Theme VI – Earthquake Predictability & Related Hypothesis Testing

Evaluating earthquake predictions and earthquake forecasts: a guide for students and new researchers

J. Douglas Zechar

Swiss Seismological Service, ETH Zurich

How to cite this article:

Zechar, J.D. (2010), Evaluating earthquake predictions and earthquake forecasts: a guide for students and new researchers, Community Online Resource for Statistical Seismicity Analysis, doi:10.5078/corssa-77337879. Available at http://www.corssa.org.

Document Information:

Issue date: 1 September 2010 Version: 1.0

Contents

Abstract In this article, I review various methods for evaluating earthquake predictions and earthquake forecasts. In many situations, you can use one or more of these methods to obtain quantitative answers to the questions: “Is this set of earthquake predictions particularly good, or could the same success be obtained by chance?”; “Is the observed record of earthquakes consistent with this earthquake forecast?”; and “Which of these two forecasts is better?” I note the primary advantages and disadvantages of each evaluation metric and describe them in general terms and with specific examples or constraints (e.g., Poisson space-rate-magnitude forecasts). Accompanying this article are source code, compiled code, and data examples, which you can use to understand the mechanics of each metric. You can modify the source code to include additional tests or to introduce other functionality. I also provide references to further reading.

1 Motivation

One of the cornerstones of science is the ability to accurately and reliably forecast natural phenomena. Unfortunately, earthquake prediction research has been plagued by controversy, and it remains an outstanding problem; for a review of some of the historical challenges, see the article by *Wyss and Michael* (in preparation) or the book by *Hough* (2009). The motivation for the work that I describe in this article is fairly self-evident: we want to know if an earthquake forecast or a set of earthquake predictions is particularly “good.” Therefore, our fundamental objectives are to define and to quantify “good.”

In this article, I emphasize the analysis of statements regarding future earthquake occurrence (i.e., characteristics such as origin time, epicenter, and magnitude) but many of the concepts discussed are applicable to other earthquake studies (i.e., probabilistic loss estimates, earthquake early warning, etc.). A broader motivation of this article is to encourage you to exercise rigorous hypothesis testing methods whenever the research problem allows.

2 Starting Point

You should have a basic understanding of probability distributions, particularly how to sample a distribution using a random number generator; see articles in Theme III of the Community Online Resource for Statistical Seismicity Analysis (CORSSA) for details. You should also have a basic understanding of seismicity catalogs. See the article by *Woessner et al.* (in preparation) for details.

What is an earthquake prediction? For the purposes of this article, a well-defined earthquake prediction is a specification of a latitude-longitude-time-magnitude range, including the magnitude scale (i.e., local magnitude, moment magnitude, etc.) and the number of earthquakes expected in this range (i.e., zero, one, at least one, etc.). Because magnitude estimates may vary from catalog to catalog, the appropriate catalog for evaluating the prediction should also be specified. An earthquake prediction may optionally include probability of success, depth range, focal mechanism, or some other measurable characteristic. What is essential is that, after the end time of the prediction, one must be able to determine objectively whether or not the prediction was successful. Making this determination should not require any interpretation. To state the preceding in more general terms, the prediction statement must be unambiguously falsifiable.

An example of a well-defined earthquake prediction is: I expect at least one earthquake between 1 March 2011 and 11 March 2011 with epicenter(s) in the latitude-longitude range 30N to 34N, 118W to 114W; qualifying earthquakes will have moment magnitude between 5.5 and 6.2 as reported in the Advanced National Seismic System earthquake catalog. Because a prediction may be correct or incorrect by "luck," we learn very little by evaluating a single prediction. Therefore, we are most interested in evaluating sets of several predictions.

I note that this working definition of an earthquake prediction differs from some preceding definitions that required that a confidence level (or probability) be assigned to each prediction (e.g., *Allen et al.* 1976). It also differs from that of *Jackson* (1996), who suggested that predictions are a special form of forecasts for which the probability of earthquake occurrence is "temporally higher than normal." In this article, these distinctions are not terribly important, and I discuss evaluation methods that are applicable to binary earthquake predictions, probabilistic earthquake predictions, and a variety of earthquake forecasts.

What is an earthquake forecast? For the purposes of this article, I emphasize discrete forecast experiments in which the geographical region of interest is subdivided into non-overlapping cells defined by a range of latitude and longitude (and, optionally, depth). For example, you might consider the state of California and subdivide the state into rectangular cells of 0.01 square degrees. The magnitude range of interest and/or the time period of interest may likewise be subdivided. In other words, the experiment space is gridded into non-overlapping "bins." In this context, the most general type of earthquake forecast is a ranking of the bins according to their expected probability to host one or more earthquakes. This ranking can be explicit—e.g., the forecaster might assign a rank to each bin and designate this ranking as the forecast. The ranking can also be implicit—e.g., the forecast might specify the expected number of earthquakes in each bin (along with an estimate of forecast uncertainty). In the latter case, you'd derive the ranking by sorting the bins according to the expected number of earthquakes.

3 Ending Point

The techniques described in this article will allow you to quantify the predictive skill of an earthquake forecast or of a set of earthquake predictions. You will be able to check if an observed set of earthquakes is consistent with a forecast, and you will have some tools to compare two forecasts. Using the accompanying code and example data, you can execute each of the test methods described in this article (see section 6).

It is important to note that you will not be able to "validate" or "invalidate" an earthquake forecast model or prediction algorithm. A set of earthquake predictability experiments may lend support to a particular hypothesis, or they may suggest evidence contrary to the hypothesis. Strictly speaking, the outcome of a particular experiment only tells us something about that particular experiment; it does not necessarily tell us how well a forecast may fare somewhere else in space, time, or magnitude. For example, a forecast may be deemed consistent with the observations of one experiment and then, owing to variation in seismicity or random systematic effects, it may be shown to be inconsistent with observations in another experiment. I include this disclaimer to remind you to be careful in reporting the outcome of a prediction experiment.

4 Theory

In this section, I briefly and abstractly review some specific concepts that are used in explaining the methods in the next section.

4.1 Normalization

To normalize a set, sum its elements and divide each element by this sum. For example, if the set is composed of elements 61, 37, and 102, the elements of the resulting normalized set are 0.305, 0.185, and 0.51.

4.2 Poisson distribution

The Poisson distribution is a discrete probability distribution that describes a Poisson process, in which the probability of an event occurring is independent of the time since the previous event, and events occur at an average rate λ. The probability that ω events will occur in a given time period is

$$\Pr(\omega \mid \lambda) = \frac{\lambda^{\omega}}{\omega!}\exp(-\lambda), \tag{1}$$

which in this article is called the Poisson likelihood of ω given λ. The Poisson likelihood is only defined for non-negative integers ω, while λ can be any non-negative number. For the Poisson process with average rate $\lambda = 11.1$, you can check that the likelihood of $\omega = 4$ is less than 1%. The Poisson cumulative distribution function gives the probability that at least ω events will occur and is defined:

$$F(\omega \mid \lambda) = \exp(-\lambda)\sum_{i=0}^{\omega}\frac{\lambda^{i}}{i!}. \tag{2}$$

For $\lambda = 11.1$, $\omega = 4$, $F(\omega|\lambda) = 1.4\%$. The Poisson inverse cumulative distribution $F^{-1}(p \mid \lambda)$ gives the smallest value of ω such that the Poisson cumulative distribution function with parameter value λ evaluated at ω equals or exceeds p, the probability of interest. (This is implemented as the `poissinv` function in MATLAB®, and a simple implementation is given in the accompanying code as `org.scec.predictionTesting.MathUtil.inverseCumulativePoisson`.) For $\lambda = 11.1$, $p = 5\%$, you can verify that $F^{-1}(\omega|\lambda) = 6$.

4.3 Poisson joint likelihood

If we consider n independent Poisson processes characterized by their respective average rates $\mathbf{\Lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_n\}$, the probability that $\mathbf{\Omega} = \{\omega_1, \omega_2, \ldots, \omega_n\}$ events, respectively, will occur is

$$\Pr(\omega_1 \mid \lambda_1)\Pr(\omega_2 \mid \lambda_2)\ldots\Pr(\omega_n \mid \lambda_n) = \prod_{i=1}^{n}\left[\frac{\lambda_i^{\omega_i}}{\omega_i!}\exp(-\lambda_i)\right], \quad (3)$$

which in this article is called the Poisson joint likelihood of $\mathbf{\Omega}$ given $\mathbf{\Lambda}$. For example, if $\mathbf{\Lambda} = \{1.2, 6.4, 3.7\}$, the likelihood of $\mathbf{\Omega} = \{1, 6, 3\}$ is (36% x 16% x 21%) = 1.2%.

4.3 Alarm function

A binned earthquake rate forecast is one that specifies the expected number of earthquakes in each forecast bin. These numeric forecasts can be converted to simple Yes/No predictions by selecting a threshold: an earthquake is expected in any bin with a rate above the threshold and earthquakes are not expected in bins with rates lower than the threshold. The resulting construct is an alarm set, and many such alarm sets can be produced by varying the threshold value. Therefore, I say that the rate forecast provides an *alarm function*. More generally, an alarm function is any forecast construct that provides a ranking from which alarm sets can be so derived. A conceptual example is shown in Fig. 1.

If the ranking is given in terms of probability—e.g., 10% of target earthquakes will fall in bin A, 70% will occur in bin B, and 20% will fall in bin C—I call this a *probabilistic alarm function*.

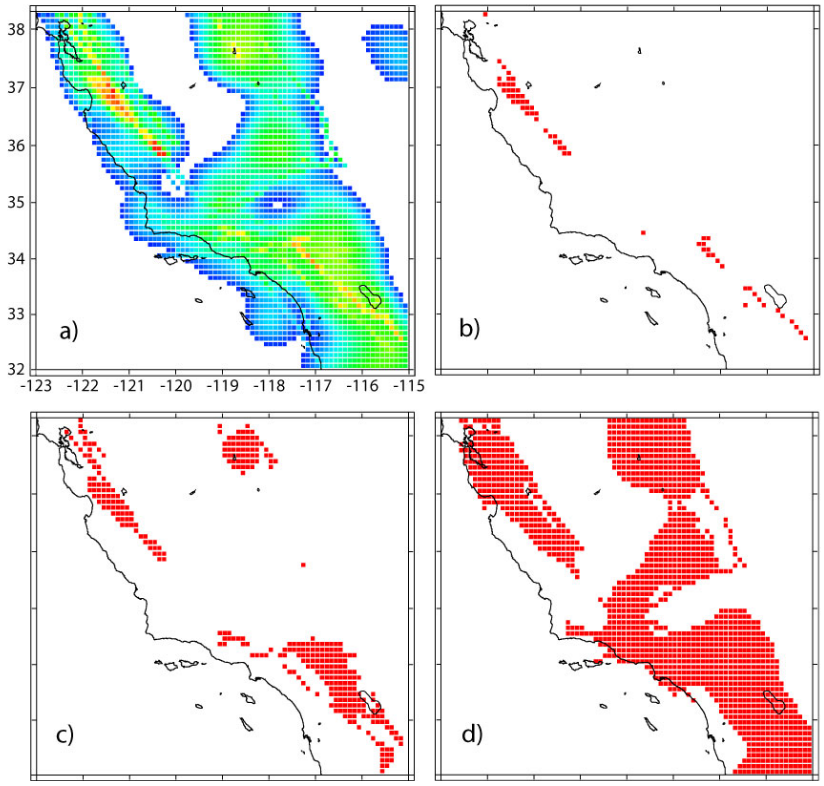

Fig. 1 Example alarm function in a) is a gridded ranking of bins in California from the USGS 2002 National Seismic Hazard Map rate model, where "warm" colors have a higher ranking than "cool" colors. When a high threshold value is chosen, the resulting alarm set is depicted in b), where red bins indicate a "Yes" prediction and all others are "No." As the threshold value is decreased, the region of "Yes" predictions continues to grow, as shown in c) and d), where medium and low thresholds, respectively, were chosen. (Modified from *Zechar and Jordan* 2008)

4.4 Simulating an observation consistent with a forecast

Many of the methods described in section 5.2 require simulated observations that are "consistent with a forecast." Consider a probabilistic alarm function with four bins in which 1.2, 3.7, 0.1, and 5 earthquakes are expected. Upon normalization, this can be interpreted as a forecast statement that the bins will host 12%, 37%, 1%, and 50% of the earthquakes, respectively. (The percentage for each bin is the number of earthquakes expected in the bin divided by the number of earthquakes expected in all bins.) The goal is to simulate observations where each simulated earthquake has a 12% chance of occurring in the first bin, a 37% of occurring in the second bin, a 1% chance of occurring in the third bin, and a 50% chance of occurring in the fourth bin.

To do this, you can construct another discrete distribution that maps random numbers to bins. The mapping distribution has the same number of bins as the forecast and is constructed by normalizing and summing the rates in each forecast bin; the bins in the mapping distribution are assigned the cumulative normalized sums. For example, the mapping bin values for the example being considered are 0.12, 0.49, 0.5, and 1. Then, simulate an earthquake by sampling the uniform distribution between 0 and 1; the simulated earthquake is placed in the bin that corresponds to the first bin in the mapping distribution which exceeds the random sample value. If the sample is 0.37, the simulated earthquake belongs to the second bin. If the sample is 0.69, it belongs to the fourth bin.

For this simulation algorithm, it is not a requirement that the forecast be a rate forecast (like I used in this example); any forecast with a probabilistic alarm function will suffice. Note that this algorithm does not necessarily produce catalogs that are physically realistic; for example, it operates under the assumption that the forecast in each bin is independent of the forecast in every other bin. Rather, this algorithm gives an idea of what seismicity would look like if a particular forecast was the correct model of seismicity.

4.5 Notation

Table 1 summarizes some of the notation used in the following sections.

Notation	Meaning	Example
$\binom{n}{k} = \frac{n!}{k!(n-k)!}$	the number of subsets with k elements that can be formed from a set with n elements	$\binom{5}{2} = \frac{5!}{2!(3)!} = \frac{5 \cdot 4}{2} = 10$
$\lvert\{A\}\rvert$	the number of elements in set A	If $A = \{a_1, a_2, \ldots, a_n\}$, $\lvert\{A\}\rvert = n$.
$\{a_i \mid a_i < b, a_i \in A\}$	the subset of elements in set A that are smaller than b	If $A = \{3,7,8,4,12\}$, b=8, the resulting subset is $\{3,7,4\}$.
$\lvert\{a_i \mid a_i < b, a_i \in A\}\rvert$	the number of elements in set A that are smaller than b	If $A = \{3,7,8,4,12\}$, b=8, the result is 3.
$[a,b]$	the range between a and b, inclusive	Considering the integers, $[4,7] = \{4, 5, 6, 7\}$

5 Available Methods

5.1 Given a set of earthquake predictions

In the simplest case, you will have a set of "Yes" or "No" earthquake predictions and a corresponding set of observations. To begin, note which predictions were successful and which were not. Call each earthquake that occurred within a "Yes" prediction a *hit*, and each earthquake that occurred within a "No" prediction a *miss*. Each "Yes" prediction in which no corresponding earthquake occurred is a *false alarm*, and each "No" prediction in which no earthquake occurred is a *correct negative*. You can construct a table with the number of each of these contingencies (e.g., Table 2). There are several metrics that depend on the number of hits and/or misses and/or false alarms and/or correct negatives; such metrics are typically said to be based on the *contingency table*. For example, you can compute the miss rate—the fraction of target earthquakes that were misses—or the false alarm rate—the fraction of "Yes" predictions that were false alarms. The trouble with using any of these measures in isolation is that they can be optimized using a simple strategy: to obtain a zero miss rate, one can specify a prediction that covers the entire experiment space-time. To address this deficiency, these metrics are often used in pairs.

		Prediction	
		Yes	No
Occurrence	Yes	Hit	Miss
	No	False alarm	Correct Negative

Table 2 Example binary prediction/binary outcome contingency table that denotes how the four contingencies are tabulated.

5.1.1 Receiver Operating Characteristic

For evaluating earthquake predictions, one popular pair of contingency table metrics is the false alarm rate, F, and the hit rate, H (which is $1 -$ miss rate); when these values are plotted together on the square [0,1] x [0,1], the resulting metric is called the *Receiver Operating Characteristic* (ROC) (*Mason* 2003, and references therein). Each set of "Yes" or "No" predictions corresponds to a single point on the ROC. To say if this set of predictions is skillful, you can compare its ROC point with the diagonal line connecting the (H, F) points (0, 0) and (1, 1)—this diagonal represents the long-term behavior of random guessing.

There are a number of metrics to establish the significance of the distance from the diagonal, but I won't mention them here because the ROC suffers from a serious problem when applied to earthquake predictions: it does not account for the fact that earthquakes are clustered in space. In particular, the implicit reference model used in ROC, what I called "random guessing," assumes that earthquakes are equally likely to occur anywhere in space, which you know from first-order observations is unrealistic. This makes the ROC a very weak tool for evaluating earthquake predictions, and it should be avoided when considering earthquake predictions with a spatial component. Indeed, many of the metrics based on the contingency table suffer the same disadvantage.

Pros: simple to compute, simple to interpret
Cons: unrealistic reference model, easy to achieve statistically significant results

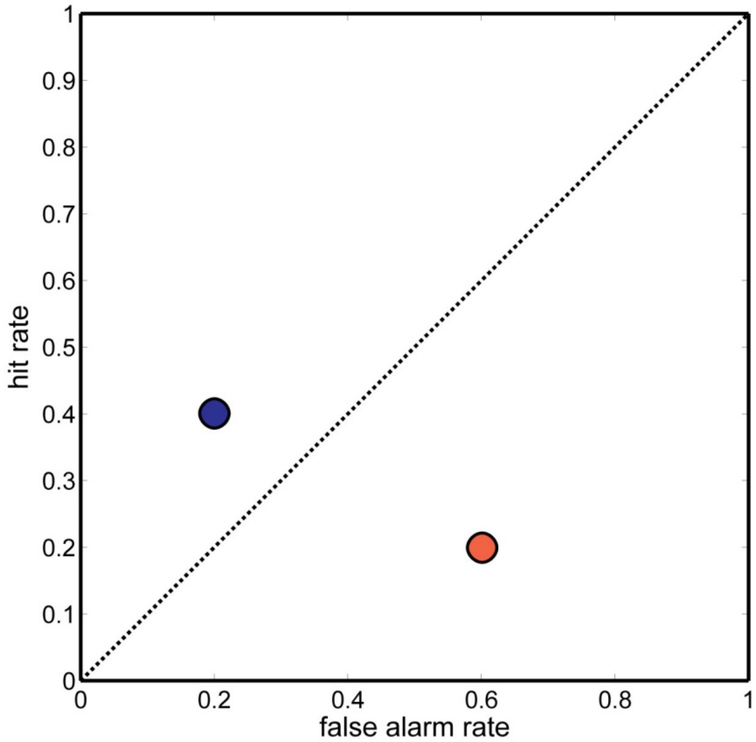

Fig. 2 Example ROC diagram indicating performance of two imaginary sets of predictions. Because the point for Alarm Set A is above the diagonal, it is suggested that the predictions are better than random guessing. Point B indicates a set of predictions that are much worse than random guessing.

5.1.2 The Molchan diagram

Closely related to the ROC is the *Molchan diagram*, which is a plot of the miss rate, υ, and the fraction of experiment space-time volume, τ, occupied by "alarms," or Yes predictions (*Molchan* 1991, *Molchan and Kagan* 1992). This second metric is what distinguishes the Molchan diagram from the ROC, and it corrects for the reference model problem mentioned above. Indeed, the Molchan diagram allows for any reference model to define the fraction of "space" occupied by alarms. Typically, "space" in this context is not geographical space (i.e., square kilometers or square degrees), but rather the reference model's probability estimate (i.e., what is the probability that an earthquake will occur in this particular space?). For most space-time predictions, the appropriate reference model is based on previous seismicity, representing the parsimonious hypothesis that future earthquakes are most likely to occur where they occurred in the past (where our knowledge of the past is typically limited by the availability of reliable data). For details on this hypothesis, see the section on smoothed seismicity in the CORSSA article by Werner *et al.* (in preparation).

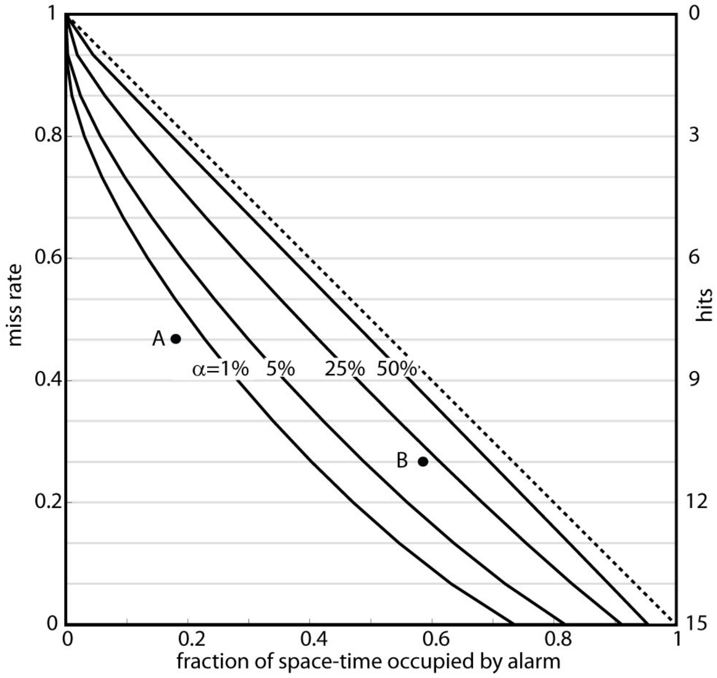

Fig. 3 Molchan diagram confidence bounds computed by solving eq. 4 for all values of h, with $N = 15$ imaginary target earthquakes. The curves are contours for $\alpha = \{1, 5, 25, 50\}\%$. Here, point A represents an imaginary alarm set that has obtained 8 hits; this point indicates that the alarm set is skillful with more than 99% confidence. The point B (11 hits from another imaginary alarm set) supports a statement of significant skill at just above 75% confidence. (Modified from *Zechar and Jordan* 2008)

In practice, a smoothed seismicity reference forecast is analogous to the real estate market, where popular locations have a higher value. A smoothed seismicity reference emphasizes regions that historically have high seismicity rates; if you make a positive prediction in a place where the seismicity rate has been high, this costs more than declaring an alarm in a region with low seismic activity. If, for example, a binary alarm set has elements 0, 1, 1, 0 and the normalized reference forecast is 0.12, 0.37, 0.01, and 0.50, $\tau = 0(0.12) + 1(0.37) + 1(0.01) + 0(0.50) = 0.38$. As with the ROC, a single alarm set corresponds to a single point on the Molchan diagram. Ideal predictions obtain points near the origin, which corresponds to maximum success ($\upsilon \rightarrow 0$) at minimum cost ($\tau \rightarrow 0$).

For a given value of υ, you can determine the statistical distribution of τ, and therefore the statistical significance of a given point on the Molchan diagram. In particular, the probability of obtaining h or more hits by chance, given that there have been N observed target earthquakes, is described by the binomial distribution (see Fig. 3):

$$\alpha = \sum_{i=h}^{N} \left[\binom{N}{i} \tau^{i} (1-\tau)^{N-i} \right]. \qquad (4)$$

Therefore, a very small α value suggests that the alarm set has high skill. *Molchan and Romashkova* (in review) have also suggested using the metric $(1 - \tau - \upsilon)$ to characterize the skill of an alarm set. In the situation where you have access only to a set of predictions, this is a reasonable way of evaluating their skill, particularly if you can form a reasonable corresponding reference model. However, in the situation where you have an alarm function and many derived alarm sets, you might obtain contradictory results: you may find that some points indicate high significance while others do not. This creates a difficulty in saying something useful regarding the predictive skill of the alarm function (as opposed to the skill of a particular alarm set).

Pros: simple to compute, allows you to specify reference model

Cons: measure of "space" is not unique and may be confusing, can yield ambiguous results for an alarm function

5.1.3 The area skill score

The area skill score directly addresses the issue of ambiguous Molchan diagram results. Such ambiguity may arise when deriving every possible alarm set from a given alarm function. By lowering the threshold value from the maximum forecast value to the minimum, you trace out a *Molchan trajectory*, which characterizes the skill of the entire alarm function. The area skill score is defined as the normalized area above the Molchan trajectory—the score varies between zero and one, and higher values are preferable. For the simplest cases, the distribution of the area skill score has relatively straightforward analytical solutions (see *Zechar and Jordan* 2010). For any reference model and/or clustered observations (i.e., more than one target earthquake in a given bin), you can estimate the distribution of the area skill score by simulating many random forecasts. The accompanying code contains one such implementation. To determine the significance of a given area skill score, you should compare that score with the distribution of scores under the reference model.

Pros: removes ambiguity of multiple Molchan trajectory points, allows straightforward hypothesis test

Cons: may result in loss of information (because vector-valued Molchan trajectory is reduced to a scalar value)

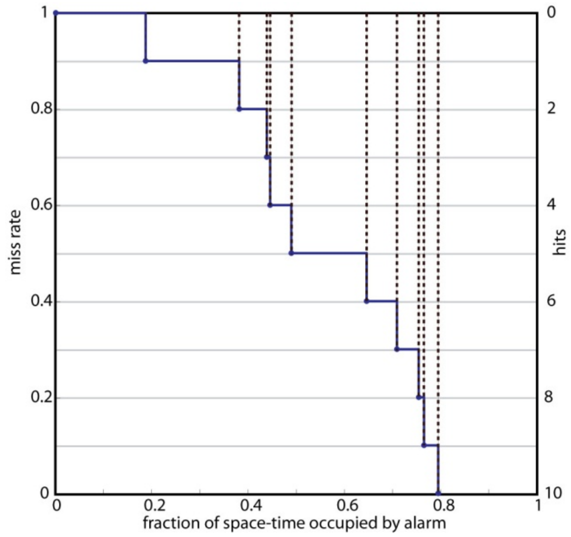

Fig. 4 Molchan diagram with hypothetical Molchan trajectory (blue line) and dashed lines showing how to compute the area skill score: sum the columns, which grow in height by $1/N$ as you move from $\tau = 0$ to $\tau = 1$. (After *Zechar and Jordan* 2010)

5.1.4 The gambling score

The previous two metrics, while correcting the spatially uniform reference model flaw of the ROC, depend on the ability to define misses. There is at least one situation in which this is not feasible: when an individual or algorithm produces only "Yes" predictions and covers the rest of the experiment space with statements of "No comment." In this situation, there are no "No" prediction statements and therefore the Molchan diagram, the area skill score, and many other contingency table measures are not informative. *Zhuang* (2010) introduced a generalized score that is analogous to gambling and which is applicable to "Yes"-only alarm sets. Moreover, it is also applicable to both fully binary and probabilistic predictions.

To understand the gambling score, imagine each prediction as a forecaster betting one credit of reputation. For the i^{th} bet, if the forecaster bets "Yes," denote this by $x_i = 1$; otherwise $x_i = 0$. If an earthquake happened inside the specified prediction, denote this by $y_i = 1$; otherwise $y_i = 0$. To compute the gambling score, one must specify a reference probability for each prediction; denote the reference probability for the i^{th} forecast by p_i^0. For such a set of binary forecasts X, binary outcomes Y, and reference probabilities P^0, the net reputation gain is

$$\Delta R\left(X,Y,P^0\right)=\sum_i\left[x_i y_i\left(\frac{1-p_i^0}{p_i^0}\right)+\left(1-x_i\right)y_i\left(-1\right)+x_i\left(1-y_i\right)\left(-1\right)+\left(1-x_i\right)\left(1-y_i\right)\left(\frac{p_i^0}{1-p_i^0}\right)\right]. \quad (5)$$

Here, the sum is performed over all predictions. The four summands in eq. 5 correspond to the reputation gain from hits, misses, false alarms, and correct negatives, respectively. If the forecasts are probabilistic, you can think of the i^{th}

forecast as a bet of p_i on "Yes" and a bet of $(1 - p_i)$ on "No." In this situation, the net reputation gain for the set of probabilistic predictions P is

$$\Delta R\left(P, Y, P^0\right) = \sum_i \left[y_i \left(p_i \left(\frac{1 - p_i^0}{p_i^0} \right) - \left(1 - p_i\right) \right) + \left(1 - y_i\right) \left(-p_i + \left(1 - p_i\right) \left(\frac{p_i^0}{1 - p_i^0} \right) \right) \right]. \quad (6)$$

In both cases, if the net reputation gain is positive, this indicates that the predictions were superior to the reference model. *Zechar and Zhuang* (2010, section 5) described a simulation method to establish the statistical significance of an observed gambling score. In the accompanying code, you can execute these simulations using a spatially inhomogeneous Poisson process as the reference model.

Pros: widely applicable
Cons: may be time-consuming to construct appropriate reference model

5.2 Given one or more rate forecasts

Another common format for earthquake forecasts is a gridded rate forecast, a forecast in which the geographical region of interest is divided into sections and the forecast specifies the expected number of earthquakes in each section. This is the format that is widely used in the Collaboratory for the Study of Earthquake Predictability (CSEP) testing centers (*Jordan* 2006, *Zechar et al.* 2010b). In particular, I consider binned space-rate-magnitude forecasts. For this class of experiments, the testing region set, $\boldsymbol{R}$, is the Cartesian product of the binned magnitude range of interest, set $\boldsymbol{M}$, and the binned spatial domain of interest, set $\boldsymbol{S}$:

$$\mathbf{R} = \mathbf{M} \times \mathbf{S} \quad (7)$$

For example, in the Regional Earthquake Likelihood Models (RELM) experiment (*Schorlemmer and Gerstenberger* 2007, *Schorlemmer et al.* 2007), the magnitude range of interest is 4.95 and greater, and the bin size is 0.1 units, with the exception of the final, open-ended bin:

$$\boldsymbol{M} = \{[4.95, 5.05), [5.05, 5.15), \ldots, [8.85, 8.95), [8.95, \infty]\} \quad (8)$$

The RELM spatial domain of interest is a polygon enclosing California (polygon coordinates can be found in Table 2 of *Schorlemmer and Gerstenberger* 2007); this area is represented as a set of latitude/longitude rectangles of 0.1° x 0.1°.
Almost all current experiments in CSEP testing centers consider binned earthquake forecasts that incorporate the assumption that the number of earthquakes in each forecast bin is Poisson-distributed and independent of those in other bins. In this case, an earthquake forecast, Λ, on $\boldsymbol{R}$ is fully specified by the expected number of events in each magnitude-space bin. If you consider the magnitude-space bin indexed by ordered pair (i, j), you can denote the number of earthquakes forecast in this bin as $\lambda(i, j)$. Using this notation, the earthquake forecast can be written as the set of each bin's forecast:

$$\Lambda = \left\{ \lambda(i, j) \mid i \in \boldsymbol{M}, j \in \boldsymbol{S} \right\}. \quad (9)$$

(Note that to simplify the notation, I employ a single index to address two-dimensional geographical space: the spatial bin i corresponds to a range of latitude and longitude.)

In the description of the metrics below, the equations are also applicable to a forecast with arbitrary independent distributions in each bin (i.e., not only Poisson). A forecast that specifies a probability distribution f_{ij} in each bin—that is, f_{ij} gives the probability of observing zero earthquakes, one earthquake, etc. in the magnitude-space bin indexed by (i, j)—can be denoted:

$$\Lambda = \left\{ f_{ij} \mid i \in \boldsymbol{M}, j \in \boldsymbol{S} \right\}. \tag{10}$$

The locations of the observed earthquakes—typically epicenters for regional experiments, but you may use hypocenters or centroid locations—are binned using the same discretization as $\boldsymbol{R}$, and the observed catalog, $\boldsymbol{\Omega}$, is represented by the set of the number of observed earthquakes in each bin, denoted $\omega(i,j)$ for the magnitude-space bin indexed by (i, j):

$$\boldsymbol{\Omega} = \left\{ \omega(i,j) \mid i \in \boldsymbol{M}, j \in \boldsymbol{S} \right\}. \tag{11}$$

All of the metrics in this section are based on the likelihood of observing the catalog given the forecast—in other words, the joint likelihood of each bin's observation given each bin's forecast. In the general case, the joint likelihood is:

$$\Pr(\omega_1 \mid \lambda_1)\Pr(\omega_2 \mid \lambda_2)\ldots\Pr(\omega_n \mid \lambda_n) = \prod_{(i,j)\in \boldsymbol{R}} f_{ij}\left(\omega(i,j)\right), \tag{12}$$

When the forecast is Poisson, the joint likelihood is given by eq. 3. Often, it is convenient to work with the natural logarithm of these joint likelihoods—the joint log-likelihood, which is the sum of each bin's log-likelihood. For the general case, this is

$$L(\boldsymbol{\Omega} \mid \Lambda) = \sum_{(i,j)\in \boldsymbol{R}} \log\left(f_{ij}\left(\omega(i,j)\right)\right), \tag{13}$$

and for the Poisson case, this is

$$L(\boldsymbol{\Omega} \mid \Lambda) = \sum_{(i,j)\in \boldsymbol{R}} \left(-\lambda(i,j) + \omega(i,j)\log\left(\lambda(i,j)\right) - \log\left(\omega(i,j)!\right)\right), \tag{14}$$

The joint log-likelihood has a negative value, and values that are closer to zero indicate a more likely observation—in other words, such a value indicates that the forecast shows better agreement with the observation.

To account for forecast uncertainty, you must usually simulate catalogs that are consistent with the forecast. In section 4.2, I described the procedure for the situation in which the number of earthquakes to simulate is known. In the general case of arbitrary independent forecast distributions, let F_{ij} be the cumulative probability density in bin (i, j). For each forecast bin, draw a random number z from the uniform distribution on $(0,1]$. The number of earthquakes to place in this bin is given by the inverse cumulative distribution at this point, $F_{ij}^{-1}(z)$. (In section 4.2, I described the Poisson inverse cumulative distribution, and, in general, any discrete inverse cumulative distribution can be solved directly by calculating the cumulative distribution function at each point and comparing with z, the probability of interest.) By iterating over all forecast bins, you will have a simulated catalog consistent with the forecast.

5.2.1 The Likelihood test (L-test)

In the L-test, you compute the observed joint log-likelihood (given by eq. 13 or 14) without any knowledge of whether this is a good score. Ask the question: if the forecast were "correct," what scores might we expect? In other words, is the observed catalog consistent with the forecast? To answer this question, simulate many catalogs consistent with the forecast using the procedure described above. In the situation where the forecast is Poisson, use the procedure described in section 4.4; the number of earthquakes to simulate for each catalog is also a random Poisson variable with expectation equal to the sum of each bin's expected rate. That is, the average rate parameter value is equal to the sum over all bins and you sample the Poisson distribution with this expectation to decide how many earthquakes to simulate. Such a sampling of the Poisson distribution is implemented in MATLAB® as `poissrnd`. Alternatively, you can draw a random number on [0,1] and execute

```
org.scec.predictionTesting.MathUtil.inverseCumulativePoisson
```

in the accompanying code.

Now you have a set of simulated catalogs $\{\hat{\Omega}\}$, where each catalog can be written

$$\hat{\Omega}_x = \{\hat{\omega}_x(i,j) \mid (i,j) \in \boldsymbol{R}\}, \tag{15}$$

where $\hat{\omega}_x(i,j)$ is the number of simulated earthquakes in bin (i, j). Here and in the following sections, the hat is used to indicate a simulated value or a value based on a simulated set of data. For each simulated catalog, compute the joint log-likelihood, forming the set $\{\hat{L}\}$ with the x^{th} member equal to the joint log-likelihood of the x^{th} simulated catalog:

$$\hat{L}_x = L(\hat{\Omega}_x \mid \Lambda), \tag{16}$$

where each member of the set is a simulated joint log-likelihood. Then compare the observed joint log-likelihood with the distribution of the simulated joint log-likelihoods. Does the observed joint log-likelihood fall in the lower tail of the distribution of $\{\hat{L}\}$? If it does, this indicates that the observation is not consistent with the forecast—in other words, the forecast is not accurate. The quantile score

γ is the fraction of simulated joint log-likelihoods less than or equal to the observed joint log-likelihood:

$$\gamma = \frac{\left|\left\{\hat{L}_x \mid \hat{L}_x \le L\right\}\right|}{\left|\left\{\hat{L}\right\}\right|}, \tag{17}$$

where $|\{A\}|$ denotes the number of elements in a set $\{A\}$ (see also Table 1). A very small value of γ indicates that the observation is inconsistent with the forecast. Indeed, this is an estimate of the significance value: you can say with 100(1-γ)% confidence that the observation is inconsistent with the forecast.

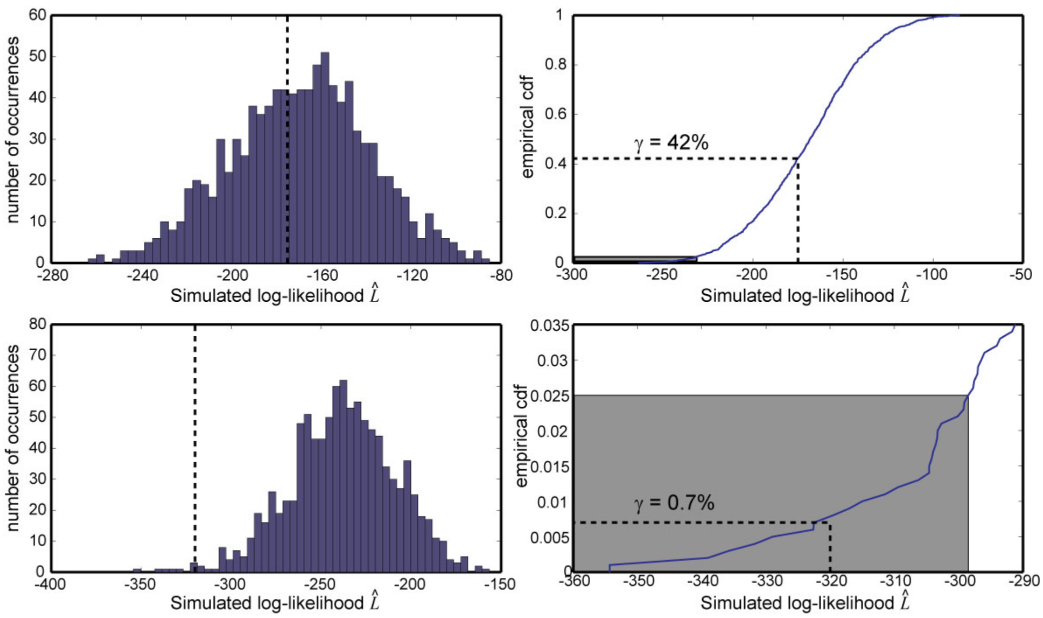

Fig. 5 Example L-test results for two different imaginary forecasts. See text for explanation.

Fig. 5 is a graphical explanation of the L-test. This figure demonstrates results for two imaginary forecasts (one in the top row and the other in the bottom row). In 5a and 5c, I show the observed joint log-likelihood (dashed black line) and a histogram of 1000 simulated joint log-likelihoods. In b) and d), I show the corresponding empirical cumulative distribution functions (solid blue line) and its intersection with the observed joint log-likelihood (dashed black line); also shown is the α=2.5% critical region (shaded box). The intersection of the black dashed line and the blue line indicates the L-test summary statistic γ. In 5a, the observed joint log-likelihood does not fall in the lower tail of simulated joint log-likelihoods; the dashed black line in 5b is not inside the shaded box. In 5c, the observed joint log-likelihood does fall in the lower tail—the dashed black line in 5d is inside the shaded box—indicating that the observed joint log-likelihood is much smaller than would be expected if the forecast in the second row were the true model of seismicity.

The L-test considers the entire space-rate-magnitude forecast and thereby blends the three components. In the following subsections, I describe three additional tests that isolate the skill of the rate forecast, magnitude forecast, and spatial forecast, respectively. Each of these tests is similar to the L-test. Heuristically, you can think of the L-test as comprising the N(umber)-test, which tests the rate forecast; the M(agnitude)-test, which tests the magnitude forecast; and the S(pace)-test, which tests the spatial forecast.

Pros: widely applicable, tests entire forecast
Cons: blends effects of spatial forecast, rate forecast, magnitude forecast

5.2.2 The Number test (N-test)

The N-test is intended to measure how well the total number of forecast earthquakes (summed over space and magnitude) matches the number of events observed; in other words, it isolates the rate component of the forecast. The question of interest, then, is as follows: is the number of observed target earthquakes consistent with the number of earthquakes forecast? The observed number of earthquakes, N_{obs}, can be written

$$N_{obs} = \sum_{(i,j)\in \boldsymbol{R}} \omega(i,j). \tag{18}$$

Does N_{obs} fall in one of the tails of the forecast rate distribution? In general, you can estimate the forecast rate distribution by simulating many catalogs with the procedure described in section 5.2. By doing this, you'll generate a set of simulated rates, $\{\hat{N}\}$, which you can use to estimate the probability i) of observing at most N_{obs} earthquakes and ii) of observing more than N_{obs} earthquakes. These probabilities are, respectively,

$$\delta_1 = \frac{\left|\{\hat{N}_x \mid \hat{N}_x \le N_{obs}\}\right|}{\left|\{\hat{N}\}\right|} \tag{19a}$$

and

$$\delta_2 = \frac{\left|\{\hat{N}_x \mid \hat{N}_x > N_{obs}\}\right|}{\left|\{\hat{N}\}\right|}. \tag{19b}$$

For a Poisson forecast, the forecast rate distribution is Poisson with expectation, N_{fore}, given by the sum over all bins:

$$N_{fore} = \sum_{(i,j)\in \boldsymbol{R}} \lambda(i,j). \tag{20}$$

Then the cumulative distribution of the forecast rate is simply the right-continuous Poisson cumulative distribution function,

$$F\left(x \mid N_{fore}\right) = \exp\left(-N_{fore}\right)\sum_{i=0}^{x}\frac{\left(N_{fore}\right)^{i}}{i!}, \tag{21}$$

and this can be used in the place of simulations. In the situation where the forecast is Poisson, the N-test metrics are:

$$\delta_1 = 1 - F\left(\left(N_{obs} - 1\right) \mid N_{fore}\right) \tag{22a}$$

and

$$\delta_2 = F\left(N_{obs} \mid N_{fore}\right). \tag{22b}$$

To interpret the N-test results, you can use a one-sided test with an effective significance value, α_{eff}, which is half of the intended significance value α; in other words, if you intend to maintain a Type I error rate of $\alpha = 5\%$, compare both δ_1 and δ_2 with a critical value of $\alpha_{eff} = 0.025$. If δ_1 is less than α_{eff}, the forecast rate is too low—an underprediction—and if δ_2 less than α_{eff}, the forecast rate is too high—an overprediction. For example, if 12.22 earthquakes were forecast (with Poisson uncertainty) and 16 earthquakes were observed, $\delta_1 = 17.2\%$ and $\delta_2 = 88.6\%$, indicating that the observation is consistent with the forecast.

Pros: isolates rate forecast, widely applicable
Cons: ignores spatial component, ignores magnitude component

5.2.3 The Magnitude test (M-test)

The objective of the M-test is to consider only the magnitude distributions of the forecast and the observation. To isolate these distributions, sum over the spatial bins and normalize the forecast so that its sum matches the observation:

$$\begin{aligned}
\mathbf{\Omega}^m &= \left\{\omega^m(i) \mid i \in \boldsymbol{M}\right\} \\
\omega^m(i) &= \sum_{j \in S}\omega(i,j) \\
\mathbf{\Lambda}^m &= \left\{\lambda^m(i) \mid i \in \boldsymbol{M}\right\} \\
\lambda^m(i) &= \frac{N_{obs}}{N_{fore}}\sum_{j \in S}\lambda(i,j)
\end{aligned} \tag{23}$$

Using these values, compute the observed joint log-likelihood just as in the L-test:

$$M = L\left(\mathbf{\Omega}^m \mid \mathbf{\Lambda}^m\right). \tag{24}$$

Here, the functional form of L is given by eq. 13 or 14, depending on the forecast format. How does the value from eq. 24 compare to the distribution of simulated joint log-likelihoods? For the M- and S-tests (see next subsection), rather than varying from simulation to simulation, the number of earthquakes to simulate, N_{sim}, is fixed at N_{obs}. (This is done to remove any effect of variations in earthquake rate.)

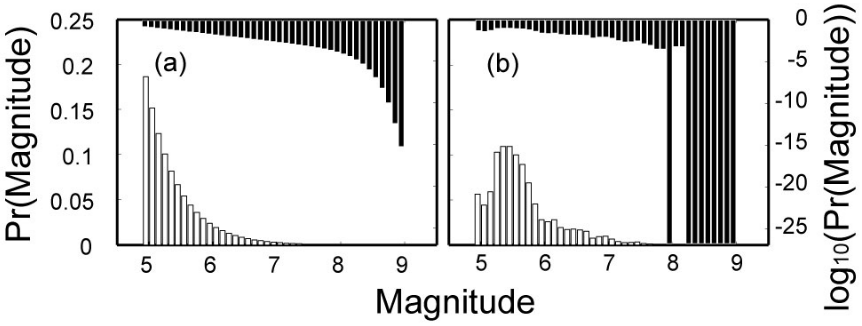

Fig. 6 Example magnitude probability distributions from RELM forecasts a) Helmstetter-Mainshock and b) Ward-Simulation. Distributions are discrete and were specified in 41 bins each having a width of 0.1 magnitude units. The white bars show the linear distribution (scale on left ordinate axis) and the black bars show the base-10 logarithm of the distribution (scale on right ordinate axis). (Modified from *Zechar et al.* 2010a)

For each simulated catalog, the joint log-likelihood is computed, forming the set $\{\hat{M}\}$ with the x^{th} member equal to the joint log-likelihood of the x^{th} simulated catalog:

$$\hat{M}_x = L\left(\hat{\Omega}_x^m \mid \Lambda^m\right). \tag{25}$$

Similar to the L-test, the M-test is summarized by a quantile score, κ,

$$\kappa = \frac{\left|\left\{\hat{M}_x \mid \hat{M}_x \le M\right\}\right|}{\left|\left\{\hat{M}\right\}\right|}. \tag{26}$$

If κ is less than the critical significance level α, this indicates that the observed magnitude distribution is inconsistent with the forecast.

Pros: isolates magnitude forecast, widely applicable
Cons: ignores spatial component, ignores rate component

5.2.4 The Space test (S-test)

The S-test is the spatial equivalent of the M-test, where only the spatial distributions of the forecast and the observation are considered. Similar to the M-test, isolate the spatial information by summing; in this case the sum is performed over magnitude bins, and the resulting forecast sum is normalized so that it matches the observation:

$$\mathbf{\Omega}^s = \{\omega^s(j) \mid j \in \boldsymbol{S}\}$$
$$\omega^s(j) = \sum_{i \in M} \omega(i,j)$$
$$\mathbf{\Lambda}^s = \{\lambda^s(j) \mid j \in \boldsymbol{S}\}$$
$$\lambda^s(j) = \frac{N_{obs}}{N_{fore}} \sum_{i \in M} \lambda(i,j) \quad . \tag{27}$$

This summing and normalization procedure removes the effect of the rate and magnitude components of the original forecast. Using these values, compute the observed joint log-likelihood just as in the L- and M-tests:

$$S = L(\mathbf{\Omega}^s \mid \mathbf{\Lambda}^s) . \tag{28}$$

Again, ask how this value compares to the distribution of simulated joint log-likelihoods. (The simulation procedure is the same as for the M-test.) For each simulated catalog, compute the joint log-likelihood, forming the set $\{\hat{S}\}$ with the x^{th} member equal to the joint log-likelihood of the x^{th} simulated catalog:

$$\hat{S}_x = L(\hat{\mathbf{\Omega}}_x^s \mid \mathbf{\Lambda}^s) . \tag{29}$$

The S-test is summarized by a quantile score ζ,

$$\zeta = \frac{\left|\{\hat{S}_x \mid \hat{S}_x \le S\}\right|}{\left|\{\hat{S}\}\right|} . \tag{30}$$

If ζ is less than critical significance value α, this indicates that the observed spatial distribution is inconsistent with the forecast.

Pros: isolates spatial forecast, widely applicable
Cons: ignores magnitude component, ignores rate component

5.2.5 The Likelihood Ratio test (R-test)

All of the preceding subsections in this section have emphasized the evaluation of a single forecast. The R-test is designed for pairwise comparisons of two forecasts, and it is based on the simple idea that a forecast with a higher joint log-likelihood is better. The likelihood ratio for two forecasts $\mathbf{\Lambda}_{\text{A}}$ and $\mathbf{\Lambda}_{\text{B}}$ is the difference of the joint log-likelihoods:

$$R_{\text{AB}} = L(\mathbf{\Omega} \mid \mathbf{\Lambda}_{\text{A}}) - L(\mathbf{\Omega} \mid \mathbf{\Lambda}_{\text{B}}) . \tag{31}$$

As with the previous likelihood-based metrics, you will simulate catalogs to establish the significance of this value. In this case, simulate catalogs consistent with Λ_A and construct a set of simulated likelihood ratios $\{\hat{R}_{AB}\}$. The corresponding metric is the quantile score

$$\alpha_{AB} = \frac{\left|\left\{\hat{R}_{AB} \mid \hat{R}_{AB} \le R_{AB}\right\}\right|}{\left|\left\{\hat{R}_{AB}\right\}\right|}. \tag{32}$$

If α_{AB} is less than the critical significance level α, this indicates that the observed catalog is inconsistent with the forecast Λ_A. This procedure is symmetric: you can compute R_{BA}, construct $\{\hat{R}_{BA}\}$, and determine α_{BA}. Keep in mind that the likelihood ratio itself—that is, the R value based on eq. 31—indicates which forecast is better, but the R-test does not provide a way to test the significance of this ratio, i.e., to quantify how much better a forecast is.

Pros: compares two forecasts while preserving the space-rate-magnitude structure
Cons: does not quantify how much better one forecast is than another

6 Illustrative Examples

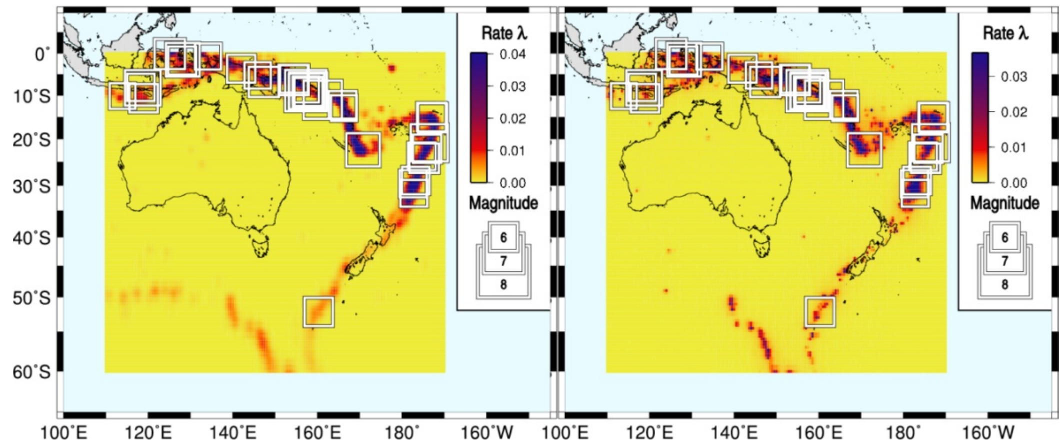

Fig. 7 Two example space-rate forecasts from the southwest Pacific CSEP testing region. White squares indicate locations and magnitudes of observed target earthquakes. These are the example forecasts in the accompanying data files (TripleS on the left, DBM on the right). (Note that the accompanying example catalog is from an earlier time period, not the one shown here.)

Along with this article, you should have downloaded a zipped file containing codes that implement the methods described in section 5 and some data files to accompany the examples. If you did not already download this zipped file, get it from www.corssa.org. These codes require the Java runtime engine, which you can download from www.java.com. All instructions for executing the examples are included in the file readme.txt.

For each test except the M-test, I have provided the two forecasts that are shown in Fig. 6. These are space-rate forecasts without magnitude discretization and are being analyzed in the US CSEP testing center. For the M-test, I provided the Helmstetter-Mainshock and Ward-Simulation forecasts from the RELM experiment (the magnitude component of these forecasts are depicted in Fig. 6) (*Helmstetter et al.* 2007, *Ward* 2007).

7 Further Reading

Mason (2003) provides an extensive overview of binary forecasts, with an emphasis on contingency table measures and their applications to, for example, tornado forecasts.

The articles of *Molchan* (1991) and *Molchan and Kagan* (1992) give the most complete treatment of the Molchan diagram (alternatively called the errors diagram). *Kossobokov* (2004, Section 7) also discussed the Molchan diagram and an empirical approach for defining the measure of space. The recent article by *Molchan and Keilis-Borok* (2008) and the article by *Molchan* (2010) consider the extension of the earthquake prediction problem to multidimensional spatial forecasts.

One of the earliest applications of likelihood metrics to earthquake forecast testing was the evaluation by *Kagan and Jackson* (1995) of a forecast by *Nishenko* (1991). The forecast method was applied to a set of spatial zones, and the probability of a target earthquake in each zone was given—in this case, the target earthquake was defined by a zone-varying minimum magnitude.

In evaluating the VAN algorithm, *Kagan* (1996) applied a simple technique in which he simulated alarms after each strong earthquake; in doing so, he showed that he could easily match the performance of the VAN algorithm. *McGuire et al.* (2005) and *McGuire* (2008) adopted a similar procedure to demonstrate the predictability of moderate earthquakes in the East Pacific Rise.

Jackson (1996, p. 3773) described a simple method for simulating target earthquakes for a set of predictions for which success probabilities are estimated.

In developing broad regional earthquake forecasts, *Jackson and Kagan* (1999) and *Kagan and Jackson* (2000) used an L-test for simulating target earthquakes for a set of predictions for which success probabilities are estimated; they fixed the number of earthquakes in the simulations, yielding a conditional L-test. They also applied a likelihood ratio comparison of forecasts.

Harte and Vere-Jones (2005) discussed an entropy score for probabilistic forecasts with several possible outcomes and considered the relationship of the entropy score to average log-likelihood and the Molchan diagram. *Harte et al.* (2007) applied these techniques (in particular, an information gain) to test the M8 algorithm in New Zealand.

Schorlemmer et al. (2007) showed an example R-test computation using forecasts from *Helmstetter et al.* (2007) and the 2002 United States National Seismic Hazard Map.

Kagan (2007, 2009) further developed the connections between Molchan diagram analysis and metrics based on likelihood.

Zechar and Jordan (2008) introduced the area skill score metric and applied it to three alarm function forecasts—two probabilistic and the other providing only a ranking. *Zechar and Jordan* (2010) further explored the area skill score and gave analytical and numerical solutions for its distribution.

Schorlemmer et al. (2010) applied the L-, N-, and R-tests to the first half of the RELM forecast experiment.

Zechar et al. (2010a) applied all of the likelihood metrics discussed in this article (save the R-Test) to the RELM forecasts and explored the stability and statistical power of the tests.

Werner et al. (in press) applied all of the likelihood metrics to the forecasts from the CSEP Italy experiment.

8 Caveats and Summary

No evaluation metric is ideal for all earthquake forecast experiments. Indeed, because the format of predictions and forecasts can vary so widely, no evaluation metric is even applicable to all experiments! My hope is that some of the methods I described in section 5 will be applicable and appropriate for your problem. But it may be that the set of predictions or forecast that you want to evaluate cannot be judged with any of these metrics. In that situation, it is likely that you can at least let these methods guide you to a custom solution. In particular, make sure that the reference model that you use, implicit or otherwise, is realistic and not overly simple. When possible, incorporate the relevant first-order observations of seismicity—for example, clustering of earthquakes in space and time.

You should note that if you apply multiple tests to a given combination of forecast and observation and you choose a nominal critical significance value for each test, the composite confidence is not the same as the significance value that you used for each test. In particular, the probability of rejecting a forecast based on at least one of the tests is higher than the assumed significance value for a given test. In other words, if you conduct an L-test with a critical value of 5% and an N-test with a critical value of 5%, the probability that a forecast will "pass" both tests is less than 5%. In this situation, you can consider applying a Bonferroni correction to the composite significance value (or, conversely, to the individual test values) (e.g., *Shaffer* 1995).

The gridded rate forecast format is also worth further consideration. In section 5, I presented solutions only for forecasts with independent bin forecasts. As seismologists, we find this unsettling, because we tend to think that earthquakes interact with and even trigger each other, i.e., bins are not independent. In principle, it is straightforward to modify the solutions if the forecast dependence is fully specified: so long as the likelihood of an observation can be computed, the metrics are easy to compute. But to specify this independence in advance of a forecast experiment is not a trivial task: for each bin, you would need to specify conditional probability distributions characterizing the relationship to all other bins. From this perspective, a shift to emphasizing short-term forecasts, which can be updated quickly after each new earthquake, may be appropriate.

Acknowledgements Portions of this article first appeared in slightly different form in *Zechar and Jordan* (2008, 2010), *Zechar and Zhuang* (2010), and *Zechar et al.* (2010a). I thank A. Michael, S. Wiemer, and an anonymous reviewer for comments that led to improvements in this article. I was supported in part by NSF grant EAR-0944202.

References

Allen, C.R., Edwards, W., Hall, W.J., Knopoff, L., Raleigh, C.B., Savit, C.H., Toksoz, M.N., Turner, R.H. (1976), Predicting earthquakes: A scientific and technical evaluation—with implications for society. *Panel on Earthquake Prediction of the Committee on Seismology' Assembly of Mathematical and Physical Sciences*, National Research Council, U.S. National Academy of Sciences, Washington, DC.

Field, E.H. (2007), Overview on the working group for the development of Regional Earthquake Likelihood Models, *Seismol. Res. Lett.* 78, 7–16.

Harte, D. and D. Vere-Jones (2005), The entropy score and its uses in earthquake forecasting, *Pure Appl. Geophys.* 162, 1229-1253.

Harte, D. D.-F. Li, M. Vreede, D. Vere-Jones, and Q. Wang (2007), Quantifying the M8 algorithm: model, forecast, and evaluation, *New. Zeal. J. Geol. Geop.* 50, 117-130.

Helmstetter, A., Y.Y. Kagan, and D.D. Jackson (2007), High-resolution time-independent grid-based forecast for M>5 earthquakes in California, *Seismol. Res. Lett.* 78, 78–86, doi: 10.1785/gssrl.78.1.78.

Hough, S.E. (2009), Predicting the unpredictable: the tumultuous science of earthquake prediction, Princeton, 272 pp.

Jackson, D.D. (1996), Hypothesis testing and earthquake prediction, *Proc. Natl. Aca. Sci. USA*, 93, 3772-3775.

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

Jordan, T.H. (2006), Earthquake predictability, brick by brick, *Seismol. Res. Lett.* 77, 3–6.

Kagan, Y.Y. (1996), VAN earthquake predictions-an attempt at statistical evaluation, *Geophys. Res. Lett.* 23(11), 1315-1318.

Kagan, Y.Y. (2007), On earthquake predictability measurement: Information score and error diagram, *Pure Appl. Geophys.* 164(2007), 1947-1962.

Kagan, Y.Y. (2009), Testing long-term earthquake forecasts: likelihood methods and error diagrams, *Geophys. J. Int.* 177, 532-542, doi: 10.1111/j.1365-246X.2008.04064.

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

Kagan, Y.Y., and D.D. Jackson (2000), Probabilistic forecasting of earthquakes, *Geophys. J. Int.* 143, 438—453.

Kossobokov, V. (2004), Earthquake prediction: basics, achievements, perspectives, *Acta Geod. Geoph. Hung.*, 39(2/3), 205-221.

Kossobokov, V.G. (2006), Testing earthquake prediction methods: "The West Pacific short-term forecast of earthquakes with magnitude MwHRV>5.8", *Tectonophys.* 413, 25-31.

Mason, I.B. (2003), Binary events, in *Forecast Verification*, pp. 37-76, eds. Jolliffe, I.T. & Stephenson, D.B., Wiley, Hoboken.

McGuire, J.J. (2008), Seismic cycles and earthquake predictability on East Pacific Rise transform faults, *Bull. Seism. Soc. Am.*, 98(3), 1067-1084.

McGuire, J.J., M.S. Boettcher, and T.H. Jordan (2005), Foreshock sequences and short-term earthquake predictability on East Pacific Rise transform faults, *Nature*, 434(7032), 457-461.

Molchan G.M. (1991), Structure of optimal strategies in earthquake prediction, *Tectonophys.* 193, 267—276.

Molchan G.M. (2010), Space-time earthquake prediction: the error diagrams, *Pure Appl. Geophys.*, doi:10.1007/s00024-010-0087-z.

Molchan, G.M. and Y.Y. Kagan (1992), Earthquake prediction and its optimization, *J. Geophys. Res.* 97, 4823—4838.

Molchan, G.M. and V.I. Keilis-Borok (2008), Earthquake prediction: Probabilistic aspect, *Geophys. J. Int.* 173, 1012-1017.

Molchan, G.M. and L.L. Romashkova (in review), Earthquake prediction analysis: the M8 algorithm, submitted to *Geophys. J. Int.*

Nishenko, S.P. (1991), Circum-Pacific seismic potential: 1989-1999, *Pure Appl. Geophys.* 135(2), 169-259.

Schorlemmer, D., and M.C. Gerstenberger (2007), RELM testing center, *Seismol. Res. Lett.* 78, 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, 17–29.

Schorlemmer, D., J.D. Zechar, M.J. Werner, E.H. Field, D.D. Jackson, and T.H. Jordan (2010), First results of the Regional Earthquake Likelihood Models experiment, *Pure Appl. Geophys.*, 167, 8/9, doi:10.1007/s00024-010-0081-5.

Shaffer, J.P. (1995), Multiple hypothesis testing, *Ann. Rev. Psych.* 46, 561–584.

Ward, S.N. (2007), Methods for evaluating earthquake potential and likelihood in and around California, *Seismol. Res. Lett.* 78, 121—133.

Werner, M.J. J.D. Zechar, W. Marzocchi, and S. Wiemer (in press), Retrospective evaluation of the five-year and ten-year CSEP-Italy earthquake forecasts, to appear in *Ann. Geophys.*

Werner, M.J., *et al.* (in preparation), Spatial estimation, *Community Online Resource for Statistical Seismology*, doi:10.5078/corssa-77337879.

Woessner, J., J Hardebeck, and E. Hauksson (in preparation), What is an instrumental seismicity catalog?, *Community Online Resource for Statistical Seismology*, doi:10.5078/corssa-38784307.

Wyss, M., and A. Michael (in preparation), Overview of earthquake predictability and related hypothesis testing, *Community Online Resource for Statistical Seismology*, doi:10.5078/corssa-53065141.

Zechar, J.D., M.C. Gerstenberger, and D.A. Rhoades (2010a), Likelihood-based tests for evaluating space-rate-magnitude forecasts, *Bull. Seis. Soc. Am.*, 100(3), 1184–1195, doi:10.1785/0120090192.

Zechar, J.D., and T.H. Jordan (2008), Testing alarm-based earthquake predictions, *Geophys. J. Int.*, 172(2), 715–724, doi: 10.1111/j.1365-246X.2007.03676.x.

Zechar, J.D., and T.H. Jordan (2010), The area skill score for evaluating earthquake predictability experiments, *Pure Appl. Geophys.*, 167, 8/9, doi:10.1007/s00024-010-0086-0.

Zechar, J.D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P.J. Maechling, and T.H. Jordan (2010b), The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science, *Concurr. Comp˒Pract.* E., doi:10.1002/cpe.1519.

Zechar, J.D., and J. Zhuang (2010), Risk and return: evaluating Reverse Tracing of Precursors earthquake predictions, *Geophys. J. Int.*, doi:10.1111/j.1365-246X.2010.04666.x.

Zhuang, J. (2010), Gambling scores for earthquake forecasts and predictions, *Geophys. J. Int.* 181, 382–390. doi: 10.1111/j.1365-246X.2010.04496.x.