Earthquake Likelihood Model Testing

D. Schorlemmer, D. D. Jackson and M. Gerstenberger

August 22, 2003

1 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 southern California. Based on differing assumptions, these models are produced both to test the validity of their assumptions and explore which models should be incorporated in seismic hazard and risk evaluation. Tests based on physical and geological criteria are useful but here we focus on statistical methods using future earthquake data only. We envision two evaluations: a self-consistency test, and comparison of every pair of models for relative consistency. Both tests are based on the likelihood ratio method, and both would be fully prospective (that is, the models are not adjusted to fit the test data). To be tested, each model must assign a probability or probability density 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 in some cases focal mechanism limits.

2 Introduction

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 the models might be assigned weights in a future consensus model or be judged as suitable for particular areas.

To test models against one another, we require that forecasts based on them can be expressed numerically in a standard format. That format is the average rate of earthquake occurrence within pre-specified limits of hypocentral latitude, longitude, magnitude, and time. For some source models there will also be bins describing depth, the inclination of P-axis (axis of maximum compression), declination of P-axis, and inclination of the T-axis (axis of least compression). There will be only a few choices for all of these limits, so that the models to be tested are not separated into too many different categories. Forecasts specified in this way are clear, easy to archive, and comparable between models.

We envision two time frames for our tests, based on the the common applications of forecasts. Quasi-static models assume that earthquake rates are relatively stable over about a year. Short-term models assume that rates vary from day to day because of stress changes and other variations possibly resulting from past earthquakes.

Quasi-static models are relevant for public policy, construction planning, and setting insurance rates and priorities for remediation, all of which require seismic hazard estimates valid for years or more. Thus the leaders of the RELM (Regional Earthquake Likelihood Models) project decided early on to develop a suite of source models for earthquakes over magnitude 5.0 in southern California over a five year period. Some quasi-static models are fundamentally time-dependent. For example, some renewal models assert that large-earthquake probability decreases substantially after a large event and only gradually recovers over a period of decades or centuries. We will evaluate the quasi-static forecasts once each year. We will also allow updates of the forecasts to exploit information gained from the yearly tests.

Short-term models are need for emergency response and public information announcements. These models incorporate probabilities that depend on time and distance from previous earthquakes, usually exploiting power-law time dependence evident in aftershock sequences. It is difficult to apply these models in any fixed time interval because of the implicit scale invariance. Because earthquake rates change so fast after an earthquake, only an automatic algorithm for updating rates would be adequate to implement the full time-dependence of these models. Automatic updating would be impractical for us, and it would make archiving more difficult. Instead we require that the forecasts from short-term models be provided in numerical form on a daily basis.

The methods for testing the forecasts will be the same for quasi-static and short-term models. In both cases the earthquake rates will be assumed to remain relatively constant over the test period, and the test will be based on the number of events forecast, and observed, in each interval of location, magnitude, time, and sometimes focal mechanism.

3 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 up to eight parameters reported in several earthquake catalogs: hypocentral latitude, longitude, depth, magnitude, origin time, P-axis inclination, P-axis declination, and T-axis inclination.

The last three describe the focal mechanism, and they would not necessarily be included in all forecasts. Strictly speaking, the P- and T- axes are not given in the catalog but can be derived unambiguously from the catalog moment tensor. The P- and T- axes define orthogonal fault and auxiliary planes, but in general we cannot determine from cataloged data which of the two is the fault plane. Thus we will use in our tests the orientations of the P- and T-axes but not of the fault plane orientation. Also, depth information is not required in issued forecasts. If no depth information is given by a forecast, the depth range will be set to 0–30 km to allow for comparisons with models providing depths information.

We will use the southern California earthquake catalog archived in the Southern California Earthquake Center Data Center (SCEC-DC). This catalog will contain focal mechanisms on a regular basis. In case no focal mechanism is provided for an earthquake above the magnitude threshold of testing, we will use the focal mechanism provided by the Harvard catalog.

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 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 probable hypocentral locations, magnitudes and focal mechanisms.

We envision two timeframes for forecast testing. Quasi-static models have to issue their forecasts for a 5 years period starting January 1st, 2004. Tests will be untertaken every year after an evaluated version of the catalog is available. We do not want to perform these tests with preliminary catalog data. Additionally, we do not want to remove any aftershock, thus decluster the catalog. This procedure is strongly debated and we do not foresee an agreement of every participant on a particular declustering algorithm and its necessary parameters. The second timeframe applies to the time-dependent models. Here we test on a daily basis, also starting January 1st, 2004. From each participant we daily request a forecast matrix definied below. The tests will be performed with the preliminary catalog as soon as all events of each particular day are processed and the catalog is made available. We want to encourage each particpant to deposit the algorithm used for creating the forecasts, so that additional replays with evaluated catalog data are possible.

In the RELM project we express a hypothesis as a forecast, which we define as a vector of earthquake rates corresponding to the specified bins. Any bin is defined by intervals of the location, time, magnitude, and focal mechanism, thus a multi-dimensional interval. The resolution of a model 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", or expected number of events within the time interval for all bins, each element of the vector corresponding to a particular bin. The expected number is just 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 they correspond directly to earthquake rates per unit area, magnitude, time, and possibly depth and orientation of angles because the bin sizes are specified. Forecasts may define a rate of 0 events for a particular bin, e. g. for magnitudes greater than a given threshold. Although this is possible, we recommend to avoid such forecasts. While testing, the hypothesis gains if no earthquakes occur in bins with a forecasted rate of 0 events but they loose completely in case an event occurs. Any number higher than 0 avoids this problem because the probabilities for any number of events are greater than 0. In case of a forecast of 0, the probabilities collapse and no other outcome than 0 events is allowed for this forecast. We want to encourage every modeller to avoid forecasted rates of 0.

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 meaning 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 will avoid using the term prediction to avoid confusion. None of the forecasts are predictions in either sense nor are they ment to be. It is all about scientific testing.

The vector of expectations has to be compared with the vector of observations, based on the same binning, to score a given forecast. The observed number of events must be integers, and for the tests envisioned here they will usually be 0 or 1.

A useful measure of the agreement between a hypothesis and an earthquake record is the likelihood, defined as the joint probability of realizing the observed number of events, given the expectations in each of the bins. By joint probability we mean the probability of realizing the observed number in bin 1 and bin 2, etc. In all of the models proposed to date, the expectations for the various bins are assumed to be independent, in which case the likelihood is the product of the probabilities. The logarithm of the joint probability, sometimes called the "log-likelihood" or "log-likelihood score," is simply the sum of the logs of the probabilities for all bins. If the expectations are not independent, the joint probability can be calculated as a product of conditional probabilities.

By comparing the observed events to a model's expectations, we derive the hypothetical probability of the observed events occurring in our model. This probability is called the likelihood and is calculated assuming a Poissonian distribution of events in each bin. The Poisson model is strictly valid only if the forecast rate is truly constant during the test interval, but it is a good approximation when the rates do not vary much within the time interval.

The log-likelihood score depends on both the earthquake record and the forecast rates, and higher values imply better agreement between the two. But how large is large enough? We answer this question with two comparisons. First, in what we call the "self-consistency test", we compare the observed likelihood score with its expected value, assuming that the hypothesis is true. Second, in the "relative consistency test", we compare the observed likelihood with the value obtain using the same data, but forecast probabilities from another hypothesis. In this project, we will compare likelihood scores from all pairs of hypotheses defined on the same bins.

Besides these strict definitions on how to test forecasts against each other to match the defined needs (quasi-static and short-term models), we want this procedure to be the skeleton for any tests, modelers envision. We only want to set requirements on additional or changed rules: The test needs overall expectations of enough earthquakes to make it meaningful. If the chance for even only one earthquake to occur is very low, this test cannot be carried out with a meaningful result because there is no chance that this test can be performed on a long enough time period. Also, we only want to accept expectation based on reliable data. Therefore, tests on magnitudes far below the completeness level do not make any sense. The last requirement is the use of meaningful objective data. This means, we only want to allow data that is published on a regular basis based on unambigious definitions.

3.1 Definitions

Expectation The forecasted number of earthquakes for any given bin, 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 of numerical estimates of the expected number of earthquakes in each bin, based on a hypothesis.

Bin A bin 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 1 and ω_2 events in bin 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 We propose two methods to test a model's forecast:

- Null hypothesis test. In this test each model (test hypothesis) competes against a given null hypothesis. We simulate earthquake rupture catalogs and follow a similar method to the standard approach used in likelihood ratio testing.

- Tournament test. For this test we test each model against all other models. The procedure is similar to the null hypothesis test however each model acts as both a null and a test hypothesis in two tests against every other model of its category. This is necessary because it is possible that all tests between a RELM model and the null hypothesis will result in rejection of the null. However, significance between two competing RELM models 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.

Results Null hypothesis test: If the null hypothesis can be rejected at the predefined significance level α, the test hypothesis becomes the new null hypothesis in this process. Otherwise the test hypothesis is rejected. In the tournament test we tabulate the significance level from each test to derive a ranking for each model against all others.

4 Computation

As outlined above, any 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, thus defining the resolution of a forecast. We denote bins with b and all bins constitute the set $\mathcal{B}$ defined as

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

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

Every forecast is issued as an expectation λ_i per bin b_i. We set up a vector Λ of all expectations as

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

Expectations have units of earthquakes per year for quasi-static, and earthquakes per day for short-term forecasts. We also set up 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 probability 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

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

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

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

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

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

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 probabilities 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. Again, it is a special case if a model supplies forecasts with cells containing $\lambda = 0$. As long as no events occur in these cells, the model is getting judged as any other model. If an event occur in a cells with expectation $\lambda = 0$, the joint log-likelihood sums up to $-\infty$. Thus, the model is rejected.

4.1 Simulation and Evaluation

How can we know the expected value of the likelihood? Furthermore, 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. The likelihood score is a statistic (i.e., a quantity measurable from any sample of the underlying forecast distribution), so it has its own probability distribution. In some simple cases the distribution of likelihood scores 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 $\hat{\Omega}_k$ (simulated values are denoted with a hat) consistent with the forecast. Then we compute the likelihood score $\hat{L}_k$ for each of these simulated records, and compare them with the likelihood score L for the observed record. From this distribution $\mathcal{L} = \{\hat{L}_1, \hat{L}_2, \ldots, \hat{L}_n\}$, we then can compute the significance as quantiles of the observed value compared to the distribution of simulated values.

For simulating values of our test statistic, we first need to define on which set of expectation this simulations should be based on, either model H^0 or H^1. 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 to a simulated number of observed events $\hat{\omega}_i$ for each given bin b_i. Iterating through all bins creates a vector of simulated events $\hat{\Omega}^j$ based on model H^j.

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

We will denote multiple simulated vectors with $\hat{\Omega}_1^j, \hat{\Omega}_2^j, \ldots, \hat{\Omega}_m^j$. The subscript of $\hat{\Omega}$ is the number of the simulation. Again, the case of $\lambda = 0$ requieres a special treatment. In this case the corresponding $\hat{\omega}$ will always be 0.

4.1.1 Data-consistency test

Consider first the data-consistency test, and assume that the test hypothesis is true. This is showing whether the observed likelihood of the test hypothesis is consistent with likelihoods obtained from simulations. A useful measure for this comparison is the quantile score γ, or the fraction of simulated likelihood values $\mathcal{L} = \{\hat{L}_1, \hat{L}_2, \ldots, \hat{L}_n\}$, $n = |\mathcal{L}|$ less than the observed likelihood L

$$\gamma = \frac{|\{\hat{L}_k | \hat{L}_k \le L, \hat{L}_k \in \mathcal{L}\}|}{|\mathcal{L}|}$$

Here $\hat{L}_k$ denotes the likelihood of the k-th simulation. If γ 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 it looks like it ought to, according to this one measure. A problem arises when considering results with a high γ. It means that the likelihood of the real observation is higher than the likelihood scores of the simulations. There are different scenarios under which this can happen. In a catalog with overall very low expectations for any event to happen, the outcome of 0 events is the most likely one. Nevertheless, the sum of all given rates may exceed 1 or even higher numbers, expecting in total some events to occur. In this case, the outcome of 0 events would show a much higher likelihood than the average simulation because the simulation will reflect the total number of expected earthquakes, distributed over the cells. In contrast, a forecast with expectations exactly matching the observations would also have a too high likelihood compared the the likelihood scores of the simulations, because every simulation will in general have different numbers of events per cell than the forecast itself. This will result in lower likelihoods for the simulations (high γ). As can been seen, a model should not be rejected based on too high likelihoods in the data-consistency test. We want to use this test only as an one-sided test, rejecting forecasts with significanctly too low likelihood compared to the simulations. Matching or too high likelihoods are not giving a measure of goodness of match of expectations with observations.

4.1.2 Introduction of possible test scenarios

In many works (e.g. Jackson and Kagan, 1994), a "test hypothesis" is compared to a "null hypothesis." The null hypothesis is presumed simpler and the test hypothesis is only accepted if an observed statistic would be very improbable 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.

In many cases the null hypothesis is similar to the test hypothesis except that it is missing one or two interesting features. Then rejecting the null hypothesis is equivalent to stating that a model must have the "interesting features" to fit the data well. A special case is when the null hypothesis is a constrained version of the test hypothesis. Both hypotheses are adjusted to fit the data, but the null hypothesis, being constrained, can fit no better than the test hypothesis, and usually it fits worse. Even if the null hypothesis were true, the test hypothesis would generally fit better, and the test measures whether that improvement is too large to be consistent with the null hypothesis. Again, the test is based on the distribution of a relevant test statistic under the null hypothesis.

It has to be mentioned here that additional parameters in a model do not correspond to additional degrees of freedom of this model, making the use of the Akaike Information Criterion (AIC) or any other related method (e. g. AICc, BIC, TIC, etc.) impossible. In these methods, the models are judged based on their likelihoods but also on the number of degrees of freedom. In all models tested in the RELM framework, the number of degrees of freedom are 0 because every forecast is issued in advance of the observation period and is not readjusted during the observation period.

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. For this reason we test each hypothesis against all others in pairs, letting one and then the other serve as the null hypothesis.

Consider for example hypotheses H^1 and H^2. First, we let the hypothesis H^1

play the role of the null hypothesis. We define a test statistic for which positive values favor hypothesis H^2, evaluate it for the observed data, and then determine by simulation the probability that the observed value would be exceeded if hypothesis H^1 were true. If that probability is low, then the hypothesis H^1 is provisionally rejected. For the test statistic, we use the log likelihood ratio of hypotheis H^2 over hypothesis H^1. We then reverse the order, negating the test statistic so that positive values favor hypothesis H^1, and determine the probability that the observed value would be exceeded if hypotheses H^2 were true. Finally we consider as the preferred model that which has the lower rejection probability when it serves as the null hypothesis.

4.1.3 Test of test hypothesis vs. null hypothesis

In the case where we want to test a (challeging) test hypothesis H^1 against a null hypothesis H^0, the test model has to reject the null hypothesis at a predifined significance level α. Usually, this level is set to $\alpha = 0.05$ or $\alpha = 0.01$. To obtain the significance level of a test model, we have to simulate observations based on the expectations of the null hypothesis Λ^0 and score both models according to the simulated observations $\hat{\Omega}^0_k$. Since this test is based on the log-likelihood-ratio, we first compute the log-likelihood-ratio R based on the true observations:

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

Given the vectors of simulated observations $\hat{\Omega}^0_k$ and the vector of expectations we can compute log-likelihood-ratios for every simulation run k. This leads to a set of log-likelihood-ratios $\mathcal{R} = \{\hat{R}_1, \hat{R}_2, \ldots, \hat{R}_m\}$ of $m = |\mathcal{R}|$ simulated observations by

$$\hat{R}_k = L(\hat{\Omega}^0_k|\Lambda^0) - L(\hat{\Omega}^0_k|\Lambda^1)$$

The percentile of the log-likelihood-ratio of the true observation R in respect to the log-likelihood-ratios of the simulated observations $\hat{R}_k$ yields the significance level α for rejecting the null hypothesis in favor of the challenging hypothesis.

$$\alpha = \frac{|\{\hat{R}_k|\hat{R}_k \leq R, \hat{R}_k \in \mathcal{R}\}|}{|\mathcal{R}|}$$

+ Interpretation for the test vs. null hypothesis

4.1.4 Hypotheses comparison

The second kind of tests covers comparisons of models (tournament test). In this kind of testing we cannot test any hypothesis against a null hypothesis and repeat this test subsequently with all hypotheses. It is most likely that the first hypothesis to test against a 'dumb' null hypothesis will win this test and the null hypothesis get rejected in favor of the tested hypothesis. Unfortunately, it will be also very 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. In the test hypothesis vs. null hypothesis test we used a simple likelihood ratio

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

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

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 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 $\mathcal{R}^{21} = \{\hat{R}_1^{21}, \hat{R}_2^{21}, \ldots, \hat{R}_m^{21}\}$ with

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

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

$$\alpha_{21} = \frac{|\{\hat{R}_k^{21}|\hat{R}_k^{21} \leq R^{21}, \hat{R}_k^{21} \in \mathcal{R}^{21}\}|}{|\mathcal{R}^{21}|}$$

So far we have focussed on H^2, but we should of course focus on H^1 as well. We derive the distribution of 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 want to use the result matrix or result table containing all computed α-values. Consider a test run with n hypothesis 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 and the result table, where the 'null hypotheses' are displayed on the left side and the 'test hypotheses' on top.

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

+ Add more here about judging on the models

5 Examples

6 Definitions

Although we do not wish to enforce any rules that impede how a model generates a forecast, it is necessary to define several rules so that we are able to compare models. We define initial bin sizes and grids for the following variables: location, time, magnitude, and focal mechanism. The bins are not limited to the predefinition, however, they must be a multiple of the default as defined in the algorithms below.

Test area We define the test area in southern California as the following cells:

Latitude range		Longitude range	
Min	Max	Min	Max
36	37	-118	-117
36	37	-117	-116
35	36	-120	-119
35	36	-119	-118
35	36	-118	-117
35	36	-117	-116
35	36	-116	-115
34	35	-121	-120
34	35	-120	-119
34	35	-119	-118
34	35	-118	-117
34	35	-117	-116
34	35	-116	-115
34	35	-115	-114
33	34	-120	-119
33	34	-119	-118
33	34	-118	-117
33	34	-117	-116
33	34	-116	-115
33	34	-115	-114
32	33	-119	-118
32	33	-118	-117
32	33	-117	-116
32	33	-116	-115
32	33	-115	-114

Grid The starting grid uses the above southern California definition with nodes centered at every whole degree. It is important to note that in a likelihood ratio test a coarser grid forecast can be resampled to a finer grid without changing the results (see Appendix). However, it is required that finer cells do not overlap cells of a coarser grid. Therefore, any modeler using higher resolution must divide each bin into 100 new equally spaced bins and so on. Equivalently, any modeler using a lower resolution must resample the results to the minimum 1° by 1° resolution.

Depth Depth binning is implemented, however, the default is no binning. If necessary, bins are defined to be 10 km, 1 km, 0.1 km, etc. This test is considering only earthquakes with depths between 0 km and 30 km.

Magnitude range We require that all models provide a forecast of events between an M_{min} and a bin containing the expectation for all earthquakes with magnitudes $M \geq M_{max}$. For quasi-static models, $M_{min} = 5$, and for time-dependent models, $M_{min} = 4$. For both types $M_{max} = 9$. The default binning is 0.01 units. It is allowed to use finer magnitude bins if necessary. In this case, the resolution in bins should be increased by a factor of 10.

Focal mechanisms Focal mechanisms are defined by 3 angles, dip direction, dip angle, and rake. Our initial binning of these angles is 30 degrees for each angle. The next binning step is 10 degrees for each angle. For further higher resolution we propose a factor of 1/10.

Time bins Though we make no clear distinction between a time-dependent and quasi-static models, we define the following time bins: 5-yearly, yearly, monthly, daily, hourly, minutely.

7 Appendix

7.1 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:

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

Since

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

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 $logx!$ 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.

8 Fileformat for forecasts

Any model forecast should be issued in this forecast definition fileformat. This format is plain 7-bit ASCII to ensure system interoperability.

9 Comparison with other test methods

References

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