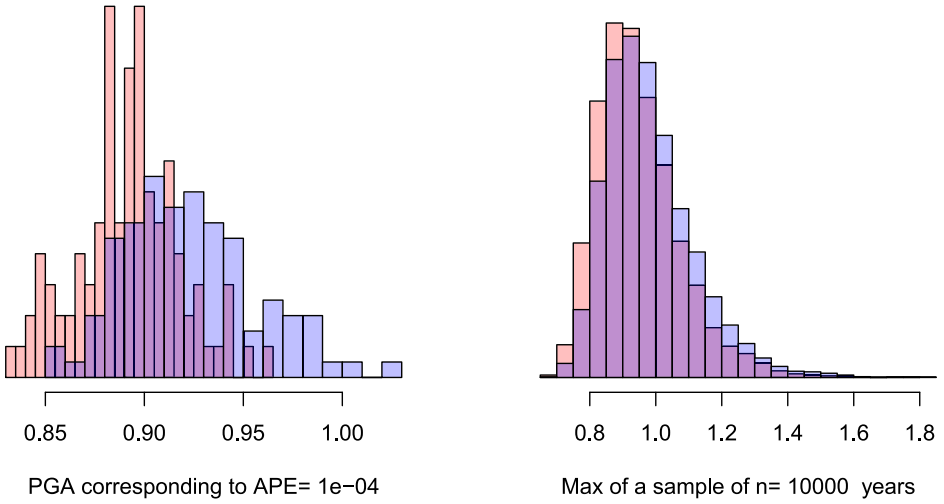


Fig. 13. Left: Distributions of PGA for a given APE; Right: Distribution of maxima of samples of size n. The colors correspond to the two models of Fig. 10-left.

et al. [9] advocate using predictive checks but only for model validation purposes. In contrast, the present study proposes an innovative extension of their application, namely the comparison of disparate hazard models.

This strategy has been demonstrated to exhibit several advantages over those proposed in the existent literature.

First, the PPP method accounts for both aleatory and epistemic uncertainty, whereas the majority of existent methods consider only epistemic uncertainty, potentially leading to overstating the difference between the models whenever aleatory uncertainty is large relative to epistemic uncertainty.

Second, in contrast to the other proposed testing methods, PPPs are predicated on a probabilistic sampling framework. This characteristic enables a less subjective interpretation of the outcome in many respects. The fact that the PPP quantifies the frequency with which a specific model would suggest a higher value of a quantity of interest relative to another model renders it a more intuitive metric in comparison to, for instance, Cohen's d. As a probability metric, a PPP falls within the interval from 0 to 1, thereby providing a natural reference point for assessing the relevance of an observed discrepancy. The absence of a definitive cutoff does not constitute a limitation; rather, it acknowledges that the response to the inquiry cannot be unequivocally binary. The transformation of a response into a dichotomous form, as frequently occurs in the context of significance tests, represents an unwarranted (and potentially risky) conclusion [27].

Third, the capacity to select a summary statistic as the basis for comparison enables the customization of the comparison to specific aspects of interest, potentially including those related to damage. This customization can be achieved by incorporating

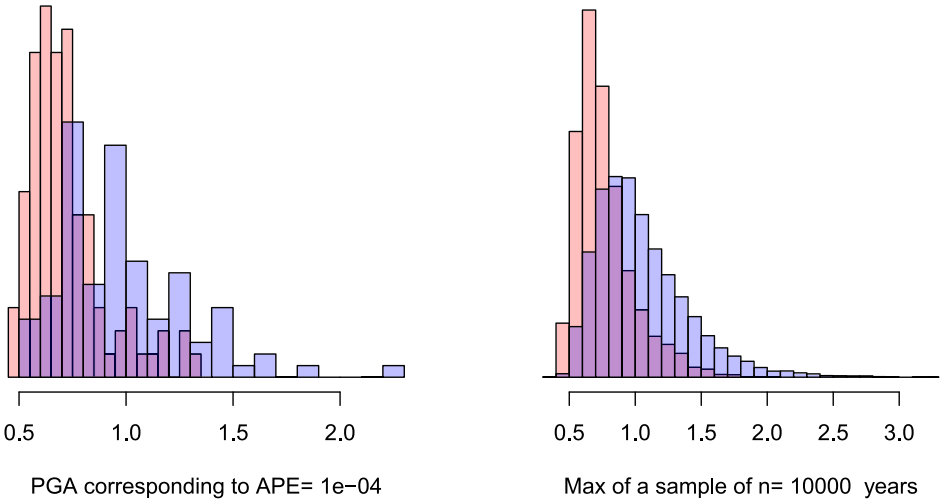


Fig. 14. Left: Distributions of PGA for a given APE; Right: Distribution of maxima of samples of size n. The colors correspond to the two models of Fig. 10-right.

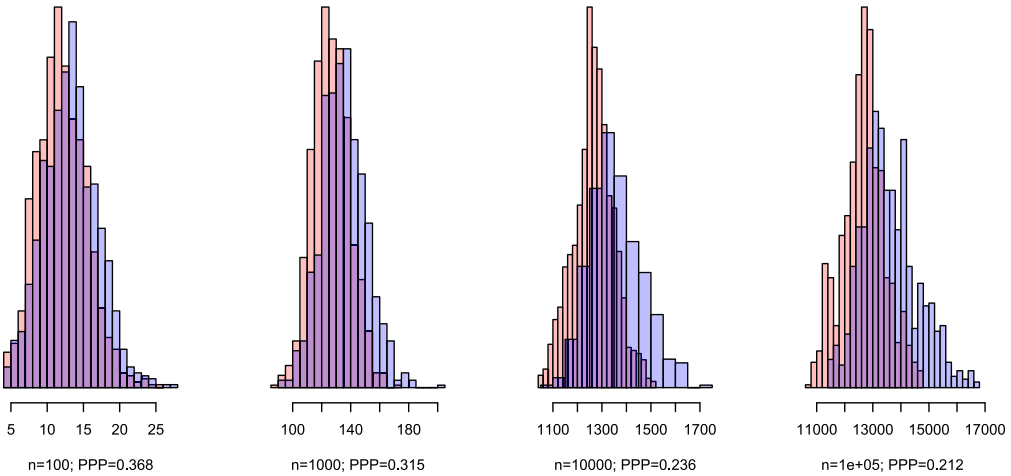


Fig. 15. Distributions of the number of exceedances of a given threshold for different sample sizes.

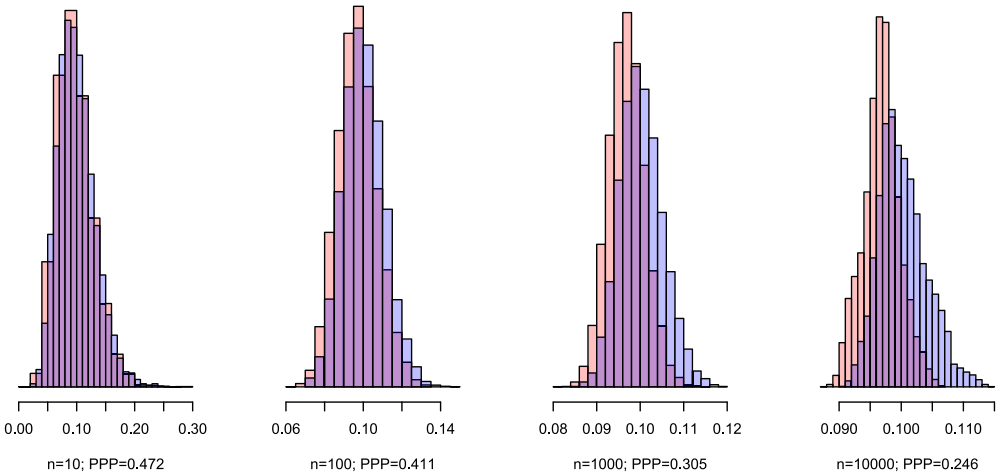


Fig. 16. Distributions of the mean for different sample sizes.

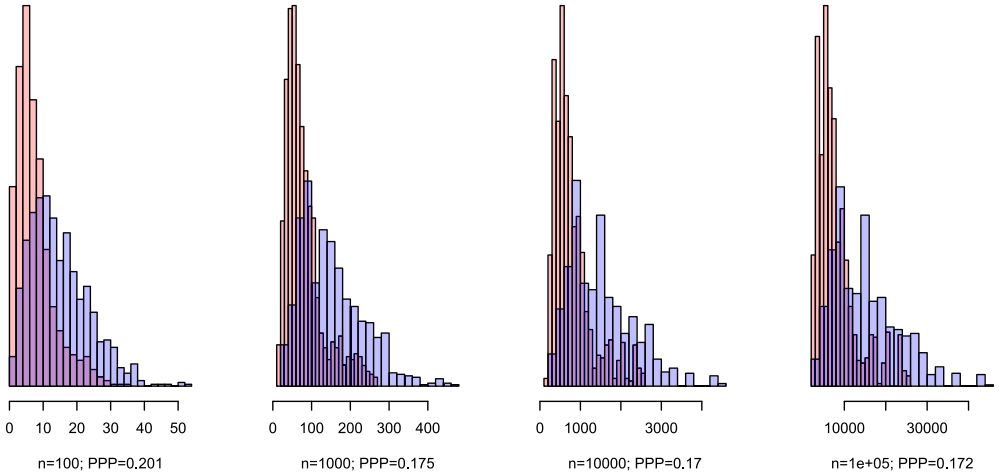


Fig. 17. Distributions of the number of exceedances of a given threshold for different sample sizes.

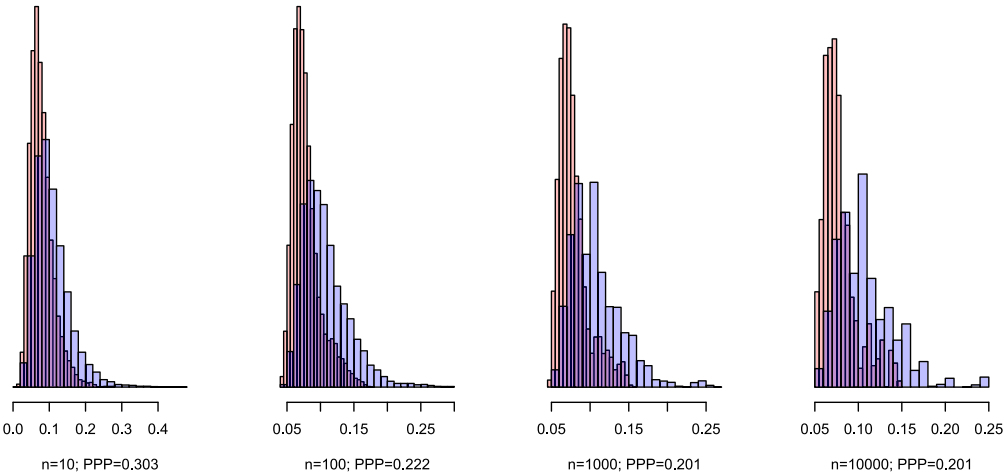


Fig. 18. Distributions of the mean for different sample sizes.

additional exposure and vulnerability layers into the model. It is conceivable that establishing a correlation with the process of decision-making, particularly in the case of multi-risk and multi-layer analysis [28] would prove to be a more efficient approach.

CRediT authorship contribution statement

Francesco Pauli: Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Stefano Parolai:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank Laurentiy Danciu and Francesco Visini for providing the output of the two models.

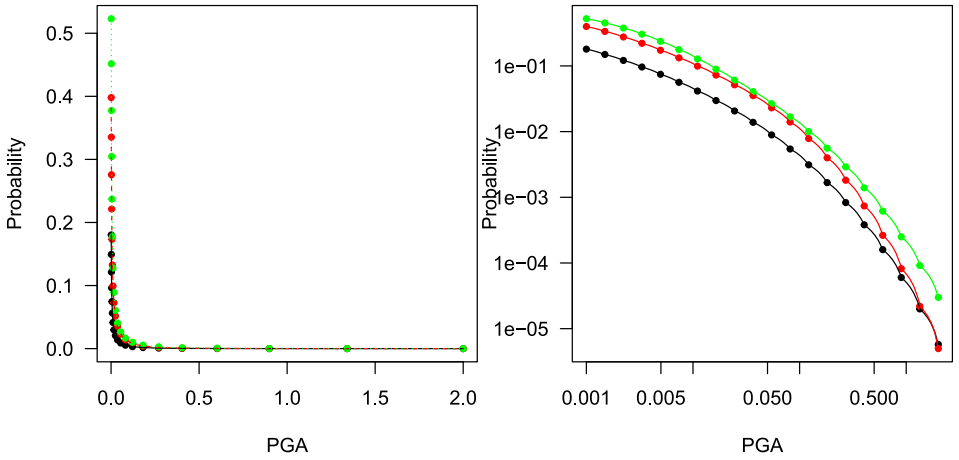


Fig. A.19. The dots represent three rows from the MPS19 for the city of Udine, linear interpolation is used to extract continuous curves.

Table A.2
Probability ranges of the available data.

	p_{mps}	q_{mps}	p_{eshm}	q_{eshm}	$1/p_{mps}$	$1/q_{mps}$	$1/p_{eshm}$	$1/q_{eshm}$
Udine	0.1803	0.0000599582	0.1574	0.0002344342	6	16 679	7	4 266
Napoli	0.0422	0.0000323991	0.1492	0.0000941778	24	30 866	7	10 619
Reggio di Calabria	0.2525	0.0000419763	0.4031	0.0006154151	4	23 823	3	1 625

Appendix A. Computation details

The output from a seismic hazard model is generally presented in the form of a table of PGAs corresponding to given probabilities, in the form depicted below.

```
        0.001 0.0014919 0.0022258  0.0033206   0.004954  0.0073908  0.0110263  0.0164501 ...
1   0.1802592 0.1493586 0.1212775 0.09627394 0.07458255 0.05633205 0.04145650 0.02969021 ...
250 0.3980317 0.3353043 0.2757313 0.22127260 0.17332360 0.13264360 0.09926863 0.07258032 ...
500 0.5230075 0.4516363 0.3773948 0.30465290 0.23699560 0.17732740 0.12776150 0.08914191 ...
```

where the first row contains the PGA values and the subsequent rows represent the probability of exceeding that PGA according to a branch of the logic tree.

The MPS19 model's available output consists of 576 branches and 20 PGA values ranging from 0.001 to 2. The ESHM model comprises 10 000 branches and 25 PGA values ranging from 0.0005 to 3.

As illustrated in Fig. A.19, the dots represent three rows from the MPS19 for the city of Udine. In order to perform the analysis outlined in the paper, it is necessary to calculate the survival functions. To that end, the points are interpolated linearly, as illustrated in Fig. A.19. Alternatively, the implementation of monotone smoothing functions could be considered, under the condition that monotone smoothing techniques are employed. This alternative was explored, and it was determined that there were negligible differences in the results.

It is important to note that the available data points do not encompass the entire 0–1 probability range. Consequently, they also do not cover the full range of possible PGA values. Fig. A.20 shows the available data from the two models (MPS in red, ESHM in blue) for a location. As already noted, ESHM data are tabulated for PGA ranging from 0.0005 to 3, while MPS data are tabulated for PGA ranging from 0.001 to 2. Any probability calculation outside this range requires extrapolation. It is reasonable, therefore, to use PGA thresholds above 0.001 and below 2 to calculate the PPPs, in order to avoid relying on extrapolated values. It should be noted, however, that extrapolation for higher thresholds would be a minor problem, since the probabilities involved are very low. A similar line of reasoning indicates that simulations should be avoided that rely on extrapolation with respect to the probability range. That is to say, probability above $\min\{p_{mps}, p_{eshm}\}$ or below $\max\{q_{mps}, q_{eshm}\}$, that is return periods below $\max\{1/p_{mps}, 1/p_{eshm}\}$ or above $\min\{1/q_{mps}, 1/q_{eshm}\}$. In Table A.2, we report these limits for each location and model.

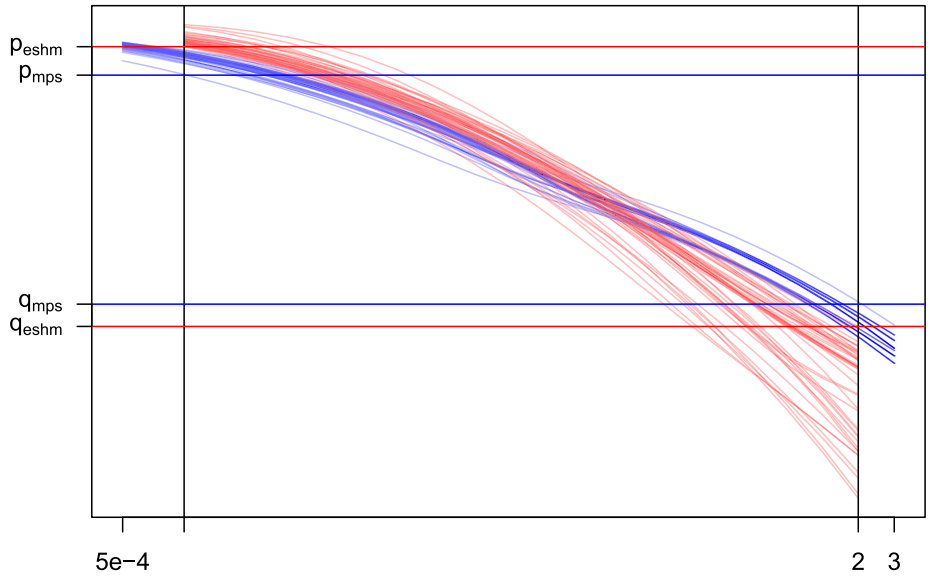


Fig. A.20. Available data from the two models (MPS in red, ESHM in blue) for a location.

Appendix B. Proposed method, algorithm description

A model comparison based on the statistic $T()$ is carried out based on of the following algorithm

1. Simulation of a sample from model 1, by first selecting a function $F^{(1)}_{\lambda^*}$ and then randomly taking a sample from it: $y^{*(1)}_1, \dots, y^{*(1)}_n \sim F^{(1)}_{\lambda^*}(.)$
2. Evaluate the statistic $T^* = T(y^{*(1)}_1, \dots, y^{*(1)}_n)$
3. Compute the probability of observing a sample corresponding to a value of the statistic higher than T^* if model 2 is correct, by (for a discrete set of models)

$$p^* = \sum_{m=1}^{M_2} P(T > T^* | H^{(2)}_m) P(H^{(2)}_m) \tag{B.1}$$

   Note that $P(T > T^* | H^{(2)}_m)$, is a (unilateral) p-value for (sub)model $H^{(2)}_m$, and, using the framework of Marzocchi and Jordan [20] is a p-value conditional on model $H^{(2)}_m$. Therefore, it is a measure of evidence against model $H^{(2)}_m$ based on the sample. The Eq. (B.1) mimics what [20] defines as an unconditional p-value: a measure of the evidence against model 2, allowing for epistemological uncertainty.
4. Repeating steps 1–3 B times, obtaining B p-values $p^*_1, \dots, p^*_B$, and average them to obtain an overall comparison of model 0 to model 1 with reference to the statistic $T()$: p.

If the two models are identical, then the expected value of p is 0.5; values near 0 or 1 indicate differences between the models.

Summarizing, we are repeatedly simulating a sample from model 1, treating it as if they were empirical observation, and comparing model 2 against that sample using the framework of [20]. Finally, we take the average of the unconditional p-value as an overall measure of the difference between model 1 and model 2.

Note that we could substitute steps 3–4 with

3 Simulating a sample from model 1, $y^{*(2)}_1, \dots, y^{*(2)}_n \sim F^{(2)}_{m^*}(.)$
4 Computing the satistic $T^{*(2)} = T(y^{*(2)}_1, \dots, y^{*(2)}_n)$
5 Comparing T^* and $T^{*(2)}$, let $I^* = |T^* > T^{*(2)}| + 0.5|T^* = T^{*(2)}|$
6 Repeat steps 1–5 B times, obtaining B values $I^*_1, \dots, I^*_B$ and average these to obtain the overall comparison p.

This second form may be more convenient if the probability in Eq. (B.1) is hard to calculate in closed form; otherwise, the former may be slightly more computationally efficient.

Data availability

The authors do not have permission to share data.

References

[1] C.A. Cornell, et al., Engineering seismic risk analysis, Bull. Seismol. Soc. Am. 58 (5) (1968) 1583–1606.
[2] K. Durham, Treating the risks in cairns, Nat. Hazards 30 (2) (2003) 251–261.
[3] G. Grünthal, A.H. Thieken, J. Schwarz, K.S. Radtke, A. Smolka, B. Merz, Comparative risk assessments for the city of Cologne–storms, floods, earthquakes, Nat. Hazards 38 (1) (2006) 21–44.
[4] H. Kreibich, P. Bubeck, M. Kunz, H. Mahlke, S. Parolai, B. Khazai, J. Daniell, T. Lakes, K. Schröter, A review of multiple natural hazards and risks in Germany, Nat. Hazards 74 (3) (2014) 2279–2304.
[5] K. Fleming, S. Parolai, A. Garcia-Aristizabal, S. Tyagunov, S. Vorogushyn, H. Kreibich, H. Mahlke, et al., Harmonizing and comparing single-type natural hazard risk estimations, Ann. Geophys. 59 (2) (2016).
[6] D. Albarello, V. D'Amico, Testing probabilistic seismic hazard estimates by comparison with observations: An example in Italy, Geophys. J. Int. 175 (3) (2008) 1088–1094.
[7] M. Giorgio, I. Iervolino, On multisite probabilistic seismic hazard analysis, Bull. Seismol. Soc. Am. 106 (3) (2016) 1223–1234.
[8] W. Marzocchi, T.H. Jordan, A unified probabilistic framework for seismic hazard analysis, Bull. Seismol. Soc. Am. 107 (6) (2017) 2738–2744.
[9] W. Marzocchi, C. Meletti, V. D'Amico, G. Lanzano, L. Luzi, F. Martinelli, B. Pace, A. Rovida, M. Taroni, F. Visini, Guidelines for the scientific evaluation of an NSHM: The legacy of the MPS19 Italian model, Bull. Seismol. Soc. Am. (2025).
[10] J. Douglas, H. Crowley, V. Silva, W. Marzocchi, L. Danciu, R. Pinho, Methods for evaluating the significance and importance of differences amongst probabilistic seismic hazard results for engineering and risk analyses: a review and insights, Bull. Earthq. Eng. 22 (6) (2024) 2769–2796.
[11] C. Meletti, W. Marzocchi, V. D'amico, G. Lanzano, L. Luzi, F. Martinelli, B. Pace, A. Rovida, M. Taroni, F. Visini, et al., The new Italian seismic hazard model (MPS19), Ann. Geophys. 64 (1) (2021).
[12] L. Danciu, S. Nandan, C. Reyes, R. Basili, G. Weatherill, C. Beauval, A. Rovida, S. Vilanova, K. Sesetyan, P.-Y. Bard, F. Cotton, S. Wiemer, D. Giardini, The 2020 update of the European Seismic Hazard Model: Model Overview, Technical Report 001, EFEHR, 2021, URL https://doi.org/10.12686/a15.
[13] P.K. Malhotra, Myth of probabilistic seismic hazard analysis, Structure (2015) 58–59.
[14] B. Lund, P. Mäntyniemi, A. Sadeghi-Bagherabadi, A. Korja, J. Lundwall, Comparing European sEismic Hazard Models: ESHM 20 Versus ESHM13 at Nuclear Power Plant Sites in Sweden And Finland, Tech. rep., Energiforsk AB, 2024.
[15] N. Abrahamson, Treatment of epistemic uncertainty in PSHA results, PSHA workshop, Lenzburg, Switzerland, 2017.
[16] P.K. Malhotra, Cost of Uncertainty in Seismic Hazard, Tech. rep, StrongMotions Inc., Sharon, 2014.
[17] J. Cohen, Statistical Power Analysis for the Behavioral Sciences, Routledge, 1988.
[18] R.K. McGuire, Precision of seismic hazard evaluations in central and eastern North America, in: Proceedings of the 15th World Conference on Earthquake Engineering, 2012, p. 5501.
[19] J. Douglas, T. Ulrich, D. Bertil, J. Rey, Comparison of the ranges of uncertainty captured in different seismic-hazard studies, Seismol. Res. Lett. 85 (5) (2014) 977–985.
[20] W. Marzocchi, T.H. Jordan, Testing for ontological errors in probabilistic forecasting models of natural systems, Proc. Natl. Acad. Sci. 111 (33) (2014) 11973–11978.
[21] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, third ed., Chapman and Hall/CRC, 2013.
[22] A. Gelman, Two simple examples for understanding posterior p-values whose distributions are far from uniform, Electron. J. Stat. 7 (none) (2013) 2595–2602, URL https://doi.org/10.1214/13-EJS854.
[23] X.-L. Meng, Posterior predictive p-values, Ann. Statist. 22 (3) (1994) 1142–1160.
[24] L.G. Fabreti, L.M. Coghill, R.C. Thomson, S. Höhna, J.M. Brown, The expected behaviors of posterior predictive tests and their unexpected interpretation, Mol. Biol. Evol. 41 (3) (2024) msae051.
[25] L. Zanetti, D. Chiffi, L. Petrini, Philosophical aspects of probabilistic seismic hazard analysis (PSHA): a critical review, Nat. Hazards 117 (2) (2023) 1193–1212.
[26] V. D'Amico, F. Visini, A. Rovida, W. Marzocchi, C. Meletti, Scoring and ranking probabilistic seismic hazard models: an application based on macroseismic intensity data, Nat. Hazards Earth Syst. Sci. 24 (4) (2024) 1401–1413.
[27] A. Gelman, H. Stern, The difference between "significant" and "not significant" is not itself statistically significant, Amer. Statist. 60 (4) (2006) 328–331.
[28] J. Zschau, Where are we with multihazards, multirisks assessment capacities? in: K. Poljanšek, M. Marín Ferrer, T. De Groeve, I. Clark (Eds.), Science for Disaster Risk Management 2017: Knowing Better and Losing Less., Publications Office of the European Union, Luxembourg, 2017, pp. 98–130.