Hindawi Publishing Corporation
International Journal of Geophysics
Volume 2012, Article ID 543482, 8 pages
doi:10.1155/2012/543482

Research Article

Evaluating the RELM Test Results

Michael K. Sachs,[1] Ya-Ting Lee,[2] Donald L. Turcotte,[3] James R. Holliday,[1] and John B. Rundle[1,3,4]

[1] *Department of Physics, University of California, Davis, Davis CA 95616, USA*
[2] *Graduate Institute of Geophysics, National Central University, Jhougli 320, Taiwan*
[3] *Department of Geology, University of California, Davis, Davis CA 95616, USA*
[4] *Theory Section, Santa Fe Institute, Santa Fe, NM 87501, USA*

Correspondence should be addressed to Michael K. Sachs, mksachs@ucdavis.edu

Received 15 July 2011; Revised 2 December 2011; Accepted 19 December 2011

Academic Editor: RodolÌfo Console

We consider implications of the Regional Earthquake Likelihood Models (RELM) test results with regard to earthquake forecasting. Prospective forecasts were solicited for $M \geq 4.95$ earthquakes in California during the period 2006–2010. During this period 31 earthquakes occurred in the test region with $M \geq 4.95$. We consider five forecasts that were submitted for the test. We compare the forecasts utilizing forecast verification methodology developed in the atmospheric sciences, specifically for tornadoes. We utilize a "skill score" based on the forecast scores λ_{fi} of occurrence of the test earthquakes. A perfect forecast would have $\lambda_{fi} = 1$, and a random (no skill) forecast would have $\lambda_{fi} = 2.86 \times 10^{-3}$. The best forecasts (largest value of λ_{fi}) for the 31 earthquakes had values of $\lambda_{fi} = 1.24 \times 10^{-1}$ to $\lambda_{fi} = 5.49 \times 10^{-3}$. The best mean forecast for all earthquakes was $\overline{\lambda}_f = 2.84 \times 10^{-2}$. The best forecasts are about an order of magnitude better than random forecasts. We discuss the earthquakes, the forecasts, and alternative methods of evaluation of the performance of RELM forecasts. We also discuss the relative merits of alarm-based versus probability-based forecasts.

1. Introduction

Earthquakes do not occur randomly in space. Large earthquakes occur preferentially in regions where small earthquakes occur. Earthquakes are complex phenomena, but they do obey several scaling laws. One example is Gutenberg-Richter frequency-magnitude scaling. The cumulative number of earthquakes N with magnitudes greater than M in a region over a specified period of time is well approximated by the relation

$$\log N = a - bM, \tag{1}$$

where b is a near universal constant in the range $0.8 < b < 1.1$ and a is a measure of the level of seismicity. Small earthquakes can be used to determine a and (1) can be used to determine the probability of occurrence of large earthquakes. Kossobokov et al. [1] utilized the number of $M \geq 4$ earthquakes in $1^\circ \times 1^\circ$ areas to map the global seismic hazard.

A question that has been studied by many groups is whether there are temporal variations in seismicity that can be used to forecast the occurrence of future earthquakes. Earthquakes on major faults (say the San Andreas in California) occur quasiperiodically. A reasonable hypothesis would be that the rate of regional seismicity would accelerate during the period between the major earthquakes. There is no evidence that this occurs systematically. Background seismicity in California appears to be stationary. With the exception of years with large aftershock sequences, Rundle et al. [2] (Figure 1) showed that seismic activity in Southern California in the magnitude range $1.5 < m < 4$ for the period 1983 to 2000 was well represented on a yearly basis by (1) taking $a = 5.4$ and $b = 1.0$.

Intermediate-term earthquake forecasting algorithms based on pattern recognition of variations in regional seismicity were developed by Keilis-Borok and colleagues [3]. These forecasts were alarm based, when a threshold of anomalous behavior was reached a warning of a time of

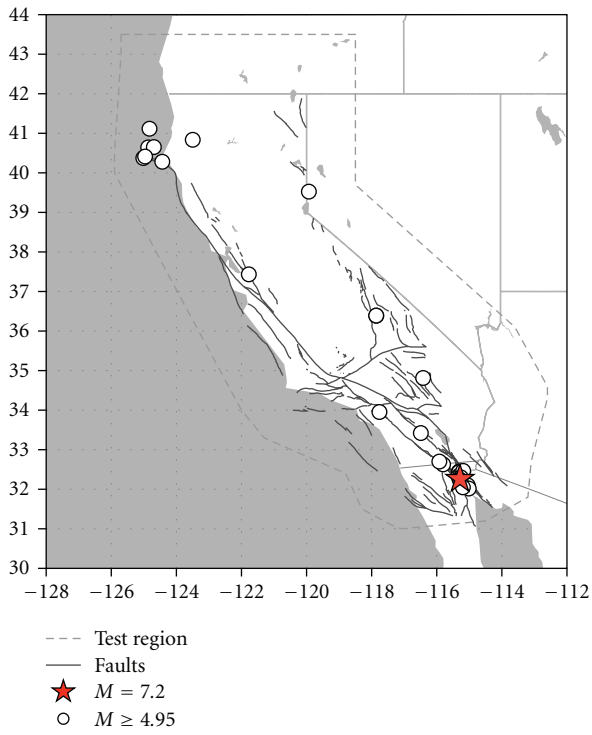

Figure 1: Map of the test region, the coast of California, major faults, and the 31 earthquakes with $M \geq 4.95$ that occurred in the test region. The earthquakes are given in Table 1.

increasing probability (TIP) of an earthquake was issued. A relatively high success rate was found including the 1988 Armenian earthquake and the 1989 Loma Prieta earthquake [4], but there were also notable false alarms and failures to predict.

The focus of this paper is to study the implications of the RELM test of earthquake forecasts in California. This was a prospective test of forecasts for $m > 5$ earthquakes during the period 2006–2010. Forecast submission was required prior to the starting date. In our study of the RELM test results we will utilize the methodology developed in the atmospheric sciences [5], specifically for tornadoes. Tornado forecasts are alarm based. Two levels of alarms are issued: (1) a tornado watch is issued for a specified area and time if atmospheric conditions appear conducive to tornados, (2) a tornado warning is issued if one or more tornados have been observed. The evaluation of tornado forecasts is based on the number of failures to predict and on the number of false alarms. A quantitative measure of success is the skill score, the skill score is unity for a perfect forecast and zero for a random (no skill) forecast. RELM forecasts were probabilistic rather than alarm based, that is a continuous range forecast probabilities were required. In an alarm-based forecast an area of high risk is specified. We will discuss the implications of the two alternative approaches.

The forecasts submitted to the RELM test were primarily based on precursory seismic activity. There are a variety of approaches to the quantification of this activity. In Section 2 of this paper we will discuss the relative intensity (RI) and pattern informatics (PI) approaches. The RI approach extrapolates the occurrence of small earthquakes during a specified precursory time window. High activity (activation) indicates high risk. The PI approach is related but includes both activation and quiescence. In Section 3 the problems with retrospective forecasts are discussed. In Section 4 the RELM test is discussed and the test earthquakes are described in Section 5. The submitted forecasts are discussed in Section 6 and are evaluated in Section 7.

An objective of this paper is to understand the relationship of the forecasts to the distribution of seismicity during the test period. We discuss what we believe is a well-defined precursory activation.

2. PI and RI

A pattern informatics (PI) approach to earthquake forecasting was proposed by Rundle et al. [2, 6] and Tiampo et al. [7]. In forecasting $M \geq 5$ earthquakes a region is divided into a grid of $0.1^\circ \times 0.1^\circ$ regions. The rates of seismicity in the regions are studied to quantify anomalous behavior. Precursory changes that include either increases or decreases in seismicity are identified during a prescribed time interval. If changes exceed a prescribed threshold hot-spots are defined. The forecast is that future $M \geq 5$ earthquakes will occur in the hot spot regions in a 10-year time window. Thus, the PI method is alarm based. Utilizing the PI method Rundle et al. [8] made a forecast of California hot spots valid for the period 2000–2010. Holliday et al. [9] reported that 16 of the 18 earthquakes that occurred during the period 2000–2005 occurred in hot spot regions. The PI forecast is time dependent because it is based on temporal changes in background seismicity.

A closely related forecasting technique is the relative intensity (RI) approach. The RI forecast is based on the direct extrapolation of the rate of occurrence of small earthquakes using (1). The RI forecast can be time dependent if the time span of the background seismicity is relatively short. The success of the PI method described above led to a discussion as to whether the PI method is significantly better than the RI method. Comparisons of these approaches have come to different conclusions regarding their validity [10, 11]. These comparisons emphasize the difficulties in evaluating the performance of seismicity forecasts.

3. Prospective versus Retrospective Forecasts

A prospective forecast is a true forecast of future earthquakes. No knowledge of these earthquakes exists. A retrospective forecast is a forecast of earthquakes that have occurred in the past (say 2000–2010) based on data available before the start of the period. The existence of the forecast earthquake is known. In principal a retrospective forecast can be carried out fairly; however, in many cases these forecasts are biased by the existence of the forecast earthquakes.

The PI forecast by Rundle et al. [8] was prospective. However, the successful forecast of 16 out of 18 earthquakes

in California led to a retrospective challenge of the results [11].

It became clear that it would be desirable to sponsor a contest in which research groups would provide prospective forecasts of earthquakes under well-defined conditions. This was the origin of the RELM test, which will be described in the next section. Some of the rules were based on the prospective forecast made by Rundle et al. [8]. The test region was California. Forecasts were made for $M > 5$ earthquakes on a grid of $0.1^\circ \times 0.1^\circ$ forecast cells. The forecast period was 1 January 2006 to 31 December 2010. The results will also be summarized in this paper.

4. RELM Test

In order to test methods for forecasting future earthquakes the Southern California Earthquake Center (SCEC) formed the working group for Regional Earthquake Likelihood Models (RELM) in 2000 [12]. For the first time a competitive test of prospective earthquake forecasts was to be carried out. Research groups were encouraged to submit forecasts of future earthquakes in California. At the end of the test period, the forecasts would be compared with the actual earthquakes that occurred.

The ground rules for the RELM test were as follows.

(1) The test region to be studied was the state of California; however the selected region extended somewhat beyond the boundaries of the state as shown in Figure 1.

(2) The objective was to forecast the largest earthquakes for which a reasonable number could be expected to occur in a reasonable time period. A five-year time period for the test was selected extending from 1 January 2006 to 31 December 2010. Earthquakes with $M \geq 5$ were to be forecast. This magnitude cutoff was chosen because at least 20 $M \geq 5$ earthquakes could be expected in this period. For $M \geq 6$, only about 2 would be expected so the 5-year period would be much too short. The applicable magnitudes were taken from the Advanced National Seismic System (ANSS) online catalog (http://www.ncedc.org/anss/anss-detail.html).

(3) Participants were required to submit the number of earthquakes expected to occur in specified spatial cells and magnitude bins during the test period. In order to do this, the test region was subdivided into $N_c = 7682$ spatial cells with dimensions $0.1^\circ \times 0.1^\circ$ (approximately 10 km × 10 km). These spatial cells were further divided into 41 magnitude bins: $4.95 \leq M < 5.05$, $5.05 \leq M < 5.15$, $5.15 \leq M < 5.25$,..., $8.85 \leq M < 8.95$, and $8.95 \leq M < \infty$. The participants were required to specify the forecast number of earthquakes N_{fmi} in magnitude bin m ($m - 0.05 < M < m + 0.05$) that would occur during the test period in cell i.

It is important to note that the RELM forecasts were continuous (probabilistic) rather than alarm based. The numbers of earthquakes expected to occur in each spatial cell and each magnitude bin was required. Continuous and alarm-based forecasts each have advantages and disadvantages. Continuous forecasts are useful for setting insurance premiums but the numbers of predicted earthquakes are so small that they have little meaning to the general public. Alarm-based forecasts specify where earthquakes are most likely to occur.

Nineteen forecasts were submitted by eight groups. Before discussing these forecasts in some detail we will discuss the earthquakes that occurred in the test region during the test period with $M \geq 4.95$.

5. The Earthquakes

During the test period 1 January 2006 to 31 December 2010, there were $N_e = 31$ earthquakes in the test region with $M \geq 4.95$. The times of occurrence, locations, and magnitudes of these earthquakes are given in Table 1. The locations of the test earthquakes are also shown in Figure 1.

The 31 earthquakes occurred in $N_{ce} = 22$ cells. The association of earthquakes with cells is given in Table 2. Five of the 22 cells had multiple earthquakes. The occurrence of five test earthquakes in cell A is not surprising since this is in the Cerro Prieto geothermal area that is recognized as having a high level of seismicity. Earthquakes occurred in 22 of the 7682 $0.1^\circ \times 0.1^\circ$ test cells in the test area.

The major earthquake that occurred during the test period was the $M = 7.2$ El Mayor-Cucapah earthquake on 4 April 2010 (event 22 in Table 1). This earthquake was on the plate boundary between the North American and Pacific plates. The epicenter was about 50 km south of the Mexico-United States border, but occurred within the test region as shown in Figure 1. Events 23, 24, 25, 26, 27, 28, 29, and 31 are well-defined aftershocks of the El Mayor-Cucapah earthquake. Events 1, 7, 8, 9, 10, 14, 16, and 19 constitute a precursory swarm of eight test earthquakes in this region in the magnitude range 4.97 to 5.80, including four in the 10-day period between 9 February and 19 February 2008 (events 7–10). These events were located some 5 km to 20 km north of the subsequent epicenter of the El Mayor-Cucapah earthquake and lie outside the primary aftershock region of that event. This swarm of earthquakes certainly cannot be considered foreshocks due to their relatively small magnitudes and early occurrence but may represent a seismic activation. We will discuss this activation in terms of AMR later in this paper.

Another swarm of earthquakes occurred in the northwest corner of the test region adjacent to Cape Mendocino. This sequence (events 23, 4, 5, 20, and 21) had magnitudes in the range 5.0 to 6.5. This is a region of high seismicity, and this concentration of events is expected. Event 21 may or may not be an aftershock of event 20. The pair of earthquakes 17 and 18 are interesting. It is very likely that the $M = 5.0$ earthquake on 1 October 2009 was a foreshock of the $M = 5.19$ earthquake on 3 October 2009.

6. Submitted Forecasts

The submitted forecasts have been discussed in some detail [13]. The nineteen forecasts submitted by eight groups are available on the RELM website (http://relm.cseptesting.org/). In order to have a common basis for comparison, we will only consider forecasts that cover the entire test region.

Table 1: Times of occurrence, locations, and magnitudes of the 31 earthquakes in the test region with $M \geq 4.95$ from 1 January 2006 until 31 December 2010. The $M = 7.2$ El Mayor-Cucapah earthquake is in bold.

No.	Origin time (UTC)	Lat.	Long.	M
1	2006/05/24 04:20:26.01	32.3067	−115.2278	5.37
2	2006/07/19 11:41:43.46	40.2807	−124.4332	5.00
3	2007/02/26 12:19:54.48	40.6428	−124.8662	5.40
4	2007/05/09 07:50:03.83	40.3745	−125.0162	5.20
5	2007/06/25 02:32:24.62	41.1155	−124.8245	5.00
6	2007/10/31 03:04:54.81	37.4337	−121.7743	5.45
7	2008/02/09 07:12:04.55	32.3595	−115.2773	5.10
8	2008/02/11 18:29:30.53	32.3272	−115.2568	5.10
9	2008/02/12 04:32:39.24	32.4475	−115.3175	4.97
10	2008/02/19 22:41:29.66	32.4325	−115.3130	5.01
11	2008/04/26 06:40:10.60	39.5253	−119.9289	5.00
12	2008/04/30 03:03:06.90	40.8358	−123.4968	5.40
13	2008/07/29 18:42:15.71	33.9530	−117.7613	5.39
14	2008/11/20 19:23:00.19	32.3288	−115.3318	4.98
15	2008/12/06 04:18:42.85	34.8133	−116.4188	5.06
16	2009/09/19 22:55:17.84	32.3707	−115.2612	5.08
17	2009/10/01 10:01:24.67	36.3878	−117.8587	5.00
18	2009/10/03 01:16:00.31	36.3910	−117.8608	5.19
19	2009/12/30 18:48:57.33	32.4640	−115.1892	5.80
20	2010/01/10 00:27:39.32	40.6520	−124.6925	6.50
21	2010/02/04 20:20:21.97	40.4123	−124.9613	5.88
22	**2010/04/04 22:40:42.15**	**32.2587**	**−115.2872**	**7.20**
23	2010/04/04 22:50:17.08	32.0972	−115.0467	5.51
24	2010/04/04 23:15:14.24	32.3000	−115.2595	5.43
25	2010/04/04 23:25:06.95	32.2462	−115.2978	5.38
26	2010/04/05 00:07:09.07	32.0180	−115.0172	5.32
27	2010/04/05 03:15:24.46	32.6282	−115.8062	4.97
28	2010/04/08 16:44:25.92	32.2198	−115.2760	5.29
29	2010/06/15 04:26:58.48	32.7002	−115.9213	5.72
30	2010/07/07 23:53:33.53	33.4205	−116.4887	5.43
31	2010/09/14 10:52:18.00	32.0485	−115.1982	4.96

Seven forecasts were submitted that gave the predicted number, N_{fmi}, for $M \geq 4.95$ earthquakes in 0.1 magnitude bins during the five-year test period for all $N_c = 7682$ $0.1^\circ \times 0.1^\circ$ cells.

The submitted forecasts are based on a variety of approaches. The Bird and Liu forecast [14] was based on a kinematic model of neotectonics. The Ebel et al. forecast [15] was based on the average rate of $M \geq 5$ earthquakes in $3^\circ \times 3^\circ$ cells for the period 1932 to 2004. The Helmstetter et al. forecast [16] was based on the extrapolation of past seismicity. The Holliday et al. forecast [17] was based on the extrapolation of past seismicity using a modification of the pattern informatics (PI) technique. The Wiemer and Schorlemmer forecast [18] was based on the asperity-based likelihood model (ALM).

We will now discuss the Holliday et al. forecast in somewhat greater detail. The basis of this RELM forecast followed the format introduced in the PI forecast methodology [7, 8]. The magnitude range $M \geq 5$ and the cell dimensions $0.1^\circ \times 0.1^\circ$ were the same. However, the PI method was alarm based. Earthquakes were forecast to either occur or not occur in specified regions (hotspots) in a specified time period. In the PI-based RELM forecast, all hotspot cells are given equal probabilities of an earthquake. For the values in Table 2, $\lambda_{fi} = 3.32 \times 10^{-2}$. Instead of being alarm based, the RELM test was based on probabilities of occurrence of an earthquake in each cell in the test region. This required a continuous assessment of risk rather than a binary, alarm-based assessment. To do this, the Holliday et al. [17] forecast introduced a uniform probability of occurrence for hotspot regions and added smaller probabilities for nonhotspot regions based on the relative intensity (RI) of seismicity in the region. A map of the Holliday et al. [17] forecast is given in Figure 2.

As stated in our description of the RELM test, each participant submitted a forecast for the number of earthquakes N_{fmi} in magnitude bin m that would occur in cell i. Thus $41 \times 7682 = 314962$ values of N_{fmi} were submitted in each

Table 2: Cell scores λ_{fi} of an earthquake with $M \geq 4.95$ for the 22 cells in which earthquakes occurred during the test period. The association of cell IDs (A–V) with the earthquake IDs (1–31) from Table 1 is given. Five submitted forecasts are given: (1) Bird and Liu (B and L), (2) Ebel et al. (Ebel), (3) Helmstetter et al. (Helm.), (4) Holliday et al. (Holl.), and (5) Wiemer and Schorlemmer (W and S). The highest (best) scores are in bold.

Cell ID	EQ ID	B and L	Ebel	Helm.	Holl.	W and S
(A)	1,7,8,16,24	$1.99e-2$	$2.20e-2$	$1.17e-1$	$3.32e-2$	**$1.24e-1$**
(B)	2	$1.41e-2$	$3.40e-2$	**$7.20e-2$**	$3.32e-2$	$4.99e-2$
(C)	3	$7.40e-3$	$6.59e-3$	$7.41e-3$	**$3.32e-2$**	$7.91e-3$
(D)	4	$3.54e-2$	$3.29e-2$	**$6.97e-2$**	$3.32e-2$	$3.59e-2$
(E)	5	**$7.23e-3$**	$1.10e-3$	$2.29e-3$	$9.72e-5$	$1.58e-7$
(F)	6	$9.37e-3$	$2.85e-2$	$3.07e-2$	$3.32e-2$	**$4.55e-2$**
(G)	9,10	$9.11e-3$	$5.49e-3$	$2.55e-2$	**$3.32e-2$**	$2.38e-2$
(H)	11	$3.42e-4$	**$5.49e-3$**	$9.15e-4$	$1.62e-4$	$2.06e-4$
(I)	12	$2.14e-3$	$1.10e-3$	$3.65e-3$	$2.05e-4$	**$9.89e-3$**
(J)	13	$1.68e-3$	$8.78e-3$	$1.11e-2$	**$3.32e-2$**	$1.13e-2$
(K)	14	$3.12e-2$	$2.20e-2$	$3.30e-2$	$3.32e-2$	**$5.90e-2$**
(L)	15	$2.07e-3$	$5.49e-3$	**$6.93e-3$**	$3.32e-3$	$2.64e-3$
(M)	17,18	$1.74e-3$	$2.20e-3$	$5.78e-3$	**$3.32e-2$**	$5.38e-4$
(N)	19	**$5.83e-2$**	$6.59e-3$	$1.49e-2$	$3.32e-2$	$7.44e-3$
(O)	20	$1.25e-2$	$1.43e-2$	$9.45e-3$	**$3.32e-2$**	$1.62e-2$
(P)	21	$6.48e-3$	$3.29e-2$	$2.71e-2$	**$3.32e-2$**	$7.46e-3$
(Q)	22,25,28	$2.88e-2$	$2.20e-2$	$2.84e-2$	$3.32e-2$	**$5.23e-2$**
(R)	23,26	**$3.06e-2$**	$1.54e-2$	$1.43e-2$	$1.73e-4$	$1.58e-2$
(S)	27	$2.13e-2$	$5.49e-3$	$1.26e-2$	**$3.32e-2$**	$1.19e-2$
(T)	29	$1.83e-2$	$1.32e-2$	$2.43e-2$	$3.32e-2$	**$4.99e-2$**
(U)	30	$1.26e-2$	$3.07e-2$	**$1.03e-1$**	$3.32e-3$	$5.16e-2$
(V)	31	$6.76e-3$	$1.54e-2$	$5.55e-3$	**$3.32e-2$**	$2.64e-3$

forecast. In order to better understand the implications of the forecasts we sum the probabilities in the magnitude bins for each spatial cell to give the number of forecast earthquakes N_{fi} in cell i with magnitude $M \geq 4.95$:

$$N_{fi} = \sum_{m=5}^{9} N_{fmi}. \tag{2}$$

The reason we carry out this sum is so that we can directly apply the "skill score" methodology developed in the atmospheric sciences. In terms of forecasting tornadoes, the question is whether a tornado occurs, not its strength. Since the RELM test was for earthquakes with $M \geq 4.95$ our scoring is whether such an earthquake occurs or does not occur in a spacial cell.

The sum of the N_{fi} over all cells is the total number of earthquakes N_f with $M \geq 4.95$ forecast to occur during the test period:

$$N_f = \sum_{i=1}^{N_c} N_{fi}, \tag{3}$$

where N_c is the total number of cells. Our objective is to separate the forecast of the total number of earthquakes from the forecast of their locations. In order to do this we introduce a cell score λ_{fi} defined by

$$\lambda_{fi} = \frac{N_{ce} N_{fi}}{N_f}, \tag{4}$$

where N_{ce} is the number of cells in which an earthquake occurred during the test period. Note that from (3) and (4) we have

$$\sum_{i=1}^{N_c} \lambda_{fi} = N_{ce}. \tag{5}$$

Thus, the sum of λ_{fi} over all cells is the same for each submitted forecast. The cell score λ_{fi} is a direct measure of the probability of occurrence of a test earthquake in cell i. A perfect forecast (a perfect skill score) would have $\lambda_{fi} \geq 1$ for the cells in which earthquakes occur and $\lambda_{fi} = 0$ for all other cells. In principal λ_{fi} can be as big as N_{ce}. However, because we are only concerned with whether an earthquake occurs in a cell, not how many occur—a point we discuss in the next paragraph—all values of $\lambda_{fi} > 1$ are just treated as 1 for that particular cell. In practice this does not occur due to the small values of N_{fmi} provided by the RELM forecasts.

Since the forecasts are for specific $0.1^\circ \times 0.1^\circ$ cells, it is necessary to consider how to handle the forecasts when more than one earthquake occurs in a cell. As stated above, in our analysis a cell in which more than one earthquake occurred is treated the same as a cell in which only one earthquake occurred. This follows the practice used in tornado forecasting. How many tornadoes occur in a region during the forecast period is not considered, only whether one or more occur. For the test earthquakes given in Table 1, events 1, 7, 8, 16, and 24 occurred in the same cell, similarly

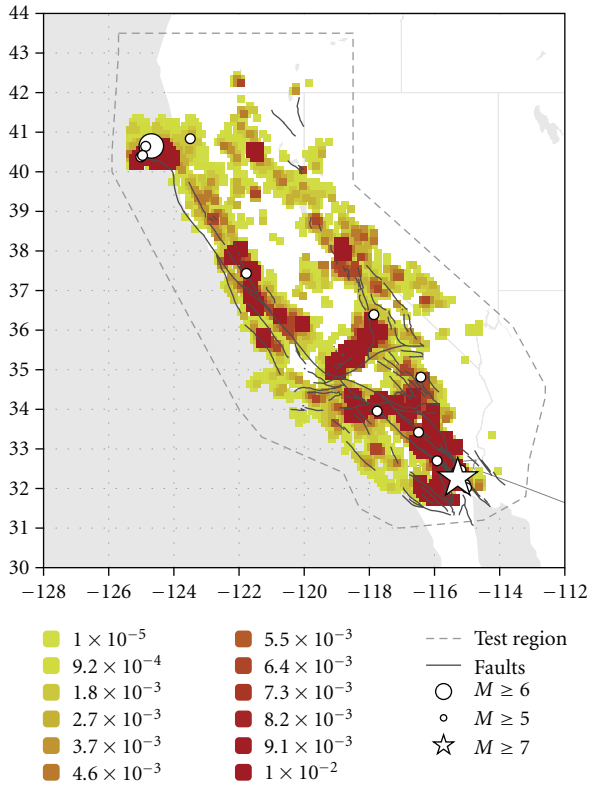

FIGURE 2: Map of the normalized probabilities λ_{fi} given for the testregion by Holliday et al. [17] using their PI-based forecast. The "hotspots" are shown in red. The test earthquakes are also shown.

TABLE 3: Comparisons of the forecasts: Column 1. the number of maximum cell scores $N_{\lambda\max}$. Column 2: the mean cell scores forecast $\bar{\lambda}_f$. Column 3: the number of earthquakes N_f predicted by each forecast. The best scores in each category are in bold.

	$N_{\lambda\max}$	$\bar{\lambda}_f$	N_f
Bird and Liu	3	$1.53e-2$	56
Ebel et al.	1	$1.51e-2$	115
Helmstetter et al.	4	**$2.84e-2$**	35
Holliday et al.	**8**	$2.45e-2$	**30**
Wiemer and Schor.	6	$2.66e-2$	24

for events 9 and 10, events 17 and 18, events 22, 25, and 28, and events 23 and 26. This multiplicity is shown in Table 2. Thus, we will consider forecasts made for 22 cells.

Taking the actual number of cells in which earthquakes occurred to be $N_{ce} = 22$ and the total number of earthquakes forecast in each submission N_f using (3), we obtained the forecast scores λ_{fi} using (4).

The seven submitted forecasts included two submissions with separate forecasts with and without aftershocks. Different numbers of events were forecast but the relative scores of locations were the same. Thus, we consider five submissions. The forecast scores λ_{fi} for each of the five submissions are given in Table 2 for the $N_{ce} = 22$ cells in which an earthquake occurred. A perfect forecast in which only the 22 cells were forecast to have earthquakes would have $\lambda_{fi} = 1$ in each of the 22 cells. A random forecast in which all $N_c = 7682$ cells were given the same $N_{fi} = a$ would yield

$$\lambda_{fi}^{\text{random}} = \frac{N_{ce}a}{N_f} = \frac{N_{ce}}{N_c} = \frac{22}{7682} = 2.86 \times 10^{-3}. \quad (6)$$

The submitted forecast scores in Table 2 have a wide range of values from $\lambda_{fi} = 1.58 \times 10^{-7}$ to $\lambda_{fi} = 1.24 \times 10^{-1}$.

7. Evaluation of Results

During the formulation of the RELM project a comprehensive testing strategy was also developed [19]. A suite of likelihood tests were proposed, which would be implemented through a testing center [20]. The approach utilized an L-test, an N-test, and an R-test. These tests were applied to the raw submitted data. This approach was applied to the first 2.5 years of RELM results by Schorlemmer et al. [13]. Zechar et al. [21] recognized a problem with the original proposed likelihood tests and proposed a modification.

This is certainly one approach to the evaluation of results, the primary purpose of this paper is to present a complementary approach. Our approach has the advantage that the evaluation of the numbers of earthquakes forecast can be separated from the forecast of their locations.

Lee et al. [22] proposed the modified approach to the evaluation of RELM test results that is used in this paper. In their short paper they compared the forecasts that had been submitted for all of California. In this paper we consider a subset of those forecasts and relate the results to the concept of alarm-based forecasts.

The results given in Table 2 can be used to compare the forecast scores for each of the cells in which earthquakes occurred. The highest scores between the models are shown in bold. Clearly there are many ways in which to evaluate the results of the forecasts. There is a tradeoff between good forecasts with large λ_{fi} and poor forecasts with small λ_{fi}. We first consider the forecasts that had the highest forecast scores. The Holliday et al. [17] forecast had the largest λ_{fi} for 8 of the 22 cells in which (target) earthquakes occurred. The Wiemer and Schorlemmer [18] forecast had 6 of the largest λ_{fi}. The Helmstetter et al. [16] forecast had 4 of the largest λ_{fi}. Finally, the Bird and Liu [14] forecast had 3 of the largest λ_{fi}. These values are also given in Table 3. The range of the highest cell scores was from $\lambda_{fi} = 1.24 \times 10^{-1}$ for event 1 to $\lambda_{fi} = 5.49 \times 10^{-3}$ for event 11.

It is also of interest to compare the mean cell forecast scores for the 22 cells in which earthquakes occurred. These values $\bar{\lambda}_f$ are given in Table 3. The Helmstetter et al. [16] forecast had the highest $\bar{\lambda}_f = 2.84 \times 10^{-2}$, the Wiemer and Schorlemmer [18] forecast had $\bar{\lambda}_f = 2.66 \times 10^{-2}$, and the Holliday et al. [17] forecast had $\bar{\lambda}_f = 2.45 \times 10^{-2}$. The Helmstetter et al. [16] forecast did the best in an average sense but did relatively poorly in providing the best cell forecasts. It should be noted that the best average forecast $\bar{\lambda}_f = 2.84 \times 10^{-2}$ is one order of magnitude better than the random (no skill) forecast $\lambda_{fi}^{\text{random}} = 2.86 \times 10^{-3}$.

As noted above, the Holliday et al. [17] forecast is primarily an alarm-based (hotspot) forecast. The PI method was used to determine the cells in which earthquakes were most likely to occur (hotspots). In the cell forecasts given in Table 2, these cells had forecast scores $\lambda_{fi} = 3.32 \times 10^{-2}$ and consisted of 8.3% of the total area of the test region (637 of the 7682 cells). Of the 22 cells in which earthquakes occurred, 17 occurred in hotspot cells. In 8 of the 17 cells, the forecast cell scores given by the Holliday et al. [17] forecast were the highest.

8. Discussion

The RELM test provides a well-defined set of prospective earthquake forecasts and a well-defined set of test earthquakes. In this paper we present a method for evaluating the RELM forecasts. We believe our approach has significant advantages but look forward to comparing our results with those obtained by other authors.

RELM forecasts provide the numbers N_{fmi} of earthquakes expected to occur in magnitude bins m and spatial cells i. The basis of our approach is

(1) to use (2) to determine the forecast number N_{fi} of earthquakes with $M \geq 4.95$ expected to occur in spatial cell i,

(2) to use (3) to determine the total forecast number N_f of earthquakes,

(3) to use (4) to determine the cell score λ_{fi}.

We first compared the actual number of earthquakes that occurred during the test period, 31 with the forecast values. The closest forecast values were those of Holliday et al. [17] with $N_f = 30$ as shown in Table 3.

We next compared the forecast scores λ_{fi} of an earthquake with $M \geq 4.95$ occurring in cell i. We noted that the values of λ_{fi} were the same for the two submissions in which both main shocks and aftershocks plus main shocks were submitted. These forecasts gave different values for the numbers N_{fmi}, N_{fi}, and N_f of earthquakes but the forecast distributions in space were identical.

In a perfect forecast the forecast score would have been $\lambda_{fi} = 1$ for each of the 22 cells in which one or more earthquakes occurred and $\lambda_{fi} = 0$ in the other 7660 cells. The mean forecast scores for the 22 cells in which earthquakes occurred for the five forecasts ranged from a high value $\overline{\lambda}_f = 2.84 \times 10^{-2}$ to a low value of $\overline{\lambda}_f = 1.53 \times 10^{-2}$. The range of values was relatively small, about a factor of two. The random (no skill) forecast assuming equal probabilities for the 7682 cells in the test region gives a forecast score $\lambda_{fi}^{\text{random}} = 2.86 \times 10^{-3}$ for all cells. The best forecast score $\lambda_{fi} = 2.84 \times 10^{-2}$ was about a factor of 10 better than the random forecast but a factor of 100 worse than a perfect forecast.

As we have previously discussed earthquake forecasts can be either probabilistic or alarm based. The submission rules for RELM were probabilistic. The only forecast that had an alarm-based distribution of forecasts was that of Holliday et al. [17]. A question of interest for future tests of earthquake forecasts is whether they should be alarm or probability based. A systematic study of alarm-based forecasts could be of considerable interest.

Another interesting question is whether the forecasts have a temporal component. Is there a time-dependent component in the data used that changes forecast probabilities significantly? As discussed previously, eight of the test earthquakes were aftershocks of the El Mayor-Cucapah earthquake and eight of the test earthquakes were associated with a precursory swarm. Thus, 17 of the 31 of the test earthquakes were associated with this earthquake. It appears reasonable to conclude that precursory activation prior to the El Mayor-Cucapah earthquake may have played a significant role in the success of forecasts.

Another swarm of 6 earthquakes during the test period adjacent to Cape Mendocino did not lead to a subsequent larger event during the test period. Swarms of activity in this region occur regularly. In terms of precursory activation this activity would lead to a false alarm. The contrast between the two regions (Cape Mendocino and El Mayor-Cucapah) is an indication of the difficulties in forecasting earthquakes utilizing precursory activation.

Glossary

M: Earthquake magnitude
m: Bin magnitude $m - 0.05 \leq M \leq m + 0.05$
N_e: Number of actual earthquakes
N_f: Number of forecast earthquakes
N_c: Number of cells
N_{ce}: Number of cells with earthquakes
N_{fi}: Number of forecast earthquakes in cell i
N_{fmi}: Number of forecast earthquakes in magnitude bin m and cell i
λ_{fi}: Forecast score, related to the probability that an earthquake with $M \geq 4.95$ will occur in cell i
$\overline{\lambda}_f$: Mean forecast score for the 22 cells in which earthquakes occurred
$\lambda_{fi}^{\text{random}}$: A random (no skill) forecast $\lambda_{fi}^{\text{random}} = 2.86 \times 10^{-3}$
$N_{\lambda \max}$: The number of maximum cell scores.

Acknowledgments

Y. T. Lee is grateful for research support from both the National Science Council (ROC) and the Institute of Geophysics (NCU, ROC). J. B. Rundle and J. R. Holliday have been supported by Nasa Grant NNXO8AF69G.

References

[1] V. G. Kossobokov, V. I. Keilis-Borok, D. L. Turcotte, and B. D. Malamud, "Implications of a statistical physics approach for earthquake hazard assessment and forecasting," *Pure and Applied Geophysics*, vol. 157, no. 11-12, pp. 2323–2349, 2000.

[2] J. B. Rundle, W. Klein, K. Tiampo, and S. Gross, "Linear pattern dynamics in nonlinear threshold systems," *Physical Review E*, vol. 61, no. 3, pp. 2418–2431, 2000.

[3] V. I. Keilis-Borok, "The lithosphere of the earth as a nonlinear system with implications for earthquake prediction," *Reviews of Geophysics*, vol. 28, pp. 19–34, 1990.

[4] V. Keilis-Borok and A. A. Soloviev, *Nonlinear Dynamicsof the Lithosphere and Earthquake Prediction*, Springer, New York, NY, USA, 2003.

[5] I. T. Jolliffe and D. B. Stephenson, *Forecast Verification*, John Wiley & Sons, Chichester, UK, 2003.

[6] J. B. Rundle, D. L. Turcotte, R. Shcherbakov, W. Klein, and C. Sammis, "Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems," *Reviews of Geophysics*, vol. 41, no. 4, p. 1019, 2003.

[7] K. F. Tiampo, J. B. Rundle, S. McGinnis, S. J. Gross, and W. Klein, "Eigenpatterns in southern California seismicity," *Journal of Geophysical Research*, vol. 107, no. B12, p. 2354, 2002.

[8] J. B. Rundle, K. F. Tiampo, W. Klein, and J. S. S. Martins, "Self-organization in leaky threshold systems: the influence of near-mean field dynamics and its implications for earthquakes, neurobiology, and forecasting," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, 1, pp. 2514–2521, 2002.

[9] J. R. Holliday, J. B. Rundle, K. F. Tiampo, and D. L. Turcotte, "Using earthquake intensities to forecast earthquake occurrence times," *Nonlinear Processes in Geophysics*, vol. 13, no. 5, pp. 585–593, 2006.

[10] J. R. Holliday, K. Z. Nanjo, K. F. Tiampo, J. B. Rundle, and D. L. Turcotte, "Earthquake forecasting and its verification," *Nonlinear Processes in Geophysics*, vol. 12, no. 6, pp. 965–977, 2005.

[11] J. D. Zechar and T. H. Jordan, "Testing alarm-based earthquake predictions," *Geophysical Journal International*, vol. 172, no. 2, pp. 715–724, 2008.

[12] E. H. Field, "Overview of the working group for the development of regional earthquake likelihood models (RELM)," *Seismological Research Letters*, vol. 78, no. 1, pp. 7–16, 2007.

[13] D. Schorlemmer, J. D. Zechar, M. J. Werner, E. H. Field, D. D. Jackson, and T. H. Jordan, "First results of the regional earthquake likelihood models experiment," *Pure and Applied Geophysics*, vol. 167, no. 8-9, pp. 859–876, 2010.

[14] P. Bird and Z. Liu, "Seismic hazard inferred from tectonics: California," *Seismological Research Letters*, vol. 78, no. 1, pp. 37–48, 2007.

[15] J. E. Ebel, D. W. Chambers, A. L. Kafla, and J. A. Baglivo, "Non-poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California," *Seismological Research Letters*, vol. 78, no. 1, pp. 57–65, 2007.

[16] A. Helmstetter, Y. Y. Kagan, and D. D. Jackson, "High-resolution time-independent grid-based forecast for $m \geq 5$ earthquakes in California," *Seismological Research Letters*, vol. 78, no. 1, pp. 78–86, 2007.

[17] J. R. Holliday, C. C. Chen, K. F. Tiampo, J. B. Rundle, D. L. Turcotte, and A. Donnellan, "A RELM earthquake forecast based on pattern informatics," *Seismological Research Letters*, vol. 78, no. 1, pp. 87–93, 2007.

[18] S. Wiemer and D. Schorlemmer, "ALM: an asperity-based likelihood model for California," *Seismological Research Letters*, vol. 78, no. 1, pp. 134–140, 2007.

[19] D. Schorlemmer, M. C. Gerstenberger, S. Wiemer, D. D. Jackson, and D. A. Rhoades, "Earthquake likelihood model testing," *Seismological Research Letters*, vol. 78, no. 1, pp. 17–29, 2007.

[20] D. Schorlemmer and M. C. Gerstenberger, "RELM testing center," *Seismological Research Letters*, vol. 78, no. 1, pp. 30–36, 2007.

[21] J. D. Zechar, M. C. Gerstenberger, and D. A. Rhoades, "Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts," *Bulletin of the Seismological Society of America*, vol. 100, no. 3, pp. 1184–1195, 2010.

[22] Y.-T. Lee, D. L. Turcotte, J. R. Holliday et al., "Results of the Regional Earthquake Likelihood Models (RELM) test of earthquake forecasts in California," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 40, pp. 16533–16538, 2011.