# Scoring and ranking probabilistic seismic hazard models: an application based on macroseismic intensity data

Vera D'Amico[1], Francesco Visini[1], Andrea Rovida[2], Warner Marzocchi[3], Carlo Meletti[1]

[1]Istituto Nazionale di Geofisica e Vulcanologia, Pisa, 56125, Italy
[2]Istituto Nazionale di Geofisica e Vulcanologia, Milan, 20133, Italy
[3]Department of Earth, Environmental, and Resources Sciences, University of Naples, Federico II, Naples, 80126, Italy

*Correspondence to*: Vera D'Amico (vera.damico@ingv.it)

**Abstract.** A probabilistic seismic hazard model consists of a set of weighted models/branches that describes the center, the body, and the range of seismic hazard. Owing to the intrinsic nature of this kind of analysis, the weight of each model/branch represents its scientific credibility. However, practical uses of this model may sometimes require the selection of one or a few hazard curves that are sampled from the whole model, that often consists of thousands of branches. Here we put forward an innovative procedure that facilitates the scoring, ranking and selection of those hazard curves to account for the requirements of a specific application. The approach consists of a careful quality check of the data used for scoring and the adoption of a proper scoring rule. To show the applicability of this approach, we present an example that consists of scoring and ranking a set of multiple models/branches constituting a recent seismic hazard model of Italy. To score these branches, hazard estimates produced by each of them are compared with time-series of macroseismic observations available in the Italian macroseismic database for a carefully selected set of localities deemed sufficiently representative, homogeneously distributed in space and complete with respect to time and intensity levels. The proper scoring parameter used for such a comparison is the logarithmic score, which can be always applied independently from the distribution of the data.

## 1 Introduction

Probabilistic Seismic Hazard Analysis (PSHA) provides basic information for the proper application of the building code. Owing to the important practical implications, PSHA models have to be widely accepted by a large scientific community. This acceptance is usually achieved by using commonly adopted procedures to calculate PSHA, and the full description of associated uncertainties is one of the key points in reliable models (Gerstenberger et al., 2020).

PSHA is usually built considering different models or branches of a logic tree, which mimics the so-called epistemic uncertainty, i.e., our ignorance of the true seismic hazard value. A critical aspect in describing quantitatively the distribution of the epistemic uncertainty is the way in which the weight of each model or branch is assigned.

Conceptually, the weighting of each model can follow two main general procedures (e.g., Albarello and D'Amico, 2015): the first one is ex-ante, that is by considering inherent properties of each competing PSHA model, i.e., its ability to take into account the current knowledge of the underlying physical process evaluated by panels of experts; the second one is ex-post,

that is by empirically scoring a set of alternative models by comparing the forecasting performance of their outcomes with available seismic observations. The first approach was the most commonly adopted in the past (e.g., Stucchi et al., 2011; Woessner et al., 2015), whereas today, thanks to the large availability of seismic data for comparisons, state-of-the-art PSHA models tend to adopt a combination of the two approaches (e.g., Danciu et al., 2021; Petersen et al., 2024). For example, in

35     the recent PSHA model for Italy called MPS19 (Meletti et al., 2021) the weight of each branch was assigned according to both ways, that is testing the performance of its components, i.e., seismicity and ground-motion attenuation models, against available observations and through the evaluation of the models by panels of experts. Worthy of note, independently from the specific scheme adopted, the weighting of each PSHA model relies on available scientific knowledge.

The use of a PSHA model for practical applications may need additional evaluations. Actually, most practical applications

40     require the choice of one (or a few) hazard curves that are sampled from the model. For instance, many current building codes use arbitrarily the mean hazard, neglecting *de facto* the dispersion of all other hazard curves. Here we propose an innovative post-processing scoring strategy that facilitates the ranking and sampling of models/branches of a PSHA model to consider specific requests from stakeholders, e.g., those responsible for planning seismic risk reduction strategies.

We introduce the procedure through an application to score and rank a set of multiple models/branches that constitute the

45     MPS19 seismic hazard model of Italy according to their fit with macroseismic intensity data available in a large set of selected sites; the aim is selecting the models/branches that minimize the difference between PSHA outcomes and macroseismic observations at these sites. The scoring procedure consists of a careful quality check of the data used for scoring and the adoption of a proper scoring rule.

MPS19 consists of 11 groups of seismicity models (each composed by a set of sub-models, for a total of 94 seismicity

50     models) combined with three Ground Motion Models (GMMs) for the active shallow crustal areas (Bindi et al., 2011; Bindi et al., 2014; Cauzzi et al., 2015), with two GMMs for the subduction zone of the Calabrian Arc (Skarlatoudis et al., 2013; Abrahamson et al., 2016) and one for the volcanic areas (Lanzano and Luzi, 2020), producing a total of 564 branches. The hazard was computed in terms of Peak Ground Acceleration (PGA) and Spectral Acceleration (SA), in the period range 0.05–4 s, for return periods from 30 to 5000 years (for more details on MPS19, see Visini et al., 2021 for seismicity models,

55     Lanzano et al., 2020 for GMMs, and Meletti et al., 2021 for the whole model).

Specifically, the scoring procedure proposed here consists of comparing the hazard of each branch of MPS19 with the time-series of macroseismic observations ("seismic histories") available in the Italian macroseismic database DBMI15 v1.5 (Locati et al., 2016; https://emidius.mi.ingv.it/CPTI15-DBMI15_v1.5/query_place/) for a set of localities deemed sufficiently complete.

60     The proper scoring parameter for such a comparison is the logarithmic score (Gneiting and Raftery, 2007), which can be always applied independently from the specific distribution of the data; when the data follow a Poisson distribution, the logarithmic score is also named Log-Likelihood score (LL):

$$LL = \sum_{i=1}^{N} \log(p_i) \tag{1}$$

where $p_i$ is the probability that each model attributes to the *i-th* observation above the $N$ available. Gneiting and Raftery (2007) show that many other metrics, such as probability, do not have these characteristics and should not be used.

The main phases of the proposed procedure are the following: i) identification of the testing localities where to compare hazard estimates of the individual models with available seismic histories, ii) building of the datasets of observed and expected macroseismic intensities for each locality, iii) comparison between estimates from each branch and observed data in terms of LL of the differences between the number of macroseismic data predicted by the model and the number of those observed for different intensity degrees, and iv) scoring and ranking of the models.

Although our application is focused on a specific PSHA model, we emphasize the generalizability to any other model and kind of observations (e.g., accelerometric data, fragile geological structures), provided they are treated with *ad hoc* procedures.

## 2 Building the datasets of observed and expected intensities

The first step of our procedure is the identification of the set of localities for evaluating the consistency of PSHA models with available observations; then, for each site, two datasets of macroseismic intensities, one of observed data and one of intensities expected according to the hazard estimates, have to be built.

### 2.1 Selection of the testing localities

The selection of the sites where to compare PSHA models' outputs with available observations represents one of the most crucial issues of the scoring procedure and thus needs great attention.

In order to have a representative set of sites to perform tests, the selected localities have to guarantee: i) a geographical coverage as dense and uniform as possible throughout the whole investigated area, in relation to both high and low hazard regions, and ii) seismic histories with a significant number of data, spanning long time periods and covering a wide range of intensity values (see the examples in Fig. 1).

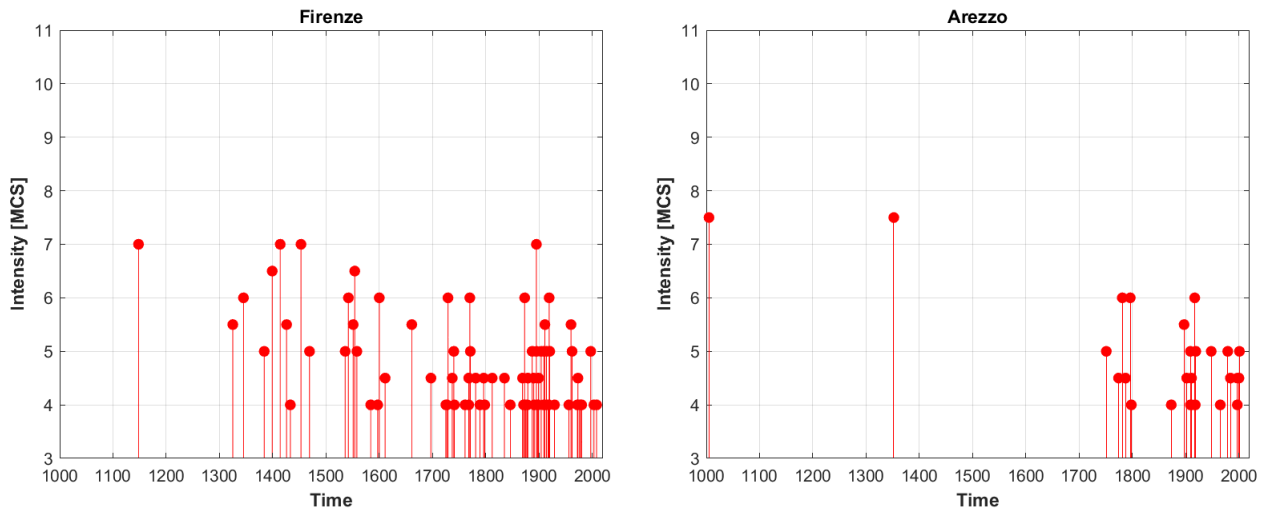



**Figure 1: Examples of seismic histories of two nearby Italian provincial capitals. The seismic history of Firenze (left) is extended and regular in time (for intensity larger than 4 MCS - Mercalli-Cancani-Sieberg scale; Sieberg, 1923), whereas the one of Arezzo (right) shows significant gaps in time (i.e., in the periods 1350–1750, and 1800–1850).**

In the application to the Italian territory described here, we first identify 133 sites corresponding to 97 provincial capitals and 36 localities selected trying to fulfil the above criteria.

We further check the representativeness of their seismic histories, provided by DBMI15 v1.5, comparing the seismic hazard estimates computed at each locality by means of the so-called "site" approach to PSHA (SASHA; D'Amico and Albarello, 2008) using: i) only the observed intensity data in DBMI15, and ii) the observed data integrated with "virtual" intensities calculated from earthquake parameters of the CPTI15 v1.5 catalogue (Rovida et al., 2016) through an intensity attenuation relationship (Pasolini et al., 2008, recalibrated by Lolli et al., 2019). High differences between the two resulting hazard estimates may indicate localities with "poor" seismic histories and/or with evident lack of data that should not be used for scoring.

On the basis of this analysis, which lead to eliminate or replace 13 localities that might bias the tests (six sites are retained to avoid large uncovered areas although they have poor seismic histories), and of the re-examination of the geographical distribution of the resulting sites, a further analysis is carried out to thin out very dense areas in Northern and Central Italy and to increase the density in some areas in the South. The final set of 124 locations selected for scoring is shown in Fig. 2.

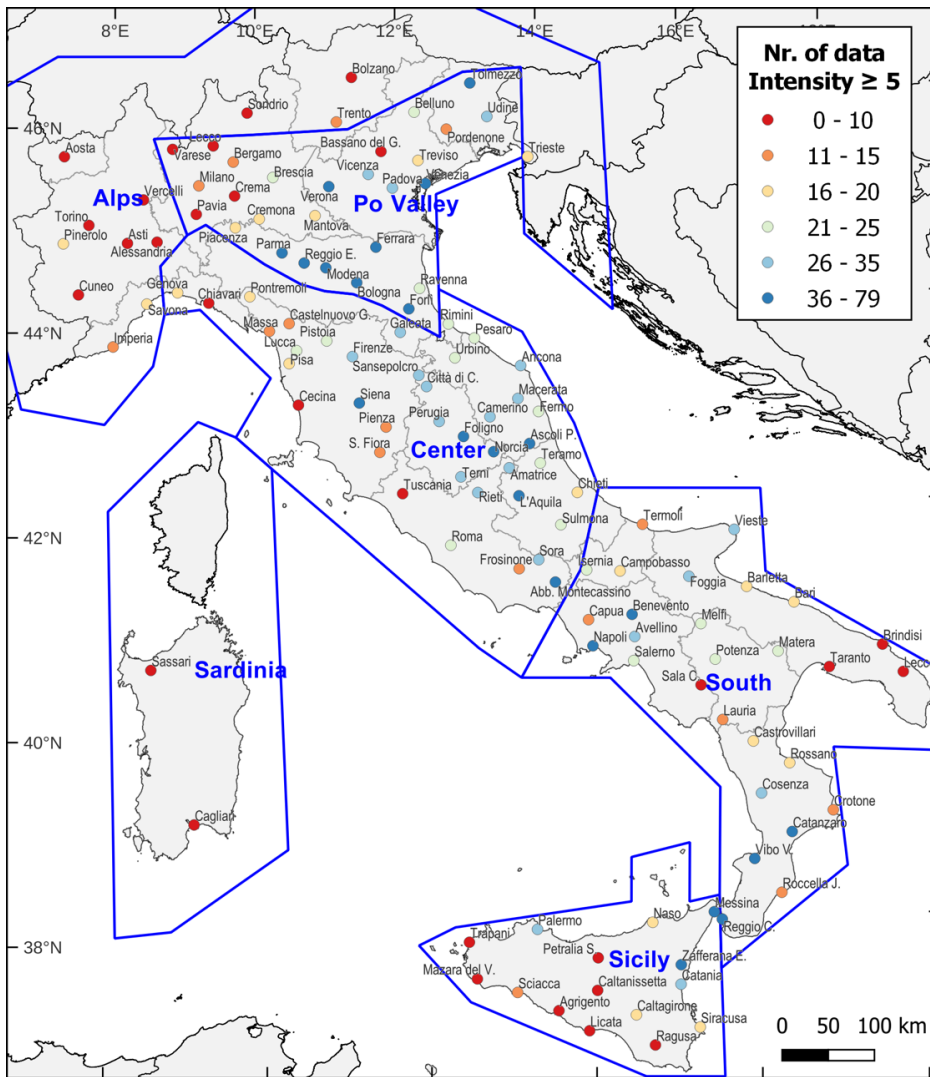



**Figure 2: Map of the 124 localities selected for scoring showing the number of intensity data ≥ 5 MCS, for each of them. The polygons identify the six macro-areas used for subsequent tests.**

110

## 2.2 Completeness periods of site seismic histories

In DBMI15, 9308 intensity data are referred to the selected localities and are associated to 2400 earthquakes spanning the period 1000–2014 and the whole range of intensity degrees, up to 10–11 MCS (Mercalli-Cancani-Sieberg scale; Sieberg, 1923).

115    However, the consistency check of the hazard estimates provided by a given model with the macroseismic observations available at the selected localities requires that the number of intensities expected from the model at each site is compared

with a complete set of observed intensities for each intensity degree. As a consequence, to calculate the number of macroseismic data, both observed and expected, at each site it is first necessary to estimate the completeness time intervals for each intensity degree, i.e., the periods in which it is reasonable to assume that all the earthquake effects above a given intensity have been actually reported in the seismic history (see Stucchi et al., 2004; Antonucci et al., 2023). For this reason, the completeness of the site seismic histories is different from the completeness of the earthquake catalogue. Indeed, the effects of an earthquake occurred in the complete period of the catalogue might not be recorded at a given site for several reasons (e.g., they were not documented, documentation exists but has not been analyzed, and so on).

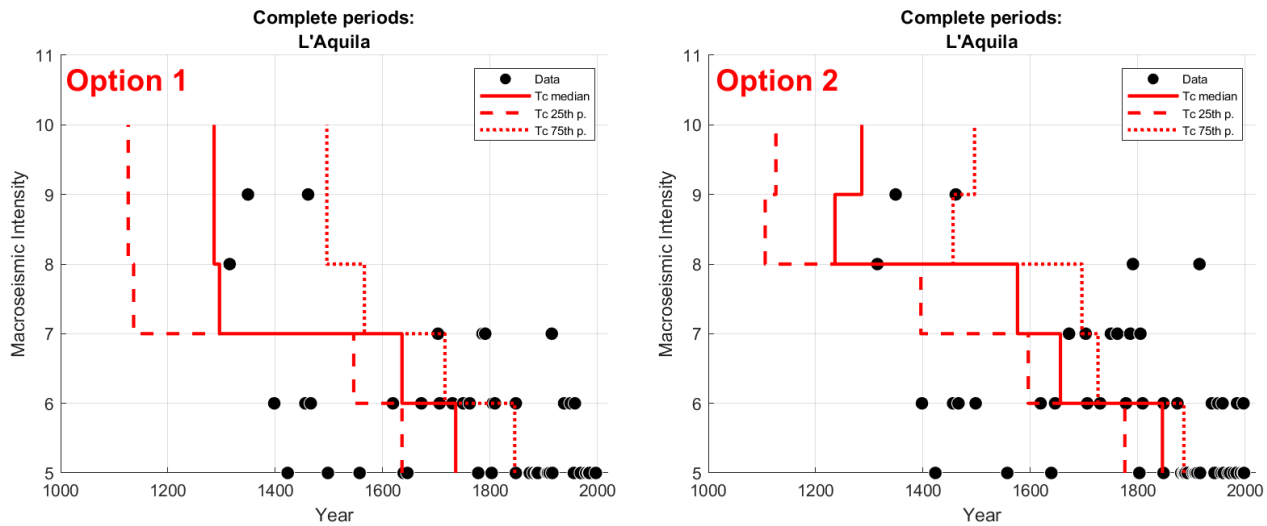
In our case study, the completeness time intervals for each site are defined using the statistical approach of Albarello et al. (2001) applied to observed data with intensity greater than or equal to 5 MCS according to the following procedure:

– intensity data related to earthquakes in CPTI15 identified as "mainshocks", according to the declustering method used in MPS19 (Gardner and Knopoff, 1974), are considered;

– only intensity data of earthquakes up to 2006 are used, because after that year the systematic collection of macroseismic data ceased and DBMI15 is incomplete (Antonucci, 2022);

– intensities expressed in DBMI15 as non-numerical values, e.g., F for "felt", HD for "heavy damage", and so on (see Rovida et al., 2020 for their complete list), are discarded;

– uncertain intensities between adjacent integer degrees (e.g., 6–7 MCS) are treated as either the lowest degree (option 1) or the highest one (option 2).

For each Macroseismic Intensity (MI) threshold, two completeness estimates are therefore obtained, in terms of the starting year of the complete period ($T_c$), referred to the two options for assigning the uncertain degrees described above. To take the uncertainty in the estimation of completeness into account, the two $T_c$ values corresponding to the median and the 75th percentile of the completeness function provided by the adopted method are considered, for a total of four $T_c$ values. The estimates of $T_c$ corresponding to the 25th percentile of the completeness function are not taken into account as they are considered unrealistic, especially for high degrees (see the example in Fig. 3). Finally, in case the completeness period of a given intensity threshold is shorter than that of the lower one (e.g., for degree 9 MCS in Fig. 3 on the right), the latter period is considered for both the thresholds.

145 **Figure 3: Example of completeness graph for the city of L'Aquila according to the two options for assigning uncertain degrees: on the left, uncertain degrees are assigned to the lowest degree (option 1); on the right, to the highest degree (option 2). The red bars indicate, for each MI threshold, the three estimates of the completeness starting year $T_c$ (the median value with solid line, the 25th and 75th percentiles with dashed lines). Black dots correspond to intensities observed up to 2006, extracted from DBMI15.**

## 2.3 Dataset of observed intensities

According to the procedures described above, the dataset of observed macroseismic intensities for each testing locality is built counting, for each MI degree, the numbers of data after the two different completeness starting years $T_c$, i.e., those corresponding to the median value and the 75th percentile of the completeness function, considering both options 1 and 2 for treating the uncertain degrees.

155 Thus, four estimates of the number of observed data for each intensity degree are obtained, corresponding to:

     i)   option 1 and the median $T_c$ value (2100 data);

     ii)  option 1 and the 75th percentile $T_c$ value (1671 data);

     iii) option 2 and the median $T_c$ value (2557 data);

     iv) option 2 and the 75th percentile $T_c$ value (2076 data).

160 (In brackets, the total number of data referred to all intensity degrees and selected localities is reported).

The resulting numbers of data are finally cumulated to obtain the observed exceedances for each considered MI degree at each locality.

## 2.4 Dataset of expected intensities

165   The number of expected intensity data at each site on the basis of the hazard estimates provided by the individual branches of MPS19 is computed as follows:

– the hazard curve for each branch is calculated assuming a value of $V_{S,30}$ equal to 600 m/s, that is for EC8 soil category B instead of A ($V_{S,30}$ = 800 m/s) to which the MPS19 model refers. This is because macroseismic intensity values quantify the earthquake effects (in particular the levels of building damage) observed in extended localities that, in

170   Italy, are generally located on class B soils ($360 \leq V_{S,30} < 800$ m/s, see e.g., Mori et al., 2020) rather than on rocky soils;

– from each hazard curve, expressed as annual rates of exceedance of different levels of shaking in terms of PGA or SA, the corresponding annual rates of occurrence are obtained. These are then converted into occurrence rates of different degrees of intensity $\lambda$(MI) through the Ground Motion Intensity Conversion Equation (GMICE) by Gomez

175   Capera et al. (2020), taking into account the associated uncertainties, as follows:

$$\lambda(\text{MI}) = \sum_{j=1}^{M} \lambda(x_j) P(\text{MI}|x_j) \qquad (2)$$

where $\lambda(x_j)$ is the annual occurrence rate of each of the $M$ levels of PGA (or SA) in the hazard curve and $P(\text{MI}|x_j)$

180   corresponds to the conditional probability distribution of the GMICE, as proposed by D'Amico and Albarello (2008);

– the rates of occurrence in intensity estimated in this way are then multiplied by the lengths of the corresponding completeness periods to obtain the number of macroseismic data expected for each intensity degree. As done for the observed intensity data, the four estimates of the completeness periods are considered (starting from the median $T_c$ value and the 75th percentile of the completeness function and for the two options for assigning the uncertain

185   degrees);

– the resulting numbers of data are finally cumulated to obtain the expected exceedances for each MI degree.

Figure 4 shows an example of the comparison between the number of observed and expected macroseismic data in the locality of Amatrice for different intensity thresholds.

190

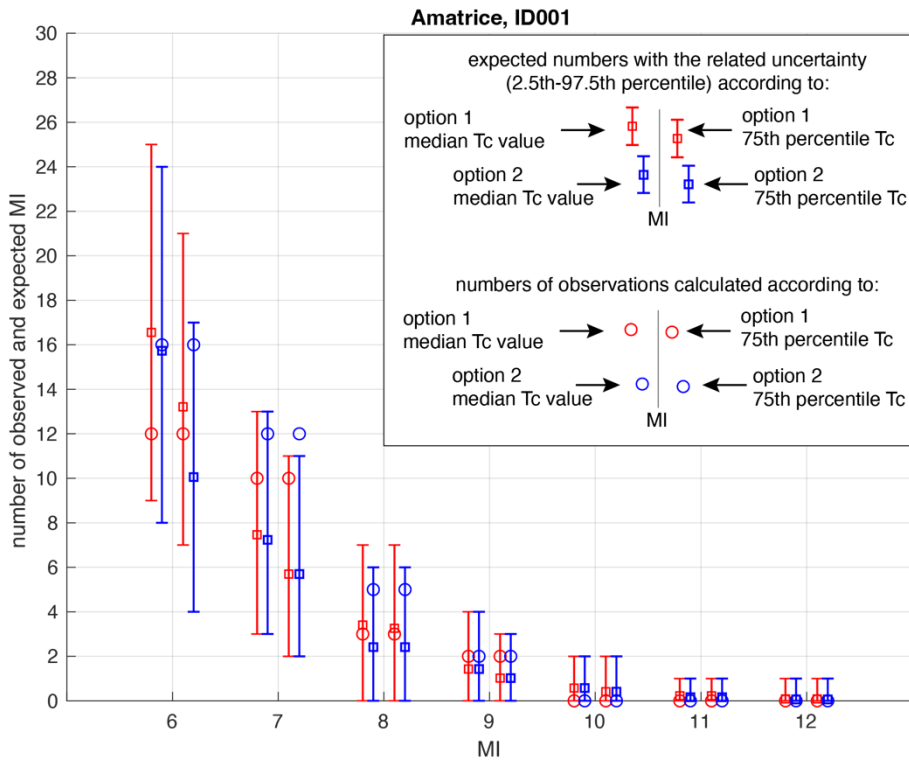



Figure 4: Comparison, for different MI thresholds, between the number of observed macroseismic data in the locality of Amatrice and the number of expected intensities from one of the branches of MPS19 (ID001, calculated for soil class B).

## 3 Consistency test between hazard estimates and macroseismic observations

The parameter used for evaluating the consistency of the predictions of a given PSHA model with the macroseismic observations available for the testing localities is the Log-Likelihood score LL (Eq. (1)). In this application, comparisons between forecasts and observations are made for individual branches (or models) of MPS19 starting from the hazard curves calculated at each testing site for soil class B ($V_{S,30}$ = 600 m/s) for PGA, SA 0.2 s and 1 s, that are considered the most relevant spectral periods for engineering purposes. The total number of analyzed branches is 282 out of the 564 of MPS19, because the hazard values estimated at the testing sites using the two alternative GMMs selected for the subduction zone are almost identical and only the branches adopting the model of Skarlatoudis et al. (2013), that obtained the highest weight, are considered.

### 3.1 Calculation of the Log-Likelihood score (LL)

205 As described in the previous section, for each considered testing locality and for each PSHA branch, four pairs of observed and expected numbers of intensity data are obtained for each MI threshold, corresponding to the four different estimates of the completeness periods, i.e., for the median value and the 75$^{th}$ percentile of the completeness function and for the two options for treating uncertain intensity degrees (see the example in Fig. 4).

For each site and branch, for each MI threshold and for each of the four pairs of observed and expected numbers of intensity

210 data, the probability $p$ of the tails of the Poisson distribution is calculated through the following algorithm (Zechar et al., 2010):

– if the number of observed data ($N_{obs}$) is greater than the number of expected ones ($N_{exp}$):

$$p = 1 - F((N_{obs} - 1); N_{exp}) \tag{3}$$

215

– if the number of observed data is lower than or equal to the number of expected ones:

$$p = F(N_{obs}; N_{exp}) \tag{4}$$

220 where $F$ is the right-continuous Poisson cumulative distribution function with expectation $N_{exp}$ evaluated at $N_{obs}$:

$$F(N_{obs}|N_{exp}) = e^{-N_{exp}} \sum_{k=0}^{|N_{obs}|} \frac{N_{exp}^{k}}{k!} \tag{5}$$

The two probabilities ($p$), defined in Eqs. (3) and (4), respectively answer one of the following questions: is the forecast too

225 low (Eq. (3)) or too high (Eq. (4)) compared to the observations?

For each site, we then calculate the weighted average of the four logarithmic scores (LL in Eq. (1)), considering the four pairs of observed and expected numbers of data for intensity greater than or equal to 6 (MI6+) and 8 (MI8+) MCS. These intensity levels correspond to the threshold of slight and structural building damage, respectively. The weighted average of observed and expected data is calculated by equally weighting the two estimates obtained from the median value and the 75$^{th}$

230 percentile of the completeness function and attributing different weights to the two options for treating uncertain degrees as follows: i) 0.75 to option 1 (i.e., uncertain degree assigned to the lower MI value), and ii) 0.25 to option 2 (i.e., uncertain degree assigned to the higher MI value). This choice is consistent with the meaning of uncertain intensity assignments described in Grünthal (1998).

The LL value calculated in this way is defined as LL$_{site}$.

## 3.2 Estimates of LL for each model

To identify the models that produce the hazard estimates most consistent with the macroseismic observations at the testing sites, we initially calculate, for each branch, the sum of the $LL_{site}$ values relating to the 124 selected localities, defined as $LL_{sum}$, for the three spectral periods (PGA, SA 0.2 s and 1 s) and the two intensity thresholds (MI6+ and MI8+) considered. Figure 5 shows the $LL_{sum}$ values for PGA for all branches: the smaller (closer to zero) is the value, the higher is the consistency between the model's outcomes and the observations.

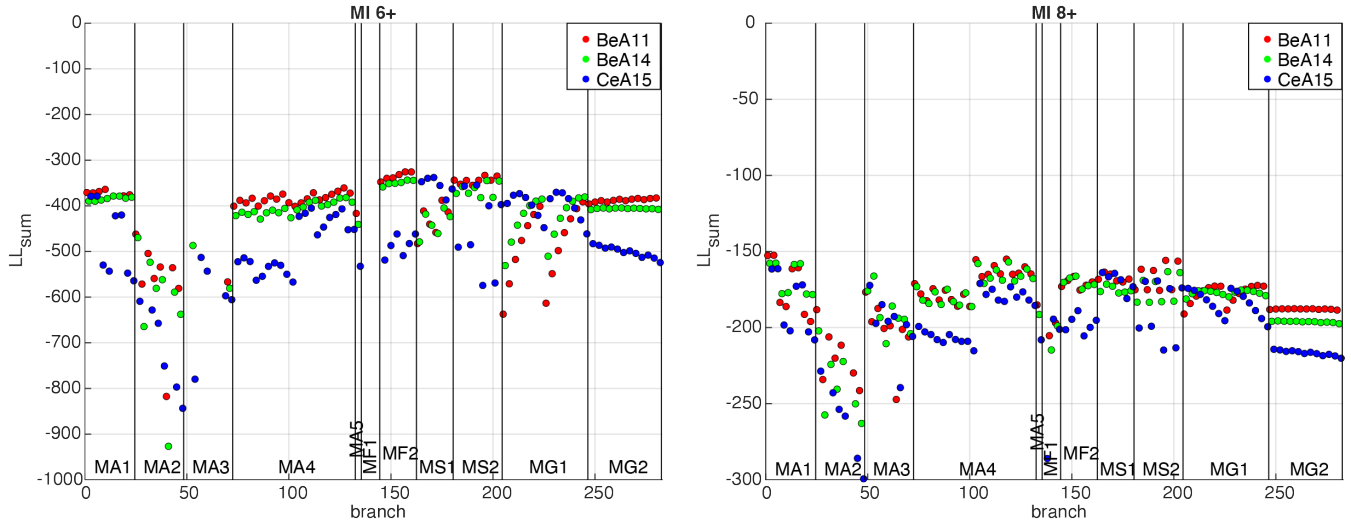



**Figure 5: Values of $LL_{sum}$ (sum of $LL_{site}$ of all localities) calculated for each of the 282 considered branches of MPS19 for PGA, MI6+ (left) and MI8+ (right). The branches are represented in abscissa from left to right grouped according to the 11 seismicity models (for the description of models, see Meletti et al., 2021; Visini et al., 2021), and colored according to the three GMMs adopted for active crustal areas, namely: "BeA11" (Bindi et al., 2011), "BeA14" (Bindi et al., 2014), "CeA15" (Cauzzi et al., 2015). Note that the y-axis for MI6+ is truncated for the purpose of visualization, as a few values tend toward negative infinity.**

Then, to test the performance of each branch over different regions of the Italian territory, we group the selected localities according to the six macro-areas defined in MPS19 to estimate the completeness of the CPTI15 catalogue, that is: Alps, Po Valley, Center, South, Sardinia, Sicily (see Fig. 2). Since these macro-areas include different numbers of sites, the average (instead of the sum) of the $LL_{site}$ values is calculated for both the entire set of 124 localities and for the localities in each macro-area (Sardinia is excluded since it has only two testing sites). Therefore, six $LL_{mean}$ values are obtained for each branch. Figure 6 shows the resulting $LL_{mean}$ values for PGA, for the two intensity thresholds MI6+ and MI8+ and for each adopted GMM.

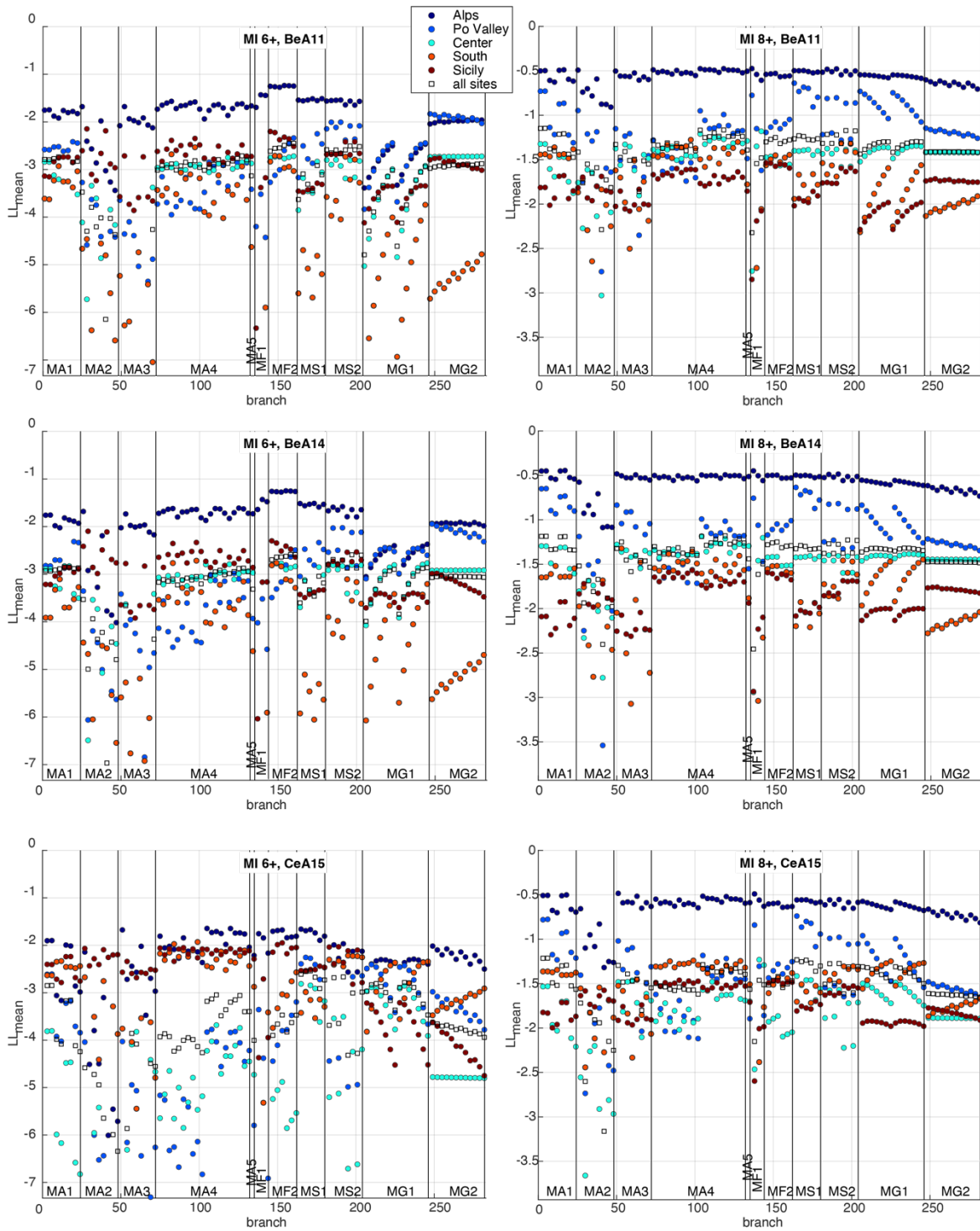



**Figure 6: Values of LL_mean for each considered branch of MPS19 for the localities in five macro-areas and for all the sites, for PGA, MI6+ (left) and MI8+ (right), and for each adopted GMM. The branches are represented in abscissa from left to right grouped according to the 11 seismicity models.**

As shown, almost all the branches seem to give a good agreement (i.e., $LL_{mean}$ values closer to zero) in the Alps macro-area, characterized by a much smaller number of sites. In the other four macro-areas, that are more significant in terms of the number of sites, the results appear to be very different depending on the different seismicity models. In particular, for the BeA11 and BeA14 GMMs, some groups of branches (e.g., MA2, MA3, MF1, MG1, MG2) show generally poorer performance in terms of $LL_{mean}$ values and a considerable geographical scatter, whereas others (e.g., MA1, MA4, MF2, MS2) show values of $LL_{mean}$ that are generally smaller and more stable in the four macro-areas. The plots of $LL_{mean}$ values for SA 0.2 s and 1 s are reported in the Supplement (Fig. S1 and S2).

In order to evaluate the stability of the performance of each branch in the different areas, we then calculate the dispersion of the $LL_{mean}$ values among the four macro-areas including the highest number of localities (Po Valley, Center, South, Sicily) as the width of the interval between the 2.5[th] and 97.5[th] percentiles. The percentiles are estimated using a non-parametric distribution of the four $LL_{mean}$ values. Obviously, a different choice of the distribution might lead to changes in the percentiles, but the aim is only to give an order of magnitude of the dispersion among the measures in the macro-areas. Figure 7 shows the dispersion values computed for PGA, for the two intensity thresholds considered (plots for SA 0.2 s and 1 s are displayed in Fig. S3 of the Supplement).

The $LL_{mean}$ values computed using the entire set of localities and the relative geographical dispersion are then used to establish a ranking of the branches.

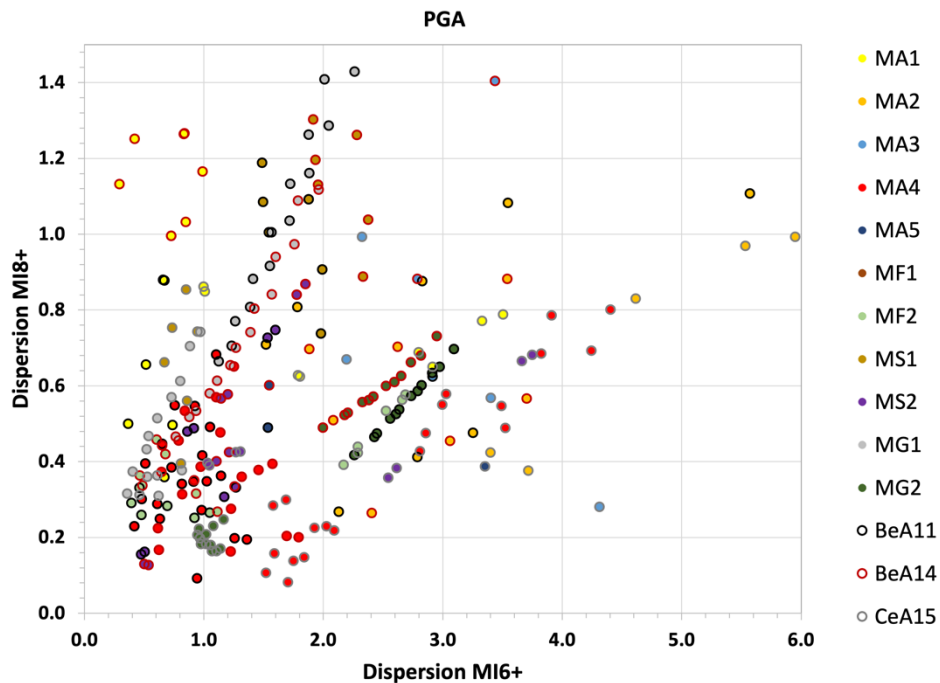



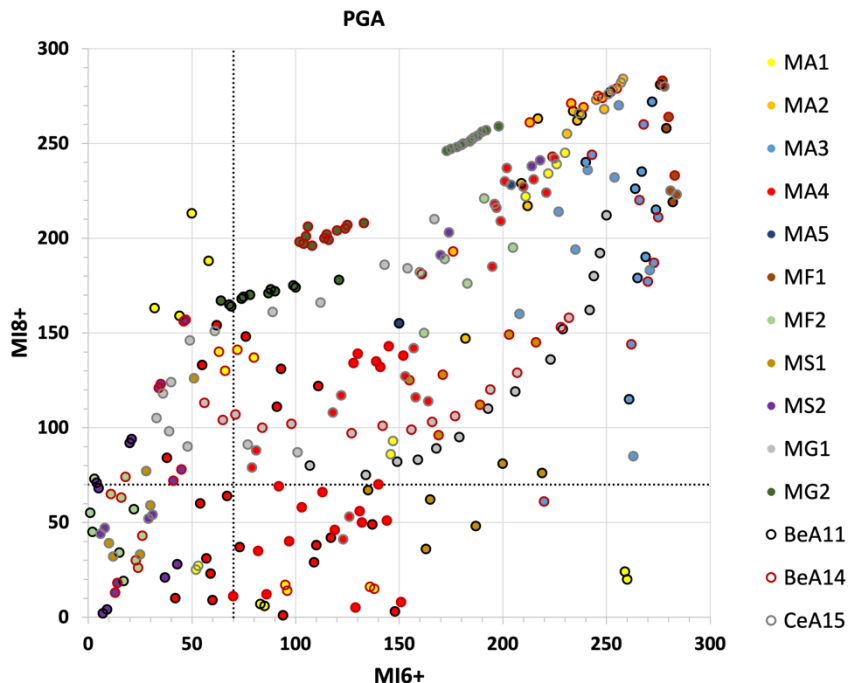
**Figure 7: Dispersion of the $LL_{mean}$ values among the four more representative macro-areas for each branch, for PGA, MI6+ and MI8+. The color of the dots indicates the seismicity model, while the color of the borders indicates the GMM used in that branch.**

### 3.3 Ranking of the models

280    Following the described procedure, a $LL_{mean}$ value for MI6+ and MI8+, for the three spectral periods (i.e., PGA, SA 0.2 s
and 1 s), is assigned to each of the 282 considered branches of MPS19, as well as an estimate of the dispersion of these
values among the four more representative macro-areas.

For each spectral period and MI threshold, the branches are then ranked according to the values of $LL_{mean}$ and the relative
geographical dispersion, assigning the 1st place to the branch with the "best" result ($LL_{mean}$ value or dispersion closest to
285    zero) and the 282nd place to the "worst" one.

Initially, comparison plots of the ranks based on $LL_{mean}$ values for the two MI thresholds are produced, focusing the attention
on those branches that fall within the 10th percentile. This choice, however subjective, result too restrictive, since none of the
branches fall within this range for all the considered spectral periods. It is then decided to expand the selection criterion.
Taking into account the first 70 positions (corresponding to the first quartile) in the $LL_{mean}$ tests, for both intensity thresholds
290    and for the three spectral periods, we select 35 branches for PGA and 37 branches for SA 0.2 s and 1 s, representing all the
GMMs used and different seismicity models of MPS19.

Figure 8 shows the placement of each branch in the $LL_{mean}$ test for PGA; the plots for SA 0.2 s and 1 s are reported in the
Supplement (Fig. S4). In all the plots, the best ranks are generally occupied by the same models, for both MI thresholds.

295    **Figure 8: Comparison between the ranking of the branches for PGA, based on the $LL_{mean}$ values for MI6+ and MI8+. The color of
the dots indicates the seismicity model, while the color of the borders indicates the GMM used in that branch. Black dotted lines
identify the 70th position in both rankings.**

The geographical dispersion of $LL_{mean}$ values among the four more representative macro-areas is then considered. For each spectral period, a ranking of the branches is made also according to this parameter and the ranks are grouped in three classes, for both MI6+ and MI8+, that is: i) rank ≤100, ii) rank 101–150, and iii) rank >150.

An overall rank (from 1, best rank, to 5) is then assigned to each selected branch, for each of the three spectral periods, based on the ranking class for the two intensity thresholds (see the abacus in Table 1).

**Table 1: Abacus built to assign an overall rank to each selected branch, for each spectral period considered, on the basis of its rank resulting from the dispersion of $LL_{mean}$ values in the four macro-areas.**

| | | Dispersion rank MI8+ | | |
|---|---|---|---|---|
| | | 1-100 | 101-150 | >150 |
| Dispersion rank MI6+ | 1-100 | 1 | 2 | 3 |
| | 101-150 | 2 | 3 | 4 |
| | >150 | 3 | 4 | 5 |

The overall rank could allow practitioners to further sample and/or (re-)weight the various branches according to the practical constraints of a specific application. Table 2 shows the overall ranks for the 20 best performing selected branches for PGA, SA 0.2 s and 1 s. As an example, if one decides to consider those models with an overall rank equal to 1 or 2 for all the spectral periods, only the first six should be selected.

Table 2: Overall ranks of the 20 best performing selected branches, according to the abacus in Table 1. The following columns report the position in every computed rank; the cells with gray background mark the branches falling in the first 70 positions in the LL$_{mean}$ ranking.

| id | Seismicity model | GMM | Overall rank PGA | Overall rank 0.2 s | Overall rank 1 s | PGA Rank_LL_MI6+ | PGA Rank_LL_MI8+ | PGA Rank_Disp_MI6+ | PGA Rank_Disp_MI8+ | SA 0.2 s Rank_LL_MI6+ | SA 0.2 s Rank_LL_MI8+ | SA 0.2 s Rank_Disp_MI6+ | SA 0.2 s Rank_Disp_MI8+ | SA 1 s Rank_LL_MI6+ | SA 1 s Rank_LL_MI8+ | SA 1 s Rank_Disp_MI6+ | SA 1 s Rank_Disp_MI8+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID255 | MA4 | BeA14 | 1 | 1 | 1 | 70 | 11 | 53 | 53 | 59 | 17 | 85 | 3 | 22 | 16 | 3 | 71 |
| ID303 | MF2 | BeA14 | 1 | 1 | 2 | 24 | 26 | 38 | 93 | 21 | 40 | 88 | 27 | 7 | 52 | 54 | 117 |
| ID297 | MF2 | BeA14 | 1 | 2 | 1 | 23 | 30 | 70 | 54 | 27 | 50 | 123 | 9 | 4 | 46 | 80 | 84 |
| ID291 | MF2 | BeA14 | 1 | 2 | 2 | 26 | 43 | 100 | 39 | 42 | 70 | 146 | 4 | 1 | 44 | 123 | 61 |
| ID391 | MS2 | BeA11 | 2 | 1 | 2 | 7 | 2 | 61 | 117 | 1 | 29 | 31 | 63 | 32 | 51 | 134 | 94 |
| ID403 | MS2 | BeA11 | 2 | 1 | 2 | 9 | 4 | 65 | 119 | 2 | 26 | 34 | 67 | 24 | 47 | 126 | 89 |
| ID393 | MS2 | BeA14 | 3 | 1 | 2 | 13 | 13 | 104 | 149 | 8 | 6 | 51 | 90 | 14 | 15 | 53 | 133 |
| ID395 | MS2 | CeA15 | 2 | 3 | 2 | 31 | 54 | 121 | 98 | 25 | 37 | 117 | 148 | 6 | 18 | 83 | 136 |
| ID407 | MS2 | CeA15 | 2 | 3 | 2 | 29 | 52 | 119 | 97 | 23 | 33 | 111 | 146 | 8 | 26 | 84 | 141 |
| ID405 | MS2 | BeA14 | 4 | 1 | 2 | 14 | 18 | 108 | 157 | 9 | 7 | 61 | 94 | 18 | 17 | 56 | 137 |
| ID367 | MS2 | BeA11 | 4 | 2 | 1 | 37 | 21 | 134 | 198 | 5 | 21 | 106 | 73 | 3 | 30 | 78 | 99 |
| ID379 | MS2 | BeA11 | 4 | 2 | 2 | 43 | 28 | 144 | 205 | 6 | 22 | 115 | 81 | 2 | 31 | 75 | 108 |
| ID249 | MA4 | BeA14 | | 1 | 1 | 82 | 35 | 31 | 71 | 55 | 34 | 24 | 17 | 48 | 42 | 34 | 96 |
| ID371 | MS2 | CeA15 | 1 | 2 | | 8 | 47 | 88 | 81 | 10 | 27 | 84 | 131 | 59 | 154 | 107 | 242 |
| ID383 | MS2 | CeA15 | 1 | 2 | | 6 | 44 | 87 | 86 | 7 | 25 | 82 | 133 | 68 | 175 | 117 | 245 |
| ID251 | MA4 | CeA15 | | 2 | 2 | 81 | 88 | 151 | 6 | 62 | 55 | 124 | 7 | 30 | 27 | 8 | 145 |
| ID175 | MA4 | BeA11 | 1 | | | 67 | 64 | 71 | 2 | 96 | 178 | 122 | 145 | 229 | 266 | 253 | 182 |
| ID223 | MA4 | BeA11 | 1 | | | 60 | 9 | 42 | 77 | 36 | 89 | 67 | 49 | 155 | 181 | 169 | 37 |
| ID247 | MA4 | BeA11 | 1 | | | 57 | 31 | 13 | 49 | 32 | 113 | 12 | 91 | 137 | 200 | 138 | 27 |
| ID253 | MA4 | BeA11 | 1 | | | 42 | 10 | 25 | 46 | 30 | 101 | 52 | 72 | 168 | 210 | 180 | 41 |

## 4 Discussion and conclusions

We have introduced a new scoring strategy that may be used to rank and sample the multiple models/branches of a PSHA model. Scoring is inherently different from testing: the former term indicates approaches devoted at ranking and eventually weighting a set of competing models, whereas testing procedures aim at evaluating the absolute predictive accuracy of each model, indicating if its outcomes are/are not compatible with observations to a given significance level threshold. Therefore, testing can allow to identify possibly wrong PSHA models, whereas scoring is aimed to compare models according to a specific metric of interest.

For the sake of example, we have scored and ranked alternative branches of the MPS19 seismic hazard model of Italy (Meletti et al., 2021) according to their fit with long-term macroseismic intensity data available in a large set of sites, with

the scope of selecting the models/branches that minimize the difference between PSHA outcomes and macroseismic observations at these sites. To properly compare the performance of the different branches, a Log-Likelihood score has been assigned to each of them based on the comparison between numbers of expected and observed intensity data at each site for
330 different shaking levels and spectral periods, not considering single return periods but the entire hazard curve.

In countries such as Italy, where the historical record is hundreds-of-years long, i.e., much more than the instrumental one, and macroseismic information covers the whole territory (Locati et al., 2022), the use of macroseismic intensity observations for scoring PSHA models could be more suitable than accelerometric recordings to consider the effects of earthquakes with large magnitudes and long return periods.

335 Of course, comparing PSHA outcomes in terms of PGA or SA with macroseismic data requires caution due to the use of GMICEs, that are empirical conversion relationships characterized by large uncertainties to be taken into account. In fact, if one simply converts the ground motion value (e.g., PGA) resulting by a PSHA model into macroseismic intensity just using the average estimates and discarding associated variance, the comparison could be severely biased. In the scoring procedure presented here, this issue is solved through the convolution of the relevant probability distributions (i.e., hazard curves and
340 GMICE), as proposed by D'Amico and Albarello (2008). Moreover, the procedure takes into account both the peculiar nature of intensity values (discrete, ordinal, range-limited) and associated uncertainties (uncertain intensity values between adjacent integer degrees, completeness of site seismic history, etc.).

A further crucial issue related to using macroseismic intensity data for empirical scoring concerns the selection of sites where to compare PSHA models' outputs with available observations. In fact, to have a representative set of localities to
345 perform tests, selected sites have to guarantee a geographical coverage as dense and uniform as possible throughout the study area (for both high and low hazard regions) as well as a significant number of macroseismic data at each site, covering long time periods and a wide range of intensity values. This clearly limits the use of macroseismic data as observables to those countries with long records of documentary information about the effects of past earthquakes at a sufficient number of sites (e.g., Fäh et al., 2011 for Switzerland; BRGM-EDF-IRSN/SisFrance, 2017 for France).

350 The presented procedure can be applied to any kind of model and set of observational data, for instance to rank and select branches of a complex PSHA model to get one outcome that better satisfies specific stakeholders' needs. In this regard, it is important to remark that our approach is based on a rigorous and quantitative procedure, although the definition of the thresholds and ranks for selecting branches is a subjective choice that depends on specific considerations and aims.

**Data availability**

355 CPTI15 v1.5 is available at https://doi.org/10.6092/ingv.it-cpti15

DBMI15 v1.5 is available at https://doi.org/10.6092/ingv.it-dbmi15

## Author contribution

All authors designed the scoring strategy and prepared the manuscript. VD wrote the draft of the paper and contributed to the selection of the testing sites, the estimation of the expected intensities and the ranking of the models. FV computed the completeness periods, the number of observed/expected intensity data and the LL values. AR performed the selection of the testing sites and contributed to the estimation of the completeness periods. WM proposed the scoring rule and CM performed the ranking of the models.

## Competing interests

The authors declare that they have no conflict of interest.

## Acknowledgements

The authors wish to thank Dario Albarello and an anonymous referee for their useful comments and suggestions. Thanks are due also to Andrea Antonucci and Matteo Taroni for their support respectively in the selection of the testing sites and in the computation of the LL values. This study benefited from funding provided by the Italian Presidenza del Consiglio dei Ministri – Dipartimento della Protezione Civile (DPC), in the framework of the DPC-INGV Agreement B1 (2020–2021), and then by the Seismic Hazard Centre (Centro di Pericolosità Sismica, CPS) of the Istituto Nazionale di Geofisica e Vulcanologia (INGV). This paper does not represent DPC official opinion and policies.

## References

Abrahamson, N., Gregor, N., and Addo, K.: BC Hydro ground motion prediction equations for subduction earthquakes, Earthq. Spectra, 32 (1), 23–44, https://doi.org/10.1193/051712EQS188MR, 2016.

Albarello, D., Camassi, R., and Rebez, A.: Detection of space and time heterogeneity in the completeness of a seismic catalog by a statistical approach: an application to the Italian area, B. Seismol. Soc. Am., 91, 1694–1703, https://doi.org/10.1785/0120000058, 2001.

Albarello, D. and D'Amico, V.: Scoring and testing procedures devoted to probabilistic seismic hazard assessment, Surv. Geophys., 36, 269–293, doi.org/10.1007/s10712-015-9316-4, 2015.

Antonucci, A.: A probabilistic approach for integrating macroseismic data and its application to estimate the data completeness, Ph.D. Thesis, University of Pisa, 222 pp. https://etd.adm.unipi.it/theses/available/etd-02222022-104750/unrestricted/Andrea_Antonucci_PhD_Thesis.pdf, 2022.

Antonucci, A., Rovida, A., D'Amico, V., and Albarello, D.: Looking for undocumented earthquake effects: a probabilistic analysis of Italian macroseismic data, Nat. Hazard Earth. Sys., 23, 1805–1816, doi.org/10.5194/nhess-23-1805-2023, 2023.

385 Bindi, D., Massa, M., Luzi, L., Ameri, G., Pacor, F., Puglia, R., and Augliera, P.: Pan-European ground-motion prediction equations for the average horizontal component of PGA, PGV, and 5%-damped PSA at spectral periods up to 3.0 s using the RESORCE dataset, B. Earthq. Eng., 12 (1), 391–430, doi: 10.1007/s10518-013-9525-5, 2014.

Bindi, D., Pacor, F., Luzi, L., Puglia, R., Massa, M., Ameri, G., and Paolucci, R.: Ground motion prediction equations derived from the Italian strong motion database, B. Earthq. Eng., 9 (6), 1899–1920, doi: 10.1007/s10518-011-9313-z, 2011.

390 BRGM-EDF-IRSN/SisFrance: Histoire et caractéristiques des séismes ressentis en France. http://www.sisfrance.net, 2017.

Cauzzi, C., Faccioli, E., Vanini, M., and Bianchini, A.: Updated predictive equations for broadband (0.01–10 s) horizontal response spectra and peak ground motions, based on a global dataset of digital acceleration records, B. Earthq. Eng., 13, 1587–1612, doi: 10.1007 / s10518-014-9685-y, 2015.

D'Amico, V. and Albarello, D.: SASHA: a computer program to assess seismic hazard from intensity data, Seismol. Res.
395 Lett., 79 (5), 663–671, doi: 10.1785/gssrl. 79.5.663, 2008.

Danciu, L., Nandan, S., Reyes, C., Basili, R., Weatherill, G., Beauval, C., Rovida, A., Vilanova, S., Sesetyan, K., Bard, P-Y., Cotton, F., Wiemer, S., and Giardini, D.: The 2020 update of the European Seismic Hazard Model: Model Overview. EFEHR Technical Report 001, v1.0.0, https://doi.org/10.12686/a15, 2021.

Fäh, D., Giardini, D., Kästli, P., Deichmann, N., Gisler, M., Schwarz-Zanetti, G., Alvarez-Rubio, S., Sellami, S., Edwards,
400 B., Allmann, B., Bethmann, F., Wössner, J., Gassner-Stamm, G., Fritsche, S., and Eberhard, D.: ECOS-09 Earthquake Catalogue of Switzerland Release 2011 Report and Database. Public catalogue, 17. 4. 2011. Swiss Seismological Service ETH Zurich, Report SED/RISK/R/001/20110417, 42 pp + Appendixes, 2011.

Gardner, J. K. and Knopoff, L.: Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian? B. Seismol. Soc. Am., 64, 5, 1363–1367, 1974.

405 Gerstenberger, M. C., Marzocchi, W., Allen, T., Pagani, M., Adams, J., Danciu, L., Field, E. H., Fujiwara, H., Luco, N., Ma, K.-F., Meletti, C., and Petersen, M. D.: Probabilistic seismic hazard analysis at regional and national scale: state of the art and future challenges, Rev. Geophys., 58, e2019RG000653, 1–49, doi: 10.1029/2019RG000653, 2020.

Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, J. Am. Stat. Assoc., 102, no. 477, 359–378, doi: 10.1198/016214506000001437, 2007.

410 Gomez Capera, A. A., D'Amico, M., Lanzano, G., Locati, M., and Santulin, M.: Relationships between ground motion parameters and macroseismic intensity for Italy, B. Earthq. Eng., 18, 5143–5164, https://doi.org/10.1007/s10518-020-00905-0, 2020.

Grünthal, G.: European Macroseismic Scale 1998 (EMS-98), Cahiers du Centre Européen de Géodynamique et de Séismologie, 13, 99 pp., 1998.

415 Lanzano, G., Luzi, L., D'Amico, V., Pacor, F., Meletti, C., Marzocchi, W., Rotondi, R., and Varini, E.: Ground Motion Models for the new seismic hazard model of Italy (MPS19): selection for active shallow crustal regions and subduction zones, B. Earthq. Eng., 18, 3487–3516, doi: 10.1007/s10518-020-00850-y, 2020.

Lanzano, G. and Luzi, L.: A ground motion model for volcanic areas in Italy, B. Earthq. Eng., 18, 57–76, doi: 10.1007/s10518-019-00735-9, 2020.

Locati, M., Camassi, R., Rovida, A., Ercolani, E., Bernardini, F., Castelli, V., Caracciolo, C. H., Tertulliani, A., Rossi, A., Azzaro, R., D'Amico, S., Conte, S., and Rocchetti, E.: Italian Macroseismic Database (DBMI15), version 1.5. Istituto Nazionale di Geofisica e Vulcanologia (INGV), https://doi.org/10.6092/INGV.IT-DBMI15, 2016.

Locati, M., Camassi, R., Rovida, A., Ercolani, E., Bernardini, F., Castelli, V., Caracciolo, C. H., Tertulliani, A., Rossi, A., Azzaro, R., D'Amico, S., and Antonucci, A.: Italian Macroseismic Database (DBMI15), version 4.0. Istituto Nazionale di Geofisica e Vulcanologia (INGV), https://doi.org/10.13127/DBMI/DBMI15.4, 2022.

Lolli, B., Pasolini, C., Gasperini, P., and Vannucci, G.: Product 4.8: Recalibration of the prediction equation by Pasolini et al. (2008), in: The seismic hazard model MPS19, Final report, Meletti, C., Marzocchi, W. (Eds.), CPS-INGV, Rome, 168 pp. + 2 Appendixes, 2019.

Meletti, C., Marzocchi, W., D'Amico, V., Lanzano, G., Luzi, L., Martinelli, F., Pace, B., Rovida, A., Taroni, M., Visini, F., and the MPS19 Working Group: The new Italian seismic hazard model (MPS19), Ann. Geophys.-Italy, 64, 1, 1–12, SE112, doi:10.4401/ag-8579, 2021.

Mori, F., Mendicelli, A., Moscatelli, M., Romagnoli, G., Peronace, E., and Naso, G.: A new $V_{s30}$ map for Italy based on the seismic microzonation dataset, Eng. Geol., 275, 105745, https://doi.org/10.1016/j.enggeo.2020.105745, 2020.

Pasolini, C., Albarello, D., Gasperini, P., D'Amico, V., and Lolli, B.: The attenuation of seismic intensity in Italy, Part II: modeling and validation, B. Seismol. Soc. Am., 98, 692–708, https://doi.org/10.1785/0120070021, 2008.

Petersen, M. D., Shumway, A. M., Powers, P. M., Field, E. H., Moschetti, M. P., Jaiswal, K. S., Milner, K. R., Rezaeian, S., Frankel, A. D., Llenos, A. L., Michael, A. J., Altekruse, J. M., Ahdi, S. K., Withers, K. B., Mueller, C. S., Zeng, Y., Chase, R. E., Salditch, L. M., Luco, N., Rukstales, K. S., Herrick, J. A., Girot, D. L., Aagaard, B. T., Bender, A. M., Blanpied, M. L., Briggs, R. W., Boyd, O. S., Clayton, B. S., DuRoss, C. B., Evans, E. L., Haeussler, P. J., Hatem, A. E., Haynie, K. L., Hearn, E. H., Johnson, K. M., Kortum, Z. A., Kwong, N. S., Makdisi, A. J., Mason, H. B., McNamara, D. E., McPhillips, D. F., Okubo, P. G., Page, M. T., Pollitz, F. F., Rubinstein, J. L., Shaw, B. E., Shen, Z-K., Shiro, B. R., Smith, J. A., Stephenson, W. J., Thompson, E. M., Thompson Jobe, J. A., Wirth, E. A., and Witter, R. C.: The 2023 US 50-State National Seismic Hazard Model: Overview and implications, Earthq. Spectra, 40 (1), 5–88, doi:10.1177/87552930231215428, 2024.

Rovida, A., Locati, M., Camassi, R., Lolli, B., and Gasperini, P.: Italian Parametric Earthquake Catalogue (CPTI15), version 1.5. Istituto Nazionale di Geofisica e Vulcanologia (INGV), https://doi.org/10.6092/INGV.IT-CPTI15, 2016.

Rovida, A., Locati, M., Camassi, R., Lolli, B., and Gasperini, P.: The Italian earthquake catalogue CPTI15, B. Earthq. Eng., 18, 2953–2984, https://doi.org/10.1007/s10518-020-00818-y, 2020.

Sieberg, A.: Geologische, physikalische und angewandte Erdbebenkunde, G. Fischer, Jena, 1923.

Skarlatoudis, A. A., Papazachos, C. B., Margaris, B. N., Ventouzi, C., Kalogeras, I., and the EGELADOS Group: Ground-Motion Prediction Equations of intermediate-depth earthquakes in the Hellenic Arc, Southern Aegean subduction area, B. Seismol. Soc. Am., 103 (3), 1952–1968, doi: 10.1785/0120120265, 2013.

Stucchi, M., Albini, P., Mirto, M., and Rebez, A.: Assessing the completeness of Italian historical earthquake data, Ann. Geophys.-Italy, 47, 2–3, https://doi.org/10.4401/ag-3330, 2004.

Stucchi, M., Meletti, C., Montaldo, V., Crowley, H., Calvi, G. M., and Boschi, E.: Seismic hazard assessment (2003-2009) for the Italian building code, B. Seismol. Soc. Am., 101, 1885–1911, doi: 10.1785/0120100130, 2011.

Visini, F., Pace, B., Meletti, C., Marzocchi, W., Akinci, A., Azzaro, R., Barani, S., Barberi, G., Barreca, G., Basili, R., Bird, P., Bonini, M., Burrato, P., Busetti, M., Carafa, M. M. C., Cocina, O., Console, R., Corti, G., D'Agostino, N., D'Amico, S., D'Amico, V., Dal Cin, M., Falcone, G., Fracassi, U., Gee, R., Kastelic, V., Lai, C. G., Langer, H., Maesano, F. E., Marchesini, A., Martelli, L., Monaco, C., Murru, M., Peruzza, L., Poli, M. E., Pondrelli, S., Rebez, A., Rotondi, R., Rovida, A., Sani, F., Santulin, M., Scafidi, D., Selva, J., Slejko, D., Spallarossa, D., Tamaro, A., Tarabusi, G., Taroni, M., Tiberti, M. M., Tusa, G., Tuvè, T., Valensise, G., Vannoli, P., Varini, E., Zanferrari, A., and Zuccolo, E.: Earthquake rupture forecasts for the MPS19 seismic hazard model of Italy, Ann. Geophys.-Italy, 64, 2, SE220, doi: 10.4401/ag-8608, 2021.

Woessner, J., Danciu, L., Giardini, D., Crowley, H., Cotton, F., Grünthal, G., Valensise, G., Arvidsson, R., Basili, R., Demircioglu, M. B., Hiemer, S., Meletti, C., Musson, R., Rovida, A., Sesetyan, K., Stucchi, M., and the SHARE consortium: The 2013 European seismic hazard model: key components and results, B. Earthq. Eng., 13, 3553–3596, doi: 10.1007/s10518-015-9795-1, 2015.

Zechar, J. D., Gerstenberger, M. C., and Rhoades, D.: Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts, B. Seismol. Soc. Am., 100 (3), 1184-1195, doi: 10.1785/0120090192, 2010.