Bulletin of the Seismological Society of America, Vol. 100, No. 3, pp. 1184–1195, June 2010, doi: 10.1785/0120090192

Likelihood-Based Tests for Evaluating Space–Rate–Magnitude Earthquake Forecasts

by J. Douglas Zechar,* Matthew C. Gerstenberger, and David A. Rhoades

Abstract The five-year experiment of the Regional Earthquake Likelihood Models (RELM) working group was designed to compare several prospective forecasts of earthquake rates in latitude–longitude–magnitude bins in and around California. This forecast format is being used as a blueprint for many other earthquake predictability experiments around the world, and therefore it is important to consider how to evaluate the performance of such forecasts. Two tests that are currently used are based on the likelihood of the observed distribution of earthquakes given a forecast; one test compares the binned space–rate–magnitude observation and forecast, and the other compares only the rate forecast and the number of observed earthquakes. In this article, we discuss a subtle flaw in the current test of rate forecasts, and we propose two new tests that isolate the spatial and magnitude component, respectively, of a space–rate–magnitude forecast. For illustration, we consider the RELM forecasts and the distribution of earthquakes observed during the first half of the ongoing RELM experiment. We show that a space–rate–magnitude forecast may appear to be consistent with the distribution of observed earthquakes despite the spatial forecast being inconsistent with the spatial distribution of observed earthquakes, and we suggest that these new tests should be used to provide increased detail in earthquake forecast evaluation. We also discuss the statistical power of each of the likelihood-based tests and the stability (with respect to earthquake catalog uncertainties) of results from the likelihood-based tests.

Online Material: Details of result stability and details of statistical power of tests.

Introduction

Several prospective earthquake forecast experiments are being conducted within testing centers of the international Collaboratory for the Study of Earthquake Predictability (CSEP) (Jordan, 2006; Schorlemmer and Gerstenberger, 2007; Zechar *et al.*, 2009). The majority of these experiments are similar to the five-year experiment designed by the Regional Earthquake Likelihood Models (RELM) working group in California (see Field, 2007, for an overview). The objective of these experiments is to quantify earthquake predictability by testing various hypotheses of earthquake occurrence and by identifying models, or model features, that can be used to improve seismic hazard estimates. In these experiments, forecasts are compared with observations using the likelihood-based testing procedures described by Schorlemmer *et al.* (2007). Based on preliminary results of the ongoing RELM experiment (Schorlemmer *et al.*, 2010), and because the tests are now being applied so widely within CSEP, it is worthwhile to reconsider these metrics. In this article, we highlight a logical error in one of the existing tests, and we propose two new tests that allow for a more comprehensive evaluation of this class of earthquake forecasts. We use the first half of the RELM experiment to illustrate these tests, and we present and discuss the results, their stability, and the statistical power of the tests.

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. To test this probabilistic forecast, Kagan and Jackson (1995) introduced three tests: the number test (N-test), the likelihood test (L-test), and the ratio test (R-test). These tests were modified slightly and became the basis for RELM experiment evaluation (Schorlemmer *et al.*, 2007). The RELM forecasts are stated in different terms than Nishenko (1991) used; rather than probabilities, these

*Also at the Swiss Seismological Service, Institute of Geophysics, ETH Zurich Sonneggstrasse 5, 8092 Zurich, Switzerland.

forecasts specify the expected number of earthquakes in bins of latitude, longitude, and magnitude. Thus, embedded in each RELM forecast are forecasts of the (1) rate, (2) spatial distribution, and (3) magnitude distribution of target earthquakes. The L-test employed in CSEP experiments, rather than being a function of spatially-varying probabilities as in Kagan and Jackson (1995), considers the entire space–rate–magnitude forecast. The N-test isolates the rate forecast. It seems logical, then, to consider also the magnitude and spatial forecasts and to test these in isolation. In a performance test of a forecasting model against a time-invariant baseline model, Evison and Rhoades (1999) compared the magnitude and spatial distributions of the target earthquakes with those expected under each model. Such a separation of the forecast into its constituent components provides extra detail and may help to identify successful features of forecast models.

Experimental Setup and Notation

For the class of experiments we consider, the testing region, $\boldsymbol{R}$, is the combination of the binned magnitude range of interest, $\boldsymbol{M}$, and the binned spatial domain of interest, $\boldsymbol{S}$:

$$\boldsymbol{R} = \boldsymbol{M} \times \boldsymbol{S}. \tag{1}$$

For example, in the RELM experiment, the magnitude range of interest is 4.95 and greater, and the bin size is 0.1 units (except for the final bin, which is open-ended):

$$\boldsymbol{M} = \{[4.95, 5.05), [5.05, 5.15), \ldots, [8.85, 8.95), [8.95, \infty]\}. \tag{2}$$

The RELM spatial domain of interest is the region within a polygon enclosing California (coordinates can be found in table 2 of Schorlemmer and Gerstenberger, 2007); this area is represented as a set of latitude–longitude rectangles of $0.1° \times 0.1°$.

Almost all current experiments in CSEP testing centers consider binned earthquake forecasts that incorporate an assumption that the number of earthquakes in each forecast bin is Poisson-distributed and independent of the number of earthquakes in other bins. In this case, an earthquake forecast, $\boldsymbol{\Lambda}$, on $\boldsymbol{R}$ is specified by the expected number of events in each magnitude–space bin:

$$\boldsymbol{\Lambda} = \{\lambda(i, j) | i \in \boldsymbol{M}, j \in \boldsymbol{S}\}. \tag{3}$$

Here, $\lambda(i, j)$ is the number of earthquakes forecast in the magnitude–space bin (i, j). Note that to simplify the notation, we employ a single index to address two-dimensional geographical space: the spatial bin j actually corresponds to a range of latitudes and longitudes.

The observed locations of the earthquakes (typically epicenters for regional experiments, but hypocenters or centroid locations may be used) are binned using the same discretization as $\boldsymbol{R}$, and the observed catalog, $\boldsymbol{\Omega}$, is represented by the set of the number of earthquakes in each bin:

$$\boldsymbol{\Omega} = \{\omega(i, j) | i \in \boldsymbol{M}, j \in \boldsymbol{S}\}. \tag{4}$$

Here $\omega(i, j)$ is the number of earthquakes observed in the magnitude–space bin (i, j).

Methods

N-Test

The N-test is intended to measure how well the forecast number of earthquakes matches the observed number of events. Because the rules of CSEP experiments thus far are such that the forecast rate in each bin is considered independent and Poisson-distributed, the overall forecast rate is also Poissonian, with expectation, N_{fore}, given by the sum over all bins:

$$N_{\text{fore}} = \sum_{(i,j) \in \boldsymbol{R}} \lambda(i, j). \tag{5}$$

Similarly, the observed number of earthquakes, N_{obs}, can be written

$$N_{\text{obs}} = \sum_{(i,j) \in \boldsymbol{R}} \omega(i, j). \tag{6}$$

The question of interest, then, is as follows: is the number of observed target earthquakes consistent with the number of earthquakes forecast? In other words, if we assume that the forecast rate distribution is correct, does N_{obs} fall in one of the tails of the distribution? Kagan and Jackson (1995) and Schorlemmer *et al.* (2007) described a simulation-based approach to answer this question, but simulations are not required when an analytical form of forecast uncertainty is used. In this case, the corresponding cumulative distribution can be used. In the limit of infinite simulations, the N-test metric that Schorlemmer *et al.* (2007) describe approaches

$$\delta = F(N_{\text{obs}} | N_{\text{fore}}), \tag{7}$$

where $F(x|\mu)$ is the right-continuous Poisson cumulative distribution function with expectation μ evaluated at x, and $F(x|\mu) = 0$ for all $x < 0$. In other words, δ is the probability that no more than N_{obs} target earthquakes are observed, given that N_{fore} are expected, and a two-sided hypothesis test is currently used within CSEP testing centers. This two-sided test considers two questions: Given the forecast, (1) what is the probability of observing at most N_{obs} earthquakes? and (2) what is the probability of observing more than N_{obs} earthquakes? The answers are δ and $(1 - \delta)$, respectively. There is, however, a subtle flaw in this approach. To illustrate this flaw, consider the case in which the total number of earthquakes forecast is 0.0015. If zero events are observed, $\delta = e^{-0.0015} \approx 0.9985$, and the forecast would be rejected at a high level of confidence. Of course, if more than zero events are observed, δ is even larger, and the forecast would also be rejected. In this case, it seems that the forecast cannot win. The problem with using equation 7 and its complement is that the questions in which we are

interested are slightly different: we should consider separately the probabilities of observing (1) at least and (2) at most N_{obs} events. These probabilities can be written respectively as

$$\delta_1 = 1 - F((N_{\text{obs}} - 1)|N_{\text{fore}}) \tag{8a}$$

and

$$\delta_2 = F(N_{\text{obs}}|N_{\text{fore}}). \tag{8b}$$

By writing the probabilities this way, we only need to be concerned with low probability outcomes. In contrast to the two-sided test applied to the outcome of equation 7, we 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 we intend to maintain a Type I error rate of $\alpha = 5\%$, we compare both δ_1 and δ_2 with a critical value of $\delta_{\text{eff}} = 0.025$. To be clear, this effective significance value is not a correction for multiple tests; rather, it is necessary because we have transformed the high tail of the distribution to a small value (equation 8a). While it is a minor technical detail, we prefer this approach because it avoids the outcome of having to reject what otherwise seems to be a high probability event. If δ_1 is very small, the forecast rate is too low (an underprediction); and, if δ_2 is very small, the forecast rate is too high (an overprediction). We note that δ_2 in equation 8 is defined to be exactly δ from equation 7 and therefore equivalent to the asymptotic result of simulations described by Schorlemmer *et al.* (2007). Considering the previous example ($N_{\text{fore}} = 0.0015$, $N_{\text{obs}} = 0$), equation 8 yields $\delta_1 = 1$ and $\delta_2 = 0.9985$, correctly indicating that the observation is consistent with the forecast.

L-Test

In this subsection, we review the L-test as described by Schorlemmer *et al.* (2007); we do this to introduce notation that will be used in subsequent sections and to point out slight differences between the description in Schorlemmer *et al.* (2007) and the CSEP implementation.

To determine how well the forecast in a single bin matches the observation in that bin, we ask: assuming that the forecast is correct, what is the likelihood of the observation? As noted by Schorlemmer *et al.* (2007), because the forecast in each bin follows a Poisson distribution, the likelihood is, by definition,

$$\Pr(\omega|\lambda) = \frac{\lambda^{\omega}}{\omega!}\exp(-\lambda). \tag{9}$$

Because the rates in each bin are assumed to be independent, the joint likelihood (for all bins) of the observation given the forecast is simply the product over all bins,

$$\Pr(\boldsymbol{\Omega}|\boldsymbol{\Lambda}) = \prod_{(i,j)\in\boldsymbol{R}} \Pr(\omega(i,j)|\lambda(i,j)), \tag{10}$$

where the likelihood for each bin is given by equation 9. Often, it is convenient to work with the natural logarithm of these respective likelihoods: the log-likelihood,

$$L(\omega|\lambda) = \log(\Pr(\omega|\lambda)) = -\lambda + \omega\log\lambda - \log(\omega!), \tag{11}$$

and the joint log-likelihood, which is the sum of each bin's log-likelihood:

$$L(\boldsymbol{\Omega}|\boldsymbol{\Lambda}) = \sum_{(i,j)\in\boldsymbol{R}} (-\lambda(i,j) + \omega(i,j)\log(\lambda(i,j)) - \log(\omega(i,j)!)). \tag{12}$$

The observed joint log-likelihood, given by equation (12), 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, catalogs that are consistent with the forecast are simulated. The simulation procedure currently used within CSEP testing centers is simpler and more efficient than that described by Schorlemmer *et al.* (2007). To begin, a discrete distribution is constructed by normalizing and summing the rates in each forecast bin, where each bin of the constructed discrete distribution is assigned the cumulative normalized sum. For a given simulation, the number of target earthquakes to simulate, N_{sim}, is determined by choosing a number at random from the Poisson distribution with expectation N_{fore}. Then, for each of the N_{sim} events, a number is chosen at random from the uniform distribution over (0, 1], and an earthquake is placed in the bin with the appropriate cumulative normalized sum value. For example, if $\boldsymbol{\Lambda} = \{1, 4, 3, 2\}$, the constructed discrete distribution has the values $\{0.1, 0.5, 0.8, 1.0\}$; if the random number drawn was 0.37, an earthquake would be placed in the second bin. This simulation process is repeated many times, yielding a set of simulated catalogs $\{\hat{\boldsymbol{\Omega}}\}$, where each catalog can be represented as

$$\hat{\boldsymbol{\Omega}}_x = \{\hat{\omega}_x(i,j)|(i,j)\in\boldsymbol{R}\}, \tag{13}$$

where $\hat{\omega}_x(i,j)$ is the number of simulated earthquakes in bin (i,j).

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

$$\hat{L}_x = L(\hat{\boldsymbol{\Omega}}_x|\boldsymbol{\Lambda}), \tag{14}$$

where we call each member of the set a simulated joint log-likelihood. We then compare the observed joint log-likelihood, $L = L(\boldsymbol{\Omega}|\boldsymbol{\Lambda})$, with the distribution of the simulated joint log-likelihoods. We are interested in the question: does L 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:

$$\gamma = \frac{|\{\hat{L}_x | \hat{L}_x \le L\}|}{|\{\hat{L}\}|}, \tag{15}$$

where $|\{A\}|$ denotes the number of elements in a set $\{A\}$. A very small value of γ indicates that the observation is inconsistent with the forecast.

The L-test, as described previously and implemented in CSEP testing centers, considers the entire space–rate–magnitude forecast and thereby blends the three components. The N-test isolates the rate portion of the forecast, but the magnitude and space components are not considered separately. Therefore, we propose two additional tests that measure the skill of the implied spatial forecast and magnitude forecast. Each of these tests is similar to the L-test. Heuristically, we can think of the L-test as comprising the N-test, which tests the rate forecast; the M-test (magnitude test), which tests the magnitude forecast; and the S-test (space test), which tests the spatial forecast.

M- and S-Tests

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

$$\begin{aligned} \boldsymbol{\Omega}^m &= \{\omega^m(i) | i \in \boldsymbol{M}\} \qquad \omega^m(i) = \sum_{j \in S} \omega(i,j) \\ \boldsymbol{\Lambda}^m &= \{\lambda^m(i) | i \in \boldsymbol{M}\} \qquad \lambda^m(i) = \frac{N_{\text{obs}}}{N_{\text{fore}}} \sum_{j \in S} \lambda(i,j). \end{aligned} \tag{16}$$

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

$$M = L(\boldsymbol{\Omega}^m | \boldsymbol{\Lambda}^m). \tag{17}$$

Again, we want to know how this value compares to the distribution of simulated joint log-likelihoods. In this case, the simulation procedure is slightly different than described in the previous 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. As with L-test simulations, a normalized distribution is computed, only in this case the values of the distribution are based on the magnitude distribution forecast values. For each simulated catalog, the joint log-likelihood is computed, forming the set $\{\hat{M}\}$ with the xth member equal to the joint log-likelihood of the xth simulated catalog:

$$\hat{M}_x = L(\hat{\boldsymbol{\Omega}}_x^m | \boldsymbol{\Lambda}^m). \tag{18}$$

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

$$\kappa = \frac{|\{\hat{M}_x | \hat{M}_x \le M\}|}{|\{\hat{M}\}|}. \tag{19}$$

If κ is very small, this indicates that the observed magnitude distribution is inconsistent with the forecast.

The S-test is the spatial equivalent of the M-test, where we consider only the spatial distribution of the forecast and the observation. Similar to the M-test, we 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:

$$\begin{aligned} \boldsymbol{\Omega}^s &= \{\omega^s(j) | j \in \boldsymbol{S}\} \qquad \omega^s(j) = \sum_{i \in M} \omega(i,j) \\ \boldsymbol{\Lambda}^s &= \{\lambda^s(j) | j \in \boldsymbol{S}\} \qquad \lambda^s(j) = \frac{N_{\text{obs}}}{N_{\text{fore}}} \sum_{i \in M} \lambda(i,j). \end{aligned} \tag{20}$$

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

$$S = L(\boldsymbol{\Omega}^s | \boldsymbol{\Lambda}^s). \tag{21}$$

Again, we want to know how this value compares to the distribution of simulated joint log-likelihoods. In this case, the simulation procedure is the same as for the M-test. For each simulated catalog, the joint log-likelihood is computed, forming the set $\{\hat{S}\}$ with the xth member equal to the joint log-likelihood of the xth simulated catalog:

$$\hat{S}_x = L(\hat{\boldsymbol{\Omega}}_x^s | \boldsymbol{\Lambda}^s). \tag{22}$$

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

$$\zeta = \frac{|\{\hat{S}_x | \hat{S}_x \le S\}|}{|\{\hat{S}\}|}. \tag{23}$$

If ζ is very small, this indicates that the observed spatial distribution is inconsistent with the forecast.

Data

To illustrate the modified N-test and the proposed M- and S-tests, we consider the first half of the ongoing five-year RELM experiment. Although the experiment is now more than halfway to completion, this period was selected so that the results can be compared directly to those presented by Schorlemmer *et al.* (2010); we also refer the reader to that article for details of the experiment and the observed target earthquakes. The RELM experiment comprises two forecast groups: one group of forecasts intended to predict independent events (the mainshock forecast group) and one group of

Table 1
L-, N-, M-, and S-Test Results for RELM Mainshock Forecasts*

Forecast	L (γ)	N (δ_1, δ_2)	M (κ)	S (ζ)
Ebel-Mainshock	0.149	0.634, 0.503	0.793	0.000
Helmstetter-Mainshock	0.723	0.726, 0.391	0.856	0.468
Holliday-PI	0.992	0.996, **0.011**	0.717	**0.025**
Kagan-Mainshock	0.974	0.982, 0.063	0.660	0.755
Shen-Mainshock	0.969	0.967, 0.107	0.654	0.931
Ward-Combo	0.998	0.999, **0.004**	0.695	0.767
Ward-Geodetic81	1.000	1.000, **0.000**	0.706	0.718
Ward-Geodetic85	0.987	0.993, 0.030	0.705	0.724
Ward-Geologic	0.998	0.998, **0.011**	0.700	0.534
Ward-Seismic	0.993	0.997, **0.014**	0.706	0.700
Ward-Simulation	0.725	0.885, 0.282	0.541	0.557
Wiemer-ALM	0.637	0.834, 0.256	0.891	**0.001**

*Quantile scores: γ, consistency of forecast with the observed space–magnitude distribution and rate of target earthquakes (values are from Schorlemmer *et al.*, 2010); δ_1 and δ_2, consistency of forecast with the total number of observed target earthquakes; κ and ζ, consistency of forecast with the observed magnitude and spatial distributions, respectively, of target earthquakes. Bold values indicate that the observed distribution is inconsistent with the forecast.

forecasts meant to predict all events with magnitude greater than or equal to 4.95 (the mainshock + aftershock group). The mainshock forecast group contains 12 forecasts, and the mainshock + aftershock group contains 5 forecasts. These forecasts are summarized by Field (2007). Because the forecasts were originally stated in terms of number of events per bin for a five-year period and we only consider the first 2.5 years of the experiment, all forecast rates in every bin of each forecast were multiplied by 1/2. We note that the results presented here are not the final results of the experiment and are subject to fluctuation.

The results we present in the following section are based on 12 earthquakes, 9 of which are considered mainshocks according to the rules of the RELM experiment (Schorlemmer and Gerstenberger, 2007). The time period considered is 1 January 2006 to 30 June 2008, inclusive. Details of the observed target earthquakes were described by Schorlemmer *et al.* (2010, their table 2). For simplicity, we do not report the results of the corrected forecast groups described by Schorlemmer *et al.* (2010).

Results

In Table 1, we present the results of the first half of the RELM experiment for the mainshock forecast group (Ebel *et al.*, 2007; Helmstetter *et al.*, 2007; Holliday *et al.*, 2007; Kagan *et al.*, 2007; Shen *et al.*, 2007; Ward, 2007; Wiemer and Schorlemmer, 2007). Table 2 lists the results for the mainshock + aftershock forecast group (Bird and Liu, 2007; Ebel *et al.*, 2007; Helmstetter *et al.*, 2007; Kagan *et al.*, 2007; Shen *et al.*, 2007). Following Schorlemmer *et al.* (2007, 2010), we declare that the observed catalog is inconsistent with a forecast if the quantile score is less than or equal to 0.025, and the quantile scores of the "rejected" forecasts are marked in boldface. Please note that the quantile scores in Tables 1 and 2 are not the final results of the five-year experiment. The final values will be different from those presented here, and a forecast that is rejected here may ultimately succeed and vice versa.

We include the L-test γ values reported by Schorlemmer *et al.* (2010) and note that no forecast fails this test. As they

Table 2
L-, N-, M-, and S-Test Results for RELM Mainshock + Aftershock Forecasts*

Forecast	L (γ)	N (δ_1, δ_2)	M (κ)	S (ζ)
Bird-Neokinema	1.000	1.000, **0.001**	0.889	0.353
	$1.00^{+0.00}_{-0.00}$	$1.00^{+0.00}_{-0.00}$, $\mathbf{0.00}^{+0.00}_{-0.00}$	$0.92^{+0.08}_{-0.27}$	$0.49^{+0.38}_{-0.29}$
	$0.96^{+0.04}_{-0.18}$	$0.98^{+0.02}_{-0.13}$, $0.04^{+0.16}_{-0.04}$	$0.78^{+0.21}_{-0.39}$	$0.15^{+0.39}_{-0.14}$
Ebel-Aftershock	1.000	1.000, **0.000**	0.815	**0.000**
	$1.00^{+0.00}_{-0.00}$	$1.00^{+0.00}_{-0.00}$, $\mathbf{0.00}^{+0.00}_{-0.00}$	$0.85^{+0.12}_{-0.31}$	$\mathbf{0.00}^{+0.00}_{-0.00}$
	$1.00^{+0.00}_{-0.06}$	$1.00^{+0.00}_{-0.00}$, $\mathbf{0.00}^{+0.00}_{-0.00}$	$0.73^{+0.24}_{-0.49}$	$\mathbf{0.00}^{+0.00}_{-0.00}$
Helmstetter-Aftershock	0.949	0.937, 0.104	0.890	0.523
	$0.97^{+0.03}_{-0.06}$	$0.96^{+0.03}_{-0.07}$, $0.06^{+0.10}_{-0.04}$	$0.92^{+0.08}_{-0.28}$	$0.48^{+0.27}_{-0.26}$
	$0.44^{+0.39}_{-0.32}$	$0.45^{+0.39}_{-0.33}$, $0.64^{+0.28}_{-0.41}$	$0.79^{+0.20}_{-0.38}$	$0.39^{+0.41}_{-0.32}$
Kagan-Aftershock	0.895	0.899, 0.193	0.901	0.793
	$0.95^{+0.04}_{-0.10}$	$0.96^{+0.03}_{-0.15}$, $0.10^{+0.09}_{-0.06}$	$0.88^{+0.11}_{-0.26}$	$0.97^{+0.02}_{-0.23}$
	$0.57^{+0.33}_{-0.42}$	$0.54^{+0.35}_{-0.43}$, $0.60^{+0.34}_{-0.40}$	$0.76^{+0.22}_{-0.49}$	$0.99^{+0.01}_{-0.20}$
Shen-Aftershock	0.896	0.854, 0.262	0.908	0.981
	$0.92^{+0.06}_{-0.12}$	$0.89^{+0.08}_{-0.16}$, $0.20^{+0.20}_{-0.14}$	$0.88^{+0.11}_{-0.27}$	$0.99^{+0.01}_{-0.10}$
	$0.43^{+0.43}_{-0.33}$	$0.45^{+0.41}_{-0.38}$, $0.69^{+0.28}_{-0.42}$	$0.76^{+0.22}_{-0.50}$	$0.92^{+0.07}_{-0.49}$

*Same as in Table 1, for RELM mainshock + aftershock forecasts. Bold values indicate that the observation is inconsistent with the forecast. For each forecast: first line, quantile scores for the observed target catalog; second line, median and 95% confidence interval for perturbed catalogs with magnitude noise parameter $\nu = 0.1$; third line, same as second line but for $\nu = 0.3$.

should, the N-test δ_2 values match the δ values reported by Schorlemmer *et al.* (2010) and indicate that the Holliday-PI, Ward-Combo, Ward-Geodetic81, Ward-Geologic, and Ward-Seismic forecasts significantly overestimated the total number of target earthquakes during the first half of the mainshock experiment. Likewise, the Bird-Neokinema and Ebel-Aftershock forecasts overpredicted during the first half of the mainshock + aftershock experiment. Based on the δ_1 values, no forecast in either group significantly underpredicted the total number of earthquakes.

The κ values in Tables 1 and 2 indicate that the observed magnitude distributions are not inconsistent with any forecast in either group. Figure 1 shows the discrete magnitude probability distribution (both linearly and in logarithmic scale) for each forecast in the mainshock group; Figure 2 presents the same for the mainshock + aftershock group. The linear plots (white bars with black borders) in these figures show examples of $\mathbf{\Lambda}^m$ from equation 16, although they have been normalized to yield a probability density function and thus have unit area. For visual comparison, Figure 3 shows the observed magnitude distribution for each group; these are examples of $\mathbf{\Omega}^m$ from equation 16. From the linear plots in Figures 1 and 2, it is evident that most of the magnitude forecasts are very similar, owing to the fact that almost all of the forecasts incorporate an exponential scaling between earthquake magnitude and probability; this scaling is based on the well-known empirical Gutenberg–Richter relation (Gutenberg and Richter, 1944). From the logarithmic plots, slight variations in the distributions are noticeable, particularly at large magnitudes. For example, for all but one of his forecasts (Ward-Simulation), Ward (2007) truncated the magnitude distribution, and this implies an absolute maximum event size. For both groups, the Helmstetter *et al.* (2007) and Kagan *et al.* (2007) models employ a tapered distribution, and the Ebel-Aftershock model indicates that all events of magnitude 7.15 and above are equally likely. Of particular note is the magnitude distribution of the Ward-Simulation mainshock model, which forecasts fewer events near M 5 than near M 5.5. This forecast is based on an earthquake simulator that does not enforce Gutenberg–Richter scaling. Rather, it simulates the magnitude distribution based on tectonic loading and fault segment interaction (Ward, 2007). The Ward-Simulation forecast obtains the lowest value of κ (0.541), but the value is not so low as to be rejected by the M-test. We note that, because this forecast does not provide values for the entire spatial extent of the testing region, the current M-test results for this forecast are based on only two observed earthquakes. If tested against a large number of earthquakes conforming to the classical Gutenberg–Richter relation, the Ward-Simulation magnitude forecast is likely to be rejected. We note that the M-test could be an effective tool for discerning between a forecast using a Gutenberg–Richter scaling and a forecast using a characteristic earthquake distribution. Moreover, given a large number of observed target earthquakes, the M-test might be able to distinguish some of the nuances of the Gutenberg–Richter tail behavior.

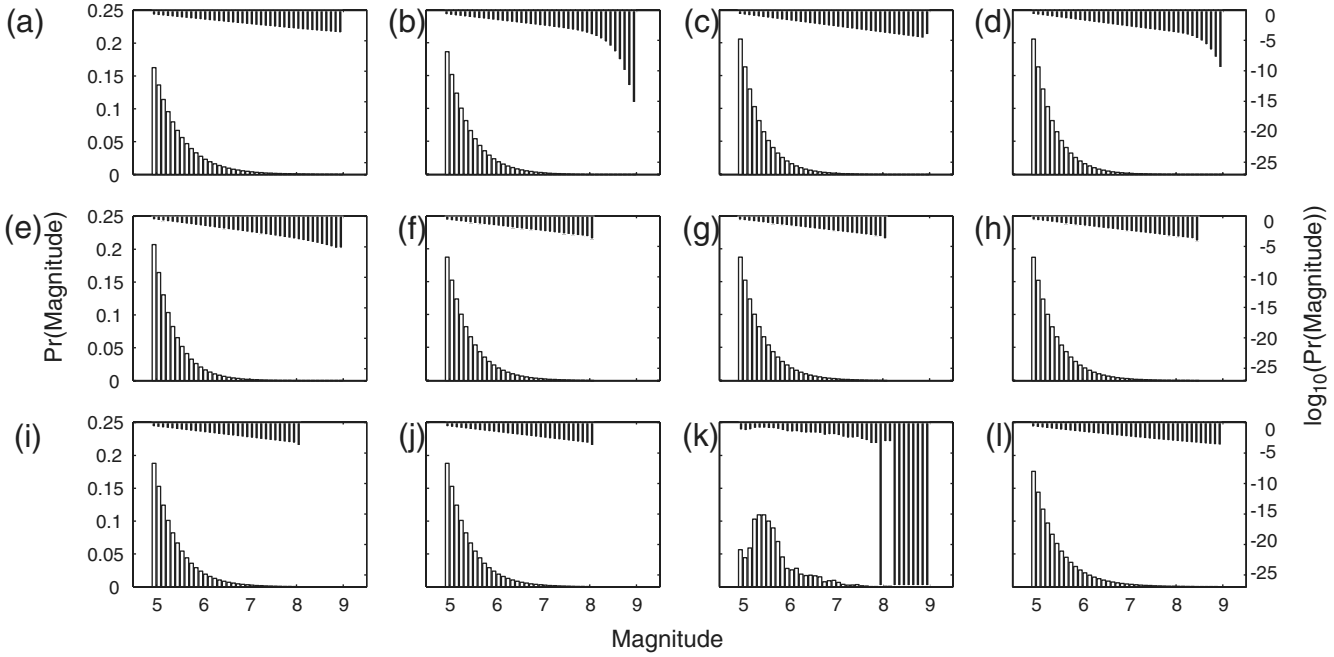

Figure 1. Magnitude probability distributions for RELM mainshock forecasts: (a) Ebel-Mainshock; (b) Helmstetter-Mainshock; (c) Holliday-PI; (d) Kagan-Mainshock; (e) Shen-Mainshock; (f) Ward-Combo81; (g) Ward-Geodetic81; (h) Ward-Geodetic85; (i) Ward-Geologic; (j) Ward-Seismic; (k) Ward-Simulation; and (l) Wiemer-ALM. Distributions are discrete and were specified in 41 bins, each having a width of 0.1 magnitude units. White bars, linear distribution (scale on left ordinate axis); black bars, base-10 logarithm of the distribution (scale on right ordinate axis).

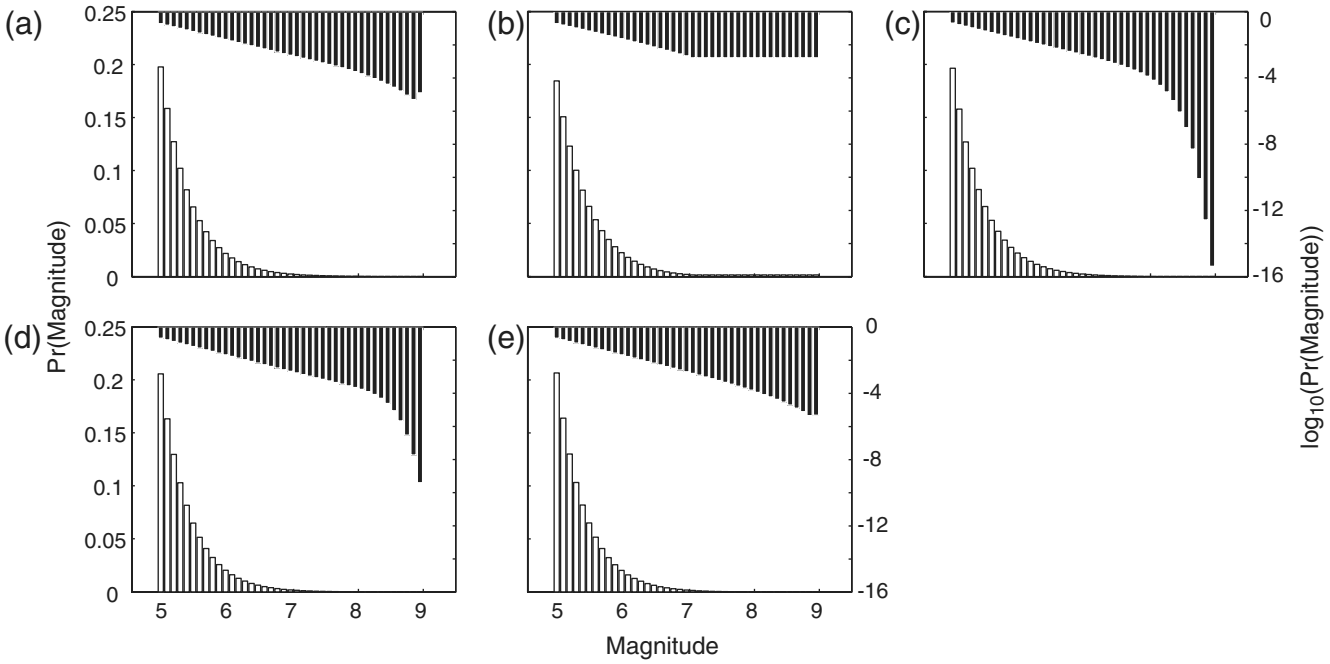

Figure 2. Same as in Figure 1, for RELM mainshock + aftershock forecasts: (a) Bird-Neokinema; (b) Ebel-Aftershock; (c) Helmstetter-Aftershock; (d) Kagan-Aftershock; and (e) Shen-Aftershock.

For plots of the RELM target earthquakes and the spatial forecasts, the latter of which are examples of $\mathbf{\Lambda}^s$ from equation 20, we refer the reader to figures 1–3 of Schorlemmer *et al.* (2010). Based on the ζ values in Table 1, the observed spatial distribution of mainshocks is inconsistent with the Ebel-Mainshock, Holliday-PI, and Wiemer-ALM forecasts. The observed spatial distribution of all M 4.95+ events in California during 2006 to 1 July 2008 is inconsistent with Ebel-Aftershock forecast. We note that, after the RELM forecast submission deadline, the Ebel-Mainshock and Ebel-Aftershock forecasts were recognized to contain systematic spatial errors, and corrected versions of these forecasts were submitted at that time. However, the strictly prospective RELM experiment continues with the original submissions, and we do not consider the corrected forecast groups here. The fact that several models fail the S-test while none fail the L-test lends weight to the assertion that the L-test blends the space, magnitude, and rate components and thereby provides a less detailed (and effectively weaker) evaluation of the forecasts. Indeed, in this regard, the example of the Wiemer-ALM model is instructive.

Figure 4 shows the distribution of simulated joint log-likelihoods for the Wiemer-ALM model for both the L-test and the S-test: $\{\hat{L}\}$ and $\{\hat{S}\}$, using the notation described previously. We plot the empirical cumulative distribution function of each set after subtracting from each member the respective observed joint log-likelihood, $L(\mathbf{\Omega}|\mathbf{\Lambda})$ or $L(\mathbf{\Omega}^s|\mathbf{\Lambda}^s)$. This figure illustrates the L-test and S-test results for Wiemer-ALM; the quantile scores γ and ζ are obtained by finding the cumulative probability when the normalized joint log-likelihood is zero. It is evident that ζ falls in the shaded critical region, indicating that the observed spatial distribution is inconsistent with the spatial forecast. For this forecast, we can understand the S-test failure; as reported by Schorlemmer *et al.* (2010), one of the target earthquakes in the mainshock experiment occurred within a spatial bin where the Wiemer-ALM forecast expected only 8.52×10^{-8} earthquakes. Because this value is so small (orders of magnitude smaller than the other spatial bins where earthquakes occurred), this earthquake dominates the observed likelihood score, and virtually no simulated catalog would contain an event in a bin with such a low forecast rate. Therefore, nearly all spatial simulated joint log-likelihoods are greater than the spatial observed joint log-likelihood, and the observation is deemed to be inconsistent with the spatial forecast.

How, then, is the Wiemer-ALM model not rejected by the L-test? Figure 4 provides a visual explanation: the distribution of simulated joint log-likelihoods is much wider

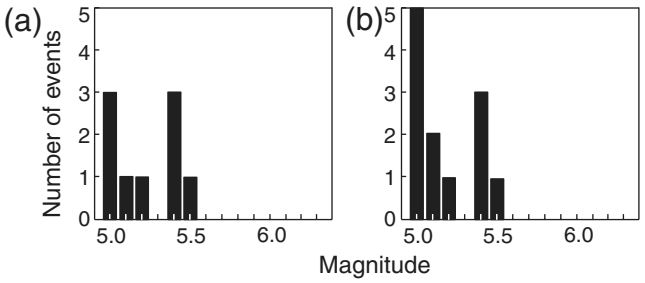

Figure 3. Observed magnitude distribution for RELM target earthquakes between 1 January 2006 and 1 July 2008. (a) Distribution of 9 events in the mainshock experiment. (b) Distribution of 12 events in the mainshock + aftershock experiment.

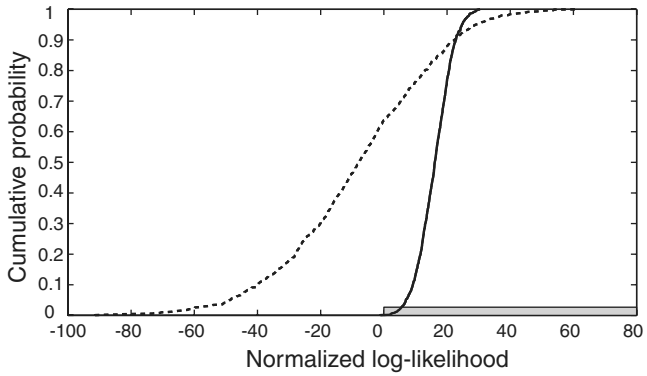

Figure 4. Comparison of L- and S-test results for the Wiemer-ALM forecast. Dashed line is the empirical cumulative distribution for normalized L-test simulated log-likelihoods, solid line is the same for S-test. The $\alpha = 0.05$ critical region is represented by the shaded box at the bottom right of the plot. The fact that the S-test curve intersects the critical region indicates that the observation is inconsistent with the forecast. The normalized observed joint log-likelihood is zero for both tests.

for the L-test than for the S-test. This is because the number of events in simulated catalogs is allowed to vary for the L-test, while it is fixed at N_{obs} in the M- and S-tests. For the case of the Wiemer-ALM model L-test simulations, it is generally true that when $N_{sim} \leq N_{obs}$, the simulated joint log-likelihood is greater than the observed joint log-likelihood; when $N_{sim} > N_{obs}$, the reverse is true. Because these cases are approximately equally likely, the variation in N_{sim} results in the observed joint log-likelihood falling roughly in the middle of the distribution of simulated joint log-likelihoods, allowing the forecast to pass the L-test. In the case of the S-test, N_{sim} is fixed and the simulated joint log-likelihoods are almost uniformly larger than the observed joint log-likelihood, leading to the forecast's rejection.

Stability of Results

Measurements of earthquake magnitude and epicenter are imperfect; and thus, because they depend on these measurements, the test results described previously are subject to uncertainty. To consider the stability of these results with respect to measurement errors, knowledge regarding the form and scale of the errors is necessary. Unfortunately, rigorous estimation of magnitude and epicenter uncertainties has not yet become part of routine processing used to construct an earthquake catalog from waveforms; we therefore made the simplifying assumption that the uncertainties for each observed earthquake could be treated identically. We assumed that epicenter errors are isotropic and normally distributed with a standard deviation $\sigma = 5$ km. Based on the catalog analysis of Werner and Sornette (2008), which indicated surprisingly large, non-Gaussian errors, we assumed that magnitude errors are well-described by a Laplace distribution with a scale parameter ν that takes a value between 0.1 and 0.3.

Under these assumptions, we explored the stability of the mainshock + aftershock test results using perturbed catalogs. We perturbed the catalog of observed events with magnitude $M_{ANSS} \geq 2.0$, origin times between 1 January 2006 and 1 July 2008, and epicenters in the RELM collection region (table 3 of Schorlemmer and Gerstenberger, 2007). The epicenter and magnitude of each earthquake were modified based on random sampling of the assumed error distribution, respectively, and the perturbed catalog was the result of filtering that preserved only the events that, after modification, had a magnitude of at least 4.95 and fell in the RELM testing region (table 2 of Schorlemmer and Gerstenberger, 2007). For each perturbed catalog, we computed the quantile score for each test (L-, N-, M-, and S-). We repeated this entire process for 200 perturbed catalogs, 100 perturbed with $\nu = 0.1$ and 100 with $\nu = 0.3$.

In Table 2, we report the median perturbed quantile scores and the intervals that contain 95% of the perturbed scores; we suggest that the size of these intervals is an apposite measure of test result stability, with smaller intervals indicating greater stability. (Ⓔ Details of result stability are available in the electronic edition of *BSSA*, and the full distributions of perturbed quantile scores are available in Figures S1 and S2.) We note that the intervals are almost always larger when $\nu = 0.3$ than when $\nu = 0.1$; this should be expected because the magnitudes in the catalogs based on $\nu = 0.3$ are more perturbed and therefore yield wider variation in the perturbed quantile scores. The effect of measurement uncertainty varies from forecast to forecast and from test to test, but we note that in the cases where the original observation was deemed inconsistent with the forecast (bold values in a forecast's first row), most frequently the perturbed catalogs would also be considered inconsistent. Conversely, in nearly every case where the observation is not inconsistent with the forecast, almost all perturbed catalogs would also be considered not inconsistent.

One could analyze similarly the stability of the mainshock test results, with the additional complication of declustering each catalog after modification and before filtering to the RELM testing region and target earthquake magnitude range.

Statistical Power of Tests

Statistical power is the probability of (correctly) rejecting an incorrect null hypothesis (Lehman and Romano, 2005). Power can be thought of as the ability of a test to distinguish between hypotheses, and, in general, power depends on the hypotheses being considered and the number of observations. Of course, in earthquake forecasting experiments, it is not known whether a given hypothesis (in this case, formulated as a space–rate–magnitude forecast) is incorrect, so therefore we have relied on simulations to estimate statistical power. Consider two forecasts, $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$: we took $\mathbf{\Lambda}_1$ to be the "true" model of seismicity and simulated 10,000 catalogs consistent with $\mathbf{\Lambda}_1$, using the

approach described in the L-Test section. We applied each of the tests (L-, M-, N-, and S-) to each simulated catalog, using $\mathbf{\Lambda}_2$ to forecast. The fraction of instances that indicate that the simulated $\mathbf{\Lambda}_1$ catalog is inconsistent with the $\mathbf{\Lambda}_2$forecast is the estimate of power. We obtained distinct estimates of power for each test and for each pair of forecasts, and we note that, except in the case of the N-test, power is not symmetric; in other words, a test may have greater power when $\mathbf{\Lambda}_1$ is the true model and $\mathbf{\Lambda}_2$ the forecast than when the opposite is true.

The statistical power of the N-test is intuitive: it is difficult to distinguish two rate forecasts with very similar expectations (low power), and it becomes increasingly easy to distinguish rate forecasts as the difference between their expectations increases (power increases). Moreover, the power of the N-test depends only on the number of earthquakes forecast by each model and does not require simulated catalogs. Consider two forecasts $\mathbf{\Lambda}_1$ and $\mathbf{\Lambda}_2$ with overall rate expectations λ_1 and λ_2, respectively. The power of the N-test is the probability of a correct rejection; this is equivalent to the probability that either δ_1 or δ_2 is smaller than α_{eff},

$$\sum_{\{i:\delta_1(i|\lambda_2)<\alpha_{\text{eff}}\}} \Pr(i|\lambda_1) + \sum_{\{j:\delta_2(j|\lambda_2)<\alpha_{\text{eff}}\}} \Pr(j|\lambda_1), \qquad (24)$$

where the summation terms are described by equation 9. We applied equation 24 to every possible pairwise comparison of forecasts; for each comparison, we considered only the bins in which both models make a forecast statement (the forecast overlap region). In Tables 3 and 4, we report the number of events expected by each forecast in each forecast overlap region for the first half of the mainshock and mainshock + aftershock experiments, respectively. In Tables 5 and 6, we report the corresponding values of N-test power, assuming $\alpha_{\text{eff}} = 0.025$. These values match intuition: for example, the N-test has very little chance of distinguishing between the Ebel-Mainshock and Wiemer-ALM forecasts, which (as seen in Table 3) expect 8.651 and 8.562 earthquakes, respectively. On the other hand, the N-test has high power with respect to the Holliday-PI and Kagan-Mainshock forecasts, which expect 10.188 and 1.544 events, respectively. Note that these results are based on the first half of the RELM experiments, and estimates of N-test power will increase as the experiment proceeds (because the forecasts are time-invariant, forecast rate differences in the first 2.5 years will be doubled at the end of the experiment); in Tables S1 and S2, we report the power of the N-test for the full five-year period (Ⓔdetails of result stability are available in the electronic edition of *BSSA*).

In Tables S3–S8 of Ⓔthe electronic edition of *BSSA*, we report the estimates of power for every pair of forecasts for each test; these estimates are also conditioned on the number of observed target earthquakes in each forecast overlap region. Here, we present one example in detail: the M-test applied to the Helmstetter-Mainshock and Ward-Simulation forecasts. In general, because most of the forecasts considered here employ some variant of the Gutenberg–Richter relation as the magnitude forecast (Figs. 1 and 2), we expect the M-test to have low power. Therefore, for this example, we chose one "typical" magnitude forecast (Helmstetter-Mainshock) and one that has a quite different visual representation (Ward-Simulation). Using the method described previously, we estimated the power of the M-test using simulations in which we varied the number of target earthquakes; the results are shown in Figure 5. It is clear from this figure that power is not symmetric; in a sense, it is easier to distinguish the forecasts if the Ward-Simulation is the correct model than if the Helmstetter-Mainshock is. As expected, power increases with the number of observations.

Discussion and Conclusions

The ongoing five-year RELM experiment in California is an ambitious attempt to improve seismic hazard analysis through a competitive process; rather than constructing a

Table 3

Expected Rates in Forecast Overlap Regions for RELM Mainshock Forecasts*

	$\mathbf{\Lambda}_2$											
$\mathbf{\Lambda}_1$	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Ebel-Mainshock	**8.651**	8.651	2.998	4.772	4.858	3.367	3.367	3.374	3.367	3.367	3.378	8.651
2. Helmstetter-Main.	7.625	**10.553**	4.178	5.872	5.948	4.296	4.296	4.296	4.296	4.296	4.296	10.553
3. Holliday-PI	11.430	14.389	**14.389**	10.188	10.662	9.391	9.391	9.395	9.391	9.391	9.397	14.389
4. Kagan-Mainshock	3.611	5.987	1.544	**5.987**	5.772	4.641	4.641	4.642	4.641	4.641	4.643	5.987
5. Shen-Mainshock	3.384	5.225	1.317	4.981	**5.225**	3.906	3.906	3.907	3.906	3.906	3.907	5.225
6. Ward-Combo	6.136	9.461	3.244	9.461	9.404	**9.461**	9.461	9.461	9.461	9.461	9.461	9.461
7. Ward-Geodetic81	7.659	12.123	3.486	12.123	12.001	12.123	**12.123**	12.123	12.123	12.123	12.123	12.123
8. Ward-Geodetic85	4.411	6.982	2.008	6.982	6.912	6.977	6.977	**6.982**	6.977	6.977	6.982	6.982
9. Ward-Geologic	5.187	8.315	2.840	8.315	8.297	8.315	8.315	8.315	**8.315**	8.315	8.315	8.315
10. Ward-Seismic	5.564	7.943	3.407	7.943	7.913	7.943	7.943	7.943	7.943	**7.943**	7.943	7.943
11. Ward-Simulation	2.443	3.718	1.418	3.718	3.702	3.716	3.716	3.718	3.716	3.716	**3.718**	3.718
12. Wiemer-ALM	8.562	11.843	4.021	6.554	5.714	3.496	3.496	3.499	3.496	3.496	3.501	**11.843**

*Bold values indicate the total number of target earthquakes expected by each $\mathbf{\Lambda}_1$ forecast.

Table 4
Expected Rates in Forecast Overlap Regions for RELM Mainshock + Aftershock Forecasts*

Λ_1	Λ_2 1.	2.	3.	4.	5.
1. Bird-NeoKinema	**27.921**	17.335	27.921	15.741	15.714
2. Ebel-Aftershock	36.362	**36.362**	36.362	19.946	20.323
3. Helmstetter-Aftershock	17.682	12.776	**17.682**	9.838	9.966
4. Kagan-Aftershock	7.982	4.815	7.982	**7.982**	7.696
5. Shen-Aftershock	7.316	4.737	7.316	6.973	**7.316**

*Bold values indicate the total number of target earthquakes expected by each Λ_1 forecast.

complex consensus model, working group participants were invited to submit testable earthquake forecasts that they considered to be worthy of study. Perhaps more importantly, the working group established rules by which all experiment participants should abide (Field, 2007); designed a set of evaluation criteria to judge the predictive skill of each forecast (Schorlemmer *et al.*, 2007); and provided a vision for a controlled computational environment in which to run the experiment (Schorlemmer and Gerstenberger, 2007). CSEP has expanded upon the RELM mission, and several testing centers are now operational worldwide. There are tens of experiments in progress within these testing centers, and most follow the RELM design. Despite the success of the RELM project in terms of inspiring simultaneous multiple-investigator studies of earthquake predictability, the experiment design is not perfect or comprehensive. In this article, we discussed one logical flaw in the N-test, and we have proposed two additional tests to evaluate the forecasts more comprehensively. We have also, for the first time, probed the stability of the results with respect to measurement uncertainty and explored the statistical power of each of the tests.

It is worth noting that the daily forecast experiments being considered within CSEP testing centers constitute a special case; depending on the testing region, it can be quite common that no target earthquake is observed during most days of the experiment. For such days, the M- and S-tests are not applicable because there is neither an observed magnitude distribution nor an observed spatial distribution. It is also worth remembering that using a single significance value of 5% for each test does not correspond to 95% confidence in the combined result. We have presented each test independently, using the RELM forecasts for illustration, rather than emphasizing any specific or combined outcome. When one has in mind a desired confidence level for the combined result of applying the four tests discussed, a correction for multiple tests, such as a Bonferoni correction, should be made.

There are variations on the tests proposed here that are equally valid and could augment the CSEP analysis of the performance of forecasting models in certain circumstances. For example, the proposed M-test compares the likelihood of the observed distribution of earthquake magnitudes with that of simulated catalogs conforming to the marginal distribution of earthquake magnitudes under the model, given the number of earthquakes in the actual catalog. An equally valid test would be to compare the likelihood of the observed distribution of earthquake magnitudes with that of simulated catalogs conforming to the conditional distribution of earthquake magnitudes under the model, given both the number and spatial distribution of earthquakes in the actual catalog. This conditional M-test differs from the proposed M-test if the magnitude distribution varies spatially under the model, and so it could be useful in cases where such variation is large. A conditional S-test could be defined analogously and would be useful under similar circumstances.

Table 5
Power of N-Test for RELM Mainshock Forecasts

Λ_1	Λ_2 1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. Ebel-Mainshock	0.024	0.066	0.816	0.054	0.119	0.151	0.346	0.037	0.035	0.035	0.056	0.024
2. Helmstetter-Mainshock		0.043	0.870	0.037	0.043	0.378	0.737	0.072	0.198	0.198	0.045	0.066
3. Holliday-PI			0.033	0.974	0.981	0.720	0.595	0.906	0.720	0.720	0.957	0.887
4. Kagan-Mainshock				0.037	0.037	0.319	0.679	0.055	0.158	0.158	0.057	0.046
5. Shen-Mainshock					0.024	0.452	0.800	0.099	0.252	0.252	0.039	0.035
6. Ward-Combo						0.032	0.092	0.100	0.063	0.063	0.603	0.538
7. Ward-Geodetic81							0.042	0.332	0.240	0.240	0.853	0.858
8. Ward-Geodetic85								0.020	0.036	0.036	0.269	0.136
9. Ward-Geologic									0.034	0.034	0.452	0.322
10. Ward-Seismic										0.031	0.400	0.322
11. Ward-Simulation											0.038	0.040
12. Wiemer-ALM												0.041

Table 6
Power of N-Test for RELM Mainshock + Aftershock Forecasts

Λ_1	Λ_2 1.	2.	3.	4.	5.
1. Bird-NeoKinema	0.037	0.951	0.595	0.608	0.702
2. Ebel-Aftershock		0.038	0.998	0.989	0.996
3. Helmstetter-After.			0.042	0.078	0.136
4. Kagan-Aftershock				0.031	0.030
5. Shen-Aftershock					0.041

Moreover, tests based on likelihood measures are not the only option for the class of forecasts considered here. Approaches based on the Molchan error diagram (Molchan, 1991, Molchan and Kagan, 1992) are also applicable; for example, Zechar and Jordan (2008, 2010) suggested a procedure using what they called the area skill score, which is derived from a Molchan trajectory. Kagan (2007, 2009) described the relationship between error diagrams, likelihood scores, and other information scores. Alternative tests should be developed further for large-scale earthquake predictability experiments, with a particular emphasis on feature identification and with the development of hybrid forecast models in mind (e.g., Rhoades and Gerstenberger, 2009).

The RELM forecast format (Poisson rates specified in bins of latitude, longitude, and magnitude) also should be reconsidered and expanded. Particularly contentious are the requirements that seismicity rates be Poisson distributed and that the rate in each bin be independent of the rates in all other bins. Kagan (1973a, 1973b) provided a theoretical basis for why the negative binomial distribution may be preferred for modeling seismicity rates, and several studies have shown that the negative binomial provides a better fit than Poisson to catalog data, even when accounting for the additional degree of freedom (e.g., Jackson and Kagan, 1999; Kagan and Jackson, 2000; Schorlemmer *et al.*, 2010). It could be, however, that rate distributions vary in space; and, more generally, there seems to be no real advantage to restricting rate forecasts to follow a particular analytical form. Therefore, we agree with Werner and Sornette (2008) that future experiments should consider forecasts that may specify an arbitrary probability distribution, discrete or otherwise, in each bin. It is straightforward to modify the evaluation metrics considered in this article to accommodate arbitrary probability distributions. The problem of bin independence is not solved so easily; it is generally thought that the dependence is conditional on earthquake occurrence. For example, we expect many small events to occur in the wake of and close to a large earthquake, but a prospective forecast experiment requires the forecast and any interbin dependence to be provided in advance of the data meant to be predicted, before the knowledge of the large earthquake is available. In light of this, perhaps more emphasis should be given to short-term forecasts that can respond to major earthquakes; another approach would be to design forecasts that are updated programmatically after every earthquake of some minimum magnitude, rather than the forecasts being updated at fixed intervals.

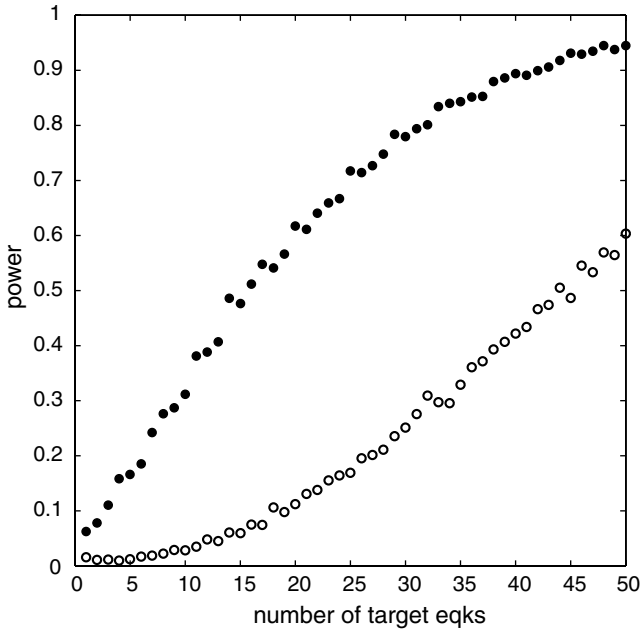

Figure 5. Power of the M-test as a function of the number of simulated target earthquakes. Empty circles, power when Helmstetter-Mainshock is taken to be the true model and Ward-Simulation is used to forecast the simulated catalogs; filled circles, power in the opposite case.

Data and Resources

All data used in this study came from published sources that are listed in the References section.

Acknowledgments

We thank Masha Liukis for assistance with computations. We thank Jochen Woessner for encouragement and a careful reading of the manuscript, and we thank Alejandro Veen and Max Werner for stimulating discussion. We thank Christine Smyth, Iain Bailey, Editor Andy Michael, and two anonymous reviewers for reading early drafts of the manuscript and providing constructive comments. This research was supported by the Southern California Earthquake Center (SCEC). The SCEC is funded by the National Science Foundation Cooperative Agreement EAR-0106924 and U.S. Geological Survey Cooperative Agreement 02HQAG0008. The SCEC contribution number for this paper is 1340.

References

Bird, P., and Z. Liu (2007) Seismic hazard inferred from tectonics: California, *Seismol. Res. Lett.* **78,** 37–48.

Ebel, J. E., D. W. Chambers, A. L. Kafka, and J. A. Baglivo (2007). Non-Poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California, *Seismol. Res. Lett.* **78,** 57–65.

Evison, F. F., and D. A. Rhoades (1999). The precursory earthquake swarm in Japan: Hypothesis test, *Earth Planets Space* **51,** 1267–1277.

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

Gutenberg, B., and C. F. Richter (1944). Frequency of earthquakes in California, *Bull. Seismol. Soc. Am.* **34,** 185–188.

Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2007). High-resolution time-independent grid-based forecast for $M \geq 5$ earthquakes in California, *Seismol. Res. Lett.* **78,** 78–86.

Holliday, J. R., C. Chen, K. F. Tiampo, J. B. Rundle, D. L. Turcotte, and A. Donnellan (2007). A RELM earthquake forecast based on pattern informatics, *Seismol. Res. Lett.* **78,** 87—93.

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. (1973a). A probabilistic description of the seismic regime, *Izvestiya Phys. Solid Earth* **4,** 213–219.

Kagan, Y. Y. (1973b). Statistical methods in the study of seismic processes, *Bull. Int. Statist. Inst.* **45,** no. 3, 437–453.

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

Kagan, Y. Y., and D. D. Jackson (1995). New seismic gap hypothesis: Five years after, *J. Geophys. Res.* **100,** no. B3, 3943–3959.

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

Kagan, Y. Y., D. D. Jackson, and Y. Rong (2007). A testable five-year forecast of moderate and large earthquakes in southern California based on smoothed seismicity, *Seismol. Res. Lett.* **78,** 94–98.

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

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

Nishenko, S. P. (1991). Circum-Pacific seismic potential: 1989–1999, *Pure Appl. Geophys.* **135,** no. 2, 169–259.

Rhoades, D. A., and M. C. Gerstenberger (2009). Mixture models for improved short-term earthquake forecasting, *Bull. Seismol. Soc. Am.* **99,** no. 2A, 636–646.

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,** nos. 8/9, (in press).

Shen, Z.-K., D. D. Jackson, and Y. Y. Kagan (2007). Implications of geodetic strain rate for future earthquakes, with a five-year forecast of M 5 earthquakes in southern California, *Seismol. Res. Lett.* **78,** 116–120.

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

Werner, M. J., and D. Sornette (2008). Magnitude uncertainties impact seismic rate estimates, forecasts, and predictability experiments, *J. Geophys. Res.* **113,** no. B08302, doi 10.1029/2007JB005427.

Wiemer, S., and D. Schorlemmer (2007). ALM: An asperity-based likelihood model for California, *Seismol. Res. Lett.* **78,** 134–140.

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

Zechar, J. D., and T. H. Jordan (2010). The area skill score statistic for evaluating earthquake predictability experiments, *Pure Appl. Geophys.*, **167,** nos. 8/9, (in press).

Zechar, J. D., D. Schorlemmer, M. Liukis, J. Yu, F. Euchner, P. J. Maechling, and T. H. Jordan (2009). The Collaboratory for the Study of Earthquake Predictability perspective on computational earthquake science, *Concurr. Comp-Pract. E.* doi 10.1002/cpe.1519.

Lamont-Doherty Earth Observatory of Columbia University
61 Route 9W
Palisades, New York 10964
zechar@ldeo.columbia.edu
(J.D.Z.)

GNS Science
P.O. Box 30368
Lower Hutt, New Zealand
m.gerstenberger@gns.cri.nz
d.rhoades@gns.cri.nz
(M.C.G., D.A.R.)

Manuscript received 27 July 2009