Bulletin of the Seismological Society of America, Vol. 102, No. 3, pp. 1195–1213, June 2012, doi: 10.1785/0120110164

A Ten-Year Earthquake Occurrence Model for Italy

by W. Marzocchi, A. Amato, A. Akinci, C. Chiarabba, A. M. Lombardi, D. Pantosti, and E. Boschi

Abstract The recent M_w 6.3 destructive L'Aquila earthquake has further stimulated the improvement of the Italian operational earthquake forecasting capability at different time intervals. Here, we describe a medium-term (10-year) forecast model for $M_w \geq 5.5$ earthquakes in Italy that aims at opening new possibilities for risk mitigation purposes. While a longer forecast yielded by the national seismic-hazard map is the primary component in establishing the building code, a medium-term earthquake forecast model may be useful to prioritize additional risk mitigation strategies such as the retrofitting of vulnerable structures. In particular, we have developed an earthquake occurrence model for a 10-year forecast that consists of a weighted average of time-independent and different types of available time-dependent models, based on seismotectonic zonations and regular grids. The inclusion of time-dependent models marks a difference with the earthquake occurrence model of the national seismic-hazard map, and it is motivated by the fact that, at the 10-year scale, the contribution of time-dependency in the earthquake occurrence process may play a major role. The models are assembled through a simple averaging scheme whereby each model is weighted through the results of a retrospective testing phase similar to the ones carried out in the framework of the Collaboratory for the Study of Earthquake Predictability. In this way, the most hazardous Italian areas in the next ten years will arise from a combination of distinct models that place more emphasis on different aspects of the earthquake occurrence process, such as earthquake clustering, historical seismic rate, and the presence of delayed faults capable of large events. Finally, we report new challenges and possible developments for future updating of the model.

Introduction

The seismic-hazard map is the primary and sole seismological piece of information used for risk mitigation in many countries, being the foundation of the building code definition (e.g., Gruppo di Lavoro MPS, 2004). Conventionally, seismic-hazard maps show the expected ground motion with a certain probability in a time horizon shown in decades (usually 50 years). However, the occurrence of the recent L'Aquila earthquake (6 April 6 2009, M_w 6.3) has brought new challenges for seismologists; in particular, it has been argued that the development of short-term (days to weeks) and medium-term (months to a few years) hazard models may provide additional information to improve and direct seismic risk mitigation in different ways and at different time intervals (Jordan and Jones, 2010; van Stiphout *et al.*, 2010; Woo, 2010; Jordan *et al.*, 2011). These new hazard models and the derived mitigation actions are meant to increase risk reduction strategies and not replace the primary information that is a sound assessment of the building code based on a long-term seismic-hazard map.

In principle, a conventional seismic-hazard model may be extrapolated at shorter time intervals (days to a few years; e.g., Mulargia, 2010), but we argue that this extrapolation could lead to inaccurate results because the time evolution of the seismicity may not be taken properly into account. While seismic-hazard maps are usually based on time-independent earthquake occurrence (henceforth, EO) models (e.g., Gruppo di Lavoro MPS, 2004), at short-time intervals the EO process is certainly time dependent, showing a marked time and spatial clustering (e.g., Ogata, 1988; Reasenberg and Jones 1989, Kagan and Jackson, 2000; Console *et al.*, 2003; Gerstenberger *et al.*, 2005; Lombardi and Marzocchi, 2010a). At medium-term intervals, seismicity still shows a significant time-space clustering (e.g., Kagan and Jackson, 1991, 2000; Parsons, 2002; Faenza *et al.*, 2003, 2008; Lombardi and Marzocchi, 2007; Marzocchi and Lombardi, 2008), and other types of time dependency may become important. Hence, we argue that each forecasting time interval requires a specific EO model able to describe the most relevant time-dependent processes at that specific time scale. These differences notwithstanding, the EO models on different forecasting time intervals must be consistent (Jordan *et al.*, 2011).

To date, many types of time-independent and time-dependent EO models for medium-term forecasts have been proposed. In order to use one, or a combination thereof, first we need to evaluate their reliability and skill, as intended by Jordan *et al.* (2011). A model is defined as reliable if its forecasts are accurate, that is, if the observations are compatible with the forecasts. Among reliable models, the more precise its forecasts, the higher the skill; in other words, the skill is a measure of the model's predictability. The evaluation of the reliability and skill of EO models is the primary aim of the experiments carried out in the framework of the Collaboratory for the Study of Earthquake Predictability (henceforth, CSEP; Jordan, 2006; Zechar, Schorlemmer *et al.*, 2010). Basically, these experiments compare the real seismicity with forecasts produced by EO models in different regions of the world (Zechar, Schorlemmer *et al.*, 2010). The recent CSEP experiment, begun in Italy in August 2009, counts 18 different models tested (Marzocchi *et al.*, 2010), with initial results to be available in 2014.

Here, we propose a first attempt to define the best medium-term model for Italy. While different procedures are possible (see Marzocchi and Zechar, 2011), we are presenting a strategy that basically consists of two steps: first, we will assess reliability and skill of all the available models through CSEP-type tests on retrospective experiments. Second, we will use these results to merge the EO models objectively, in order to define the best EO model. In the last section of the paper (A Closer Look at the Best EO Model), we describe the best EO map, region by region, emphasizing the most critical points for each area, while comparing the results with the present knowledge of active fault activity and Global Positioning System (GPS)–based deformation patterns not yet used in the EO maps.

The intrinsic scientific interest notwithstanding, we remark that this attempt has a clear practical objective, because this map will be used to mitigate the medium-term seismic risk in Italy. In fact, at the beginning of 2010, the Italian government decided to provide funding for reducing seismic risk based on medium-term time scales (*Interventi per la prevenzione del rischio sismico*; see Data and Resources section). The planned government mitigation actions consist of retrofitting buildings and vulnerable structures over the next few years, in those areas that contain the highest seismic risk, while the Istituto Nazionale di Geofisica e Vulcanologia (INGV) has been appointed to prepare a national EO model on this time interval.

The Selected Models

Scientific literature reports many different kinds of EO models, based on different assumptions and scientific components. The range of models reflects the large epistemic uncertainty still existing in the modeling of the earthquake process (e.g., Marzocchi *et al.*, 2010). We argue that any so-called best EO model must properly take into account such a large variability. While the number of models considered here is limited due to some practical constraints, we aim at presenting an approach that may be applied elsewhere and, in future applications, when a larger number of models become available. Here we choose the set of models according to pragmatic and mandatory rules. In order to be taken into consideration, an EO model must produce forecasts for the whole Italian territory because the risk reduction program should act on a national scale. Moreover, the model must be in a specific format in order to facilitate the comparison with other models and eventually their averaging. For this purpose, we have taken advantage of the CSEP experiment that started on 1 August 2009 (Marzocchi *et al.*, 2010) and the S2 project (named "Development of a Dynamical Model for Seismic Hazard Assessment at National Scale") funded in the framework of the agreement for the years 2007–2009 between INGV and the Dipartimento della Protezione Civile. One of the objectives of the S2 project was to develop experiments applied to available EO models in order to quantify their reliability and skill (see The Retrospective Tests of the Models: Reliability and Skill). The format of the EO forecasts in CSEP and S2 initiatives is roughly the same, and it consists of providing a number of events/probability in a specific spatial grid (Fig. 1) that has been used for the CSEP experiment in Italy. A last requirement is that each EO model candidate must be based on reasonable physical assumptions.

This first selection has identified a set of models that satisfy these requirements. A full description of these models can be found in the cited references. Here, we briefly describe only the main features of two time-independent (Akinci, 2010; Gruppo di Lavoro MPS, 2004) and four time-dependent (Lombardi and Marzocchi, 2010b; Falcone *et al.*, 2010, Akinci *et al.*, 2009) models considered in this study.

Time-Independent Models

HAZGRIDX (Akinci, 2010). The model is based on the assumption that future earthquakes will occur near locations of historical earthquakes; it does not take into account any information from tectonic, geological, or geodetic data. It is built on a spatially grid-based format using two catalogs: the parametric catalog of Italian earthquakes (Catalogo Parametrico dei Terremoti Italiani [CPTI], CPTI04; Gruppo di Lavoro CPTI, 2004), which contains the larger earthquakes (M_w 7.0 and larger) since 1100, and the catalog of Italian seismicity (Catalogo della Sismicità Italiana, CSI 1.1), which contains smaller earthquakes down to a local magnitude of M_L 1.0, with a maximum of M_L 5.9 over the past 22 years (1981–2003). The model assumes that earthquakes follow the Gutenberg–Richter law, with a constant *b*-value. Seismicity rates in the model are determined by counting earthquakes with magnitude $M_w \geq 5.5$ in each cell, with dimensions 0.1° longitude × 0.1° latitude of a grid adopted by the CSEP experiment in Italy (Fig. 1; see also Schorlemmer, Christophersen, *et al.* 2010). The gridded 10^a values are computed for each catalog using a maximum-likelihood method (Weichert, 1980), and they are spatially smoothed

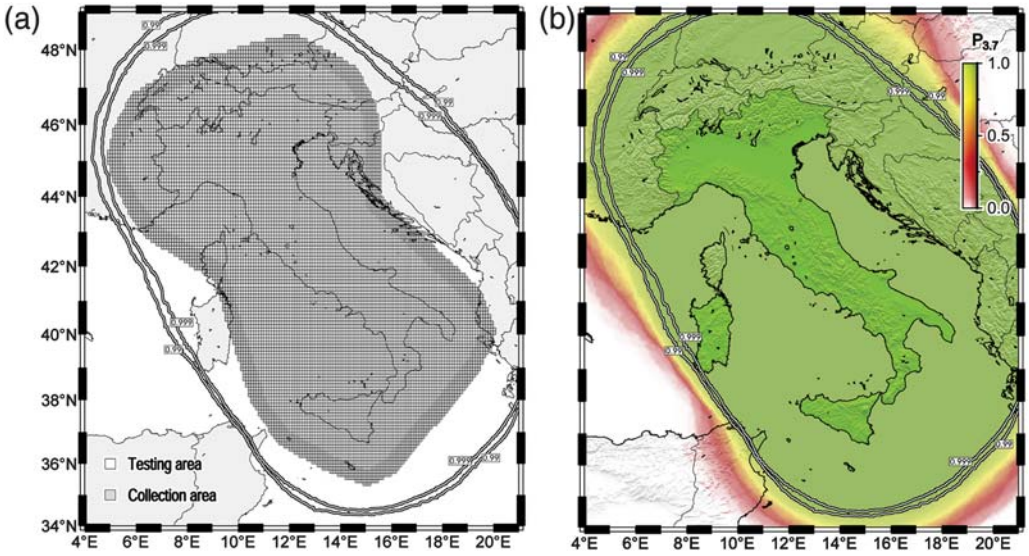

Figure 1. (a) Grid of the test area in this work and in the CSEP experiment (Schorlemmer, Christophersen, *et al.* 2010). (b) Probability of detecting an earthquake with magnitude 3.7 or larger in the testing area (Schorlemmer, Mele, and Marzocchi, 2010).

using a two-dimensional Gaussian function with a 15-km correlation distance (for more details, see Akinci, 2010). The error in the epicenter location is assumed to be three times larger than the correlation distance (Frankel, 1995).

MPS04 (Gruppo di Lavoro MPS, 2004). This model is the reference model for the seismic-hazard map of Italy. The EO model is composed of the temporal Poisson process applied to a seismotectonic zonation (Meletti *et al.*, 2008). The frequency–magnitude distribution varies across the zones, being a truncated Gutenberg–Richter distribution with different maximum magnitudes derived from the historical seismicity (Gruppo di Lavoro CPTI, 2004) and tectonic considerations (Meletti *et al.*, 2008; see Data and Resources section). The parameters of the original model are set using a declustered seismic catalog (Gruppo di Lavoro CPTI, 2004). Here, all calculations are run adding a correction factor that takes into account possible aftershocks. This correction increases each rate by a factor of 1.25 (Faenza and Marzocchi, 2010). In any case, we anticipate that the inclusion or not of this correction factor will not change the final results significantly.

Time-Dependent Models

DBM (Lombardi and Marzocchi, 2010b). The double-branching model is a stochastic time-dependent model that assumes that each earthquake can generate other earthquakes through different physical mechanisms (Marzocchi and Lombardi, 2008). Specifically, it consists of a sequential application of two branching processes (Daley and Vere-Jones, 2004) in which any earthquake can trigger a family of later events on different space–time scales. The first part of our model consists of a well-known epidemic-type aftershock sequence (ETAS) model (Ogata, 1988), describing the short-term clustering of earthquakes due to coseismic stress transfer. The second branching process works at larger space–time intervals (typically about 30 years; Marzocchi and Lombardi, 2008) with respect to the smaller domains involved in short-term clustering, and it describes possible further correlations between earthquakes not associated with coseismic stress-perturbations.

DISS-BPT. The Database of Individual Seismogenic Sources (DISS)–Brownian Passage Time (BPT) time-dependent model has been set in the framework of the S2 project with a clear and specific scientific goal: project participants aim to verify whether a model fully based on a Brownian passage time (BPT) model and applied to faults using the best available knowledge currently available is close to providing reliable large earthquake forecasts for the Italian region. The scientific importance of this model is linked to the use of similar EO models to forecast the largest earthquakes in California (Field *et al.*, 2009). Basically, the model is based on a BPT model (Matthews *et al.*, 2002) applied to the individual seismogenic sources (ISS) included in the DISS3 database (see Data and Resources section) and follows a characteristic earthquake model for the frequency–magnitude distribution. ISS are determined on the basis of geological and geophysical data and characterized by a set of geometric, kinematic, and seismological parameters. Each contains 115 individual sources, sixteen of which are not associated with any earthquake. Most of the fault slip rates are not well constrained, and they are set to a conventional range of 0.1–1.0 mm/yr. Only 44 out of 115 faults have slip rates estimated from local geological data. The input parameters of the BPT and the characteristic magnitude of each fault are taken from the DISS3 database without critical review. We use the mean value slip rate given by DISS3. For those 16

faults in which the latest earthquake is classified as Unknown in the database, we assume 1500 years of elapsed time; this value has been arbitrarily chosen in accordance with the maximum length of the Italian historical seismic catalog. The value of aperiodicity (coefficient of variation, COV) assumed is 0.5 for each fault, according to Ellsworth *et al.* (1999). The probability estimated for each fault is homogeneously partitioned over the cells of the grid that overlap the projection of the same fault at the surface. These two last features mark the difference between the DISS-BPT model used here and the version developed in S2 by other researchers. We anticipate that the results reported here do not depend on these changes.

LTST (Falcone et al.*, 2010).* The long-term stress transfer (LTST) model is obtained by a statistical renewal model that mimics the recurrent (characteristic) behavior of faults and by a physical-based model that considers fault interaction. The recurrent behavior of the faults is based on the BPT model (Matthews *et al.*, 2002) that quantifies the interevent time probability distribution of earthquakes on individual sources. This model is applied to seismogenetic sources, reported in the DISS3 database (Data and Resources), for which the date of the last event is known (104 faults). The fault interaction is represented by the coseismic static permanent Coulomb stress change (Stein *et al.*, 1997) caused by each earthquake reported in the catalog. This stress change can increase or decrease earthquake probability based on a simple renewal model. The final model accounts for the effects of all past earthquakes that induced stress variations on the selected faults. The final forecasts are obtained by adding a time-independent background seismicity to the cells in which no seismogenic sources have been identified; background seismicity is estimated by past activity using the procedure described by Frankel (1995).

HAZFX-BPT (Akinci et al.*, 2009).* This model incorporates both smoothed historical seismicity ($M_w < 5.5$) and geological information on faults ($M_w \geq 5.5$). In the background model, the seismicity rates are obtained for earthquakes $M_w < 5.5$ following the procedure described by Akinci (2010). The earthquake recurrence rates of individual fault for $M_w \geq 5.5$ events are derived in a straightforward manner from slip rate and magnitude, using the technique known as the conservation of the seismic moment rate (Field *et al.*, 1999). Time dependency is introduced through a BPT model (Matthews *et al.*, 2002). The conditional probability of each individual faults, P_{cond}, in a given T years is expressed in terms of an effective Poisson annual rate, R_{eff}, such that

$$R_{eff} = -\ln(1 - P_{cond})/T. \quad (1)$$

These rates are then converted into maps of earthquake rate density by using an isotropic Gaussian filter with 15 km of correlation distance (for more details see Akinci, 2010). In this way, this procedure merges the earthquake potential of adjacent faults and turns a fault-based estimate into a grid-based estimate.

For individual fault sources, we use the integrated dataset from DISS3 that includes both ISS and macroseismic well-constrained sources (MWS; Data and Resources). The detailed information of the fault geometry (dip, length, width) and its seismic behavior (slip rate, maximum magnitude) are taken by Basili *et al.* (2008), Akinci (2010), and Akinci *et al.* (2010). According to Ellsworth *et al.* (1999), the value of COV is set equal to 0.5 for each fault. For the 16 out of 232 sources of DISS3 that are not associated with any earthquake (latest earthquake referred to as "unknown"), the elapsed time is imposed to 1500 yr, assuming that the last event occurred around 500 A.D. (Akinci *et al.*, 2009).

The Retrospective Tests of the Models: Reliability and Skill

The models described in the section The Selected Models are based on different assumptions and account for different geological and geophysical information. In order to gain some insight into how these models may perform in forecasting future large events, we run a retrospective CSEP-type testing experiment (Werner *et al.*, 2010). The goal of this testing experiment is to assess the (retrospective) reliability and skill of each model. More details about the CSEP testing procedures may be found in Schorlemmer *et al.* (2007) and Zechar, Gerstenberger, and Rhoades (2010).

CSEP experiments are truly prospective, in that they are meant to compare forecasts and future seismicity. Of course, this testing phase may require many years to be completed (as noted previously, the first CSEP results for Italy will be available in 2014). Meanwhile, we can get some preliminary information about the models that produce the most reliable and skillful forecasts through a retrospective test (Werner *et al.*, 2010), that is, using earthquakes that already occurred. Retrospective tests do not represent the optimum strategy to test models and forecasts because some sort of retrospective (conscious or unconscious) adjustment or overfit can never be ruled out. In other terms, this means that the results of retrospective tests may represent an upper limit or, even worse, an optimistic estimation of the real forecast capability of the models. This testing phase aims at showing unreliable models (models that produce inaccurate forecasts) and the retrospective skill of each model. In this way, we may assign a weight to each model depending on how these models have performed in this experiment. Most of the work presented here has been done in the framework of the S2 project.

As for prospective tests, retrospective experiments also require the definition of the rules of the game (see Schorlemmer, Christophersen, *et al.*, 2010, for the rules of the game for the Italian CSEP experiment). These are

- *Set of models under testing:* MPS04, HAZGRIDX, HAZFX-BPT, DBM, DISS-BPT, and LTST.
- *Testing period:* 1 January 1950 to 31 December 2009.

- *Testing area:* the area is the same as that considered during the real CSEP experiment in Italy (see Fig. 1). In the next figures, we report the results only for the Italian territory.
- *Authorized seismic catalog for the testing phase*: the CPTI08 catalog used for the CSEP experiment (provided by the Gruppo di Lavoro CPTI Working Group) integrated for the last part of the testing period with the data of the Centroid Moment Tensor (CMT) catalog. In contrast to previous versions, CPTI08 is not declustered; and, in accordance with the CSEP experiment (Marzocchi *et al.*, 2010; Schorlemmer, Christophersen, *et al.*, 2010), we consider all earthquakes for the testing phase, without removing possible aftershocks.
- *Forecasting time intervals*: 10 years. The first forecast spans from 1 January 1950 to 31 December 1959; the second forecast goes from 1 January 1960 to 31 December 1969, and so on, for a set of six forecasts per model.
- *The forecast*: each model has to provide the probabilities of 1, 2, and 3 or more earthquakes, or P(1), P(2), and P(3+), respectively, in each specific space–time-magnitude bin. Each bin always has a time length of 10 years, while the magnitude considered is only one class ($M_w \geq 5.5$), and the spatial dimension is each cell of the CSEP grid reported in Figure 1 (see Schorlemmer, Cristophersen, *et al.*, 2010; Schorlemmer, Mele, and Marzocchi, 2010). Note that this aspect marks a difference with respect to the CSEP experiments in which the seismic rate and different magnitude bins are taken into consideration. While this difference does not lead to substantially different interpretation of the results, the use of probability instead of the seismic rate has an important aspect that deserves to be mentioned: the use of probability implies that the uncertainty associated with the expected number of events is already incorporated in the calculations by the EO model.
- *Learning information for calibrating the models*: ideally, all models should use only information available before the forecast starting date (i.e., for the first forecast, before 1 January 1950). In practice, this is impossible for models that use geological information or a seismotectonic zonation (MPS04, HAZFX-BPT, DISS-BPT, and LTST). Unfortunately, this objective advantage is neither quantifiable nor reducible; on the other hand, a rigid application of the prospective rules would lead us to not consider the national seismic-hazard map (MPS04), which is hard to justify for practical applications. For this reason, we keep a pragmatic attitude, simply keeping in mind the objective advantage of these models when they are evaluated and compared to the others. Hopefully, future applications of this procedure using the results of the CSEP experiments (truly prospective) may reduce this subjectivity as new data for the testing phase become available.

We do not use the results of the retrospective analysis of time-independent EO models carried out by Werner *et al.* (2010) for two reasons. First, the S2 experiment uses a slight different testing procedure/format with respect to CSEP, making the results of the two experiments difficult to compare. Second, because time dependency may play an important role in a 10-year forecast, the S2 experiment also includes time-dependent models, while the Werner *et al.* (2010) paper does not.

The six forecasts for each model are reported in Figure 2. The models present a large variability being based on quite different physical assumptions and geological components. The variation of the forecasts of each time-independent model is only due to the inclusion of more data to calculate the expected seismic rate. The largest differences are observed for the time-dependent models (DBM, LTST, HAZFX-BPT, and DISS-BPT), because the forecasts may significantly change through time, even on a wide area, as a consequence of recent seismicity; for instance, the DBM model shows forecasts that vary through space and time, due to the effects of long-term earthquake clustering. All forecasts are compared with the real seismicity using the tests described in the Appendix. Here, we just remark that the *N*- and *L*-tests are basically goodness-of-fit tests that check the forecast reliability for each model, whereas the *R*-test checks if the forecast of one (reliable) model is significantly better (higher skill) than the forecast of another (reliable) model.

In Figure 3, we show the log-likelihood for the different models as a function of the six forecasts and, in the legend, the cumulative values. From Figure 3 we see that one model (DISS-BPT) has a log-likelihood equal to infinity (i.e., an earthquake occurred when the model forecasted zero probability). The LTST model has lower finite log-likelihood values with respect to the other models. This is likely due to the smaller spreading of the spatial distribution of forecasts given by the LTST model, which fail to predict the location of some real events (see also Fig. 2).

In Figure 4, we compare observed earthquakes for each EO model with the expected number of earthquakes. The average number of earthquakes for each model is calculated as

$$\langle N \rangle \approx -\sum_i \ln[1 - P_i(0)], \tag{2}$$

where $P_i(0) = 1 - [P_i(1) + P_i(2) + P_i(3+)]$ for the i-th cell. This approximation holds reasonably well when probabilities and the expected number of events are small (and consequently similar), as in the present case. In Figure 4, we see that LTST and DISS-BPT grossly fail to predict the number of observed events. The LTST's forecasts lead to a much higher average number of expected events, while DISS-BPT goes in the opposite direction, underestimating the expected number of earthquakes. As implicit in their nature, the expected number of earthquakes presents no significant variations through time for time-independent EO models, while different kinds of time-variation are seen for time-dependent EO models. These varying trends reflect basic differences in the physical and geological components included in the time-dependent models.

The agreement between forecasts and observations in terms of number of events—that is, spatial bins that have experienced one or more earthquakes—is quantitatively evaluated by the *N*-test (see Appendix). In Figure 5, we show the values of δ (percentage of simulations generated by the models in which the number of events is less or equal than the observed number) for each forecast, and in the legend we report the cumulative values of δ for the entire testing period. Forecasts of the model are not reliable if $\delta < 0.025$ (the number of observed events is significantly less than expected) or $\delta > 0.975$ (the number of observed events is significantly more than expected). The cumulative *N*-test confirms that the LTST and the DISS-BPT are not able to predict the number of events for the entire testing period. Looking at each forecast, DISS-BPT does not pass the *N*-test for two out of six forecasts. On the other hand, the *N*-test rejects the LTST model in the first testing time interval and shows that the LTST model constantly forecasts a higher number of earthquakes for all time intervals. The low reliability of DISS-BPT and LTST explains most of the large variability in the expected

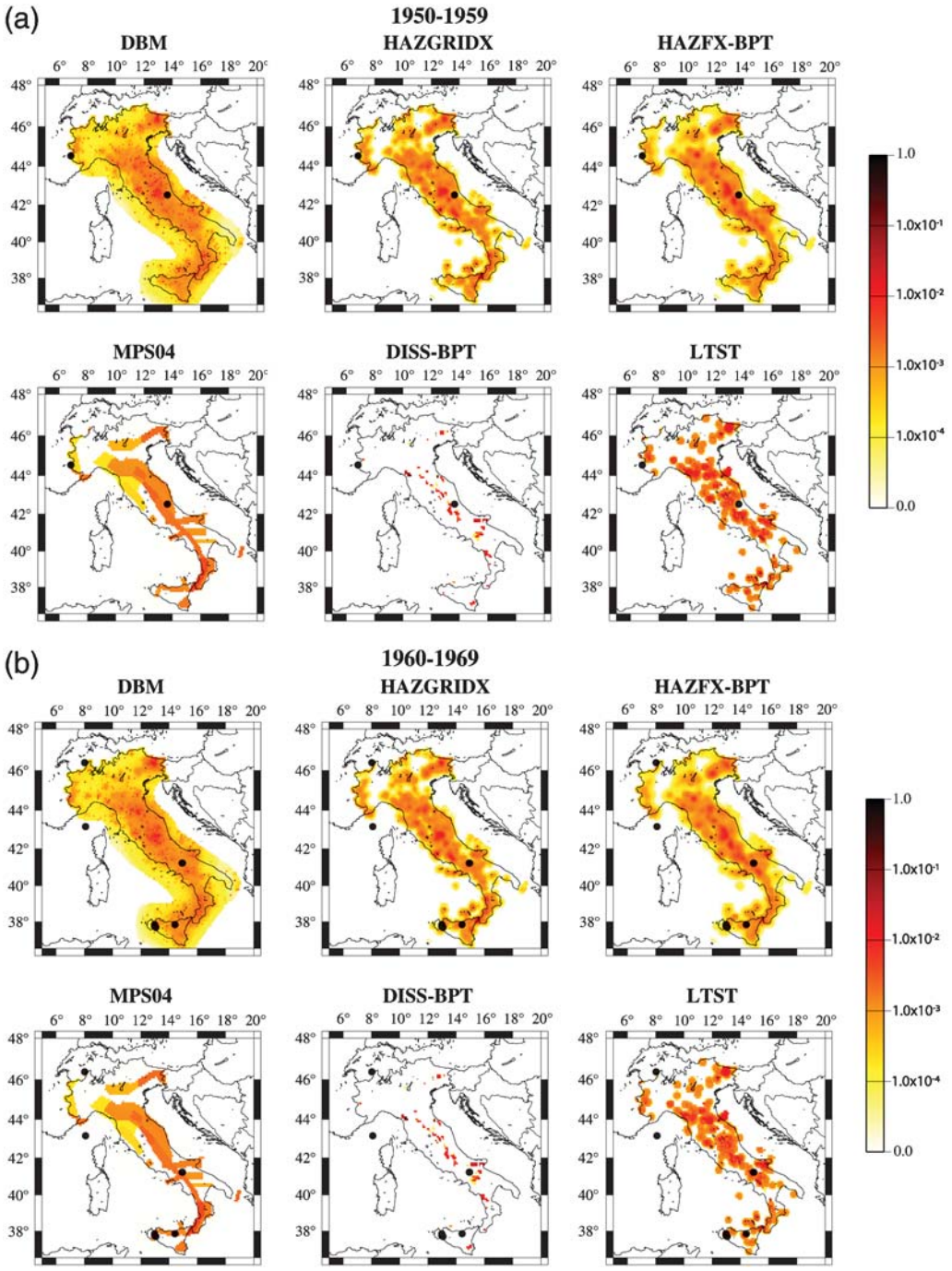

Figure 2. Forecasts of the six models for six decades: 1950s–2000s. (*Continued*)

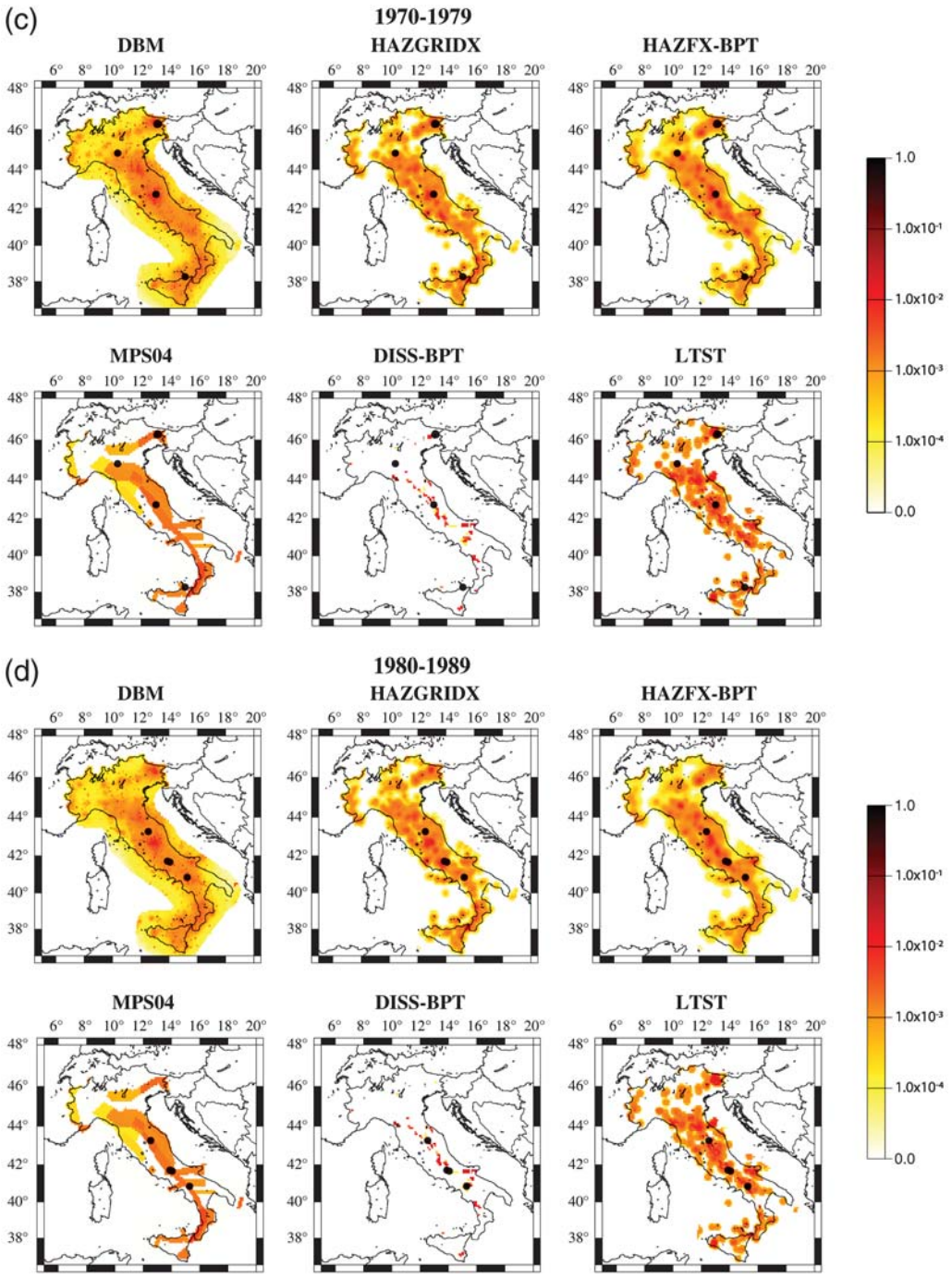

Figure 2. Continued.

number of earthquakes reported in Figure 4. The reliable models present variability, in terms of the expected number of earthquakes, which is less than what would be expected by pure chance from a Poisson process. In particular, if we correct the rates of the MPS04 model to account for the presence of possible aftershocks (see the description of the model in the MPS04 [Gruppo di Lavoro MPS, 2004] section), the total variability among the four reliable models is of a few expected earthquakes.

In Figure 6, we report the results of the *L*-test, both for each specific forecast and the entire testing period. The *L*-test, as described in the Appendix, also considers the spatial component of forecasts, as well as the number of events taken into account by the *N*-test (see also Zechar, Gerstenberger, and Rhoades, 2010). The results show a scarce fit (when $\gamma < 0.05$) of the spatial distribution for MPS04 and DISS-BPT. The LTST model passes the *L*-test, although it shows a low spatial fit of the observed events (Fig. 2). These results might appear

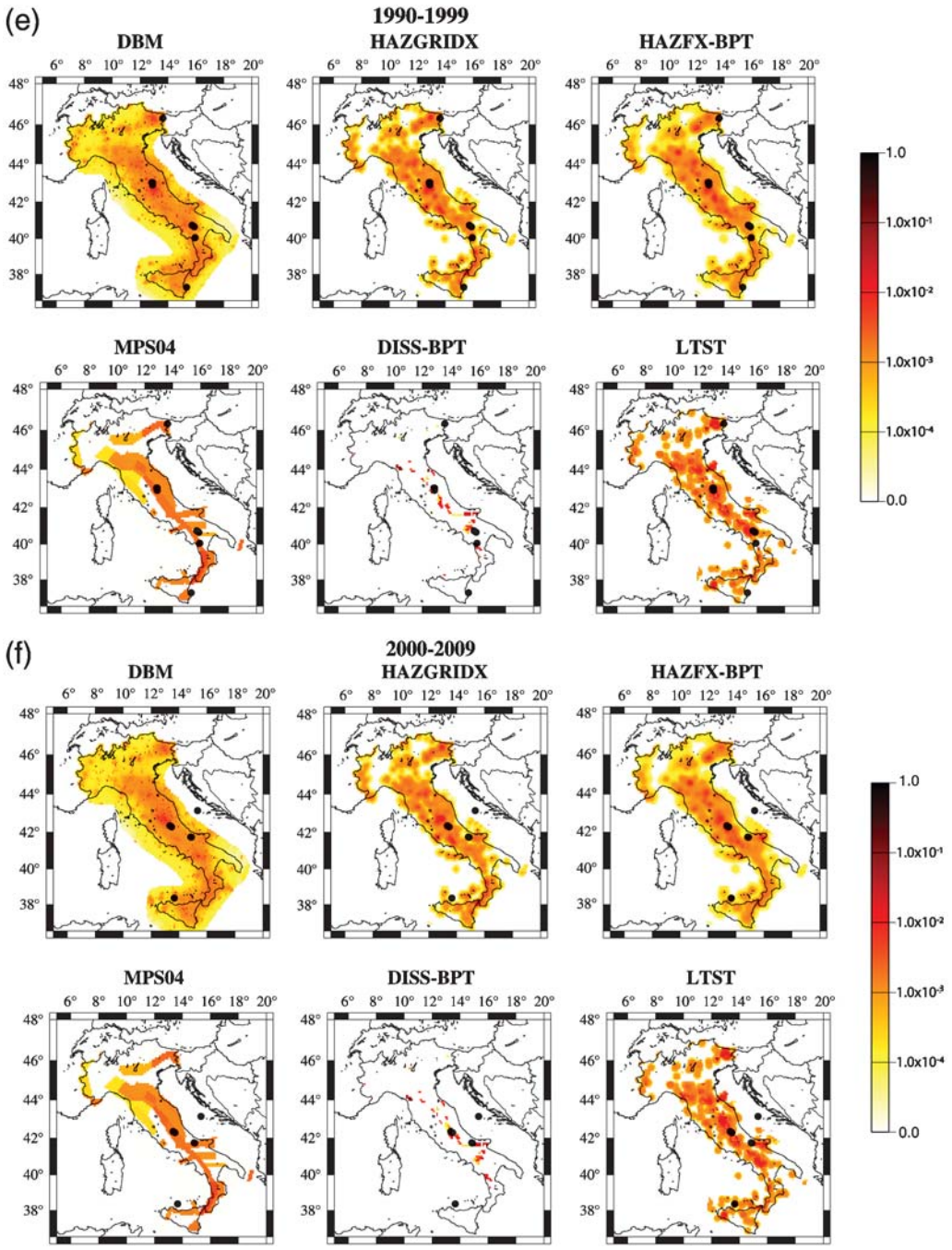

Figure 2. Continued.

surprising, because LTST has low log-likelihood scores (Fig. 3) and it has been rejected by the *N*-test (Fig. 5); actually, this apparent inconsistency can be explained as an effect of the systematic overestimation of the number of events given by the LTST model (Fig. 4; also Schorlemmer *et al.*, 2007).

Only models DBM, HAZFX-BPT, and HAZGRIDX passed both tests at a 0.01 significance level. The failure of model MPS04 deserves specific attention, being the basis of the national seismic-hazard map (MPS04). In particular, the failure of the *L*-test is due to earthquakes that occurred offshore, where MPS04 provides a very low probability. However, offshore earthquakes were not the original targets of MPS04 because the national seismic-hazard map is used to set the building code; as a general consideration, we may say that target earthquakes inland or offshore should have a different weight in the testing procedures if the model is oriented toward seismic risk mitigation. Hence, we run the tests in a smaller area covering only onshore areas. In this case, MPS04 passes both tests; therefore, we have decided to keep this model as well for the next stages of the analysis.

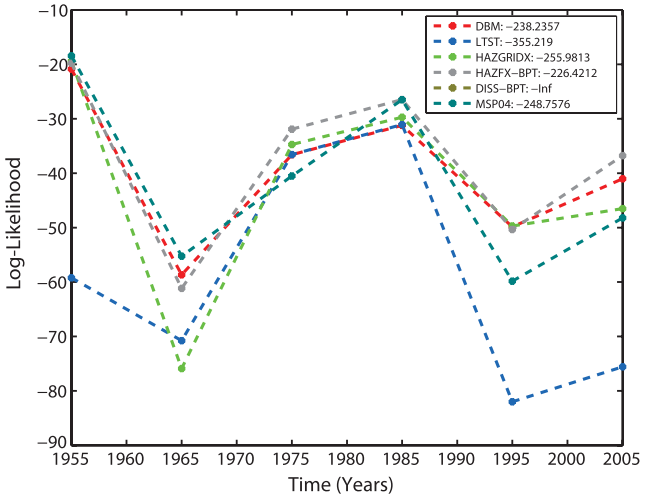

Figure 3. Log-likelihood test of the six EO models considered in this study for each forecast (10 years each). The legend reports the cumulative log-likelihood; -Inf refers to an infinite negative likelihood.

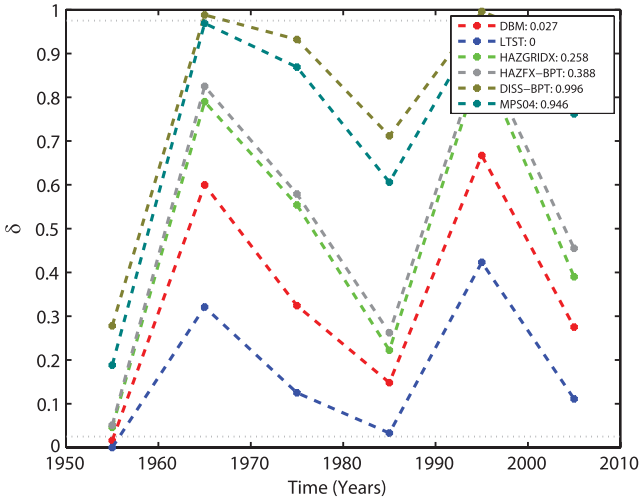

Figure 5. *N*-test results. The variable δ represents the percentage of simulations generated by the models in which the number of events is less or equal than the observed number. The two horizontal dotted lines indicate the 0.05 significance level of the test. The legend reports the cumulative significance level of the test for the whole testing period.

The two models mostly based on faults, LTST and DISS-BPT, fail at least one of two tests. The same results are found considering only inland earthquakes. It may be surprising that the DISS-BPT model fails the *L*-test, because the fault distribution on the DISS3 database has been defined by also including the most recent earthquakes that are the target events in the testing procedure. This is probably due to: (1) an erroneous geometry of the inferred seismic sources; (2) epicenter uncertainties move some earthquakes off the source projection at the surface; and (3) some $M_w \geq 5.5$ sources are still unknown—and thus the source model is incomplete. Failure of the *L*-test notwithstanding, the most remarkable point is that this model fails the *N*-test. In other words, even somehow fixing the problem of spatial distribution, this model forecasts many fewer earthquakes than were observed. The failure of LTST and DISS-BPT to forecast the right average number of earthquakes may be a symptom of our still-limited knowledge of many relevant active faults and paleoseismological data (e.g., completeness of the fault database, realistic modeling of the seismic source from the geometry of the faults, accurate estimation of the relevant fault parameters such as slip rate, slip per event, timing of paleoearthquakes, etc.).

To summarize, only HAZGRIDX, DBM, HAZFX-BPT, and MPS04 pass to the next step of the testing phase, the

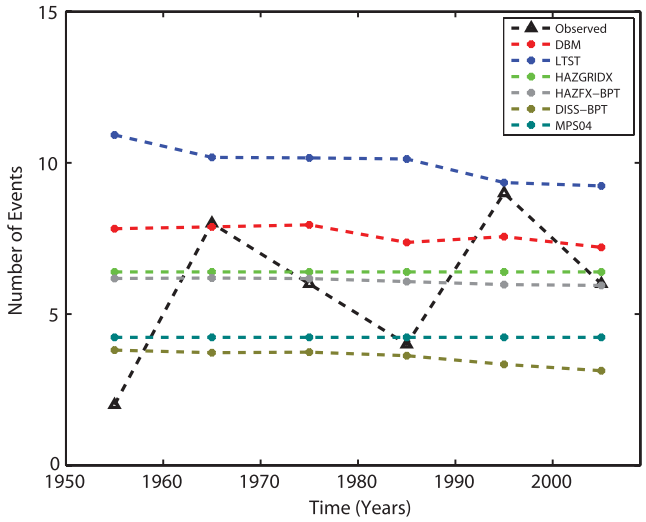

Figure 4. Observed (black line) and the average number of events expected by the models.

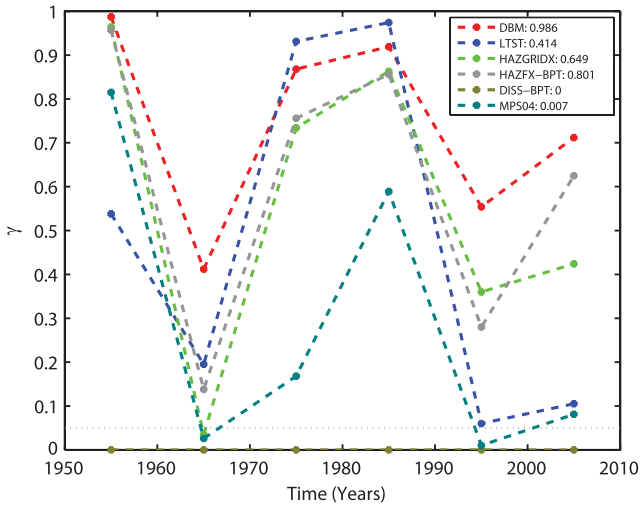

Figure 6. *L*-test results. The variable γ represents the percentage of simulations generated by the models in which the simulated log-likelihood is smaller than the observed log-likelihood. A value of $\gamma < 0.05$ indicates a significant difference between forecast and model.

R-test (see Table 1). The *R*-test is used to see if a model H_1 performs significantly better than a model H_2 (or not) under specific assumptions (see Appendix; also Schorlemmer *et al.*, 2007). A positive log-likelihood difference means that H_1 performs better than H_2. The fourth column reports whether the log-likelihood difference is statistically significant, according to the Schorlemmer *et al.* (2007) interpretation of the *R*-test. In this view, a model performs significantly better than another if the *p*-value of the *R*-test is greater when the log-likelihood is positive and smaller when we compare the same models in the opposite order (negative log-likelihood). More details can be found in Schorlemmer *et al.* (2007) and the Appendix of this paper.

The *R*-test results of Table 1 indicate that HAZFX-BPT has the best performances because its results are always better than the other three models in terms of log-likelihood. Models DBM, MPS04, and HAZGRIDX follow in order of performances, with DBM being better than MPS04 and HAZGRIDX and MPS04 being better than HAZGRIDX. The last column of the table reports the probability gain per event (PGpe), defined as

$$PGpe = \exp\left(\frac{l_{H_1} - l_{H_2}}{N}\right), \tag{3}$$

where $l_{H_1} - l_{H_2}$ is the difference between the two log-likelihoods and N is the number of target earthquakes. This quantity gives an immediate picture of how model H_1 performs better than model H_2. The intrinsic subjectivity of any weight assignment procedure notwithstanding, we argue that PGpe may represent a natural weight for each model.

Despite the fact that HAZFX-BPT appears to be the best performing, it is worth noting that differences in terms of log-likelihood (Fig. 3) are not statistically significant. Moreover, we have to keep in mind that HAZFX-BPT has the objective advantage of including information from faults that are well known and mapped because they produced large earthquakes in the testing period. In fact, from Figure 3, we can see that HAZFX-BPT performs better than DBM in the three decades in which the largest earthquakes occurred (Friuli, 1976; Irpinia, 1980; L'Aquila 2009). MPS04 has similar advantages because the zonation has been made by also taking into account the occurrence of the earthquakes in the testing period. The small PGpe difference among the models corroborates the view that no model is significantly better than the others. From a scientific point of view, we note that the two time-dependent models (HAZFX-BPT and DBM) are the best performing, even though the differences with the time-independent MPS04 are not statistically significant using the *R*-test, in the form suggested by Schorlemmer *et al.* (2007). On the other hand, as noted before, MPS04 has some objective advantage with respect to DBM that is fully prospective. For example, it only uses information available before the date in which the forecast is issued. Such an advantage smooths the real difference in the forecasting capability of DBM and MPS04. Using a different procedure, Lombardi and Marzocchi (2009) showed that DBM is significantly better than a time-independent smoothed seismicity, but the results reported here do not yet prove the significance and importance of the time-dependency in a forecasting time interval of 10 years. Hopefully, future tests with a larger dataset may bring much more robust conclusions on this important issue.

Defining and Building the Best Model

In theory, the concept of best model may sound easy to understand, but, in practice, what "best" means is far from obvious. The selection of the best model is very common in many fields, and there is no unique strategy to accomplish this goal in a rational way. To this purpose, it is worth remarking the recent example of two important initiatives, namely the Collaboratory for the Study of Earthquake Predictability (CSEP) experiments and the Unified California Earthquake Rupture Forecast (UCERF), that follow different attitudes in obtaining—or suggesting the best procedure to define—the best model. Marzocchi and Zechar (2011) describe in detail this important issue. Here, we report only the main features. The basic distinction of the CSEP and UCERF approach is mostly related to a key issue, how much of what we know about small-to-moderate earthquakes can be

Table 1
Results of the *R*-Test for the EO Models

Model H_1	Model H_2	Log-Likelihood Difference $[l_{H_1} - l_{H_2}]$*	Is $[l_{H_1} - l_{H_2}]$ Statistically Significant?†	PGpe‡
DBM	HAZGRIDX	+ 17.75	NO	1.66
DBM	MPS04	+ 10.52	NO	1.35
HAZFX-BPT	DBM	+ 11.81	NO	1.40
HAZFX-BPT	HAZGRIDX	+ 29.56	YES	2.33
HAZFX-BPT	MPS04	+ 22.33	NO	1.89
MPS04	HAZGRIDX	+ 7.22	NO	1.23

*The log-likelihood difference $[l_{H_1} - l_{H_2}]$ between the models for the whole testing period.
†Results of the statistical test (significance level of 0.01) according to the interpretation of the *R*-test made by Schorlemmer *et al.* (2007).
‡PGpe, probability gain per event of the first model with respect to the second one.

extrapolated from the largest earthquakes. The largest earthquakes are not frequent, therefore the CSEP experiments mostly rank EO models by taking into account smaller events. On the other hand, UCERF implicitly assumes that the largest earthquakes may have peculiarities (for instance, they might occur only in specific areas) that make them different from smaller events. Because only very little data of large earthquakes are available, UCERF builds the best model by adopting a significant amount of expert judgments. Expert judgment is necessary when data are weak and in the presence of large uncertainties, yet, contrarily, it usually moves the model against parsimony and testability, which are two main concepts of science.

In order to achieve the best EO model for the Italian case, there are a couple of relevant constraints/conditions that we must consider. First, the available models are in the CSEP format, and we have enough past target earthquakes to test the models retrospectively. Second, the still limited number of known active faults and paleoseismological and other possibly relevant data prevent us from seeking a broader consensus model through some sort of expert judgment procedure. These constraints and conditions lead us to the decision of achieving the best model through a weighted average of those models that produced reliable forecasts in the retrospective experiments (see The Retrospective Tests of the Models: Reliability and Skill). In this case, the best model is implicitly defined as the model that produces the best forecasts, according to a metric measurement established by the statistical tests adopted. The weighted average has several similarities with the logic-tree approach widely used in seismic-hazard assessment. Here, the weight of each model is not arbitrarily assigned, as is usually done in logic-tree applications, but it is assessed using the skill of the forecasts in the retrospective test. This procedure is similar to what Rhoades and Gerstenberger (2009) proposed: to mix two short-term models that were focused on different aspects of the earthquake occurrence process.

In general, the use of the weighted average of the models has some features and limitations that deserve mention. First, the model average does not necessarily imply being closer to reality. For instance, if we assume that the set of models used to calculate the average also includes the true model (assuming that a "true" model may exist), each kind of average will produce forecasts that are worse than the forecasts of the single true model. In other words, in practical applications we never know which is the true model because of the ubiquitous presence of epistemic uncertainty, but we emphasize that the models' average is not necessarily better than each single model. Second, and related to the previous point, if the epistemic uncertainty plays a major role, the variability among the forecasts becomes of great interest because it may be considered a measure of epistemic uncertainty. Third, the results of the average depend on the selection of the models. For instance, if we consider a large number of models based on very similar components, the final importance of these components in the average map tends to be overweighted. Ideally, the selected models should be representative of the different views of the earthquake occurrence process or account for the correlation among the forecasts. Fourth, the average among models may account for different aspects of the earthquake generation process (see Rhoades and Gerstenberger, 2009). For instance, some models are able to well describe the space–time clustering at different scales, while other models may incorporate time recurrence on major faults. In this case, the average model may or may not provide more exhaustive forecasts that account for these different aspects. Fifth, the combination of probabilities through the average operator, in some cases, might pose conceptual problems both from a theoretical and philosophical point of view. Nonetheless, when probabilities are very small, as in the present case, the difference between probability and the expected number of earthquakes becomes negligible; therefore, the average of probabilities can be seen as the average of the expected number of events.

In the section The Retrospective Tests of the Models: Reliability and Skill, we have seen that four out of six models, HAZFX-BPT, DBM, HAZGRIDX, and MPS04, show reliable forecasts. The *R*-test indicates that all models perform similarly. Hence, we conclude that, at the present state of knowledge, all EO models may be considered equally skilled. Therefore we define the best model as the average model (named AVE) that represents the average of the four EO models HAZFX-BPT, DBM, MPS04, and HAZGRIDX. In order to check whether the model AVE is more skilled than any of the single models, we run the same experiment carried out for each model and described in section The Retrospective Tests of the Models: Reliability and Skill. The results of the *N*- and *L*-test are reported in Figure 7; the plot shows that the model AVE produces accurate forecasts for the past six decades. The results of the *R*-test are reported in Table 2, showing the higher skill of the model AVE with respect to the other models.

The AVE model has a small increase of skill with respect to HAZFX-BPT. This can be due to the fact that HAZFX-BPT has the clear advantage of using information of the testing phase (i.e., the location and geometry of the faults that produced the largest earthquakes during the past 60 years); this leads the HAZFX-BPT model to perform particularly well in forecasting the largest earthquakes of the testing phase (Fig. 3).

In Figure 8, we show a map of future seismicity forecasts in Italy (2010–2019) predicted by the AVE model. In Figure 9, we report the (COV of the map, the ratio of the standard deviation and average of the models' forecasts. This map highlights the areas where forecasts of individual models are similar or different. In particular, COV = 0 means that the forecasts of all EO models are identical; COV = 1 means that the standard deviation of the forecasts of all EO models is comparable to the average (areas with a large epistemic uncertainty). The most striking evidence one can see from the maps is that a large part of central Italy presents low COV values, while the southern part is characterized by higher COV values. This spatial variability indicates areas where EO models have a strong or weak consensus. For example, some

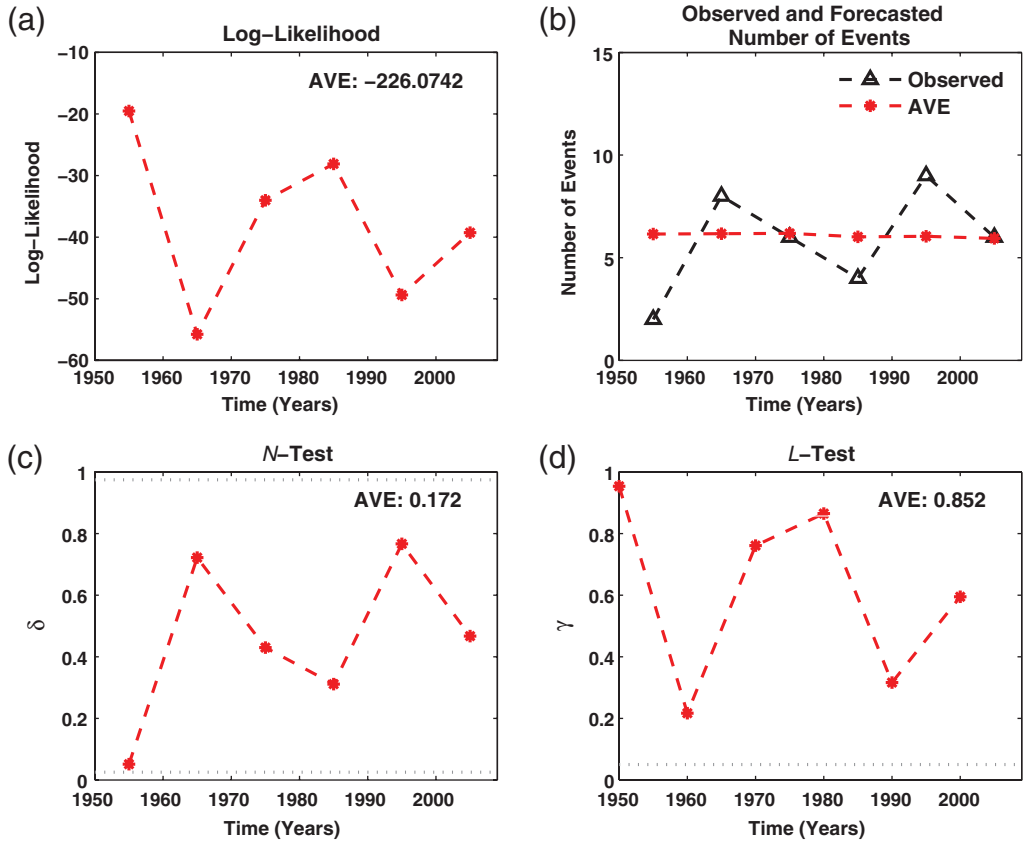

Figure 7. Retrospective analysis of the AVE model: (a) log-likelihood; (b) number of predicted vs. observed earthquakes; (c) *N*-test results; and (d) *L*-test results. The legend reports the cumulative log-likelihood in (a) and the cumulative significance level of the *N*-test and *L*-test for the whole testing period in (c) and (d), respectively.

high values, as noted in the Puglia region and offshore from Sicily, are probably consequences of the seismotectonic zonation adopted by MPS04. In reality, the MPS04 assigns very small probabilities in these regions because they are outside of the tectonic zonation, while other models do not. Other high values of COV in southern Italy may be explained by the use of time-independent and time-dependent models. In fact, while southern Italy has been historically characterized by a high level of seismicity (high probability in time-independent models), a time-dependent model based on clustering (e.g., DBM) tends to provide comparatively higher probabilities in other areas that have been recently characterized by higher earthquake frequency, like central Italy (see Faenza *et al.*, 2003; Cinti *et al.*, 2004; Lombardi and Marzocchi, 2009). Note that the scarce number of reliable models (4) prevents us from providing a more complete characterization of the epistemic uncertainty, for instance, through quantiles or an empirical distribution.

We emphasize the importance of considering the spatial variability of the agreement among forecasts for risk mitigation actions. For instance, let us consider two cities with exactly the same low probability of large earthquake occurrences (i.e., the same value in the AVE map), with the second having a larger uncertainty; this means that the latter may have a much greater probability of experiencing a large shock with respect to the first one.

Table 2

Results of the *R*-Test for the EO Models Relative to the Comparison of the Average (AVE) Model, with Respect to the Single Models*

Model H_1	Model H_2	Log-Likelihood Difference $[l_{H_1} - l_{H_2}]$	Is $[l_{H_1} - l_{H_2}]$ Statistically Significant?	PGpe
AVE	HAZGRIDX	+ 29.91	YES	2.35
AVE	MPS04	+ 22.68	YES	1.91
AVE	DBM	+ 12.16	YES	1.42
AVE	HAZFX-BPT	+ 0.347	NO	1.01

*Column descriptions are the same as for Table 1.

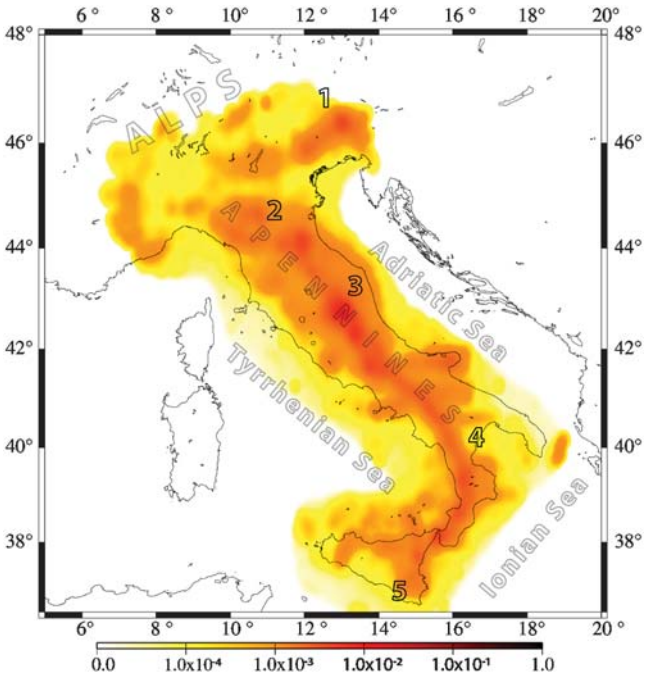

Figure 8. Occurrence probabilities of $M > 5.5$ earthquakes in the AVE model for 2010–2019. This model is derived as an equally weighted average of the four models that passed the tests (see text for explanation).

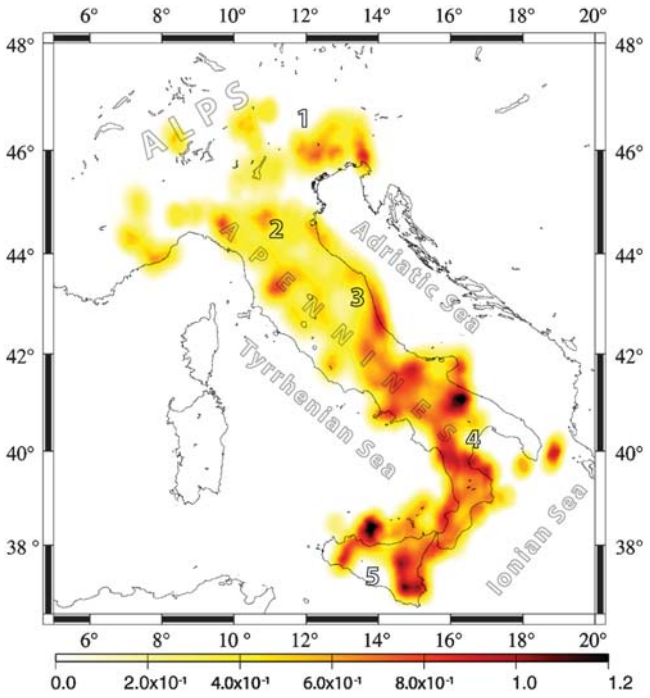

Figure 9. Plot of the spatial coefficient of variation (COV) of the four models used (see text for explanation).

A Closer Look at the Best EO Model

In this section we examine the average EO map in the framework of the Italian tectonics. We also discuss the regions of the map showing the highest probability of occurrence, including information and evaluations based on data not considered in the present EO models. Figure 10 reports the EO map of Figure 8, and the details of the area discussed in this section. The main goal of this section is to describe in detail the different tectonic regions of Italy, where the average model shows either high EO probabilities or large variability among the different models.

Eastern Alps and Prealps

This belt shows an EO probability that is clearly high, with respect to the rest of the Alpine belt (area 1 in Figs. 8 and 10). This is consistent with what is known from the active tectonics of the central Mediterranean region, where the Adriatic plate moves northward, creating compression along the Alps. GPS data confirm these kinematics, although it must be noted that a detailed estimate of the strain rate along the whole Alpine belt is not yet available, due to poor network coverage. Good constraints are available only for the eastern Alps, thanks to the regional GPS network FredNet (D'Agostino *et al.*, 2005).

Individual seismogenic sources are relatively well defined in this region, with a continuous belt where $M_L > 6$ earthquakes are expected from Friuli to Garda Lake (Fig. 8).

The high occurrence probability in this area comes from all the models, with HAZFX-BPT suggesting a higher potential for the western sector. This includes the Montello fault system where an M_L 6.5 is expected by some authors (e.g., DISS, see Data and Resources).

Northern Apennines

In this area (area 2 in Figs. 8 and 10), M_L 5–6 earthquakes (or even slightly larger) are frequent in the historical catalog. Nonetheless, in the last 60 years only a few $M_L >$ 5.5 events occurred (Parma region, 1971, and the Colfiorito earthquakes in 1997). This area includes both compressional earthquakes of the external front (Parma, Reggio Emilia, Bologna, Forlì, Adriatic coast, etc.), occurring mostly at a 15–20-km depth, and shallow (5–10 km) normal faulting events in the internal sector of the Apennines, that is, Umbria and Tuscany (Mugello, Garfagnana, High Tiber Valley, etc.). This latter sector was very active at the beginning of the twentieth century (Monterchi 1917, Mugello 1919, Garfagnana 1914 and 1920) but did not experience significant events in the decades to follow until the Umbria earthquakes in 1979 (Norcia), 1984 (Perugia–Gubbio), and 1997 (Colfiorito).

Our results show that the highest probability of occurrence is in the Apennines close to Forlì, where the four models agree. The area near Parma and Garfagnana have probabilities slightly higher than the surrounding regions (Fig. 8).

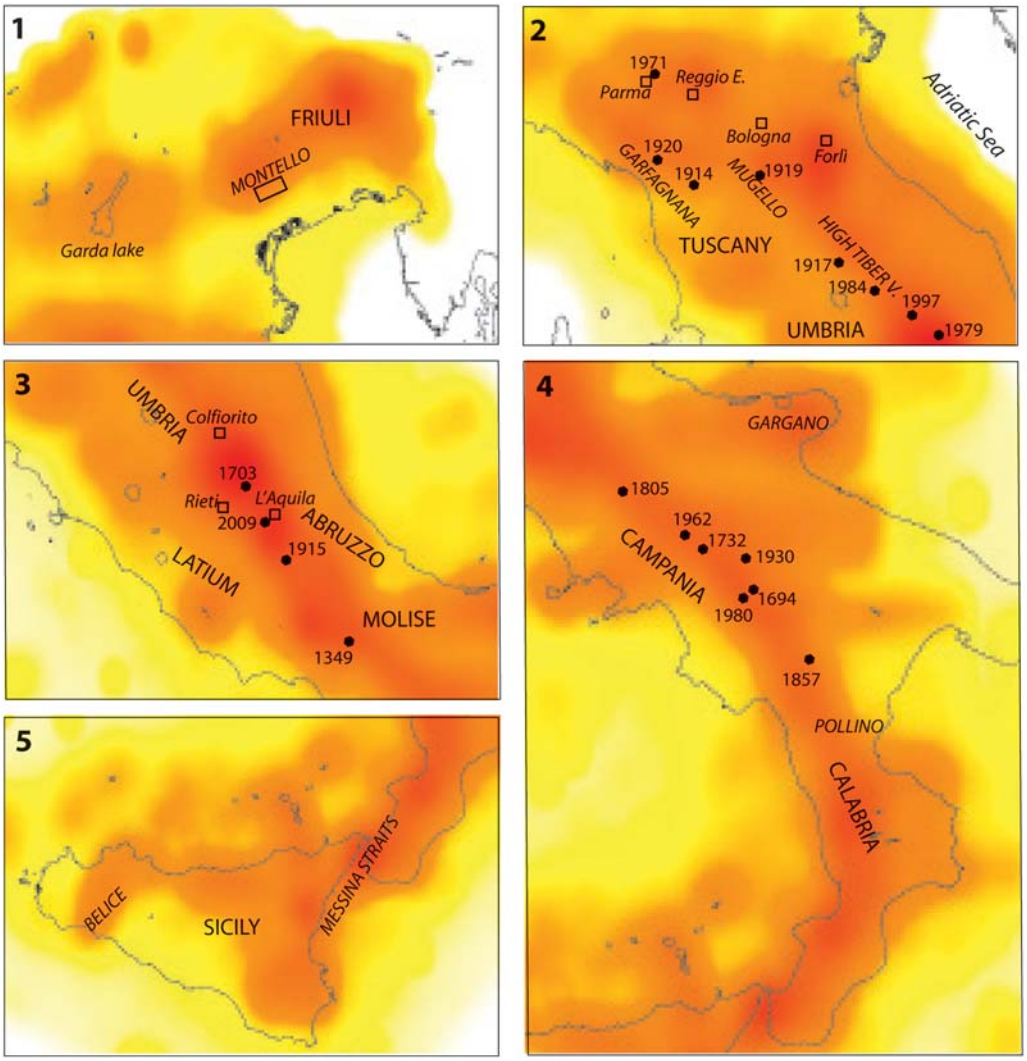

Figure 10. Larger-scale view of selected areas from Figure 8, discussed in the section A Closer Look at the Best EO Model. Locations, cities, and earthquakes mentioned in the text are shown. Open squares, city locations; full circles, major earthquakes.

Central Apennines

This Apennine range (area 3 in Figs. 8 and 10) shows a generally higher EO probability compared to the adjacent peri-Tyrrhenian and peri-Adriatic areas. All models consistently define the axial belt of the central Apennines (south of Colfiorito to L'Aquila) as one of the most hazardous regions in Italy. This derives from both the frequent M_L 5–6 earthquakes that have occurred in the last few decades and from some large ($M_L \sim 6.5$–7) infrequent events (1703, 1915). Notably, this region also contains the best active faults mapping in Italy, leading to the development of some local hazard models (e.g., Pace *et al.*, 2006; Akinci *et al.*, 2010) that use more geological information with respect to the national models. According to some time-dependent models, after the L'Aquila earthquake of 2009, the probabilities of events in adjacent regions have increased. This includes the area called "reatino" (from Rieti, Latium) where the microseismicity rate has increased significantly after the L'Aquila event.

In the southern part of the central Apennines, we note a region with a high probability of occurrence (southern Latium, Abruzzo, and Molise), where a long seismic swarm started after the L'Aquila earthquake, and some major active faults are known. These faults have been silent for many centuries, at least since the large 1349 earthquake.

Gargano

The Gargano promontory appears to be an area of high probability (Fig. 8 and area 4 in Fig. 10) mainly due to the seismic sources used in HAZFX-BPT and (indirectly) in MPS04. On the other hand, it also represents an area where models show some uncertainty (Fig. 9). This may partially be due to the fact that, as described earlier in this paper, the knowledge of the seismic sources reported in DISS3 is limited, especially for repeat times and the elapsed time since the last event. Interestingly, GPS data (Avallone *et al.*, 2010) show low strain accumulation in this area and suggest that the tectonic loading of this area is low. This could imply that

repeat times of large earthquakes in this area are very long, likely more than 1000 years.

Southern Apennines and Calabria

Apart from the southern Latium area described previously, the whole southern Apennine region shows relatively low probabilities in the AVE model for the next 10 years (area 4 in Figs. 8 and 10). However, Figure 9 shows that this is the region with the largest uncertainty, due to the sizable differences among the four models. Actually, both the standard MPS04 map and the fault-based HAZFX-BPT model show high probabilities in local patches of the seismic belt. The averaging tends to lower these values, due to a low recent seismic activity that tends to diminish the expected number of events for the DBM model. It is worth noting that many large earthquakes (M_L 6.5–7) of past centuries have occurred in this sector of the Apennines (Matese 1805; Irpinia 1694, 1732, 1930, 1962, and 1980; Val d'Agri, 1857) and that the knowledge of active faults in this area is incomplete and debated. For instance, according to DISS3, the 1857 earthquake also ruptured a northern fault segment (see Burrato and Valensise, 2008) that, according to other studies, is a kind of seismic gap (Moro *et al.*, 2007). These two different interpretations lead to two completely different scenarios in terms of probability of occurrence, and this also validates the large uncertainty shown by the AVE model. A similar circumstance occurs for the Molise sector, where high probabilities are found by HAZFX-BPT and the AVE model has lower values.

More to the south, Calabria shows some areas with high probabilities in the AVE model, including the Messina Straits, for which all models agree. It is worth noting that the Pollino area (at the boundary between the southern Apennines and Calabria), often depicted as a seismic gap area (Cinti *et al.*, 2002), does not show up as a critical area in the AVE model. Accordingly, GPS data does not show a significant deformation in the area south of the 1857 fault (Avallone *et al.*, 2010).

Sicily

Apart from the Messina Straits region, described with Southern Apennines and Calabria, eastern Sicily shows medium probability values in all four models (area 5 in Figs. 8 and 10). It is worth remarking that at least two large historical earthquakes did occur in this area, but knowledge of the active faults in this area is still poor, with some of them probably hidden in the Ionian Sea. Central and western Sicily, where some moderate events occurred in the Tyrrhenian offshore and in the Belice area (as in 1968), do not show high probabilities for the next decade.

Discussion and Final Remarks

The main goal of the paper is to present an original earthquake occurrence map of Italy for the next 10 years. The map benefits from recent activities, like the start of the CSEP experiment in Italy and the results of the S2 project funded by INGV/Dipartimento della Protezione Civile (2009–2010), in which different earthquake occurrence models have been tested to check their retrospective reliability and skill to reproduce the occurrence of past large earthquakes in Italy. The golden rule for evaluating EO models is through real prospective tests (e.g., Jordan, 2006). Nonetheless, prospective tests take time to be accomplished. In Italy, the first results of the CSEP experiment begun in 2009 will not be available until 2014. In order to provide a first reliable ranking of the forecasting performances of the available EO models, we have run a retrospective CSEP-type analysis (e.g., Werner *et al.*, 2010; Woessner *et al.*, 2011). Specifically, the retrospective analysis presented in this paper is applied to the EO models available in a specific format to make the models comparable. Many other models are already available, but they cover only part of the Italian territory (e.g., Pace *et al.*, 2006), have a nonhomogeneous format (e.g., Rotondi, 2010), and their retrospective tests (see Werner *et al.*, 2010) are not straightforwardly comparable to the ones presented here. Nevertheless, the upcoming prospective CSEP results, to include other tectonic areas, will certainly have a major influence on this type of initiative.

The final map is a weighted average of the EO models that have produced reliable results in the past decades for the entire Italian territory. In principle, the models may be ranked according to their retrospective skill, but in the present cases all reliable models seem to be almost equally skilled. We also provide a map of uncertainty that is of fundamental importance for decision makers to select optimal risk-mitigation actions. The average map, being time dependent, requires regular updates and is to include future developments of models, knowledge, and data. In this respect, we emphasize that the models used here only marginally account for fault distribution, and other information, like the deformation field, is not included at all. We think that significant improvements in the skill of the forecasts might be achieved using new reliable geological information. In particular, geological information may be of paramount importance to constrain significantly better the spatial distribution of large earthquakes and the frequency–magnitude distribution of each area. At the present state of knowledge, although representing a unique effort and useful basis, the present nationwide reference compilation of $M \geq 5.5$ seismic sources (DISS3; Data and Resources) is still largely incomplete and mostly contains sources that produced an earthquake in historical times. Sources that may be responsible for future ones are possibly lacking. Besides the identification of the sources, many efforts must still be made to estimate the source parameters (slip rates, elapsed time, frequency–magnitude distribution, etc.) and understand how the geometry observed in the field may be used to forecast future large events.

Data and Resources

Legge 24 giugno 2009, n. 77, articolo 11: Interventi per la prevenzione del rischio sismico (Law 77, article 11) is

available from http://www.camera.it/parlam/leggi/09077l.htm (last accessed March 2012). Funding from this law has been available since 2010.

Plots were made using the Generic Mapping Tools, version 4.2.0 (www.soest.hawaii.edu/gmt, last accessed April 2007; Wessel and Smith, 1998). Three out of six earthquake occurrence models have used the Italian fault database DISS3, freely available at http://diss.rm.ingv.it/diss/ (last accessed December 2011). The CPTI08 seismic catalog was used for the testing phase and Collaboratory for the Study of Earthquake Predictability (CSEP) experiment (provided by the Catalogo Parametrico dei Terremoti Italiani [CPTI] Working Group) and integrated for the last part of the testing period with the data of the Centroid Moment Tensor (CMT) catalog (http://www.globalcmt.org/CMTsearch.html, last accessed January 2011). The CPTI08 catalog has been released only for CSEP activities. The new catalog (CPTI11, Rovida *et al.*, 2011) is available at http://emidius.mi.ingv.it/CPTI11/ (last accessed March 2012). Tectonic zonation is from http://zonesismiche.mi.ingv.it/documenti/App2.pdf (last accessed March 2004; in Italian).

Acknowledgments

This work has been carried out in the framework of an initiative of the Dipartimento della Protezione Civile (DPC) and the Italian government, aimed at reducing the seismic risk during the 2011–2017 time period, with funding to retrofit vulnerable buildings. We are grateful to all Istituto Nazionale di Geofisica e Vulcanologia (INGV)–DPC S2 project participants and, in particular, to the co-coordinators, E. Faccioli, Dario Albarello, Bruno Pace, and Laura Peruzza. Discussions during the project meetings have been very helpful and shaped some of the ideas developed in this paper. Specific thanks go to G. Falcone, M. Murru, and R. Console for providing us with the retrospective forecasts of the LTST model. We also thank L. Malagnini and L. Valensise for their internal review that help to improve this paper. N. D'Agostino, G. Selvaggi, and A. Piersanti have participated in some of the activities of the INGV team. We also thank two anonymous reviewers and the associate editor for helping to improve the manuscript.

References

Akinci, A, F. Galadini, D. Pantosti, M. Petersen, L. Malagnini, and D. Perkins (2009). Effect of time dependence on probabilistic seismic-hazard maps and deaggregation for the central Apennines, Italy, *Bull. Seismol. Soc. Am.* **99,** 585–610.

Akinci, A., D. Perkins, A. M. Lombardi, and R. Basili (2010). Uncertainties in the estimation of the probability of occurrence of strong earthquakes from individual seismological sources in the Apennines, Italy, *J. Seismol.* **14,** 95–117.

Akinci, A. (2010). HAZGRIDX: Earthquake forecasting model for $M_L \geq 5.0$ earthquakes in Italy based on spatially smoothed seismicity, *Ann. Geophys.* **53,** no. 3, 51–61, doi 10.4401/ag-4811.

Albarello, D., and V. D'Amico (2008). Testing probabilistic seismic hazard estimates by comparison with observations: An example in Italy, *Geophys. J. Int.* **175,** 1088–1094.

Avallone, A., G. Selvaggi, E. D'Anastasio, N. D'Agostino, G. Pietrantonio, F. Riguzzi, E. Serpelloni, M. Anzidei, G. Casula, and G. Cecere *et al.* (2010). The RING network: Improvements of a GPS velocity field in the central Mediterranean, *Ann. Geophys.* **53,** 39–54, doi 10.4401/ag-4549.

Basili, R., G. Valensise, P. Vannoli, P. Burrato, U. Fracassi, S. Mariano, M. M. Tiberti, and E. Boschi (2008). The Database of Individual Seismogenic Sources (DISS), version 3: Summarizing 20 years of research on Italy's earthquake geology, *Tectonophysics* **453,** 20–43, doi 10.1016/j.tecto.2007.04.014.

Burrato, P., and G. Valensise (2008). Rise and fall of a hypothesized seismic gap: Source complexity in the M_w 7.0 16 December 1857 southern Italy earthquake, *Bull. Seismol. Soc. Am.* **98,** 139–148, doi 10.1785/0120070094.

Cinti, F. R., M. Moro, D. Pantosti, L. Cucci, and G. D'Addezio (2002). New constraints on the seismic history of the Castrovillari fault in the Pollino gap (Calabria, southern Italy), *J. Seismol.* **6,** 199–217, doi 10.1023/A:1015693127008.

Cinti, F. R., L. Faenza, W. Marzocchi, and P. Montone (2004). Probability map of the next $M \geq 5.5$ earthquakes in Italy, *Geochem. Geophys. Geosys.* **5,** Q11003, doi 10.1029/2004GC000724.

Console, R., M. Murru, and A. M. Lombardi (2003). Refining earthquake clustering models, *J. Geophys. Res.* **108,** 2468 pp.

D'Agostino, N., D. Cheloni, S. Mantenuto, G. Selvaggi, A. Michelini, and D. Zuliani (2005) Strain accumulation in the southern Alps (NE Italy) and deformation at the northeastern boundary of Adria observed by CGPS measurements, *Geophys. Res. Lett.* **32,** L19306, doi 10.1029/2005GL024266.

Daley, D. J., and D. Vere-Jones (2004). Scoring probability forecasts for point processes: The entropy score and information gain, in *Stochastic Methods and Their Applications*, J. Gani and E. Seneta (Editors), special issue, *J. Appl. Prob.* **41A,** 297–312.

Ellsworth, W. L., M. V. Matthews, R. M. Nadeau, S. P. Nishenko, P. A. Reasenberg, and R. W. Simpson (1999). A physically-based earthquake recurrence model for estimation of long term earthquake probabilities, *U. S. Geol. Surv. Open-File Rept. 99-522*, 23 pp.

Faenza, L., W. Marzocchi, and E. Boschi (2003). A nonparametric hazard model to characterize the spatio-temporal occurrence of large earthquakes; an application to the Italian catalogue, *Geophys. J. Int.* **155,** 521–531.

Faenza, L., W. Marzocchi, P. Serretti, and E. Boschi (2008). On the spatio-temporal distribution of M 7.0+ worldwide seismicity with a non-parametric statistics, *Tectonophysics* **449,** 97–104, doi 10.1016/j.tecto.2007.11.066.

Faenza, L., and W. Marzocchi (2010). The proportional hazard model applied to the CSEP testing area in Italy, *Ann. Geophys.* **53,** 77–84.

Falcone, G., R. Console, and M. Murru (2010). Short-term and long-term earthquake occurrence models for Italy: ETES, ERS and LTST, *Ann. Geophys.* **53,** 41–50.

Field, E. H., D. D. Johnson, and J. F. Dolan (1999). A mutually consistent seismic-hazard source model for southern California, *Bull. Seismol. Soc. Am.* **89,** 559–578.

Field, E. H., T. E. Dawson, K. R. Felzer, A. D. Frankel, V. Gupta, T. H. Jordan, T. Parsons, M. D. Petersen, R. S. Stein, R. J. Weldon III, and C. J. Wills (2009). Uniform California Earthquake Rupture Forecast, version 2 (UCERF 2), *Bull. Seismol. Soc. Am.* **99,** 2053–2107, doi 10.1785/0120080049.

Frankel, A. (1995). Mapping seismic hazard in the central and eastern United States, *Seismol. Res. Lett.* **66,** 8–21.

Gerstenberger, M., S. Wiemer, L. M. Jones, and P. A. Reasenberg (2005). Real-time forecasts of tomorrow's earthquakes in California, *Nature* **435,** 328–331.

Gruppo di Lavoro CPTI (2004). Catalogo Parametrico dei Terremoti Italiani, versione 2004 (CPTI04), Istituto Nazionale di Geofisica e Vulcanologia, Bologna, http://emidius.mi.ingv.it/CPTI04/ (last accessed October 2011).

Gruppo di Lavoro MPS (2004). Redazione della mappa di pericolosità sismica prevista dall'Ordinanza PCM 3274 del 20 marzo 2003, *rapporto conclusivo per il Dipartimento della Protezione Civile*, INGV, Milano-Roma, aprile 2004, 65 pp. + 5 appendici (in Italian), http://zonesismiche.mi.ingv.it/documenti/rapporto_conclusivo.pdf (last accessed October 2008).

Jordan, T. H. (2006). Earthquake predictability, brick by brick, *Seismol. Res. Lett.* **77,** 3–6.

Jordan, T. H., and L. M. Jones (2010). Operational earthquake forecasting: Some thoughts on why and how, *Seismol. Res. Lett.* **81,** 571–574, doi 10.1785/gssrl.81.4.571.

Jordan, T. H., Y.-T. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, G. Papadopoulos, G. Sobolev, K. Yamaoka, and J. Zschau (2011). Operational earthquake forecasting: State of knowledge and guidelines for implementation, *Ann. Geophys.* **54,** 315–391, doi 10.4401/ag-5350.

Kagan, Y. Y., and D. D. Jackson (1991). Long-term earthquake clustering, *Geophys. J. Int.* **104,** 117–133.

Kagan, Y. Y., and D. D. Jackson (1994). Long-term probabilistic forecasting of earthquakes, *J. Geophys.Res.* **99,** 13,685–13,700.

Kagan, Y. Y., and D. D. Jackson (2000). Probabilistic forecasting of earthquakes, *Geophys. J. Int.* **143,** 438–453.

Lombardi, A. M., and W. Marzocchi (2007). Evidence of clustering and nonstationarity in the time distribution of large worldwide earthquakes, *J. Geophys. Res.* **112,** no. B02303, doi 10.1029/2006JB004568.

Lombardi, A. M., and W. Marzocchi (2009). Double branching model to forecast the next $M \geq 5.5$ earthquakes in Italy, *Tectonophysics* **475,** 514–523, doi 10.1016/j.tecto.2009.06.014.

Lombardi, A. M., and W. Marzocchi (2010a). A stochastic point process for daily earthquake forecasting; application to the Italian seismicity, *Ann. Geophys.* **53,** 155–164.

Lombardi, A. M., and W. Marzocchi (2010b). The double branching model (DBM) applied to forecasting Italian seismicity in CSEP experiment, *Ann. Geophys.* **53,** 31–39.

Lombardi, A. M., and W. Marzocchi (2010c). The assumption of Poisson seismic rate variability in CSEP/RELM experiments, *Bull. Seismol. Soc. Am.* **100,** 2293–2300.

Marzocchi, W., and A. M. Lombardi (2008). A double branching model for earthquake occurrence, *J. Geophys. Res.* **113,** no. B08317, doi 10.1029/2007JB005472.

Marzocchi, W., D. Schorlemmer, and S. Wiemer (2010). Preface in *An Earthquake Forecast Experiment in Italy*, special issue, *Ann. Geophys.* **53,** III–VIII.

Marzocchi, W., and J. D. Zechar (2011). Earthquake forecasting and earthquake prediction: Different approaches for obtaining the best model, *Seismol. Res. Lett.* **82,** 442–448.

Matthews, M. V., W. L. Ellsworth, and P. A. Reasenberg (2002). A Brownian model for recurrent earthquakes, *Bull. Seismol. Soc. Am.* **92,** 2233–2250, doi 10.1785/0120010267.

Meletti, C., F. Galadini, G. Valensise, M. Stucchi, R. Basili, S. Barba, G. Vannucci, and E. Boschi (2008). A seismic source zone model for the seismic hazard assessment of the Italian territory, *Tectonophysics* **450,** 85–108.

Moro, M., L. Amicucci, F. R. Cinti, F. Doumaz, P. Montone, S. Pierdominici, M. Saroli, S. Stramondo, and B. Di Fiore (2007). Surface evidence of active tectonics along the Pergola-Melandro fault: A critical issue for the seismogenic potential of the southern Apennines, Italy, *J. Geodyn.* **44,** 19–32.

Mulargia, F. (2010). Opinion: Extending the usefulness of seismic hazard studies, *Seismol. Res. Lett.* **81,** 423–424.

Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes, *J. Am. Stat. Assoc.* **83,** 9–27.

Pace, B., L. Peruzza, G. Lavecchia, and P. Boncio (2006). Layered seismogenic source model and probabilistic seismic-hazard analyses in central Italy, *Bull. Seismol. Soc. Am.* **96,** 107–132, doi 10.1785/0120040231.

Parsons, T. (2002). Global Omori law decay of triggered earthquakes: Large aftershocks outside the classical aftershock zone, *J. Geophys. Res.* **107,** 2199, doi 10.1029/2001JB000646.

Reasenberg, P. A., and L. M. Jones (1989). Earthquake hazard after a mainshock in California, *Science* **243,** 1173–1176.

Rhoades, D. A., and M. C. Gerstenberger (2009). Mixture models for improved short-term earthquake forecasting, *Bull. Seismol. Soc. Am.* **99,** 636–646.

Rotondi, R. (2010). Bayesian nonparametric inference for earthquake recurrence time distributions in different tectonic regimes, *J. Geophys. Res.* **115,** no. B01302, doi 10.1029/2008JB006272.

Rovida, A., R. Camassi, P. Gasperini, and M. Stucchi (Editors) (2011). CPTI11, the 2011 version of the Parametric Catalogue of Italian Earthquakes, http://emidius.mi.ingv.it/CPTI (last accessed March 2012), Milano, Bologna.

Schorlemmer, D., and M. C. Gerstenberger (2007). RELM testing center, *Seismol. Res.Lett.* **78,** 30–36.

Schorlemmer, D., M. C. Gerstenberger, S. Wiemer, D. D. Jackson, and D. A. Rhoades (2007). Earthquake likelihood model testing, *Seismol. Res. Lett.* **78,** 17–29.

Schorlemmer, D., A. Christophersen, A. Rovida, F. Mele, M. Stucchi, and W. Marzocchi (2010). An earthquake forecast experiment in Italy, *Ann. Geophys.* **53,** 1–9.

Schorlemmer, D., F. Mele, and W. Marzocchi (2010). A completeness analysis of the National Seismic Network of Italy, *J. Geophys. Res.* **115,** no. B04308, doi 10.1029/2008JB006097.

Stein, R. S., A. A. Barka, and J. H. Dieterich (1997). Progressive failure on the north Anatolian fault since 1939 by earthquake stress triggering, *Geophys. J. Int.* **128,** 594–604.

van Stiphout, T., S. Wiemer, and W. Marzocchi (2010). Are short-term evacuations warranted? Case of the 2009 L'Aquila earthquake, *Geophys. Res. Lett.* **37,** L06306, doi 10.1029/2009GL042352.

Weichert, D. H. (1980). Estimation of the earthquake recurrence parameters for unequal observation periods for different magnitudes, *Bull. Seismol. Soc. Am.* **70,** 1337–1346.

Werner, M., J. D. Zechar, W. Marzocchi, and S. Wiemer (2010). Retrospective evaluation of the five-year and ten-year CSEP-Italy earthquake forecasts, *Ann. Geophys.* **53,** 11–30.

Wessel, P., and W. H. F. Smith (1998). New, improved version of the Generic Mapping Tools released, *Eos Trans. AGU* **79,** 579.

Woessner, J., S. Hainzl, W. Marzocchi, M. J. Werner, A. M. Lombardi, F. Catalli, B. Enescu, M. Cocco, M. C. Gerstenberger, and S. Wiemer (2011). A retrospective comparative forecast test on the 1992 Landers sequence, *J. Geophys. Res.* **116,** no. B05305, doi 10.1029/2010JB007846.

Woo, G. (2010). Operational earthquake forecasting and risk management, *Seismol. Res. Lett.* **81,** 778–782.

Zechar, J. D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P. J. Maechling, and T. H. Jordan (2010). The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science, *Concurrency Comput. Pract. Ex.* **22,** 1836–1847, doi 10.1002/cpe.1519.

Zechar, J. D., M. C. Gerstenberger, and D. A. Rhoades (2010). Likelihood-based tests for evaluating space–rate–magnitude earthquake forecasts, *Bull. Seismol. Soc. Am.* **100,** 1184–1195.

Appendix

The Test of the Earthquake Occurrence Models

The strategy adopted here is along the lines of the testing procedures established by Regional Earthquake Likelihood Models (RELM)/Collaboratory for the Study of Earthquake Predictability CSEP projects (Schorlemmer and Gerstenberger, 2007; Schorlemmer *et al.*, 2007), introducing specific changes in order to optimize the methodology to test EO models suitable for seismic-hazard assessment. The main difference with the CSEP tests is the forecast format. Here, the models produce the probability to have 0, 1, 2, and 3 or more earthquakes in a defined space–time–magnitude bin, while CSEP experiments use the expected number of earthquakes per bin.

We perform two evaluations: (a) a goodness-of-fit test of consistency of each EO model with observed data, and (b) a comparison test of all pairs of EO models consistent with observations, for quantitatively evaluating the relative performance. Note that sharing the same RELM/CSEP philosophy leads to the possibility of comparing the results obtained in the S2 project with other similar initiatives carried out under the CSEP umbrella. All tests will be applied to a well-defined testing area reported in Figure 1 (Schorlemmer, Christophersen, *et al.*, 2010; Schorlemmer, Mele, and Marzocchi, 2010). Both area and grid are the same as were used in the Italian CSEP experiment that started in August 2009.

The hypothesis under testing is expressed as a forecast of earthquake occurrences per specified bins $B_{s=(i,j,k)} = [T_i, S_j, M_k]$ in time (T_i), space (S_j), and magnitude (M_k) of the whole spatiotemporal–magnitude window under testing. For each bin, we define a binary variable e_s that assumes either the value of 0 or, in the case that at least one earthquake occurs inside the bin B_s, the value of 1. Hereinafter, we distinguish the terms "earthquake" and "event," the latter representing the occurrence of one or more earthquakes in a specific bin. Each EO model (named H_m) provides a probability $P(e_s|H_m)$, that is, the probability to have the outcome e_s (at least one or no earthquake) for any specific bin.

The test of consistency with observations and relative performance of EO models is based on a direct comparison of probabilistic forecasts (given by each EO model) and observed occurrences in a testing time interval, independent of the time period (learning) used to set up each EO model under testing.

The Number-of-Events Test

The number-of-events test (*N*-test) is based on the comparison of expected and observed frequency of events (Schorlemmer *et al.*, 2007). The null hypothesis is that these two numbers come from the same distribution. We formalize this procedure as follows.

By following the procedure proposed by Albarello and D'Amico (2008), for each EO model H_m, we simulate L synthetic realizations of events e_s in spatiotemporal–magnitude space, according to the discrete probabilistic function $P(e_s|H_m)$. Specifically, each synthetic dataset consists of a sequence of stochastic realizations of variable e_s, generated by the probability function $P(e_s|H_m)$. Then, for each simulation we compute the number of sites that have experienced one or more earthquakes in each temporal bin T_i, $(\hat{N}_1^{i,H_m}, \ldots, \hat{N}_L^{i,H_m})$, by summing the values of variable e_s in all spatiomagnitude bins $[S_j, M_k]$. Finally, we evaluate an eventual disagreement between the observed number N_i^* (i.e., the number of sites in which at least one earthquake occurred during the time period T_i) and the theoretical distribution $(\hat{N}_1^{i,H_m}, \ldots, \hat{N}_L^{i,H_m})$ by computing the quantile score $\delta_i^{H_m}$ given by

$$\delta_i^{H_m} = \frac{|\hat{N}_k^{i,H_m} / \hat{N}_k^{i,H_m} \le N_i^*; k = 1, \ldots, L|}{L}. \tag{A1}$$

Model H_m can be rejected at significance level α for time bin T_i if $\delta_i^{H_m}$ is lower than $\alpha/2$ or larger than $1 - \alpha/2$.

Note that, unlike the *N*-test developed for RELM/CSEP evaluation (Schorlemmer *et al.*, 2007), this test does not check the number of earthquakes that have occurred in each temporal bin, but considers events as previously defined (i.e., the reliability to have at least one earthquake). This implicitly means that we focus our attention on the reliability of having earthquakes in each interval time, regardless of the possibility of having more than one event. In this way we neglect the physical mechanisms, such as the co-seismic stress transfer, responsible for the space–time short-term clustering of earthquakes. This choice is particularly suitable for seismic-hazard models, for which the emphasis is more on the occurrence (or not) of large earthquakes rather than on the exact number occurring in a fixed time period. This test does not account for the spatial distribution of the forecasts and earthquakes; this means that a model may pass the *N*-test if it accurately predicts the number of events but not their spatial location. The spatial location will be considered by the *L*-test (see The Likelihood Test).

We emphasize that the use of $P(e_s|H_m)$ instead of the rate of events, as in CSEP experiments, also has another important advantage. Specifically, we do not need to assume that the number of events in each bin follows a Poisson distribution, like in Schorlemmer *et al.* (2007). As a matter of fact, this assumption may lead to significant biases in the results, in particular for the time-dependent clustering models (see Lombardi and Marzocchi, 2010c).

The Likelihood Test

Likelihood computations are a very common statistical tool for the evaluation of a statistical model. Unlike the *N*-test, the likelihood test (*L*-test) permits evaluation of the agreement of magnitude–spatial features, beside the temporal evolution, between forecasts and observations.

The formulation of the *L*-test, presented in the following discussion, assumes the statistical independence of events e_s (see Schorlemmer *et al.*, 2007).

For each time bin T_i, the likelihood of independent events e_s for the EO model H_m is given by

$$L_{i,H_m} = \prod_{(s \in B^*)} P(e_s = 1|H_m) \prod_{(s \in \tilde{B}^*)} [1 - P(e_s = 1|H_m)], \tag{A2}$$

or, in logarithmic scale, by

$$l_{i,H_m} = \sum_{(s \in B^*)} \log[P(e_s = 1|H_m)] + \sum_{(s \in \tilde{B}^*)} \log[1 - P(e_s = 1|H_m)], \tag{A3}$$

where B^* is the set of spatiotemporal–magnitude bins in which at least one earthquake occurs and $\tilde{B}^*$ is the complementary set, consisting of bins with no earthquakes.

By following Kagan and Jackson (1994), we compare the time history of observed log-likelihood values l^*_{i,H_m} with the corresponding values $(\hat{l}^l_{i,H_m}, \ldots, \hat{l}^l_{i,H_m})$ computed on L synthetic stochastic realizations of variables e_s, simulated in agreement with model H_m and similarly to the N-test. Model H_m can be quantitatively tested by quantile

$$\gamma_i^{H_m} = \frac{|\hat{l}^k_{i,H_m} / \hat{l}^k_{i,H_m} \leq l^*_{i,H_m}; k = 1, \ldots, L|}{L} \tag{A4}$$

and rejected at a significance level α, if $\gamma_i^{H_m} < \alpha$ (the observed log-likelihood is much smaller than what would be expected by the model). If $\gamma_i^{H_m}$ is very high, the observed likelihood is much higher than the mean value expected by the model. Because this situation leaves room for the possibility that a model is good but smooths its forecast too much (see Schorlemmer *et al.*, 2007, for details), high quantile values are not considered for model rejection (one-tailed test). The L-test takes into account both the spatial component of forecasts and the rate (Zechar, Gerstenberger, and Rhoades, 2010). It is generally true that the simulated log-likelihoods $\hat{l}^k_{i,H_m}$ are greater or smaller than the observed log-likelihood l^*_{i,H_m}, when the numbers of simulated events are more or less than the observed events, respectively. This allows the forecasts to pass the L-test, whenever the spatial distribution of forecasts fails to predict the occurrence of real events (see Zechar, Gerstenberger, and Rhoades, 2010).

It is worth highlighting a main assumption of this test—that there is spatial independence among bins. In other words, it is assumed that the occurrence of earthquakes in a specific bin during the testing phase does not alter the conditional probability in adjacent bins. This assumption may be debatable, in particular for the occurrence of small-magnitude earthquakes. Here, we assume that the effect on the L-test may be considered negligible when large-magnitude earthquakes are considered.

Likelihood-Ratio Test

The previous two tests permit the evaluation of the goodness of fit of an EO model using observations. By performing these tests on two competitive models, H_1 and H_2, it cannot be ruled out that both models are consistent with the observed data. In this case, it is advisable to investigate their comparative performance, using a further test.

A measure that can be used to compare the performance of two models on a testing dataset is the likelihood-ratio test, outlined by Kagan and Jackson (1994). This measure consists of a pairwise comparison between models that are consistent with both the N- and L-tests. The statistic of the likelihood-ratio test is given by the difference in log-likelihoods (computed on whole temporal intervals) of two competitive forecasts, H_1 and H_2:

$$R_{12} = l_{H_1} - l_{H_2}. \tag{A5}$$

This indicates which model fits the data better: if $R_{12} > 0$, then H_1 provides more likely forecasts; if $R_{12} < 0$, then H_2 performs better. Supposing that $R_{12} > 0$, we have to judge the statistical significance of a better performance by model H_1. Originally, the likelihood-ratio test was conceived for nested models. In comparing earthquake forecast models, most of the times the models are not nested, complicating the interpretation of the test. Schorlemmer *et al.* (2007) suggested one possible strategy to test the significance of the R-test results. Specifically, L synthetic realizations of events e_s are produced assuming that H_1 is the right model and the likelihood ratios $(\hat{R}_{12}^{H_1,1}, \ldots, \hat{R}_{12}^{H_1,L}$ are calculated for the simulated catalogs. Then, the same procedure is followed, assuming that H_2 is the right model and the likelihood-ratios $(\hat{R}_{21}^{H_2,1}, \ldots, \hat{R}_{21}^{H_2,L})$ are calculated for the simulated catalogs. Using this procedure, the model H_1 is significantly better than H_2 if the fraction of synthetic likelihood-ratios $(\hat{R}_{12}^{H_1,1}, \ldots, \hat{R}_{12}^{H_1,L})$ is large (for instance, higher than 0.10) and the fraction of synthetic likelihood-ratios $(\hat{R}_{21}^{H_2,1}, \ldots, \hat{R}_{21}^{H_2,L})$ is small (for instance, less than 0.01).

Istituto Nazionale di Geofisica e Vulcanologia
via di Vigna Murata 605
00143 Roma
Italy

Manuscript received 26 May 2011