The Assumption of Poisson Seismic-Rate Variability in CSEP/RELM Experiments

A.M. Lombardi & W. Marzocchi

Istituto Nazionale di Geofisica e Vulcanologia, Via di Vigna Murata 605, 00143 Rome, Italy.

Revised version, submitted to BSSA

2010

Abstract

Evaluating the performances of earthquake forecasting/prediction models is the main rationale behind some recent international efforts like the Regional Earthquake Likelihood Model (RELM) and the Collaboratory for the Study of Earthquake Predictability (CSEP). Basically, the evaluation process consists of two steps: 1) to run simultaneously all codes to forecast future seismicity in well-defined testing regions; 2) to compare the forecasts through a suite of statistical tests. The tests are based on the likelihood score and they check both the time and space performances. All these tests rely on some basic assumptions that have never been deeply discussed and analyzed. In particular, models are required to specify a rate in space-time-magnitude bins, and it is assumed that these rates are independent and characterized by Poisson uncertainty. In this work we have explored in detail these assumptions and their impact on CSEP testing procedures when applied to a widely used class of models, i.e., the Epidemic-Type Aftershock Sequence (ETAS) models. Our results show that, if an ETAS model is an accurate representation of seismicity, the same "right" model is rejected by the current CSEP testing procedures a number of times significantly higher than expected. We show that this deficiency is due to the fact that the ETAS models produce forecasts with a variability significantly higher than that of a Poisson process, invalidating one of the main assumption that stands behind the CSEP/RELM evaluation process. Certainly, this shortcoming does not negate the paramount importance of the CSEP experiments as a whole, but it does call for a specific revision of the testing procedures to allow a better understanding of the results of such experiments.

1. Introduction

The success of operational forecast indispensably depends on the use of reliable and skillful models (ICEF, 2009). In a nutshell, a model has to produce forecasts/predictions compatible with the future seismicity, and the forecasts/predictions have to be precise enough to be usable for practical purposes (i.e., they need a good skill). Moreover, if a set of reliable models is available, it is important to know what is the "best" one(s), i.e., the one(s) with the highest skill.

The evaluation of these pivotal features characterizing each forecasting/prediction model is the primary goal of the Collaboratory for the Study of Earthquake Predictability (CSEP hereinafter; Jordan 2006; http://www.cseptesting.org).

CSEP provides a rigorous framework for an empirical evaluation of any forecasting and prediction model. CSEP can be considered the successor of the Regional Earthquake Likelihood Model (RELM) experiment (Schorlemmer and Gerstenberger, 2007). While RELM was focusing on California, CSEP extends this focus to many other regions (New Zealand, Italy, Japan, North- and South-Western Pacific, and the whole World) as well as global testing centers (New Zealand, Europe, Japan). The coordinated international experiment has two main advantages: the evaluation process is supervised by an international scientific committee, not only by the modelers themselves, and the cross-evaluation of a model performances in different regions of the world can facilitate its evaluation in a much shorter period of time (see also Zechar et al., 2009).

All CSEP experiments performed in each testing region are truly prospective tests. In other words, each experiment compares forecasts produced by several models under testing with real data observed in the corresponding testing region after the forecasts have been produced. The forecasts are generated in the testing center independent of the modelers. The testing procedure adopted can be summarized in two subsequent steps: 1) to measure the *reliability* of each model; 2) to quantify the relative *skill* among the set of reliable models. In the first step, the forecasts/predictions made by each model are compared to the real seismicity through one or more goodness-of-fit tests. If the seismicity observed is compatible with the output of the model and the model-based variability, then the performance of the models can be contrasted with other models in the second step of the analysis. Specifically, the second step of the analysis compares quantitatively the forecasting/prediction capabilities of the models in order to establish a hierarchy of best performing models.

In this paper, we explore the performances of the CSEP/RELM testing procedure for two classes of forecasting models, Poisson and ETAS, that are largely represented in CSEP/RELM experiments (for the reliability of the prediction models see, e.g., Marzocchi et al., 2003; Zechar & Jordan, 2008, and references therein).

2. The CSEP/RELM suite of tests

The CSEP/RELM suite of tests is originally composed of three different tests (Schorlemmer et al., 2007; see also Kagan and Jackson 1994; 1995). The *L*-test (*Data-consistency test*) and *N*-test (*Number of events test*) are intended to check the goodness-of-fit of the model, while the *R*-test (*Hypotheses comparison*) compares the forecasting performances of different models.

The *L*-test and *R*-test are based on the well-known concept of conditional likelihood that is one of most used statistical tools to check and compare the performance of one or more models on data. The formulation of these tests requires the definition of bins that are specified intervals in space, magnitude and time. Using the same symbols of *Schorlemmer et al. (*2007), we define:

ω_i	number of earthquakes occurred in the *i*-th bin
λ_i^j	rate of earthquake occurrence for the *i*-th bin and *j*-th model.
$L_i^j = L(\omega_i \mid \lambda_i^j)$	log-likelihood calculated for the *i*-th bin and *j*-th model

The joint log-likelihood for the *j*-th model is calculated as

$$L^j = \sum_{i=1}^{n} L(\omega_i \mid \lambda_i^j) \qquad (1)$$

where *n* is the number of bins.

In order to get numbers from equation (1) $L(\omega_i / \lambda_i^j)$ must be defined. The basic assumption that stands behind the CSEP/RELM testing procedure is that earthquakes are assumed to occur in each bin according to a Poisson process with the rate specified by the model (Schorlemmer et al., 2007).

Note that this assumption is associated with the CSEP/RELM testing procedure not with the loglikelihood tests that can manage any kind of arbitrary distributions. Therefore, equation (1) becomes

$$L^j = \sum_{i=1}^{n} L(\omega_i \mid \lambda_i^j) = \sum_{i=1}^{n} \left(-\lambda_i^j + \omega_i \, ln \, \lambda_i^j - ln \, \omega_i! \right) \tag{2}$$

This assumption is crucial and a careful evaluation of its validity is mandatory to fully understand the CSEP/RELM tests. This assumption means that the bins are spatially and temporally independent, and the number of earthquakes in time has a variance equal to the average. Although some authors have already categorized such assumptions as "unlikely" and foresee possible inconsistencies of the tests (e.g., Werner and Sornette, 2008), the consequences have never been explored in detail. Moreover, we argue that the current use of this testing procedure in CSEP experiments may lead to think that the departures from this hypothesis could be considered as negligible.

The log-likelihood obtained by equation (2) is used to get the significance level of the tests through simulations. The *L*-test compares the observed log-likelihood value (see equation (2)) with a prefixed number of synthetic values obtained under the Poisson assumption for each bin, i.e., simulating records where each bin has a number of earthquakes generated according to a Poisson process with the rate given by the model. The quantile score γ^j for the *j*-th model is the fraction of simulated likelihood values that are less or equal to the observed *L*. This quantile score can be considered the p-value of the test. Note that, compared to the analyses performed by Schorlemmer et al. (2007) and Werner and Sornette (2008), here we do not consider the inclusion of *uncertainties*, because we aim to explore the tests in an optimal situation, i.e., with negligible uncertainty in the observations.

Schorlemmer et al. (2007) discussed the case in which a model can pass the *L*-test even if it is wrong. For this reason, the authors proposed a second test, the *N*-test, that checks if the total number of forecasted events is compatible with the observed number. In this case the quantile score,

δ^j, is the probability to have no more than the observed number of events by a Poisson process with a rate given by the model. In this case the test is two-sided, checking both possible over-prediction and under-prediction. To summarize, a model is "good" (*reliable*) if it is not rejected by both *L* and *N* tests. Only if the model passes these tests, then it is considered in the *R*-test, where it is compared to other reliable models. In the next section, we explore the performances of the L- and N-tests applied to synthetic catalogs. The goal is to check, in a controlled experiment, if the proportion of rejections of the "right" model is comparable to the significance level of the test. We anticipate that possible departures may point to inconsistencies of the Poisson variability for each bin assumed in the CSEP/RELM testing procedure.

3. Application of the CSEP/RELM testing procedure to synthetic catalogs.

In order to evaluate quantitatively the performances of CSEP/RELM testing procedure, we use these tests in a controlled experiment where we know exactly the model that generates earthquakes. The experiment can be described in three steps:

1. We generate 100 synthetic catalogues that we call "pseudo-real catalogs". Specifically we simulate two sets of 100 pseudo-real catalogs: one is consistent with a stationary non-homogeneous Poisson process, and another that is consistent with the well-known Epidemic-Type Aftershocks Sequence (ETAS; e.g., Ogata 1998) model. The generation of the ETAS pseudo-real catalogs is described in Appendix A and mimics the 1992 Landers sequence.

2. We generate one-day forecasts for a period of 10 days after the mainshock using exactly the same models and relative parameters that generate the pseudo-real catalogs. After each one-day forecast, the history is updated to take into account all events that occurred before the starting time of the next forecast. The forecasts are computed and evaluated in terms of expected number of events with magnitude above M_l 3.0 in each cell C_i of a grid, with a spacing of 0.1°x0.1° and covering the target

region [-117.5°W/33.25°N − -115.5°W/35.5°N]. Specifically for each cell C_i and for each time window T_j we compute the relative forecast rate λ_i^j by the formula

$$\lambda_i^j = \int_{T_j} \iint_{C_i} \int_{M_c}^{M_{max}} \lambda(t,x,y,m/H_t)dtdxdydm \qquad (3)$$

where $\lambda(t,x,y|\mathcal{H}_t)$ is the space-time conditional intensity defined by Poisson and ETAS models (see Appendix A), M_c and M_{max} are the minimum and maximum magnitude considered. The seismic history H_t, i.e. the information coming from the events that occurred before the time t, is crucial for time-dependent models, such as the ETAS model. On the other hand, the Poisson rate is independent of H_t and the time t. For the ETAS model we include in seismic history H_t the parameters of earthquakes that occurred before the time window T_j. To take into account the expected triggering effect of events that occurred during T_j, we simulate 1000 different stochastic realizations of the model inside the time window T_j and then we calculate for each bin the mean and the variance of predictions λ_i^j coming from each of these synthetic realizations.

3. We compare each one-day forecast with each pseudo-real catalog for both classes of models (Poisson and ETAS). For each of 100 pseudo-real catalogs we apply the N and L tests in order to verify the agreement between observations and forecasts. In this case, the model is certainly right; therefore we expect to see a number of rejections by both tests comparable to the significance level used.

In Figure 1 we show the fraction of rejections of both L (one tail test) and N-tests (two tails test) on 100 ETAS pseudo-real catalogs at significance level 0.05, for daily and cumulative tests, and for each time window T_j. The plots show that the proportions of rejections of N-test are above 30% (see Figure 1a), much larger than the theoretical fraction (i.e., 5%). Similar results are found for the L-test (see Figure 1b), computed on whole region, for which the fraction of rejections is above 20%. In order to verify the spatial distribution of L-test failures we show in Figure 2 the maps of quantile scores γ_j^i for each time window. The figure shows that the failures are mainly near Landers and Big Bear locations, where the number of events is larger and the spatial clustering is more evident.

The same analyses on Poissonian catalogues show that the fractions of rejections for both tests are in perfect agreement with the significance level (0.05) adopted (see Figure 3).

To explain one of the possible reasons for this discrepancy, we report in Figure 4, the ratio between mean and variance of the number of events recorded into 1000 synthetic ETAS catalogues, simulated following the same rules used for the 100 pseudo-real catalogs (see Appendix A). This ratio is much smaller than the unity, the value that characterizes the Poisson distribution (see Figure 4). This proves that the variability of the number of events is much larger than that expected in the case of a Poisson process. By performing a Chi-squared test, the Poisson distribution is rejected for all time-windows at a significance level of 0.01, and this independent of how the data are regrouped to compare expected and observed distributions.

To quantify the differences between the variability of the seismic rate due to Poisson and ETAS distributions, we plot in Figure 5 the differences of their 95% confidence bounds. Specifically, for each pseudo-real ETAS catalog and for each day, we compute the variability of the seismic rate $\Delta\alpha_{95\%}^{POISSON}$ expected by the Poisson distribution and assumed by CSEP tests; this value is compared with the empirical variability $\Delta\alpha_{95\%}^{ETAS}$ of the ETAS distribution that has been calculated numerically by the 1000 synthetic rates used for producing forecasts. Figure 5 shows the average of the differences $\Delta\alpha_{95\%}^{ETAS} - \Delta\alpha_{95\%}^{POISSON}$ calculated for 100 pseudo-real catalogs. The positive differences mean that the variability for the ETAS model is much larger than the variability of the Poisson distribution. Interestingly, this difference decreases with time, implying that this difference becomes less serious when the seismic rate tends to decrease.

3. Discussion and conclusions.

In this paper we show that part of the CSEP/RELM testing procedure does not perform correctly for a widely used class of models, i.e., the ETAS models. Specifically, by reproducing the CSEP experiment on “pseudo-real” ETAS catalogs – for which we know the right model – we find that the rejections are much more than expected. We identify one main reason for this deficiency: the assumption that the number of earthquakes per bin has a Poisson distribution does not hold for ETAS models. The latter have a variability of occurrences much larger than what predicted by a Poisson distribution. The underestimation of the variability made under the Poisson hypothesis unavoidably leads to a high rejection frequency during the CSEP experiments, at least for the ETAS class of models. It is worth noting that a higher variability compared to what assumed by the Poisson hypothesis is also observed on real catalogs (e.g., Saichev and Sornette, 2007; Kagan, 2009 and references therein) possibly (but not necessarily) leading to a wide generalization of the conclusions reported in this paper (see also Schorlemmer et al., 2010). These results may be generalized in this way: forecasting models that produce a higher variability of the seismic rates compared to the Poisson process may be rejected too often also when they represent an accurate representation of the observed spatio-temporal evolution of the seismicity. On the other hand, we also foresee that forecasting models producing a variability of the seismic rates smaller than that expected in the case of a Poisson process may be not rejected often enough even in case they do not represent an accurate representation of the seismicity. Figure 2 shows also another interesting departure from the Poisson distribution. Rejected bins appear clustered in space. The Poisson distribution assumes that the seismic variability per bin is conditioned only by the seismic rate of the model. Actually, the observed rate in a bin is also conditioned by the seismicity occurred in the adjacent bins during the forecasting time window; this component is neglected in the testing phase and it may play an important role on the results of the L-test.

In this paper we have investigated a strongly clustered sequence (pseudo real catalogs mimicking an aftershock sequence) that is characterized by bins with a large number of events. In other cases, such as the one-day forecasts during a quiet period or the forecast of large events ($M \geq 5.0$) in a 5-year time period, the expected number of events is probably much smaller. In these cases, the bias may be less serious as showed by Figure 1 (cf. the rejection rates for M3+ and M4+ events) and

Figure 5, and also as expected by the theory of hypothesis testing (basically, the fewer the data, the more difficult is to reject an hypothesis).

Although these results indicate a bias of the current testing procedures of the CSEP experiments, we stress that these experiment remain of paramount importance and they are unavoidable if we wish to maintain earthquake forecasting in a scientific domain that requires formulation of hypothesis and testing. The lesson to be learned is that some of the CSEP/RELM testing procedures should be improved and/or implemented. Specifically, in order to get reliable results, we argue that the CSEP/RELM suite of tests needs a significant revision. We identify three possible strategies that could be implemented for current and future experiments:

1. Each forecasting model has to provide the likelihood function. This allows the likelihood tests to be applied correctly because the Poisson assumption of the seismic rate variability is no longer necessary, and other goodness-of-fit tests and skill measures may be applied, like the residuals analysis (Ogata 1998; Marzocchi and Lombardi, 2009) and the Information Gain (e.g., Daley and Vere-Jones, 2003). Notably, this approach would also avoid potential biases in the testing phase due to the spatial correlation of the rejected bins (see figure 2). This is maybe the optimal choice from a statistical point of view, but it is not applicable to models that do not have a likelihood function, such as many pattern recognition algorithms.

2. The forecasts have to be described by a distribution of the expected number of earthquakes (see also Werner and Sornette 2008), not by a single value as now. For example, the forecasts may be composed by 1000 expected number of events, from which a central value and the dispersion can be easily retrieved (see Marzocchi and Lombardi, 2009). This strategy is in principle applicable to every model, but it would require a change in the CSEP procedures. In our mind, this option is probably the easiest to implement for future experiments, but it is inapplicable to the present forecasts that are composed just by one single expected number of earthquakes. Moreover, being still based on binning forecasts, we remark that this strategy would not avoid possible biases induced by the spatial correlation of the bins; a careful analysis of such potential bias is required.

3. The model-based variability of the number of earthquakes in each bin may be set by some empirical rules that take into account the higher variability that characterize many models. This is widely applicable for all models and all experiments so far completed or running, but certainly it raises important technical problems. The first one is the introduction of a key parameter (i.e. the dispersion) *after* the forecasts have been made. This would corrupt the prospective philosophy of the experiments. Second, the choice of the empirical adjustment rule becomes critical for the evaluation process. Unavoidably, this choice would raise a lot of debate about what is the best adjustment rule, and if different rules should be applied to different models. In any case, it may be difficult to establish these rules objectively and independently from the modelers.

Data and Resources

The Landers earthquake data were obtained from Southern California Earthquake Data Center, website (http://data.scec.org/research/altcatalogs.html). The maps were made using the Generic Mapping Tools (www.soest.hawaii.edu/gmt). The MATLAB GNU codes used in the present work to run the N and L tests have been provided by the Southern California Earthquake Center CSEP software development team.

Acknowledgments

The researches that motivated this study have been promoted in the framework of EC project Network of Research Infrastructures for European Seismology, NERIES, www.neries-eu.org. A.M. Lombardi has been supported by the EC project NERIES contract 026130. We thank Jeremy Zechar and Max Werner for useful discussions. We thank Jochen Woessner for advising on the CSEP tests.

References

Daley D.J. and D. Vere-Jones (2003), *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York, 2-nd ed., Vol. 1, pp. 469.

ICEF (International Commission on Eathquake Forecasting for Civil Protection) (2009). Operational Earthquake Forecasting: State of Knowledge and Guidelines for Utilization. Report for Italian Civil Protection .

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

Kagan Y.Y. (2009). Statistical distributions of earthquake numbers: consequence of branching process, *Geophys. J. Int*., submitted.

Kagan Y.Y., D.D. Jackson (1994). Long-term probabilistic forecasting of earthquakes, J. Geophys. Res., 99, 13685-13700.

Kagan Y. Y., D. D. Jackson (1995). New seismic gap hypothesis: Five years after. J. Geophys. Res., 100, 3943-3959.

Marzocchi W., L. Sandri, E. Boschi (2003). On the validation of earthquake-forecasting models: the case of pattern recognition algorithms. *Bull. Seismol. Soc. Am*., 93, 1994-2004.

Marzocchi W., A.M. Lombardi (2009). Real-time forecasting following a damaging earthquake. *Geophys. Res. Lett*., 36, L21302, doi:10.1029/2009GL040233.

Ogata, Y. (1998). Space-Time Point-Process Models for Earthquake Occurrences, *Ann. Inst. Statist. Math.* 50(2), 379-402.

Saichev A., D. Sornette (2007). Power law distribution of seismic rates, *Tectonophysics,* 431*,* 7-13

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

Schorlemmer, D., M.C. Gerstenberger, S. Wiemer, D.D. Jackson and D.A. Rhoades (2007). Earthquake Likelihood Model Testing, *Seismol.Res.Lett.,* 78(1)*,* 17-29.

Schorlemmer D., J.D. Zechar, M.J. Werner, E.H. Field, D.D. Jackson, T.H. Jordan (2010). First results of the Regional Earthquake Likelihood Models experiment, *Pure Applied Geophys*., in press.

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

Zechar J.D., T.H. Jordan (2008), Testing alarm-based earthquake predictions, *Geophys. J. Int.*, 172, 715-724.

Zechar J.D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P.J. Macheling, and T.H. Jordan (2009), The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science, Concurrency and Computation: Practice and Experience, doi: 10.1002/cpe.1519

Figure Captions

Figure 1: Fractions of rejections of the daily *L* and *N*-test on 100 pseudo-real ETAS catalogs.

Figure 2: Spatial distribution of fractions of rejections on 100 pseudo-real ETAS catalogs for *L*-test conducted on 10 time windows.

Figure 3: The same of Figure 1 but for pseudo-real Poisson catalogs.

Figure 4: Ratio between mean and variance of events recorded in 1000 ETAS pseudo-real catalogs for 10 time windows.

Figure 5: Difference between the 95% confidence intervals of the ETAS and Poisson distributions as a function of the forecasting time window; each point represents the average of the differences calculated for the 100 ETAS pseudo-real catalogs used for *L* and *N*-tests.

APPENDIX A. Generating the pseudo-real synthetic catalogs

In this appendix, we report the strategy adopted to generate ETAS and Poissonian pseudo-real catalogs.

The total space-time conditional intensity $\lambda(t,x,y/\mathcal{H}_t)$ of the ETAS model (i.e. the probability of an earthquake occurring in the infinitesimal space-time volume conditioned to all past history) is defined by equation:

$$\lambda(t,x,y,m/\mathcal{H}_t) = \left[\nu u(x,y) + \sum_{t_i<t} \frac{K}{(t - t_i + c)^p} e^{\alpha(M_i - M_c)} \frac{c^i_{d,q,\gamma}}{\left[r_i^2 + \left(d e^{\gamma(M_i - M_c)} \right)^2 \right]^q} \right] \beta e^{\beta(m - M_c)} \qquad (A1)$$

where $\mathcal{H}_t = \{(t_i,x_i,y_i,M_i);\ t_i < t\}$ is the observation history up time t, M_c is the completeness magnitude of the catalog, $u(x,y)$ is the spatial probability density function (PDF) of background events, $c^i_{d,q,\gamma} = \frac{q-1}{\pi}\,[(d e^{\gamma(M_i - M_c)})^2\,]^{q-1}$ is the normalization constant of the spatial PDF for triggered events, and r_i is the distance between location (x,y) and the epicenter of i-th event (x_i,y_i) (Lombardi et al., 2009). Finally $\beta = b \cdot ln(10)$ is the parameter of the well-known Gutenberg-Richer Law (Gutenberg and Richter, 1954), assumed as distribution for magnitude of all events.

The set of parameters $\Theta=(\nu,K,c,p,\alpha,d,q,\gamma,\beta)$ of the model, for the events occurred within a time interval $[T_1,T_2]$ and a region R, can be estimated by maximizing the log-likelihood function (Daley and Vere-Jones, 2003), given by

$$logL(\Theta) = \sum_{i=1}^{N} log\,\lambda(t_i,x_i,y_i,m_i / H_{t_i}) - \int_{T_1}^{T_2}\int_R \int_{M_c}^{M_{max}} \lambda(t,x,y,m / H_t)\,dtdxdydm \qquad (A2)$$

A careful method to obtain the best parameters of the model is the iteration algorithm developed by Zhuang et al. (2002), providing also an estimation of the PDF $u(x,y)$ for background events.

Our pseudo-real ETAS catalogs are simulated in agreement with the ETAS model estimated for the region hit by the Landers earthquake. Specifically we use the relocated data set (Hauksson and Shearer, 2005) recorded by the California Institute of Technology/U.S. Geological Survey (CIT / USGS) Southern California Seismic Network and available at the SCEDC (Southern California Earthquake Data Center) website (http://data.scec.org/research/altcatalogs.html). We consider earthquakes with a depth less than 30 km and a magnitude above 3.0, occurred from Jan 1 1984 to Dec 31 2004 and located in the region [-119.0°W/32.5°N − -115.0°W/36.5°N] (5757 events). The parameters estimated by using the procedure proposed by Zhuang et al. (2002) are listed in Table

A1. We perform simulations by including in the past history the real observed seismicity above magnitude 3.0, occurred before July 1 1992, 3 days after the $M_L 7.3$ Landers mainshock. In this way we take into account knowledge coming from the initial phase of the sequence, including also the $M_L 6.4$ Big Bear aftershock.

We simulate the Poisson pseudo-real catalogs by imposing a rate of 60 day^{-1} and adopting the PDF $u(x,y)$, estimated for the ETAS model, for the spatial distribution of events. All pseudo-real catalogs recover a time period of 10 days. We remark that we intend to perform simulations by reproducing the type of forecasts usually tested in CSEP laboratories, no matter the specific region or time period we consider.

In order to verify the reliability of our pseudo-real catalogs, we analyze their residuals. The residual analysis is a common diagnostic technique for stochastic point processes based on transformation of the time axis t into a new scale τ by the increasing function

$$\tau = \Lambda(t) = \int_{T_{start}}^{t} dt \int_{R} dxdy \int_{M_c}^{M_{max}} dm\, \lambda(t,x,y,m/\mathcal{H}_t) \qquad (A3)$$

where T_{start} is the starting time of the observation history H_t (Ogata, 1998). The random variable τ represents the expected number of occurrences in time period $[T_{start}, t]$. If a model with conditional intensity $\lambda(t,x,y,m/\mathcal{H}_t)$ describes the temporal evolution of the process, the transformed data $\tau_i = \Lambda(t_i)$, known in statistical seismology with the name of *residuals*, are expected to behave like a stationary Poisson process with the unit rate (Ogata, 1998); i.e. the values $\Delta\tau_i = \tau_{i+1}-\tau_i$ are independent and exponentially distributed (with mean equal to 1) random variables. We check this hypothesis for residuals by means of two nonparametric tests: the Runs test, to verify the reliability of the independence property, and the one-sample Kolmogorov-Smirnov (KS1) test, to check the standard exponential distribution (Gibbons and Chakraborti, 2003; Lombardi and Marzocchi, 2007). Specifically the Runs-test can be used to test if a process is not auto-correlated and consists in testing the randomness of runs, i.e. of uninterrupted subsequences of values above or below the mean (see Gibbons and Chakraborti, 2003; Lombardi and Marzocchi, 2007 for details). We use both tests because all goodness-of-fit tests (as KS1) are ineffective to check the presence of a memory in the time series. Hence, any discrepancy of residuals by Poisson hypothesis, identified by just one or both tests, is a sign of inadequacy of ETAS model to explain all basic features of analyzed seismicity. We stress that this check analysis is similar to the RELM/CSEP *N*-test. As the *N*-test, it consists in a comparison between the observed and the expected total number of events and it is directed to highlight under or over-prediction. On the other side the residual analysis does not need the discretization of the temporal scale in time bins. As explained along the text, this is a crucial point of RELM/CSEP tests. In Figure A1 we show the empirical cumulative function of p-values of KS1 and Runs tests, for the 100 pseudo-real ETAS catalogs, together with the 99% confidence bounds. The confidence level is calculated assuming that for each point of the curve the expected fraction of rejection is given by the p-value reported on the x-axis, and the variability (1 sigma) is given by $\sqrt{p(1-p)/N}$. Note that, for both tests the cumulative distribution is inside the 99% confidence interval.

References

Daley, D.J. and D. Vere-Jones (2003). *An Introduction to the Theory of Point Processes*, Springer-Verlag, New York, 2-nd ed., Vol. 1, pp. 469.

Gibbons, J.D. and S. Chakraborti (2003). Non-parametric Statistical Inference, 4th ed., rev. and expanded, New York: Marcel Dekker, 645 pp.

Gutenberg, B. and C.F. Richter (1954). *Seismicity of the Earth and Associated Phenomena*, Princeton, pp. 273.

Hauksson, E. and P. Shearer (2005). Southern California Hypocenter Relocation withWaveform Cross-Correlation, Part 1: Results Using the Double-Difference Method, *Bull. Seismol. Soc. Am.*, **95 (3)**, 896–903, doi:10.1785/0120040167.

Lombardi, A. M., and W. Marzocchi (2007). Evidence of clustering and nonstationarity in the time distribution of large worldwide earthquakes, *J. Geophys. Res.*, **112**, B02303, doi:10.1029/2006JB004568.

Lombardi, A.M., M. Cocco and W. Marzocchi (2009). On the increase of background seismicity rate during the 1997-1998 Umbria-Marche (central Italy) sequence: apparent variation or fluid-driven triggering? Submitted to *Bull. Seismol. Soc. Am.*

Ogata, Y. (1998). Space-Time Point-Process Models for Earthquake Occurrences, *Ann. Inst. Statist. Math.* **50(2)**, 379-402.

Zhuang, J., Y. Ogata and D. Vere-Jones (2002). Stochastic declustering of space-time earthquake occurrence, *J. Am. Stat. Assoc.,* **97**, 369-380.

Figure Captions

Figure A1: Cumulative function of the empirical p-values (solid black lines) for KS1 (panel a) and RUNS (panel b) Test applied to Residuals of 100 simulated ETAS catalogues. Dashed gray lines mark the 99% confidence bounds.

Parameter	Value
ν	0.10 ± 0.004 (day^{-1})
K	0.043 ± 0.002 (day^{p-1})
p	1.20 ± 0.01
c	0.030 ± 0.004 (day)
α	1.20 ± 0.03 (mag^{-1})
d	0.30 ± 0.01 (km)
q	$\equiv 1.5$
γ	0.60 ± 0.03 (mag^{-1})
Log-likelihood	-21277.5

TableA1: Maximum Likelihood parameters (with relative errors) and log-likelihood of ETAS model for Landers region seismicity [-119.0° W/32.5° N − -115.0° W/36.5° N] (M_c = 3.0; Jan 1 1984 – Dec 31 2004; 5757 events)

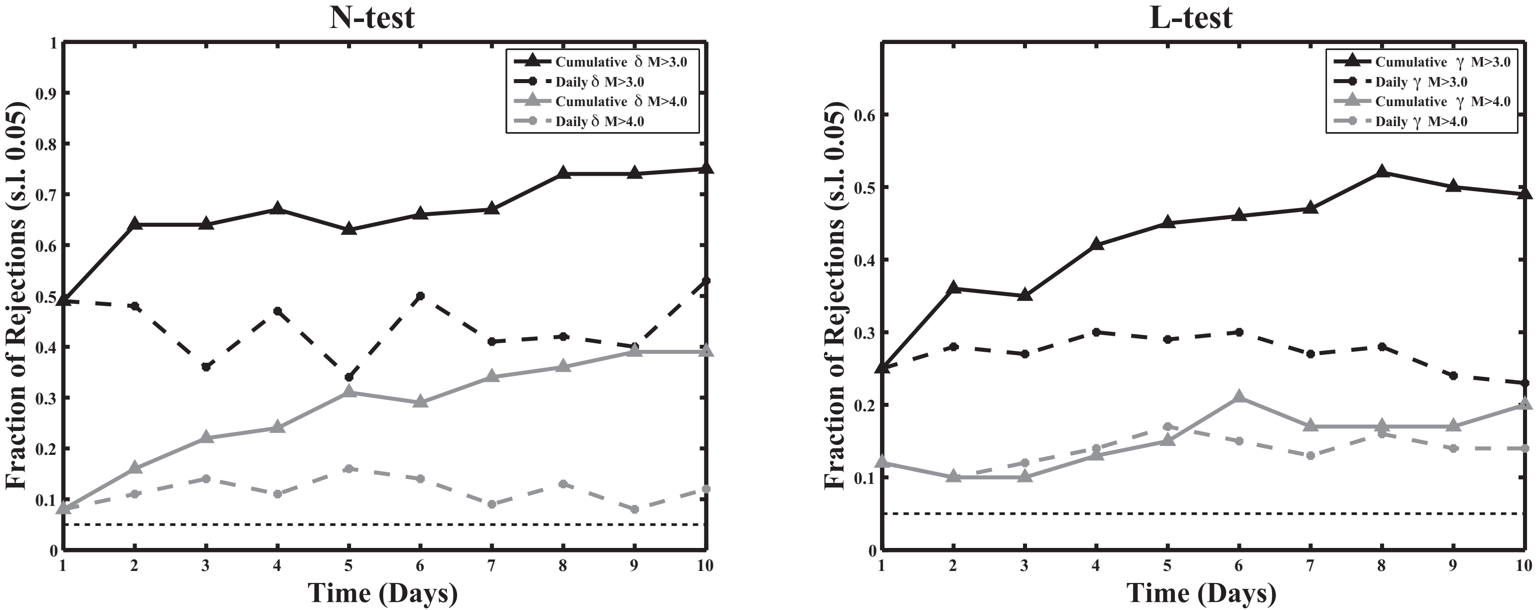

N-test
Cumulative δ M>3.0
Daily δ M>3.0
Cumulative δ M>4.0
Daily δ M>4.0
Fraction of Rejections (s.l. 0.05)
Time (Days)
L-test
Cumulative γ M>3.0
Daily γ M>3.0
Cumulative γ M>4.0
Daily γ M>4.0
Fraction of Rejections (s.l. 0.05)
Time (Days)

Figure 1

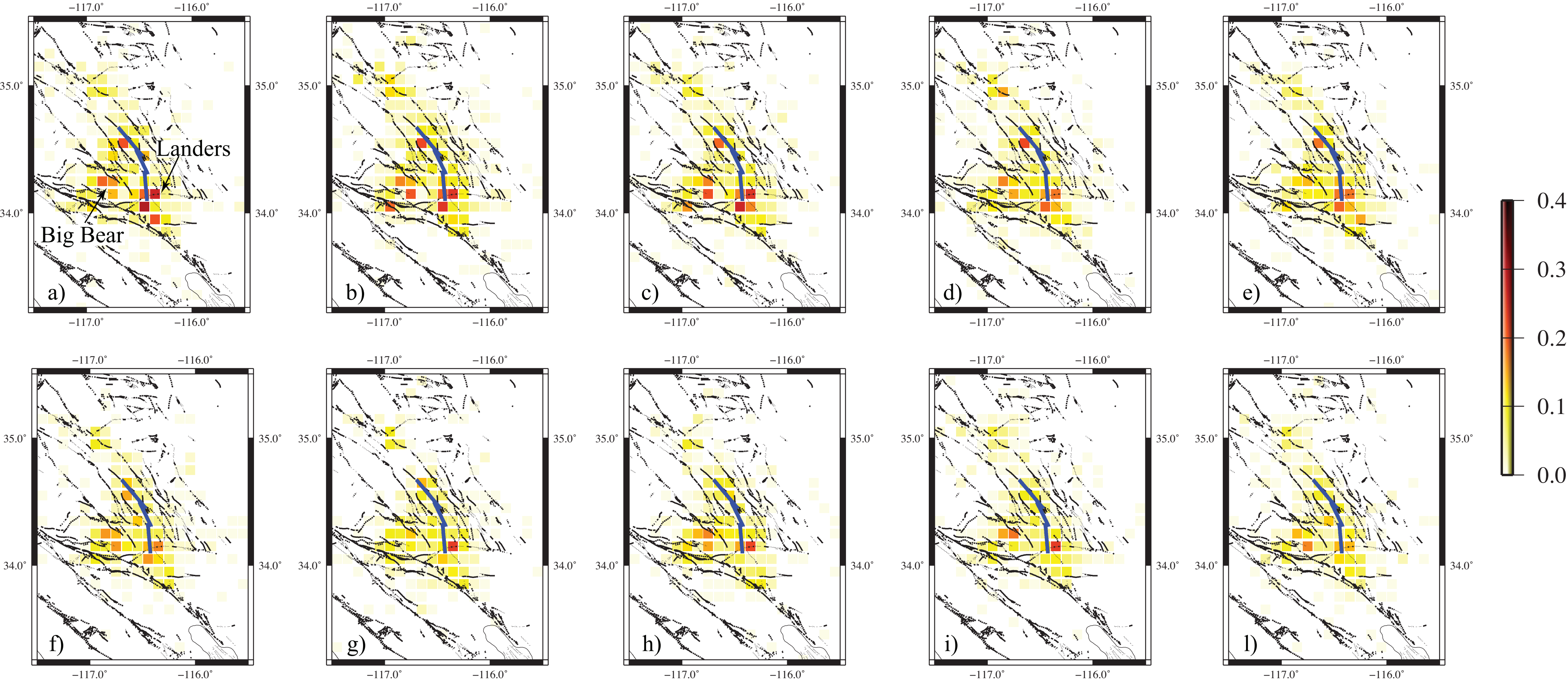


Figure 2

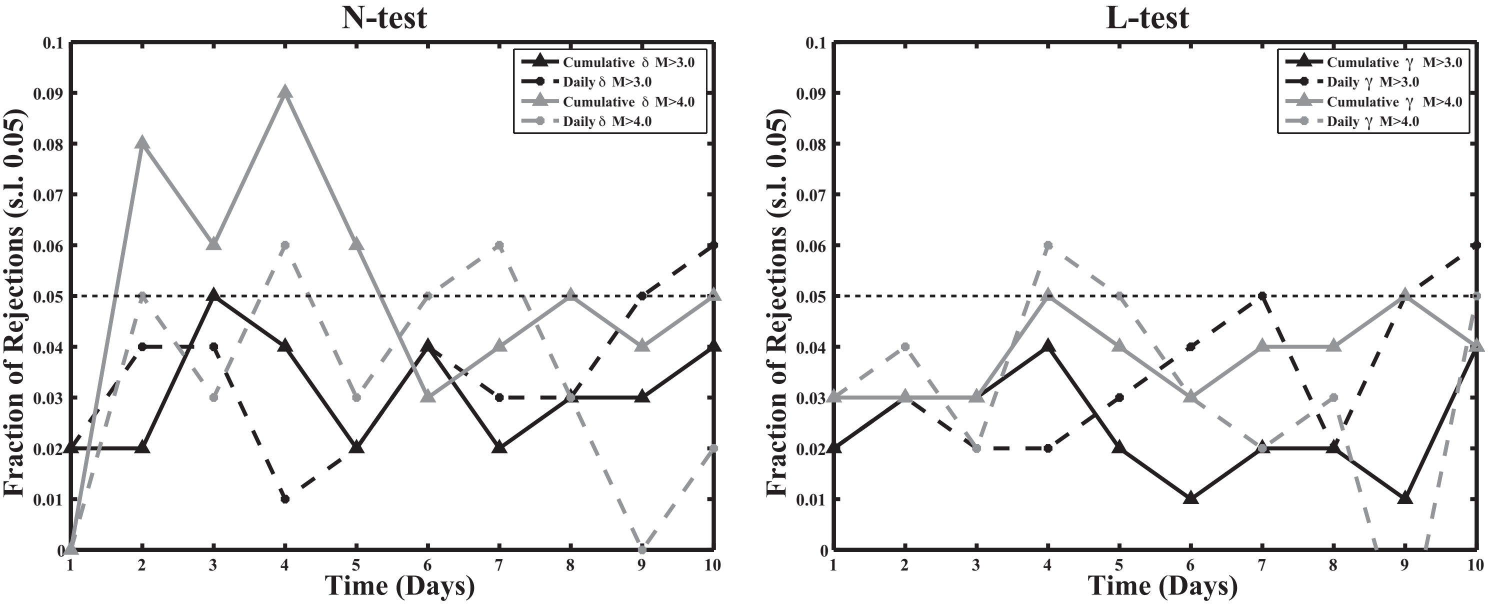

N-test
L-test
Cumulative δ M>3.0
Daily δ M>3.0
Cumulative δ M>4.0
Daily δ M>4.0
Cumulative γ M>3.0
Daily γ M>3.0
Cumulative γ M>4.0
Daily γ M>4.0
Fraction of Rejections (s.l. 0.05)
Time (Days)
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
1
2
3
4
5
6
7
8
9
10

Figure 3

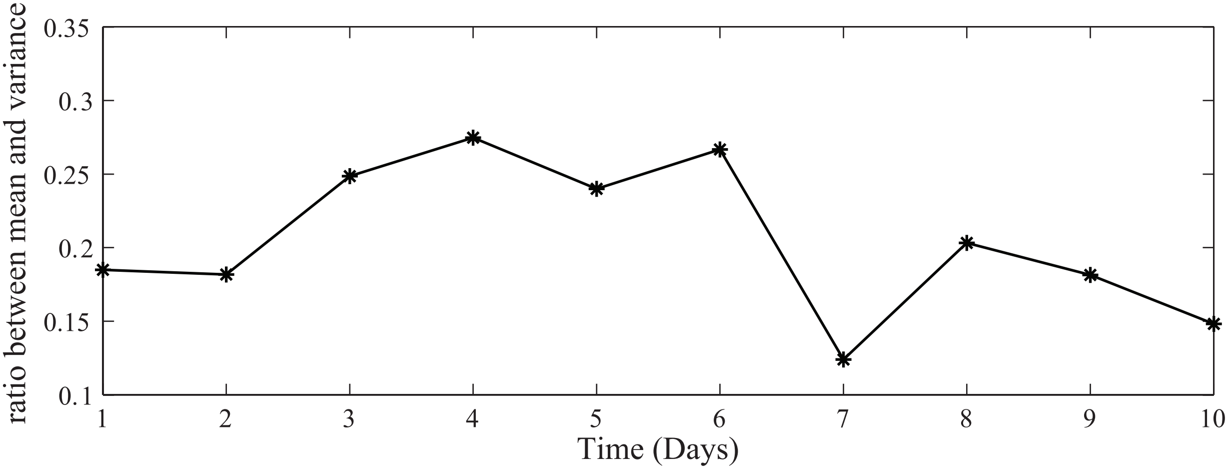
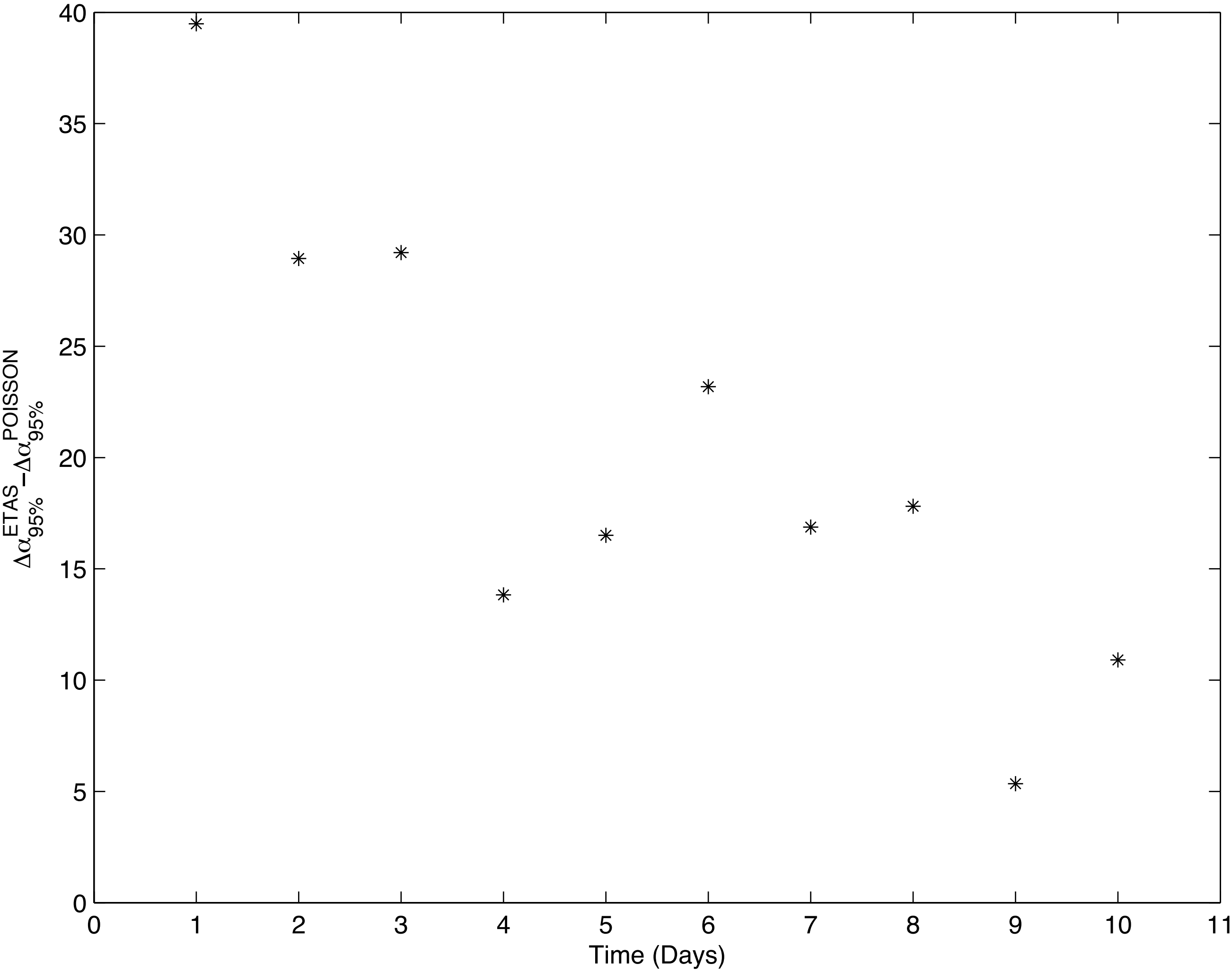

ratio between mean and variance
0.35
0.3
0.25
0.2
0.15
0.1
1
2
3
4
5
6
7
8
9
10
Time (Days)

Figure 4



$\Delta\alpha_{95\%}^{ETAS}-\Delta\alpha_{95\%}^{POISSON}$
Time (Days)
0
1
2
3
4
5
6
7
8
9
10
11
5
10
15
20
25
30
35
40

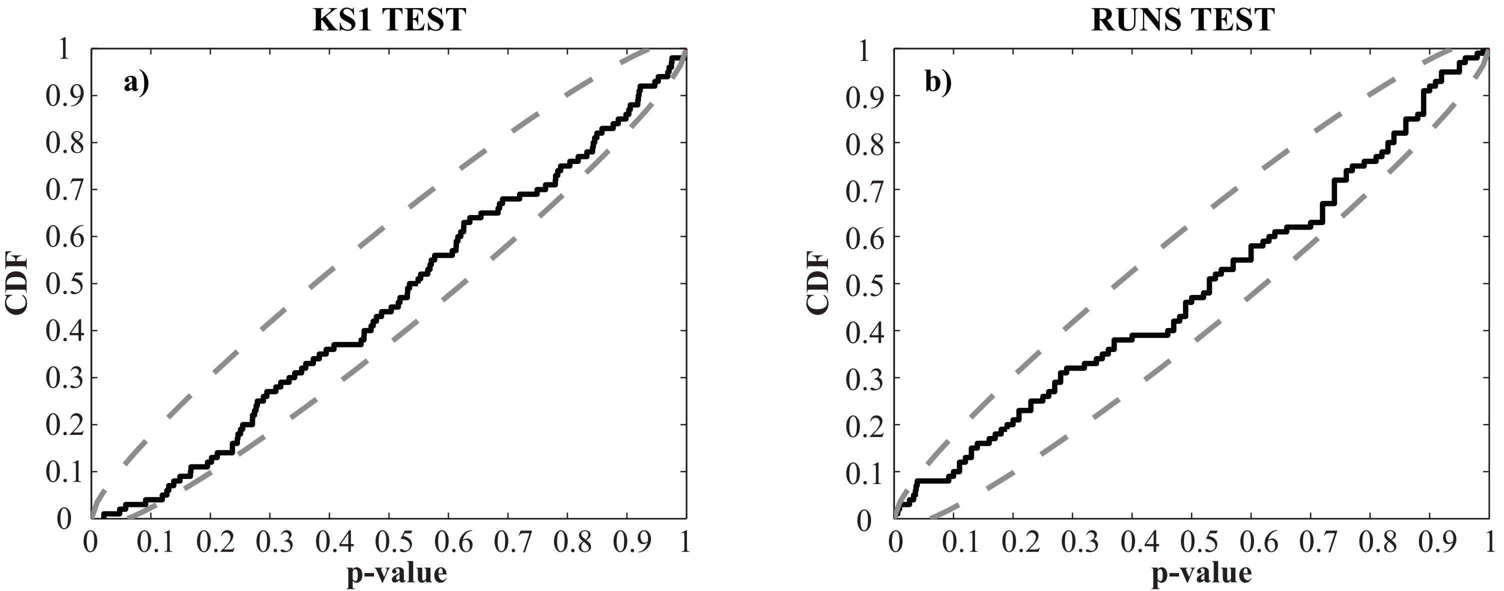

KS1 TEST
a)
CDF
p-value
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
RUNS TEST
b)
CDF
p-value

Figure A1