Retrospective evaluation of the five-year and ten-year CSEP-Italy earthquake forecasts

Maximilian J. Werner[1,*], J. Douglas Zechar[1,2], Warner Marzocchi[3], Stefan Wiemer[1] and the CSEP-Italy Working Group[4]

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

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

[3] *Istituto Nazionale di Geofisica e Vulcanologia, sezione di Roma, Italy*

[4] *See Acknowledgements section*

Article history
Received March 3, 2010; accepted August 19, 2010.
Subject classification:
Probabilistic forecasting, Earthquake predictability, Hypothesis testing, Likelihood.

ABSTRACT

On August 1, 2009, the global Collaboratory for the Study of Earthquake Predictability (CSEP) launched a prospective and comparative earthquake predictability experiment in Italy. The goal of this CSEP-Italy experiment is to test earthquake occurrence hypotheses that have been formalized as probabilistic earthquake forecasts over temporal scales that range from days to years. In the first round of forecast submissions, members of the CSEP-Italy Working Group presented 18 five-year and ten-year earthquake forecasts to the European CSEP Testing Center at ETH Zurich. We have considered here the twelve time-independent earthquake forecasts among this set, and evaluated them with respect to past seismicity data from two Italian earthquake catalogs. We present the results of the tests that measure the consistencies of the forecasts according to past observations. As well as being an evaluation of the time-independent forecasts submitted, this exercise provides insight into a number of important issues in predictability experiments with regard to the specification of the forecasts, the performance of the tests, and the trade-off between robustness of results and experiment duration. We conclude with suggestions for the design of future earthquake predictability experiments.

1. Introduction

On August 1, 2009, a prospective and competitive earthquake predictability experiment began for the region of Italy [Schorlemmer et al. 2010a]. The experiment followed the design proposed by the Regional Earthquake Likelihood Model (RELM) working group in California, USA [Field 2007, Schorlemmer et al. 2007, Schorlemmer and Gerstenberger 2007, Schorlemmer et al. 2010b], and it falls under the global umbrella of the Collaboratory for the Study of Earthquake Predictability (CSEP) [Jordan 2006, Zechar et al. 2010b]. Eighteen five-year forecasts that express a variety of scientific hypotheses about earthquake occurrence were submitted to the European CSEP Testing Center at ETH Zurich. In this study, we present the results from the retrospective testing of these forecasts on seismicity data from two Italian earthquake catalogs.

The rationale for performing these retrospective tests is as follows:

1. to verify that the submitted forecasts are as intended by the modelers;
2. to provide a «sanity check» of the forecasts before the end of the five-year and ten-year experiments;
3. to provide feedback to each of the modelers about the performance of their model in retrospective tests, and to encourage model improvements;
4. to better understand the tests and performance metrics;
5. to have explicit, pedagogical examples of plausible observations and results; and
6. to understand the relationships between the duration of predictability experiments and the robustness of the outcomes.

Nevertheless, retrospective tests also come with significant caveats:

(i) We have only evaluated the time-independent models. To fairly test the time-dependent models on past data would require that the model software is installed at the testing center, so that hindcasts can be generated. We identified the long-term forecasts from time-dependent models, as described in Section 2, and we did not analyze these forecasts.

(ii) Past data might be of lower quality than the data used for prospective testing (e.g., greater uncertainties in magnitudes and locations, or missing aftershocks, which can potentially show systematic bias).

(iii) There are different versions of the past data, in the form of the several earthquake catalogs that are available. In an attempt to address this issue, we tested the data with respect to two catalogs (see Section 4).

Model	Forecast number of earthquakes	Reference
Time-independent		
AKINCI–ET–AL.HAZGRIDX	11.46	Akinci [2010]
CHAN–ET–AL.HZATI	14.76	Chan et al. [2010]
GULIA–WIEMER.ALM	8.58	Gulia et al. [2010]
GULIA–WIEMER.HALM	9.53	Gulia et al. [2010]
MELETTI–ET–AL.MPS04	15.60	MPS Working Group [2004]
NANJO–ET–AL.RI	2.78	Nanjo [2010]
SCHORLEMMER–WIEMER.ALM	12.74	Gulia et al. [2010]
WERNER–ET–AL.CSI	6.21	Werner et al. [2010b]
WERNER–ET–AL. CPTI	6.52	Werner et al. [2010b]
ZECHAR–JORDAN.CPTI	14.38	Zechar and Jordan [2010b]
ZECHAR–JORDAN.CSI	5.88	Zechar and Jordan [2010b]
ZECHAR–JORDAN.HYBRID	13.23	Zechar and Jordan [2010b]
Time dependent		
CHAN–ET–AL.HZATD	14.87	Chan et al. [2010]
CONSOLE–ET–AL.LTST	10.98	Falcone et al. [2010]
FAENZA–ET–AL.PHMGRID	6.64	Faenza and Marzocchi [2010]
FAENZA–ET–AL.PHMZONE	6.30	Faenza and Marzocchi [2010]
LOMBARDI–MARZOCCHI.DBM	9.06	Lombardi and Marzocchi [2010a]
PERUZZA–ET–AL.LASSCI*	1.90	Pace et al. [2010]

*Only covered 7.09% of the study region forecast area; all others covered 100%.

Table 1. Five-year and ten-year CSEP-Italy forecasts that are being evaluated within the European CSEP Testing Center at ETH Zurich. The forecasts were submitted before August 1, 2009.

(iv) All of the forecasts considered here are in some way based on past observations; e.g., parameters of the models were typically optimized on part of or all of the data against which we tested the models retrospectively. Therefore, positive retrospective test results can simply reveal that a model adequately fits the data on which it was calibrated, and might not be indicative of future performance on independent data.

A study beyond the scope of this report would be required to decide which of the retrospective data can be regarded as out-of-sample for each model. On the other hand, poor performance of a time-independent forecast in these retrospective experiments indicates that the forecast cannot adequately explain the available data. Therefore, one aim of this study was to identify forecasts of time-independent models that consistently fail in retrospective tests, thereby separating ineffective time-independent models from potentially good ones.

Poor performance of a time-independent forecast might result from one or more of several factors, such as: technical errors (i.e., errors in software implementation); misunderstanding of the required object to be forecast; calibration with low-quality data; evaluation with low-quality data; statistical type II errors; or incorrect hypotheses of earthquake occurrence. The CSEP modelers sought to minimize the chances of each of these effects, except for the final one: that a forecast is rejected because its underlying hypotheses about earthquake occurrence are incorrect.

This study is accompanied by an electronic supplement (available online at http://www.annalsofgeophysics.eu/index.php/annals/rt/suppFiles/4840/0); the reader can find additional figures and a table of information gains that aid in the evaluation of the forecasts considered.

2. Overview of the time-independent models

Each of the forecasts submitted to the five-year and ten-year CSEP-Italy experiments can be broadly grouped into one of two classes: those derived from time-independent models; and those derived from time-dependent models (see Table 1). The forecasts in the former class are considered to be suitable for any time translation, and they depend only on the length of the forecasting time interval (at least over a reasonable time interval, where the models are assumed to be time-independent). Therefore, these forecasts can be tested on different target periods. In contrast, the forecasts derived from time-dependent models depend on the initial time of the forecast. Because the methodologies for calculating the forecasts (i.e. the model software) were not available to us, we were not able to generate hindcasts from these models that could be meaningfully evaluated. We therefore did not consider time-dependent models in this study. Below, we provide a brief summary of each time-independent model.

The AKINCI–ET–AL.HAZGRIDX model contains the assumption that future earthquakes will occur close in space to the locations of historical $M_w \geq 4$ mainshocks. No tectonic, geological or geodetic information was used to calculate the forecast. The model is based on the method of Weichert [1980] to estimate the seismic rate from declustered earthquake catalogs where the magnitude completeness threshold varies with time. The forecast uses a Gutenberg-Richter law with a uniform *b*-value.

CHAN–ET–AL.HZATI considers a specific bandwidth function to smooth the past seismicity and to evaluate the spatial seismicity density of the earthquakes. The model smoothes both spatial locations and magnitudes. The smoothing procedure is applied to a coarse seismotectonic zonation that is based on a large-scale geological structure. The expected rate of earthquakes is obtained from the average historical seismicity rate.

Each asperity likelihood model (ALM) – GULIA–WIEMER.ALM, GULIA–WIEMER.HALM, and SCHORLEMMER–WIEMER.ALM – hypothesizes that small-scale spatial variations in the *b*-value of the Gutenberg-Richter relationship have a central role in forecasting the future seismicity [Wiemer and Schorlemmer 2007]. The physical basis of these models is the concept that the local *b*-value is inversely proportional to the applied shear stress. Thus low *b*-values ($b < 0.7$) are believed to characterize the locked patches of faults (asperities) from which future mainshocks are more likely to be generated, whereas high *b*-values ($b > 1.1$), e.g., seen in creeping sections of faults, suggest a lower probability of large events. The *b*-value variability is mapped on a grid. The local *a* and *b*-values in the GULIA–WIEMER.ALM and GULIA–WIEMER.HALM forecasts were obtained from the observed rates of declustered earthquakes from 1981 to 2009, using the Reasenberg declustering method [Reasenberg 1985] and the entire-magnitude-range method for completeness estimation of Woessner and Wiemer [2005] [see also Schorlemmer and Woessner 2008]. In the GULIA–WIEMER.HALM model (Hybrid ALM), which is a "hybrid" between a grid-based and a zoning model, the Italian territory was divided into distinct regions that depended on their main tectonic regime and the local *b*-value variability, and was thus mapped using independent *b*-values for each tectonic zone. In the SCHORLEMMER–WIEMER.ALM model, which is derived from the original ALM [Wiemer and Schorlemmer 2007], the input catalog (2005-2009) for $M_w \geq 2$ was declustered using the method of Gardner and Knopoff [1974], and the node-wise rates of the declustered catalog were smoothed with a Gaussian filter. Completeness values for each node were taken from the analysis of Schorlemmer et al. [2010c] using the probability-based magnitude of completeness method [Schorlemmer and Woessner 2008]. The resulting forecast was calibrated according to the observed average number of events with $M_w \geq 4.95$.

The MELETTI–ET–AL.MPS04 model [MPS Working Group 2004, http://zonesismiche.mi.ingv.it] is the reference model for seismic hazard in Italy. MELETTI–ET–AL.MPS04 is derived from the standard approach to probabilistic seismic hazard assessment of Cornell [1968], in which a Poisson process is assumed. The model distributed the seismicity into a seismotectonic zonation, and through a logic tree structure it considered the historical catalog using two different methods (historical and statistical) to estimate its completeness. The models also assumed that each zone was characterized by its own Gutenberg-Richter law, with varying truncation.

The relative intensity model of NANJO–ET–AL.RI is a pattern recognition model that was based on the main assumption that future large earthquakes tend to occur where the seismic activity has had a specific pattern in the past (usually a higher seismicity). In its first version, the relative intensity code was «alarm-based»; i.e., the code made a binary statement about the occurrence of earthquakes. For the CSEP-Italy experiment, the code was modified to estimate the expected number of earthquakes in a specific time-space-magnitude bin.

The models of WERNER–ET–AL.CSI and WERNER–ET–AL.CPTI are based on smoothed seismicity. Future earthquakes were assumed to occur with higher probabilities in areas where past earthquakes have occurred. The locations of the past mainshocks were smoothed using an adaptive power-law kernel, i.e. little in regions of dense seismicity, more in sparse regions. The degree of smoothing was optimized via retrospective tests. The magnitude of each earthquake was independently distributed according to a tapered Gutenberg-Richter distribution with corner magnitude 8.0. The model used small magnitude $M_w \geq 2.95$ earthquakes, when the data could be trusted, to better forecast future large events. The two WERNER–ET–AL.CSI and WERNER–ET–AL.CPTI forecasts were obtained by calibrating the model according to two different earthquake catalogs.

The forecasts of ZECHAR–JORDAN.CPTI, ZECHAR–JORDAN.CSI and ZECHAR–JORDAN.HYBRID are derived from the Simple Smoothed Seismicity (Triple-S) model, which is based on Gaussian smoothing of past seismicity. Past epicenters have smoothed contributions to earthquake density estimation, where the epicenters were smoothed using a fixed lengthscale σ; and σ was optimized by minimizing the average area skill score misfit function in a retrospective experiment [Zechar and Jordan 2010a]. The density map was scaled to match the average historical rate of seismicity. The two forecasts of ZECHAR–JORDAN.CPTI and ZECHAR–JORDAN.CSI were optimized according to two different catalogs, while ZECHAR–JORDAN.HYBRID is a hybrid forecast.

3. Specification of CSEP-Italy forecasts

We use the term «seismicity model» to mean a system of hypotheses and inferences that is presented as a mathematical, numerical and simplified description of the process of seismicity. A «seismicity forecast» is a statement about some observable aspect of seismicity that derives from a seismicity model. In the context of the CSEP-Italy experiment, a seismicity forecast is a set of estimates of the expected number of future earthquakes in each bin, where the bins are specified by intervals of location, time and magnitude within the multi-dimensional testing volume [see also Schorlemmer et al. 2007]. More precisely, the CSEP-Italy modelers agreed (within the official "rules of the game" document) to provide a numerical

estimation of the likelihood distribution of observing any number of earthquakes within each bin. Moreover, this discrete distribution, which specifies the probability of observing zero, one, two or more earthquakes in a bin, is given by a Poisson distribution (defined below, in Section 4.3) that is uniquely defined by the expected number of earthquakes. The distribution of each bin is assumed to be independent of the distribution in other bins, and the observed number of earthquakes in a given bin is compared with the forecast for that bin.

4. Data used for retrospective testing

For these prospective tests of the submitted forecasts, the Italian seismic bulletin (Bollettino Sismico Italiano, BSI; http://bollettinosismico.rm.ingv.it/) recorded by INGV was used [see Schorlemmer et al. 2010a]. We did not use the BSI for retrospective evaluations of the forecasts because it is only available in its current form from April 2005. Instead, we used two alternative Italian earthquake catalogs provided by the INGV, which were also provided as tools for the modelers for model learning and calibration: the parametric catalog of Italian earthquakes (Catalogo Parametrico dei Terremoti Italiani, version CPTI08) [Rovida and the CPTI Working Group 2008] and the catalog of Italian seismicity (Catalogo della Sismicità Italiana, CSI 1.1) [Castello et al. 2007, Chiarabba et al. 2005]. Schorlemmer et al. [2010a] discussed these catalogs in detail, so here we only provide a brief overview. Both of these datasets are available for downloading from http://www.cseptesting.org/regions/italy.

4.1. The CSI 1.1 1981-2002

The CSI spans the time period from 1981 to 2002, and it reports local magnitudes (M_L), in agreement with the BSI magnitudes that are used during these prospective evaluation of forecasts. Schorlemmer et al. [2010a] indicated a clear change in earthquake numbers per year in 1984 that was due to the numerous network changes in the early 1980s, and have therefore recommended the use of the CSI data only from July 1, 1984, onwards. For this retrospective evaluation, we selected earthquakes with $M_L \geq 4.95$ from 1985 to the end of 2002. To mimic the durations of the prospective experiments, we selected three non-overlapping five-year periods (1998-2002, 1993-1997, 1988-1992). To test the robustness of the results, we also used the entire 18-year span of reliable data, from 1985 to 2002. We selected shocks as test data if they occurred within the CSEP-Italy testing region [see Schorlemmer et al. 2010a].

4.2. The CPTI08 1901-2006

The CPTI covers the period from 1901 to 2006, and it is based on both instrumental and historical observations [for details, see Schorlemmer et al. 2010a]. This catalog lists moment magnitudes (M_w) that were estimated either from macroseismic data or were calculated using a linear regression relationship between moment magnitude and surface, body wave or local magnitudes. Because the prospective experiment uses M_L, we converted the M_w to M_L using the same regression equation that was used to convert the original M_L to M_w for the creation of the CPTI, as given by Equation (1) [MPS Working Group 2004, Schorlemmer et al. 2010a]:

$$M_L = 1.231\,(M_w - 1.145)\,. \tag{1}$$

Schorlemmer et al. [2010a] estimated a conservative completeness magnitude of $M_L = 4.5$, to justify the use of the entire period from 1901 to 2006 for the retrospective evaluation. However, we have focused mainly on the data since the 1950s, because it appears to be of higher quality [Schorlemmer et al. 2010a]. We divided the period into non-overlapping ten-year periods to mimic the duration of the prospective experiment, although we also evaluated the forecasts on a 57-year time span, from 1950 to 2006, and on the 106-year period, from 1901 to 2006. As for the CSI, we only selected shocks within the testing region. Some of the earthquakes, which were mostly from the early part of the CPTI, were not assigned depths. We included these earthquakes as observations within the testing region because it is not very likely that they were deeper than 30 km [see also Schorlemmer et al. 2010a].

4.3. The distribution of the number of earthquakes

In this section, we consider the distribution of the number of observed events in the five-year and ten-year periods that are relevant to the Italian forecasts. Analysis of the empirical distribution can test the assumption (made by all of the time-independent forecasts) that the Poisson distribution approximates well the observed variation in the number of events in each cell and in the entire testing region (the CSEP-Italy modelers forecast all of the earthquakes, and not only the so-called mainshocks – see Section 4.4).

The Poisson distribution is defined by its discrete probability mass function, according to Equation (2):

$$p(n \mid \lambda) = \lambda^n \frac{\exp(-\lambda)}{n!}\,, \tag{2}$$

where $n = 0, 1, 2, \ldots$, and λ is the rate parameter, the only parameter that is needed to define the distribution. The expected value and the variance σ^2_{POI} of the Poisson distribution are both equal to λ.

Because the span of time over which the CSI is reliable is short, we used the CPTI for the seismicity-rate analysis. The sample variance of the distribution of the number of observed earthquakes in the CPTI over non-overlapping five-year periods from 1907 to 2006 (inclusive) was $\sigma^2_{5yr} = 23.73$, and the sample mean was $\mu_{5yr} = 8.55$. For non-overlapping ten-year periods of the CPTI, the sample variance was $\sigma^2_{10yr} = 64.54$, and the sample mean was $\mu_{10yr} = 17.10$. Because

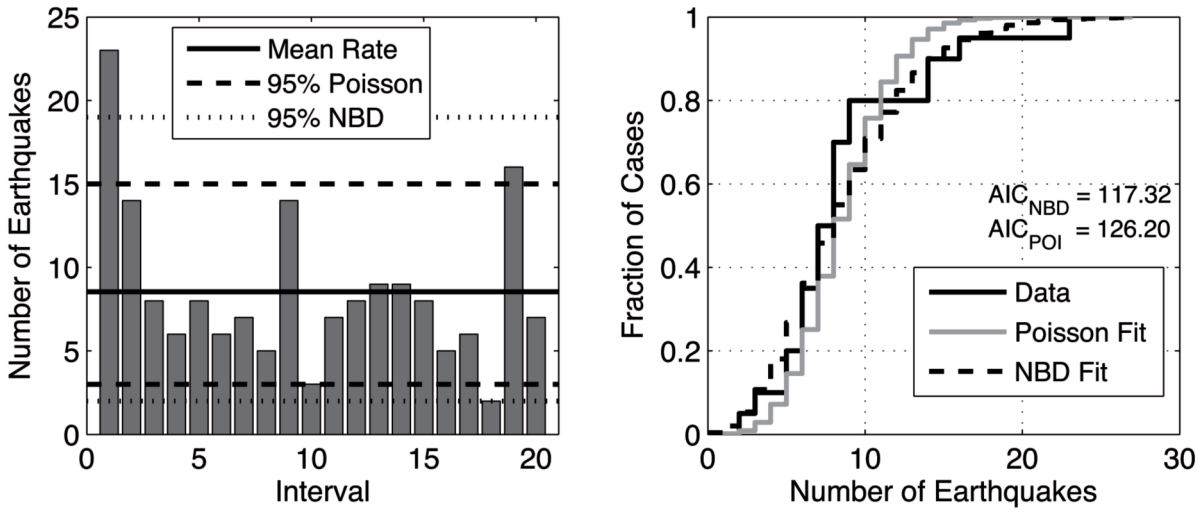

Figure 1. Left: Observed number of earthquakes in the 20 non-overlapping five-year intervals in the CPTI, from 1907 to 2006 (inclusive). Solid line, mean number of observed events; dashed lines enclose the 95% confidence interval of the Poisson distribution; dotted lines, the 95% confidence interval of the negative binomial distribution (NBD). Right: Cumulative distribution functions. Solid black line, observations; solid gray line, data fit with a Poisson distribution; dashed black line, data fit with a NBD. The Akaike Information Criterion (AIC) values of the fitted distributions are also shown.

the sample variance was so much larger than the sample mean on the five-year and ten-year timescales, it is clear that the seismicity rate varied more widely than expected by a Poisson distribution.

Figure 1 shows the number of observed earthquakes in each of the 20 non-overlapping five-year intervals, along with the empirical cumulative distribution function. The Poisson distribution with $\lambda = \mu_{5yr} = 8.55$ and its 95% confidence limits are also shown. One in 20 of the data points would be expected to fall outside the 95% confidence interval; however, four are seen to, one of which lies outside the 99.99% quantile.

We compared the goodness of fit of the Poisson distribution with that of a negative-binomial distribution (NBD), due to studies that have suggested the use of a NBD based on empirical and theoretical considerations [Vere-Jones 1970, Kagan 1973, Jackson and Kagan 1999, Kagan 2010, Schorlemmer et al. 2010b, Werner et al. 2010a]. The discrete negative-binomial probability mass function is described as in Equation (3):

$$p(n \mid \tau, \nu) = \frac{\Gamma(\tau + n)}{\Gamma(\tau) n!} \nu^{\tau} (1 - \nu)^{n} , \qquad (3)$$

where $n = 0, 1, 2, \ldots$, Γ is the gamma function, $0 \leq \nu \leq 1$, and $\tau > 0$. The average of the NBD is given by Equation (4):

$$\mu_{\mathrm{NBD}} = \tau \frac{1 - \nu}{\nu} , \qquad (4)$$

while the variance is given by Equation (5):

$$\sigma^2_{\mathrm{NBD}} = \tau \frac{1 - \nu}{\nu^2} , \qquad (5)$$

implying that $\sigma^2_{\mathrm{NBD}} \geq \sigma^2_{\mathrm{POI}}$.

Kagan [2010] discussed different parameterizations of the NBD. For simplicity, we used the parametrization in Equation (3) and maximum likelihood parameter value estimation. We found $\tau_{5yr} = 6.49$ and $\nu_{5yr} = 0.43$, with the 95% confidence limits given by (–0.39, 13.37) and (0.17, 0.70), respectively. These large uncertainties reflect the small sample size of 20. For the ten-year intervals, we estimated $\tau_{10yr} = 9.24$ and $\nu_{10yr} = 0.35$, with 95% confidence limits given by (–2.74, 21.22) and (0.05, 0.65), respectively. Figure 1 shows the 95% confidence limits of the fitted NBD in the number of observed events (left panel), and the NBD cumulative distribution function (right panel).

Because the NBD is characterized by two parameters, while the Poisson distribution has only one, we used the Akaike information criterion (AIC) of Equation (6) [Akaike 1974] to compare the fits:

$$\mathrm{AIC} = 2p - 2\log(L) , \qquad (6)$$

where L is the likelihood of the data given the fitted distribution, and p is the number of free parameters.

For the five-year and ten-year intervals, the NBD provided a better fit of the data than the Poisson distribution, despite the penalty of the extra parameter: for the five-year intervals, $\mathrm{AIC_{NBD}} = 117.32$ and $\mathrm{AIC_{POI}} = 126.20$; and for the ten-year intervals, $\mathrm{AIC_{NBD}} = 70.05$ and $\mathrm{AIC_{POI}} = 77.56$. To test the robustness of this better fit of the NBD over the Poisson distribution, we also checked the distribution of the number of events in one-year, two-year and three-year intervals for both of the catalogs. In all cases, the NBD provided a better fit than the Poisson distribution, despite the penalty of the extra parameter.

4.4. Implications for the CSEP-Italy experiment

Several previous studies have shown that the distribution of the number of earthquakes in any finite time period is not well approximated by a Poisson distribution, and that a better fit is provided by a NBD [Kagan 1973, Jackson and Kagan 1999, Schorlemmer et al. 2010b, Werner et al. 2010a] or a

heavy-tailed distribution [Saichev and Sornette 2006]. The implications for the CSEP-Italy experiment, and indeed for all of the CSEP experiments to date, are important.

The only time-independent point process is the Poisson process [Daley and Vere-Jones 2003]. Therefore, a non-Poissonian distribution of the number of earthquakes in a finite time period implies that if a point process can model earthquakes well, this process must be time-dependent (although there might be other, non-point-process classes of models that are time-independent and generate non-Poissonian distributions). Therefore, the Poisson point-process representation is inadequate, even on five-year and ten-year timescales for large $M_w \geq 4.95$ earthquakes in Italy, because the rate variability of the time-independent Poisson forecasts is too small, and they will fail more often than expected. As a result, the agreement of the CSEP-Italy modelers to use a Poisson distribution should be viewed as problematic for time-independent models, because no Poisson distribution that their model could produce will ever pass the tests with the expected 95% confidence limits. On the other hand, time-dependent models with varying rates might generate a NBD over a longer time period (the empirical NBD can even be used as a constraint on the model).

Solutions to this problem have been discussed previously. The traditional approach has been to filter the data via declustering (deletion) of so-called aftershocks (e.g., as used in the RELM mainshock experiment [Field 2007, Schorlemmer et al. 2007]). However, the term «aftershock» is model-dependent and can only be applied retrospectively. A more objective approach is to forecast all of the earthquakes, allowing for time-dependence and non-Poissonian variability. In theory, each model can predict its own distribution for each space-time-magnitude bin [Werner and Sornette 2008], and future predictability experiments should consider allowing modelers to provide such a comprehensive forecast (see also Section 7).

A third ad-hoc solution [see Werner et al. 2010a] is more practical for time-independent models in the current context. Based on an empirical estimation of the observed variability of past earthquake numbers, the original Poisson forecasts of time-independent models can be reinterpreted to create forecasts that are characterized by a NBD. All of the tests (defined below in Section 5) can be performed using the original Poisson forecasts, and the tests can be repeated with so-called NBD forecasts.

We created NBD forecasts for the total number of observed events using the mean of each forecast and an imposed variance that was identical for all of the models, which we estimated either directly from the CPTI or from extrapolation, assuming that the observed number of events were uncorrelated. Appendix A describes this process in detail. Because the resulting NBD forecasts are tested on the same data that were used to estimate the variance, the NBD forecasts would be expected to perform well. The broader NBD results in less specificity, but also fewer unforeseen observations. We re-examine this ad-hoc solution in the discussion in Section 7.

5. Tests

To follow the agreed-upon rules of the prospective CSEP-Italy experiment, we used the statistical tests proposed for the RELM experiment and more recent ones that have been implemented within CSEP [Schorlemmer et al. 2007, Schorlemmer et al. 2010b, Zechar et al. 2010a]. These included: (i) the N(umber)-Test, based on the consistency between the total number of observed and expected earthquakes; (ii) the L(ikelihood)-Test, based on the consistency between the observed and expected joint log-likelihood scores of the forecast; (iii) the S(pace)-Test, based on the consistency between the observed and expected joint log-likelihood scores of the spatial distribution of the earthquakes; and (iv) the M(agnitude)-Test, based on the consistency between the observed and expected joint log-likelihood scores of the magnitude distributions of earthquakes.

The L-Test was proposed by Schorlemmer et al. [2007], and it is a relatively uninformative test, because the expected likelihood score is influenced by both the entropy of the model and the expected number of earthquakes. As the expected number increases, the expected likelihood score decreases. Therefore, a model that overpredicts the number of earthquakes will tend to underpredict the likelihood score. Because the L-Test is one-sided, i.e., a forecast is not rejected if the observed likelihood score is underpredicted, the models that overpredict the number of earthquakes might pass the L-Test not because they predicted seismicity correctly, but solely because they overpredicted the number of earthquakes [Zechar et al. 2010a, pp. 1190-1191, discussed a particular instance of this property of the L-Test in the context of the RELM experiment]. As a remedy, we also used a conditional L-Test [Werner et al. 2010a], in which the observed likelihood score was compared with the expected likelihood scores conditional on the number of observed quakes. In contrast to the S-Tests and M-Tests, the conditional L-Test assessed the joint space–magnitude forecast, but it overcame the sensitive dependence of the expected likelihood scores on the number of expected events.

6. Results

6.1. Testing five-year forecasts on the CSI

In Figure 2, we show the results of the N-, L-, S- and M-Tests as applied to the time-independent forecasts for the most recent five-year target period from 1998-2002 of the CSI. We discuss each of these test results below. As a summary of all of the results presented here and below, Table 2 and Table 3 list all of the tests that the forecasts failed for each of the considered target periods of the CSI and CPTI, respectively.

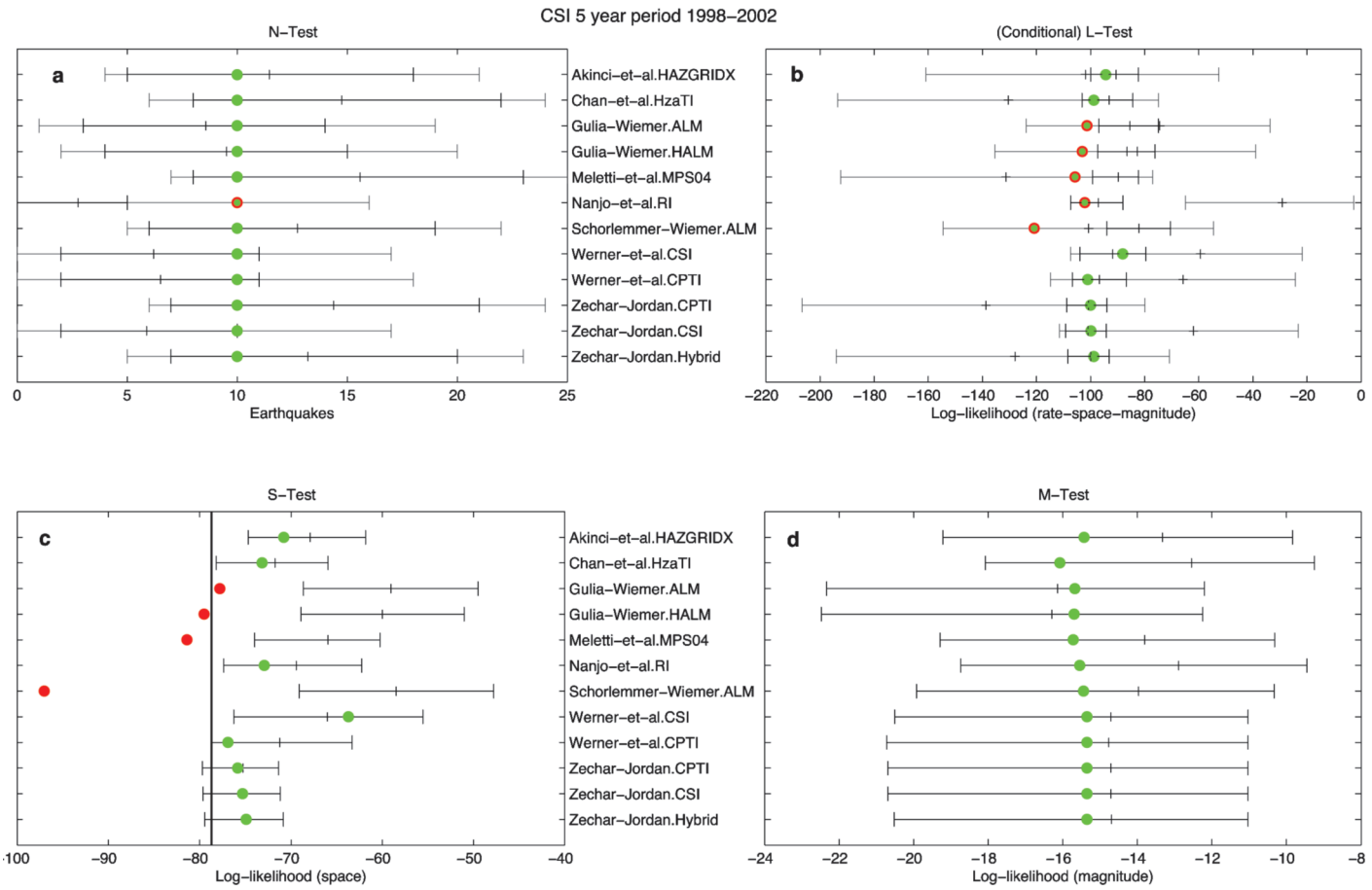

Figure 2. Results of the (a) N-Test, (b) unconditional/conditional L-Tests, (c) S-Test, and (d) M-Test of the 5-year time-independent forecasts using the 5-year target period from 1998 to 2002 of the data from the CSI. Red and green symbols, rejected and passed forecasts, respectively. In (a), green symbols with red edges, Poisson forecast was rejected while NBD forecast was passed. In (b), green symbols with red edges, only one of the two L-Tests was passed. Black crosses: (a) expected number of earthquakes; (b) expected unconditional/conditional log-likelihood score; (c) expected spatial log-likelihood score; (d) expected magnitude log-likelihood score, assuming the forecast was correct. Black bars: 95% confidence limits of the model forecasts assuming a Poisson distribution. In (a), gray bars, 95% confidence limits of the model forecast assuming a negative binomial distribution. In (b), gray/ black bars, 95% confidence limits of the unconditional/ conditional likelihood scores. In (c), vertical line, likelihood score of a spatially uniform model.

6.1.1. N-Test results

The N-Test results in Figure 2a show that only one forecast (NANJO–ET–AL.RI) was rejected assuming the Poisson confidence limits, because significantly more earthquakes were observed than expected. Using NBD uncertainties, none of the forecasts were rejected, because the confidence limits are wider (typically by several earthquakes on both sides).

6.1.2. L-Test results

In Figure 2b, we show the results of the unconditional and conditional L-Tests applied to the original Poisson forecasts. We did not try to apply NBD uncertainty to the rate forecasts in each space–magnitude bin, and therefore did not simulate the likelihood values based on a NBD forecast.

Only one forecast failed the unconditional L-Test, while four failed the conditional L-Test. The confidence limits of the unconditional L-Test were much larger because the number of simulated earthquakes was allowed to vary, thereby increasing the spread of the simulated likelihood scores. The impact of the expected number of earthquakes on the expected unconditional likelihood score was particularly visible for the forecasts of MELETTI–ET–AL.MPS04 and NANJO–ET–AL.RI. The MELETTI–ET–AL.MPS04 forecast expected more earthquakes than were observed during this period (although not significantly more), and therefore it also expected a likelihood score that is lower than observed. Moreover, the additional variability due to the increased number of events broadened the confidence limits and the model thus passed the L-Test. However, the forecast failed the conditional L-Test, because given the number of observed earthquakes, the observed likelihood score was too small to be consistent with the forecast. Meanwhile, the NANJO–ET–AL.RI forecast underpredicted the number of quakes (assuming Poisson variability), and therefore overpredicted the likelihood score and failed the unconditional L-Test. However, conditional on the number of observed earthquakes, the observed likelihood score was consistent with the forecast.

To summarize, the conditional L-Test reveals information that is separate from the N-Test results and presents a stricter evaluation of the forecasts. In the remainder of this report, we only consider the more informative conditional L-Test results. From the results of the 1998-2002 target period, we

can conclude that the joint distribution of the locations and magnitudes of the observed earthquakes were inconsistent with the group of ALM forecasts and the MELETTI–ET–AL.MPS04 forecast.

6.1.3. Reference forecast from a «model of most information»

To quantify the ability of the present time-independent forecasts to accurately predict the locations and magnitudes of the observed earthquakes, the likelihood score of an ideal earthquake forecast can be calculated (what might be called a successful prediction of the observed earthquakes – naturally, with the benefit of hindsight – or a forecast from a «model of most information», as opposed to a «model of least information» [Evison 1999], as discussed next). For instance, working within the constraints of a Poisson distribution of events in each bin, it is possible to calculate the likelihood score of a forecast that assigns an expected rate in each space-magnitude bin that is equal to the number of observed shocks within that bin. If one earthquake at most occurs per bin, then the observed log-likelihood score of such a perfect forecast is the negative number of observed events. The score is only slightly smaller if more than one event occurs in a given bin. In Figure 2b, the observed likelihood scores of the forecasts are evidently «far» from the score of a perfect forecast, which would roughly equal −10. The typical scores of the forecasts lie in the region of −100, which implies that the likelihood of the data under the perfect forecast is about 10^{39} times more likely than under a typical CSEP-Italy forecast. The information gain per earthquake [Harte and Vere-Jones 2005] of the perfect forecast over a typical forecast is of the order of 10^4.

These numbers help to quantify the differences between a perfect «prediction» within the current CSEP experimental design and a typical probabilistic earthquake forecast. Potentially, this index of earthquake predictability can be tracked to quantify the progress of the community of earthquake modelers towards better models. However, the primary goal of the CSEP experiments is to test and evaluate hypotheses about earthquake occurrence, and the observed degree of predictability is sufficient to carry this out.

6.1.4. Reference forecast from a «model of least information»

A forecast can equally be constructed from a «model of least information» [Evison 1999], which is often called the null hypothesis, and which might be based on a uniform spatial distribution, a total expected rate equal to the observed mean over a period prior to the target period, and a Gutenberg-Richter magnitude distribution with a *b*-value of 1.0. Because several of the models have already assumed that (i) magnitudes are identically and independently distributed according to the Gutenberg-Richter magnitude distribution, and (ii) the total expected rate is equal to the mean number of observed shocks, the only real difference between these models and a forecast that is not informative lies in the spatial distribution. We therefore included the likelihood score of a spatially uniform forecast only in the S-Test results. In Table S1 of the electronic supplement, we also provide the information gains per earthquake [Kagan and Knopoff 1977, Harte and Vere-Jones 2005] for each spatial forecast over a spatially uniform forecast for all of the considered target periods.

6.1.5. S-Test and M-Test results

The S-Test and M-Test results are shown in Figure 2c, d, and they suggest that the weaknesses of the group of ALM forecasts and the MELETTI–ET–AL.MPS04 forecast lie in forecasting the spatial distribution of the earthquakes: all four of these forecasts failed the S-Test with very small *p*-values, while all of these models passed the M-Test. Additionally, the forecasts of GULIA–WIEMER.HALM, MELETTI–ET–AL.MPS04 and SCHORLEMMER–WIEMER.ALM obtained scores that are lower than the score of a uniform model of least information.

In the case of the ALM group of forecasts, the low spatial likelihood scores that led to the S-Test failures have a common origin. In roughly one half of all of the spatial bins, the three forecasts expected an extremely small constant number of earthquakes per spatial bin, which indicates that a constant background rate was set in these cells. The GULIA–WIEMER.ALM and GULIA–WIEMER.HALM forecasts expected on the order of 10^{-8} earthquakes in each spatial background bin, while the SCHORLEMMER–WIEMER.ALM forecast expected an even smaller 10^{-15} earthquakes per bin. Accordingly, the probability of observing one earthquake in these bins is of the same order of magnitude. However, earthquakes do occur in some of these bins, and their occurrences in such low-probability (background) bins resulted in very low likelihood scores. Because these losses against a normalized uniform forecast, which expects roughly 10^{-3} earthquakes per bin to add to the 10 observed earthquakes, are not compensated for by equal or greater gains from earthquakes in the regions where the forecasts are higher, the forecasts obtained extremely small spatial likelihood scores and failed the S-Test.

During the 1998-2002 period, the GULIA–WIEMER.ALM and GULIA–WIEMER.HALM forecasts failed the S-Test because of one M_L 5.4 earthquake, located offshore, north of Sicily at 39.06° N and 15.02° E, which occurred in such a background rate bin. Similarly, the SCHORLEMMER–WIEMER.ALM forecast failed the S-Test because of a M_L 5.1 earthquake at 37.93° N and 17.55° E on the south-eastern boundary of the testing region. Apart from two other events, the remaining seven earthquakes during this target period occurred in cells where the ALM forecasts expected more earthquakes than the uniform forecast.

The distribution of rates of the MELETTI–ET–AL.MPS04 forecast showed a similar background rate, although it was larger (10^{-4} earthquakes per bin) than the background rates of the ALM forecasts. The occurrence of an earthquake in a

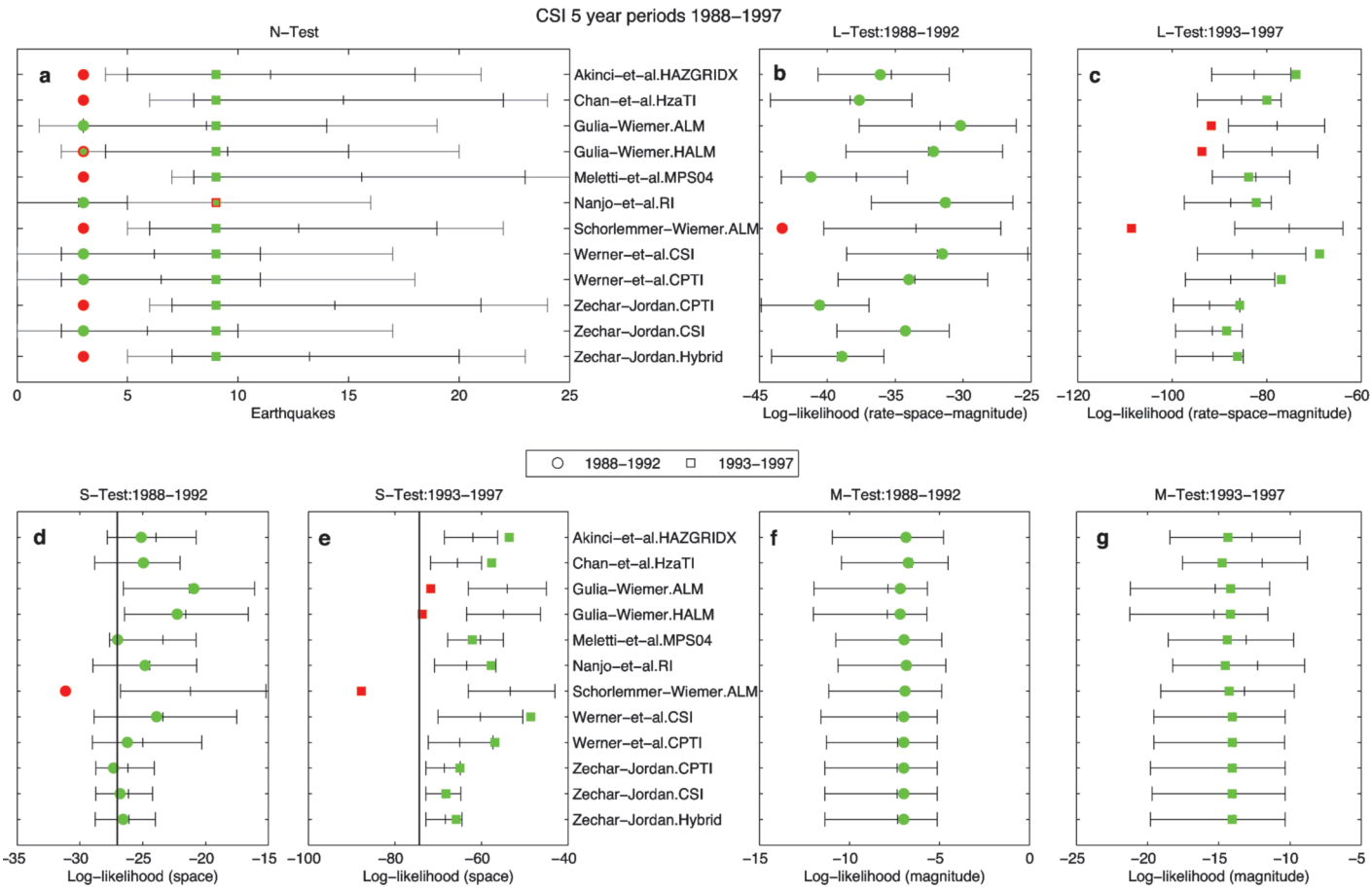

Figure 3. Results of the (a) N-Test, (b-c) conditional L-Test, (d-e) S-Test, and (f-g) M-Test of the 5-year time-independent forecasts using two separate 5-year target periods of the data from the CSI. Symbols and bars as given in the legend to Figure 1.

background bin can therefore be more easily compensated for by gains achieved from other earthquakes. However, during the 1998-2002 period, five earthquakes occurred in such background bins, and the losses were not masked by the gains. These five earthquakes included all four offshore earthquakes during this period (including the two events that caused the ALM forecasts to fail), along with one additional shock of magnitude M_L 5.3 at 46.697° N and 11.07° E in northern Italy.

6.1.6. Results from the other five-year target periods of the CSI

In Figure 3, we show the results of two further, separate, five-year target periods from the CSI: 1988-1992 and 1993-1997. In combination with Figure 2, this provides insight into the variability of the five-year test results that are due to natural fluctuations in seismicity.

For 1988-1992, only three target earthquakes occurred. Although this number is small, it falls within the 95% confidence limits of historical fluctuations (see Figure 1). Six forecasts were rejected by the N-Test because they overpredicted the number of observed events. These forecasts were: AKINCI–ET–AL.HAZGRIDX, CHAN–ET–AL.HZATI, MELETTI–ET–AL.MPS04, SCHORLEMMER–WIEMER.ALM, ZECHAR–JORDAN.CPTI and ZECHAR–JORDAN.HYBRID. As the results from longer target periods below confirm, this group consistently overpredicted the total rate. The modelers of the AKINCI–ET–AL.HAZGRIDX, CHAN–ET–AL.HZATI, MELETTI–ET–AL.MPS04, SCHORLEMMER–WIEMER.ALM, ZECHAR–JORDAN.CPTI and ZECHAR–JORDAN.HYBRID forecasts indicated to us that they calibrated their models on the M_w scale, rather than the M_L scale used for prospective testing, which led to an overprediction of the number of earthquakes with $M_L \geq 4.95$. This error in the calibration complicated the interpretation of the N-Test results for this group of models.

As before, we observed differences in the results from the NBD and Poisson N-Tests. For 1988-1992, the GULIA–WIEMER.HALM forecast was rejected by the N-Test assuming Poisson confidence limits, but the more realistic NBD uncertainties allowed the forecast to pass. Similarly, for 1993-1997, the NANJO–ET–AL.RI forecast failed the Poisson N-Test, but passed the NBD N-Test.

The conditional L-Test results indicated that for the SCHORLEMMER–WIEMER.ALM forecast, the three earthquakes during 1988-1992 were enough to reject the model. The results from the 1993-1997 period again showed rejections for the ALM group of forecasts. However, in contrast to the 1998-2002 period, the MELETTI–ET–AL.MPS04 forecast passed both periods. The results from the longer target periods that are presented below are necessary to judge this forecast conclusively.

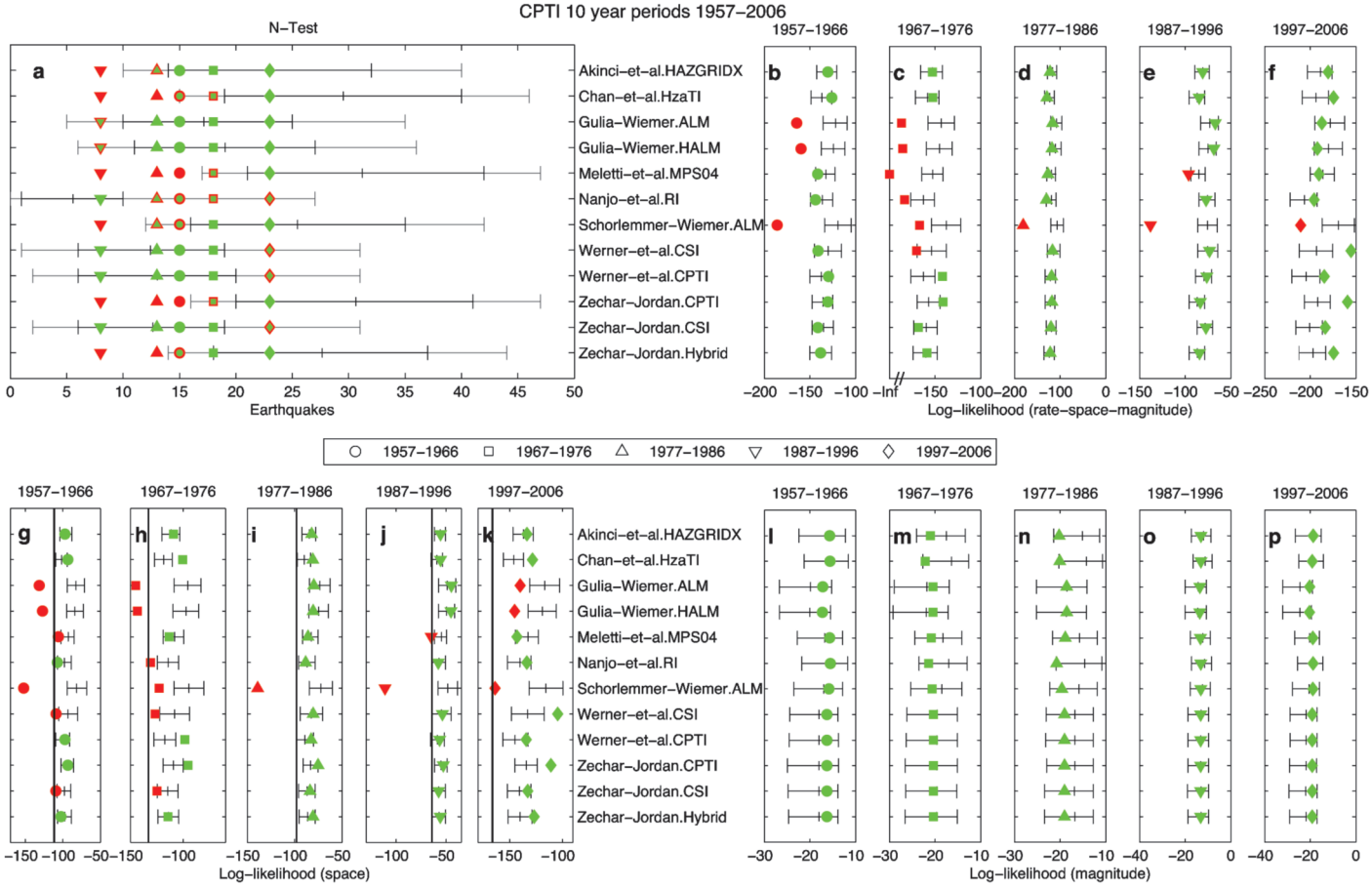

Figure 4. Results of the (a) N-Test, (b-f) conditional L-Test, (g-k) S-Test, and (l-p) M-Test of the 10-year time-independent forecasts using five separate 10-year target periods of data from the CPTI. Symbols and bars as given in the legend to Figure 1.

The combined S-Test and M-Test results again located the source of the ALM rejections in the spatial dimension of the forecast. Moreover, SCHORLEMMER–WIEMER.ALM continued to perform worse than a uniform model during both of the target periods. For the 1993-1997 target period, the forecasts failed because of a M_L 5.8 earthquake in 1994 at 39.398° N and 15.21° E, offshore of the north of Sicily, which occurred in a background bin. The large resulting likelihood loss cannot be compensated for by the gains achieved from the other eight observed earthquakes. For the 1988-1992 target period, the GULIA–WIEMER.ALM and GULIA–WIEMER.HALM forecasts passed the S-Test, although the SCHORLEMMER–WIEMER.ALM forecast received a low likelihood score because of an uncompensated likelihood loss due to a M_L 5.4 earthquake in 1990 at 37.33° N and 15.24° E offshore and east of Mount Etna, which occurred in a low-probability (but not background) bin. Additionally, the ZECHAR–JORDAN.CPTI forecast scored marginally less than a uniform forecast, although the score was consistent with the forecast expectation.

The M-Test results thus far, and for all but the longest of the target periods considered below, were not very informative: there were no rejections. The individual model distributions were very similar, which indicated that the differences between the predicted magnitude distributions were small. The differences between the observed likelihood scores were equally small.

To summarize, some of the test results varied for the considered five-year target period, while others were robust. SCHORLEMMER–WIEMER.ALM consistently showed poor performance in the spatial forecasts, while the other two ALM forecasts were rejected in two of the three target periods. MELETTI–ET–AL.MPS04 failed the conditional L-Tests and S-Tests for one of the three five-year target periods.

6.2. Testing ten-year forecasts on the CPTI

In Figure 4, we summarize the results of the N-, conditional L-, S- and M-Tests for the time-independent models and five non-overlapping ten-year target periods of the CPTI. These results mimicked the prospective ten-year experiment and helped gauge the variabilities of the results. The online material that accompanies this report (available at http://www.annalsofgeophysics.eu/index.php/annals/rt/suppFiles/4840/0) provides additional figures of the forecasts, maps of their likelihood ratios against a uniform forecast, and concentration diagrams [Rong and Jackson 2002, Kagan 2009] for the entire CPTI dataset from 1901 to 2006. Because the figures were based on the longest target period, which we consider explicitly in Section 6.3, they included all of the earthquakes observed during the ten-year target periods, and provided an informative visual presentation of the results.

6.2.1. N-Test results

The N-Test results are shown in Figure 4a. The numbers of observed shocks during the five ten-year periods were 15, 18, 13, 8 and 23. For the remainder of this report, we do not discuss the N-Test results from the group of models that were incorrectly calibrated on the M_w scale (see Section 6.1.6). Of the remaining six forecasts, none of them forecast all five observations within the 95% confidence limits of the Poisson distribution. Five forecasts – GULIA–WIEMER.ALM, GULIA–WIEMER.HALM, WERNER–ET–AL.CSI, WERNER–ET–AL.CPTI and ZECHAR–JORDAN.CSI – were rejected only for one of the five periods when assuming Poisson confidence limits, and were not rejected at all when considering the confidence limits based on a NBD.

The NANJO–ET–AL.RI forecast expected far fewer shocks than the other forecasts, and consistently underpredicted the number of earthquakes. Assuming the original Poisson variability in the number of shocks, the forecast was rejected for four of the five target periods. However, the forecast were not rejected at all when the NBD confidence limits were used.

6.2.2. Conditional L-Test results

The conditional L-Test results are shown in Figure 4b-f. The only robust result was the continued failure of the SCHORLEMMER–WIEMER.ALM forecast. The forecasts of GULIA–WIEMER.ALM and GULIA–WIEMER.HALM failed the test for two periods, while NANJO–ET–AL.RI and WERNER–ET–AL.CSI were both rejected for 1967-1976. The reasons for these rejections are discussed in the context of the S-Test and M-Test results below.

The MELETTI–ET–AL.MPS04 forecast obtained an observed joint-log-likelihood score of negative infinity for the target period 1967-1976. This score occurred because one earthquake occurred in a space-magnitude bin in which the forecast rate was zero. A zero forecast is equivalent to saying that target earthquakes are not possible in the bin, and if an event does occur in this bin, the forecast is automatically rejected. The earthquake in question was the 1968 Belice earthquake, which occurred on January 15, 1968, in western Sicily at 37.76° N and 12.98° E with a magnitude M_L 6.39, which caused several hundred fatalities. According to this forecast, however, earthquakes larger than $M_L = 6.25$ are impossible in this spatial bin, because the forecast rates in the magnitude bins are non-zero only for magnitudes up to $M_L = 6.25$. This forecast rejection implies that the maximum magnitude set for this location was too small; the discrepancy might be due to the wrong magnitude calibration reported above and/or the discrepancy might indicate that the maximum magnitude might require a modification in this area. (The MELETTI–ET–AL.MPS04 forecast did not fail the S-Test because the forecast in this particular spatial cell was non-zero when summed over the individual magnitude bins.)

6.2.3. S-Test and M-Test results

The S-Test results are shown in Figure 4g-k. Five (spatial) forecasts were not rejected by the S-Test for any of the five target periods. These forecasts were AKINCI–ET–AL.HAZGRIDX, CHAN–ET–AL.HZATI, WERNER–ET–AL.CPTI, ZECHAR–JORDAN.CPTI and ZECHAR–JORDAN.HYBRID.

The WERNER–ET–AL.CSI and ZECHAR–JORDAN.CSI forecasts, which were optimized on the CSI, both fared well during the target periods that are also at least partially covered by the CSI, i.e., from 1981 onwards. However, the two forecasts were rejected during the two earliest target periods, which can be considered as out-of-sample tests for these two forecasts. For the 1957-1966 period, these forecasts failed to predict several diffuse earthquakes in northern Italy and two offshore earthquakes between the Ligurian coast and Corsica. The 1967-1976 period contained the 1968 western Sicily earthquake sequence (including the above-mentioned $M_L = 6.39$ Belice earthquake), which occurred in spatial cells with low expected rates. Evidently, the CSI contained little seismicity in these regions from which these models could have anticipated the occurrence of these earthquakes.

Of interest, the NANJO–ET–AL.RI forecast, which was also calibrated on the CSI data, only failed for the 1967-1977 period (again, due to the western Sicily sequence in 1968); it passed for the 1957-1966 interval. The model used a relatively coarse grid to forecast the earthquakes (see Figure S6 of the electronic supplement), and this characteristic helped it to forecast the offshore quakes north of Corsica better than the WERNER–ET–AL.CSI and ZECHAR–JORDAN.CSI forecasts.

The three ALM-based forecasts continued to forecast the spatial distribution of the observed earthquakes poorly. For the 1957-1966 target period, the two above-mentioned earthquakes north of Corsica and a shock in northern Italy occurred in the background bins of all three of these ALM forecasts, leading to their S-Test failures. For the 1967-1976 target periods, the GULIA–WIEMER.ALM and GULIA–WIEMER.HALM forecasts failed because of three earthquakes in the background bins: two shocks that occurred as part of the 1968 western Sicily earthquake sequence, and one in central Italy at 44.81° N and 10.35° E. While none of these events (nor any others) occurred in the background bins of the SCHORLEMMER–WIEMER.ALM forecast for this period, two earthquakes of the 1968 western Sicily sequence, as well as an earthquake at 41.65° N and 15.73° E, did incur unexpectedly low likelihood scores, which resulted in the S-Test rejection. Indeed, SCHORLEMMER–WIEMER.ALM failed all of the considered ten-year target periods. Whenever the spatial likelihood score fell below a uniform forecast, at least one earthquake occurred in a so-called background cell.

The MELETTI–ET–AL.MPS04 forecast was rejected twice by the S-Test. For the period 1957-1966 the forecast failed because of the two recurring offshore earthquakes north of Corsica in July 1963, and because of two earthquakes in

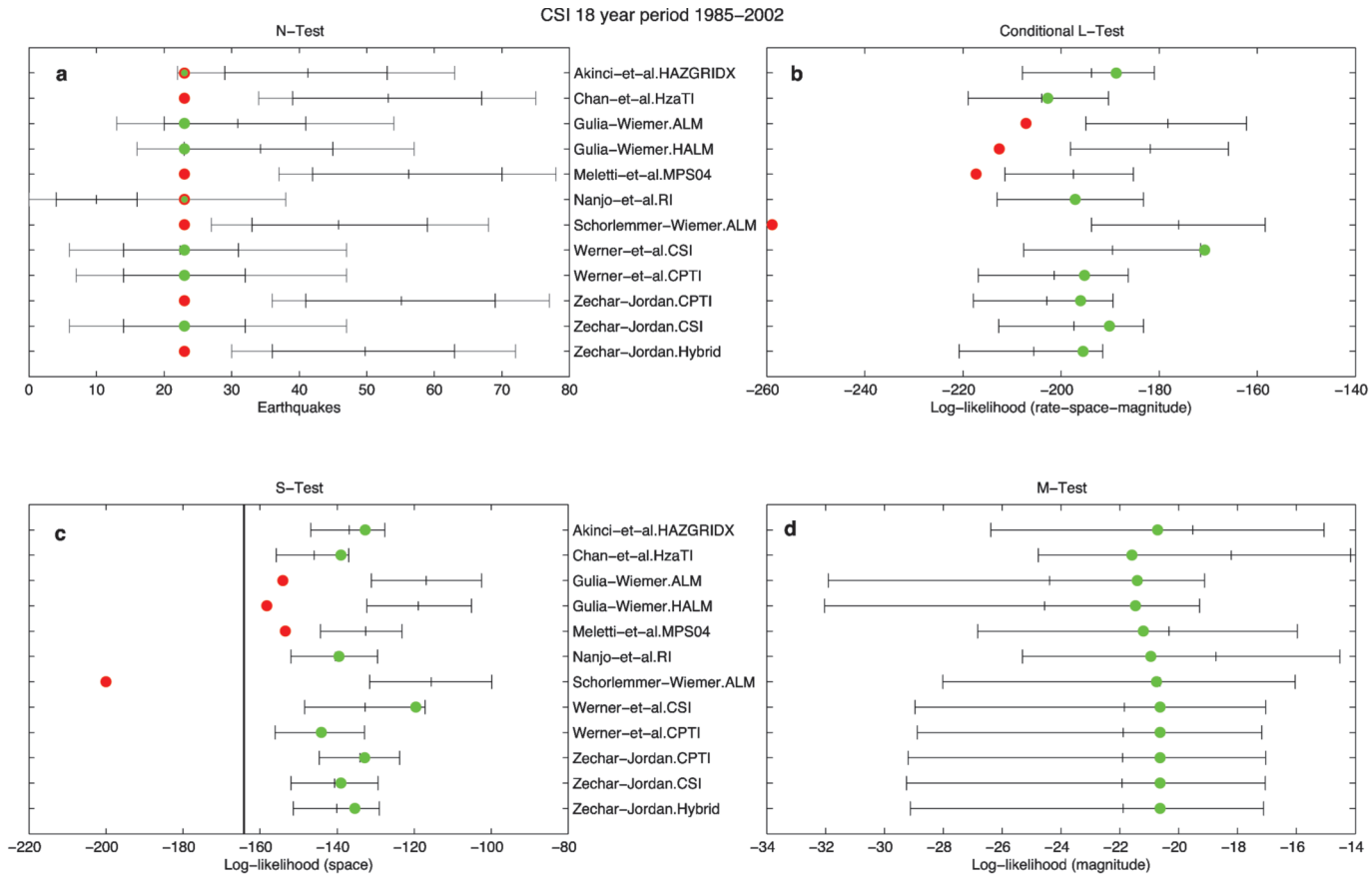

Figure 5. Results of the (a) N-Test, (b) conditional L-Test, (c) S-Test, and (d) M-Test of the scaled 10-year time-independent forecasts using the 18-year target period from 1985 to 2002 of the data from the CSI. Symbols and bars as given in the legend to Figure 1.

northeastern Italy, all of which occurred in background bins. For 1987-1996, three earthquakes also occurred in background bins: (i) an offshore earthquake on April 26, 1988, at 42.21° N and 16.66° E; (ii) an M_L 5.43 aftershock of the Potenza, southern Italy, earthquake of May 5, 1990; and (iii) an M_L 5.54 offshore earthquake on December 13, 1990, east of Mount Etna in the Sea of Sicily.

6.3. Test results from longer periods

The long-term forecasts submitted for the CSEP-Italy experiment were calculated for five-year and ten-year periods. Because the forecasts are time-independent and characterized by Poisson uncertainty, suitably scaled versions of the forecasts can be tested on longer time periods: 18 years (duration of the reliable part of the entire CSI, from 1985 to 2002), 57 years (duration of the most reliable part of the CPTI, from 1950 to 2006), and 106 years (the entire CPTI). In this section, we present the results of testing these scaled forecasts. The accompanying online material includes further figures for these forecasts, likelihood ratios, and concentration diagrams based on the 106-year target period.

The test results of the 18-year period from 1985 to 2002 of the CSI are shown in Figure 5. Twenty-three earthquakes occurred during this period. The N-Test results revealed the same features as already observed previously, with a group of models that overpredicted the number of earthquakes: AKINCI–ET–AL.HAZGRIDX, CHAN–ET–AL.HZATI, MELETTI–ET–AL.MPS04, SCHORLEMMER–WIEMER.ALM, ZECHAR–JORDAN.CPTI and ZECHAR–JORDAN.HYBRID. While the confidence limits of the negative binomial distribution remained substantially wider than the limits based on the Poisson distribution, there were only two forecasts for which the test results were ambiguous: AKINCI–ET–AL.HAZGRIDX and NANJO–ET–AL.RI. The ALM forecasts and the MELETTI–ET–AL.MPS04 forecast failed the conditional L-Test and the S-Test, with SCHORLEMMER–WIEMER.ALM scoring less than a uniform spatial forecast. These failures are due to the earthquakes discussed above, which occurred either in background bins or in locations with low expected rates.

Increasing the duration of the retrospective tests to the 57 most recent years of the CPTI (1950-2007) yielded 83 earthquakes and led to similar results, although with greater statistical significance (Figure 6). In addition to the rejections mentioned in the preceding paragraph, the N-Test now unequivocally rejected the AKINCI–ET–AL.HAZGRIDX and NANJO–ET–AL.RI forecasts, and even when the confidence limits of a NBD were considered. The conditional L-Test rejected the MELETTI–ET–AL.MPS04 forecast because of a likelihood score of negative infinity (discussed in Section 6.2.2). The S-Test results showed that the NANJO–ET–AL.RI forecast was rejected in addition to the ALM forecasts and MELETTI–ET–AL.MPS04. No forecasts were rejected by the M-Test, despite 57 years of data.

The longest period over which we evaluated the scaled

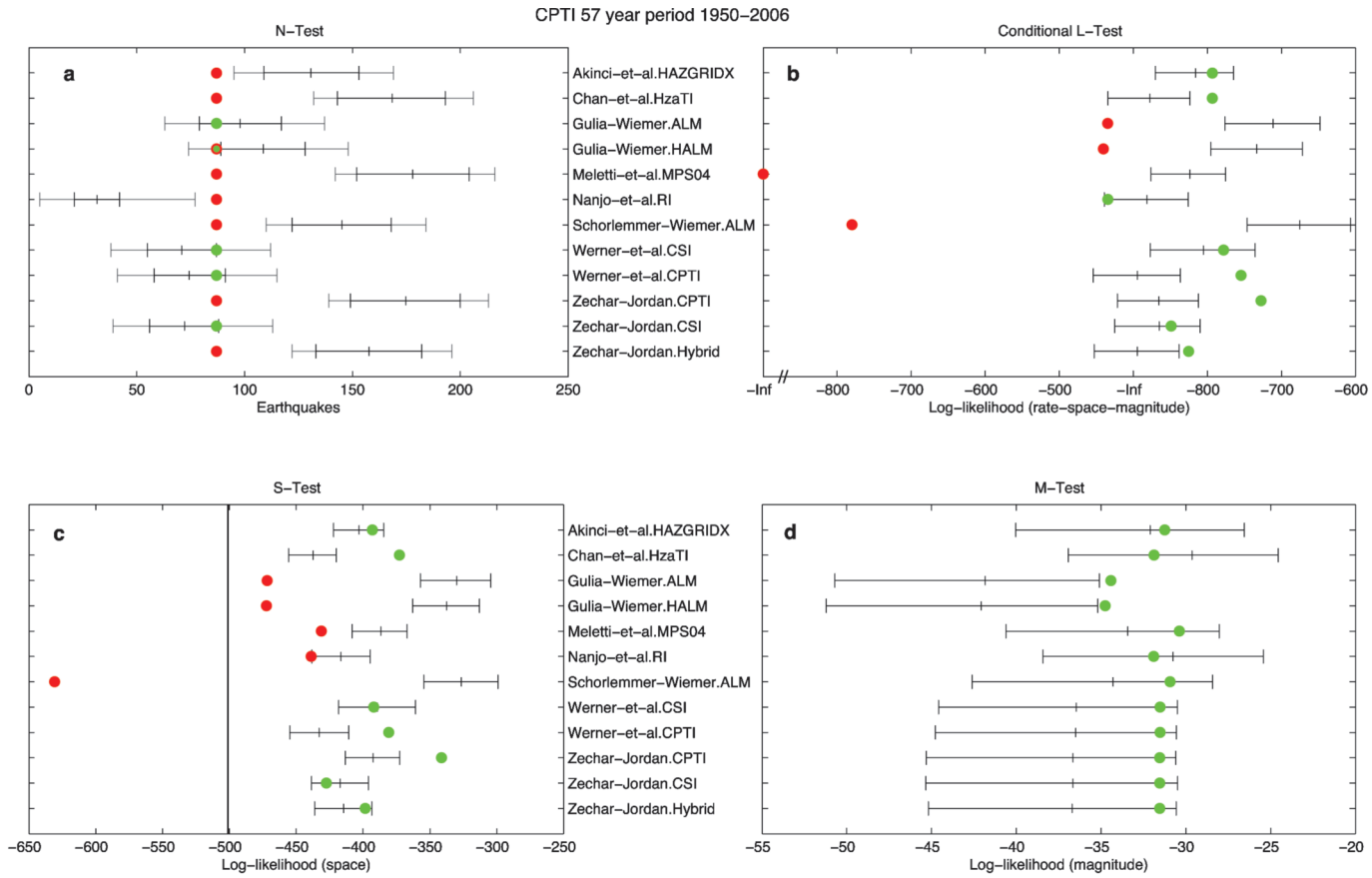

Figure 6. Results of the (a) N-Test, (b) conditional L-Test, (c) S-Test and (d) M-Test of the scaled 10-year time-independent forecasts using the 57-year target period from 1950 to 2006 of the data from the CPTI. Symbols and bars as given in the legend to Figure 1.

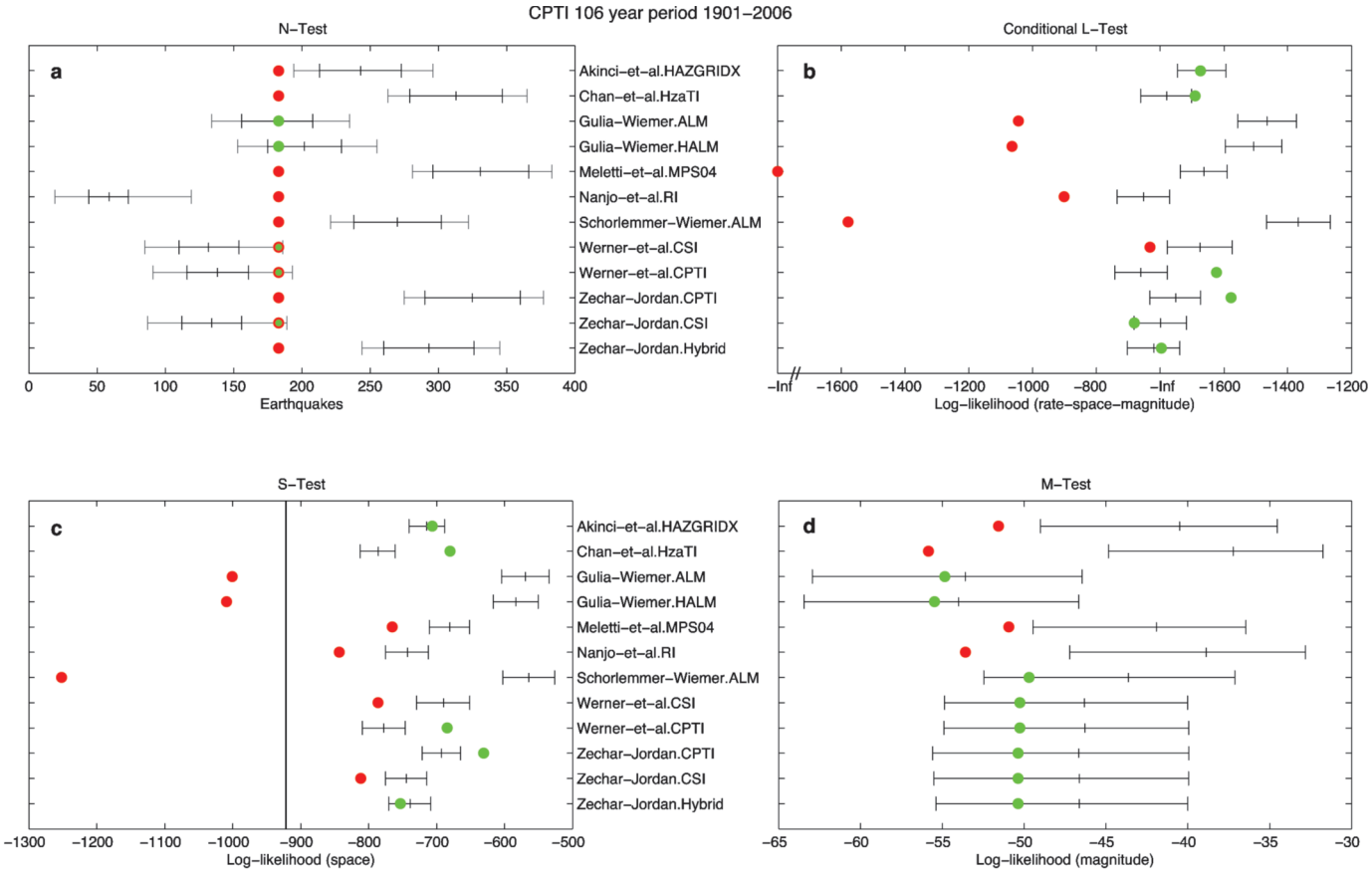

Figure 7. Results of the (a) N-Test, (b) conditional L-Test, (c) S-Test and (d) M-Test of the scaled 10-year time-independent forecasts using the 106-year target period from 1901 to 2006 of the data from the CPTI. Symbols and bars as given in the legend to Figure 1.

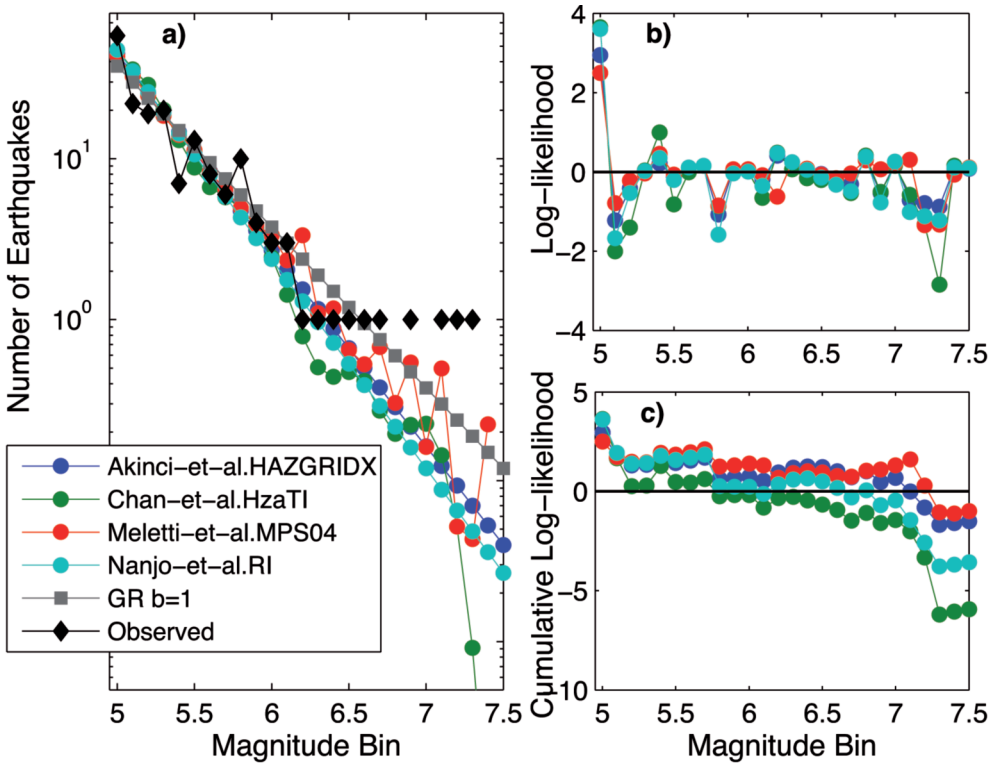

Figure 8. Magnitude distributions and likelihood scores for the four forecasts that failed the M-Test for the 106 year (1901-2006) CPTI target period, and, for comparison, a pure Gutenberg-Richter (GR) distribution with *b*-value equal to 1.0 ($b = 1$), which passed the test. (a) Observed and predicted magnitude histograms, pure GR law. (b, c) Bin-wise log-likelihood ratios (b) and cumulative log-likelihood ratios (c) of the forecasts against the pure GR law.

forecasts was 106 years, which spanned the full duration of the CPTI08 and contained 183 earthquakes (Figure 7; see online material for maps of the forecasts, likelihood ratios and concentration diagrams). The N-Test results now showed a clear separation between the group of forecasts that consistently overpredicted, the NANJO–ET–AL.RI forecast that underpredicted, and the forecasts that were not rejected when assuming confidence limits based on either a Poisson or a NBD. Application of the conditional L-Test additionally rejected the NANJO–ET–AL.RI and WERNER–ET–AL.CSI forecasts, while the S-Test now also failed for WERNER–ET–AL.CSI and ZECHAR–JORDAN.CSI.

Interestingly, four forecasts failed the M-Test: AKINCI–ET–AL.HAZGRIDX, CHAN–ET–AL.HZATI, MELETTI–ET–AL.MPS04 and NANJO–ET–AL.RI. In Figure 8, we have compared the observed magnitude distributions with those predicted. For reference, we added a pure Gutenberg-Richter (GR) distribution with *b*-value of 1.0, which passed the M-Test. The magnitude distributions predicted by AKINCI–ET–AL. HAZGRIDX and NANJO–ET–AL.RI were close to exponential, but with *b*-values larger than 1.0. As a result, large earthquakes were predicted as less likely, and the forecasts were thus penalized for the occurrence of three $M_L > 7$ earthquakes. The magnitude distribution of the CHAN–ET–AL.HZATI forecast appeared to reflect its non-parametric kernel estimation method (see Section 2), and it also underpredicted the rate of large shocks. Finally, the magnitude distribution of MELETTI–ET–AL.MPS04 was non-monotonic: several characteristic magnitude bulges were seen. However, the largest events occurred between the bulges, for which the forecast was penalized.

7. Discussion and conclusions

7.1. The role of the Poisson distribution in the forecast specification

The assumption of Poisson rate variability in the CSEP-Italy experiments (as well as other CSEP experiments, including RELM [Field 2007, Schorlemmer et al. 2007]) has certain advantages. In particular, this is a simplifying assumption: because the Poisson distribution is defined by a single parameter, the forecasts do not require a complete probability distribution in each bin. Moreover, Poisson variability has often been used as a reference model against which to compare time-varying forecasts, and it provides intuitive understanding.

Despite these advantages, however, this assumption is questionable, and the method of forcing each forecast to be characterized by the same uncertainty is not the only solution [see also the discussion in Zechar et al. 2010a]. Werner and Sornette [2008] commented that most forecast models generate their own likelihood distribution, and this distribution depends on the particular assumptions of the model; moreover, there is no reason to force every model to use the same form of likelihood distribution. The effect of this forcing is probably stronger for time-dependent, e.g. daily,

forecasts [Lombardi and Marzocchi 2010b], and it is difficult to judge the quality of approximating each model-dependent distribution to a Poisson distribution (without the help of modelers). On the other hand, whether or not the Poisson assumption is appropriate with respect to observations can be checked. In Section 4.3, we showed that the target earthquake rate distribution was better approximated by a NBD than by a Poisson distribution. Therefore, time-independent forecasts that predict Poisson rate variability necessarily fail more often than expected at the 95% confidence limits because the distribution observed differs from the model distribution. To improve time-independent forecasts, the (non-Poissonian and potentially negative binomial) marginal rate distribution over long timescales needs to be estimated. However, the parameter values of the NBD rate change as a function of the temporal and spatial boundaries of the study region over the available observational periods [Kagan 2010]. Whether a stable asymptotic limit exists (loosely speaking, whether seismic rates are stationary) remains an open question. For time-dependent models, on the other hand, there are several classes that can produce a NBD rate over finite time periods that include branching processes [Kagan 2010] and Poisson processes with a stochastic rate parameter distributed according to the Gamma distribution.

Despite this criticism, it is unlikely that the Poisson distribution will be replaced by a model-dependent distribution that is substantially different, particularly for long-term models. Therefore, although the *p*-values of the test statistics used in the N-, L-, S- and M-Tests might be biased towards lower values, they do provide rough estimates. Nevertheless, it needs to be borne in mind that a quantile score that is outside the 95% confidence limits of the Poisson distribution might be within the acceptable range if a model-dependent distribution were used. As an example, and to explore the effects of the Poisson assumption in these experiments, we created a set of modified forecasts with the rate variability estimated from the observed history. The width of the 95% confidence interval of the total rate forecast increased, and in certain cases substantially so. Several forecasts were rejected if a Poisson variability was assumed, while they passed the test under the assumption of a NBD. Overall, however, the *p*-values (quantile scores) of the test statistics based on the Poisson approximation often gave good approximate values. Only in borderline cases did the Poisson assumption lead to (potentially) false rejections of forecasts.

The modified forecasts based on a NBD are not an entirely satisfactory solution to the problem, however. First, the model distribution in each bin should arise naturally from the hypotheses of a model, rather than from an empirical adjustment made by those evaluating the forecast. Second, even if a negative binomial distribution adequately represents the distribution of the total number of observed events in an entire testing region, the parameter values for each bin should be specified to make the non-Poisson forecasts amenable also to the L-, S- and M-Tests. Therefore, future experiments should allow forecasts that are not characterized by Poisson rate uncertainty.

More generally, future experiments can consider other forecast formats and additional model classes. For example, stochastic point process models provide a continuous likelihood function that can characterize conditional dependence in time, magnitude and space (and focal mechanisms, etc.). As a result, full likelihood-based inference for point processes and tools for model diagnostics are applicable to this class of models [e.g., Ogata 1999, Daley and Vere-Jones 2003, Schoenberg 2003]. However, when considering new classes of forecasts, it should be borne in mind that a major success of the RELM and CSEP experiments was the homogenization of forecast formats to facilitate comparative testing.

7.2. Performance and utility of the tests

We explored the results from the N-, L-, S- and M-Tests in this study because they are the «staple» CSEP tests. Other metrics for evaluating forecasts should certainly be considered, especially with regard to alarm-based tests [e.g., Molchan and Keilis-Borok 2008, Zechar and Jordan 2008] and further conditional likelihood tests [Zechar et al. 2010a]. Overall, the N-, L-, S- and M-Tests are intuitive and relatively easy to interpret. However, we demonstrated a weakness in the L-Test, and replaced it with a conditional L-Test that better assessed the quality of the forecasts [see also Werner et al. 2010a]. Among the metrics, the S-Test results were the most helpful in tracking down the weak features of forecasts, because the biggest differences between time-independent models lie in their spatial forecasts.

The M-Test results were generally not informative. Because the magnitude distributions considered here were so similar, this result is not surprising; indeed, it is in agreement with the statistical power exploration of Zechar et al. [2010a]. No forecast could be rejected for target periods ranging from 5 to 57 years. Different tests, such as the traditional Kolmogorov-Smirnov test, should be compared with the current likelihood-based M-Test, particularly in terms of statistical power.

The current status quo in the CSEP experiments is to reject a forecast if it fails a single test at 95% confidence. As we discussed above, the actual *p*-values provided more meaningful assessments than a simple binary yes/no statement, because the assumed confidence limits may not accurately represent the model uncertainty. Furthermore, as the suite of tests grows, we should be concerned with the joint confidence limits of the ensemble of tests, rather than the individual significance levels of each test. Joint confidence limits can be obtained from model simulations. A global confidence limit for the multiple

	CSI*			
Model	1988-1992	1993-1997	1998-2002	1985-2003
AKINCI-ET-AL.HAZGRIDX	N^+			N_p^+
CHAN-ET-AL.HZATI	N^+			N^+
GULIA-WIEMER.ALM		$\hat{L}$, S	$\hat{L}$, S	$\hat{L}$, S
GULIA-WIEMER.HALM	N_p^+	$\hat{L}$, S	$\hat{L}$, S	$\hat{L}$, S
MELETTI-ET-AL.MPS04	N^+		$\hat{L}$, S	N^+, $\hat{L}$, S
NANJO-ET-AL.RI		N_p^-	N_p^-, L	N_p^-
SCHORLEMMER-WIEMER.ALM	N^+, $\hat{L}$, S	$\hat{L}$, S	$\hat{L}$, S	N^+, $\hat{L}$, S
WERNER-ET-AL.CSI				
WERNER-ET-AL.CPTI				
ZECHAR-JORDAN.CPTI	N^+			N^+
ZECHAR-JORDAN.CSI				
ZECHAR-JORDAN.HYBRID	N^+			N^+

Table 2. Summary results of the forecast tests obtained using the CSI. *For each model and each experiment time period, the tests that the forecast failed are indicated, according to a 5% critical significance value. For the N-Test, N^+ indicates that the forecast overpredicted the observed rate, N^- indicates underprediction; the subscript *p* indicates that the forecast only failed when assuming a Poisson uncertainty; otherwise it failed under both the Poisson and NBD.

tests can then be established. A similar question will arise when forecasts from the same model are tested within nested regions, as will be the case when considering the performance of a model forecast for Italy with that for the entire globe.

Finally, future experiments can consider developing tests that address particular characteristics of a forecast [see also discussion in Zechar et al. 2010a]. For example, a forecast might be a reflection of the hypothesis that the magnitude distribution varies as a function of the tectonic setting. In this context, an M-Test conditioned on the spatial distribution of the observed earthquakes would provide a more powerful test.

7.3. Overall performance of the forecasts

A summary of all of the results is given in Tables 2 and 3. The Poisson N-Test is possibly the strictest test within the present context, because none of the forecasts pass every N-Test of the different periods. On the other hand, five forecasts pass all of the N-Tests with confidence limits based on a negative binomial distribution: GULIA–WIEMER.ALM, GULIA–WIEMER.HALM, WERNER–ET–AL.CSI, WERNER–ET–AL.CPTI and ZECHAR–JORDAN.CSI. As mentioned above, several of the modelers indicated to us that their forecasts were calibrated on the M_w scale. As a result, it was difficult to interpret their overpredictions, beyond the obvious statement that the forecasts were poorly calibrated. The NANJO–ET–AL.RI forecast is the only forecast that expects substantially fewer earthquakes than the observed sample mean, although the forecast failed the NBD N-Test only for the longest of the considered target periods. The forecasts that expected the same number of shocks as the sample mean over their calibration period predicted the number of earthquakes well, as should be expected.

With one important exception, the results from the conditional L-Test largely reflected the S-Test results, because the predicted magnitude distributions were consistent with observations from all but the 106-year target period. The exception concerns the occurrence of an earthquake in a space-magnitude bin in which an earthquake should have been impossible according to the forecast: the 1968 M_L 6.39 Belice earthquake was in a spatial cell in which the MELETTI–ET–AL.MPS04 forecast set a maximum magnitude of M_L 6.25. This discrepancy might be explained by the wrong magnitude conversion that was adopted and/or it suggests that the model assumptions regarding the spatial variation of maximum magnitudes need to be revised. However, if we had tested this forecast against the M_w of the Belice earthquake (M_w 6.33, according to the CPTI), the forecast would have still failed, thus indicating the latter explanation.

The S-Test results provided the most insight into the weaknesses of the forecasts. Only five forecasts passed all of the S-Tests: AKINCI–ET–AL.HAZGRIDX, CHAN–ET–AL.HZATI, WERNER–ET–AL.CPTI, ZECHAR–JORDAN.CPTI and ZECHAR–JORDAN.HYBRID. These forecasts fit the spatial distributions of the CSI and CPTI well, although they might overfit and perform poorly in the future. The models are also among the simplest, especially when compared to the MELETTI–ET–AL.MPS04 forecast. However, the WERNER–ET–AL.CSI and ZECHAR–JORDAN.CSI forecasts, which were calibrated on the CSI data, did not adequately forecast the spatial locations of earthquakes during the period before the CSI data began. This might indicate that the models are not smooth enough and do not sufficiently anticipate that quiet regions can become active.

The ALM group of forecasts (GULIA–WIEMER.ALM, GULIA–WIEMER.HALM and SCHORLEMMER–WIEMER.ALM) consistently failed the S-Tests, and often performed worse than a uniform forecast, because isolated earthquakes occurred in extremely low-probability «background» bins that covered roughly 50% of the region. Among these earthquakes that occurred in background bins, we could not identify a common characteristic. The likelihood losses incurred were not compensated for by the gains achieved by adequately forecasting the majority of the earthquakes.

	CPTI*						
Model	57-66	67-76	77-86	87-96	97-06	1950-2006	1901-2006
AKINCI-ET-AL.HAZGRIDX			N_p^+	N^+		N^+	N^+
CHAN-ET-AL.HZATI	N_p^+	N_p^+	N^+	N^+		N^+	N^+
GULIA-WIEMER.ALM	$\hat{L}$, S	$\hat{L}$, S		N_p^+	S	$\hat{L}$, S	$\hat{L}$, S
GULIA-WIEMER.HALM	$\hat{L}$, S	$\hat{L}$, S		N_p^+	S	N_p^+, $\hat{L}$, S	$\hat{L}$, S
MELETTI-ET-AL.MPS04	N^+, S	N_p^+, $\hat{L}$	N^+	N^+, $\hat{L}$, S		N^+, $\hat{L}$, S	N^+, $\hat{L}$, S
NANJO-ET-AL.RI	N_p^-	N_p^-, $\hat{L}$, S	N_p^-		N_p^-	N^-, S	N^-, $\hat{L}$, S
SCHORLEMMER-WIEMER.ALM	N_p^+, $\hat{L}$, S	$\hat{L}$, S	N_p^+, $\hat{L}$, S	N^+, $\hat{L}$, S	$\hat{L}$, S	N^+, $\hat{L}$, S	N^+, $\hat{L}$, S
WERNER-ET-AL.CSI	S	$\hat{L}$, S			N_p^-		N_p^-, $\hat{L}$, S
WERNER-ET-AL.CPTI					N_p^-		N_p^-
ZECHAR-JORDAN.CPTI	N^+	N_p^+	N^+	N^+		N^+	N^+
ZECHAR-JORDAN.CSI	S	S			N_p^-		N_p^-, S
ZECHAR-JORDAN.HYBRID	N_p^+		N^+	N^+		N^+	N^+

Table 3. Summary results of the forecast tests obtained using the CPTI. *For each model and each experiment time period, the tests that the forecast failed are indicated, according to a 5% critical significance value. For the N-Test, N^+ indicates that the forecast overpredicted the observed rate, N^- indicates underprediction; the subscript *p* indicates that the forecast only failed when assuming a Poisson uncertainty; otherwise it failed under both the Poisson and NBD.

These results suggest that the ALM forecasts are overly optimistic in ruling out earthquakes in their background bins, i.e. the models are not smooth enough.

The MELETTI–ET–AL.MPS04 forecast also often failed the S-Test, because of a minority of earthquakes that occurred in low-probability regions. Almost all of the earthquakes that incurred likelihood losses were located offshore. However, while the forecast performed substantially better onshore, a few surprising onshore earthquake locations remained. Poor performance of a forecast for offshore earthquakes potentially raises the problem of the «weight» of each earthquake in the testing procedure. Specifically, if a model is intended for the practical purpose of seismic hazard assessment, then a rejection of its forecast due to offshore earthquakes might not have the same importance as a rejection due to earthquakes in regions of higher exposure and/or vulnerability.

Eight forecasts passed all of the M-Tests: GULIA–WIEMER.ALM, GULIA–WIEMER.HALM, SCHORLEMMER–WIEMER.ALM, WERNER–ET–AL.CSI, WERNER–ET–AL.CPTI, ZECHAR–JORDAN.CPTI, ZECHAR–JORDAN.CSI and ZECHAR–JORDAN.HYBRID. Five of these are based on a simple Gutenberg-Richter distribution with a uniform *b*-value of 1.0: WERNER–ET–AL.CSI, WERNER–ET–AL.CPTI, ZECHAR–JORDAN.CPTI, ZECHAR–JORDAN.CSI and ZECHAR–JORDAN.HYBRID. This suggests that the hypothesis of a universally applicable, uniform Gutenberg-Richter distribution with a *b*-value of 1.0 [e.g., Bird and Kagan 2004] cannot be ruled out for the region of Italy.

Four of the forecasts failed the M-Test for the 1901-2007 target period of the CPTI. The magnitude distributions of the forecasts of AKINCI–ET–AL.HAZGRIDX, NANJO–ET–AL.RI, CHAN–ET–AL.HZATI and MELETTI–ET–AL.MPS04 did not adequately forecast the largest magnitudes, and the three observed $M_L > 7$, in particular. For the AKINCI–ET–AL.HAZGRIDX and NANJO–ET–AL.RI forecasts, the reason appears to be a *b*-value of the Gutenberg-Richter distribution that is too large. The non-parametric estimation of CHAN–ET–AL.HZATI also decayed too quickly. The magnitude distribution of MELETTI–ET–AL.MPS04 revealed several characteristic magnitude values with elevated rates, but earthquakes also occurred between them in extremely-low-probability bins. However, these results should be interpreted cautiously, because the same magnitude forecasts passed the 1950-2007 period, and because the greater uncertainty of the data prior to 1950 arguably influenced the results.

7.4. Value of retrospective evaluation

The initial submission deadline for long-term earthquake forecasts for CSEP-Italy was July 1, 2009. Because the formal experiment was not intended to start until August 1, 2009, there was a brief period for initial analysis and quality control of the forecasts submitted. We provided a quick summary of the features of the forecasts and preliminary results of a retrospective evaluation to the modelers during this period. As a result, six of the 18 time-independent and time-dependent long-term forecasts were modified and resubmitted before the final deadline of August 1, 2009. This initial quality control period was therefore extremely useful, and future experiments should consider expanding and formalizing the initial quality control period.

The short one-month period was, however, too short to evaluate the forecasts retrospectively in the detail that we now present here. During the course of this study, the problem of the wrong magnitude scaling was discovered. Because at least four of the 18 forecasts were affected, a second round of submissions was requested for November 1, 2009, and 15 revisions (and two new forecasts) were submitted. This suggests that the feedback provided to modelers based on the present study was useful and informative. The task of converting even a relatively simple hypothesis into a testable, probabilistic earthquake forecast should not be underestimated, and we suggest that future experiments include some form of

retrospective testing prior to final submission.

The retrospective evaluation also showed that at least the time-independent forecasts can be evaluated in a meaningful manner, and that useful information about the models can be extracted. This information is critical for the development of better forecasts and for the evaluation of the underlying hypotheses of earthquake occurrence.

At the same time, retrospective evaluations cannot replace prospective tests with zero degrees of freedom. Given the relative robustness of the results from the retrospective evaluation, we anticipate that prospective experiments will provide further useful and more definite information about the quality of these forecasts. Most importantly, if the second round of forecast submissions contains significantly improved forecasts with fewer technical errors, we expect to see real progress in our understanding of earthquake predictability.

Data and sharing resources

We used two earthquake catalogs for this study: the parametric catalog of Italian earthquakes (Catalogo Parametrico dei Terremoti Italiani, CPTI08) [Rovida and the CPTI Working Group 2008] and the catalog of Italian seismicity (Catalogo della Sismicità Italiana, CSI 1.1) [Chiarabba et al. 2005, Castello et al. 2007]. The particular versions of these catalogs that we used are available at http://www.cseptesting.org/regions/italy.

Acknowledgements. We thank the following for their contributions to the CSEP-Italy Working Group: A. Akinci, C.-H. Chan, A. Christophersen, R. Console, F. Euchner, L. Faenza, G. Falcone, M. Gerstenberger, L. Gulia, A. Helmstetter, M. Liukis, A.M. Lombardi, C. Meletti, M. Murru, K. Nanjo, B. Pace, L. Peruzza, D. Rhoades, D. Schorlemmer, M. Stucchi and J. Woessner. MJW was supported by the EXTREMES project of ETH Competence Center for the Environment and Sustainability (CCES). The tests were performed in the European CSEP Testing Center at ETH Zurich, which is funded in part through the European Union project NERIES. MJW thanks the Southern California Earthquake Center (SCEC) for travel support. SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. The SCEC contribution number for this paper is 1436.

References

Akaike, H. (1974). A new look at the statistical model identification, IEEE Trans. Automatic Control, AC-19, 716-723.

Akinci, A. (2010). HAZGRIDX: earthquake forecasting model for $M_L \geq 5.0$ earthquakes in Italy based on spatially smoothed seismicity, Annals of Geophysics, *53*, *3* (present issue).

Bird, P. and Y.Y. Kagan (2004). Plate-tectonic analysis of shallow seismicity: Apparent boundary width, beta, corner magnitude, coupled lithosphere thickness, and coupling in seven tectonic settings, Bull. Seismol. Soc. Am., *94*, 2380-2399; doi: 10.1785/0120030107.

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.

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, Annals of Geophysics, *53*, *3* (present issue).

Chiarabba, C., L. Jovane and R. Stefano (2005). A new view of Italian seismicity using 20 years of instrumental recordings, Tectonophysics, *395*, 251-268.

Cornell, C.A. (1968). Engineering seismic risk analysis, Bull. Seismol. Soc. Am., *58* (5), 1583-1606.

Daley, D.J. and D. Vere-Jones (2003). An Introduction to the Theory of Point Processes, vol. I, New York, USA.

Evison, F.F. (1999). On the existence of earthquake precursors, Annals of Geophysics, *42* (5), 763-770.

Faenza, L. and W. Marzocchi (2010). The Proportional Hazard Model as applied to the CSEP forcasting area in Italy, Annals of Geophysics, *53*, *3* (present issue).

Falcone, G., R. Console and M. Murru (2010). Short-term and long-term earthquake occurrence models for Italy: ETES, ERS and LTST, Annals of Geophysics, *53*, *3* (present issue).

Field, E.H. (2007). A summary of previous Working Groups on California Earthquake Probabilities, Bull. Seismol. Soc. Am., *97* (4), 1033-1053, doi: 10.1785/0120060048.

Gardner, J.K. and L. Knopoff (1974). Is the sequence of earthquakes in Southern California, with aftershocks removed, Poissonian?, Bull. Seis. Soc. Am., *64* (5), 1363-1367.

Gulia, L., S. Wiemer and D. Schorlemmer (2010). Asperity-based earthquake likelihood models for Italy, Annals of Geophysics, *53*, *3* (present issue).

Harte, D. and D. Vere-Jones (2005). The entropy score and its uses in earthquake forecasting, Pure and Applied Geophysics, *162*, 1229-1253; doi:10.1007/s00024-004-2667-2.

Jackson, D.D. and Y.Y. Kagan (1999). Testable earthquake forecasts for 1999, Seismol. Res. Lett., *70* (4), 393-403.

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

Kagan, Y.Y. (1973). Statistical methods in the study of seismic processes, Bull. Int. Stat. Inst., *45* (3), 437-453 (with discussion).

Kagan, Y. and L. Knopoff (1977). Earthquake risk prediction as a stochastic process, Physics of the Earth and Planetary Interiors, *14*, 97-108; doi: 10.1016/0031-9201(77)90147-9.

Kagan, Y.Y. (2009). Testing long-term earthquake forecasts: likelihood methods and error diagrams, Geophys. J. Intern., *177* (2), 532-542; doi: f10.1111/j.1365-246X.2008.04064.xg.

Kagan, Y.Y. (2010). Statistical distributions of earthquake numbers: Consequence of branching process, Geophys. J. Intern., *180*, 1313-1328; doi: 10.1111/j.1365-246X.2009.04487.x.

Lombardi, A.M. and W. Marzocchi (2010a). A double-branching model applied to long-term forecasting of Italian seismicity ($M_L \geq 5.0$) within the CSEP project, Annals of Geophysics, *53*, *3* (present issue).

Lombardi, A.M. and W. Marzocchi (2010b). Exploring the per-

formances and usability of the CSEP suite of tests, Bull. Seismol. Soc. Am., in review.

Molchan, G. and V. Keilis-Borok (2008). Earthquake prediction: probabilistic aspect, Geophysical Journal International, *173*, 1012-1017; doi: 10.1111/j.1365-246X.2008.03785.x.

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, Annals of Geophysics, *53*, *3* (present issue).

Ogata, Y. (1999). Estimating the hazard of rupture using uncertain occurrence times of paleoearthquakes, J. Geophys. Res., 104 (B8), 17,995-18,014.

Pace, B., L. Peruzza and F. Visini (2010). LASSCI2009.2: layered earthquake rupture forecast model for central Italy, submitted to the CSEP project, Annals of Geophysics, *53*, *3* (present issue).

Reasenberg, P. (1985). Second-order moment of central California seismicity, 1969-82, J. Geophys. Res., 90, 5479-5495.

Rong, Y.F. and D.D. Jackson (2002). Earthquake potential in and around China: Estimated from past earthquakes, Geophys. Res. Lett., *29* (16); doi: f10.1029/2002GL015297g.

Rovida, A. and the CPTI Working Group (2008). Catalogo Parametrico dei Terremoti Italiani, 1901-2006, versione 2008 (CPTI08), INGV, Milano-Pavia; http://www.cseptesting.org/regions/italy.

Saichev, A. and D. Sornette (2006). Power law distribution of seismic rates: theory and data analysis, Eur. Phys. J. B, *49*, 377-401; doi: 10.1140/epjb/e2006-00075-3.

Schoenberg, F.P. (2003). Multidimensional residual analysis of point process models for earthquake occurrences, J. Am. Stat. Assoc., 98 (464), 789-795; doi: 10.1198/016214503000000710.

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

Schorlemmer, D., M.C. Gerstenberger, S. Wiemer, D.D. Jackson and D.A. Rhoades (2007). Earthquake likelihood model testing, Seismol. Res. Lett., 78 (1), 17.

Schorlemmer, D. and J. Woessner (2008). Probability of Detecting an Earthquake, Bull. Seismol. Soc. Am., 98 (5), 2103-2117; doi: 10.1785/0120070105.

Schorlemmer, D., A. Christophersen, A. Rovida, F. Mele, M. Stucci and W. Marzocchi (2010a). Setting up an earthquake forecast experiment in Italy, Annals of Geophysics, *53*, *3* (present issue).

Schorlemmer, D., J.D. Zechar, M.J. Werner, E. Field, D.D. Jackson and T.H. Jordan (2010b). First results of the regional earthquake likelihood models experiment, Pure and Appl. Geophys.: The Frank Evison Volume, 167 (8-9); doi: 10.1007/s00024-010-0081-5.

Schorlemmer, D., F. Mele and W. Marzocchi (2010c). A completeness analysis of the National Seismic Network of Italy, J. Geophys. Res. Solid Earth, 115; doi: 10.1029/2008JB006097.

Vere-Jones, D. (1970). Stochastic models for earthquake occurrence, J. Roy. Stat. Soc. Series B (Methodological), *32* (1), 1-62 (with discussion).

Weichert, D.H. (1980). Estimation of the earthquake recurrence parameters for unequal observation periods for different magnitudes, Bull. Seismol. Soc. Am., 70 (4), 1337-1346.

Werner, M.J. and D. Sornette (2008). Magnitude uncertainties impact seismic rate estimates, forecasts and predictability experiments, J. Geophys. Res. Solid Earth, 113; doi: 10.1029/2007JB005427.

Werner, M.J., A. Helmstetter, D.D. Jackson and Y.Y. Kagan (2010a). High resolution long- and short-term earthquake forecasts for California, Bull. Seismol. Soc. Am., in revision; preprint available at http://arxiv.org/abs/0910.4981.

Werner, M.J., A. Helmstetter, D.D. Jackson, Y.Y. Kagan and S.Wiemer (2010b). Adaptively smoothed seismicity earthquake forecasts for Italy, Annals of Geophysics, *53*, *3* (present issue); preprint available at http://arxiv.org/abs/1003.4374.

Wiemer, S. and D. Schorlemmer (2007). ALM: An Asperity-based Likelihood Model for California, Seismological Research Letters, 78 (1), 134-140; doi:10.1785/gssrl.78.1.134.

Woessner, J. and S. Wiemer (2005). Assessing the quality of earthquake catalogues: Estimating the magnitude of completeness and its uncertainty, Bull. Seismol. Soc. Am., *95* (2), 684-698.

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); doi: 10.1007/s00024-010-0086-0.

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

Zechar, J.D., M.C. Gerstenberger and D. Rhoades (2010a). Likelihood-based tests for evaluating the spatial and magnitude component of earthquake rate forecasts, Bull. Seismol. Soc. Am., 100 (3); doi: 10.1785/0120090192.

Zechar, J.D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P.J. Maechling and T.H. Jordan (2010b). The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science, Concurrency and Computation: Practice and Experience, 22, 1836-1847; doi: 10.1002/cpe.1519.

Appendix A

Negative-binomial forecasts

To create NBD forecasts, we used the total expected rate of each forecast as the average of the distribution, and we fixed the variance of the forecast equal to the observed sample variance from the CPTI (estimated in Section 4.3). Thus, for the five-year experiments, we used $\sigma^2_{5\mathrm{yr}} = 23.73$, while for ten-year experiments, we used $\sigma^2_{10\mathrm{yr}} = 64.54$.

For longer time periods (e.g., the durations of the CSI and CPTI), for which we cannot estimate the sample variance directly, we used the property that the variance of a finite sum of uncorrelated random variables is equal to the sum of their variances. We treated the numbers of the observed earthquakes as uncorrelated random variables, meaning that we assumed that the numbers of the observed earthquakes in adjacent time intervals were independent of each other. This is likely to be a better approximation for the ten-year intervals. We computed the variance $\sigma^2(T)$ over some finite interval of T years from the reference variance $\sigma^2_{10\mathrm{yr}}$ using Equation (A1):

$$\sigma^2(T) = \frac{T}{10}\,\sigma^2_{10\mathrm{yr}} \qquad \text{(A1)}$$

Table 4 lists the estimated and calculated variances for the various time intervals used in this study. If needed, the NBD parameters can be estimated from Equations (4) and (5). Because the direct estimate of $\sigma^2_{10\mathrm{yr}}$ is more than twice that of $\sigma^2_{5\mathrm{yr}}$, it appears that there might be correlations at the five-year time scale. Alternatively, the sample size might be too small, because the 95% confidence intervals are large.

Time interval T [yrs]	Estimated $\sigma^2(T)$
5	23.73*
10	64.54*
18	116.17
57	367.88
106	684.12

Table 4. Estimated variance of the numbers of observed earthquakes for the different time intervals. *The variance was estimated directly from the catalog. The others values were computed using Equation (A1).

*Corresponding author: Maximilian J. Werner,
ETH Zurich, Swiss Seismological Service, Zurich, Switzerland;
email: mwerner@sed.ethz.ch

Electronic supplement (additional figures of earthquake forecasts, likelihood ratios and concentration diagrams) available at:
http://www.annalsofgeophysics.eu/index.php/annals/rt/suppFiles/4840/0