Earthquake Likelihood Model Testing

Article *in* Seismological Research Letters · January 2007

DOI: 10.1785/gssrl.78.1.17

CITATIONS
283

READS
676

5 authors, including:

Stefan Wiemer
ETH Zurich
413 PUBLICATIONS **18,798** CITATIONS
SEE PROFILE

David D. Jackson
University of California, Los Angeles
184 PUBLICATIONS **12,047** CITATIONS
SEE PROFILE

D. A. Rhoades
GNS Science
148 PUBLICATIONS **4,760** CITATIONS
SEE PROFILE

Earthquake Likelihood Model Testing

D. Schorlemmer[1], M. Gerstenberger[2], S. Wiemer[1], and D. Jackson[3]

[1] Swiss Seismological Service, ETH Zürich, Schafmattstr. 30, 8093 Zürich, Switzerland.

[2] United States Geological Survery, 525 S. Wilson Ave., Pasadena, CA 91106, USA. Now at: Institute of Geological and Nuclear Sciences, 69 Gracefield Road, Lower Hutt, New Zealand.

[3] University of California Los Angeles, Dept. Earth and Space Sciences, 595 Young Drive East, Los Angeles, CA 90095-1567, USA.

Abstract

The Regional Earthquake Likelihood Models (RELM) project aims to produce and evaluate alternate models of earthquake potential (probability per unit volume, magnitude, and time) for California. Based on differing assumptions, these models are produced to test the validity of their assumptions and to explore which models should be incorporated in seismic hazard and risk evaluation. Tests based on physical and geological criteria are useful but we focus on statistical methods using future earthquake catalog data only. We envision two evaluations: a test of consistency with observed data, and a comparison of all pairs of models for relative consistency. Both tests are based on the likelihood method, and both are fully prospective (i. e., the models are not adjusted to fit the test data). To be tested, each model must assign a probability to any possible event within a specified region of space, time, and magnitude. For our tests the models must use a common format: earthquake rates in specified "bins" with location, magnitude, time, and focal mechanism limits.

Introduction

Seismology cannot yet deterministically predict earthquake occurrence; however, it should seek for the best possible models for forecasting earthquake occurrence. This paper describes the statistical rules of an experiment to examine and test earthquake forecasts. The primary purposes of the tests described below are to evaluate physical models for earthquakes, assure that source models used in seismic hazard and risk studies are consistent with earthquake data, and provide quantitative measures by which models can be assigned weights in a consensus model or be judged as suitable for particular regions.

In this paper, we develop a statistical method used for testing earthquake likelihood models. A companion paper [*Schorlemmer and Gerstenberger*, in this volume] discusses the actual implementation of these tests in the framework of the Regional Earthquake Likelihood Model (RELM) initiative.

Statistical testing of hypotheses is a common task and a wide range of possible testing procedures exist. *Jolliffe and Stephenson* [2003] present different forecast verifications from atmospheric science, among them, likelihood testing of probability forecasts and testing the occurrence of binary events. Testing binary events requires that for each forecasted event, the spatial, temporal and magnitude limits be given. Although major earthquakes can be considered binary events, the models within the RELM project express their forecast on a spatial grid and in 0.1 magnitude units, thus a distribution of rates over space and magnitude. These forecasts can be tested with likelihood tests.

In general, likelihood tests assume a valid null hypothesis against which a test hypothesis is tested. The outcome is either a rejection of the null hypothesis in favor of the test hypothesis or a non-rejection, meaning the test hypothesis cannot outperform the null hypothesis at the given significance level. Within RELM, there is no accepted null hypothesis and thus the likelihood test needs to be expanded such that it allows comparable testing of equipollent hypotheses.

To test models against one another, we require that forecasts are expressed in a standard format: the average rate of earthquake occurrence within prespecified limits of hypocentral latitude, longitude, depth, magnitude, time period, and focal mechanisms. Focal mechanisms should either be described as the inclination of P-axis, declination of P-axis, and inclination of the T-axis, or as strike, dip, and rake angles. In [*Schorlemmer and Gerstenberger*, in this volume], we design classes of these parameters such that similar models will be tested against each other. These classes make the forecasts comparable between models. Additionally, we are limited to testing only what is precisely defined and consistently reported in earthquake catalogs. Therefore it is currently not possible to test such information as fault rupture length or area, asperity location, etc. Also, to account for data quality issues, we allow for location and magnitude uncertainties as well as the probability that an event is dependent on another event.

As we mentioned above, only models with comparable forecasts can be tested against each other. Our current tests are designed to examine grid-based models. This requires that any fault-based model be adapted to a grid before testing is possible. While this is a limitation of the testing, it is an inherent difficulty in any such comparative testing.

The testing suite we present, consists of three different tests: L-Test, N-Test, and R-Test. These tests are defined similarily to *Kagan and Jackson* [1995]. The first two tests examine the consistency of the hypotheses with the observations while the last test compares the spatial performances of the models.

Basic Ideas and Definitions

We refer to a model as a concept of earthquake occurrence, composed of theories, assumptions, and data. Models can be rather general and need not be testable in a practical sense. A hypothesis is a more formal, testable statement derived from a model. The hypothesis should follow directly from 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.

For tests described here, we treat earthquakes as point sources with eight parameters: hypocentral latitude and longitude, depth, magnitude, origin time, focal mechanism. Depending on the models, not all parameters are required.

Of course, earthquakes are too complex to be fully described as point sources with eight parameters. Some earthquake models, especially those based on active faults, describe likely rupture length, area, end points, asperity location, etc. However, at this stage we use only the eight hypocentral parameters and their full error distribution because other qualities are not precisely defined nor consistently reported in earthquake catalogs. Adopting the eight parameter description means that fault-based models must be adapted to express grid-based probable hypocentral locations, magnitudes, and focal mechanisms.

A forecast is defined as a vector of earthquake rates corresponding to all specified bins. Any single bin is defined by intervals of the location, time, magnitude, and focal mechanism. Thus a bin is a multi-dimensional interval. The resolution of a forecast corresponds to the bin sizes; the smaller the bins, the higher the resolution.

From the rates specified in each forecast we calculate a vector of expectations, the expected number of events within the time interval for all bins; each element of the vector corresponds to a particular bin. The expected number is the earthquake rate multiplied by the volume in parameter space of the bin. An expectation need not be a whole number nor must it be less than 1. The expectations are dimensionless, but correspond directly to earthquake rates per unit area, magnitude, time, and depth and orientation of angles if specified. The vector of expectations is compared with the vector of observations, based on the same binning, to score a given forecast. The observed number of events will always be integers.

In some texts the expectation is referred to as the prediction or predicted number of events for the bin. While the term prediction has a fairly standard meaning in statistics, it has a different understanding in earthquake studies. Earthquake prediction usually refers to a single earthquake and implies both high probability and imminence. We consider earthquake prediction as a special case of a forecast in which the forecast rate is temporarily high enough to justify an exceptional response beyond that appropriate for normal conditions. One can also adopt the definition of prediction by *Main* [1999]. We do not use the term prediction to avoid confusion.

A useful measure of the agreement between a hypothesis and an earthquake record is the joint likelihood, defined as the probability of realizing the observed

number of earthquakes, given the expectations in each of the bins. In all of the models proposed to date, the expectations for the various bins are assumed to be independent. In this case the joint likelihood is the product of the likelihoods of each bin. The logarithm of the joint likelihood, sometimes called the log-likelihood or log-likelihood score, is simply the sum of the logs of the likelihoods for all bins.

By comparing the observed events to a model's expectations, we derive the likelihood of the observed events occurring in our model. We calculate this likelihood by assuming a Poissonian distribution of events in each bin. While the Poisson model is strictly valid only if the forecast rate is truly constant during the test interval, it is a good approximation when the rates do not vary much within the time interval [*Borradaile*, 2003].

The log-likelihood score depends on both the earthquake record and the forecast rates; higher values imply better agreement between the two. How large is large enough? We answer this question with two comparisons: First, in the consistency test, we compare the observed likelihood score with Poissonian simulations of expectations based on the hypothesis; Second, in the relative consistency test, we compare two hypotheses using the observed likelihood and the simulated expectations obtained using the forecast probabilities from the alternative hypothesis. In this project, we will compare likelihood scores from all pairs of hypotheses defined using the same bins.

Computation

As outlined above, a hypothesis is expressed as a forecast of earthquake rates per specified bin. Any bin is defined by intervals of location (volume), magnitude, time, and focal mechanism angles. We denote bins with b and all bins constitute the set $\mathcal{B}$ defined as

$$\mathcal{B} := \{b_1, b_2, \ldots, b_n\}, n = |\mathcal{B}|$$

where n is the number of bins b_i in the set $\mathcal{B}$.

A forecast of a model j is issued as expectations λ_i^j per bin b_i. We set up a vector Λ^j of all expectations as

$$\Lambda^j = \begin{pmatrix} \lambda_1^j \\ \lambda_2^j \\ \vdots \\ \lambda_n^j \end{pmatrix}, \lambda_i^j := \lambda_i^j(b_i), b_i \in \mathcal{B}$$

Expectations have units of earthquakes per time. We also design the vector Ω of observations ω_i per bin b_i based on the same binning as the vector Λ to be

$$\Omega = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{pmatrix}, \omega_i = \omega_i(b_i), b_i \in \mathcal{B}$$

Assuming that earthquakes are independent, the likelihood of observing ω events in a bin with an expectation λ is the Poissonian probability p

$$p(\omega|\lambda) = \frac{\lambda^\omega}{\omega!} e^{-\lambda}$$

In case of $\lambda = 0$ the probability p is given as

$$p(\omega|0) = \begin{cases} 0, & \omega > 0 \\ 1, & \omega = 0 \end{cases}$$

The log-likelihood L for observing ω earthquakes at a given expectation λ is defined as the logarithm of the probability $p(\omega|\lambda)$, thus

$$L(\omega|\lambda) = \log p(\omega|\lambda) = -\lambda + \omega \log \lambda - \log \omega!$$

and for a model j at a bin b_i

$$L(\omega_i|\lambda_i^j) = -\lambda_i^j + \omega_i \log \lambda_i^j - \log \omega_i!$$

Applying the logarithm to the probabilities p in case of $\lambda = 0$ gives

$$L(\omega|0) = \begin{cases} 1, & \omega > 0 \\ -\infty, & \omega = 0 \end{cases}$$

The joint likelihood is the product of the individual bin likelihoods, so its logarithm $L(\Omega|\Lambda^j)$ is the sum of $L(\omega_i|\lambda_i^j)$ over all bins b_i

$$L^j = L(\Omega|\Lambda^j) = \sum_{i=1}^{n} L(\omega_i|\lambda_i^j) = \sum_{i=1}^{n} -\lambda_i^j + \omega_i \log \lambda_i^j - \log \omega_i!$$

To compare the joint log-likelihoods of two models we compute the log-likelihood-ratio, defined as

$$R = L(\Omega|\Lambda^0) - L(\Omega|\Lambda^1) = L^0 - L^1$$

where Λ^0 denotes the vector of expectations of model H^0, Λ^1 denotes the vector of expectations of model H^1. L^0 and L^1 are the joint likelihoods of models H^0 and H^1, respectively. If the log-likelihood-ratio R is less than 0, model H^1 provides a more likely forecast; if $R > 0$, model H^0 performs better.

Special attention must be paid to forecasts with any cell containing $\lambda = 0$. As long as zero events occur in these cells, the model is judged as any other model; however, if an event does occur in a cells with expectation $\lambda = 0$, the joint log-likelihood sums up to $-\infty$. Thus, the model will be rejected.

Uncertainties in Earthquake Parameters

None of the observed earthquake parameters (location, focal time, etc.) can be estimated without uncertainties. Therefore, each parameter uncertainty is included in the testing. Additionally, for testing stationary models, every observed event is assigned an independence probability p_I of not being associated

with an earthquake cluster; we calculate these probabilities by using a Monte Carlo approach to the *Reasenberg* [1985] declustering algorithm. The lower p_I the more likely the event is an aftershock.

To account for these uncertainties, we generate s simulated "observed catalogs" using modified observations of each event based on its uncertainty. We draw random numbers according to the given uncertainty distribution for each parameter of each earthquake to obtain a modified parameter. Also, for each simulated event, we draw a random number from a uniform distribution between 0 and 1 to decide whether each event will be considered independent, given its independence probability p_I. If not, it will be deleted from the record. This gives a modified observation vector $\widetilde{\Omega}$ (Modified observations are denoted with a tilde).

$$\widetilde{\Omega} = \begin{pmatrix} \widetilde{\omega}_1 \\ \widetilde{\omega}_2 \\ \vdots \\ \widetilde{\omega}_n \end{pmatrix}, \widetilde{\omega}_i = \widetilde{\omega}_i(b_i), b_i \in \mathcal{B}$$

Repeating this procedure s times yields a set of modified observations of the event record $\{\widetilde{\Omega}_1, \widetilde{\Omega}_2, \ldots, \widetilde{\Omega}_s\}$, representing its uncertainty and its possible realizations. It should be noted that parameter uncertainties may cause events to be associated with different bins while event independence probabilities p_I may change the total number of events in a record $\widetilde{\Omega}$.

To represent the uncertainties of earthquake parameters in the results, we compute s times the log-likelihoods L^j and the log-likelihood-ratios R using the modified observations $\widetilde{\Omega}$, obtaining sets of log-likelihoods $\widetilde{\mathcal{L}}^j = \{\widetilde{L}_1^j, \widetilde{L}_2^j, \ldots, \widetilde{L}_s^j\}$, total numbers of events $\widetilde{\mathcal{N}} = \{\widetilde{N}_1, \widetilde{N}_2, \ldots, \widetilde{N}_s\}$, and log-likelihood-ratios $\widetilde{\mathcal{R}} = \{\widetilde{R}_1, \widetilde{R}_2, \ldots, \widetilde{R}_s\}$. The log-likelihood of a model j is the mean value of $\widetilde{\mathcal{L}}^j$ and its standard deviation is given by the second moment of $\widetilde{\mathcal{L}}^j$. Corresponding to that, the log-likelihood-ratio between two models is the mean value of $\widetilde{\mathcal{R}}$ and the standard deviation is given accordingly.

Simulation and Evaluation

Testing based on likelihood raises two questions: How can we know the expected value of the likelihood? and, if the likelihood for the observed earthquake record exceeds the expected value, how can we know whether the result is truly significant rather than accidental? To answer these questions, we need to derive a probability distribution for the likelihood score. In some simple cases the distribution might be derived analytically from the rates in the forecast; however, the analytic solution is not practical here, so we derive the distribution of expected likelihood scores by simulation. That is, we draw random numbers according to the probabilities implied by the forecast to generate random earthquake records $\widehat{\Omega}_k$ (simulated values are denoted with a hat) consistent with the forecast. Then we compute the likelihood score $\widehat{L}_k^j$ for each of these simulated records, obtaining the set $\widehat{\mathcal{L}}^j = \{\widehat{L}_1^j, \widehat{L}_2^j, \ldots, \widehat{L}_m^j\}$. From this distribution, we then can compute

the significance as quantiles of the observed values compared to the distribution of simulated values.

To create the simulated observations, we draw random numbers from a uniform distribution in the interval $[0; 1]$, for every bin and every simulation run. We use this random number as the probability of the inverse cumulative Poissonian probability density function. This yields a simulated number of observed events $\widehat{\omega}_i^j$ for each given bin b_i assuming the expectations λ_i^j of model H^j. Iterating through all bins creates a vector of simulated events $\widehat{\Omega}^j$ based on model H^j.

$$\widehat{\Omega}^j = \begin{pmatrix} \widehat{\omega}_1^j \\ \widehat{\omega}_2^j \\ \vdots \\ \widehat{\omega}_n^j \end{pmatrix}, \widehat{\omega}_i^j = \widehat{\omega}_i^j(b_i), b_i \in \mathcal{B}$$

We will denote multiple simulated vectors with $\widehat{\Omega}_1^j, \widehat{\Omega}_2^j, \ldots, \widehat{\Omega}_m^j$. The subscript of $\widehat{\Omega}$ is the number of the simulation.

Again, the case of $\lambda = 0$ requires a special treatment; in this case the corresponding $\widehat{\omega}$ will always be 0.

Data-consistency test or L-Test

Consider the data-consistency test, and assume that the hypothesis is true. This test shows whether the observed likelihood of the hypothesis is consistent with likelihoods obtained from simulations or not. A useful measure for this comparison is the quantile score γ_q, or the fraction of simulated likelihood values $\widehat{\mathcal{L}}^j = \{\widehat{L}_1^j, \widehat{L}_2^j, \ldots, \widehat{L}_m^j\}$, $m = |\widehat{\mathcal{L}}^j|$ less than the observed likelihoods $\widetilde{\mathcal{L}}^j$

$$\gamma_q^j = \frac{|\{\widehat{L}_k^j | \widehat{L}_k^j \leq \widetilde{L}_q^j, \widehat{L}_k^j \in \widehat{\mathcal{L}}^j, \widetilde{L}_q^j \in \widetilde{\mathcal{L}}^j\}|}{|\widehat{\mathcal{L}}^j|}$$

Here $\widehat{L}_k^j$ denotes the log-likelihood of the k-th simulation and $\widetilde{L}_q^j$ the log-likelihood of the q-th modification of the event record. Thus, we perform this computation s times iterating through all modifications of the event record. This results in a distribution of quantile scores $\{\gamma_1^j, \gamma_2^j, \ldots, \gamma_s^j\}$. The quantile score γ^j is the mean of this distribution and its standard deviation is given as second moment of this distribution.

If γ^j is low, then the observed likelihood score is less than most of the simulated values, and the record is not consistent with the forecast. If the observed likelihood is in the middle of the simulated values, then, according to this test, the forecast is consistent with the data. A problem arises when considering results with a high γ^j. It means that the likelihood of the real observation is higher than the likelihood scores of the simulations based on the model itself. There are several scenarios when this can happen. First, in a catalog with very low total expectations, the outcome of 0 events is the most likely one; nevertheless, the sum of all given rates may exceed 1 and

therefore some events will be forecasted. In this case, the outcome of 0 events would have a much higher likelihood than the average simulation because the simulations will reflect the total number of expected earthquakes, distributed over the cells (> 0). In contrast, a forecast with expectations exactly matching the observations would also have a higher likelihood when compared to the likelihood scores of the simulations. This is because every simulation will in general add poissonian scatter to the expectations, thus generating observations that do not match the expectations. This will result in lower likelihoods for the simulations (high γ^j). As can been seen, a model should not be rejected based on high likelihoods in the data-consistency test. Accordingly, the test is one-sided, rejecting forecasts with a significantly low likelihood when compared to the simulations. However, models with high likelihoods may be inconsistent with the observed data; therefore, we will additionally apply the number test (N-Test) to determine if this is the case (see next section).

Number test or N-Test

The N-Test also tests the consistency of a model with the observation. However, instead of comparing the observed likelihoods with likelihoods obtained from simulations, the N-Test compares the observed total number of events with the number of events in the simulated catalogs; no spatial or magnitude information is contained in this test. Again, we use a quantile score δ_q for this comparison. The total number N^j of expected events of a model j is simply the sum over all expectations λ_i^j

$$N^j = \sum_{i=1}^{n} \lambda_i^j$$

while the total number of observed events N is the sum over all ω_i

$$N = \sum_{i=1}^{n} \omega_i$$

Simulating earthquake records according to the probabilities of model j, as done in the L-Test, leads to a set of total numbers of earthquakes records $\widehat{\mathcal{N}}^j = \{\widehat{N}_1^j, \widehat{N}_2^j, \ldots, \widehat{N}_m^j\}$. The quantile score δ^j is defined as the fraction of $\widehat{\mathcal{N}}^j$ smaller than the observed number of events $\widetilde{N}$.

$$\delta_q^j = \frac{|\{\widehat{N}_k^j | \widehat{N}_k^j \leq \widetilde{N}_q, \widehat{N}_k^j \in \widehat{\mathcal{N}}^j, \widetilde{N}_q \in \widetilde{\mathcal{N}}\}|}{|\widehat{\mathcal{N}}^j|}$$

As in the L-Test, we require that the observed number of events N is in the middle of the distribution $\widehat{\mathcal{N}}^j$ for a model j to be considered consistent with the observations. Because this test examines only the total number of events it is weaker than the L-Test; however, the N-Test is necessary to overcome the problem of underpredicting which can be missed in the L-Test. If a model is underpredicting the total number of events, it may not be rejected in the L-Test, but it will fail in the N-Test. If we can reject models in the L-Test, the N-Test

is not necessary. If a model cannot be rejected in the L-Test, the N-Test may show that a model is underpredicting events and can be rejected.

Hypotheses Comparison or R-Test

In many studies (e. g. [*Kagan and Jackson*, 1994]), a "test hypothesis" is compared to an established "null hypothesis." The null hypothesis is the more simple hypothesis and the test hypothesis is only accepted if an observed statistic would have a significantly low probability under the null hypothesis. Evaluating that probability requires knowledge of the distribution, usually estimated by simulation, of the relevant test statistic under the null hypothesis.

Our study differs from most textbook cases because all models we consider are fully specified in advance. Some hypotheses may be derived by using more degrees of freedom during the "learning" period, but these parameters are then fixed before the test, so all hypotheses have exactly the same number of free parameters: none. Furthermore we have no null hypothesis that we believe should be accepted over others in case of doubt. Nevertheless, we wish to exploit the methods used for testing against null hypotheses, without necessarily choosing a favorite a priori.

It should be mentioned here that additional parameters in a model do not correspond to additional degrees of freedom of models in our experiment; this makes the use of the Akaike Information Criterion (AIC) [*Akaike*, 1973, 1974] or any other related method (e. g. $\mathrm{AIC_c}$, BIC, etc. [*Chow*, 1981]) impossible. In all models tested in the RELM framework, the number of degrees of freedom is 0 because every forecast is issued in advance of the observation period and is not readjusted during the observation period.

If we were to select a single "simple" null hypothesis and repeat this test subsequently with all hypotheses, it is likely that the first hypothesis to test against the "simple" null hypothesis would win and the null hypothesis would be rejected in favor of the tested hypothesis. Unfortunately, it will also be possible, or even likely, that none of the remaining hypotheses will be able to beat the new null hypothesis at the given significance level. Therefore, we compare all hypotheses against the others with a different definition of the test statistic to avoid the first model tested becoming the default winner.

In the test hypothesis against null hypothesis one uses a simple likelihood ratio

$$R = L(\Omega|\Lambda^0) - L(\Omega|\Lambda^1)$$

and obtains the significance level α by computing log likelihood ratios $\widehat{R}_k$ of simulated observation $\widehat{\Omega}_k$.

Now consider the comparative likelihood test, in which we commit to accept one hypothesis and reject the other. Suppose we use the same observed record to compute likelihood scores for two hypothesis, say H^1 and H^2. We call these likelihood scores $L^1 = L(\Omega|\Lambda^1)$ and $L^2 = L(\Omega|\Lambda^2)$, and let the log-likelihood-ratio $R^{21} = L^2 - L^1$. If R^{21} is large it would seem to support H^2, but how can we know whether the result is significant? The likelihood ratio is a statistic, as

	H^1	H^2	...	H^n
H^1	α^{11}	α^{21}	...	α^{n1}
H^2	α^{12}	α^{22}	...	α^{n2}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
H^n	α^{1n}	α^{2n}	...	α^{nn}

Table 1: Table of results of the R-Test of n Hypotheses.

described above, and we can derive its probability distribution by simulation. We assume H^2 is correct, generate many synthetic records, and score each using both Λ^1 and Λ^2 separately (as we did for the observed record), obtaining the set $\widehat{\mathcal{R}}^{21} = \{\widehat{R}_1^{21}, \widehat{R}_2^{21}, \ldots, \widehat{R}_m^{21}\}$ with

$$\widehat{R}_k^{21} = L(\widehat{\Omega}_k^2|\Lambda^2) - L(\widehat{\Omega}_k^2|\Lambda^1)$$

Let α^{21} be the fraction of simulated values of $\widehat{R}_k^{21}$ less than the observed $\widetilde{R}^{21}$. Large values support H^2.

$$\alpha_q^{21} = \frac{|\{\widehat{R}_k^{21} | \widehat{R}_k^{21} \leq \widetilde{R}_q^{21}, \widehat{R}_k^{21} \in \widehat{\mathcal{R}}^{21}, \widetilde{R}_q^{21} \in \widetilde{\mathcal{R}}^{21}\}|}{|\widehat{\mathcal{R}}^{21}|}$$

We perform this computation s times iterating through all modifications of the event record. This results in a distribution of quantile scores $\{\alpha_1^{21}, \alpha_2^{21}, \ldots, \alpha_s^{21}\}$. The quantile score α^{21} is the mean of this distribution and its standard deviation is given as second moment of this distribution.

So far we have focussed on H^2, but we should focus on H^1 as well. We derive the distribution $\widehat{\mathcal{R}}^{12}$ assuming that H^1 is correct by simulating records using H^1, then score them using both Λ^1 and Λ^2 separately as above. Let $R^{12} = L^1 - L^2$ for both observed and simulated catalogs, and compare the observed and synthetic using α^{12} (fraction of synthetics less than observed).

The advantage of this approach is its symmetry in respect to the models. When swapping H^1 and H^2, simply α^{21} and α^{12} are swapped. For interpretation of the outcome of this test we will provide a result table containing all computed α-values. Consider a test run with n hypotheses H^1, H^2, ..., H^n. Each of these hypotheses will play the role of a null hypothesis against the others as well as the role of a test hypothesis against the others. Performing the aforementioned test will lead to a set of α-values as shown in Table 1.

Evaluation

We have described the procedure of three tests: 1) the relative performance test (R-Test); 2) the data-consistency test in likelihood space (L-Test); 3) the data-consistency test in number space (N-Test).

If we were to try to define the best performing model, this model must never be rejected in the R-Test and must show data-consistency in the L- and N-Tests. This would be a statement about the model's performances over the full magnitude range and the entire testing area.

Additionally, we propose more detailed investigations of a model's performance by testing smaller spatial and magnitude ranges. Therefore, we compute and retain the significances for each model for each bin. This results in maps from which areas can be identified for which certain models show a strong or weak performance. This kind of secondary tests can help understanding how and why models perform as they do.

Examples

To illustrate possible testing scenarios and to give a feeling of test performances, we have undertaken a few example tests with models which potentially will be part of the RELM testing framework. We performed all tests (L-Test, N-Test, and R-Test) with two models. The first model H^0 is derived from the USGS 1996 long-term hazard model [*Frankel et al.*, 1996] and the second model H^1 is a stationary model designed by *Helmstetter et al.* [submitted]. We tested these models against the earthquake catalog ($M \geq 4.7$) provided by Yan Kagan (`http://moho.ess.ucla.edu/~kagan/relm_index.html`). In one set of tests we use a catalog $\mathrm{C_M}$ containing only events with independence probability $p_I = 1$, thus only main shocks. In the second set we use a different catalog $\mathrm{C_A}$ which includes all events. Thus we create modifications of the observation based on the independence probabilities. We tested using three different time frames, a 20-year period (1981–2000), a 70-year period (1932–2001), and a 5-year period (1998–2002). For all tests, we used a significance level of 0.1.

Model $\mathbf{H^0}$

We adapted the USGS 1996 model [*Frankel et al.*, 1996] to fit the requirements of the test; we have derived the daily background rate from the long-term rate of this model, interpolated the characteristic fault information onto our grid, and extrapolated all rates Λ^0 into $\Delta M = 0.1$ magnitude bins down to magnitude $M = 5$ using the Gutenberg-Richter relationship. The characteristic fault information was interpolated to the given grid by: 1) projecting the fault to the surface, 2) distributing a large number of points over the surface projection of the fault, 3) counting the number of points that fell within a grid node, and 4) assigning the appropriate percentage of rates to each grid node, based on the percentage of overall points that fell within the node. The nodewise sums of expectations λ_i^0 over all magnitude bins for a 20-years period are shown in Figure 1.

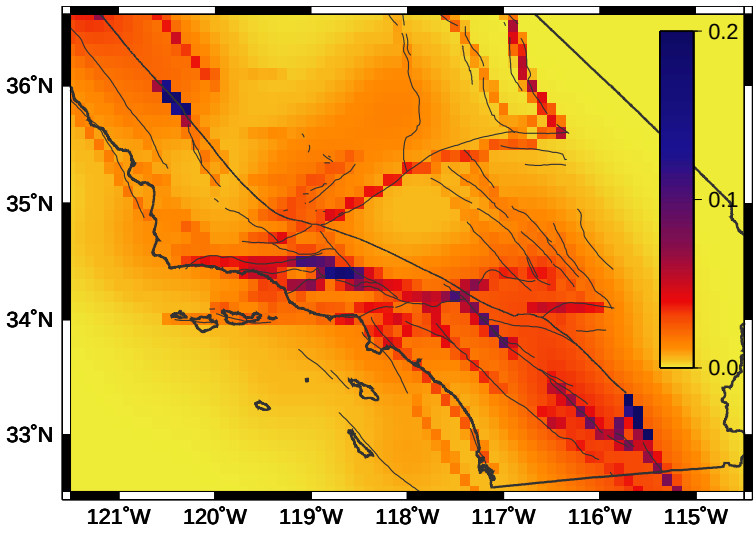

Figure 1: Nodewise sum of the expectations λ_i^0 over all magnitude bins for a 20-year period of the model H^0 (based on the USGS 1996 model [*Frankel et al.*, 1996]).

Consistency of Model H^0 with the Observation

We first performed the L-Test to show whether H^0 is consistent with the observation for the 20-year period 1981–2000 using the catalog C_M ($p_I = 1$). We performed 10,000 simulations obtaining $\widehat{\Omega}$. The result is shown in Figure 2A. As can be seen, the model is consistent with the observations. The curve of log-likelihoods based on simulated observations $\widehat{\Omega}$ (green curve in Figure 2A) intersects the log-likelihood of the real observation (black vertical line) at $\gamma^0 = 0.389$.

In the N-Test the model shows almost the same consistency with the observation ($\delta^0 = 0.314$). The total number of events in the simulated records range from 17 to 60 while 33 events have been observed. Therefore, we can state that model H^0 is consistent with the observations in the given period.

The result changes if we use catalog C_A and modify the observations (10,000 runs) based on the independence probability p_I, thereby introducing additional events into the catalog. Figure 3A shows that almost all log-likelihoods computed using simulated observations $\widehat{\Omega}$ are higher than any of the log-likelihoods computed using the modifications $\widetilde{\Omega}$ of the observation record. The $\widehat{L}$ range from -409.58 to -126.56 while the $\widetilde{L}$ span only the range from -419.93 to -335.72. This results in a low $\gamma^0 = 0.0018 \pm 0.0016$. The N-Test gives an explanation for this result (Figure 3B). Model H^0 is underpredicting the total number of

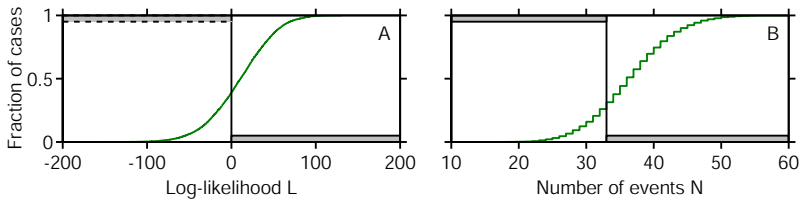

Figure 2: Result of the data-consistency test of model H^0 for the period 1981–2000 using catalog C_M. The gray patches mark the rejection bars. (A) L-Test. The green curve indicates the cumulative distribution of log-likelihoods based on simulated events $\widehat{\Omega}$. The vertical black line indicates the log-likelihood of the real observation Ω ($\gamma^0 = 0.389$). (B) N-Test. The green curve indicates the cumulative distribution of numbers of simulated events $\widehat{\Omega}$. The vertical black line indicates the observed events Ω ($\delta^0 = 0.314$).

events ($\delta^0 = 0.990 \pm 0.007$), thus showing higher likelihoods in the simulations than expected considering the real observation. The number of events in the modified observation records ranges from 44 to 57, while the total expectation of the model is 36.52.

Therefore, we can state that model H^0 is consistent with the observation in catalog C_M of events with $p_I = 1$ while when including the uncertainties of events being main shocks or aftershocks, the model underpredicts the total number of events. This results in too high log-likelihoods.

Model H^1

The second model H^1 [*Helmstetter et al.*, submitted] has the same total expectation as model H^0 (i. e. the exact total number of forecasted events) but a different spatial distribution of expectations. The nodewise expectations are shown in Figure 4. The expectations are more concentrated in areas of active seismicity and less smoothed over the area.

Consistency of Model H^1 with the Observation

Repeating the same tests for model H^1, shows a very similar distribution (Figure 5). Using catalog C_M, the model is consistent with the observation, while with catalog C_A, the model underpredicts the total number of events. This results in too high log-likelihoods as we observed in the consistency test for model H^0.

To quantify the result, we computed the γ- and δ-values of the tests. Performing the test with catalog C_M gives $\gamma^1 = 0.531$. This shows the consistency of the model with the observation. Because the total expectation of this model is the same as of model H^0, the N-Test gives the same result of $\delta^1 = 0.314$. Using catalog C_A results in $\gamma^1 = 0.009 \pm 0.006$ and $\delta^1 = 0.990 \pm 0.007$.

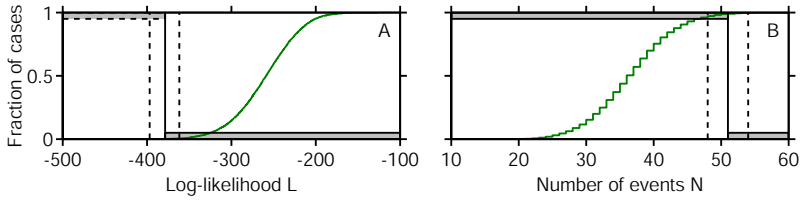

Figure 3: Result of the data-consistency test of model H^0 for the period 1981–2000 using catalog C_A. The gray patches mark the rejection bars. (A) L-Test. The green curve indicates the cumulative distribution of log-likelihoods based on simulated events $\widehat{\Omega}$. The vertical solid black line indicates the median of the log-likelihoods computed with the modifications of the observation record $\widetilde{\Omega}$. The vertical dashed lines indicate the 5 and 95 percentile of log-likelihoods computed with the modifications of the observation record $\widetilde{\Omega}$. (B) N-Test. The green curve indicates the cumulative distribution of numbers of simulated events $\widehat{\Omega}$. The vertical solid black line indicates the the median of the numbers of events in the modified observation records $\widetilde{\Omega}$. The vertical dashed lines indicate the 5 and 95 percentile of the numbers of events in the modified observation records $\widetilde{\Omega}$.

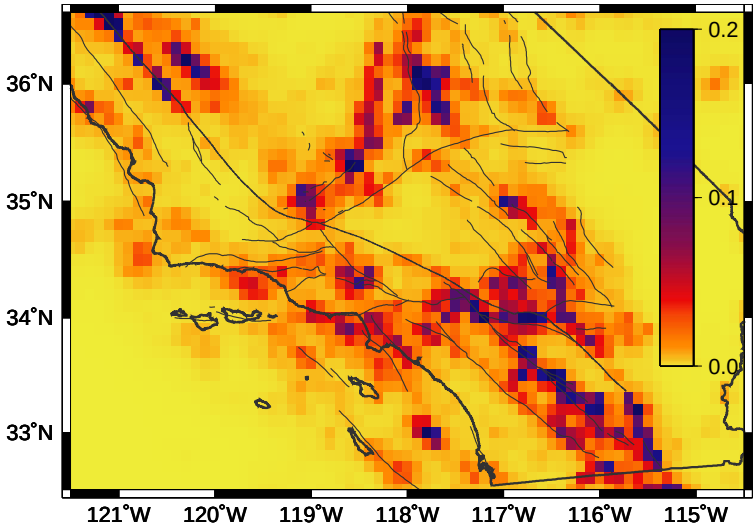

Figure 4: Nodewise sum of the expectations λ_i^1 over all magnitude bins for a 20-years period of the model H^1 [*Helmstetter et al.*, submitted].

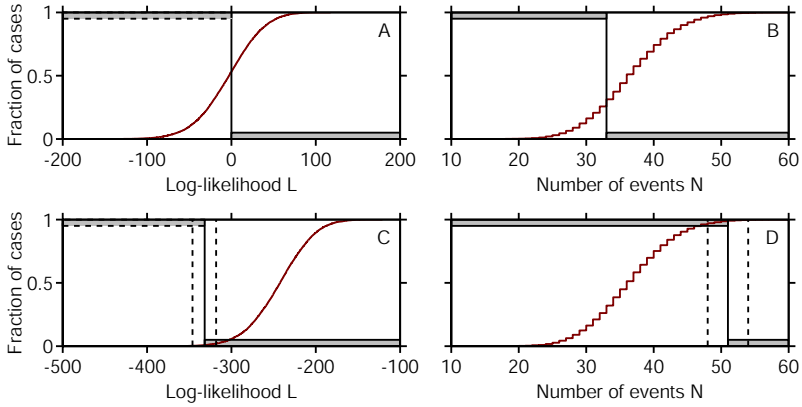

Figure 5: Result of the data-consistency test of model H^0 for the period 1981–2000 using both catalogs. See description of Figures 2 and 3. The red color indicates results of model H^1. (A) L-Test using catalog C_M ($\gamma^1 = 0.389$). (B) N-Test using catalog C_M ($\delta^1 = 0.314$). (C) L-Test using catalog C_A ($\gamma^1 = 0.009 \pm 0.005$). (D) N-Test using catalog C_A ($\delta^1 = 0.990 \pm 0.007$).

Model comparison

We have seen that both models are consistent with the observation when using catalog C_M, but inconsistent using catalog C_A. So far, we can not decide which model has a better forecast performance. Here we want to investigate their comparative spatial performance using the R-Test.

Figure 6 shows the results using both catalogs. In both cases, log-likelihood-ratios based on the expectations of model H^1 are in the range of the observed log-likelihood-ratio, giving $\alpha^{10} = 0.179$ and $\alpha^{10} = 0.321 \pm 0.049$. At the same time, model H^0 can be rejected in favor of the alternative model H^1 at the given significance level, because $\alpha^{01} = 0$ and $\alpha^{01} = 0 \pm 0$.

Evaluating our results for the given time period, we reject model H^0 in favor of model H^1 due to its spatial performance. We also state, that both models forecast the total number of events equally well or badly. Using catalog C_A, both models fail to forecast the average seismicity while with catalog C_M both model's forecast are consistent with the average seismicity.

Tests over different time periods

We repeated the tests for two additional time periods of the catalogs. Figures 7 and 8 show the distributions of all tests and Tables 2 and 3 give the quantitative results. The period of 70 years shows a similar result as the previous results for the 20-years period. The L-Test and N-Test show consistency of both models with catalog C_M (Figures 7A and 7B) and an underprediction of events with catalog C_A (Figures 7C and 7D). Here again, we reject model H^0 in favor of model H^1 due to their spatial performance in the R-Test (Figures 7E and 7F).

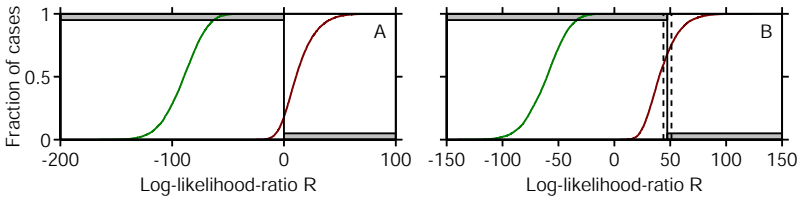

Figure 6: Result of the R-Test comparing the performance of models H^0 and H^1 for the period 1981–2000 using both catalogs. The green and red curves indicate the cumulative distribution of log-likelihood-ratios computed with simulated observations based on Λ^0 and Λ^1, respectively. The gray patches mark the rejection bars. (A) R-Test using catalog C_M ($\alpha^{10} = 0.179$, $\alpha^{01} = 0$). The vertical line marks the observed log-likelihood-ratio. (B) R-Test using catalog C_A ($\alpha^{10} = 0.321 \pm 0.049$, $\alpha^{01} = 0 \pm 0$). The vertical solid black line indicates the median of the log-likelihood-ratios computed with the modified observation records $\widetilde{\Omega}$. The vertical dashed lines indicate the 5 and 95 percentile of the log-likelihood-ratios computed with the modified observation records $\widetilde{\Omega}$.

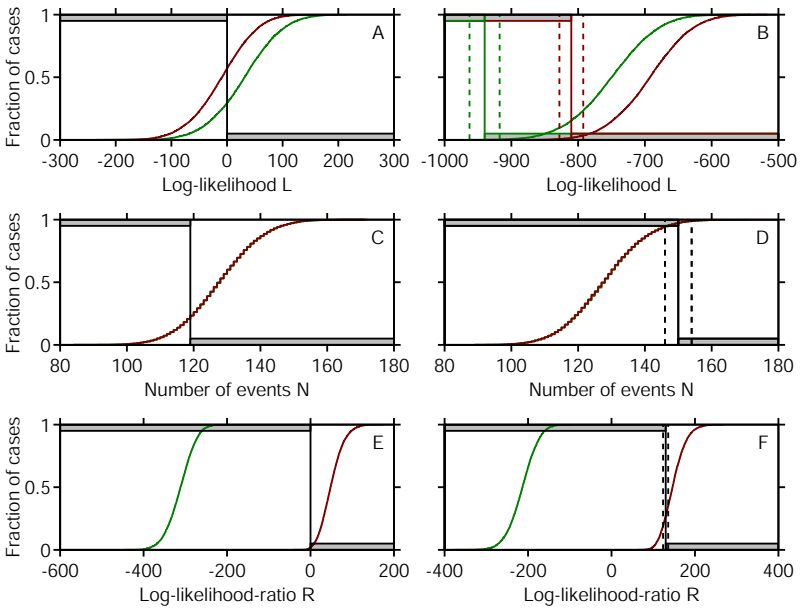

Figure 7: Results of all tests for the 70-years period 1932–2001. (Left column) Catalog C_M. (Right column) Catalog C_A. (A, B) L-Test. The median, 5, and 95 percentile lines are drawn in the color corresponding to the model. (C, D) N-Test. (E, F) R-Test. Quantitative results are listed in Table 2.

Test	Catalog C_M	Catalog C_A
L-Test	$\gamma^0 = 0.295$	$\gamma^0 = 0.001 \pm 0.001$
	$\gamma^1 = 0.569$	$\gamma^1 = 0.018 \pm 0.009$
N-Test	$\delta^0 = 0.235$	$\delta^0 = 0.974 \pm 0.013$
	$\delta^1 = 0.233$	$\delta^1 = 0.974 \pm 0.014$
R-Test	$\alpha^{01} = 0$	$\alpha^{01} = 0 \pm 0$
	$\alpha^{10} = 0.019$	$\alpha^{10} = 0.731 \pm 0.050$

Table 2: Results of all tests for the 70-years period 1932–2001. Distributions are shown in Figure 7.

Test	Catalog C_M	Catalog C_A
L-Test	$\gamma^0 = 0.948$	$\gamma^0 = 0.771 \pm 0.074$
	$\gamma^1 = 0.962$	$\gamma^1 = 0.816 \pm 0.054$
N-Test	$\delta^0 = 0.047$	$\delta^0 = 0.230 \pm 0.069$
	$\delta^1 = 0.048$	$\delta^1 = 0.229 \pm 0.070$
R-Test	$\alpha^{01} = 0.002$	$\alpha^{01} = 0 \pm 0$
	$\alpha^{10} = 0.128$	$\alpha^{10} = 0.719 \pm 0.114$

Table 3: Results of all tests for the 5-years period 1998–2002. Distributions are shown in Figure 8.

In the 5-year period 1998–2002 the results look quite different. While both model overpredict the number of events in catalog C_M (Figure 8C), they are consistent with catalog C_A (Figures 8B and 8D). The comparative R-Test again shows that model H^1 produces a better forecast than model H^0. H^0 can be rejected in favor of H^1 at the given significance level.

This last tests over the 5-years period are a good test case for the forecast model testing in the RELM framework; they show that it is possible to distinguish the forecast capabilities of models after a 5-year period. This is further emphasized by the fact that the two models used here have the same total expectation of events. In the RELM framework, the models will not necessarily exhibit the same total expectation, making their differences even bigger.

Discussion

In this paper, we define the scientific foundation for testing earthquake likelihood models in the framework of the RELM project. We believe that the outlined tests are well suited for the task of analyzing the forecast ability of models in California and other regions. The statistical tools can be used in a diagnostic sense, allowing seismologist to quantitatively compare models, identifying their

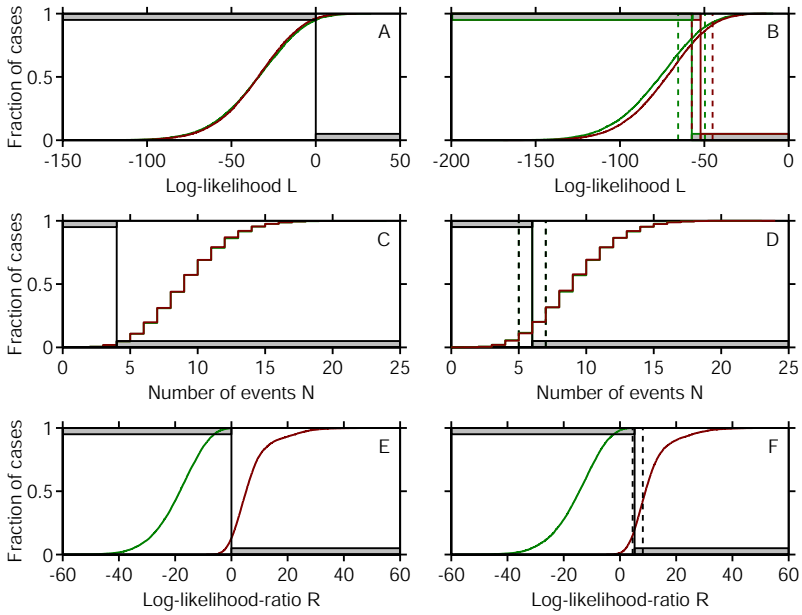

Figure 8: Results of all tests for the 5-years period 1998–2002. (Left column) Catalog C_M. (Right column) Catalog C_A. (A, B) L-Test. The median, 5, and 95 percentile lines are drawn in the color corresponding to the model. (C, D) N-Test. (E, F) R-Test. Quantitative results are listed in Table 3.

strong and weak points. This will allow for an improved understanding of the physical processes at work and ultimately help to improve probabilistic seismic hazard assessment.

Multiple forecast models are available within RELM, each of them covering different aspects of the physics of earthquakes or their pattern of occurrence. To improve our forecasting ability, we must evaluate these models without prejudice or bias. The best, and possibly only way to achieve this goal is to test the forecasts of all models in a truly prospective test against observed seismicity. During a 5-year period, the RELM project will undertake testing of a suite of forecasting models. The performance of all competing models will be determined and evaluated. The logistical details of how this test is implemented at a RELM test center are described in *Schorlemmer and Gerstenberger* [in this volume].

Although, the described tests cover data-consistency evaluation as well as spatial comparative performance tests, they lack the ability to judge models on only their performance for large ($M \geq 7$) events. This problem is not due to insufficient sophistication of the procedure but inherent to tests covering only a short period of time; a 5-year period is not enough time to make significant statements for $M \geq 7$ events in California. In the case of the two tested models (Figures 1 and 4), the total expectation for any event of magnitude $M \geq 7$ is 0.436. This means, that even less than one event of this magnitude range is expected in a 10-year period. Trading time versus space, one can extend forecast models on a global scale ([*Jackson*, 1996; *Kagan and Jackson*, 1991, 1994, 1995, 2000]) and thus have enough events per year for testing. However, most of the models proposed within RELM depend on information not available on a global scale (e. g., microseismicity, deformation, fault databases). This is a clear limitation in testing hazard related forecasting abilities, where much of the hazard stems from the largest events. It is likely that we will not be able to make a statement about which model has the highest performance in forecasting large events. However, we showed in this paper that for magnitude $M \geq 5$, sufficient events do occur within a 5-year time frame to evaluate forecast models (Figure 8). Unless one assumes that the moderate to large main shocks responsible for the majority of the hazard ($M \geq 6$) are not related to the occurrence rate of M5+ events, the RELM testing will be able to offer some insight into the relative and absolute performance of forecast models. Therefore, for practical reasons we are initially conducting our tests for only a 5 year period.

Appendix

Likelihood ratio independence on bin-sizes

Let P be the likelihood for observing x events for a given expectation (rate) λ:

$$\log P = -\lambda + x \log \lambda - \log x!$$

Let there be a cell C with a given rate λ and one observed event. The likelihood for this observation is

$$\log P = -\lambda + \log \lambda - \log 1 = -\lambda + \log \lambda$$

Now lets divide the cell C into n equally sized subcells $C_1, C_2, \ldots, C_n$. Since the event can only happen in one of the subcells, the likelihood of the observation is:

$$\log P = 1(-\lambda^* + \log \lambda^* - \log 1) + (n-1)(-\lambda^* - \log 1)$$

Because

$$\lambda^* = \frac{\lambda}{n}$$

and $\log 1 = 0$, we can write the likelihood of the observation as

$$\log P = (-\frac{\lambda}{n} + \log \frac{\lambda}{n}) + (n-1)(-\frac{\lambda}{n})$$

Rearranged:

$$\begin{aligned} \log P &= -\frac{\lambda}{n} + (n-1)(-\frac{\lambda}{n}) + \log \frac{\lambda}{n} \\ &= n(-\frac{\lambda}{n}) + \log \frac{\lambda}{n} \\ &= -\lambda + \log \frac{\lambda}{n} \\ &= -\lambda + \log \lambda - \log n \end{aligned}$$

The likelihood changed only by the term $\log n$. Thus, in the likelihood ratio this term will vanish because it does not depend on the λ and the likelihood ratio will be the same for the case with one cell as well as for the case with n cells.

Now let us assume m observed events. The likelihood for the case of only one cell is

$$\log P = -\lambda + m \log \lambda - \log(m!)$$

Regardless of the distribution of these m events over the given n subcells, the likelihood will be

$$\log P = -\lambda^* + m \log \lambda^* - X$$

where X is based on the original term $\log x!$ and reflects the distribution of the m events of the n cells. The likelihoods of all possible cases may differ but in the likelihood ratio the term X vanishes, making the likelihood ratio the same as in the one-cell case.

Definitions

Expectation The forecasted number λ of earthquakes for any given bin b, equal to the earthquake rate times the binsize.

Model The methodology used to express a scientific idea.

Hypothesis A model with all functions, parameters, etc. completely specified. In the framework of RELM a hypothesis must generate a well defined forecast of future earthquakes including location, magnitude and time.

Forecast A set Λ of numerical estimates of the expected number of earthquakes in each bin, based on a hypothesis.

Bin A bin b is defined by intervals of the location, time, magnitude, and focal mechanism, thus a multi-dimensional interval.

Likelihood The joint probability of observing ω_1 events in bin b_1 and ω_2 events in bin b_2, etc., given the expectations λ_1, λ_2, etc.

Likelihood ratio The ratio of likelihood values for two forecasts evaluated using the same catalog, or two catalogs using the same forecast.

Test Contrary to the standard null hypothesis tests, where a test hypothesis competes against a given null hypothesis, we test each hypothesis against all other hypotheses. Hereby, each hypothesis acts as both a null and a test hypothesis in two tests against every other hypothesis of its category. This is necessary because it is possible that all tests between a hypothesis of a RELM model and the null hypothesis will result in rejection of the null hypothesis. However, significance between two competing RELM hypotheses may be much more difficult to establish. Therefore, without this test, the first model to test against the null hypothesis could become the de facto null hypothesis even if it does not forecast significantly better than later models. We simulate earthquake rupture catalogs and follow a similar method to the standard approach used in likelihood-ratio testing to obtain the significances of our results.

References

Akaike, H., Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory*, pp. 267–281, Akademiai Kaidó, Budapest, 1973.

Akaike, H., A new look at the statistical model identification, *IEEE Trans. Automatic Control*, *19*, 716–723, 1974.

Borradaile, G., *Statistics of Earth Science Data*, Springer, Berlin, Heidelberg, New York, 2003.

Chow, G. C., A comparison of the information and posterior probability criteria for model selection, *Journal of Econometrics*, *16*, 21–33, 1981.

Frankel, A., C. Mueller, T. Barnhard, D. Perkins, E. V. Leyendecker, N. Dickman, S. Hanson, and M. Hopper, National seismic hazard maps, *United States Geological Survey Open-File Report 96-532*, 1996.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson, Comparison of short-term and long-term earthquake forecast models for southern california, *Bull. Seismol. Soc. Am.*, submitted.

Jackson, D. D., Hypothesis testing and earthquake prediction, *Proc. Natl. Acad. Sci. USA*, *93*, 3772–3775, 1996.

Jolliffe, I. T., and D. B. Stephenson (Eds.), *Forecast Verification*, John Wiley & Sons Ltd, 2003.

Kagan, Y. Y., and D. D. Jackson, Seismic gap hypothesis - 10 years after, *J. Geophys. Res.*, *96*(B13), 21,419–21,431, 1991.

Kagan, Y. Y., and D. D. Jackson, Long-term probabilistic forecasting of earthquakes, *J. Geophys. Res.*, *99*(B7), 13,685–13,700, 1994.

Kagan, Y. Y., and D. D. Jackson, New seismic gap hypothesis: Five years after, *J. Geophys. Res.*, *100*(B3), 3943–3959, 1995.

Kagan, Y. Y., and D. D. Jackson, Probabilistic forecast of earthquakes, *Geophys. J. Int.*, *143*, 438–453, 2000.

Main, I., Nature debate: Is the reliable prediction of individual earthquakes a realistic scientific goal?, `http://www.nature.com/nature/debates/earthquake/equake_contents.html`, 1999.

Reasenberg, P., Second-order moment of central california seismicity, 1969–1982, *J. Geophys. Res.*, *90*(B7), 5479–5495, 1985.

Schorlemmer, D., and M. Gerstenberger, Relm testing center, *Seismol. Res. Letts.*, in this volume.