Seismological Research Letters
Prospective CSEP evaluation of 1-day, 3-month, and 5-year earthquake forecasts for Italy
--Manuscript Draft--

Manuscript Number:	SRL-D-18-00031R2
Full Title:	Prospective CSEP evaluation of 1-day, 3-month, and 5-year earthquake forecasts for Italy
Article Type:	Focus Section - CSEP: New Results and Future Directions
Corresponding Author:	Matteo Taroni INGV Rome, Rome ITALY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	INGV
Corresponding Author's Secondary Institution:	
First Author:	Matteo Taroni
First Author Secondary Information:	
Order of Authors:	Matteo Taroni
	Warner Marzocchi
	Daniel Schorlemmer
	Maximilian Werner
	Stefan Wiemer
	Jeremy Douglas Zechar
	Lukas Heiniger
	Fabian Euchner
Order of Authors Secondary Information:	
Manuscript Region of Origin:	ITALY
Suggested Reviewers:	Matt Gerstenberger m.gerstenberger@gns.cri.nz
	David Rhoades D.Rhoades@gns.cri.nz
	Anne Strader astrader@ucla.edu
	David Jackson djackson@g.ucla.edu
	Takahiro Omi omi@sat.t.u-tokyo.ac.jp
	Jiancang Zhuang zhuangjc@ism.ac.jp
	Naoshi Hirata
Opposed Reviewers:	

Prospective CSEP evaluation of 1-day, 3-month, and 5-year earthquake forecasts for Italy

M. Taroni[1], W. Marzocchi[1], D. Schorlemmer[2], M. J. Werner[3], S. Wiemer[4], J.D. Zechar[5], L. Heiniger[4], F. Euchner[4].

1- Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy.
2- GFZ Potsdam, Potsdam, Germany
3- University of Bristol, United Kingdom.
4- ETH Zurich, Switzerland
5- AXIS Capital, Zürich, Switzerland

Abstract

In 2009, the global Collaboratory for the Study of Earthquake Predictability (CSEP) launched three experiments to forecast the distribution of earthquakes in Italy in the subsequent five years. CSEP solicited forecasts for seismicity tomorrow, in the next three months, and for the entire five years. In those 5 years, INGV recorded 83 target earthquakes with local magnitude $3.95 \leq M < 4.95$, and 14 larger shocks. The results show that: 1-day forecasts are consistent with the number and magnitudes of the target earthquakes, and one version of the ETAS model is also consistent with the spatial distribution; ensemble forecasts, which we created for the 1-day experiment, are consistent with the number, locations, and magnitudes of the target earthquakes, and they perform as well as the best model; none of the 3-month time-independent models produce consistent forecasts; the best 5-year models account for the fault distribution and the historical seismicity; and 5-year models based on instrumental seismicity and b-value spatial variation show poor forecasting performance.

Introduction

The Collaboratory for the Study of Earthquake Predictability (CSEP; Jordan, 2006; Zechar et al., 2010a) is an international infrastructure that promotes assessing scientific hypotheses about earthquake occurrence within a standardized environment and following community-endorsed procedures and metrics. CSEP conducts prospective (i.e., zero degrees of freedom) and reproducible experiments, which compare the forecasts of a set of models automatically running in a testing center with the observed seismicity in a testing region (Schorlemmer and Gerstenberger, 2007). To carry out reproducible and transparent experiments in a controlled environment, CSEP defines, *a priori*, unambiguous rules, such as: the definition of the testing region, characterized by high-quality seismic recordings; an exact description of the expected forecast format; and exact definition of the earthquake data (from an independent and authoritative source). The objective of these experiments is to quantify, for each forecast model, predictive skill (relative performance of a model with respect to others) and the consistency with the observations, with a broader goal of evaluating models and their underlying hypotheses about earthquakes.

In 2009, Italy joined California, Japan, New Zealand, the Western Pacific, and the entire globe as a CSEP testing region. This was feasible thanks to the high quality of seismic monitoring by the Istituto Nazionale di Geofisica e Vulcanologia (INGV) (Schorlemmer et al., 2010a; 2010b). The European CSEP testing center at ETH Zurich is conducting three prospective CSEP-Italy experiments, which began on August 1, 2009 with an expected first evaluation after five years (Marzocchi et al., 2010). Each model forecasts the expected number of target earthquakes (i.e., earthquakes with magnitude above a pre-defined threshold) in small time-space-magnitude bins covering the CSEP Italy testing region (Schorlemmer et al., 2010a; 2010b). The forecasts are based on—and tested against—observations provided by the official seismic bulletin of the INGV. Each CSEP-Italy experiment (or testing class) is distinguished by its forecast horizons:

1. 5-year forecasts (19 models submitted): models forecast local magnitude M = 4.95 and above. The experiment started January 1, 2010 and all forecasts were submitted to the testing center before this day and retrospectively evaluated by Werner et al. (2010a) for so-called "sanity checks."
2. 3-month forecasts (3 models submitted): models forecast M = 3.95 and above, and the duration of each time bin is three months. The experiment started October 1, 2009. The models were implemented in the European CSEP testing center as software codes.

3. 1-day forecasts (4 models submitted): models forecast M = 3.95 and above, and the duration of each time bin is one day. The experiment started August 1, 2009, at midnight UTC. The 1-day models were also installed at the testing center as software codes.

Here, we show the results of these first experiments and discuss the scientific lessons learned. Specifically, we explore the strengths and weaknesses of the forecasting models, their skill with respect to the other models, and the importance and the limitations of the model evaluations. Besides the scientific interest, these results also have a significant practical impact. In particular, the seismic hazard center (Centro di Pericolosità Sismica, CPS) at INGV provides, as part of a pilot phase, operational earthquake forecasts for time windows of one week to the Department of Civil Protection (Marzocchi et al., 2014; 2017). This forecasting system (OEF_ITALY) is entirely based on models that are currently under test in CSEP testing centers. Hence, the results of these experiments are essential to assess the current forecasting capabilities of models that may be applied for practical purposes.

Finally, we discuss the forecasting performance of different flavors of ensemble models for the 1-day testing class. Ensemble models are an emerging field of research applied to many different kinds of forecasts (e.g. Ranjan and Gneiting, 2010); their main goal is to improve the forecasting skill by combining available models. In seismology, ensemble modeling is used by OEF_ITALY (Marzocchi et al., 2014). In this paper we use the three different types of ensembles introduced by Marzocchi et al. (2012a), namely the Bayesian model averaging (BMA), score model averaging (SMA), and generalized score model averaging (gSMA). In short, BMA, gSMA and SMA ensembles (in this order) weight the models with decreasing emphasis on past forecasting performance (see equations 16, 20, and 18 in Marzocchi et al., 2012a, respectively).

Models and data

The models under evaluation are summarized in Table 1. They are described in detail in a special issue of the Annals of Geophysics (references in Table 1); here, we only mention the main features. Marzocchi et al. (2010) showed maps of the 16 5-year forecasts (their Figure 2).

Table 1. List of models under test in the first CSEP experiment in Italy.

Model name	Testing class	Main features	Reference
ETAS_LM	1-day	Epidemic-type aftershocks sequence (ETAS) model that was calibrated using the Italian instrumental catalog from 2005 to 2009, both for the background and triggering part.	Lombardi and Marzocchi, 2010a
STEP_NG, STEP_LG	1-day	Short-term earthquake probability (STEP) model with specific parametrization for Italy (NG) and original global parameters (LG). For the triggering part, these models use a spatial extension of the Reasenberg and Jones aftershock model. The model was calibrated on a merged Italian instrumental catalog, covering the period from 1981 to 2007.	Woessner et al., 2010
ETES	1-day	ETAS model with a spatial decay of the triggering activity independent from the magnitude of the shock, and with the aftershock productivity parameter α fixed to 1. The model was calibrated on a merged Italian instrumental catalog from 1987 to 2009.	Falcone et al., 2010
TripleS_CPTI, TripleS_CSI, TripleS_Hyb	3-month, 5-years	Time-independent smoothed seismicity, using a historical, instrumental and merged (hybrid) catalog, respectively; catalogs are not declustered. The models use a two-dimensional isotropic Gaussian smoothing kernel with a single parameter, re-estimated for each forecast generation. TripleS_CSI is used for 3-month and 5-year forecasts, while TripleS_CPTI and TripleS_Hyb only contribute 5-year forecasts.	Zechar and Jordan, 2010
RI_L, RI_S, RI	3-month, 5-year	Time-independent smoothed seismicity model, which assumes that future earthquakes are more likely to occur where historical seismicity has been relatively high. The different versions refer to a different smoothing parameter (RI_L is smoother than RI_S). All versions use a spatially uniform b-value equal to 1.2. The models were calibrated on two merged instrumental catalogs, from 1985 to 2002 and from 2005 to 2009.	Nanjo, 2010
HAZFX_BPT	5-year	Time-dependent model based on Brownian-Passage-Time recurrence on the Italian individual seismogenic sources and well-constrained macroseismic sources. A time-independent background seismicity was added by smoothing the historical seismicity.	Marzocchi et al., 2012b
MPS04, MPS04_AFTER	5-year	Official time-independent model of the national seismic hazard model for Italy, and the same model corrected for clustering: the total original rate was multiplied by a factor 1.25 to adjust for local magnitude rates. These models are composed by different sub-regions, each one with its own b-value and seismic rate (spatially uniform inside the sub-region).	MPS Working Group, 2004
HRSS_m1, HRSS_m2	5-year	Time-independent smoothed seismicity calibrated on instrumental seismicity since 1981, and merged historical/ instrumental seismicity since 1900, respectively. The models use an adaptive power-law kernel to smooth the seismicity, and a tapered Gutenberg-Richter.	Werner et al., 2010b

DBM	5-year	Time-dependent two-layer clustering model, based on the idea that there are two branching processes that describe the seismicity: a short-term one (that lasts days to months) and a long term one (typically a few decades).	Lombardi and Marzocchi, 2010b
HZA_TI, HZA_TD	5-year	Time-independent and time-dependent versions of a smoothed seismicity model, which also includes Coulomb Failure Stress interaction among faults and a rate-and-state friction law. The time-dependent seismicity rate changes were estimated by the rate-and-state model, which considers all M4+ earthquakes that occurred since 2007.	Chan et al., 2010
LTST	5-years	Time-dependent recurrence model on faults with Coulomb Failure Stress interaction computed for the Italian individual seismogenic sources added to time-independent background seismicity for the cells without seismogenic sources; the smoothed seismicity was computed using the Frankel (1995) method.	Falcone et al., 2010
HALM, ALM, ALM_IT	5-year	Different parametrizations of the time-independent model with spatial variations of the b-value. These models assume that small-scale spatial variations in the b-value of the Gutenberg-Richter relationship are useful to recognize the zones with bigger probability to have major events in the future. Then, lower b-values characterize asperities, where future mainshocks are more likely to occur.	Gulia et al., 2010
PHM_GRID, PHM_ZONE	5-years	Spatial grid and tectonic zonation applied to a time-dependent clustering model. These models uses a common empirical time clustering decay, which is independent of the magnitude of the earthquakes.	Faenza and Marzocchi, 2010
HAZGRIDX	5-year	Time-independent smoothed seismicity based both on historical and instrumental catalogs; this model use an isotropic spatial smoothing kernel and the Weichert (1980) method to estimate b-values and seismic rates.	Akinci et al., 2010

FIGURE 1 HERE

Figure 1. Target earthquakes in the five years CSEP experiments in Italy. Yellow dots are earthquakes of magnitude $3.95 \leq M < 4.95$*, red dots for* $M \geq 4.95$*. The inset shows a zoom for the 2012 Emilia seismic sequence.*

The target earthquakes that occurred during the testing period are shown in Figure 1, and in Table S1 in the electronic supplement. During the five years of the experiments, 97 target earthquakes (83 earthquakes with magnitude in the range $3.95 \leq \mathrm{M} < 4.95$, and 14 with magnitude $\mathrm{M} \geq 4.95$) were recorded, including a significant sequence in the Emilia region in 2012, which caused 27 fatalities and severe disruption to one of the most economically productive areas in Italy. The CSEP Italy experiment participants decided against declustering target earthquakes (Schorlemmer et al., 2010a).

Model evaluation

The CSEP testing methods are continuously evolving and strengthening. Changes to the suite of tests do not impair the validity of the CSEP experiments that are rooted in the principles of transparency and reproducibility. Whenever a new test becomes available, the experiment can be re-run at any time without weakening the principle of the independence between the data used for testing and the forecasts of the models because the models were submitted before the start of the testing period. Certainly, some modelers might balk at having unanticipated metrics applied to their forecasts; analogously, basketball player Wilt Chamberlain, well-known for chasing individual stats rather than wins, would have played basketball differently if he had foreseen the advanced metrics used in today's NBA. But CSEP introduces metrics to provide new insights into model strengths and weaknesses, and not for the purpose of penalizing models.

The suite of tests used by CSEP are described in detail in Schorlemmer et al. (2007), Zechar et al. (2010b), Rhoades et al. (2011), and Zechar and Zhuang (2014); below, we summarize the main features of the tests. The consistency tests assess if (i) the observed number of earthquakes (*N*-test), (ii) their spatial distribution (S-test), and (iii) their magnitude distribution (*M*-test), are consistent with a forecast. If the *P*-value of a test is below a preselected threshold, the model "fails" to describe satisfactorily the observed data. We interpret such occurrences as a potentially meaningful discrepancy between a forecast and observations.

The *N*-test evaluates if the sum of predicted earthquakes in all time-space-magnitude bins (N_{fore}) is consistent with the number of target earthquakes observed (N_{obs}) over the entire testing region, over any set of forecasting time windows (one forecasting time window for the 5-year experiment; the sum of all time windows for the cumulative *N*-test of the 1-day experiment), and of any magnitude above the threshold used to define the target earthquakes. The (two-tailed) *P*-value of the test is calculated assuming that the target earthquakes follow the Poisson distribution with average N_{fore} and a two-tailed hypothesis test is appropriate.

The *S*-test evaluates the consistency of the spatial occurrence of target earthquakes regardless of their magnitudes with a model's normalized spatial forecast, isolating the spatial component of the forecast. After normalizing the forecast with the total number of observed target earthquakes, the *S*-test is summarized by a quantile score, ζ, which is also the *P*-value of the test. This quantile score is the fraction of simulated synthetic catalogs of target earthquakes with spatial log-likelihoods smaller than the observed spatial log-likelhood calculated with the observed target earthquakes; a small *P*-value means the model is fitting the data less well than expected if the model were generating the data. The *M*-test checks if the observed magnitude distribution is consistent with the magnitude distribution forecast by the model. This test is analogous to the *S*-test, but it isolates the magnitude distribution of target earthquakes and normalizes it to the observed number, hereby neglecting the spatial distribution. Here we show the results of the cumulative tests, while the plot of the incremental *N*- and *S*-tests are reported in Figure S1 in the electronic supplement.

All of these tests are based on the Poisson assumption that earthquakes occur in time and space independently from previous earthquakes, a distribution characterized only by one parameter that is the forecast of the model in each space-time-magnitude bin. This assumption has been largely discussed in the literature (e.g. Werner and Sornette, 2008; Lombardi and Marzocchi, 2010c): it only roughly approximates the earthquake occurrence distribution, in particular for earthquakes with a small-to-moderate minimum magnitude. In general, the number of target earthquakes is expected to follow overdispersed distributions (e.g. Kagan, 2017; Lombardi and Marzocchi, 2010c). For testing purposes, the Poisson assumption makes the tests more severe, i.e., it may lead to 'rejecting' models that would capture the overdispersed distributions, like most ETAS models (Lombardi and Marzocchi, 2010c). We account for this effect by using a significance level of 0.01, instead of the customary 0.05 originally adopted by CSEP, and we show the *P*-value of the test, which is a graded measure of the strength of evidence against the null Poisson hypothesis (Amrhein et al., 2017).

To evaluate the skill of the forecasts, we use three metrics: i) the log-likelihood per event (LLe); ii) the information gain (IG), which is the difference of the log-likelihood of a model and the log-likelihood of a reference model for Italy; and iii) the parimutuel gambling score (PGS), which compares the model and the reference Italian model in a gambling and betting framework. The larger the IG or PGS, the more skilled the model with respect to the reference model. Importantly, LLe, IG and PGS are proper scores (Gneiting and Raftery, 2007), meaning that they tend to maximize if the model is the data-generator ("true"). The use of different metrics like IG and PGS is justified by the fact that they are highlighting different aspects of the model performance. For example, PGS assumes that each model can win or loose almost the same amount with respect to the reference model. On the contrary, IG is based on log-likelihood which is heavily asymmetrical, because it has an upper bound (zero), but it does not have a lower bound (see Taroni et al., 2014, for more details). In practice, IG tends to reward models that never fail, while PGS tends to reward models that perform better on average than the reference model.

A common reference model for all testing classes allows us to compare how much better (or worse) any model (short-term or long-term) is with respect to the reference. For this experiment, we define the seismicity rate model of the national seismic hazard model, corrected for clustering (MPS04_AFTER), as reference model. This model has been submitted to the 5-year testing class with target earthquakes M4.95+. To extend this model as reference for IG and PGS in all other testing classes which consider target earthquakes M3.95+, we multiply the seismicity rate of MPS04_AFTER in each bin by ten (i.e. assuming a global b-value equal to 1) and scale the rate to the length of the forecast horizon. We deem this approximation reasonable; in fact, MPS04_AFTER is based on a seismotectonic zonation with seismicity rates distributed according to truncated Gutenberg-Richter distributions, and the tests are focused on the average behavior of the model (so an average Gutenberg-Richter law is expected to hold), not on the behavior in each specific region. Conversely, this approximation precludes the possibility to consider the magnitude bins, so the log-likelihood accounts only for the spatial and temporal domains. (This choice is not critical in this experiment because 24 of 26 models pass the *M*-test.) Technically, the log-likelihood for the IG is calculated summing up all magnitude bins, and it is not the same log-likelihood used to calculate LLe; hence, they carry different information.

Results for the 1-day models

In the five years of the 1-day model experiment, 97 target earthquakes (M3.95+) were recorded. The forecasts are updated at 00:00 UTC of each day from August 1, 2009 (resulting in 1,826 1-day forecasts). In Table 2 we show the results of this experiment.

Table 2. Cumulative results (P-value and rank) of the 1-day experiment. A white model cell indicates that all tests are passed (all P-values ≥ 0.01), light gray that two tests out of three have P-values ≥ 0.01, and dark gray otherwise. P-values in bold show values below 0.01.

Model	***N*-test** (N_{fore} / N_{obs})	***S*-test**	***M*-test**	**LLe (rank)**	**IG (rank)**	**PGS (rank)**
ETAS_LM	0.22 (1.14)	0.17	0.50	- 12.18 (1)	290.03 (1)	33.95 (1)
STEP_NG	0.14 (0.86)	**0.001**	0.58	- 14.07 (2)	106.70 (2)	26.19 (2)
STEP_LG	0.03 (1.24)	**<0.001**	0.60	- 14.08 (3)	105.73 (3)	11.64 (3)
ETES	**0** (342)	**<0.001**	0.81	- 348.03 (4)	- 32291.30 (4)	-14579.10 (4)

The ETAS_LM model passes all tests and leads in ranking according to all scoring metrics. The difference between the predicted and observed number of target earthquakes is +14% for ETAS_LM, -14% for STEP_NG and +24% for STEP_LG. The ETES model shows a significant overprediction. The spatial distribution is well captured only by the ETAS_LM model, while all other models fail the *S*-test. All models pass the *M*-test.

The reason for the overprediction of the ETES model was recognized as a bug in the software code submitted to the testing center. Subsequently, the code was corrected during the Emilia sequence, but the rules of the CSEP experiments do not allow us to consider this new version for the present tests. Incidentally, the new ETES model codes seems to be performing well: this is supported by prospective applications during the two most recent major seismic sequences in Italy, the Emilia sequence in 2012 (Marzocchi et al., 2012c), and the Amatrice-Norcia sequence in 2016-2017 (Marzocchi et al., 2017).

FIGURE 2 HERE

Figure 2. Upper panel: cumulative IG for the 1-day models and ensemble models shown in the legend. Lower panel: the daily number of target earthquakes.

In Figure 2 we show the trend of IG as a function of time for the 1-day models and for BMA, SMA and gSMA ensemble models. We do not consider the ETES model because its log-likelihood immediately goes out of scale. The figure shows that the ETAS_LM model outperforms the other models after few months. As expected, the biggest increase in IG for all models is in proximity of the beginning of the 2012 seismic sequence, because the (time-independent) reference model cannot track the marked space-time evolution of the sequence. The performance of the ETAS_LM model is particularly better at the time of the Emilia sequence as shown by the larger increase in the cumulative IG. The STEP_NG and STEP_LG models perform similarly in terms of IG, but the STEP_NG model is more consistent and has greater skill measured by PGS, in agreement to what was found in the retrospective analysis (Woessner et al., 2010). Finally, we notice that the model ranking with time is stable, and that all models outperform the reference time-independent model MPS04_AFTER.

FIGURE 3 HERE

Figure 3. The left panels show the daily number of earthquakes of the STEP_NG model in the time bins with at least one target earthquake (63 days). Red bars denote S-test

failures and green bars denote S-test passes. The right panel shows the histogram of simulated spatial loglikelihood scores of the STEP_NG model, and the vertical red bar is the observed spatial log-likelihood; the vertical dashed line is the value of the spatial log-likelihood associated with the significance level of the test (0.01).

The STEP_NG model does not describe the spatial distribution of earthquakes well because its spatial aftershock distribution decays too quickly. In Figure 3, we show the *S*-test results for the STEP_NG model on days with at least one earthquake. The STEP_NG model fails the *S*-test on particularly active days; the same holds also for the STEP_LG model. The reason can be seen in Figure 4, which shows the 1-day forecast for May 29, 2012, when the second large peak of the Emilia sequence occurred. The figure shows that the forecast in the aftershock zone of both STEP models decays much faster than the forecast of the ETAS_LM model. The STEP models' spatial clustering zone is too small with respect to the observed triggering.

FIGURE 4 HERE

Figure 4. 1-day forecasts for May 29, 2012. The left, central and right panels show the ETAS_LM, the STEP_NG, and the STEP_LG forecasts, respectively. The black points are the target earthquakes. The color palette on the right of the map shows the logarithm in base 10 of the expected number of target earthquake per each spatial cell. The lower panels are close-up versions of the corresponding upper panels.

Finally, for this testing class we also assess the performance of the BMA, SMA and gSMA ensemble models. In Figure 2 we show the evolution of the IG for all 1-day models and the different ensembles. We do not consider the ETES model, because the consistency tests have shown a bug in the code. Two ensemble models perform about as well as the best individual model (ETAS_LM), while the SMA ensemble has a slightly lower IG than the best individual model. This is due to the fact that BMA and gSMA, more than SMA, tend to have superior forecasting performances when one of the models outperform significantly all the others as in the present case (Marzocchi et al., 2012a); this is because BMA and gSMA weight the best performing model more strongly, as it can be seen in Figure 5. Moreover, all ensemble models pass the *N*- and *S*-tests. This result confirms the importance of ensemble modeling for practical purposes; in fact, we cannot know at the beginning of an experiment which model will be the best one, but increasing evidence from the scientific literature suggests that the ensemble model will perform about as well as the best individual model, and in some cases even better (e.g., Taroni et al., 2014).

FIGURE 5 HERE

Figure 5. Weight of each model as a function of time in the different ensembles. The BMA and gSMA ensembles give much greater weight to the best performing model (ETAS_LM) than the SMA ensemble.

Results for the 3-month models

This testing class considers the same target earthquakes of the 1-day class (93 M3.95+ earthquakes in the five years of the experiment). The models have been updated every three months, on 1 Jan., 1 Apr., 1 Jul., and 1 Oct.

Table 3. As for Table 2, but for the 3-months experiment in Italy.

Model	N-test (N_{fore} / N_{obs})	S-test	M-test	LLe (rank)	IG (rank)	PGS (rank)
TripleS_CSI	0.01 (0.77)	**<0.001**	0.51	- 10.91 (3)	- 12.61 (3)	14.55 (3)
RI_L	**0.0002 (0.67)**	**<0.001**	**0.005**	- 10.77 (1)	4.85 (1)	17.46 (2)
RI_S	**0.0002 (0.67)**	**<0.001**	**0.005**	- 10.89 (2)	- 6.79 (2)	19.40 (1)

From Table 3 we can see that the RI_L and RI_S 3-month models fail all consistency tests, while TripleS_CSI fails only the *S*-test. The RI_L and RI_S models fail the *M*-test because they assume a *b*-value (1.2) for the Gutenberg-Richter distribution that is much higher than the *b*-value (0.97) of the observed earthquakes. The ranking is almost stable across the different metrics. The poor forecasting performance of all models can be explained by the fact that all models are (quasi) time-independent (Table 1), so they are not particularly suitable to track the evolution of seismic sequences. Only TripleS_CSI may partially do this, because, even though the model is time-independent, it re-estimated the parameters before each forecast, introducing an implicit capability to cope with time-dependency. However, this experiment is particularly challenging for forecasting, because the major seismic sequence occurred in late May and June of 2012, i.e., at the end of one forecasting time window, in a region of small long-term seismicity rate.

Results for the 5-years models

During the testing period, 14 M4.95+ target earthquakes occurred. Most of them (10 of 14) are related to the Emilia earthquake in 2012, and 5 out of 14 are in the same spatial bin, so this dataset clearly does not meet the Poisson assumption, and it may affect the outcomes of tests based on the Poisson hypothesis. Moreover, the limited number of target earthquakes suggests caution in interpreting the results (Strader et al., 2017). To strengthen the robustness of the results, CSEP already planned a new analysis for a 10-year experiment on January 1, 2020; this will allow a comparison of the results obtained after 5 and 10 years, and a check of the stability of the results.

That said, we now outline the most interesting features of the results that are reported in Table 4. No models fail the M-test. All models underpredict the number of target earthquakes, with some of them failing the N-test. The universal underprediction by all forecasts is due to the large percentage of triggered target earthquakes (9 of 14), i.e., to the large time clustering observed. The best performing model in this aspect is the MPS04 model after it has been corrected for declustering (Faenza et al., 2010; Marzocchi and Taroni, 2014).

Several models fail the *S*-test, i.e., target earthquakes did not occur where expected by the models. The generally poor spatial performance is due to the fact that 5 out of 14 target earthquakes occur in a single spatial bin that has a small long-term seismicity rate and this is hardly accommodated by any model under the Poisson assumption. In Table S2 of the electronic supplement we show that some models fail the S-test due to very poor spatial forecasts for some target earthquakes (ALM, ALM_IT, HALM and PHM_ZONE), while others fail because of average poor performance in spatial bins where no target earthquakes occurred (e.g. PHM_GRID). The best performing models in space take into account faults (HAZFX_BPT), and the historical seismicity (TRIPLES_CPTI); in particular, the T/W test (Rhoades et al., 2011) –which allows us to test the statistical significance of the IG difference of two models under the null hypothesis of models equally informative– shows that HAZFX_BPT outperforms all other models. Conversely, most models based on smoothing recent instrumental seismicity (e.g. TripleS_CSI, HRSS_m1, and HAZGRIDX) are among the worst performing. This can be explained by the fact that the large majority of target earthquakes occur in a low-seismicity area, and/or, more generally, by the fact that a short instrumental catalog cannot provide a good description of the spatial variability of the seismicity (Werner et al., 2010).

Models using the variability of the *b*-value do not perform well (consistently with the retrospective tests carried out by Werner et al., 2010), even though the discrepancy in their PGS and IG scores seem to suggest that the HALM model does a good job in forecasting most target earthquakes, but fails severely in forecasting a few of them (see Table S1 in the electronic supplement).

The models that inform the national seismic hazard model (MPS04 and MPS04_AFTER) perform well with respect to the other models (most of the models have negative IG and PGS with respect to MPS04_AFTER. In Table S1 (electronic supplement) we show that the poor spatial performance (*S*-test) of these two models is likely due to a poor performance of these models in spatial cells where no target earthquake occurred (i.e., given the number of earthquakes, the models would have placed greater likelihood in cells other than those in which earthquakes occurred).

Finally, if we compare the log-likelihood per event (LLe) with the same quantity obtained for the CSEP experiment in California (Zechar et al., 2013; Strader et al., 2017), we notice that the average predictive skill of Italian models is worse in this first 5-year experiment; in fact, while in Table 4, the LLe ranges mostly from -9 and -10, the same quantity is smaller in California ranging mostly from -7 and -9 (Tables 2 and 3 in Zechar et al., 2013), and even better in a more recent experiment (Strader et al., 2017). We conjecture this worse performance of the Italian experiment might be due to the marked clustering of target earthquakes in low-seismicity area. Moreover, whereas in California, the adaptive smoothing model by Helmstetter et al. (2007), calibrated on recent high quality locations of small to moderate events, performed better than other methods, here we observe a different trend: both the fixed-radius and adaptive smoothing methods (TripleS_CSI, HRSS_m1, respectively) calibrated to the modern era of the Italian network (mid 80s to mid/late 2000s) performed near the bottom of the group.

Table 4. Results (P-values) of the 5-year experiment in Italy.

Model	***N*-test (N_{fore} / N_{obs})**	***S*-test**	***M*-test**	**LLe (rank)**	**IG (rank)**	**PGS (rank)**
HAZFX_BPT	0.02 (0.54)	0.20	0.61	- 8.24 (1)	10.78 (1)	4.48 (1)
MPS04_AFTER	0.51 (0.88)	**0.001**	0.66	- 9.07 (2)	0 (4)	0 (4)
TripleS_CPTI	0.04 (0.58)	0.36	0.62	- 9.09 (3)	- 0.84 (6)	- 0.14 (6)
MPS04	0.16 (0.71)	**0.004**	0.66	- 9.11 (4)	- 0.7 (5)	- 0.28 (7)
HRSS_m2	0.06 (0.61)	0.06	0.62	- 9.12 (5)	- 1.12 (7)	- 0.42 (9)
DBM	0.11 (0.67)	0.01	0.62	- 9.23 (6)	- 2.80 (8)	- 1.12 (12)
RI	**0.0004 (0.35)**	0.02	0.58	- 9.34 (7)	- 4.9 (10)	- 0.56 (10)
TripleS_Hyb	0.03 (0.56)	0.04	0.65	- 9.36 (8)	- 4.48 (9)	- 1.12 (13)
HZA_TI	0.09 (0.65)	0.10	0.73	- 9.42 (9)	0.42 (2)	0.056 (3)
HZA_TD	0.09 (0.65)	0.06	0.68	- 9.56 (12)	0.056 (3)	- 0.056 (5)
LTST	0.06 (0.62)	**<0.001**	0.60	- 9.66 (13)	- 8.68 (13)	- 1.96 (17)
HALM	0.05 (0.60)	**<0.001**	0.72	- 12.05 (17)	- 40.74 (17)	1.82 (2)
ALM	0.04 (0.58)	**<0.001**	0.72	- 12.96 (18)	- 54.18 (18)	- 3.92 (8)
ALM_IT	0.06 (0.62)	**<0.001**	0.56	- 20.70 (19)	- 164.5 (19)	- 3.08 (19)
PHM_GRID	**0.006 (0.46)**	**0.002**	0.60	- 9.45 (10)	- 0.41 (11)	- 0.06 (11)
HAZGRIDX	**0.004 (0.44)**	**<0.001**	0.63	- 9.52 (11)	- 0.50 (12)	- 0.10 (14)
HRSS_m1	**0.004 (0.44)**	**<0.001**	0.62	- 9.71 (14)	- 0.68 (14)	- 0.11 (15)
TripleS_CSI	**0.002 (0.42)**	**0.002**	0.63	- 9.78 (15)	- 0.75 (15)	- 0.12 (16)
PHM_ZONE	**0.004 (0.44)**	**<0.001**	0.63	- 10.29 (16)	- 1.26 (16)	- 0.15 (18)

Discussion and conclusions

We provided the results of the first CSEP experiments in Italy. The limited number of target earthquakes warrants caution in drawing firm conclusions from the 5-year experiment, but the results reported in this paper nonetheless provide some indications worth considering.

- The ranking of the 1-day models is stable in time (Figure 2), suggesting that testing 1,826 1-day forecasts provides useful information about the consistency and skill of the models.
- The 1-day models (other than ETES) are consistent with the number and magnitudes of the target earthquakes; ETAS_LM is also spatially consistent, while STEP models predict a triggering region that is too small to adequately model the 2012 Emilia sequence. The IG shows that 1-day models outperform the time-independent reference model even in times of low seismic activity (see the trend in Figure 2). These results are consistent with results of the prospective tests carried out during other seismic sequences in Italy (Marzocchi and Lombardi, 2009; Marzocchi et al., 2012c; 2017), and it vindicates the scientific robustness of these models for potential uses in an operational earthquake forecasting perspective (Jordan et al., 2011; Marzocchi et al., 2014).
- Ensemble models perform well in the 1-day testing class; they pass the consistency N- and S-tests and show IG comparable to the best performing model (ETAS_LM).
- The 3-month models do not pass most of the consistency tests, showing that they were unable to track the space-time evolution of the seismicity.
- Even though the clustering of large earthquakes can last for a few years, the best performing models in the 5-year testing class are time-independent. As shown also by Taroni et al. (2014) in testing 1-year models at global scale, time-dependent models may offer advantages only if the forecasts can be updated after significant earthquakes and/or sequences occur. This is not (currently) accommodated in CSEP experiments, so this testing framework is more suited to assessing the spatial skill of the forecasting models rather than the time dependence.
- The 5-year experiment has only 14 target earthquakes, and they are strongly clustered. This is a very challenging feature for any forecasting model. The results show that models only based on instrumental seismicity fail to describe the spatial distribution of target earthquakes, while the inclusion of historical seismicity and the fault distribution improves the spatial forecasting significantly. Moreover, models based on the spatial variability of the b-value perform poorly in forecasting some target earthquakes, but at least one such model (HALM) forecasts the spatial distribution of the other target earthquakes well.

Data and resources

Data and models are available in the CSEP European testing center at ETH in Zurich.

Acknowledgments

This research was supported by the Southern California Earthquake Center (Contribution No. 8022). SCEC is funded by NSF Cooperative Agreement EAR-1033462 & USGS Cooperative Agreement G12AC20038 and the King Abdullah University of Science and Technology (KAUST) research grant URF/1/2160-01-01. Part of the work has been carried out within the Seismic Hazard Center (Centro di Pericolosità Sismica, CPS) at the Istituto Nazionale di Geofisica e Vulcanologia (INGV).

References

Akinci, A. (2010). HAZGRIDX: earthquake forecasting model for ML ≥ 5.0 earthquakes in Italy based on spatially smoothed seismicity, *Ann. Geophys.* **53(3)**, 51-61.

Amrhein, V., F. Korner-Nievergelt, and T. Roth (2017). The earth is flat (p> 0.05): Significance thresholds and the crisis of unreplicable research, *PeerJ* **5**, e3544.

Chan, C. H., M. B. Sørensen, D. Stromeyer, G. Grünthal, O. Heidbach, A. Hakimhashemi, and F. Catalli (2010). Forecasting Italian seismicity through a spatio-temporal physical model: importance of considering time dependency and reliability of the forecast, *Ann. Geophys.* **53(3)**, 129-140.

Faenza, L. and W. Marzocchi (2010). The Proportional Hazard Model as applied to the CSEP forecasting area in Italy, *Ann. Geophys.* **53(3)**, 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(3)**, 41-50.

Gneiting, T. and A. E. Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102(477)**, 359-378.

Gulia, L., S. Wiemer and D. Schorlemmer (2010). Asperity based earthquake likelihood models for Italy, *Ann. Geophys.* **53(3)**, 63-75.

Jordan, T., 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(4)**, 315-391.

Jordan, T.H. (2006). Earthquake predictability, brick by brick, *Seismol. Res.Lett.* **77(1)**, 3-6.

Kagan, Y. Y. (2017). Earthquake number forecasts testing, *Geophys. J. Int.* **211(1)**, 335-345.

Lombardi, A. M., and W. Marzocchi (2010a). The ETAS model for daily forecasting of Italian seismicity in the CSEP experiment, *Ann. Geophys.* **53(3)**, 155-164.

Lombardi, A. M., and W. Marzocchi (2010b). A double-branching model applied to long-term forecasting of Italian seismicity (ML ≥ 5.0) within the CSEP project, *Ann. Geophys.* **53(3)**, 31-39.

Lombardi, A. M., and W. Marzocchi (2010c). The assumption of Poisson seismic-rate variability in CSEP/RELM experiments, *Bull. Seism. Soc. Am.* **100(5A)**, 2293-2300.

Marzocchi, W., and A. M. Lombardi (2009). Real-time forecasting following a damaging earthquake, *Geophys. Res. Lett*. **36(21)**, L21302.

Marzocchi, W., D. Schorlemmer and S. Wiemer (2010). Preface, *Ann. Geophys.* **53(3)**, III-VIII.

Marzocchi, W., J.D. Zechar, and T.H. Jordan (2012a). Bayesian forecast evaluation and ensemble earthquake forecasting, *Bull. Seismol. Soc. Am.*,**102**, 2574-2584.

Marzocchi, W., A. Amato, A. Akinci, C. Chiarabba,A. M. Lombardi,D. Pantosti, and E. Boschi (2012b). A Ten-Year Earthquake Occurrence Model for Italy, *Bull. Seism. Soc. Am.* **102(3),** 1195-1213.

Marzocchi, W., M. Murru, A. M. Lombardi, G. Falcone and R. Console (2012c). Daily earthquake forecasts during the May-June 2012 Emilia earthquake sequence (northern Italy), *Ann. Geophys.* **55(4)**, 561-567.

Marzocchi, W., and M. Taroni (2014). Some thoughts on declustering in probabilistic seismic-hazard analysis, *Bull. Seism. Soc. Am.* **104(4)**, 1838-1845.

Marzocchi, W., A. M. Lombardi, and E. Casarotti (2014). The establishment of an operational earthquake forecasting system in Italy, *Seismol. Res.Lett.* **85(5)**, 961-969.

Marzocchi, W., M. Taroni and G. Falcone (2017). Earthquake forecasting during the complex Amatrice-Norcia seismic sequence, *Sci. Adv*. **3(9)**, e1701239.

MPS Working Group (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, April 2004 (MPS04), 65 pp. + 5 appendices; http://zonesismiche.mi.ingv.it

Nanjo, K. Z. (2010). Earthquake forecast models for Italy based on the RI algorithm, *Ann. Geophys.* **53(3)**, 117-127.

Ranjan, R., and T. Gneiting (2010). Combining probabilistic forecasts. *J. R. Statist. Soc. B* **72**, 71-91.

Rhoades, D. A., D. Schorlemmer, M.C. Gerstenberger, A. Christophersen, J. D. Zechar and M. Imoto (2011). Efficient testing of earthquake forecasting models, *Acta Geophys*. **59(4)**, 728-747.

Schorlemmer, D., and M. C. Gerstenberger (2007). RELM testing center, *Seismol. Res. Lett.* **78(1),** 30-36.

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

Schorlemmer, D., A. Christophersen, A. Rovida,F. Mele, M. Stucchi, and W. Marzocchi (2010a). Setting up an earthquake forecast experiment in Italy, *Ann. Geophys.* **53(3)**, 1-9.

Schorlemmer, D., F. Mele, and W. Marzocchi (2010b). A completeness analysis of the National Seismic Network of Italy, *J. Geophys. Res. B Solid Earth Planets* **115(B4)**, B04308.

Strader, A., M. Schneider, and D. Schorlemmer (2017). Prospective and retrospective evaluation of five-year earthquake forecast models for California, *Geophys. J. Int.* **211(1)**, 239-251.

Taroni, M., J. D. Zechar, and W. Marzocchi (2014). Assessing annual global M 6+ seismicity forecasts, *Geophys. J. Int.* **196(1)**, 422-431.

Werner, M. J., and D. Sornette (2008). Magnitude uncertainties impact seismic rate estimates, forecasts, and predictability experiments, *J. Geophys. Res. B Solid Earth* **113(B8)**, B08302.

Werner, M.J., J.D. Zechar, W. Marzocchi, and S. Wiemer (2010a). Retrospective evaluation of the five-year and ten-year CSEP-Italy earthquake forecasts. *Ann. Geophys.* **53(3)**, 11-30.

Werner, M. J., A. Helmstetter, D. D. Jackson, Y. Y. Kagan, and S. Wiemer (2010b). Adaptively smoothed seismicity earthquake forecasts for Italy, *Ann. Geophys.* **53(3)**, 107-116.

Woessner, J., A. Christophersen, J. D. Zechar, and D. Monelli (2010). Building self-consistent, short-term earthquake probability (STEP) models: improved strategies and calibration procedures, *Ann. Geophys.* **53(3)**, 141-154.

Zechar, J. D., and J. Zhuang (2014). A parimutuel gambling perspective to compare probabilistic seismicity forecasts, *Geophys. J. Int*. **199(1),** 60-68.

Zechar, J. D. and T. H. Jordan (2010). Simple smoothed seismicity earthquake forecasts for Italy, *Ann. Geophys.* **53(3)**, 99-105.

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

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

Zechar, J.D., D. Schorlemmer, M.J. Werner, M.C. Gerstenberger, D.A. Rhoades, and T.H. Jordan (2013). Regional Earthquake Likelihood Models I: First-Order Results. *Bull. Seismol. Soc. Am.* **103**, 787-798.

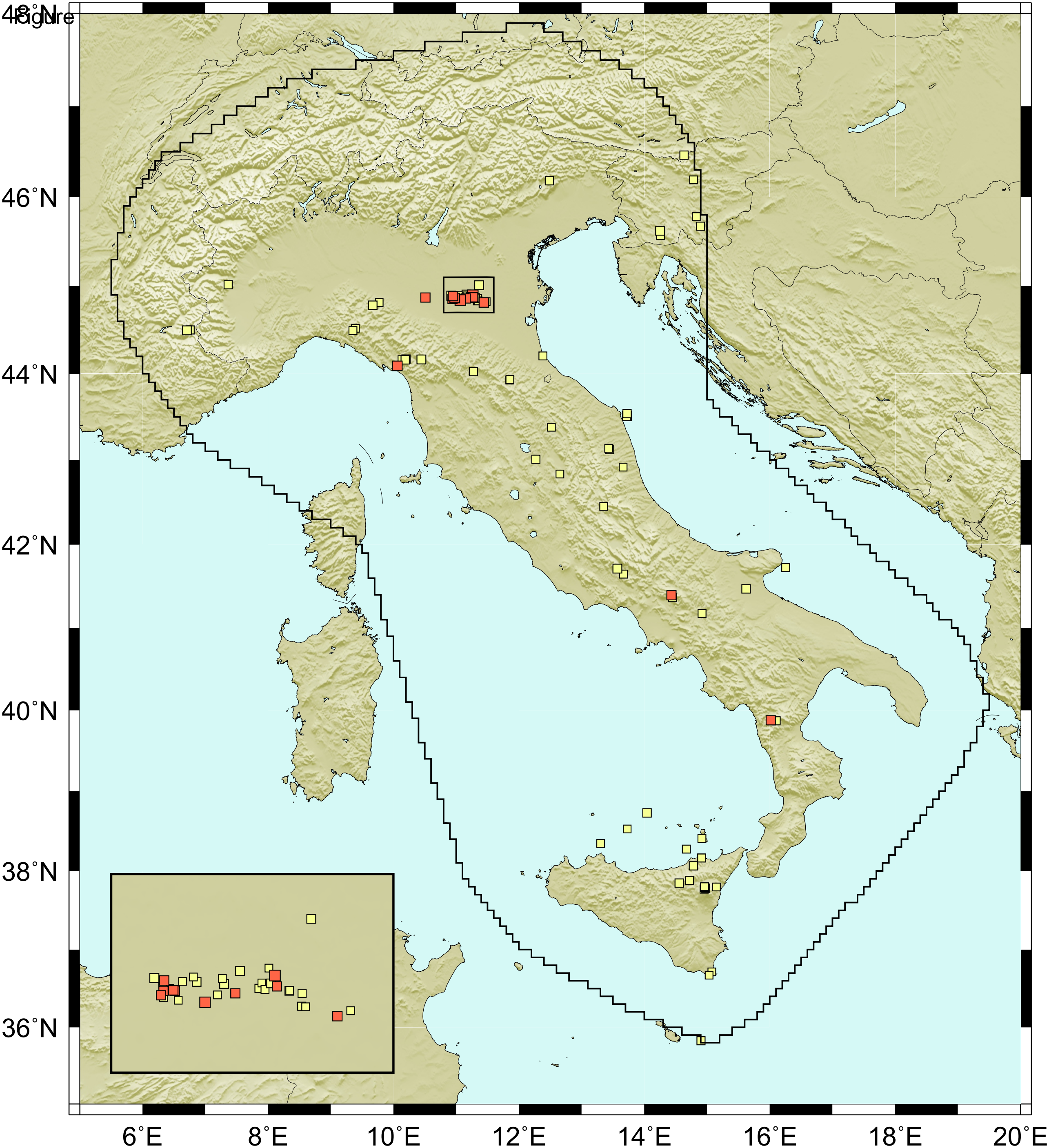
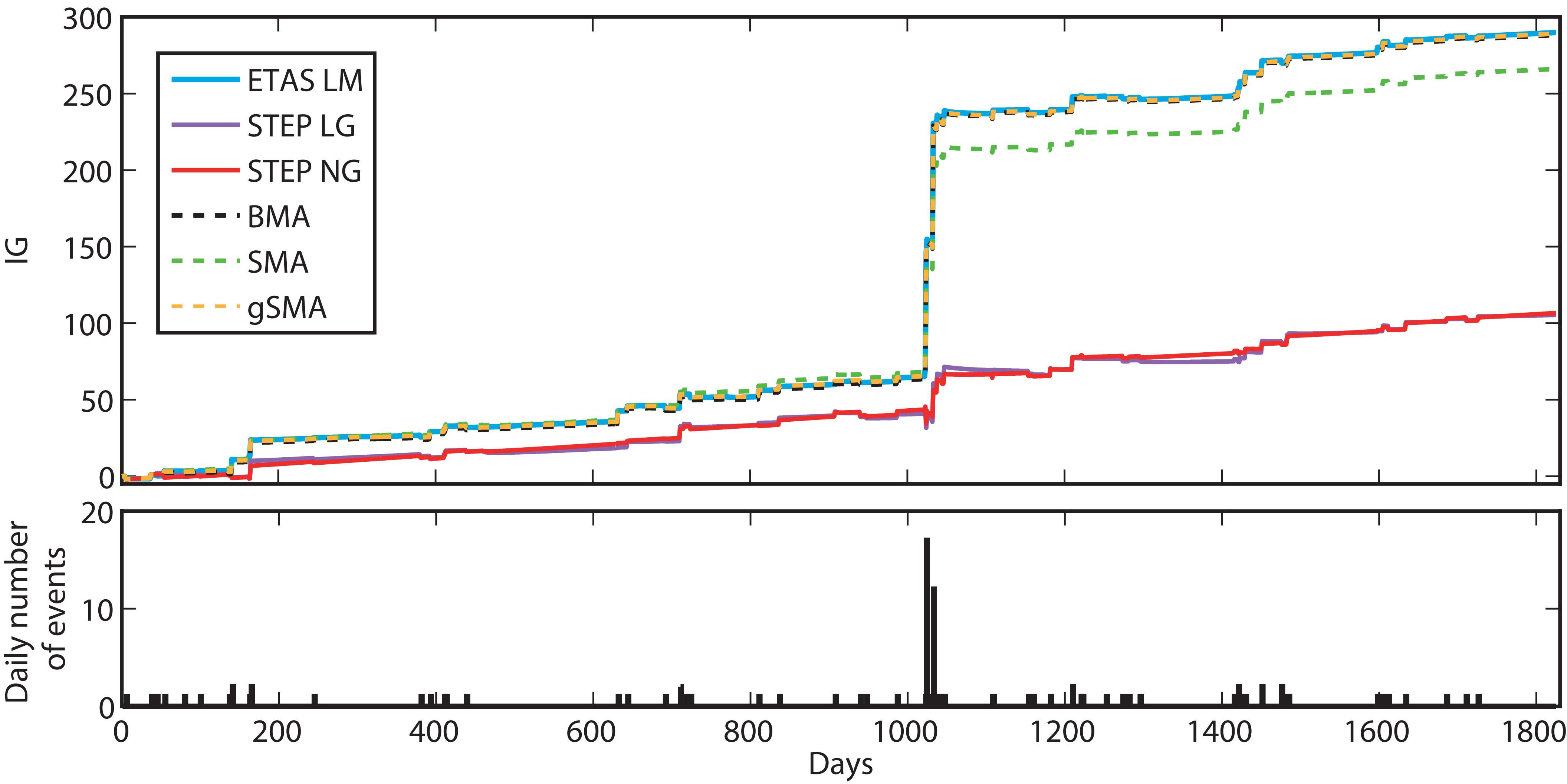


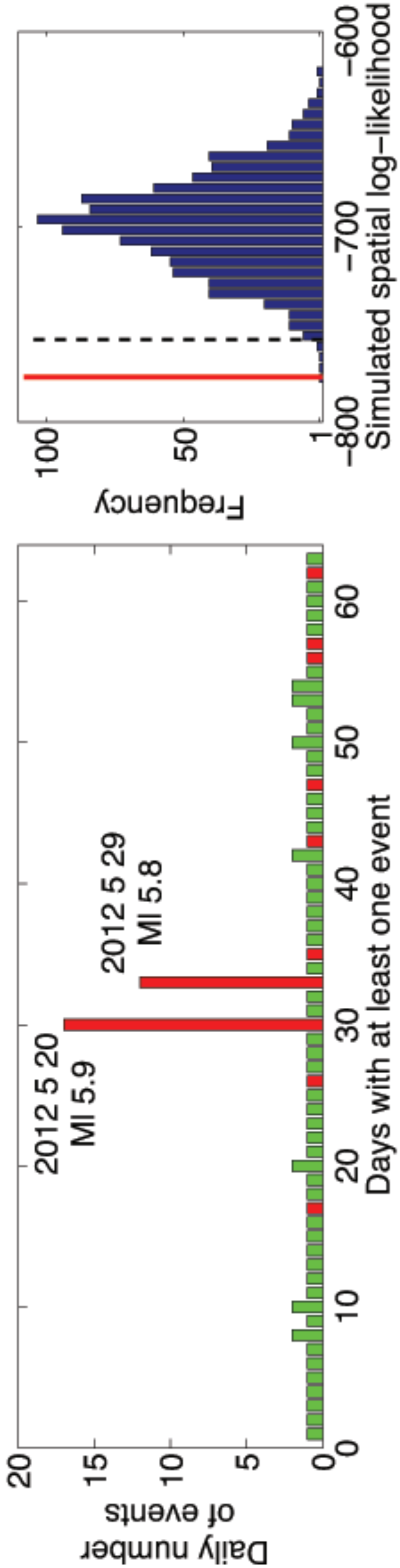
Figure
48°N
46°N
44°N
42°N
40°N
38°N
36°N
6°E
8°E
10°E
12°E
14°E
16°E
18°E
20°E

Figure



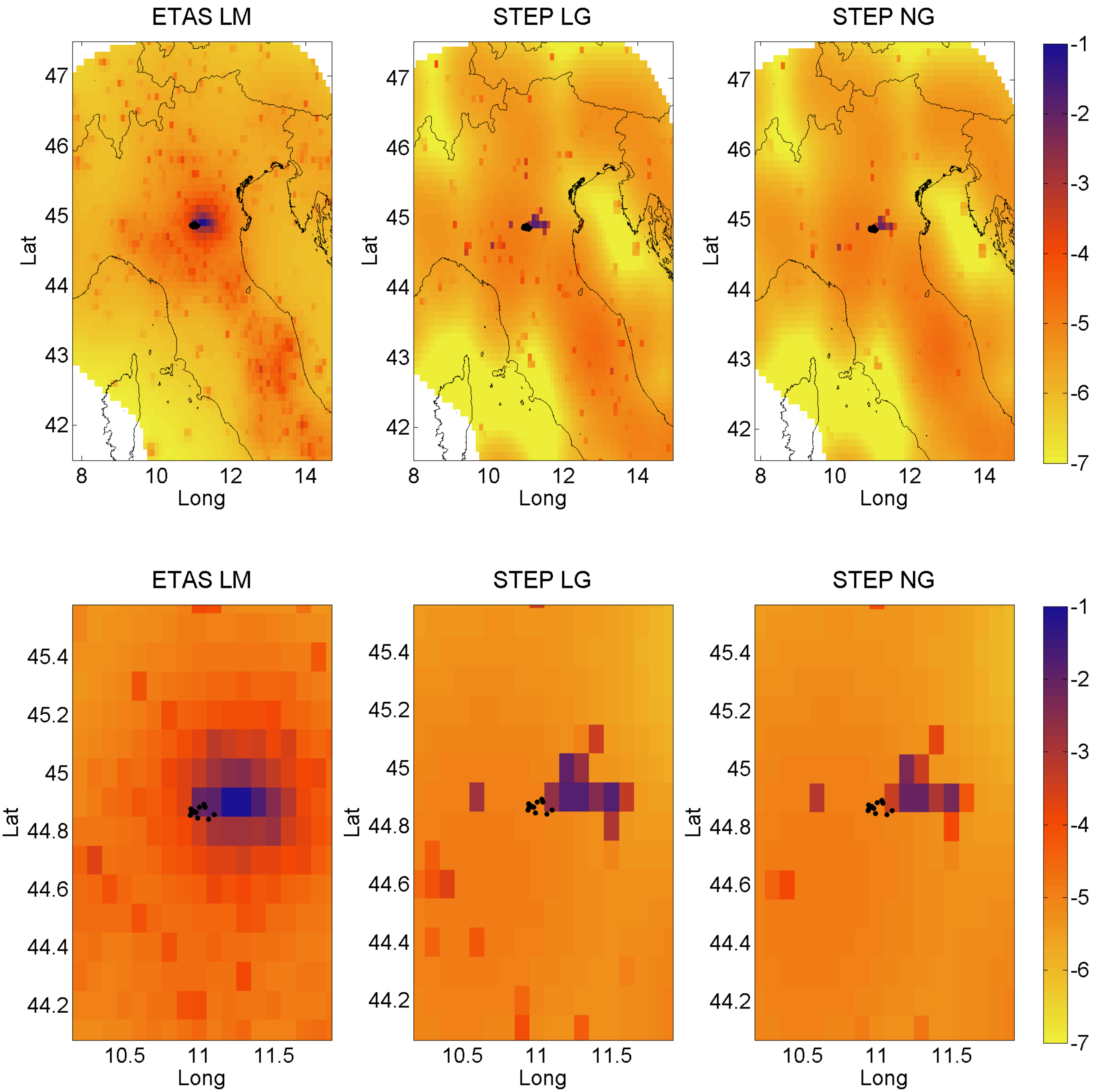
300
250
200
150
100
50
0
IG
ETAS LM
STEP LG
STEP NG
BMA
SMA
gSMA
20
10
0
Daily number of events
0
200
400
600
800
1000
1200
1400
1600
1800
Days

Figure 3



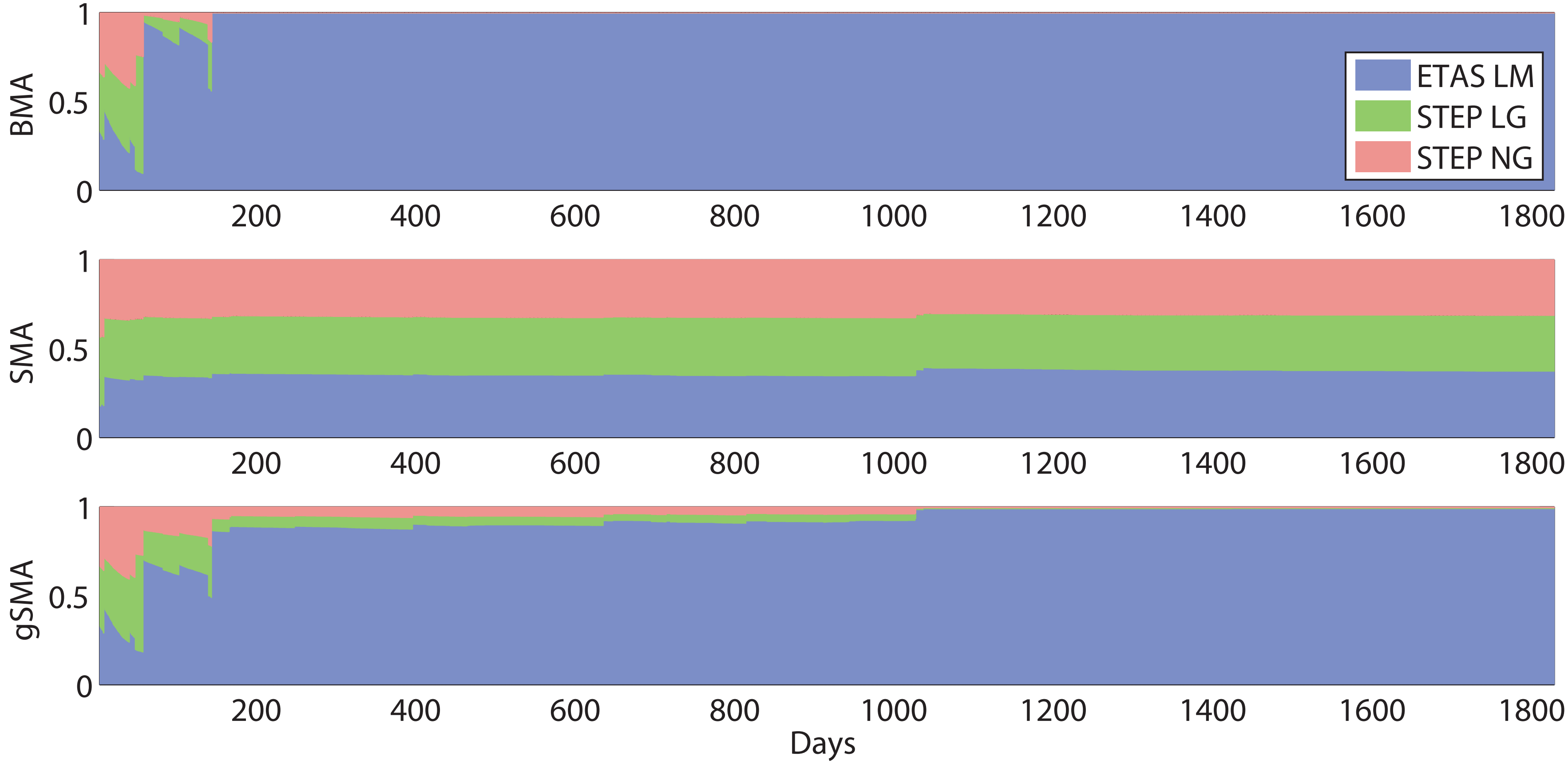
Daily number of events
0
5
10
15
20
0
10
20
30
40
50
60
2012 5 20
Ml 5.9
2012 5 29
Ml 5.8
Days with at least one event
Frequency
1
50
100
−800
−700
−600
Simulated spatial log−likelihood

Figure



ETAS LM
STEP LG
STEP NG
Lat
Long
8
10
12
14
42
43
44
45
46
47
10.5
11
11.5
44.2
44.4
44.6
44.8
45
45.2
45.4
-1
-2
-3
-4
-5
-6
-7

Figure



1
0.5
0
BMA
200
400
600
800
1000
1200
1400
1600
1800
ETAS LM
STEP LG
STEP NG
1
0.5
0
SMA
200
400
600
800
1000
1200
1400
1600
1800
1
0.5
0
gSMA
200
400
600
800
1000
1200
1400
1600
1800
Days

Electronic supplement

Prospective CSEP evaluation of 1-day, 3-month, and 5-year earthquake forecasts for Italy

M. Taroni[1], W. Marzocchi[1], D. Schorlemmer[2], M. Werner[3], S. Wiemer[4], J.D. Zechar[5], L. Heiniger[3], F. Euchner[3].

1- Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy.

2- GFZ, Potsdam, Germany

3- University of Bristol, United Kingdom.

4- ETH Zurich, Switzerland

5- AXIS Capital, Zürich, Switzerland

Table S1 contains the list of the target earthquakes of the experiments. In Figure S1 we show the incremental *N*- and *S*-test results of the 1-day experiment (in the manuscript we show only the cumulative results); the aim is to show when models start failing the tests. In Table S2 we show the log-likelihood of each model in the spatial bins where target earthquakes occurred; the goal is to show which model (if any) misses the spatial occurrence of one or more target earthquakes.

Table S1. Target earthquakes for the Italian experiment since 2009. In bold the target earthquakes for the 5-year experiment.

Figure S1. The *P*-value of the incremental *N*- and *S*-tests for each 1-day model as a function of the day since August 1, 2009. ETES has been excluded for the reasons reported in the manuscript. The horizontal dashed line is the 0.01 significance level. From the figure we can see that STEP_NG and STEP_LG failed both tests at the time of the Emilia sequence (around day 1000). Afterwards the *N*-test recovers while the *S*-test does not (see Table 2 in the manuscript).

Table S2. The log-likelihood of each 5-year model in the spatial bin where target earthquakes occurred. The number of target earthquakes that occurred in each spatial bin are reported in the first row. From the table we can notice that some models (ALM_IT, ALM; HALM; and PHM_zone) failed the *S*-test because they score very low log-likelihood (grey cells) in at least one spatial bin where target earthquakes occurred. Conversely, the other models that failed the *S*-test (see Table 4 in the paper) performed poorly in the spatial bins without target earthquakes. This can be noted, for example, by comparing the HZA_TD and PHM_grid models. The first model passes the consistency tests while the second one does not; however, they have similar log-likelihoods in the spatial bins where target earthquakes occurred. Finally, we notice that MPS04 and MPS04_after models – which inform the seismicity rate model of the national seismic hazard model – do not perform well for the first and the second-to-last target earthquakes, but the poor performance in the S-test is likely due to a poor performance on the spatial cells where no target earthquakes occurred.

Long	Lat	Year	Month	Day	Hour	Minute	Second	Magnitude	Depth
13.67	41.65	2009	8	6	15	36	44.44	4.2	15.7
14.04	38.73	2009	9	7	21	26	29.69	4.5	25.5
11.28	44.02	2009	9	14	20	4	31.3	4.3	7
13.35	42.45	2009	9	24	16	14	57.56	4.1	16.3
9.772	44.81	2009	10	19	10	8	49.64	4.0	23.6
14.56	37.85	2009	11	8	6	51	16.41	4.4	7.6
12.27	43.01	2009	12	15	13	11	58.98	4.3	8.8
14.95	37.77	2009	12	19	5	36	28.79	4.3	24.7
14.97	37.78	2009	12	19	9	1	16.46	4.4	26.9
13.45	43.12	2010	1	10	8	33	35.64	4.0	16.9
13.45	43.12	2010	1	12	8	25	11.32	4.1	17.1
13.43	43.13	2010	1	12	13	35	45.29	4.2	18.1
15.15	37.8	2010	4	2	20	4	45.1	4.0	3.4
14.92	38.41	2010	8	16	12	54	47.5	4.2	16.9
12.65	42.83	2010	8	28	7	8	3.25	4.1	6.7
14.26	45.57	2010	9	15	2	23	13.75	4.0	10
15.62	41.47	2010	9	17	12	20	17.75	4.5	6
12.38	44.2	2010	10	13	22	43	14.74	4.2	26.5
14.9	35.83	2011	4	24	13	2	12.3	4.3	9.7
14.96	37.79	2011	5	6	15	12	35.5	4.2	28.1
14.78	38.06	2011	6	23	22	2	46.71	4.4	7.3
11.86	43.93	2011	7	12	6	53	22.47	4.0	7.6
11.86	43.93	2011	7	12	7	15	8.33	4.0	8.2
11.37	45.01	2011	7	17	18	30	27.31	4.8	2.4
7.365	45.02	2011	7	25	12	31	20.46	4.3	11
9.393	44.52	2011	10	20	6	11	18.86	4.0	5.1
14.67	38.27	2011	11	15	4	59	0.36	4.2	8.4
10.51	**44.87**	**2012**	**1**	**25**	**8**	**6**	**37.9**	**5.0**	**29**

6.759	44.5	2012	2	26	22	37	55.92	4.3	10.4
9.354	44.49	2012	3	5	15	15	6.99	4.2	10.8
13.3	38.34	2012	4	13	6	21	32.63	4.1	9.2
11.25	44.91	2012	5	19	23	13	25.62	4.1	9.3
11.26	**44.9**	**2012**	**5**	**20**	**2**	**3**	**50.17**	**5.9**	**9.5**
11.12	44.88	2012	5	20	2	6	12.5	4.8	5
11.16	44.91	2012	5	20	2	6	26.47	4.8	4.3
11.27	**44.87**	**2012**	**5**	**20**	**2**	**7**	**28.95**	**5.0**	**6.1**
11.34	44.83	2012	5	20	2	9	48.35	4.3	4.9
11.34	44.86	2012	5	20	2	11	45.55	4.3	10.9
11.22	44.87	2012	5	20	2	12	40.47	4.3	6.7
10.95	44.85	2012	5	20	2	20	56.52	4.2	5
11.12	44.89	2012	5	20	2	21	50.49	4.1	4.9
11.48	44.83	2012	5	20	2	35	32.44	4.0	25.9
11.23	44.88	2012	5	20	2	39	7.41	4.0	6.6
11.15	**44.86**	**2012**	**5**	**20**	**3**	**2**	**47.9**	**5.0**	**9.1**
11.24	44.87	2012	5	20	9	13	18.49	4.2	7.2
11.44	**44.81**	**2012**	**5**	**20**	**13**	**18**	**1.77**	**5.1**	**3.4**
11.35	44.83	2012	5	20	13	21	5.31	4.1	8.3
11.31	44.87	2012	5	20	17	37	14.14	4.6	5.4
11.25	44.88	2012	5	20	17	38	14.38	4.6	3.7
11.31	44.87	2012	5	21	16	37	31.36	4.1	3.6
16.1	39.87	2012	5	28	1	6	26.83	4.3	8.3
11.07	**44.84**	**2012**	**5**	**29**	**7**	**0**	**2.88**	**5.8**	**8.1**
10.99	44.85	2012	5	29	7	7	20.91	4.0	3.5
10.95	**44.87**	**2012**	**5**	**29**	**8**	**25**	**51.48**	**5.0**	**7.9**
11.04	44.88	2012	5	29	8	27	22.65	4.6	6
10.97	44.87	2012	5	29	8	40	57.44	4.2	4.1
10.95	44.88	2012	5	29	8	41	42.33	4.1	6.5

11	44.88	2012	5	29	9	29	37.9	4.1	6.4
11.1	44.86	2012	5	29	10	3	25.76	4.0	2.5
10.98	**44.87**	**2012**	**5**	**29**	**10**	**55**	**56.55**	**5.3**	**4.4**
10.94	**44.86**	**2012**	**5**	**29**	**11**	**0**	**1.68**	**5.0**	**8.7**
10.98	**44.87**	**2012**	**5**	**29**	**11**	**0**	**22.99**	**5.1**	**7.2**
11.03	44.89	2012	5	29	11	7	4.63	4.0	8
10.95	**44.89**	**2012**	**6**	**3**	**19**	**20**	**43.39**	**5.1**	**8.7**
12.49	46.18	2012	6	9	2	4	56.6	4.4	6.9
10.92	44.89	2012	6	12	1	48	36.14	4.9	8.3
16.25	41.73	2012	8	12	1	21	36.8	4.2	29.1
13.73	38.53	2012	8	13	7	30	51.89	4.0	26.8
14.92	41.18	2012	9	27	1	8	22.65	4.2	10.3
9.67	44.78	2012	10	3	14	41	29.36	4.5	23.8
16.02	**39.88**	**2012**	**10**	**25**	**23**	**5**	**24.73**	**5.0**	**9.7**
14.96	37.8	2012	11	22	9	10	41.53	4.0	24.4
14.96	37.8	2012	11	22	11	25	51.67	4.1	27.3
14.79	46.19	2012	12	3	4	36	0.66	4.1	7.3
13.66	42.91	2012	12	5	1	18	20.29	4.0	17.5
14.72	37.88	2013	1	4	7	50	6.8	4.3	15.1
10.45	44.16	2013	1	25	14	48	18.27	4.8	19.8
14.63	46.46	2013	2	2	13	35	34.28	4.4	10
13.57	41.71	2013	2	16	21	16	9.29	4.7	17.1
14.83	45.78	2013	6	16	20	5	0	4.2	10
10.06	**44.09**	**2013**	**6**	**21**	**10**	**33**	**56.7**	**5.3**	**5.7**
10.14	44.16	2013	6	21	12	12	39.66	4.0	8.1
10.2	44.17	2013	6	23	15	1	33.86	4.4	9.2
10.19	44.16	2013	6	30	14	40	8.48	4.4	6.1
13.72	43.51	2013	7	21	1	32	24.24	4.9	7.9
13.72	43.5	2013	7	21	3	7	24.44	4.0	8.6

14.91	38.16	2013	8	15	23	4	58.47	4.2	25.6
14.91	38.16	2013	8	15	23	6	51.2	4.2	24.8
13.72	43.54	2013	8	22	6	44	51.58	4.4	8.9
15.08	36.71	2013	8	24	17	18	18.77	4.0	8.7
15.03	36.67	2013	12	15	3	57	34.1	4.1	10.5
12.52	43.38	2013	12	22	10	6	35.69	4.0	8.6
14.43	**41.4**	**2013**	**12**	**29**	**17**	**8**	**43.23**	**5.0**	**20.4**
14.45	41.37	2014	1	20	7	12	40.1	4.3	17.2
14.9	45.67	2014	3	13	17	31	59.48	4.3	8.3
6.707	44.5	2014	4	7	19	26	59.79	4.7	11.1
14.25	45.62	2014	4	22	8	58	27.42	4.7	10

Table S2

Model Name	#1 2012/ 1/25 Ml 5.0	#2 2012/ 5/20; Ml 5.9 Ml 5.0	#1 2012/ 5/20; Ml 5.0	#1 2012/ 5/20; Ml 5.1	#1 2012/ 5/29; Ml 5.8	#5 4 2012/5/29; 1 2012/6/30; Ml 5.3 2 Ml 5.1 2 Ml 5.0	#1 2012/ 10/25; Ml 5.0	#1 2013/ 6/21; Ml 5.3	#1 2013/ 12/29; Ml 5.0
HAZGRIDX	-9.6	-24	-11.6	-12.8	-13.6	-51.5	-11	-12.2	-10.8
HAZFX_BPT	-10.2	-21.5	-10.1	-12.2	-12	-45.1	-9.9	-10.8	-8.9
HZA_TD	-10.8	-24.1	-11.8	-11.9	-13.7	-51.3	-13.6	-12.2	-10.9
HZA_TI	-10.7	-23.9	-11.6	-11.7	-13.7	-50.5	-13.3	-12.1	-10.7
LTST	-9.9	-25.7	-12.6	-10.1	-14.7	-53	-11.3	-12.7	-11.4
PHM_grid	-11.4	-23.9	-11.6	-11.5	-13.3	-51.6	-10.6	-11.9	-10.7
PHM_zone	-16.9	-25.3	-12.3	-13.6	-14.1	-55.9	-10.1	-10.5	-9.2
ALM	-11	-22.5	-10.6	-12.3	-14.1	-102.8	-12.5	-10.7	-10.6
HALM	-11.1	-20.6	-9.8	-11.5	-13.1	-95.1	-12.5	-10.7	-10.4
DBM	-10.7	-23.7	-11.5	-11.9	-13.3	-50.9	-11.3	-11.3	-11.2
MPS04	-13.8	-22.9	-10.9	-11.2	-13	-49.3	-10.4	-12.4	-10.4
MPS04_after	-13.8	-22.9	-10.9	-11.2	-13	-49.3	-10.4	-12.4	-10.4
RI	-10.4	-24.2	-11	-12.6	-13	-48.3	-11	-11.5	-10.6
ALM_IT	-37.6	-26.4	-11.2	-12.9	-13.8	-182.8	-9.4	-12.3	-9.6
HRSS_m1	-10.9	-23.7	-11.5	-12.3	-13.6	-53.3	-11	-12.2	-11.4
HRSS_m2	-10.9	-23.2	-11.2	-11.6	-13	-49.7	-11.7	-11.4	-11
TripleS_CPTI	-11.1	-22.8	-11.1	-11.3	-12.9	-49.8	-11.2	-12	-10.8
TripleS_CSI	-11.8	-23.9	-11.7	-11.8	-13.5	-52.9	-11	-12.6	-11.2
TripleS_Hyb	-11.4	-23.3	-11.3	-11.5	-13.2	-51.2	-11.1	-12.3	-11.1

Electronic supplement

Prospective CSEP evaluation of 1-day, 3-month, and 5-year earthquake forecasts for Italy

M. Taroni[1], W. Marzocchi[1], D. Schorlemmer[2], M. Werner[3], S. Wiemer[4], J.D. Zechar[5], L. Heiniger[3], F. Euchner[3].

1- Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy.

2- GFZ, Potsdam, Germany

3- University of Bristol, United Kingdom.

4- ETH Zurich, Switzerland

5- AXIS Capital, Zürich, Switzerland

Table S1 contains the list of the target earthquakes of the experiments. In Figure S1 we show the incremental *N*- and *S*-test results of the 1-day experiment (in the manuscript we show only the cumulative results); the aim is to show when models start failing the tests. In Table S2 we show the log-likelihood of each model in the spatial bins where target earthquakes occurred; the goal is to show which model (if any) misses the spatial occurrence of one or more target earthquakes.

Long	Lat	Year	Month	Day	Hour	Minute	Second	Magnitude	Depth
13.67	41.65	2009	8	6	15	36	44.44	4.2	15.7
14.04	38.73	2009	9	7	21	26	29.69	4.5	25.5
11.28	44.02	2009	9	14	20	4	31.3	4.3	7
13.35	42.45	2009	9	24	16	14	57.56	4.1	16.3
9.772	44.81	2009	10	19	10	8	49.64	4.0	23.6
14.56	37.85	2009	11	8	6	51	16.41	4.4	7.6
12.27	43.01	2009	12	15	13	11	58.98	4.3	8.8
14.95	37.77	2009	12	19	5	36	28.79	4.3	24.7
14.97	37.78	2009	12	19	9	1	16.46	4.4	26.9
13.45	43.12	2010	1	10	8	33	35.64	4.0	16.9
13.45	43.12	2010	1	12	8	25	11.32	4.1	17.1
13.43	43.13	2010	1	12	13	35	45.29	4.2	18.1
15.15	37.8	2010	4	2	20	4	45.1	4.0	3.4
14.92	38.41	2010	8	16	12	54	47.5	4.2	16.9
12.65	42.83	2010	8	28	7	8	3.25	4.1	6.7
14.26	45.57	2010	9	15	2	23	13.75	4.0	10
15.62	41.47	2010	9	17	12	20	17.75	4.5	6
12.38	44.2	2010	10	13	22	43	14.74	4.2	26.5
14.9	35.83	2011	4	24	13	2	12.3	4.3	9.7
14.96	37.79	2011	5	6	15	12	35.5	4.2	28.1
14.78	38.06	2011	6	23	22	2	46.71	4.4	7.3
11.86	43.93	2011	7	12	6	53	22.47	4.0	7.6
11.86	43.93	2011	7	12	7	15	8.33	4.0	8.2
11.37	45.01	2011	7	17	18	30	27.31	4.8	2.4
7.365	45.02	2011	7	25	12	31	20.46	4.3	11
9.393	44.52	2011	10	20	6	11	18.86	4.0	5.1
14.67	38.27	2011	11	15	4	59	0.36	4.2	8.4
10.51	**44.87**	**2012**	**1**	**25**	**8**	**6**	**37.9**	**5.0**	**29**
6.759	44.5	2012	2	26	22	37	55.92	4.3	10.4
9.354	44.49	2012	3	5	15	15	6.99	4.2	10.8
13.3	38.34	2012	4	13	6	21	32.63	4.1	9.2
11.25	44.91	2012	5	19	23	13	25.62	4.1	9.3
11.26	**44.9**	**2012**	**5**	**20**	**2**	**3**	**50.17**	**5.9**	**9.5**
11.12	44.88	2012	5	20	2	6	12.5	4.8	5
11.16	44.91	2012	5	20	2	6	26.47	4.8	4.3
11.27	**44.87**	**2012**	**5**	**20**	**2**	**7**	**28.95**	**5.0**	**6.1**
11.34	44.83	2012	5	20	2	9	48.35	4.3	4.9
11.34	44.86	2012	5	20	2	11	45.55	4.3	10.9
11.22	44.87	2012	5	20	2	12	40.47	4.3	6.7
10.95	44.85	2012	5	20	2	20	56.52	4.2	5
11.12	44.89	2012	5	20	2	21	50.49	4.1	4.9
11.48	44.83	2012	5	20	2	35	32.44	4.0	25.9
11.23	44.88	2012	5	20	2	39	7.41	4.0	6.6
11.15	**44.86**	**2012**	**5**	**20**	**3**	**2**	**47.9**	**5.0**	**9.1**
11.24	44.87	2012	5	20	9	13	18.49	4.2	7.2
11.44	**44.81**	**2012**	**5**	**20**	**13**	**18**	**1.77**	**5.1**	**3.4**
11.35	44.83	2012	5	20	13	21	5.31	4.1	8.3
11.31	44.87	2012	5	20	17	37	14.14	4.6	5.4
11.25	44.88	2012	5	20	17	38	14.38	4.6	3.7
11.31	44.87	2012	5	21	16	37	31.36	4.1	3.6
16.1	39.87	2012	5	28	1	6	26.83	4.3	8.3

11.07	**44.84**	**2012**	**5**	**29**	**7**	**0**	**2.88**	**5.8**	**8.1**
10.99	44.85	2012	5	29	7	7	20.91	4.0	3.5
10.95	**44.87**	**2012**	**5**	**29**	**8**	**25**	**51.48**	**5.0**	**7.9**
11.04	44.88	2012	5	29	8	27	22.65	4.6	6
10.97	44.87	2012	5	29	8	40	57.44	4.2	4.1
10.95	44.88	2012	5	29	8	41	42.33	4.1	6.5
11	44.88	2012	5	29	9	29	37.9	4.1	6.4
11.1	44.86	2012	5	29	10	3	25.76	4.0	2.5
10.98	**44.87**	**2012**	**5**	**29**	**10**	**55**	**56.55**	**5.3**	**4.4**
10.94	**44.86**	**2012**	**5**	**29**	**11**	**0**	**1.68**	**5.0**	**8.7**
10.98	**44.87**	**2012**	**5**	**29**	**11**	**0**	**22.99**	**5.1**	**7.2**
11.03	44.89	2012	5	29	11	7	4.63	4.0	8
10.95	**44.89**	**2012**	**6**	**3**	**19**	**20**	**43.39**	**5.1**	**8.7**
12.49	46.18	2012	6	9	2	4	56.6	4.4	6.9
10.92	44.89	2012	6	12	1	48	36.14	4.9	8.3
16.25	41.73	2012	8	12	1	21	36.8	4.2	29.1
13.73	38.53	2012	8	13	7	30	51.89	4.0	26.8
14.92	41.18	2012	9	27	1	8	22.65	4.2	10.3
9.67	44.78	2012	10	3	14	41	29.36	4.5	23.8
16.02	**39.88**	**2012**	**10**	**25**	**23**	**5**	**24.73**	**5.0**	**9.7**
14.96	37.8	2012	11	22	9	10	41.53	4.0	24.4
14.96	37.8	2012	11	22	11	25	51.67	4.1	27.3
14.79	46.19	2012	12	3	4	36	0.66	4.1	7.3
13.66	42.91	2012	12	5	1	18	20.29	4.0	17.5
14.72	37.88	2013	1	4	7	50	6.8	4.3	15.1
10.45	44.16	2013	1	25	14	48	18.27	4.8	19.8
14.63	46.46	2013	2	2	13	35	34.28	4.4	10
13.57	41.71	2013	2	16	21	16	9.29	4.7	17.1
14.83	45.78	2013	6	16	20	5	0	4.2	10
10.06	**44.09**	**2013**	**6**	**21**	**10**	**33**	**56.7**	**5.3**	**5.7**
10.14	44.16	2013	6	21	12	12	39.66	4.0	8.1
10.2	44.17	2013	6	23	15	1	33.86	4.4	9.2
10.19	44.16	2013	6	30	14	40	8.48	4.4	6.1
13.72	43.51	2013	7	21	1	32	24.24	4.9	7.9
13.72	43.5	2013	7	21	3	7	24.44	4.0	8.6
14.91	38.16	2013	8	15	23	4	58.47	4.2	25.6
14.91	38.16	2013	8	15	23	6	51.2	4.2	24.8
13.72	43.54	2013	8	22	6	44	51.58	4.4	8.9
15.08	36.71	2013	8	24	17	18	18.77	4.0	8.7
15.03	36.67	2013	12	15	3	57	34.1	4.1	10.5
12.52	43.38	2013	12	22	10	6	35.69	4.0	8.6
14.43	**41.4**	**2013**	**12**	**29**	**17**	**8**	**43.23**	**5.0**	**20.4**
14.45	41.37	2014	1	20	7	12	40.1	4.3	17.2
14.9	45.67	2014	3	13	17	31	59.48	4.3	8.3
6.707	44.5	2014	4	7	19	26	59.79	4.7	11.1
14.25	45.62	2014	4	22	8	58	27.42	4.7	10

Table S1. Target earthquakes for the Italian experiment since 2009. In bold the target earthquakes for the 5-year experiment.

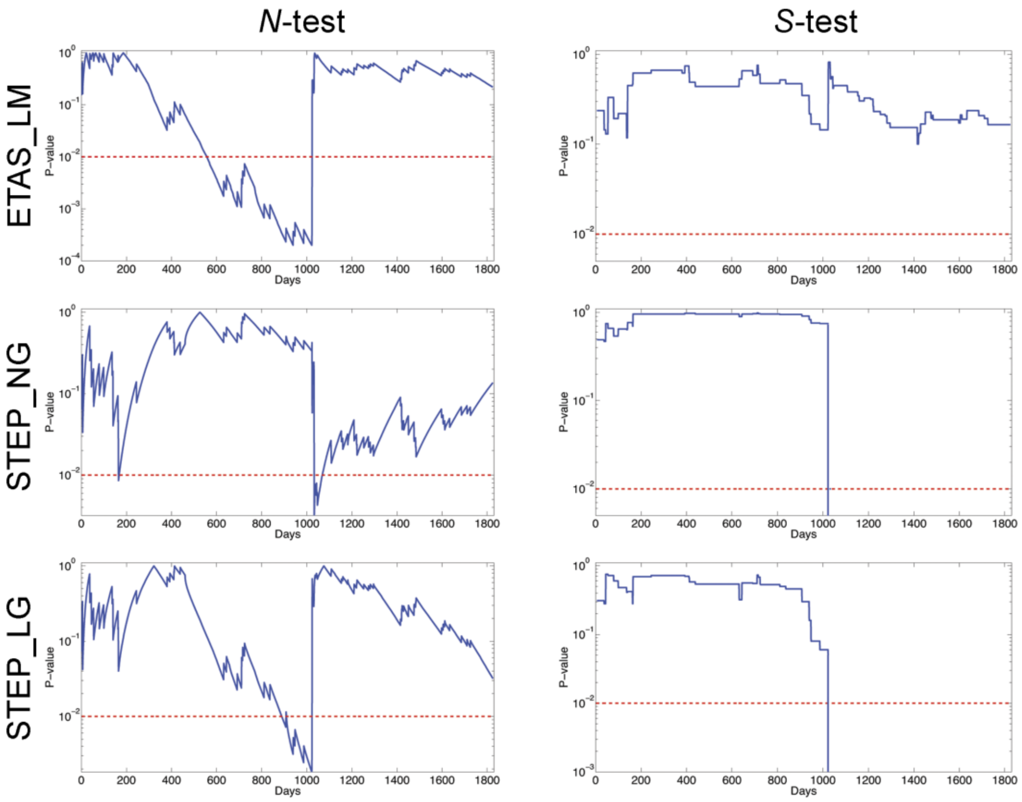


Figure S1. The *P*-value of the incremental *N*- and *S*-tests for each 1-day model as a function of the day since August 1, 2009. ETES has been excluded for the reasons reported in the manuscript. The horizontal dashed line is the 0.01 significance level. From the figure we can see that STEP_NG and STEP_LG failed both tests at the time of the Emilia sequence (around day 1000). Afterwards the *N*-test recovers while the *S*-test does not (see Table 2 in the manuscript).

Model Name	#1 2012/ 1/25 MI 5.0	#2 2012/ 5/20; MI 5.9 MI 5.0	#1 2012/ 5/20; MI 5.0	#1 2012/ 5/20; MI 5.1	#1 2012/ 5/29; MI 5.8	#5 4 2012/5/29; 1 2012/6/30; MI 5.3 2 MI 5.1 2 MI 5.0	#1 2012/ 10/25; MI 5.0	#1 2013/ 6/21; MI 5.3	#1 2013/ 12/29; MI 5.0
HAZGRIDX	-9.6	-24	-11.6	-12.8	-13.6	-51.5	-11	-12.2	-10.8
HAZFX_BPT	-10.2	-21.5	-10.1	-12.2	-12	-45.1	-9.9	-10.8	-8.9
HZA_TD	-10.8	-24.1	-11.8	-11.9	-13.7	-51.3	-13.6	-12.2	-10.9
HZA_TI	-10.7	-23.9	-11.6	-11.7	-13.7	-50.5	-13.3	-12.1	-10.7
LTST	-9.9	-25.7	-12.6	-10.1	-14.7	-53	-11.3	-12.7	-11.4
PHM_grid	-11.4	-23.9	-11.6	-11.5	-13.3	-51.6	-10.6	-11.9	-10.7
PHM_zone	-16.9	-25.3	-12.3	-13.6	-14.1	-55.9	-10.1	-10.5	-9.2
ALM	-11	-22.5	-10.6	-12.3	-14.1	-102.8	-12.5	-10.7	-10.6
HALM	-11.1	-20.6	-9.8	-11.5	-13.1	-95.1	-12.5	-10.7	-10.4
DBM	-10.7	-23.7	-11.5	-11.9	-13.3	-50.9	-11.3	-11.3	-11.2
MPS04	-13.8	-22.9	-10.9	-11.2	-13	-49.3	-10.4	-12.4	-10.4
MPS04_after	-13.8	-22.9	-10.9	-11.2	-13	-49.3	-10.4	-12.4	-10.4
RI	-10.4	-24.2	-11	-12.6	-13	-48.3	-11	-11.5	-10.6
ALM_IT	-37.6	-26.4	-11.2	-12.9	-13.8	-182.8	-9.4	-12.3	-9.6
HRSS_m1	-10.9	-23.7	-11.5	-12.3	-13.6	-53.3	-11	-12.2	-11.4
HRSS_m2	-10.9	-23.2	-11.2	-11.6	-13	-49.7	-11.7	-11.4	-11
TripleS_CPTI	-11.1	-22.8	-11.1	-11.3	-12.9	-49.8	-11.2	-12	-10.8
TripleS_CSI	-11.8	-23.9	-11.7	-11.8	-13.5	-52.9	-11	-12.6	-11.2
TripleS_Hyb	-11.4	-23.3	-11.3	-11.5	-13.2	-51.2	-11.1	-12.3	-11.1

Table S2. The log-likelihood of each 5-year model in the spatial bin where target earthquakes occurred. The number of target earthquakes that occurred in each spatial bin are reported in the first row. From the table we can notice that some models (ALM_IT, ALM; HALM; and PHM_zone) failed the *S*-test because they score very low log-likelihood (grey cells) in at least one spatial bin where target earthquakes occurred. Conversely, the other models that failed the *S*-test (see Table 4 in the paper) performed poorly in the spatial bins without target earthquakes. This can be noted, for example, by comparing the HZA_TD and PHM_grid models. The first model passes the consistency tests while the second one does not; however, they have similar log-likelihoods in the spatial bins where target earthquakes occurred. Finally, we notice that MPS04 and MPS04_after models – which inform the seismicity rate model of the national seismic hazard model – do not perform well for the first and the second-to-last target earthquakes, but the poor performance in the S-test is likely due to a poor performance on the spatial cells where no target earthquakes occurred.