A scoring test on probabilistic seismic hazard estimates in Italy

D. Albarello¹, L. Peruzza², and V. D’Amico³

¹Dipartimento di Scienze Fisiche, della Terra e dell’Ambiente, Università degli Studi di Siena, Siena, Italy
²OGS (Istituto Nazionale di Oceanografia e di Geofisica Sperimentale), Trieste, Italy
³Istituto Nazionale di Geofisica e Vulcanologia, Pisa, Italy

Received: 1 July 2014 – Accepted: 3 August 2014 – Published: 8 September 2014

Correspondence to: L. Peruzza (lperuzza@inogs.it)

Published by Copernicus Publications on behalf of the European Geosciences Union.
Abstract

Probabilistic estimates of seismic hazard represent a basic element for planning seismic risk reduction strategies and they are key elements of seismic regulation. Due to its importance, it is mandatory to select most effective estimates among the available ones. A possible empirical scoring strategy is described here and is applied to a number of time-independent hazard estimates available in Italy both at national and regional scale. Scoring is based on the comparison of outcomes provided by available computational models at a number of accelerometric sites where observations are available for 25 years. This comparison also allows identifying computational models providing outcomes that contrast observations and thus should be discarded. The analysis shows that most of hazard estimates so far proposed for Italy do not contrast with observations and some computational models perform significantly better than the others do. Furthermore, one can see that, at least locally, older estimates can perform better that the most recent ones. Finally, since the same computational model can perform differently depending on the region considered and on average return time of concern, no single model can be considered as the best performing one. This implies that time-by-time, the most suitable model must be selected by considering the specific problem of concern.

1 Introduction

Seismic hazard assessment is a basic tool for risk estimates necessary to develop effective preventive strategies against seismic damage. Being in essence a forecasting of future ground shaking, uncertainty is a basic element of seismic hazard and this requires specific formalizations based on a probabilistic formalism (probabilistic seismic hazard assessment or PSHA) to manage available information by providing likelihood estimates for each possible ground-shaking level (hazard curve). Information considered on purpose includes deterministic (e.g., geometry of...
seismogenic structures or seismic waves propagation patterns) and statistical (e.g. average seismicity rates) elements. The last ones aim at managing the lack of information about important elements of seismic hazard (e.g., seismogenic activity of the faults) that results into an intrinsic aleatory character of seismic occurrences (the so called “aleatory uncertainty”). Actually, many PSHA procedures exist that mainly differentiate for the relative roles played by deterministic and statistical elements. Procedures span from purely deterministic approaches assuming a nearly complete knowledge of the seismic process (e.g., Peresan et al., 2011) to purely statistical analyses assuming a nearly complete ignorance of underlying physical processes (e.g., Kagan and Jackson, 1994; Frankel, 1995; Albarello and Mucciarelli, 2002), including balanced combinations of deterministic and statistical elements to manage aleatory uncertainty (e.g., Cornell, 1968; McGuire, 1978). Outcomes of these approaches may present strong differences and this makes mandatory any evaluation of the respective heuristic value and effectiveness. Arrogating ageless Shakespeare’s words “Shall I compare thee to a summer’s day”, comparison of subjects with different nature is always difficult. Actually, effectiveness of any considered procedure (which includes both computational aspects and data used to feed the model) is uncertain (an “epistemic uncertainty”) and this is managed by associating to each procedure a degree of “belief” (again in the form of a probability). Being hazard estimates ultimately the combination of relevant uncertainties (and complementarily of lack of uncertainty about deterministic elements), both aleatory and epistemic uncertainties have to be considered and contribute to the estimate of the hazard curve.

While each PSHA procedure is on purpose determined to manage the relevant aleatory uncertainty via probabilistic modelling, assessment and management of epistemic uncertainty are more controversial topics. Given any $i$th PSHA model $H_i$, epistemic uncertainty can be defined as the probability $P(H_i)$ expressing the degree of belief in the effectiveness of that model. This formalization allows the management of epistemic uncertainty within a coherent frame (Albarello and D’Amico, 2014). A key aspect is the way to assess $P(H_i)$, i.e., scoring $H_i$. Two general approaches
exist on purpose. The first one relies on “ex-ante” expert evaluations of the actual reliability (in terms of internal robustness or coherency with current knowledge about underlying physical process) of the elements constituting the relevant PSHA procedure \( H_i \) (e.g., the geometry of considered seismic sources, the ground-motion attenuation relationship, etc.): these evaluations are combined in the frame of a “logic tree” (e.g., SSHAC, 1997). The second kind of approach is “ex-post” and considers a comparison of procedure outcomes (“forecasts”) with observations. Examples of empirical testing procedures have been provided by Mucciarelli et al. (2000), Albarello and D’Amico (2005, 2008), Beauval (2011) and recently by Tasan et al. (2014). Ex-ante and ex-post approaches can be seen as complementary in the frame of a Bayesian view (Albarello and D’Amico, 2014).

Only the ex-post approach only will be considered here to score on empirical basis a number of PSHA models available for the Italian area by the use of a simple procedure described in the next section. Then the data set of observations used for scoring the PSHA models will be described. The scoring test has been performed in the frame of the research agreement between the National Civil Defence Department (DPC) and the National Institute of Geophysics and Volcanology (INGV), namely the S2-2012 Project (https://sites.google.com/site/ingvdpc2012progettos2/home), after that observed maximum peak ground acceleration (PGA) values at a subset of available accelerometric stations were provided for a long time window (Pacor et al., 2013), and a repository of released PSHA results has been compiled too (Faccioli and Vanini, 2013). The data used in this study are given as supplementary material to stimulate alternative analyses and/or methodological comparisons.

2 Empirical scoring and testing

The bulk of any empirical scoring procedure is the evaluation of a probability \( L = P(E|H_i) \) that expresses the degree of belief (likelihood \( L \)) that a set of observed seismic occurrences \( E \) (“evidence”) will verify in the case that the PSHA computational
model $H_i$ provides a correct hazard estimate (Albarello and D’Amico, 2008). Given the model $H_i$ and the set of sites $E_{\Delta t^*}$ where ground shaking has been monitored during the control interval $\Delta t^*$ of duration equal to the hazard exposure time $\Delta t$, the model’s likelihood $L_i$ can be estimated from the control sample $E_{\Delta t^*}$. If the seismic occurrences $e_s$ are mutually independent (in the PSHA computational model considered) and if, over the duration of the control interval, a total of $N^*$ out of $S$ sites have experienced ground shaking above any threshold $g_0$, then we have

$$L_i = P(E|H_i, \Delta t) = \left\{ \prod_{s=1}^{N^*} P(e_s|H_i, \Delta t) \right\} \left\{ \prod_{s=N^*+1}^{S} [1 - P(e_s|H_i, \Delta t)] \right\},$$

where each value $P(e_s|H_i)$ is the hazard estimated (i.e., the exceedance probability for $g_0$) by the $i$th model at the $s$th site for the exposure time $\Delta t = \Delta t^*$. In the case that time stationarity is assumed in the relevant PSHA model, the overall duration of the exposure time is of concern only; this is not true when time-dependent PSHA models are considered. Of course one should account that several possible combinations sites/events may exist that result in the same configuration of the available evidence: all sites characterized in $H_i$ by the same exceedance probability are equivalent. It is worth to note, however, that likelihood value in Eq. (1) also depends on the number of sites considered and on the $P$ values of concern: this implies that comparison among different models by using respective likelihoods should be performed by considering the same values for $S$ and $P$. When this is not the case, any kind of “rescaling” is necessary. This rescaling could be performed by considering instead of Eq. (1) the “support” function $l$ that is the log-likelihood ratio as defined by Edwards (1972) in the form

$$l_i = \left\{ \sum_{s=1}^{N^*} \ln [P(e_s|H_i, \Delta t)] + \sum_{s=N^*+1}^{S} \ln [1 - P(e_s|H_i, \Delta t)] \right\} - r [P(e_s|H_i, \Delta t), S],$$

where $r$ is a reference log-likelihood value computed as in Appendix A as a function of $P(e_s|H_i)$ and $S$. 

5725
It can be seen (Kagan and Jackson, 1994) that probability distribution of the support \( l \) is nearly normal. This formulation allows using the reference value in the Appendix A and the relevant standard deviation to compute a studentized form of \( l \) as

\[
Z_i = \left| \frac{l_i}{\sigma_i(P,S)} \right|
\]

(3)

where the denominator is provided in Eq. (A5). In general, values of \( Z_i \) near 0 indicate best performing models while \( Z_i > 2 \) indicate models providing outcomes significantly different from observations. In this case, the model should be considered as “unreliable”. In this frame, the value \( Z_i \) can be considered as the “score” of the \( i \)th model: smaller is \( Z \) better is the computational model.

Other possibilities exist for testing any PSHA procedure against the evidence \( E \) (e.g., Schorlemmer and Gerstenberger, 2007; Schorlemmer et al., 2007). Counting is one of these procedures. In this case, a binary variable \( e_s(g_0) \) is defined which assumes the value of 1 if during the control interval \( \Delta t^* \) (which has the same extension as the hazard exposure time \( \Delta t \)) at least one earthquake occurred producing a ground motion in excess of \( g_0 \) at the \( s \)th site; otherwise \( e_s(g_0) = 0 \). We define the “control sample” \( E_{\Delta t^*} \) as the set of \( S \) realizations of the variable \( e_s(g_0) \) at \( S \) sites. The \( i \)th considered PSHA computational model \( H_i \) provides a probability \( P_{si} \) for the event \( e_s(g_0) = 1 \) given by

\[
P_{si} = P(e_s|H_i)
\]

(4)

where the dependence on \( g_0 \) and \( \Delta t \) is omitted to simplify the notation. Expectation \( \mu_{si} \) and standard deviation \( \sigma_{si} \) relative to the Bernoulli variable \( e_s \) result to be

\[
\mu_{si} = \mu(e_s|H_i) = P_{si}
\]

(5)

and

\[
\sigma_{si} = \sqrt{P_{si}(1 - P_{si})}
\]

(6)
respectively. The number $N^*$ of sites out of the $S$ sites considered for testing that experienced at least one earthquake during $\Delta t^*$ with ground shaking greater than $g_0$ is

$$N^* = \sum_{s=1}^{S} e_s.$$  

(7)

In terms of probabilistic forecasts provided by the $H_i$ PSHA computational model, $N^*$ is a random variate with expectation

$$\mu_i(N^*) = \sum_{s=1}^{S} \mu_i(e_s) = \sum_{s=1}^{S} P(e_s|H_i).$$  

(8)

In the hypothesis that $e_s$ are independent realizations of the stochastic process modelized in the PSHA computations, one can assume that

$$P_i(e_s|e_z) = P_i(e_s),$$  

(9)

where $e_s$ and $e_z$ are the realizations of the Bernoulli variable defined above at two generic $s$th and $z$th sites. In this case, the standard deviation of the random variable $N^*$ is

$$\sigma_i(N^*) = \sqrt{\sum_{s=1}^{S} P_{si}(1 - P_{si})}.$$  

(10)

When $S$ is relatively large, the Lyapunov variant of the Central Limit Theorem (e.g., Gnedenko, 1976) implies that

$$\text{prob} \left[ |N^* - \mu_i(N^*)| \geq 2\sigma_i(N^*) \right] \approx 0.05.$$  

(11)

Equation (11) allows us to evaluate whether a potential disagreement between the experimental value $N^*$ and the “forecast” $\mu_i(N^*)$ is statistically significant, thus making the $H_i$ PSHA computational model “not confirmed” by the set of $S$ observations.
3 Evidence: long lasting accelerometric recordings

Selection of observed data to be used for scoring PSHA models is a key element of the present analysis. The Italian accelerometric database ITACA (Luzi et al., 2008; Pacor et al., 2011), reporting records of accelerometric stations operating in Italy since 1974, has been considered for our purpose. A number of accelerometric sites continuously operating for long times were selected by considering available station books (courtesy of F. Pacor and R. Puglia from INGV-Milan and A. Gorini, Dept. of National Civil Defence, Fig. 1). Finally, we selected 71 stations operating during the time span 1979–2004. These time boundaries were chosen in order to maximize the number of stations contemporaneously active. These stations are unevenly distributed all over the Italian area (Fig. 2) and are located both on rock or different kind of soils classified by following the National Seismic Code NTC08 (NTC, 2008). In particular: 25 stations are located on A type soil ($V_{s30} > 800 \text{ m s}^{-1}$, where $V_{s30}$ is the average shear-wave velocity in the uppermost 30 m of underground), 30 on B type soil ($V_{s30}$ in the range $360–800 \text{ m s}^{-1}$), 13 on C type soil ($V_{s30}$ in the range $180–360 \text{ m s}^{-1}$), 1 on D type soil ($V_{s30} < 180 \text{ m s}^{-1}$) and 2 on E type soil (i.e., C and D type soils, but with seismic bedrock at a depth in the range 3–20 m from the surface). Most of the stations (52) lay on a flat outcrop (T1 type topography by following the NTC08 code), 14 on smooth morphology (T2 type, i.e., on a surface dipping in the range 15–30°) and 5 on rough topography (T3 type, i.e. surface dipping more than 30°). Note that the site classification based on $V_{s30}$ mimics the one adopted by the EuroCode EC8. One can see that most of stations lay on stratigraphic/geomorphological configurations different from the “reference” one (i.e., flat outcrop of a rigid bedrock with $V_{s30} > 800 \text{ m s}^{-1}$) generally considered for seismic hazard estimates. This implies that, in order to compare hazard outcomes with observations, some correction terms should be considered to “reduce” observed accelerometric data to “reference” values. In the present study, such a correction term has been assumed equal to the amplification factor stated by the NTC08 regulation code, which includes both stratigraphic and topographic effects: it assumes values
depending on the soil type and topographic class at the site, but also on the hazard estimated on the reference outcrop. The relevant correction coefficients computed at the 71 accelerometric stations considered for testing (Fig. 2) are reported in the Supplement File A; details are given in Sect. 3.2.3 of NTC08. These coefficients represent a first approximation to site-specific hazard, coherent with the common practice for buildings that do not require specific studies; they have been used to correct maximum PGA values observed on horizontal components in the time interval 1979–2004.

About available recordings, on 71 stations, 12 have no records at all, for this 25 year long period. In these cases, we assumed the sensitivity threshold of the early deployed accelerometers (0.01 g, i.e. 9.8 cm s\(^{-2}\)) as the maximum “observed” value. We checked possible problems with data completeness that nevertheless we acknowledge they cannot be properly fixed. For this purpose, PGA values expected at all the sites due to the occurrence of nearby earthquakes have been computed (synthetic “observations”), on the basis of epicentral information (CPTI11 earthquake catalogue, Rovida et al., 2011) and the Ground Motion Prediction Equation (GMPE) ITA10 by Bindi et al. (2011). In particular, synthetic PGA values have been considered potential observations at the relevant site if they exceed the sensitivity trigger threshold. Synthetic PGA values obtained at ALT (Auletta, Salerno) and PTL (Pietralunga, Perugia) stations are plotted in Fig. 3, and compared with effective recorded data: note that even if some data are possibly missing (blue circles in Fig. 1 correspond to values above the sensitivity threshold of the accelerometric network), on average, the maximum observed values in the time window considered for the analysis is coherent with the expectations. We estimated that missing maximum PGA should have occurred on less than 5 % of stations, thus not affecting the results obtained. This analysis also suggested that completeness problems should be critical if testing were performed on the whole sequence of observations instead of only maximum observed PGA.

On the subset of selected stations, the observed maximum PGA values span from about 1 cm s\(^{-2}\) for \(M < 4.5\) earthquakes at about 40–60 km distance (e.g. at ARI Ariano
Irpino, Avellino) to the 490 cm s\(^{-2}\) at NCR (Nocera Umbra, Perugia) for the main shock of the long lasting Colfiorito Umbria-Marche sequence (26 September 1997, \(M_w = 6.0\) at 11 km distance). Station code, coordinates, site conditions and the maximum registered PGA values are given in the Supplement File A.

4 Models: PSHA in Italy

Italy has three maps, or groups of maps, of PSHA which have been turned into regulation acts, therefore having an impact on society: as shown in Fig. 4, these maps were released in 1979, 1996–1999 and 2004, and they were adopted by laws with some delays from their release, always after deadly earthquakes.

The 1979 map (Gruppo di Lavoro Scuotibilità, 1979) is expressed in terms of macroseismic intensity and belongs to the so-called generation of “historical probabilism”: essentially an earthquake catalogue, a given relationship for attenuating intensity (without uncertainties) and Gumbel type I statistics of shakings at the sites (Gumbel, 1958). The map was translated into seismic categories with given prescription rules after the 1980 Irpinia \(M_w = 6.9\) earthquake (about 3000 casualties), and municipalities entered into regulation by a series of acts, from 1981 to 1984 (Petrini et al., 1980; Slejko, 1993).

The other maps belong conceptually to the generation of the “seismotectonic probabilism” (Muir Wood, 1993). The second group of maps was released in 1996 (Slejko et al., 1998), and refined in 1998–1999 (Albarello et al., 2000): they are maps in terms of macroseismic intensity and PGA (for additional details refer to Table 1, Project frame GNDT). The refinements, mostly due to changes in seismicity rates interpolation and GMPEs, came after the long and highly damaging Umbria-Marche sequence (known as Colfiorito sequence, 11 casualties) in 1997–1998. These maps were the basis of the revision of seismic code approved in 2003, after the collapse of a school in San Giuliano of Puglia (2002 Molise earthquake) that killed 27 pupils.
and their teacher. The same code (Ord. 3274/03) stated the rules for preparing a new reference national hazard map.

The third and ultimate map was released in 2004 (Gruppo di Lavoro MPS, 2004) and it was provided by supplementary elaborations (maps for PGA and spectral accelerations for several return periods) in the following years (see MPS04 and S1 2004–2006 in Table 1, Montaldo et al., 2007; Stucchi et al., 2011); it became the official reference document for seismic reclassification in 2006 (Ord. 3519/06), and later it was fully embedded together with the supplementary elaborations in the building code NTC08 (NTC, 2008). After some years of partial application, compulsory rules for its adoption have been stated after the 2009 L'Aquila $M_w = 6.3$ earthquake (about 300 deaths).

Both the 1996–1999 and 2004 results are based on the Cornell’s approach (Cornell, 1968) to PSHA and area source models; these models entered into the European hazard maps too, respectively ZS4 (Meletti et al., 2000) in GSHAP (Giardini et al., 1999) and ESC-Sesame (Jimenez et al., 2001) elaborations, and ZS9 (Meletti et al., 2008) with some modifications in the SHARE model (Giardini et al., 2013). After 2004, other maps and prototypal elaborations have been realized and published, referring to similar datasets but using alternative methodologies (e.g. seismic site histories, D’Amico and Albarello, 2008; fault-based time dependency, Peruzza, 2006), and updating the model components (databases, GMPEs or seismogenic sources), mostly at the regional scale (e.g. Pace et al., 2006; Akinci et al., 2009).

In the frame of the S2-2012 annual project funded under the decennial agreement of DPC and INGV, a research team (Politecnico of Milan, Faccioli and Vanini, 2013) acted to collect and to make available the PSHA results to public accessibility; selected data are stored after the compilation of an online form, by the upload of electronic files or worksheets; data, summary list and a short report are freely available at https://sites.google.com/site/ingvdpc2012progettos2/deliverables/d1-1.

The PSHA outcomes are mostly provided in terms of PGA values and it is the only shaking parameter considered for scoring by now. Models are released by referring to
5732 one or a limited number of return periods (i.e. thresholds of exceedance probability in given exposure time). Figure 5 shows the comparison of expected PGA values for the models having approximately the same return period (i.e., 475 years, or 10% probability of exceedance in 50 years) at two localities, in Northern (Modena) and Southern Italy (Potenza); remarkably, the Po Plain and Southern Apennines have been set as priority regions by the first year of DPC-INGV research agreement. As time-dependent models (blue labels in Fig. 5) refer to origin time set up in 2010, they cannot be used in our retrospective testing. In Tables 1 and 2 the list of selected models and their references are given; a synoptic graphical representation of results referred to the whole Italy is given in Fig. 6. Individual pictures are given in the Supplement File B.

The SHARE model (Giardini et al., 2013, represented in Fig. 6 by ID 9 frame) has not been stored in the repository of S2-2012 project. SHARE results have been progressively released since 2013, and are available at the SHARE Portal http://www.efehr.org:8080/jetspeed/portal/hazard.psml.

All the PGA values used for the scoring test are given in an excel file (Supplement File A). The values refer to the computation node nearest to the selected accelerometric sites previously described. These data are provided for stimulating additional testing by the scientific community.

5 Results

In order to compare observations and predictions provided by each PSHA model, the time span covered by both should be the same. In general, PSHA outcomes have the form of a PGA value \( g_0 \) characterized by a fixed exceedance probability in a time span of duration \( \Delta t \) (the exposure time) at the \( s \)th site (see above). Actually, being all the considered PSHA models based on the assumption that the seismogenic process is Poissonian, the following relation holds

\[
P_{si} = P\left( e_s | H_i, \Delta t \right) = 1 - e^{-\lambda_{si}(g_0)\Delta t},
\]
where $\lambda_{si}(g_0)$ is the annual rate of exceedance for the threshold $g_0$ and $P$ is the exceedance probability at the $s$th site, for the relevant exposure time $\Delta t$ and acceleration threshold $g_0$ if the $i$th model is considered.

In the case we are considering, $\Delta t$ lasts 25 years (i.e., the time span contemporary covered by 71 accelerometric observations, see above). However, most of the PSHA models provide hazard values for a different exposure time $\Delta t'$ (in general 30 or 50 years), i.e. $P(e_s|H_i, \Delta t)$. Thus, in order to apply Eqs. (1)–(11), some conversion tool is necessary to compare hazard estimates and observations. This conversion takes advantage of the stationary Poissonian character of seismic occurrences assumed by considered PSHA models. In this case, in fact, one has that

$$P'_{si} = P(e_s|H_i, \Delta t') = 1 - e^{-\frac{\ln[1-P(e_s|H_i, \Delta t)]}{\Delta t}} \Delta t'. \quad (13)$$

The above formula can be used to compute the exceedance probability relative to the acceleration threshold $g_0$ for a given exposure time ($\Delta t'$) when the exceedance probability is supplied for another exposure time ($\Delta t$). The value $P'_{si}$ is then considered for testing.

Since some models also provide $g_0$ values corresponding to different exceedance probabilities, they were scored by considering each realization as an independent “forecast”. In general, since in the same model lower exceedance probabilities correspond to longer return times and to higher $g_0$ values, different scoring can be attributed to different parts of the hazard curve.

Thus, for each PSHA model, a set of $P'_{si}$ values is computed for the sites considered for testing, for $\Delta t = 25$ years. Consequently, the binary variable $e_S(g_0)$ is computed: equal to 1 when at the $s$th site the value $g_0$ was exceeded in the time interval 1979–2004, and to 0 otherwise.

On this basis, the score $Z$ (Eq. 3) was computed for each PSHA model. This value is considered as the empirical score of the model: as lower is $Z$ as most effective results the relevant model. The overall number of exceedances (Eq. 6) was also compared with the values expected in the relevant PSHA model (Eq. 8): if this difference exceeds
two times the relevant standard deviation (Eq. 10), the PSHA model is considered to be not compatible with observations (Eq. 11).

5.1 Scoring models at the national scale

Except ID 7 and 8, all models have nation-wide coverage, thus allowing the scoring on the full set of 71 selected accelerometric stations. Some models have been given for different return periods; they give a final set of 12 realizations from 7 models. Comparison of expected vs. observed occurrences, are shown in Fig. 7; models are sorted accordingly to the relevant return period.

Despite of the fact that some models tend to slightly underestimate the observed number of exceedances, in all the cases these discrepancies are not significant by following Eq. (11). This, however, does not mean that all the models equally fit observations. In fact, when the score factor $Z$ is considered (Fig. 7b), one can see that significant differences exist in the performances of the considered models at the different return times.

The best performing model is the 1996 GNDT model at intermediate return time (ID 1, RT = 284 years) followed by the MPS04-like area-based source model using Cauzzi and Faccioli (2008) GMPE (ID 5) for 984 year return period; notably, models obtained under different theoretical assumptions or computational choices behave nearly the same: as an example one can see the results provided by the ID6 model (smoothed seismicity approach by Akinci et al., 2010), the ID5 one (the one provided by Meletti et al., 2009 with the standard Cornell-McGuire approach, by considering the same single Ground Motion Prediction Equation used in ID5), and ID9 (produced in the frame of the SHARE project). On the other hand, same model performs in different ways at different return times of 94 years: e.g. see the ID 1 best performing at a return time of 284 years and providing a worse performance at a shorter return time (Fig. 7b). As models that explore different parts of the hazard curve have controversial scoring (i.e., different scores for different return times), it is not easy to identify a single “best” performing model.
5.2 Scoring models at the regional scale

The same test has been performed at the regional scale, for the two selected priority regions of Southern Apennines and Po Plain, for which ad hoc regional PSH estimates have been released during the S2-2012 project (Fig. 8). Thus, the same subset of accelerometric stations on national and regional PSHA models have been manually selected and controlled to exclude sites not considered for the relevant hazard estimate.

In Southern Apennines, all the models provide results that are compatible with observations that refer to 21 sites (Fig. 9a). When the score factor $Z$ is considered (Fig. 9b) the best performing models at about 480 years are the ones derived from smoothed seismicity model (ID 6), and the MPS04 model (ID 3). Similarly, the best performing model at the shortest times of 30 years is the one provided by Akinci et al., 2013 (ID 7). Thus, some PSHA evaluations seem to be more adequate to represent the observed shakings on that Southern region. Note the scoring positions of the long-term predictions of the MPS04 model (ID 4) T2475 now aligned with the ones provided by Akinci et al., 2013 (ID 7).

The same analysis performed for the Po Plain area uses only 12 stations; again it indicates that all except the ID 8 computational models provide results that are compatible with observations (Fig. 9c). In this case the scoring indicates several best performing (i.e., MPS04, ID 4, at RT = 2475 years; smoothed seismicity and MPS04-like models ID 6 and ID 5, at RT = 984 years; SHARE, MPS04-like and GNDT 1996 results, ID 9, 6 and 1, at RT = 475 years). Note also the underestimate of results released in the frame of the S2-2012 project (ID 8) gives $Z$ values higher than 3.

In order to better visualize the impact of scoring in hazard estimates, the PGA values provided by the models at different return times are labelled with their relevant $Z$ values (Fig. 10), for the cities of Potenza (Southern Apennines) and Modena (Po Plain). We believe this kind of analyses should help in defining a comprehensive PSHA, no more
based on logic tree procedures, or expert elicitation, but on the strength of observation data.

6 Conclusions

Nowadays, the scientific community has been looking for a coherent, formal and robust procedure for testing probabilistic seismic hazard estimates. As for weather forecast, the availability of observational data of the last years is not comparable with the previous decades and probably it is going to produce faster changes of approaches to seismic hazard than ever before. An extensive empirical test of seismic hazard estimates in Italy has been carried out by evaluating quantitatively their performances. In particular, an empirical scoring procedure has been applied to a number of PSHA computational models in the frame of the DPC-INGV S2-2012 seismological research project; many of the considered models provided outcomes that were included in the Italian Seismic Regulation Code and this fact strengthens the importance of evaluating their actual reliability. Twelve realizations from seven time-independent PSHA models available at the national scale plus six maps from two models at the regional scale have been collected; a set of accelerometric stations continuously operating in the time interval 1979–2004 has been analysed, using the maximum observed PGA at each station like the testing parameter. Site-specific corrections were applied for PGA values at accelerometric stations where possible amplification effects are expected due to the local soil conditions. These correction coefficients are the ones set up in the Italian seismic code (NTC08). The scoring results obtained suggest some preliminary conclusions as it follows:

1. Nearly all the considered models provide outcomes that are compatible with available observations;

2. The most recent models are not necessarily the best performing ones;
3. None of the considered models can be considered as the best performing at all the considered return times;

4. Testing done on sub-regions reveals different features with respect to the national scale, but the reasons should be investigated with other cases;

5. GMPEs could hamper the differences of alternative source modeling, but this is not a general rule.

This study focuses some open questions to be addressed in the future, like:

1. Site-specific PSHA or calibrated amplification functions at the accelerometric stations are necessary to avoid the over-simplification here adopted; they may play a key role in scoring results: specific activities have been planned on these subject in the prosecution of S2 Project started in 2014 (see Tasks 2 and 4, at https://sites.google.com/site/ingvdpc2014progettos2/);

2. Completeness of accelerometric records relative to accelerometric sites is a critical aspect for validation; we overcome the problems by considering the maximum PGA in a quite long time period, but further analyses are needed to fully exploit the observations provided by the actual Italian databases.

This study shows that the likelihood estimates accompanied by other testing procedures are able to provide useful indications about performances of competing models and could represent a basic tool for driving new researches devoted to a best practice for hazard assessment.

Appendix

Being $r$ the reference log-likelihood computed by considering a set of $S$ observations relative to sites characterized by an exceedance probability $P$, one has

$$r = N^* \ln(P) + (S - N^*) \ln(1 - P), \quad (A1)$$
where $N^*$ is the number of sites where the ground-motion threshold characterized by the exceedance probability $P$ has been exceeded during the control interval considered. One has has 

\[ r = N^* (\ln(P) - \ln(1 - P)) + S \ln(1 - P). \]  

(A2)

Sampling properties of $r$ only depend on the random variate $N^*$ (being all the other parameters fixed). This variable $N^*$ is the sum of $S$ realization of a binomial variable characterized by a probability $P$ of occurrence. The expected value of $N^*$ is then $SP$, while its variance is $P(1 - P)$.

Thus, the expected value $\mu(r)$ of $r$ is

\[ \mu(r) = SP \ln(P) + S(1 - P) \ln(1 - P) = S[P \ln(P) + (1 - P) \ln(1 - P)]. \]  

(A3)

One can see that $\mu(r)$ monotonically increases with $S$, is a symmetric function of $P$ with a minimum for $P = 0.5$ and values 0 for $P = 1$ and $P = 0$, respectively. The relevant sampling variance is

\[ \sigma^2(r) = SP (1 - P)[\ln(P) - \ln(1 - P)]^2 \]  

(A4)

with a standard deviation equal to

\[ \sigma(r) = [\ln(P) - \ln(1 - P)] \sqrt{SP(1 - P)}. \]  

(A5)

The Supplement related to this article is available online at doi:10.5194/nhessd-2-5721-2014-supplement.

Acknowledgements. This study has benefited from funding provided by the Italian Presidenza del Consiglio dei Ministri – Dipartimento della Protezione Civile (DPC), Project S2-2012. This paper does not necessarily represent DPC official opinion and policies.
References


Cauzzi, C. and Faccioli, E.: Broadband (0.05 to 20 s) prediction of displacement response spectra based on worldwide digital records, J. Seismol., 12, 453–475, 2008.


Table 1. List of selected PSHA models.

<table>
<thead>
<tr>
<th>Project frame</th>
<th>Model ID</th>
<th>Filename of PSHA results</th>
<th>Description*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PS4_1996_PGA_10–30-50y</td>
<td>First PSH map for Italy using seismotectonic probabilism. Catalogue of declustered events till 1980 (NT4.1), area sources (ZS4), GMPE on undifferentiated soil condition (Amb95). PGA values computed on a 0.1° grid. Exceedance prob. of 10 % in 10, 30 and 50 years</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SSN-GNDT99_PGA</td>
<td>Consensus map refining the previous model; logic tree for GMPE (Amb96, SP96). PGA given on irregular grid (communes), exceedance probability of 10 % in 50 years, 50 percentile</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Appennino_Meridionale_MPS04_ag_002, Pianura_Padana_MPS04_ag_002</td>
<td>Italian PSH map developed on rules stated by law (Ord. 3274/03). PGA values computed on a 0.02° step grid. Catalogue of declustered events till 2002 (CPTI04), area sources (ZS9), logic tree including alternative GMPEs (Amb96, SP96, REG.A, REG.B). Exceedance prob. of 10 % in 50 years, percentile 16, 50 and 84. Data points collected by S2-2012 Project refer only to the priority areas of Po Plain and Southern Apennines: the data sampled on a 0.05° grid on the whole country here used are available at <a href="http://zonesismiche.mi.ingv.it/">http://zonesismiche.mi.ingv.it/</a></td>
<td></td>
</tr>
<tr>
<td>MPS04 – S1 2004–2006</td>
<td>S1_2004-2006_SA_0.0s_D2_39-2</td>
<td>Same approach and input data of previous model 3, MPS04, additional probabilities of exceedance in 50 years have been computed during the project S1 (2004–2006): 2 and 39 % are selected in this analysis. PGA values on a 0.05° step grid for all Italy. 16, 50 and 84 percentile</td>
<td></td>
</tr>
</tbody>
</table>
**Table 1. Continued.**

<table>
<thead>
<tr>
<th>Project frame time</th>
<th>Model ID</th>
<th>Filename of PSHA results</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>S2 2008–2010</td>
<td>5</td>
<td>S2_2008-2010_SA_0.0s_MPS04_D2.1</td>
<td>MPS04-like model using different software (CRISIS vs. SEISRISK III used by mod.1–4) and GMPE (CF08, no logic tree). Gridded seismicity based on ZS9. PGA (SA at $T = 0s$) values on a 0.1° step grid for the whole of Italy, for 3 and 6% exceedance probability in 30 years.</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>S2_2008–2010_SA_0.0s_HAZGRID_D2.2</td>
<td>Zone free smoothed seismicity, based on CPT04 and instrumental datasets: same GMPE, sampling and return period of the previous model 5.</td>
</tr>
<tr>
<td>S2 2012</td>
<td>7</td>
<td>S2_2012-2013_SA_0-1_TimeIndep_AppMerid_D5.2</td>
<td>PSH estimates developed by the S2-2012 project, for priority area Southern Italy. Combination of smoothed seismicity approach (CPT111, instrumental datasets) and characteristic model on faults (DISS3.1.1), under Poissonian assumption, GMPE logic tree (AB10, BA08, ITA10, CF08). Spectral acceleration at 0 (PGA) and 1 s. Probability of exceedance of 2, 5, 10 and 81% in 50 years.</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>S2_2012-2013_SA_0-2_PianuraPadana_D4.1</td>
<td>PSH estimates developed by the S2-2012 project, for priority area Po Plain. It derives from model 5 for several spectral accelerations (0–2 s). Rock and site specific conditions, implemented by regulation amplification factors on 1:100 000 scale soil map.</td>
</tr>
<tr>
<td>SHARE</td>
<td>9</td>
<td>Latest PSHA for Europe, first regional project in GEM initiative (<a href="http://www.globalquakemodel.org/">http://www.globalquakemodel.org/</a>). New European historical and instrumental catalogues, full logic-tree of GMPE set on tectonic regionalization, combination of area sources, distributed seismicity and larger events concentrated on faults, with new maximum magnitude scheme for the whole region. Results progressively released via the SHARE portal.</td>
<td></td>
</tr>
</tbody>
</table>

* For the explanation of acronyms refer to the references listed in Table 2.
**Table 2. List of references for the selected PSHA models.**

<table>
<thead>
<tr>
<th>Model ID</th>
<th>References</th>
</tr>
</thead>
</table>
| 1        | Slejko et al. (1998)  
NT4.1 = Camassi and Stucchi (1997)  
ZS4 = Meletti et al. (2000)  
Amb95 = Ambraseys (1995) |
| 2        | Albarello et al. (2000)  
Amb96 = Ambraseys et al. (1996)  
SP96 = Sabetta and Pugliese (1996) |
| 3        | Gruppo di Lavoro MPS (2004); Stucchi et al. (2011)  
CPTI04 = Gruppo di Lavoro CPTI (2004)  
ZS9 = Meletti et al. (2008)  
REG.A, REG.B = two combinations of regionalized GMPEs (i.e., Malagnini et al., 2000, 2002; Morasca et al., 2006; De Natale et al., 1988; Patanè et al., 1994). |
| 4        | Meletti and Montaldo (2007) |
| 5        | Meletti et al. (2009)  
CRISIS = Ordaz et al. (2013)  
SEISRISKIII = Bender and Perkins (1987)  
CF08 = Cauzzi and Faccioli (2008) |
| 6        | Akinci (2010) |
| 7        | Akinci (2013) and annexed files at  
https://sites.google.com/site/ingvdpc2012progettos2/deliverables/d5_2  
CPTI11 = Rovida et al. (2011)  
DISS3.1.1 = DISS Working Group (2010)  
AB10 = Akkar and Bommer (2010)  
BA08 = Boore and Atkinson (2008)  
ITA10 = Bindi et al. (2011) |
| 8        | Task 4 Working Group (2013) and annexed files at  
https://sites.google.com/site/ingvdpc2012progettos2/deliverables/d4-1 |
| 9        | Giardini et al. (2013)  
http://www.efehr.org:8080/jetspeed/portal/hazard.psml  
Data downloaded in September 2013 |
Figure 1. ON-OFF status for accelerometric stations declared to be continuously operating for at least 30 years, as reported on ITACA V1.1 database (Pacor at al., 2011, 2013). Blue circles show potential triggering conditions, computed as (mean PGA + 1 SD) > 0.01g, using CPTI11 earthquake catalogue (Rovida et al., 2011) and ITA10 GMPE (Bindi et al., 2011); black plusses are the effective recordings available. Data acquired on continuous-mode recording in 2009–2011 have not been considered for the existing time gap between the analogical and new digital equipments.
**Figure 2.** Location of the accelerometric stations considered for empirical testing. Dotted pins refer to stations on A or A* (* for hypothesized conditions) type soil in Eurocode classification, pins colour represents a simplified amplification factor for PGA (NTC, 2008) used to accomplish stratigraphic and topographic site response: green = 1, yellow = 1.2–1.35, violet = 1.5. See Appendix A and Supplement for details. The stations in small turquoise pins, in the list of stations in Fig. 1, have been discarded in this analysis as they do not satisfy the ON status in the 1979–2004 time window.
Figure 3. Observed and synthetic PGA values at two stations in the time span 1975–2005. (a) observed (b/w symbols, values as reported in ITACA 1.1. database and Pacor et al., 2013) and computed PGA values at ALT (Auletta, Salerno) station, in Southern Apennines; (b) PTL (Pietralunga, Perugia) station in Central Italy. Mean ± 1 standard deviation computed values, represented respectively by green circle, red and blue triangles, have been obtained by ITA10 GMPE applied to CPTI11 earthquake catalogue; synthetic PGA are plotted if MeanPGA+1sd is greater than 0.01 g, as in Fig. 1.
Figure 4. Timeline of PSHA maps in Italy relevant for regulation; orange symbols represent deadly earthquakes occurred in the last 40 years.
Figure 5. Comparison at two sites of expected PGA values (with 10% probability of exceedance in 50 years) from collected PSHA models (redrawn from Faccioli and Vanini, 2013). Modena is located in the Po Plain, at about 20–30 km distance from the main earthquakes of the 2012 Emilia sequence; Potenza is at about 90 km distance from the recursive sequences that affected the border of Calabria and Basilicata Regions, in Southern Apennines, since 2011. Time-dependent models listed in this graph (labels in blue) have not been used in this analysis.
Figure 6. Synoptic view of PSHA maps collected by S2-2012 Project at the national scale. Model ID refers to Table 1, the vertical axis shows approximately the return period the elaborations refer to; on x axis, a rough timeline of results release (from 1996 till 2013). The colour scale is automatically adjusted on values (in particular the SHARE model have extremely high values out of the represented are); all the maps and graphic details are given in Supplement File B.
Figure 7. Results of scoring for national based PSHA models. **(a)** Observed vs. computed number of stations exceeding the predicted PGA values $g_0$. Results are sorted according to the return period and subordinately on model IDs. **(b)** Final scores, the $y$ axis represents the absolute value of $Z$ score as given in Eq. (3); the lower the best. Model IDs are given in Table 1; $T_{year}$ indicates the mean return time the elaboration refers to.
Figure 8. Synoptic view of regional PSHA maps collected by S2-2012 project. Model IDs refer to Table 1; the vertical axis shows approximately the return period the elaborations refer to. The colour scale is automatically adjusted on values; all the maps and graphic details are given in Supplement File B.
Figure 9. Results of scoring for sub-regions in Italy, as in Fig. 7: (a) and (b) Southern Apennines; (c) and (d) Po Plain.
Figure 10. PSHA and scores at two selected sites. (a) Modena, in the Po Plain; (b) Potenza in the Southern Apennines. Red dots and b/w diamonds represent the national and regional models, black and blue labels respectively the absolute $Z$ values (Eq. 3) on the whole set and regional subsets of stations.