Building self-consistent, short-term earthquake probability (STEP) models: improved strategies and calibration procedures

Jochen Woessner[1,*], Annemarie Christophersen[1,3], J. Douglas Zechar[1,2], Damiano Monelli[1]

[1] *ETH Zurich, Swiss Seismological Service, Zurich, Switzerland*

[2] *Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, USA*

[3] *GNS Science, Avalon, Lower Hutt, New Zealand*

Article history
Received October 23, 2009; accepted April 13, 2010.
Subject classification:
Earthquake statistics, Earthquake forecasting, Likelihood testing, Aftershock model, Seismicity analysis.

ABSTRACT

We present two self-consistent implementations of a short-term earthquake probability (STEP) model that produces daily seismicity forecasts for the area of the Italian national seismic network. Both implementations combine a time-varying and a time-invariant contribution, for which we assume that the instrumental Italian earthquake catalog provides the best information. For the time-invariant contribution, the catalog is declustered using the clustering technique of the STEP model; the smoothed seismicity model is generated from the declustered catalog. The time-varying contribution is what distinguishes the two implementations: 1) for one implementation (STEP-LG), the original model parameterization and estimation is used; 2) for the other (STEP-NG), the mean abundance method is used to estimate aftershock productivity. In the STEP-NG implementation, earthquakes with magnitude up to $M_L = 6.2$ are expected to be less productive compared to the STEP-LG implementation, whereas larger earthquakes are expected to be more productive. We have retrospectively tested the performance of these two implementations and applied likelihood tests to evaluate their consistencies with observed earthquakes. Both of these implementations were consistent with the observed earthquake data in space: STEP-NG performed better than STEP-LG in terms of forecast rates. More generally, we found that testing earthquake forecasts issued at regular intervals does not test the full power of clustering models, and future experiments should allow for more frequent forecasts starting at the times of triggering events.

Introduction

Constructing effective real-time, short-term earthquake forecasts remains one of the most challenging problems for seismologists. To provide useful information, a model must capture earthquake physics with a complex physical and/or statistical understanding of spatial and temporal clustering. The difficulties are compounded as these clustering processes operate on different scales in the space, time and magnitude domains. Currently, only statistical forecast model frameworks such as epidemic type aftershock sequence (ETAS) models and short-term earthquake probability (STEP) models are used for automated, near-real-time applications [e.g., Console et al. 2003, Gerstenberger et al. 2005, Helmstetter et al. 2006, Marzocchi and Lombardi 2008]. Both of these frameworks can adapt to ongoing earthquake sequences by re-estimating model parameter values and automatically generating forecasts that account for the most recent seismicity. Physics-based models that combine calculations of stress changes with a rate-and-state friction model to determine seismicity rates [e.g., Hainzl et al. 2009, Cocco et al. 2010] are not yet applicable in near real-time. These models require additional seismological, geological, and tectonic information that is often not immediately available. There are indications from retrospective testing experiments that these models can perform as well as the statistical models only when uncertainties of the physical model are included stochastically [Woessner et al. 2009].

Motivated by (1) the upcoming prospective daily seismicity forecast experiment to be performed at the ETH Zurich Testing Center as part of the Collaboratory Study for Earthquake Predictability (CSEP) initiative, (2) recent results from retrospective comparative testing experiments [Woessner et al. 2009], and (3) methodological improvements in estimating triggering effects of earthquakes [Christophersen and Smith 2008, Christophersen and Gerstenberger 2010], we have developed two implementations of the STEP model for the Italian testing region. Additionally, we suggest that with some regionalization, these two implementations can be tested within other CSEP testing regions. Moreover, dependent on the forthcoming results of the prospective tests of these implementations, time-varying hazard estimates can be generated for Italy, similar to those already publicly available in real-time for the state of California, USA (http://earthquake.usgs.gov/eqcenter/step/).

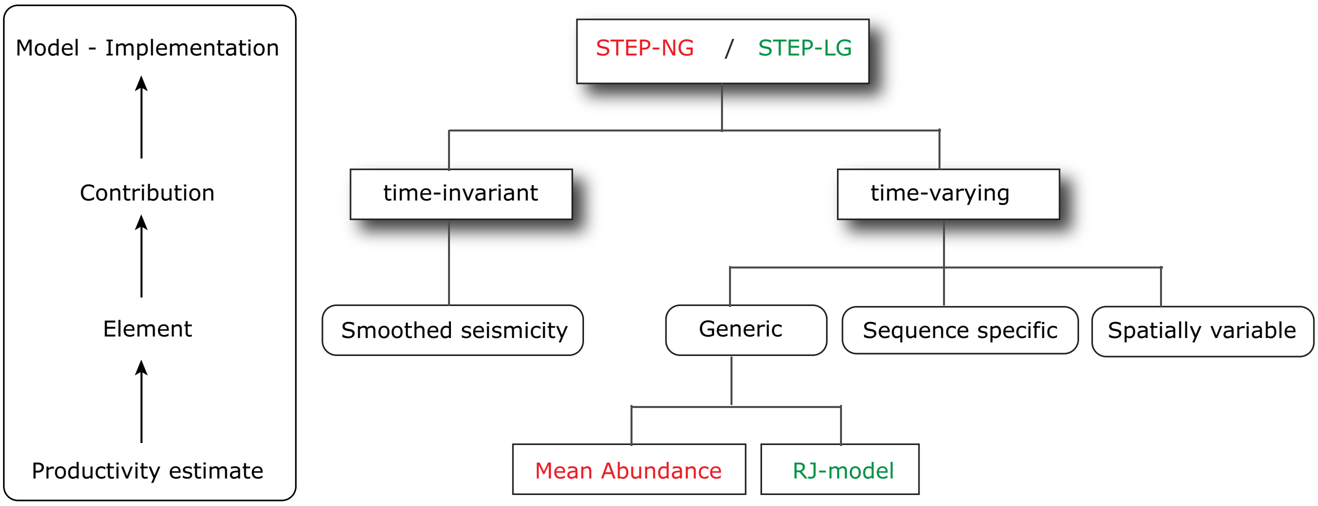

Figure 1. STEP model hierarchy. The model implementations (STEP-NG/STEP-LG) are composed of model contributions (time-invariant/time-varying), which consist of model elements. The elements of the time-varying contribution were weighted using an AICc criterion [Gerstenberger et al. 2005]. The generic element can be based on the mean abundance model by Christophersen and Smith [2008] or the model by Reasenberg and Jones [1989] (RJ-model). For the latter, we used parameters determined by Lolli and Gasperini [2003] and Gasperini and Lolli [2006].

When building an earthquake forecast based on a STEP model, a time-invariant contribution is computed from previous seismicity and, if desired, other information. The time-varying contribution is based on estimating parameter values from the analysis of previous seismicity. The two contributions are estimated independently; therefore, to build a self-consistent model, the same underlying assumptions are made. Our implementations combine a time-invariant contribution based on a simple smoothed seismicity approach [Zechar and Jordan 2010b] with two slightly different time-varying contributions (Figure 1); these different paths are what distinguish the two implementations. Each time-varying contribution is itself a combination of three elements: (1) a generic element based on the average statistical behavior of Italian aftershock sequences; (2) an element based on the temporal behavior of the particular aftershock sequence in progress; and (3) a spatially variable element where the aftershock behavior is mapped in space. One time-varying contribution closely follows the methodology of Reasenberg and Jones [1989, 1990, 1994], using parameter values estimated by Lolli and Gasperini [2003] and Gasperini and Lolli [2006]. The other time-varying contribution uses a new generic element based on Christophersen and Gerstenberger [2010]. We describe the latter model in detail and discuss its capabilities and limitations. Each of these contributions is then combined with the time-invariant contribution, forming the STEP-LG (Lolli-Gasperini) and the STEP-NG (new generic) model. When combining the contributions, the final forecast rate at each point is taken to be the greater of the time-invariant and time-varying contributions.

In the following sections, we describe the data used, the time-invariant contribution, and the two distinct time-varying contributions. We report the performance of the two implementations in a retrospective daily forecast experiment of 906 days starting on January 1, 2007. The performance of each implementation is analyzed with the CSEP likelihood(L)-Test [Schorlemmer et al. 2007], a modified number(N)-Test, and a normalized L-Test and a space(S)-Test [Werner et al. 2009, Zechar et al. 2010]. As these tests were developed for evaluating long-term forecasts, we have adjusted them to the needs of daily forecasts. We discuss the test results, state our expectations for prospective testing, comment on how the experiment set-up influences the model performance, and outline possible improvements to the current model implementations.

Data

We used seismic catalog data provided by the CSEP EU Testing Center (http://www.cseptesting.org/regions/italy) for the time period from January 1, 1981, to December 31, 2002 (CSI 1.1 catalog) [Castello et al. 2007]. We used the Italian seismic bulletin (Bollettino Sismico Italiano; http://bollettinosismico.rm.ingv.it/) for the period from 2002 to March 31, 2009, by adding the data into the periods that were not available from the CSEP Testing Center in Zurich. From April 1, 2009, to June 25, 2009, we used data from the Italian Seismic Instrumental and parametric Data-basE (ISIDe; http://iside.rm.ingv.it), knowing that this data had not been reviewed entirely.

The catalog originally contained 75,520 earthquakes in the magnitude range $0.1 \leq M_L \leq 5.9$ with depths less than 30 km. The catalog included 436 events with $M_L \geq 3.95$, 406 of which were located in the collection area of the CSEP-Italy testing region. The $M_L \geq 3.95$ seismicity is shown in Figure 2A, together with the cumulative number of events versus time (Figure 2B). To minimize possible edge effects, earthquakes in the catalog that fell outside the collection area were included when estimating the time-invariant parameter values. Figure 2A shows 279 events in the learning period for

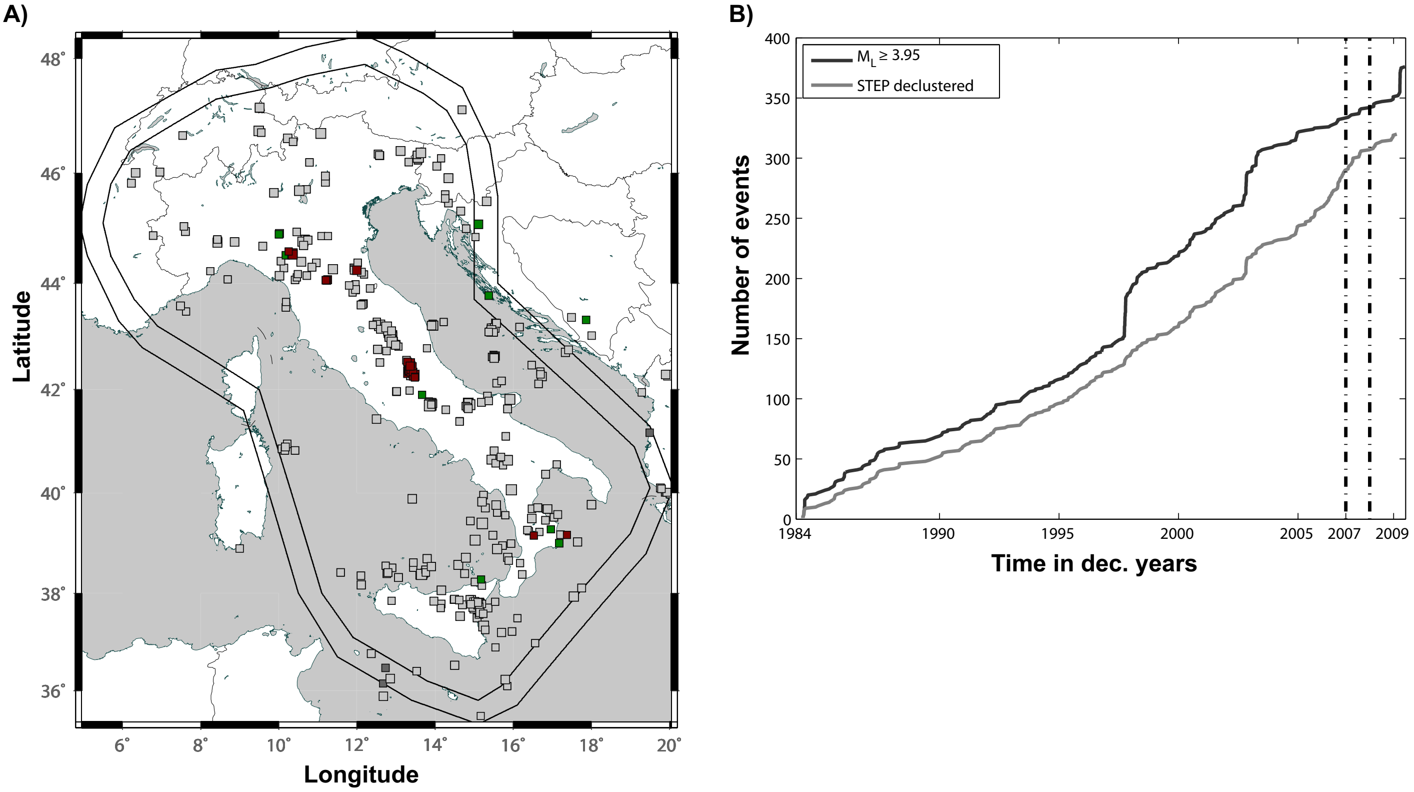

Figure 2. (A) Seismicity map for events in the STEP-declustered catalog (light gray squares) and entire catalog (dark gray squares) for the period 1984 to 2006. Events in target period for the smoothed seismicity map (green squares: January 1, 2007, to December 31, 2007) and target period of retrospective testing (red squares: January 1, 2007, to June 25, 2009). The inner polygon represents the testing area, the outer polygon, the collection area, for the prospective CSEP testing experiments. (B) Cumulative number of events *versus* time for the entire catalog (black line) and for the STEP-declustered catalog (gray line). Only events with $M_L \geq 3.95$ are plotted.

the time-invariant parameters, and 31 in the target period of January 1, 2007, to December 31, 2007. There were 37 events in the target period for the retrospective daily experiment from January 1, 2007, to June 25, 2009.

Model description

In building our implementations, we strove to use self-consistent assumptions and data that were well suited for estimating the parameter values for each model contribution. Nevertheless, our implementations are not completely free of subjective choices, as the parameters were not estimated in one single, simultaneous procedure.

We modified several aspects of the STEP model compared to the version implemented for California, USA [Gerstenberger et al. 2005]: (1) We regionalized the model by using the Italian testing region defined for the 1-day model class by the testing center [Schorlemmer et al. 2010, this issue]. (2) We computed the time-invariant model contribution using different assumptions than Gerstenberger et al. [2005]: rather than using the seismicity rates of the national Italian seismic hazard map [Meletti et al. 2008], we independently estimated the time-invariant contributions using a smoothed seismicity approach. (3) We implemented two distinct time-varying contributions: (a) one based on the work of Reasenberg and Jones [1994], with parameter values estimated by Lolli and Gasperini [2003] and Gasperini and Lolli [2006], essentially following the procedure used by Gerstenberger et al. [2005]; (b) another based on the mean abundance model of Christophersen and Smith [2008] and Christophersen and Gerstenberger [2010], to estimate aftershock productivity.

Time-invariant contribution

For the generation of the time-invariant contribution, we estimated seismic background rates using a smoothed seismicity approach, assuming that the background seismicity rates vary in space and time [Marzocchi and Lombardi 2008]. The smoothed seismicity approach was adopted from Zechar and Jordan [2010b] using a smoothing kernel with an isotropic two dimensional Gaussian function governed by a single length scale parameter. By smoothing the recent seismic activity, we assumed to better approximate the current state of background seismicity, as opposed to using long-term rates derived in the national seismic hazard map that rely on seismic source zonation [Meletti et al. 2008]. In this sense, each contribution was based only on the earthquake catalog; smoothing the seismicity rates and the resulting implementations are built on self-consistent assumptions.

To compute the time-invariant contribution, we first declustered the catalog containing events in the period January 1, 1981, to December 31, 2007. For comparison, we applied several different clustering approaches and show the

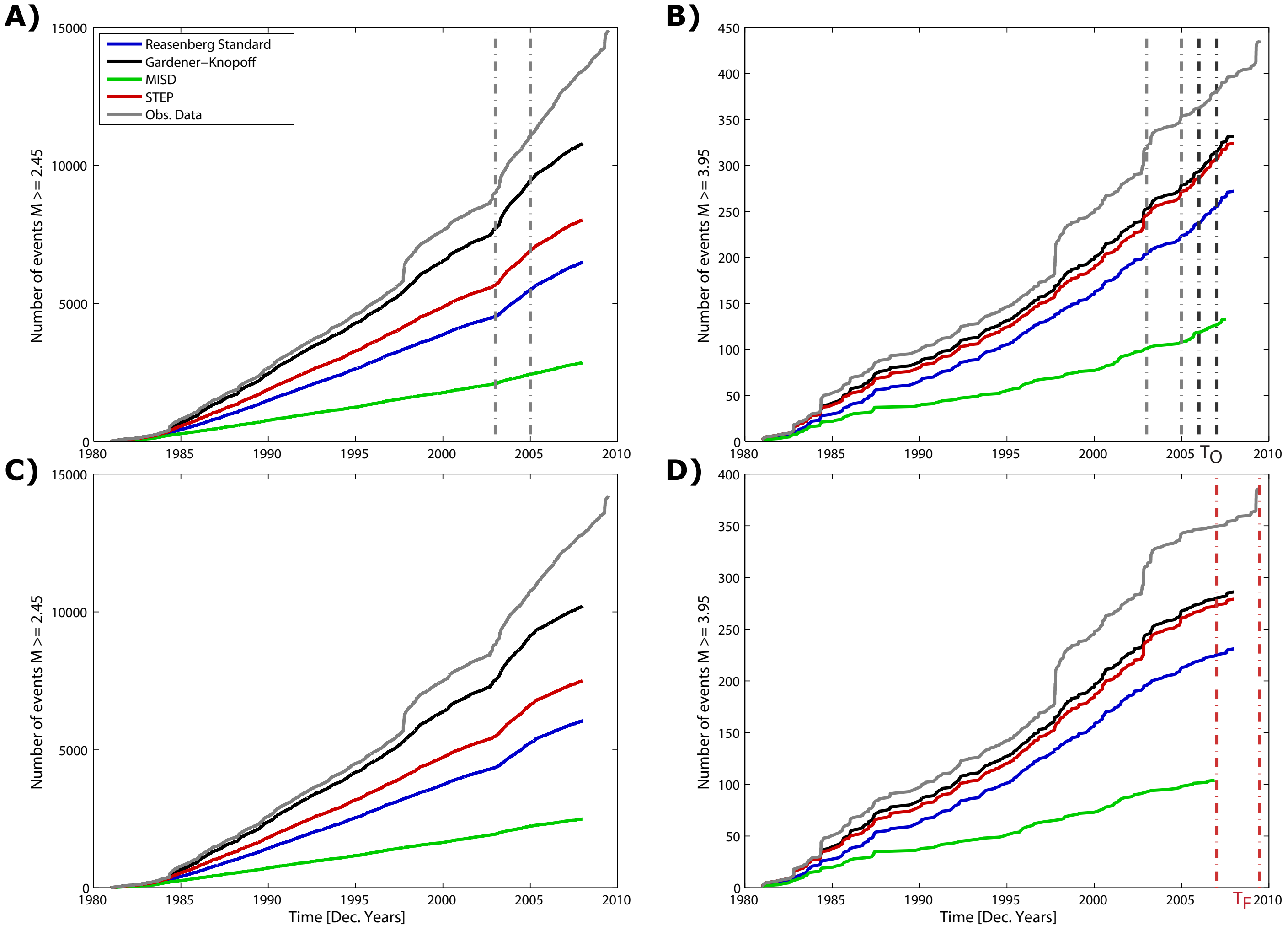

Figure 3. Cumulative number of earthquakes *versus* time of the entire observed data (gray) and resulting from four (de-)clustering approaches applied to the catalog from January 1, 1981, to December 31, 2007: for the entire study region (see Figure 2) with (A) $M_L \geq 2.45$ and (B) $M_L \geq 3.95$; for the testing region with (C) $M_L \geq 2.45$ and (D) $M_L \geq 3.95$. The dashed gray lines in (A) and (B) indicate the period of 2003 to 2005 for which the magnitude determinations appeared to be different, although not affecting magnitudes $M_L \geq 3.95$. Dash-dotted black lines in (B) denote the period T_O when the time-invariant model contribution was optimized. Dash-dotted red lines in (D) denote the period of retrospective testing T_F. Note the smaller slope in this period compared to previous periods.

inherent variability (Figure 3). We generated four declustered catalogs using: (1) a windowing approach based on Gardner and Knopoff [1974]; (2) the clustering approach by Reasenberg [1985], with the originally proposed parameters for California, USA; (3) the model-independent approach of Marsan and Lengliné [2008]; and (4) the windowing approach as used in the time-varying STEP contribution. In approach 4, any earthquake that followed another within a certain space-time range was considered to be an aftershock of the first event. The time window was 30 days, and it was extended by another 30 days if another aftershock occurred. The spatial window was a circular area, centered on the potential main shock with radius $r(m)$:

$$r(m) = \begin{cases} 5 & : m \leq 5.3 \\ 10^{059m - 2.44} & : m > 5.3 \end{cases} \qquad (1)$$

where m is the local magnitude M_L of the potential main shock, and $r(m)$ is measured in km. This relation is based on Wells and Coppersmith [1994], and involves a minimum 5 km radius to select events taking into consideration location uncertainties; this implies that for $M_L \leq 5.3$ the search radius is constant and thus the smaller the magnitude is, the more likely are potential background earthquakes to be included in a cluster.

We emphasize that the choice of the declustering procedure did affect the time-invariant contribution in terms of total forecast rate and the resulting spatial forecast. For consistency of the two contributions, the implementations presented here were based on the STEP windowing approach (approach 4). The cumulative numbers of earthquakes for the entire catalog and for each declustered catalog as a function of time are shown in Figure 3A, B, respectively; for the testing area, as indicated in Figure 1, this is shown in Figure 3C, D. Data from the entire catalog were used to compute the time-invariant model. The total number of events varied between 117 and 304 for the entire catalog, and 89 and 259 for the testing area (Table 1). This implies a daily seismicity rate for the entire area of between 0.0134 and 0.0338, and for the testing area, 0.010 and 0.0296. For the selected STEP-declustered catalog, 296 events with $M_L \geq 3.95$ remained,

Total in the area catalog spans $N\ (M \geq 3.95)$	Yearly rate $N_{Obs}(M \geq 3.95)$ / [year]	Daily rate $N_{Obs}(M \geq 3.95)$ / [day]	Reference
296/ 252	12.333/ 10.500	0.0338/ 0.0288	STEP windowing
304/ 259	12.667/ 10.792	0.0347/ 0.0296	Reasenberg [1985]
250/ 210	10.417/ 8.750	0.0285/ 0.0240	Gardner and Knopoff [1974]
117/ 89	4.875/ 3.708	0.0134/ 0.010	Marsan and Lengliné [2008]

Table 1. Total number of events, yearly rates ($M \geq 3.95$) and daily rates ($M \geq 3.95$) after declustering for the collection/testing area of the INGV catalog in the period 1984-2007.

which indicated a total daily rate of 0.0338; in the testing area, a yearly/daily rate of 10.5/0.0288 events was expected.

The overall seismicity rate for the period 2003 to 2005 was higher than in the other periods; however, the magnitude range $M_L \geq 3.95$ was only slightly affected compared to the range of smaller magnitude events. In Figure 3, a change in the slope can also be seen around March 29, 2003. On this day, the magnitude $M_L = 5.4$ Jabuka island event occurred in the central Adriatic sea, off the coast of Croatia; this was one of the strongest events ever recorded within the Adriatic microplate [Herak et al. 2005].

In Figure 3D, the vertical lines depict the period of the retrospective daily forecast tests (January 1, 2007, to June 25, 2009). The slope of the observed number of earthquakes in the cumulative number plot for the observed data was slightly less steep than in the periods before the 1997 Umbria-Marche earthquake and in other periods without large aftershock sequence contributions. This might be due to the natural temporal fluctuation of seismicity, or it might be due to changes in policies regarding magnitude determination. This is important to remember when interpreting the results of the retrospective testing experiment described in the following sections.

Parameter	STEP-LG	STEP-NG
a-value	−1.84 ±0.12	−1.84 ±0.12
b-value	0.98 ±0.03	0.98 ±0.03
c-value	0.09 ±0.27	0.09 ±0.27
p-value	0.92 ±0.06	0.92 ±0.06
α-value		1.31 ±0.21
M_1		5.39 ±0.11
I_{OU}		5.468
M_c correction	0.2	0.2
$N_{min,SS}$	100	100

Table 2. Model parameters and standard deviations for STEP-LG and STEP-NG. All other models were used as retrospective comparisons. M_c correction, adjustment factor for data quality based on Woessner and Wiemer [2005]. $N_{min,SS}$, minimum number to estimate parameter values for the sequence specific and spatially variable time-varying model elements.

The STEP-declustered catalog served as the input to the smoothing algorithm of Zechar and Jordan [2010b]. Similar approaches have been used for national seismic hazard maps in the USA and New Zealand [Frankel et al. 2002, Stirling et al. 2002]. In contrast to the adaptive kernel estimation method by Stock and Smith [2002a, 2002b], in which the kernel width varied in space, here the standard deviation σ of the Gaussian kernel was uniform over the domain of interest. The novelty of the method with respect to past implementations, such as Frankel et al. [2002], is that the kernel width was optimized by performing a set of retrospective tests with different smoothing length scales and by measuring the performance of each test in terms of the area skill score (ASS) misfit statistic [Zechar and Jordan 2008, Zechar and Jordan 2010a]. In other words, the kernel width that gave the best performance in terms of the ASS misfit statistic χ was chosen as the optimal one.

To determine the optimal kernel width, we used data from the declustered catalog over the period from January 1, 1984, to December 31, 2006, as the learning period, and the period from January 1, 2007, to December 31, 2007, as the target period. For magnitude $M_L \geq 3.95$, the learning period contained 296 events, and the target period, 19 events. The kernel width σ was varied between 5 km and 200 km, and $\chi(\sigma)$ was calculated (Table 2). The optimization procedure suggested an optimal smoothing length scale of 30 km. The seismicity rates of the smoothed time-invariant model for the testing region are shown in Figure 4.

Time-varying contribution

The STEP model includes a spatial extension of the simple aftershock model of Reasenberg and Jones [1989, 1990, 1994]:

$$\lambda(t, M) = \frac{10^{a' + b(M_m - M_{th})}}{(t + c)^p} \tag{2}$$

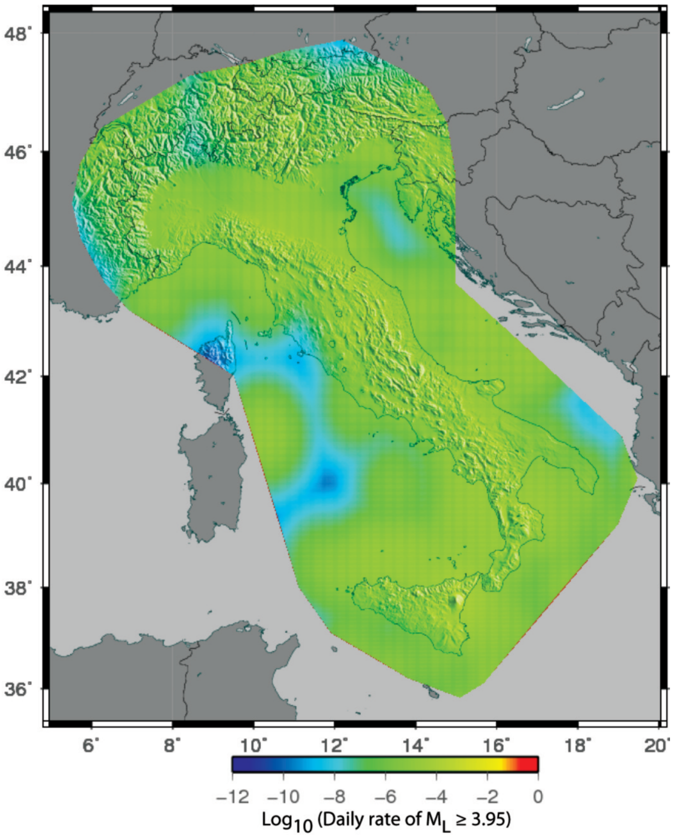

Figure 4. Log$_{10}$ (daily rate of $M_L \geq 3.95$) for the time-invariant model, derived from the smoothed seismicity approach of Zechar and Jordan [2010b]. The optimum smoothing kernel width was 30 km (Table 3), derived from the STEP-declustered catalog for the period of January 1, 1984, to December 31, 2006, as learning, and the period of January 1, 2007, to December 31, 2007, as forecasting.

Kernel width σ [km]	Area skill score misfit, χ
5	0.21252891
10	0.17705183
20	0.10374542
25	0.07435822
30	0.04899255
50	0.07346651
75	0.10409969
100	0.12572704
200	0.21312013

Table 3. Kernel estimates for increasing kernel widths (σ, standard deviation). The lowest value of χ indicates the best solution.

where λ (t, M) is the rate of aftershocks with magnitudes greater than the magnitude threshold M_{th} and occurring at time t, M_m denotes the main shock magnitude. The constants a' and b were derived from the Gutenberg-Richter relationship [Gutenberg and Richter 1944], and p and c resulted from the Omori-Utsu law [Utsu 1961, Ogata 1983]. As aftershock sequences progress, the model parameter values were re-estimated.

We applied the corrected Akaike information criterion (AICc) to construct the best-fitting model from the three elements in the time-varying contribution (generic, sequence specific, spatially varying; see Figure 1). The AICc is a likelihood-based metric designed for model selection [Kenneth et al. 2002]. The calculation took into account the number of free parameters and the number of observed data; to be preferred in the final scoring, a model with more free parameters must fit the data better than models with fewer parameters. Rather than selecting only a single model element, we used AICc weighting, where a relative weight for each model was based on its AICc score, and the final model was a weighted sum of the three element contributions; details of the STEP model are described in Gerstenberger et al. [2004]. One essential component is the spatial smoothing of the aftershock productivity parameter a', which defines the total rate of aftershocks in a sequence. We applied a smoothing of $1/r^2$, with r as the distance from the point of interest to the point source in the generic model, or the empirically parameterized fault in the sequence-specific and the spatially variable model. Thus, for the generic model, the rates were smoothed radially outwards, while they were smoothed away from the fault in the other model elements. For the CSEP Italy implementation, where the model generates daily forecasts of earthquakes with magnitude $M_L \geq 3.95$, we assumed that only $M_L \geq 3$ events can trigger events of interest. In the current model, with the generic parameters for the Italian catalog, there is a 4% chance that an earthquake of $M_L = 3$ will trigger an event with $M_L \geq 3.95$ (for $M_L = 2.5$ and $M_L = 2$; the corresponding probabilities are 1% and 0.1%, respectively). Leaving the smaller events out, we probably underestimated the rate of occurrence; however, this covers the intrinsic uncertainty of the model forecasting ability well.

We implemented two variations of the time-varying contribution: STEP-LG and STEP-NG. The two implementations differed only in the description of the average aftershock productivity in the generic model element (Figure 1); all other features were in common. For each sequence analyzed, we estimated the magnitude of completeness, M_c, using the maximum curvature approach without bootstrapping [Woessner and Wiemer 2005]. We added a correction factor to each estimate of $M = 0.2$ units, because the maximum curvature approach tended to yield unreasonably low M_c values. If at least 100 events with magnitudes larger than M_c were available for an ongoing sequence, the sequence-specific estimates of Gutenberg-Richter and Omori-Utsu parameter values were computed using maximum likelihood estimators [Bender 1983, Ogata 1983].

STEP-LG model: Reasenberg-Jones model with parameters by Lolli-Gasperini

Equation 2 introduced the generic model element for the STEP model [Gerstenberger et al. 2005], for which the productivity parameter k in the Omori-Utsu law is replaced, to describe an average productivity that increases with mainshock magnitude. Lolli and Gasperini [2003] used combined earthquake catalog data from 1960 to 1996 to estimate the values of these model parameters. They applied the clustering algorithm by Reasenberg [1985], as well as a window method, to define aftershocks. They estimated parameter values from about 40 aftershock sequences. Finally, the parameter a' was determined by equating the numerator of Equation 2 to k and solving for a':

$$a' = \log k - b(M_m - M_{th}) \tag{3}$$

The median values were: $a' = 1.84 \pm 0.12$, $p = 0.92 \pm 0.06$, $c = 0.09 \pm 0.27$, and $b = 0.98 \pm 0.03$ (Table 2).

STEP-NG model: new generic model

The STEP-LG implementation assumed that aftershock productivity increases with main shock magnitude and depends on the b-value of the frequency-magnitude distribution. Christophersen and Gerstenberger [2010] derived an alternative description for the average productivity as a function of main shock magnitude based on mean abundance, $N_{ma}(m)$, the mean number of aftershocks for a main shock with magnitude m. Estimating mean abundance values has its own challenges and here we briefly describe how we determined our mean abundance for Italy.

To estimate the mean abundance parameter values, we used data covering the time period of January 1981 to December 2002, as covered by the CSI 1.1 catalog [Castello et al. 2007]. Near the beginning of this period, the seismic network changed significantly. As a result, the detection threshold decreased from around $M_L = 4.0$ to $M_L = 3.0$ and below by the middle of 1984. Therefore, we analyzed the complete catalog with a threshold magnitude $M_{th} = 4$, and also considered the period 1984.5 to 2002 with threshold magnitude $M_{th} = 3$.

We clustered the earthquakes according to the STEP windowing approach, as outlined above, with an upper time limit of 30 days. Once the earthquake clusters were defined, the largest earthquake in a cluster was considered to be the main shock. If two or more earthquakes within one cluster had the same magnitude, the earliest of these was taken to be the main shock. To determine mean abundance, the number of main shocks in each 0.1 magnitude bin was counted; the total number of aftershocks within a chosen time period was counted; and the number of aftershocks per main shock magnitude bin was divided by the observed number of main shocks. To address issues of completeness, we started counting aftershocks 0.1 days (2 hours and 24 minutes) after the main shock, to avoid missing smaller aftershocks that might be hidden in the coda of the main shock.

It has been shown that, in general, mean abundance grows exponentially with main shock magnitude M_m:

$$N_{ma}(M_m, M_{th}) = 10^{\alpha(M_m - M_1(M_{th}))} \tag{4}$$

where α is the growth exponent, and $M_1(M_{th})$ is the magnitude, which on average has one aftershock above the threshold magnitude M_{th} [Christophersen and Smith 2008].

Figure 5A shows the results from the mean abundance analysis for two sub-sets of the catalog for the period 1981 to 2002: The gray rectangles are the mean abundance in the time period from 1984.5 with $M_{th} = 3.0$. Each data point was scaled by 0.1 to match a threshold magnitude of 4.0 for the lower target magnitude of the daily CSEP testing class. This scaling corresponded to a magnitude unit difference with a b-value of $b = 1.0$. The black triangles are data for the complete time period, with $M_{th} = 4.0$. We report the best-fitting α and M1 values and their respective 95% confidence intervals for the two periods: in the first case ($M_{th} \geq 4.0$), $\alpha = 1.31 \pm 0.21$ and $M_1(M_{th} = 4.0) = 5.39 \pm 0.1$; in the second case, we obtained $\alpha = 0.87 \pm 0.12$ and M_1 $(M_{th} = 4.0) = 5.63 \pm 0.13$. Due to the fixed search radius of 5 km for all earthquakes below magnitude $M_L = 5.3$, background events could well have been included, biasing the data towards higher abundance. Therefore, although the dataset with magnitude cut-off $M_{th} = 4.0$ had fewer sequences and more scatter than the events with magnitudes in the range 3-4, we took the parameter value estimates based on the $M_L \geq 4.0$ set because the data were less likely to be affected by possible background events.

Mean abundance can be related to the Omori-Utsu k-value [Utsu, 1961, Ogata 1983] by integrating over the time interval used to estimate mean abundance N_{ma}:

$$N_{ma} = \int_S^T n(t)\, dt = k(M_{th}) \int_S^T (t + c)^{-p}\, dt = k(M_{th}) I_{OU}(S, T) \tag{5}$$

Here, S and T are the start and end times of the period analyzed, respectively. We call I_{OU} (S,T) the Omori-Utsu integral, with:

$$I_{OU} = \begin{cases} 1n\left[\dfrac{T + c}{S + c}\right] & : p = 1 \\ \dfrac{(T + c)^{1-p} - (S + c)^{1-p}}{1 - p} & : p \neq 1 \end{cases} \tag{6}$$

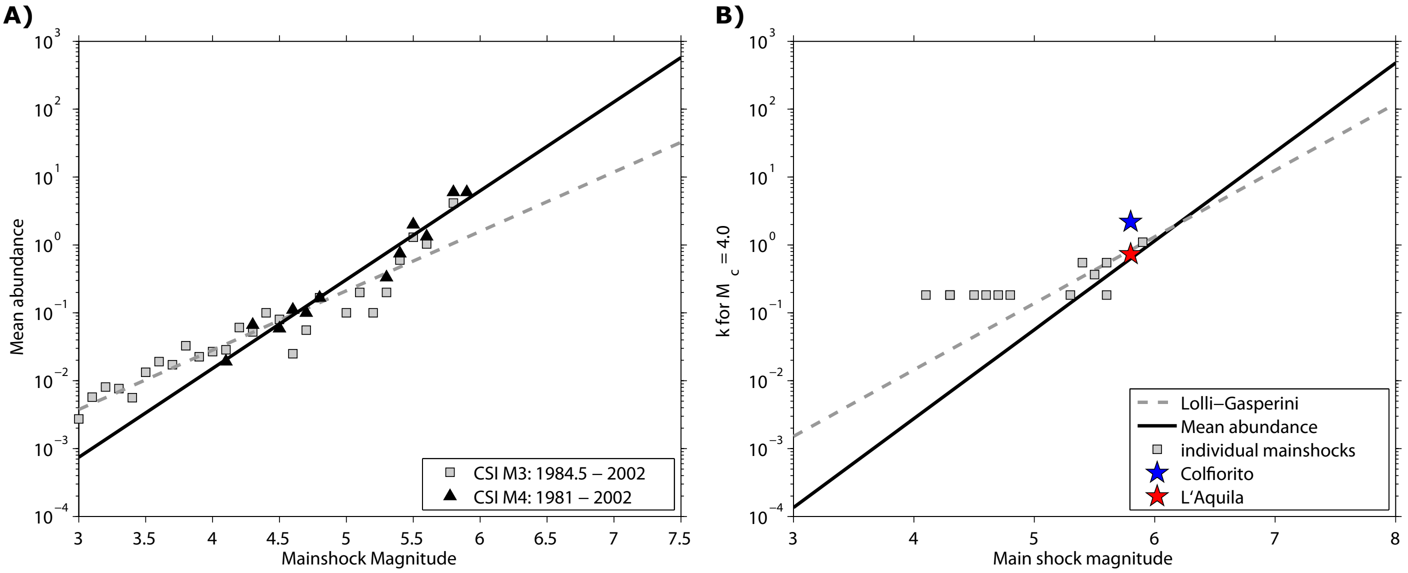

Figure 5. Mean abundance analysis. (A) Mean abundance N_{ma} as a function of magnitude for the period of 1981 to 2002 with threshold magnitude $M_{th} = 4$ (black triangles) and the period 1984.5 to 2002 with threshold magnitude $M_{th} = 3$ (gray squares). The solid black line and the gray dashed line show the fit of the mean abundance model for to the data, respectively. (B) The *k*-value for threshold magnitude $M_{th} = 4$ as a function of magnitude. The gray dashed line and solid black line show model predictions based on Gasperini and Lolli [2006] and the mean abundance model, respectively. Gray squares show productivity of sequences available, those with zero events in the magnitude range are not shown. Blue and red stars: 1997 Colfiorito and 2009 L'Aquila sequences.

For the Lolli-Gasperini parameter values $p = 0.92$ and $c = 0.09$, for the mean abundance time interval of 0.1 to 30 days, we obtained $I_{OU} = 5.486$. Figure 5B shows the mean abundance estimates of *k* obtained by dividing N_{ma} by I_{OU}. The individual main shock data were computed by dividing the number of aftershocks in the time interval of 0.1 to 30 days by I_{OU}. The data for the 1997 Colfiorito and the 2009 L'Aquila earthquakes are highlighted by blue and red stars, respectively. If only one aftershock was observed, the *k*-value was 0.20. If no aftershocks were observed, the main shock was not shown in Figure 5B. As a consequence, the squares fall above the model line, which also includes the main shocks without aftershocks in the time and magnitude interval analyzed. The final parameter values of the STEP-NG implementation are listed in Table 2.

Figure 5B indicates that the 1997 Colfiorito sequence was particularly productive, while the 2009 L'Aquila sequence fell in between the predictions of the two models. In the magnitude range 5.3 to 5.8, the data were scattered around both models, and both models agreed within the scatter of the data. Due to the difference in slope (0.98 for the RJ-model, and 1.31 for the mean abundance model), the models deviated for smaller and larger main-shock magnitudes. For a magnitude $M_L = 3$ event, the *k*-value for the RJ-model was about one order of magnitude larger than the *k*-value for the mean abundance model. Because these earthquakes are used for forecasting, and because there are many more of these than larger earthquakes, the STEP-LG implementation will predict an overall higher rate of seismicity than the STEP-NG implementation. The productivity of the main shocks based on Gasperini and Lolli [2006] was higher for events with magnitudes up to $M_L \leq 6.2$. For larger events, the mean abundance model yielded higher productivity estimates.

Retrospective testing

We performed retrospective daily tests for both models for the period January 1, 2007, to June 25, 2009 (906 days). We forecast 24-hour seismicity rates for the Italian testing region in the magnitude range $4 \leq M_L \leq 8$, with the last bin including rates up to $M_L = 9$. The first forecast estimated seismicity rates from midnight on January 1, 2007, to midnight on January 2, 2007, the second forecast covered the following 24 hours, and so on. We measured the performance of the models with the modified N-Test [Zechar et al. 2010], the CSEP L-Test [Schorlemmer et al. 2007], a normalized L-Test [Werner et al. 2009] and an S-Test [Zechar et al. 2010]. With retrospective testing, we hoped to identify weaknesses in the forecasts and gain some insight as to what we should expect from prospective forecast experiments. We considered various tests because each analyzed different features of the forecast.

N-Test and modified N-Test

The two-sided N-Test of Schorlemmer et al. [2007] contains a subtle flaw: in some cases, the test rejects forecasts with low rates when there are zero earthquakes observed, although such a forecast should be considered consistent. Therefore, we followed the suggestion by Zechar et al. [2010] and applied the modified N-Test, summarized by the metrics:

$$\delta_1 = 1 - F(N_{Obs} - 1 \mid N_F) \text{ , and} \tag{7}$$

$$\delta_2 = F(N_{Obs} \mid N_F) \tag{8}$$

Here, $F(x\,|\,\mu)$ is the right-continuous Poisson cumulative distribution function with expectation μ evaluated at x. Using this approach, we can answer two questions separately: (1) What is the probability of observing at least N_{Obs} earthquakes (δ_1); and (2) What is the probability of observing N_{Obs} at most earthquakes (δ_2).

The forecast rate distribution for all CSEP rate forecasts was assumed to be Poisson, in which case it was appropriate to perform a two-sided hypothesis test to determine whether the forecast rate was too high or too low. When the probabilities were written as in Equations 7 and 8, we instead used a one-sided test with a scaled critical value. For example, a critical value of 0.025 corresponds to a hypothesis test for which we have 95% confidence.

The flaw in the original N-Test was critically important for daily forecasts, where the number of observed events on any given day was often zero. For our implementation, the daily rate forecast of the time-invariant contribution was N_F = 0.0288 (Table 1) for the entire testing region. On most testing days, there will be no magnitude $M_L \geq 3.95$ event ($N_{Obs} = 0$). For these days, we obtained $\delta_1 = 1$ and $\delta_2 = 0.9716$, indicating that the forecast was consistent with the observation. Applying the original approach of a two-sided N-Test, this forecast would be very close to rejection starting at $\delta = 0.975$.

Consistency tests in space: L-Test, normalized L-Test, and S-Test

We considered three different tests to analyze the consistency of the retrospective forecasts with the data in the spatial and magnitude distribution. We applied the CSEP L-Test (γ- score) and a normalized L-Test (γ_{norm}). The γ-statistic suggested by Schorlemmer et al. [2007] measures the consistency of a forecast space-rate-magnitude distribution with the observations. The γ-score is dependent on the number of observed events (N_{Obs}). Therefore, a forecast that is consistent in terms of the spatial and magnitude distribution with the observed earthquakes might be rejected by the L-Test simply because it does a poor job at forecasting the overall rate of seismicity. To account for this scenario, we also applied the normalized L-Test suggested by Werner et al. [2009], which normalizes the test results by the number of observed target earthquakes. Low values of γ indicate that a space-rate-magnitude forecast is not consistent with the observed distribution, whereas low values of γ_{norm} indicate that a forecast is not consistent in terms of the space-magnitude distribution. The S-Test (ζ-score) isolates the spatial component of the forecast and compares this with the spatial distribution of target earthquakes [Zechar et al. 2010]; low values of ζ indicate a forecast spatial distribution is inconsistent with the observation.

Results

To provide a first overview, Figure 6 shows for both models the cumulative numbers of the forecast earthquakes, the time-invariant model contribution, and the cumulative numbers of observed earthquakes versus time. The time-invariant contribution (dash-dotted line) has a steeper slope as compared to the observed data (light gray line). As indicated in the data section, the rate of observed seismicity was lower for the testing period as compared to similar periods, not including prominent aftershock sequences (see Figure 3D). The reason for this might be natural rate fluctuations, the result of optimizing the model only for a one-year period, or the change in network policies to determine magnitudes. We consider the last possibility least likely, as most of the events occur in the testing area after the known change in network policies in April 2005.

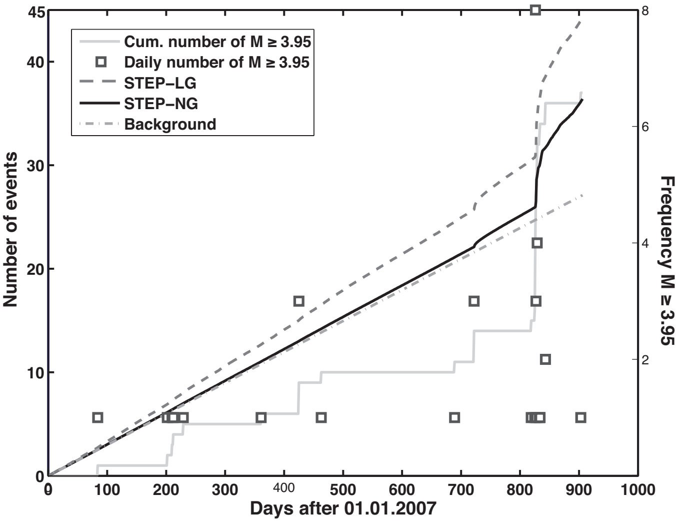

Figure 6. Number of observed and forecasted events versus time from January 1, 2007, to June 25, 2009. Cumulative number of observed events (light gray line), STEP-NG model (black line), STEP-LG model (gray dashed line), and rate of the time-invariant contribution (gray dash-dotted). Squares show the frequency of the observed number of events with $M_L \geq 3.95$ per day (y-axis, right).

There were 751 earthquakes with $M_L \geq 3$ that contributed to forecast rates during the retrospective daily testing period (January 1, 2007, to June 25, 2009). All of the triggering events fell into the magnitude range $3 \leq M_L \leq 6.2$, with the largest earthquake being the April 6, 2009, $M_L = 5.8$ L'Aquila earthquake. The STEP-LG implementation forecasts higher rates of seismicity than STEP-NG, indicated by the steeper slope in Figure 5A.

The forecast rates of STEP-NG match the total observed number of earthquakes better in the cumulative test than for STEP-LG. Figure 7 shows the 906 daily and cumulative results of the N-Test scores $\delta_1(t)$ and $\delta_2(t)$. Daily N-Test scores $\delta_1(t)$ (Figure 7, red triangles) were rejected with at least 95% confidence when at least two events occurred and there was no contribution from ongoing earthquake sequences. No daily forecast was ever rejected for underpredicting. Considering the cumulative tests, we find that STEP-NG was rejected on 30.1% of the test days, and STEP-LG was rejected on 48.6%. The models were rejected

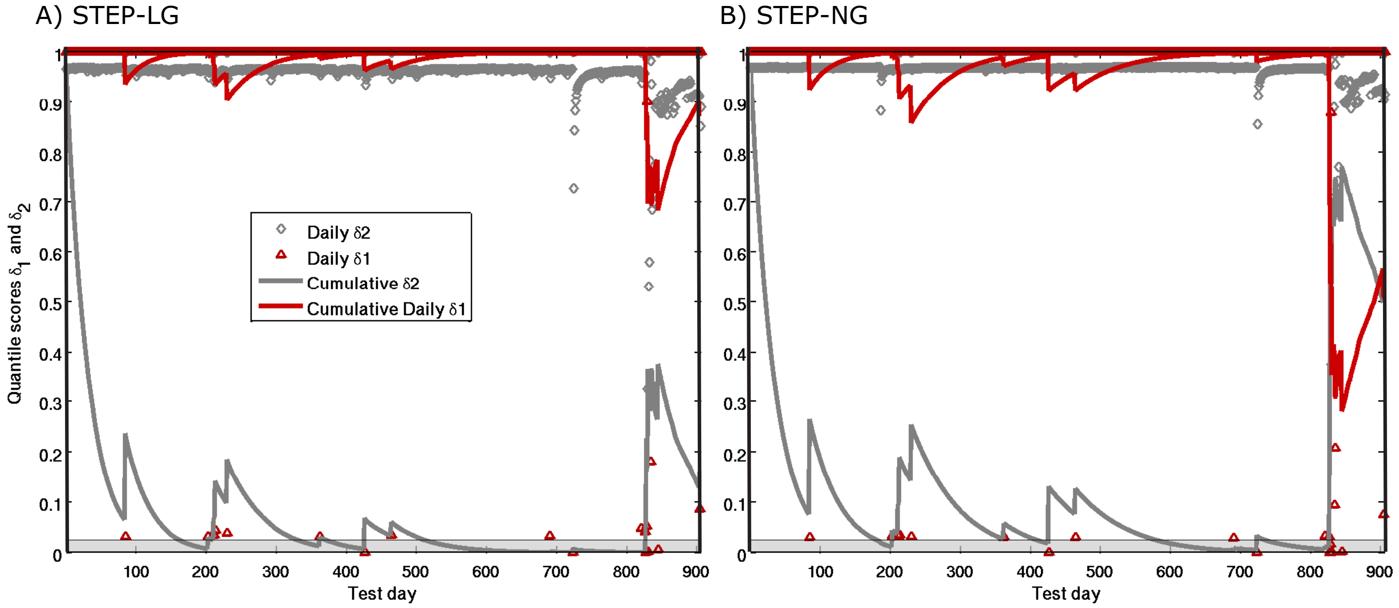

Figure 7. Retrospective N-Test results for daily and cumulative consistency. Daily and cumulative $\delta_1(t)$-scores (red) of observing at least N_{Obs} events were generally not rejected at the α_{eff} significance level (gray bar). Note that cumulative $\delta_2(t)$-scores (gray line) for observing at most N_{Obs} events were rejected in about 48.6% of the test days for model STEP-LG, and in about 30.1% for model STEP-NG.

due to the constant rate increase caused by the time-invariant contribution of the model. The N-Test result was strongly influenced by each event occurring. As soon as a large sequence occurs, such as the L'Aquila sequence (Figure 6) that started on test day 826, the forecasts performed well overall in the N-Test.

To test the overall consistency of each implementation with the daily observations, we applied the L-Test [Schorlemmer et al. 2007] daily and cumulatively. Both models were never rejected in the cumulative tests (Figure 8, red line) and only on a few days in the day-by-day tests (Figure 8, black triangles). For example, the log-likelihood score LL to observe $N_{Obs} = 0$ events for the first days forecast of model STEP-NG $N_F = 0.0302$ is $LL = -0.031$. Using $N_{Sim} = 10000$ simulations for the L-Test, we obtained a score on the first test day of $\gamma\ (t = 1) = 0.036$. The quantile score improves over time as there are more and more events occurring in the regions where events are expected.

The influence of the forecast number of events on the original CSEP L-Test is illustrated well for these models, as the spatial distribution of rates was exactly the same (Figure 8, red boxes). On test day $t = 831$, for example, STEP-LG forecast 0.0431 target earthquakes, and STEP-NG forecast 0.0326 target earthquakes; $N_{Obs} = 0$ events were observed. This led to a small difference in the γ-scores, of 0.334 and 0.313, respectively.

The normalized L-Test removes the dependency on the forecast number of events with a factor expressing the ratio of the observed and forecasted events: N_{Obs}/N_F [Werner et al. 2009], while keeping the space and magnitude information. The S-Test isolated the spatial distribution by summing the entire rates. For both of these tests, there is no difference between our two implementations, as they were using the same *b*-value for distributing in the frequency magnitude domain. We only show the results of STEP-NG in Figure 9, as through the normalization the differences of the two implementations were removed.

The tests were developed and applied retrospectively for long-term forecasts (5 years), for which the number of observed events is typically non-zero. However, in daily experiments, the number of observed events is often equal to zero, and thus the normalization was not applicable. We therefore applied the S-Tests and normalized L-Tests only on the forecast days on which at least one earthquake occurred. The daily test results displayed as the quantile scores (Figure 9, gray triangles) show that the models were consistent in the space-magnitude domain as the γ_{norm}-scores are larger than 0.025 (Figure 9a), the effective significance level for the test. Similarly, the ζ-scores of the S-Test were larger than 0.025 for every day of the experiment, implying that the model is consistent with the spatial distribution of the seismicity observed (Figure 9a).

Discussion and conclusion

We have present two implementations of the STEP-model, STEP-LG and STEP-NG, which can generate daily forecasts of seismicity calibrated on Italian seismicity. Both implementations have the same spatial distribution of the seismicity rates, but they vary in terms of the total forecast rate. For the generic element of the time-varying contribution, STEP-LG incorporates rate estimates based on the Reasenberg-Jones model [Reasenberg and Jones 1989] with parameters from Gasperini and Lolli [2006], while STEP-NG defines rates based on the mean abundance model [Christophersen and Gerstenberger 2010] (Figure 1).

We emphasize that the two implementations described

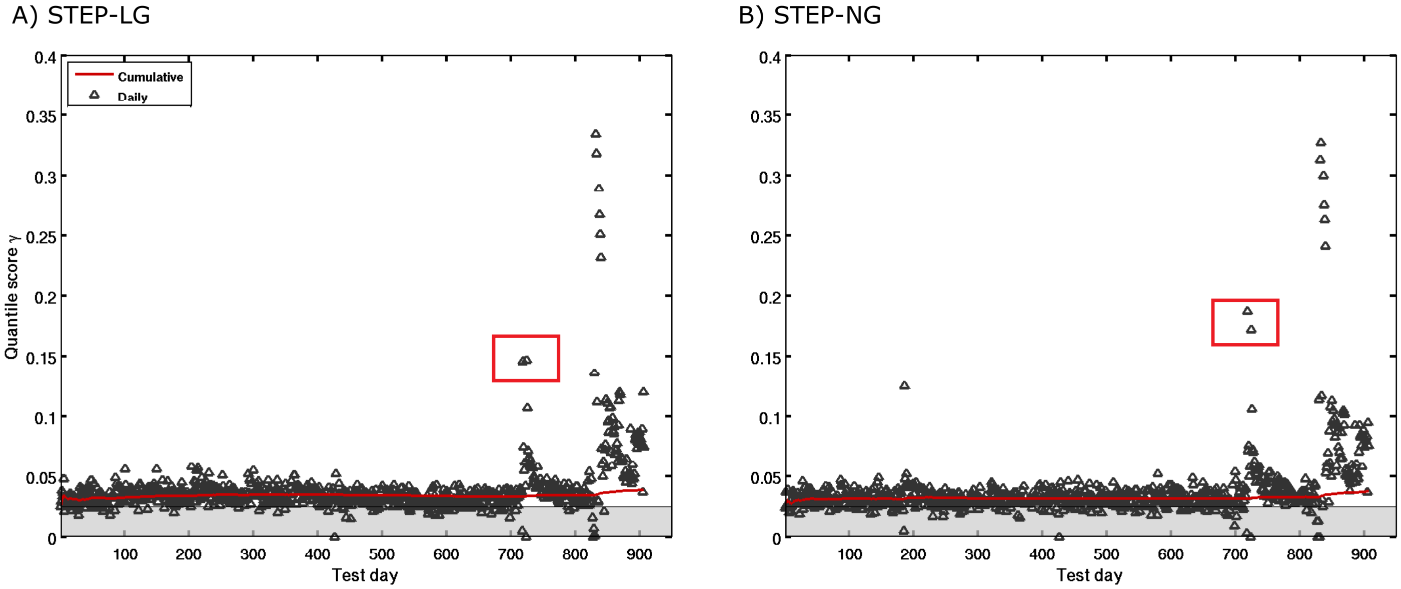

Figure 8. Retrospective L-Test results for daily (triangles) and cumulative (red lines) tests for (A) STEP-LG and(B) STEP-NG. The models were not rejected at the 0.05 significance level, indicated by the gray bar (effective value of 0.025). The spatial consistency improved with time, as indicated by increasing cumulative γ-scores. Slight differences in the daily tests were observed, due to the difference in the number of forecasted events; examples are seen highlighted in red rectangles.

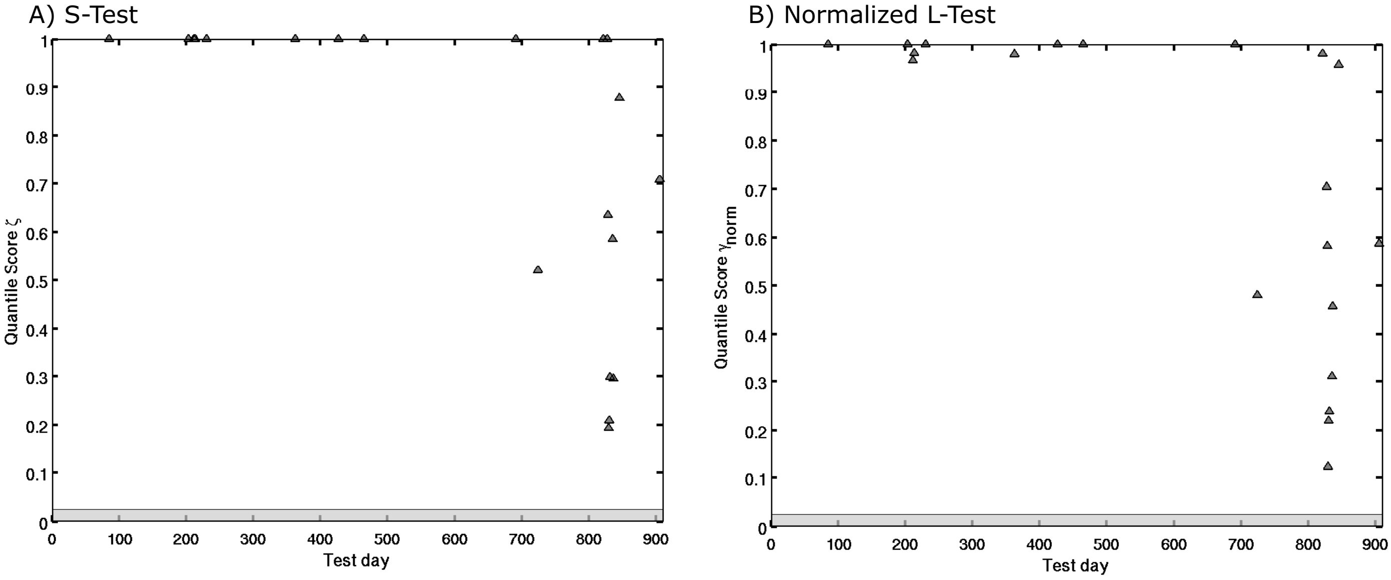

Figure 9. Retrospective (A) S-Test and (B) normalized L-Test results of STEP-NG for daily forecasts. The models were never rejected at the 0.05 significance level (for quantile scores ζ and γ_{norm} equal or less than the effective significance level 0.025, indicated by the gray bar), implying spatial consistency throughout the testing period. The test results for STEP-LG were the same as for STEP-NG.

here are different from the implementation of the STEP model submitted to the CSEP Testing Center at the Southern California Earthquake Center and the model that is implemented at US Geology Survey (http://earthquake.usgs.gov/eqcenter/step). Our guiding principle was to generate a self-consistent model for both contributions: time-invariant and time-varying. We used a time-invariant contribution based on the smoothed seismicity approach following Zechar and Jordan [2010b], using only instrumental seismicity. This is a change in philosophy from the approach of Gerstenberger et al. [2005]: we assumed that the seismicity of the last 25 years was a well defined proxy for background seismicity varying on the scales of tenths of years, better than a long-term hazard map with all its assumptions [Meletti et al. 2008]. We have also explored a new generic model element based on the mean abundance model (STEP-NG). This implementation forecast fewer events following small to moderate earthquakes ($M_L \leq 6.2$) than the STEP-LG implementation, and it forecast more following larger events.

Although the stationary rate estimates do influence the performance of our implementations (see Table 1), a detailed comparison of different smoothed seismicity approaches is beyond the scope of this study. We used the window approach to cluster events and to estimate the time-invariant contribution based on this result, because this is consistent with the approach used to compute the time-varying contribution; nevertheless we recognize the intrinsic uncertainty in such an estimate.

The issue of deciding the best way to estimate the stationary component for a time-varying daily forecast model is probably best approached by prospectively testing long-term time-invariant forecasts, as was followed in the 5-year CSEP experiments. We have chosen an approach that is simple, although probably imperfect. The novelty of this method with respect to a past implementation [Frankel et al. 2002] was optimization of the kernel width by performing a set of retrospective tests with different smoothing length scales. The kernel width that gave the best performance in terms of the ASS misfit statistic [Zechar and Jordan 2008] was chosen as the optimal one (Table 2). The choice of the kernel width was objective; the choice of the data was subjective.

For both of the STEP-LG and STEP-NG implementations, the triggering main shock magnitude was set to $M_L = 3.0$, as the probability of a foreshock to trigger an earthquake of one magnitude larger increases by up to 4%, and might be smaller by a factor of one hundred, depending on which model is used [Michael and Field 2009]. The parameters for both implementations were based on the analysis of earthquake sequences with main shocks in the range $4.0 \leq M_L \leq 7.0$ (Figure 5A, B). Thus, the extrapolation for $k(M_m) \leq 4.0$ contained uncertainty that was not well constrained. The parameter value estimates for each implementation also contained uncertainties due to the methods used to determine the completeness and declustering of the original catalog. Gasperini and Lolli [2006] applied the Reasenberg [1985] approach, while we used the STEP clustering approach. A sensitivity study of the influence of these uncertainties on the forecast results, as well as a thorough uncertainty study on the parameter value estimation, has yet to be performed.

We chose the parameter value estimates based on the period from 1981 to 2002, considering only sequences above $M_L \geq 4$ (Figure 5) because the results of the clustering process of the sequences in the magnitude range $3 \leq M_L \leq 4$ might be biased; there are probably many sequences that should be associated with the background seismicity, thus increasing the productivity parameter k, although it should be smaller.

From the magnitude dependent k-value (Figure 5B), we expected the STEP-LG model to provide in general higher seismicity rates than the STEP-NG model. Small magnitude contributions are more frequent and thus have a stronger influence for long periods on the daily forecast than larger magnitude events. In the case where a large main shock occurs, its influence on the large event prevails for some time, as can be seen in the retrospective testing results (Figure 6).

We retrospectively tested the implementations for a period of 906 test days, starting on January 1, 2007, in accordance with the procedures that are used in the one-day-testing class. We applied multiple testing procedures, as used in the CSEP test centers, and also test statistics that can be included in the testing center software [Werner et al. 2009, Zechar et al. 2010]. The modified N-Test results showed that overall, the rate forecasts of STEP-NG were superior to those of STEP-LG (Figures 6, 7). Both of the models were spatially consistent with the observed data; forecasts were not rejected in the S-Tests (Figure 9). The models were also consistent with the data when considering the space-magnitude distributions for the normalized and non-normalized L-Test. (Figure 8).

The 2009 L'Aquila sequence illustrated the influence of a moderate earthquake on the performance of the models. The largest shock occurred on test day 827, which was clearly seen in Figures 6 and 7, with the burst of $M_L \geq 3.95$ events. Both of the models showed increased rates due to the foreshock activity that started about 7 days before (test day 820) with a sequence of $M_L \geq 3.95$ events, but both underestimated the occurrence of successive earthquakes for this sequence (Figure 6). The low forecast rates were due to the definition of the retrospective testing class, restricting the models to update only every 24 hours.

The L'Aquila event occurred on April 6, 2009, at 1.32 a.m. (GMT). The forecast based on the seismicity to test day 826 did not include this information, only the information of the foreshocks. The forecast for day 827 did include rates due to the largest shock, but it only forecast the rates from midnight on April 6, 2009, to midnight on April 7, 2009. The models in essence missed the chance to forecast seismicity in the period of April 6, 2009, 1.32 a.m. (GMT) to 0.00 a.m. on April 7, 2009, which is actually the period where most of the events should occur according to the Omori-Utsu law. For testing forecasts, it is thus important at what time an event occurs. In other words, the forecasts are sensitive to when the forecast period starts [Helmstetter et al. 2006]. For the existing implementations, a best case scenario is a strong event occurring shortly before midnight of any day; a worst case scenario is the occurrence of a strong event shortly after midnight. Updating forecasts more frequently or after each earthquake that is considered to trigger further events would reveal the actual capabilities of short-term forecast models. However, testing these forecasts becomes more complicated as the duration of the forecasts and the updating periods would become variable.

Both implementations provided short-term rate forecasts that can serve as the basis for time-varying hazard information. To provide up-to date information on the probability of exceeding a specific ground motion measure, a full hazard computation is required [Gerstenberger et al. 2005]. We suggest that for best performance this type of short-term forecast should be updated more frequently, and ideally immediately after an earthquake has been detected by the monitoring system. If communicated to decision-makers, the media, and the public in the appropriate language, information on seismicity rates and seismic hazards can promote a better understanding of the time-varying earthquake hazard before and during a strong earthquake sequence.

Acknowledgements. We thank M. Werner for discussing all aspects of this study, and especially the retrospective testing. We thank M.C. Gerstenberger for providing the initial STEP code, F. Euchner for critical code review, and T. van Stiphout and S. Wiemer for discussions, encouragement and reviews on an early version of the manuscript. We acknowledge the comments and suggestions from two reviewers and the editor, W. Marzocchi, that helped us to significantly improve our manuscript. This work is partly funded by the EU-FP6 projects SAFER (contract number 036935) and NERIES (contract number 26130), as well as the EU-FP7 project SHARE (contract number 226967).

References

Bender, B. (1983). Maximum likelihood estimation of *b*-values for magnitude grouped data, Bull. Seismol. Soc. Am., 73, 831-851.

Castello, B., M. Olivieri and B. Selvaggi (2007). Local and duration magnitude determination for the Italian earthquake catalog, 1981-2002, Bull. Seismol. Soc. Am., 97, 128-139; doi: 10.1785/0120050258.

Christophersen, A. and E.G. Smith (2008). Foreshock rates from aftershock abundance, Bull. Seismol. Soc. Am., 98, 2133-2148; doi: 10.1785/0120060143.

Christophersen, A. and M.C. Gerstenberger (2010). A new generic model for aftershock decay in earthquake forecasting, Seis. Res. Lett., 81, 316.

Cocco, M., F. Catalli, S. Hainzl, B. Enescu and J. Woessner (2010). Sensitivity study of forecasts based on Coulomb stress calculation and rate-state frictional response, J. Geophy. Res., 115, B05307; doi: 10.1029/2009JB006838.

Console, R., M. Murru and A.M. Lombardi (2003). Refining earthquake clustering models, J. Geophy. Res., 108 (B10); doi: 10.1029/2002JB002130.

Frankel, A.D., M.D. Petersen, C.S. Mueller, K.M. Haller, R.L. Wheeler, E.V. Leyendecker, R.L. Wesson, S.C. Harmsen, C.H. Cramer, D.M. Perkins and K.S. Rukstales (2002). Documentation for the 2002 Update of the National Seismic Hazard Maps, Open-file Report 02-420., US Geological Survey.

Gardner, J.K. and L. Knopoff (1974). Is the sequence of earthquakes in southern California, with aftershocks removed, Poissonian?, Bull. Seismol. Soc. Am., 64, 1363-1367.

Gasperini, P. and B. Lolli (2006). Correlation between parameters of the aftershock rate equation: implications for the forecasting of future sequences, Phys. Earth Planet. Inter., 156, 41-58.

Gerstenberger, M.C., S. Wiemer and L. Jones (2004). Real-time forecast of tomorrow's earthquakes in California: a new mapping tool, Tech. Rep. Open-File Report 2004-1390, U.S. Geological Survey.

Gerstenberger, M.C., S. Wiemer, L.M. Jones and P.A. Reasenberg (2005). Real-time forecasts of tomorrow's earthquakes in California, Nature, 435 (7040), 328-331; doi: 10.1038/nature03622.

Gutenberg, B. and C.F. Richter (1944). Frequency of earthquakes in California, Bull. Seismol. Soc. Am., 34, 185-188.

Hainzl, S., B. Enescu, F. Catalli, M. Cocco, R. Wang, F. Roth and J. Woessner (2009). Aftershock modeling based on uncertain stress calculations, J. Geophy. Res., 114, B05309; doi: 10.1029/2008JB006011.

Helmstetter, A., Y.Y. Kagan and D.D. Jackson (2006). Comparison of short-term and time-independent earthquake forecast models for southern California, Bull. Seismol. Soc. Am., 96 (1); doi: 10.1785/0120050067.

Herak, D., M. Herak, E. Prelogovic, S. Markusic and Z. Markulin (2005). Jabuka island (central Adriatic Sea) earthquakes of 2003, Tectonophysics, 398 (3-4), 167-180; doi: 10.1016/j.tecto.2005.01.007.

Kenneth, P., K. Burnham and D.R. Anderson (2002). Model Selection and Multimodel Inference: A Practical Information -Theoretical Approach, 2nd ed., New York, 496 pp.

Lolli, B. and P. Gasperini (2003). Aftershock hazard in Italy Part I: Estimation of time-magnitude distribution model parameters and computation of probabilities of occurrence, J. of Seismology, 7, 235-257.

Marsan, D. and O. Lengliné (2008). Extending Earthquakes' Reach Through Cascading, Science, 319 (5866), 1076-1079; doi: 10.1126/science.1148783.

Marzocchi, W. and A.M. Lombardi (2008). A double branching mode for earthquake occurrence, J. Geophy. Res., 113, B08317; doi: 10.1029/2007JB005472.

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 (1-4), 85-108; doi: 10.1016/j.tecto.2008.01.003.

Michael, A.J. and E.H. Field (2009). Short-term earthquake probabilities based on long-term probability mode, Seism. Res. Let., 80 (2).

Ogata, Y. (1983). Estimation of the parameters in the modified Omori formula for aftershock frequencies by the maximum likelihood procedure, J. Phys. Earth, 31, 115-124.

Reasenberg, P.A. (1985). Second-order moment of central California seismicity,1969-1982, J. Geophy. Res., 90 (B7), 5479-5495.

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

Reasenberg, P.A. and L.M. Jones (1990). California aftershock hazard forecast, Science, 247, 345-346.

Reasenberg, P.A., and L.M. Jones (1994). Earthquake aftershocks: update, Science, 265, 1251-1252.

Schorlemmer, D., M.C. Gerstenberger, S. Wiemer, D. Jackson and D.A. Rhoades (2007). Earthquake likelihood model testing, Seism. Res. Let., 87, 17-29.

Schorlemmer , D., A. Christophersen, A. Rovida, F. Mele, M.

Stucchi and W Marzocchi (2010). Setting up an earthquake forecast experiment in Italy, Annals of Geophysics, 53, 3 (present issue).

Stirling, M.W., G. McVerry and K. Berryman (2002). A new seismic hazard model for New Zealand, Bull. Seismol. Soc. Am., 92 (5), 1878-1903.

Stock, C. and E.G.C. Smith (2002a). Adaptive kernel estimation and continuous probability representation of historical earthquake catalogs, Bull. Seismol. Soc.Am., 92 (3), 904-912.

Stock, C. and E.G.C. Smith (2002b). Comparison of seismicity models generated by different kernel estimations, Bull. Seismol. Soc. Am., 92 (3), 913-922.

Utsu, T. (1961). A statistical study of the occurrence of aftershocks, Geophys. Mag., 3, 521-605.

Wells, D.L., and K.J. Coppersmith (1994). New empirical relationships among magnitude, rupture length, rupture width, rupture area, and surface displacement, Bull. Seismol. Soc. Am., 84 (4), 974-1002.

Werner, M.J., A. Helmstetter, D. Jackson and Y. Kagan (2009). Long-term earthquake forecasts for California and Italy, Geophys. Res. Abstracts, 12 (EGU2009-12045).

Woessner, J. and S. Wiemer (2005). Assessing the quality of earthquake catalogs: Estimating the magnitude of completeness and its uncertainties, Bull. Seismol. Soc. Am., 95 (2); doi: 10.1785/0120040007.

Woessner, J., A.M. Lombardi, M.J. Werner and W. Marzocchi (2009). Testing the predictive power of Coulomb stress on aftershock sequences, Eos Trans. AGU, 90 (52), abstract S22C-08.

Zechar, J.D. and T.H. Jordan (2008). Testing alarm-based earthquake predictions, Geophys. J. Int., 172 (2), 715-724; doi: 10.1111/j.1365-246X.2007.03676.x.

Zechar, J.D. and T.H. Jordan (2010a). The area skill score statistic for evaluating earthquake predictability experiments, Pure and Appl. Geophys., 167 (8/9), 893-906; doi: 10.1007/s00024-010-0086-0

Zechar, J.D. and T.H. Jordan (2010b). Simple smoothed seismicity earthquake forecasts for Italy, 53, 3 (present issue).

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

*Corresponding author: Jochen Woessner,
ETH Zurich, Swiss Seismological Service, Zurich, Switzerland;
email: jochen.woessner@sed.ethz.ch