Earth, Planets and Space
a SpringerOpen Journal

FULL PAPER Open Access

Some reasoning on the RELM-CSEP likelihood-based tests

Anna Maria Lombardi

Abstract

The null hypothesis is the essence of any statistical test: this is basically a comparison of what we observe with what we would expect to see if the null hypothesis was true. In this work, I explore the suitability of the null hypothesis of likelihood-based tests (LBTs), which are often adopted by the laboratories of the Collaboratory for the Study of Earthquake Predictability (CSEP), to check earthquake forecast models. First, I discuss the LBT in the wider context of classical statistical hypothesis testing. Then, I present some cases in which the null hypothesis of LBT is not appropriate for determining the merits of earthquake forecast models. I justify these results from a theoretical point of view, within the framework of point process theory. Finally, I propose a possible upgrade of LBT to enable the correct assessment of the forecasting capability of earthquake models. This study may provide new insights to the CSEP LBT.

Keywords: Statistical tests; Earthquake forecast; Point processes

Background

The increasing interest of the seismological community in earthquake forecasting has highlighted the need for a proper evaluation of forecast models. This has motivated the birth of the working group on Regional Earthquake Likelihood Models (RELM, Schorlemmer and Gerstenberger 2007) and of the Collaboratory for the Study of Earthquake Predictability (CSEP, Jordan 2006), both designed to evaluate the quality of forecast models. The protocol adopted by RELM/CSEP is based on classical statistical hypothesis testing (Schorlemmer et al. 2007). This is then finalized to reject or accept the null hypothesis (hereinafter H_0) on the basis of a numerical summary of the data. RELM/CSEP working groups adopt two main types of testing methods: likelihood-based tests (LBTs) (Schorlemmer et al. 2007; Zechar et al. 2010) and alarm-based tests (ABTs) (Zechar and Jordan 2008). In this study, I focus on LBTs and specifically on N and L tests (Schorlemmer et al. 2007).

The RELM/CSEP working groups formalized the LBT to test hypotheses that 'should follow directly the model, so that if the model is valid, the hypothesis should be consistent with data used in a test. Otherwise, the hypothesis, and the model on which it was constructed, can be rejected' (Schorlemmer et al. 2007). Actually, as I discuss below, this intent was not attained (Lombardi and Marzocchi 2010a; Schorlemmer et al. 2007, 2010a; Werner et al. 2010).

The CSEP testing centers use the N and L tests to check the consistency of expected ($\Lambda = \{\lambda_{(i,j)}\}$) and observed ($\Omega = \{\omega_{(i,j)}\}$) values of variables $X_{(i,j)}$, representing the number of earthquakes with magnitude above a threshold M_F, in nonoverlapping bins $\{(T_i, R_j); T_i \in \mathcal{T}, R_j \in \mathcal{R}\}$ of a predetermined spatio-temporal space $\mathcal{S} = \mathcal{R} \times \mathcal{T}$ (Jordan 2006; Zechar et al. 2010). A model is represented by forecasts Λ, which are the only values provided by the modelers. The correct calculation of the p values of the LBT requires the probability distribution of $X_{(i,j)}$ given by the model and specifically the probabilities

$$p_n^{ij} = P\{X_{(i,j)} = n\} \qquad \text{for } n = 0, 1, 2, \ldots \tag{1}$$

As this information is not available to modelers, the LBT assumes, as the null hypothesis H_0, that the variables $X_{(i,j)}$ are independent and follow a Poisson distribution with mean $\lambda_{(i,j)}$. Therefore, the set of probabilities p_n^{ij} are substituted for the probabilities

$$q_n^{ij} = \frac{[\lambda_{(i,j)}]^n}{n!} \exp^{-\lambda_{(i,j)}} \qquad \text{for } n = 0, 1, 2, \ldots \tag{2}$$

and the p values of the LBT are computed accordingly (Schorlemmer et al. 2007).

Correspondence: annamaria.lombardi@ingv.it
Istituto Nazionale di Geofisica e Vulcanologia, via di Vigna Murata, 605, Rome 00143, Italy

Specifically, the N test measures the probability of observing $N_i^O = \sum_j \omega_{(i,j)}$ events, for each forecast time period T_i. The p values of the N test are given by the probabilities (Zechar et al. 2010):

$$\delta_1 = P(X_i \geq N_i^O) \qquad \delta_2 = P(X_i \leq N_i^O), \tag{3}$$

where $X_i = \sum_j X_{(i,j)}$. The RELM/CSEP protocol rejects a model if δ_1 or δ_2 is too small, meaning that the model overpredicts or underpredicts the observed seismicity. Under H_0, X_i is a Poisson variable with expectation $N_i^F = \sum_j \lambda_{(i,j)}$ (and PDF $q_n^i = [N_i^F]^n e^{-N_i^F}/n!$), and the percentiles δ_1/δ_2 are computed by this distribution (see Schorlemmer et al. 2007).

The L-test measures the probability of the joint log-likelihood $L(\Omega_i|\Lambda)$ of observing Ω, given the forecast Λ. Under H_0, $L(\Omega_i|\Lambda)$ is given by:

$$L(\Omega_i|\Lambda) = \sum_j \left[\ln \frac{(\lambda_{(i,j)})^{\omega_{(i,j)}}}{\omega_{(i,j)}!} - \lambda_{(i,j)} \right]. \tag{4}$$

The p value of the L test is estimated by comparing $L(\Omega_i|\Lambda)$ with a predetermined number N of synthetic likelihood values $L(\Omega_i^S|\Lambda) = \{L(\Omega_i^{S_l}|\Lambda), l = 1, \ldots, N\}$, computed by Equation 4, of simulated catalogs 'consistent with the forecast' (Schorlemmer et al. 2007). This means that the forecast grids $\Omega_i^{S_l}$ are simulated according to the Poisson hypothesis supposed by H_0, and the p value of the L test is given by the proportion of simulated log-likelihoods below the value $L(\Omega_i|\Lambda)$:

$$\gamma = \frac{\left| \{ L(\Omega_i^{S_l}|\Lambda) \;\middle|\; L(\Omega_i^{S_l}|\Lambda) \leq L(\Omega_i|\Lambda); l = 1, \ldots, N \} \right|}{N}. \tag{5}$$

This shows that the LBT does not check the hypothesis that a forecast model has merit with the given data (marked hereinafter by Hyp_1). Actually, the LBT tests whether $\{\omega_{(i,j)}\}$ are independent random variables, from a Poisson population with mean $\{\lambda_{(i,j)}\}$ (marked hereinafter by Hyp_2). When a model is not consistent with Hyp_2, i.e., when the set of probabilities $\{p_n^{ij}\}$ is significantly different from $\{q_n^{ij}\}$, the specific computation of the p values of the LBT is misleading, causing a potentially unjustified rejection of the model itself (Lombardi and Marzocchi 2010a).

The CSEP laboratories still systematically use the LBT, but a process of revision has begun. This study is intended to provide a contribution to this process.

Methods

A suitable revision of the LBT requires the full recognition and quantification of the causes and effects of the present inefficiencies. For this purpose, I apply the N and L tests to two classes of 1,000 simulated forecast grids, generated by different spatio-temporal magnitude models. In this way, the data are perfectly known, and the rejection of H_0 cannot mean the failure of the model being tested.

First, I generate two sets of synthetic catalogs. Each catalog covers a time period of 1 month (January 1 to 31, 2012), the Italian collecting region, and a magnitude range of [2.5, 9.0], as chosen by CSEP (Schorlemmer et al. 2010b).

The first class of simulations is consistent with a version of the epidemic-type aftershocks sequence (ETAS) model (Ogata 1998), submitted to the CSEP-Italy testing region (Lombardi and Marzocchi 2010b). The rate of the model at time t, with location (x, y) and magnitude m, is given by:

$$\lambda_1(t, x, y, m/\mathcal{H}_t) = \left\{ \mu \cdot u(x, y) + \sum_{T_i < t} \frac{K e^{\alpha(M_i - M_0)}}{(t - T_i + c)^p} \frac{c^i_{dq\gamma}}{[r_i^2 + (d e^{\gamma(M_i - M_0)})^2]^q} \right\} \times \frac{b 10^{b(m - M_0)}}{1 - 10^{b(M_{\max} - M_0)}} \tag{6}$$

where $\{\mu, K, c, p, \alpha, d, q, \gamma, b\}$ are the model parameters, M_0 and $M_{\max}$ are the minimum and maximum magnitudes, $\mathcal{H}_t = \{(T_i, X_i, Y_i, M_i); T_i < t\}$ is the history (i.e., the information relative to past events) up to time t, and r_i is the distance between location (x, y) and the epicenter of the ith event (X_i, Y_i) (see Lombardi and Marzocchi 2010b, for details). To compute the rate $\lambda_1(t, x, y, m/\mathcal{H}_t)$, I include in the history the seismic bulletin of the Istituto Nazionale di Geofisica e Vulcanologia (INGV) from April 16, 2005 to December 31, 2011. Moreover, I add a synthetic event $(T_{\text{ms}}, X_{\text{ms}}, Y_{\text{ms}}, M_{\text{ms}})$ at time 00:00:00 on January 1, 2012 (T_{ms}), with magnitude $M_{\text{ms}} = 6.0$ and coordinates $(X_{\text{ms}}, Y_{\text{ms}}) = (13.384°E, 42.346°N)$. The parameter values used in this study are $\mu = 0.7$, $K = 0.026$, $p = 1.15$, $c = 0.01$, $\alpha = 1.4$, $d = 0.7$, $q = 1.5$, $\gamma = 0.3$, $b = 1.0$, $M_0 = 2.5$, and $M_{\max} = 9.0$.

To generate the ETAS forecasts for day T_i and catalog C_k, I mimic the CSEP real-time experiment: specifically, I include the triggering rate for events with history $\mathcal{H}_{T_i}$ of C_k and average the triggering rates of 1,000 simulated realizations of the process inside T_i (see Lombardi and Marzocchi 2010b, for details).

The second class of simulations follows a nonstationary poisson (NP) process. Specifically, the rate $\lambda_2(t, x, y, m)$ is given by a stationary background and the triggering effect of event $(T_{\text{ms}}, X_{\text{ms}}, Y_{\text{ms}}, M_{\text{ms}})$. The rate of the NP model is as follows:

$$\lambda_2(t, x, y, m) = \left\{ \mu u(x, y) + \frac{K e^{\alpha(M_{\text{ms}} - M_0)}}{(t - T_{\text{ms}} + c)^p} \frac{c_{dq\gamma}}{\left[r^2 + (d e^{\gamma(M_{\text{ms}} - M_0)})^2\right]^q} \right\} \times \frac{b 10^{b(m - M_0)}}{1 - 10^{b(M_{\max} - M_0)}} \tag{7}$$

where r is the distance between (x, y) and $(X_{\rm ms}, Y_{\rm ms})$. The parameters used here are $\mu = 0.7$, $K = 0.1$, $p = 0.9$, $c = 0.02$, $\alpha = 1.4$, $d = 0.7$, $q = 1.5$, $\gamma = 0.3$, $b = 1.0$, $M_0 = 2.5$, $M_{\rm max} = 9.0$.

The simulations represent the average seismicity of the first month of a sequence (following a shock with magnitude 6.0), as predicted by the ETAS and NP models. The basic difference between the models is that the rate of the ETAS model depends on the whole history $\mathcal{H}_t$ (i.e., information relative to past events), whereas the rate of the NP model depends on the coordinates of only one event $(T_{\rm ms}, X_{\rm ms}, Y_{\rm ms}, M_{\rm ms})$. Thus, the rate of the NP model is deterministic and decreasing in time from $T_{\rm ms}$, whereas the rate of the ETAS model has a random nonmonotonic time evolution, depending on history $\mathcal{H}_t$.

For each synthetic catalog, I compute the 1-day binned forecast grids Λ ($M_F = 2.5$) by integrating (in time, space, and magnitude) the rate of the model used to generate the catalog. The forecast grids Λ cover a period of 1 month (starting from January 1, 2012) and the test spatial grid adopted for the CSEP Italian laboratory (Schorlemmer et al. 2010b). Finally, I apply the CSEP/RELM N and L tests (with significance level $\alpha = 0.05$ and $M_F = 2.5$) on all simulated catalogs, using the forecast grids previously computed.

In this paper, I propose an obvious upgrade of LBT, which does without the Poisson distribution. First, the discrete log-likelihood function $L(\Omega_i|\Lambda)$ of variables X_i (Equation 4) is substituted for the continuous-time log-likelihood function (hereinafter, CLF). This is a proper measure of the agreement between model and data, taking into account the features of a model. For a spatio-temporal magnitude earthquake model, this is given by

$$\mathrm{CLF} = \sum_{i=1}^{N_{\mathcal{R}x\mathcal{T}x[M_0 M_{\max}]}} \ln \lambda(t_i, x_i, y_i, m_i) - \int_{\mathcal{T}} \int_{\mathcal{R}} \int_{M_0}^{M_{\max}} \lambda(t, x, y, m) dt dx dy dm \qquad (8)$$

where $\lambda(t, x, y, m)$ is the rate of the model (Daley and Vere-Jones 2003) and $N_{\mathcal{R}x\mathcal{T}x[M_0 M_{\max}]}$ is the number of events inside the spatio-temporal magnitude space $\mathcal{R}x\mathcal{T}x[M_0 M_{\max}]$.

Second, the percentiles of the distributions of both the variables X_i and the CLF are derived directly by the model. This information allows the computation of more reliable p values for the tests (Werner and Sornette 2008; Schorlemmer et al. 2010a).

In brief, the new testing procedure presented here consists of the following steps:

1. For each forecast period T_i, the number of events (Ω_i) and the CLF ($\mathrm{CLF}_{M,i}$) of model M being tested are computed.
2. For each T_i, N catalogs given by model M are simulated; the occurrences $\Omega^S_{M,i} = \left\{\Omega^{S_l}_{M,i}, l = 1, \ldots, N\right\}$ and the likelihood $\mathrm{CLF}^S_{M,i} = \left\{\mathrm{CLF}^{S_l}_{M,i}, l = 1, \ldots, N\right\}$ are computed for all catalogs.
3. The percentiles of the empirical distributions generated in the previous step, used to perform a test at the 95% confidence level, are estimated. Specifically, the 2.5th and 97.5th percentiles $\left(P^{\Omega}_{M,i}[2.5\%] \text{ and } P^{\Omega}_{M,i}[97.5\%]\right)$ of values Ω^S_i and the 5th percentile $\left(P^{\mathrm{CLF}}_{M,i}[5\%]\right)$ of quantities $\mathrm{CLF}^S_{M,i}$ are identified.
4) The observed values Ω_i and $\mathrm{CLF}_{M,i}$ are compared with the percentiles computed in the previous step. In this way, model M is rejected or retained for T_i. Specifically, model M is rejected if $\Omega_i < P^{\Omega}_{M,i}[2.5\%]$ or $\Omega_i > P^{\Omega}_{M,i}[97.5\%]$ or if $\mathrm{CLF}_{M,i} \leq P^{\mathrm{CLF}}_{M,i}[5\%]$.

In this procedure, the percentiles of model M are estimated by simulations because it is often not possible to derive them analytically. However, the use of simulations is not mandatory for modelers, of course.

Results

First, I apply the CSEP LBT to two classes of ETAS and NP simulations. Figure 1a shows the fraction of rejections F_R (i.e., the proportion of catalogs for which H_0 is rejected) of the N and L tests as a function of time. As shown in Lombardi and Marzocchi (2010a), F_R for the ETAS simulations is well above 5%, which is the threshold justifiable by chance. On the other hand, F_R for the NP simulations is close to or below 5%, suggesting that Hyp_2 is consistent with the NP model.

To investigate whether previous results depend on M_F or on the average seismic rate of the region, I apply the procedure described above to 1,000 new catalogs, reproducing the average seismicity of Japan (which has a seismic rate two orders of magnitude higher than that of Italy). These datasets are simulated by using an *ad hoc* ETAS model of this region. In this experiment, I consider a forecast time span T_i of 3 months, an overall time period of 10 years, and $M_F = 4.0$. This last value is the threshold magnitude adopted by the Japanese CSEP laboratory for short-term forecasting experiments (Nanjo et al. 2011; Tsuruoka et al. 2012). I find that F_R is equal to 40% *to* 50% and 60% *to* 75% for the N and L tests, respectively (see Figure 1b).

I apply the new testing procedure described previously to the simulated Japanese catalogs. This gives the values of F_R in Figure 1b. The improvement, with respect to the CSEP version of the N and L tests, is clear: F_R is close to or below 0.05 for both tests. To clearly compare the CSEP methodology and the new testing procedures,

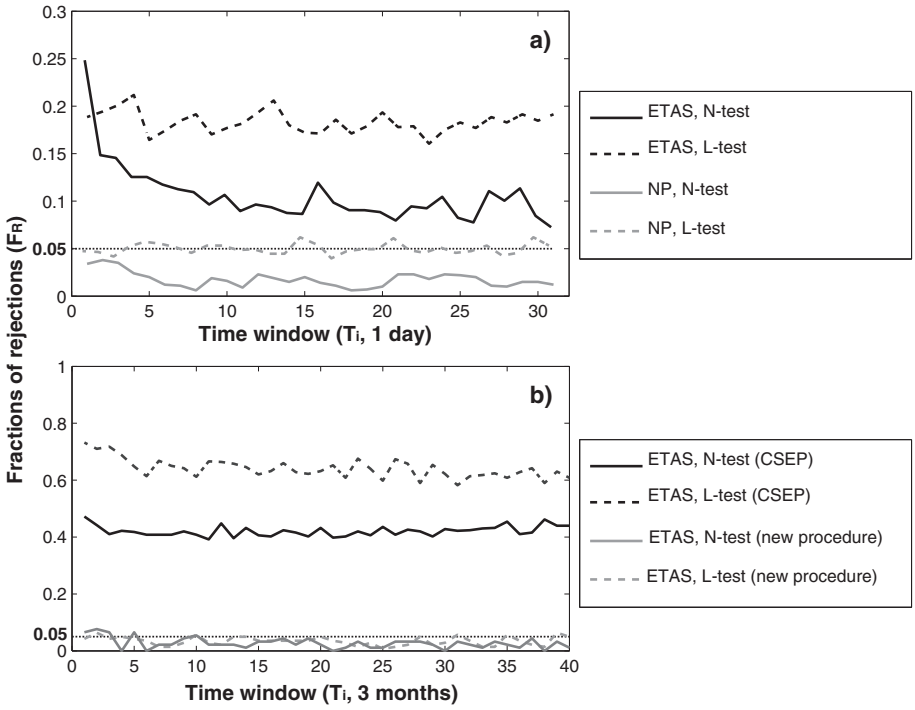

Figure 1 Fraction of rejections. Application of CSEP/RELM LBT and the proposed testing procedure on simulated catalogs. **(a)** F_R of daily CSEP *N* and *L* tests, for ETAS and NP simulations of Italian seismicity and $M_F = 2.5$. **(b)** Comparison of F_R values of testing procedure proposed here with those obtained by CSEP/RELM LBT, for ETAS simulations of Japanese seismicity ($M_F = 4.0$), with a forecast time span of 3 months.

Figure 2 shows the PDF of occurrences and log-likelihoods computed by the CSEP LBT and the proposed procedure for the first ETAS simulated Japanese catalog. The observed occurrences (solid black line, Figures 2a,b) are well above or below the confidence bounds (dashed black lines, Figure 2a) of the Poisson PDF (Equation 1) supposed by Hyp_2. This is because the distribution expected by the ETAS model (contour plot, Figure 2b), estimated by the empirical PDF of $\Omega^S_{\mathrm{ETAS},i}$, has a long/heavy tail, which is clearly not consistent with Hyp_2. Similar results are found for the log-likelihood. The log-likelihoods $L(\Omega_i|\Lambda)$ computed by Equation 4 are well below the values of $L(\Omega^S_i|\Lambda)$ expected by Hyp_2 (contour plot, Figure 2c). However, the log-likelihoods $\mathrm{CLF}_{\mathrm{ETAS},i}$ (Equation 8) are fully consistent with the log-likelihoods $\mathrm{CLF}^S_{\mathrm{ETAS},i}$ expected by the ETAS model (contour plot, Figure 2d).

Discussion

The rejection of the null hypothesis of a statistical test can be due to chance because it is really false or because it is probabilistically inadequate (Stark 1997; Luen and Stark 2008). The null hypothesis H_0 of the RELM/CSEP LBT supposes that $X_{(i,j)}$ are independent (in time and space) and Poisson random variables, with mean $\lambda_{(i,j)}$, given by the model. The CSEP protocol interprets the rejection of H_0 as the failure of the model being tested. However, this procedure is misleading because H_0 is not consistent with any model (Lombardi and Marzocchi 2010a).

The above findings may be explained with the help of stochastic point process theory (Daley and Vere-Jones 2003); this is the natural context in which stochastic earthquake models may be discussed. A point process is fully represented by its 'conditional intensity function' (CIF) $\lambda(t,\vec{x}/\mathcal{H}_t)$, i.e., the probability of observing an event in the instant $t \in \mathcal{T}$ and with additional variables (called marks) $\vec{x} \in \vec{\mathcal{X}}$, given the realization $\mathcal{H}_t$ of the process before t (Daley and Vere-Jones 2003). The CIF of the models described in the previous section are given by Equations 6 and 7; the marks are locations and magnitudes. In the case of an NP process, the CIF is a deterministic function of time and marks, but it is independent of the past history (i.e., $\lambda(t,\vec{x}/\mathcal{H}_t) = \lambda(t,\vec{x})$). Therefore, the events in nonoverlapping subsets of $\mathcal{T} \times \vec{\mathcal{X}}$ are independent and Poisson random variables (Daley and Vere-Jones 2003), as supposed by the RELM/CSEP LBT. In the most general case, the CIF is also a function of history $\mathcal{H}_t$, and the variables $X_{(i,j)}$ are not Poisson, unless the history is fully known (Meyer 1971; Papangelou 1972a; 1972b; Daley and Vere-Jones 2003). In a real-time forecast experiment, the history inside the forecast time window T_i is unknown; therefore, for such history-dependent models, such as ETAS, Hyp_2 is inadequate.

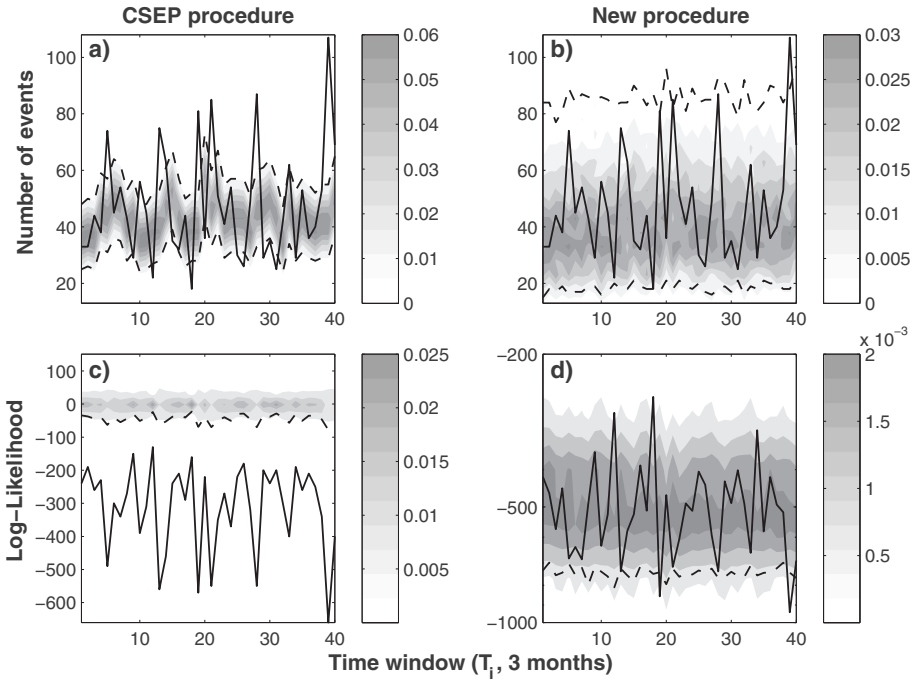

Figure 2 Distribution of the number of events and of likelihoods for ETAS simulations. Contour plot of probability density as a function of time interval T_i of the number of events and likelihood for the first ETAS Japanese simulated catalog. **(a)** Contour plot of probabilities q_n^i predicted by Poisson hypothesis Hyp_2. Solid black line marks the observed number of events. Dashed black lines mark the 2.5th and 97.5th percentiles of distribution. **(b)** The same as **(a)** but for the distribution expected by ETAS model. Specifically, dotted lines mark the values $P^{\Omega}_{\text{ETAS},i}[\,2.5\%]$ and $P^{\Omega}_{\text{ETAS},i}[\,97.5\%]$. **(c)** Contour plot of PDF of log-likelihoods $L(\Omega_i|\Lambda)$ predicted by Poisson hypothesis Hyp_2 (Equation 4). Solid black line marks the observed log-likelihoods. Dashed black lines mark the 5.0th percentile expected by Poisson distribution. **(d)** The same as (c) but for the CLF (see Equation 8). Solid black line marks the observed values $\text{CLF}_{\text{ETAS},i}$. Dotted black line marks the percentile $P^{\text{CLF}}_{\text{ETAS},i}[\,5.0\%]$.

The hypothesis Hyp_2 has been questioned in several studies (Schorlemmer et al. 2010a; Werner et al. 2010) and, in the specific context of ETAS modeling, by Lombardi and Marzocchi (2010a). Here, I examine the causes and effects of the failure of the LBT. Specifically, I show that the failure of the LBT may be significant for high values of M_F and that it has heavy consequences for long forecast time windows. This is because the longer the forecast time window T_i, the greater the randomness of forecasts (due to the effect of the unknown history inside T_i) and the lower the reliability of Hyp_2. This result contradicts the statement that the Poisson distribution is a good approximation of the forecast variability when M_F is large (Werner et al. 2010).

The process of revising the LBT has begun inside the scientific community. Some people have proposed replacing the Poisson distribution with a negative binomial distribution (Werner et al. 2010) to compute the p values of the tests. However, this solution does not significantly improve the LBT because the negative binomial distribution (as for the Poisson or any other distribution) is not consistent with all models. Inside the CSEP community, some suggest updating the forecasts more regularly, leaving the LBT unchanged (personal communication). I do not think this is the best way to resolve the inefficiencies of the LBT, as these do not derive from the regularity of the forecast calculations.

The procedure described above is an obvious upgrade of the N and L tests. It accounts for the actual variability of the $X_{(i,j)}$ given by the model being tested. Moreover, it uses the CLF, which is a better tool for checking the agreement between models and data than the discrete log-likelihood (Equation 4) used by the CSEP L test and based on Hyp_2 (Schorlemmer et al. 2007).

This study has focused on short-term forecasts, without analyzing the dependence of results on the size of the forecast window. From a theoretical point of view, LBT might also fail for long-term forecasts because of dissimilarities between the sets of probabilities $\{p_n^{ij}\}$ and $\{q_n^{ij}\}$ (see Equations 1 and 2) or, in other words, the unsuitability of Hyp_2. This study is not relevant to models that are explicitly supposed to be time-invariant, such as the models tested in the 5-year mainshock RELM experiment (Schorlemmer et al. 2010a; Zechar et al. 2013). However, the failure of the LBT might be significant for medium long-term forecast models with strong time-dependent

components, especially in testing regions with a high seismic rate. In other words, the present study does not invalidate most of the results of the first RELM/CSEP forecast experiments, which focus on long-term time-invariant models. However, the inclusion of different forecast time-spans and time-dependent models in new CSEP experiments requires both an urgent revision of the testing procedure and an effort by modelers to provide full distributions of the variables being tested.

Conclusions

The main goal of this study was to interpret the failures of the CSEP/RELM LBT and to propose a possible upgrade of the N and L tests. The main findings can be summarized as follows:

1. All LBTs are based on classical statistical hypothesis testing; therefore, they are intended to reject or not reject a null hypothesis H_0. The null hypothesis of the LBT is that the variables $X_{(i,j)}$ are independent and Poisson-distributed, with the rate given by forecasts. Therefore, the LBT is inadequate for checking the merits of a forecast model that is inconsistent with Hyp_2.
2. Specifically, Hyp_2 is not adequate for history-dependent models, such as ETAS, because the unknown history inside the forecast period means that $X_{(i,j)}$ do not follow a Poisson distribution.
3. In these cases, the LBT may fail for large values of M_F, especially for large forecast time windows, as the effect of the unknown history is greater.
4. I propose a revised version of the LBT that (1) adopts the CLF and (2) requires the percentiles of the distributions of X_i and $\mathrm{CLF}_{M,i}$.
5. The points discussed in this study highlight the need to revise the testing procedure for present and future experiments, which include many time-dependent models. However, they have a relative effect on the first RELM/CSEP experiments, mainly focused on long-term time-independent models.

Competing interests
The author declares that she has no competing interests.

Acknowledgements
The author is grateful to W. Marzocchi (INGV) for stimulating discussions on the topics presented in this paper. The suggestions made by D.D. Jackson (UCLA) and two anonymous referees have significantly improved the quality of the paper. The Italy earthquake data were obtained from the seismic bulletin of the Istituto Nazionale di Geofisica e Vulcanologia (INGV, http://iside.rm.ingv.it). The Japan earthquake data were extracted by the Earthquake Catalog of the Japan Meteorological Agency (JMA, http://www.jma.go.jp/en/quake). Information on CSEP is available at www.cseptesting.org.

Received: 15 October 2013 Accepted: 14 April 2014
Published: 1 May 2014

References

Daley DJ, Vere-Jones D (2003) An introduction to the theory of point processes. Second Ed, Vol. 1. Springer, New York, pp. 469

Jordan TH (2006) Earthquake predictability: Brick by brick. Seism Res, Lett 77(1): 3–6

Lombardi AM, Marzocchi W (2010a) Exploring the performances and usability of the CSEP suite of tests. Bull Seismol Soc Am 100: 2293–2300

Lombardi, AM, Marzocchi W (2010b) The ETAS model for daily forecasting of Italian seismicity in the CSEP experiment. Ann Geophys 53: 155–164

Luen B, Stark PB (2008) Testing earthquake predictions. IMS Lecture Notes Monograph Series. Probability and Statistics: Essays in Honor of David A. Freedman. Institute for Mathematical Statistics Press, Beachwood. 302-315

Meyer P (1971) Demonstration simplifiée d'un thèoréme de Knight In: Sèminaire de, Probabilitès V. Univ. Strasbourg, Lecture Notes in Math. vol 191, pp. 191–195

Nanjo KZ, Tsuruoka H, Hirata N, Jordan TH (2011) Overview of the first earthquake forecast testing experiment in Japan. Earth Planets Space 63(3): 159–169

Ogata Y (1998) Space-time point-process models for earthquake occurrences. Ann Inst Statist Math 50(2): 379–402

Papangelou F (1972a) Summary of some results on point and line processes, in Lewis P.A.W. Stochastic Point Processes. Wiley, New York. pp. 522–532

Papangelou F (1972b) Integrability of expected increments of point processes and a related random change of scale. Trans Amer Math Soc 165: 483–506

Schorlemmer D, Gerstenberger MC (2007) RELM Testing Center. Seismological Res, Lett 78(1): 30–36

Schorlemmer D, Gerstenberger MC, Wiemer S, Jackson DD, Rhoades DA (2007) Earthquake likelihood model testing. Seism Res Lett 78(1): 17–29

Schorlemmer D, Zecher JD, Werner MJ, Field EH, Jackson DD, Jordan TH (2010a) First results of the Regional Earthquake likelihood models experiment. Pure Appl Geophys 167: 859–876

Schorlemmer D, Christophersen A, Rovida A, Mele F, Stucchi M, Marzocchi W (2010b) Setting up an earthquake forecast experiment in Italy. Ann Geophys 53: 1–9

Stark PB (1997) Earthquake prediction: the null hypothesis. Geophys J Int 131: 495–499

Tsuruoka H, Hirata N, Schorlemmer D, Euchner F, Nanjo KZ, Jordan TH (2012) CSEP Testing Center and the first results of the earthquake forecast testing experiment in Japan. Earth Planets Space 64(8): 661–671

Werner MJ, Sornette D (2008) Magnitude uncertainties impact seismic rate estimates, forecasts, and predictability experiments. J Geophys Res 113: B08302. doi:10.1029/2007JB005427

Werner MJ, Zechar JD, Marzocchi W, Wiemer S (2010) Retrospective evaluation of the five-year and ten-year CSEP-Italy earthquake forecasts. Ann Geophys 53(3): 11–30. doi:10.4401/ag-4840

Zechar JD, Jordan TH (2008) Testing alarm-based earthquake predictions. Geophys J Int 172: 715-724. doi:10.1111/j.1365-246X.2007.03676.x

Zechar JD, Gerstenberger MC, Rhoades DA (2010) Likelihood-based tests for evaluating space-rate-magnitude earthquakes forecasts. Bull Seism, Soc Am 100(3): 1184–1195. doi:10.1785/0120090192

Zechar JD, Schorlemmer D, Werner MJ, Gerstenberger MC, Rhoades DA, Jordan TH (2013) Regional earthquake likelihood models I: first-order results. Bull Seism, Soc Am 103(2A): 787–798. doi.10.1785/0120120186

doi:10.1186/1880-5981-66-4
Cite this article as: Lombardi: **Some reasoning on the RELM-CSEP likelihood-based tests.** *Earth, Planets and Space* 2014 **66**:4.