Geophysical Journal International

Geophys. J. Int. (2014) **199,** 60–68 doi: 10.1093/gji/ggu137
GJI Seismology

A parimutuel gambling perspective to compare probabilistic seismicity forecasts

J. Douglas Zechar[1,2] and Jiancang Zhuang[3]

[1]*Swiss Seismological Service, ETH Zurich, Sonneggstrasse* 5, *NO H* 3, 8092 *Zurich, Switzerland. E-mail: jeremy.zechar@sed.ethz.ch*
[2]*Department of Earth Sciences, University of Southern California,* 3651 *Trousdale Pkwy, Los Angeles, CA* 90089, *USA*
[3]*Institute of Statistical Mathematics,* 10-3 *Midori-Cho, Tachikawa-Shi, Tokyo* 190-8562, *Japan*

Accepted 2014 April 7. Received 2014 April 4; in original form 2013 August 6

SUMMARY

Using analogies to gaming, we consider the problem of comparing multiple probabilistic seismicity forecasts. To measure relative model performance, we suggest a parimutuel gambling perspective which addresses shortcomings of other methods such as likelihood ratio, information gain and Molchan diagrams. We describe two variants of the parimutuel approach for a set of forecasts: head-to-head, in which forecasts are compared in pairs, and round table, in which all forecasts are compared simultaneously. For illustration, we compare the 5-yr forecasts of the Regional Earthquake Likelihood Models experiment for $M4.95+$ seismicity in California.

Key words: Probabilistic forecasting; Probability distributions; Earthquake interaction, forecasting and prediction; Seismicity and tectonics; Statistical seismology.

1 INTRODUCTION

In 1997, Geller *et al.* published their controversial landmark article claiming that 'Earthquakes cannot be predicted'. Seventeen years later, the statement remains true, but this is not to say that nothing has changed. One of the primary points that Geller *et al.* made was that earthquake predictions—that is, predictive statements about individual earthquakes—were not being expressed as unambiguously falsifiable statements. To a large extent, this is no longer true: many researchers have since followed the early suggestions made by Kagan & Knopoff (1987), Evison & Rhoades (1993) and Kagan & Jackson (1994) and begun constructing probabilistic forecasts related to the distribution of earthquakes, or seismicity forecasts. The resulting proliferation of forecast experiments has highlighted a second-order problem: how do you assess the performance of earthquake and seismicity forecasts in general? And, more specifically, how do you compare seismicity forecasts, whether they be derived from different models or from one model with different parameter values? These questions have scientific as well as practical implications: researchers can use seismicity forecasts to test hypotheses related to seismogenesis, earthquake clustering and earthquake triggering; and seismicity forecasts are also the basis for seismic hazard assessments that influence building codes, insurance rates and preparatory exercises.

Researchers have applied several methods to compare the performance of seismicity forecasts. To directly compare two probabilistic forecasts, you can calculate the likelihood ratio (Kagan & Jackson 1995) or the information gain per earthquake (Kagan & Knopoff 1977; Harte & Vere-Jones 2005; Rhoades *et al.* 2011). The likelihood ratio indicates which forecast is more likely to have generated the observed distribution of earthquakes and the 'null observations' where no earthquakes occurred. The information gain emphasizes the forecast probabilities where earthquakes occurred and the total number of earthquakes expected by each forecast. The Molchan diagram method (Molchan 1991, 1997, 2010, 2011; Molchan & Kagan 1992; Molchan & Keilis-Borok 2008) applies to a wider class of earthquake and seismicity forecasts, and it reduces each forecast to a set of binary earthquake predictions, or alarms (Kagan 2007; Zechar & Jordan 2008). You calculate the fraction of earthquakes that did not occur during alarms and the fraction of space occupied by alarms, where space is measured according to a reference model. Kossobokov (2004, 2006) noted that such alarms could be thought of as wagers in a game of what he called seismic roulette, where Nature controls the wheel. Zhuang (2010) expanded this notion with a comparison measure that we call the fixed-odds gambling score; using this method, an earthquake forecast is viewed as a series of wagers, and the forecast is pitted against a reference model that functions as the house, or the odds-maker. The fixed-odds gambling score method applies to models producing binary predictions (Zhuang & Jiang 2012) and probabilistic forecasts, as well as point process models that generate continuous (i.e. without gridding) forecasts.

Each of these methods for comparing earthquake forecasts has drawbacks. The likelihood ratio is sensitive to the occurrence of low-probability events, and a single earthquake can strongly affect a forecast comparison. For example, we recall the situation described by Holliday *et al.* (2005, p. 969): suppose that Forecast A had very large probabilities for 99 of 100 earthquakes and an extraordinarily small probability for the 100th earthquake, while Forecast B had intermediate probabilities for all 100 earthquakes.

Instinctually, we would say that Forecast A is better. However, because the joint likelihood is sensitive to small values, the likelihood ratio will favour Forecast B, a result that is intuitively unsatisfying. The information gain has the same problem. Moreover, the current tests used to establish statistical significance of the observed information gain require an assumption about the information gain distribution (Rhoades *et al.* 2011): to use the T-test, you assume that the information gains are normally distributed and for the W-test, you make the weaker assumption that the information gains are symmetrically distributed. Both of these assumptions may be violated in practice (Eberhard *et al.* 2012; Taroni *et al.* 2014). One stumbling block associated with the Molchan diagram and the fixed-odds gambling score is the choice or construction of a reference model: depending on the format of the candidate model, what reference model to use can be controversial (Stark 1997; Kossobokov *et al.* 1999; Marzocchi *et al.* 2003; Luen & Stark 2008; Marzocchi & Zechar 2011). Moreover, the parametrization of the reference model and the parameter value choices can have a strong effect on the resulting assessment (Molchan & Romashkova 2010), and therefore they must be carefully justified. There is no perfect reference model for a particular candidate model, let alone a panacea.

Another, more subtle, problem with the fixed-odds gambling score is that it is not symmetric. Imagine that you take some Forecast A as the candidate model and some Forecast B as the reference model, and you calculate the net return of Forecast A. If you then switch the roles of A and B, there is no guarantee that B's net return will be the same size with opposite sign as when A was the candidate model. When this asymmetry occurs and the candidate model's net return is positive in either case, it indicates that the model performances are, in a sense, nested—A has some virtues that B lacks, and vice versa. Molchan & Romashkova (2011) discussed the potential instability of the fixed-odds gambling score: successfully predicting an event that the reference model says is very unlikely has a large impact on the gambling score and might make a candidate model appear better than it really is. If we follow the gambling analogy, the candidate model can win big on a single bet.

In this paper, we propose an alternative method for comparing earthquake and seismicity forecasts: the parimutuel gambling score. It, too, has many analogues to gambling and therefore the score and the interpretation of results are intuitive. The main difference between the parimutuel gambling score and the fixed-odds gambling score is the lack of a specific reference model. The parimutuel gambling score addresses the drawbacks discussed above and applies to a wide variety of earthquake forecasts.

In the following section, we introduce mathematical notation, review the relevant features of the fixed-odds gambling score and describe the parimutuel gambling score. In Section 3, we describe the 5-yr Regional Earthquake Likelihood Models (RELM) earthquake forecast experiment in California, and in Section 4 we report the results of analysing the RELM experiment using the parimutuel gambling score method. In Section 5, we discuss the relationship between likelihood ratio and parimutuel gambling and describe how the parimutuel gambling score can be used for model optimization. In Section 6, we summarize our findings and mention other possible applications of our method.

2 METHODS

Zhuang (2010) introduced the fixed-odds gambling score and described its application to three types of candidate forecasts: alarm-based, probabilistic, and continuous (i.e. generated by point process models). Forecasts from each of these classes can be thought of as a series of wagers, and once a reference model is chosen or constructed, a gambling return can be calculated for each wager. Let p be the probability for an event to happen according to the candidate forecast and p_0 be the corresponding reference model probability. This event could be the occurrence of a single earthquake within a well-defined space–time–magnitude volume, or it could be the occurrence of zero earthquakes, or more than one earthquake, in a volume. You can think of the candidate forecast as two complementary bets: a wager of p on the event happening and $(1 - p)$ on the event not happening. If the event occurs, the forecast will lose $(1 - p)$ and win an amount that is proportional to p and p_0. The gambling return if the event occurs is

$$\Delta R = -(1 - p) + p\frac{1 - p_0}{p_0}. \tag{1}$$

If the event does not occur, the return is

$$\Delta R = (1 - p)\frac{p_0}{1 - p_0} - p. \tag{2}$$

Zhuang (2010) showed these returns guarantee that, if the reference model is correct, the expected return for any betting strategy is zero. This method applies when you want to evaluate the skill of a candidate model relative to a reference model, where the choice of the reference model is well-justified. However, the game is not symmetric: exchanging the roles of the candidate model and the reference model does not yield an equal and opposite return. For example, if the candidate model probability $p = 0.2$, the reference model probability $p_0 = 0.5$ and the event happens, the return is -0.6. (The candidate model assigned a lower probability of the event happening and therefore a loss is sensible). If the roles of the models are reversed, the gambling return is 1.5. It can also happen that, given a series of wagers, the sign of the net return is the same regardless of each model's role as candidate or reference. This paradox is particularly counter-intuitive because it seems to indicate that each model is better (or worse) than the other. One explanation is that the fixed-odds gambling score is a partial score: it does not punish random guesses or 'no-comment' predictions. If you know a shortcoming of the reference model, you can win with a corresponding betting strategy. For example, if you somehow determine that the reference model consistently underestimates the earthquake probability under certain conditions, you can win by betting with certainty ($p = 1$) when these conditions are met and otherwise not bet at all. However, detecting such biases is not trivial, and neither is constructing or choosing an appropriate reference model. To address the role reversal paradox, and rather than treating one model as the reference and allowing it to set the odds, you can take the view of an alternative betting paradigm: parimutuel gambling.

Instead of thinking about a candidate model and a reference model, consider two candidate forecasts, Λ_1 and Λ_2, with respective probabilities for an event to happen p_1 and p_2, and the size of the 'pot' is two. (Because the bets are probabilities, each forecast bets a total of one unit.) If the event happens, the forecasts, or bettors, will split the pot in a way that reflects their relative skill. In particular, the return for each bettor is the ratio of the amount that the bettor wagered on the outcome to the total amount wagered on the outcome, multiplied by the size of the pot. The net return

Table 1. Excerpt of an example forecast specified by a range of longitudes, latitudes and magnitudes, the expected number of earthquakes in this bin and a masking bit (1 indicates that it should be included in analyses).

Min. lon.	Max. lon.	Min. lat.	Max. lat.	Min. mag.	Max. mag.	Forecast	Mask
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−119.2	−119.1	33.7	33.8	5.35	5.45	0.37	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
−118.9	−118.8	38.3	38.4	6.15	6.25	0.00	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

is simply the total initial wager (here, one) subtracted from the return:

$$\Delta R_1 = 2\frac{p_1}{p_1 + p_2} - 1 \qquad \Delta R_2 = 2\frac{p_2}{p_1 + p_2} - 1. \tag{3}$$

If the event does not happen, p_1 and p_2 in (3) are simply replaced by $1 - p_1$ and $1 - p_2$, respectively. This equation describes a zero-sum game: in gambling terms, it is parimutuel gambling without a vigorish (the fee that the house charges to play). Over a series of wagers, parimutuel gambling leads to a redistribution of wealth, where the redistribution is driven by skill. And the parimutuel gambling approach can be simultaneously applied to multiple forecasts: we generalize to the situation of k forecasts with respective probabilities $p_1, p_2, \ldots, p_k$. The net return for the jth forecast is

$$\Delta R_j = -1 + kp_j \frac{1}{\sum_{m=1}^{k} p_m}. \tag{4}$$

Rather than thinking of p_j as the probability for an event to happen according to the jth forecast, it is useful to instead interpret it as the probability of the observation conditional on the jth forecast. In other words, if Λ_1 stated that an event would occur with probability 0.3 and the event did not occur, $p_1 = 0.7$. Under this interpretation, (4) applies regardless of the outcome.

The relationship between the parimutuel gambling score and the fixed-odds gambling score is straightforward: the (implicit) reference model for parimutuel gambling is the average model. If we let $p_0 = (p_1 + p_2 + \ldots + p_k)/k$, (1) and (4) are identical. In other words, if there is no commonly accepted reference model, we use the average model as the reference instead. When there are many forecasters, the average model becomes a reference model based on common sense. In this way, the parimutuel gambling approach also solves the stability problem that plagues the fixed-odds gambling score. With fixed odds, if a reference model estimates a vanishingly small probability of a particular outcome, the candidate model stands to gain a large amount on a single wager; indeed this amount could dominate the total return over all wagers. This was the situation we (Zechar & Zhuang 2010) encountered when evaluating the Reverse Tracing of Precursors (RTP) algorithm (Keilis-Borok & Shebalin 2004): if some contentious predictions with small reference model probabilities were ignored, the RTP net return was greatly diminished and the overall RTP performance was no longer statistically significant. While a candidate forecast gambling against a reference model can gain an unlimited amount of wealth with a single wager, it stands to lose at most one unit for each wager. For a single wager in parimutuel gambling, each forecast can lose at most one unit, but the maximum return is finite and governed by the number of candidate forecasts (i.e. the maximum return is $k - 1$).

Suppose that, under the true model, the probability for the forecasted event to happen is p^*, and let $p_0 = \frac{1}{k}\sum_{i=1}^{k} p_i$ be the reference probability. Then, the expected return for Forecast 1 is

$$\begin{aligned}
&\mathbf{E}[\Delta R_1] \\
&= p^* \left(\frac{k\, p_1}{\sum_{i=1}^{k} p_i} - 1 \right) + (1 - p^*) \left[\frac{k\,(1 - p_1)}{\sum_{i=1}^{k} (1 - p_i)} - 1 \right] \\
&= \frac{p^*\, p_1}{p_0} + \frac{(1 - p^*)(1 - p_1)}{1 - p_0} - 1 \\
&= \frac{(p_1 - p_0)(p^* - p_0)}{p_0(1 - p_0)}.
\end{aligned} \tag{5}$$

Then, the expected value of ΔR_1 is positive only when p_1 is positively correlated to p^*/p_0, that is, if $p^* > p_0$, it must be that $p_1 > p_0$. In summary, the parimutuel gambling score prefers the model that is most closely correlated to the true model, relative to the average model.

Currently, the most common format for seismicity forecasts is the space–rate–magnitude format that was designed for the RELM experiment (Schorlemmer *et al.* 2007): the forecaster specifies the expected number of earthquakes to occur within bins that are defined by a range of latitude, longitude, magnitude and time. Table 1 shows an example hypothetical forecast. Certainly, these are not deterministic forecasts: the number of earthquakes specified by each forecast in each bin is an expected value and represents a Poisson distribution. In other words, each bin contains a Poisson probability distribution for the number of events. We note that the forecast need not be Poisson in every bin; it is only required that a probability mass function is given for each bin so the probability of any observation can be calculated. Specifying a probability mass function is analogous to spreading your chips across all the possible outcomes in a game of roulette, where the sum of your chips is unity. After the wheel spins, you lose all the chips not placed on the winning number and win your fair share of the pot—this is exactly what (4) describes.

At the end of an experiment, we use the probability distribution in each bin to calculate the probability of the observation in each bin. For example, for the forecast in Table 1, imagine that zero earthquakes occurred in the first bin that is shown. The probability for this bin is $f(0|0.37) = 0.69$, where f is the Poisson probability mass function. Each bin represents a separate wager, so the total amount won/lost by the jth candidate forecast is obtained by summing over all n bins:

$$\Delta r_j^i = -1 + kp_i^j \sum_{m=1}^{k} \frac{1}{p_m^i} \qquad \Delta R_j = \sum_{i=1}^{n} \Delta r_j^i \tag{6}$$

We use (6) to quantify the skill of multiple probabilistic space–rate–magnitude forecasts. In the next section, we describe two such sets of forecasts from the 5-yr RELM experiment in California.

3 DATA

The RELM working group designed a 5-yr experiment to forecast M4.95+ earthquakes in and around California. In preparing for the experiment, the working group: developed several probabilistic seismicity forecasts; precisely delineated the space–time–magnitude region of interest and the earthquake catalogue to use for observations; and proposed several tests to assess forecast performance. These efforts were documented in the 2007 January/February special issue of *Seismological Research Letters* and were summarized by Schorlemmer *et al.* (2010b). RELM forecasts were constructed using the format described in the previous section: they specified a probability mass function for the number of earthquakes to occur during the 5-yr period from 2006 January 1 to 2010 December 31 (inclusive) in latitude–longitude–magnitude bins ($0.1^\circ \times 0.1^\circ \times 0.1$ units). The RELM working group created two forecast classes: one would forecast all seismicity, and the other would forecast only mainshock seismicity, where mainshocks would be identified *ex post facto* by a pre-determined algorithm (Reasenberg 1985). The working group developed 11 mainshock forecasts and 6 mainshock+aftershock forecasts. One other feature of RELM forecasts is that bins could be masked—any forecast in a masked bin should be ignored. From the gambling perspective, masking is equivalent to sitting out a round or abstaining from gambling. The forecasts were developed independently and without knowledge of each other: unlike some games, no bettor could adjust wagers based on the wagers of other bettors.

The scientific bases of the RELM forecasts have been widely discussed elsewhere and are not the emphasis of this paper, so we refer you to the RELM special issue (Bird & Liu 2007; Ebel *et al.* 2007; Helmstetter *et al.* 2007; Holliday *et al.* 2007; Kagan *et al.* 2007; Shen *et al.* 2007; Ward 2007; Wiemer & Schorlemmer 2007) and the summary articles by Schorlemmer *et al.* (2010b) and Zechar *et al.* (2013). For map-view representations of each forecast and the testing region, see Zechar *et al.*'s (2013) figs 1 and 2. We note that the majority of the forecasts used masking extensively: some forecasts were thereby limited to southern California, and others have irregular holes.

During the RELM experiment, 31 target earthquakes occurred, and 20 of them were deemed to be mainshocks; details are shown in Fig. 1 and Table 2. We provide the forecasts and the catalogues of observed earthquakes in the Supporting Information. Of particular interest for this study is the large number of bins used in the experiment: 314 962 (7682 spatial cells each having 41 magnitude bins).

4 RESULTS

In Fig. 2, we present the results of a parimutuel gambling analysis of the RELM experiment. The forecasts are ordered by overall net return [ΔR_j from (6); filled circles in the figure], and we also plot the returns from only the bins in which earthquakes occurred (i.e. disregarding those bins where no earthquakes occurred; hollow diamonds in the figure). In the mainshock class, the forecast developed by Helmstetter *et al.* is the best, which matches the conclusions of Zechar *et al.* (2013) based on pairwise information gain analyses. We note that the Helmstetter *et al.* mainshock forecast is also best when considering only the bins where target earthquakes occurred, but to look at only these bins could be misleading. For example, the mainshock+aftershock forecast by Ebel *et al.* had the highest net return when only looking at bins where earthquakes occurred; because no bin had an expectation greater than unity, the high gambling return suggests that the Ebel *et al.* mainshock+aftershock forecast had higher expectations than the other forecasts. However, it also has high expectations in bins where earthquakes did not occur, resulting in a negative net return when considering all bins. Zechar *et al.* (2013) also noted that such high rates meant that the Ebel *et al.* mainshock+aftershock forecast significantly overpredicted the number of target earthquakes. In the mainshock+aftershock class, the Holliday *et al.* forecast obtained the largest net return and Helmstetter *et al.* the second largest. In a direct pairwise information gain analysis, Zechar *et al.* (2013) found the reverse, but concluded that the difference was not statistically significant.

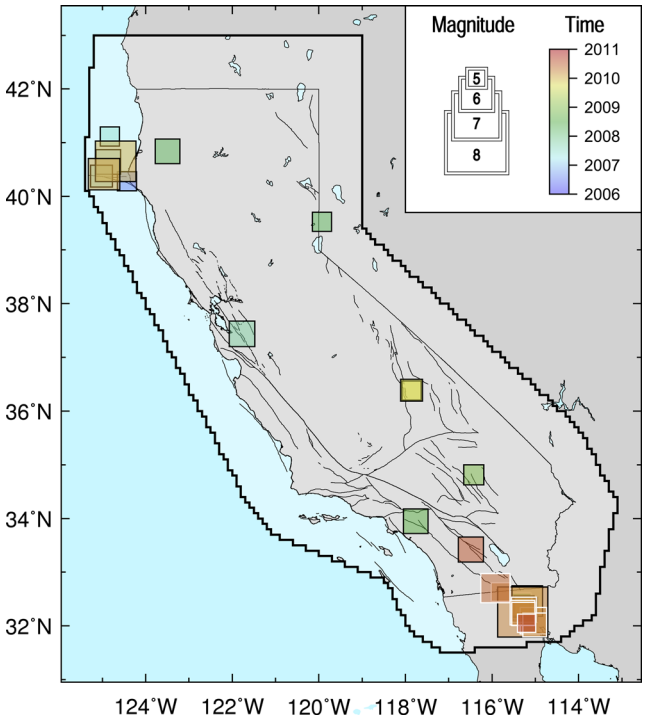

Figure 1. Catalogue of RELM target earthquakes, modified from fig. 1 in Zechar *et al.* (2013). The colour of each square represents the earthquake occurrence time. The squares with white borders are, according to the RELM definition, aftershocks.

In Fig. S1 and Tables S1 and S2, we report the returns for each bin where an earthquake occurred. This detailed breakdown emphasizes that, when masking is allowed, the total pot is not the same for every bin. This type of analysis could provide insight in the event that particular earthquakes are of special interest: for example, if you wanted to emphasize large earthquakes (such as the El Mayor-Cucapah M7.2 earthquake).

Maps showing spatial gambling returns can provide additional insight, and we present two example maps in Figs 3 and 4. We note that the largest gambling returns are for the cells where earthquakes occurred; the probability of observing zero target earthquakes is of the same order of magnitude for almost all forecasts, meaning there is not much 'action' in bins without earthquakes. Nevertheless, we note a slightly rosy tint to the Ebel *et al.* mainshock+aftershock map; this tint, indicating a negative gambling return, corresponds to the overall overprediction mentioned above. Parimutuel gambling maps also reveal features that corresponding maps of likelihood cannot. [We include maps of net returns and likelihood for each forecast in the Supporting Information; see Clements *et al.* (2011) for additional examples of graphical methods for forecast evaluation and comparison.] For example, consider the main shock+aftershock

Table 2. Catalogue of RELM target earthquakes, same as Table 1 in Zechar *et al.* (2013). Italics indicate aftershocks, final column is the magnitude indicated in the Advanced National Seismic System (ANSS) catalogue.

	Origin time (UTC)	Latitude	Longitude	M_{ANSS}
1	2006 May 24, 4:20	32.31	−115.23	5.37
2	2006 July 19, 11:41	40.28	−124.43	5.00
3	2007 February 26, 12:19	40.64	−124.87	5.40
4	2007 May 9, 7:50	40.37	−125.02	5.20
5	2007 June 25, 2:32	41.12	−124.82	5.00
6	2007 October 31, 3:04	37.43	−121.77	5.45
7	2008 February 9, 7:12	32.36	−115.28	5.10
8	*2008 February 11, 18:29*	*32.33*	*−115.26*	*5.10*
9	*2008 February 12, 4:32*	*32.45*	*−115.32*	*4.97*
10	*2008 February 19, 22:41*	*32.43*	*−115.31*	*5.01*
11	2008 April 26, 6:40	39.53	−119.93	5.00
12	2008 April 30, 3:03	40.84	−123.50	5.40
13	2008 July 29, 18:42	33.95	−117.76	5.39
14	2008 November 20, 19:23	32.33	−115.33	4.98
15	2008 December 6, 4:18	34.81	−116.42	5.06
16	2009 September 19, 22:55	32.37	−115.26	5.08
17	2009 October 1, 10:01	36.39	−117.86	5.00
18	2009 October 3, 1:16	36.39	−117.86	5.19
19	2009 December 30, 18:48	32.46	−115.19	5.80
20	2010 January 10, 0:27	40.65	−124.69	6.50
21	2010 February 4, 20:20	40.41	−124.96	5.88
22	2010 April 4, 22:40	32.26	−115.29	7.20
23	*2010 April 4, 22:50*	*32.10*	*−115.05*	*5.51*
24	*2010 April 4, 23:15*	*32.30*	*−115.26*	*5.43*
25	*2010 April 4, 23:25*	*32.25*	*−115.30*	*5.38*
26	*2010 April 4, 0:07*	*32.02*	*−115.02*	*5.32*
27	*2010 April 5, 3:15*	*32.63*	*−115.81*	*4.97*
28	*2010 April 8, 16:44*	*32.22*	*−115.28*	*5.29*
29	*2010 June 15, 4:26*	*32.70*	*−115.92*	*5.72*
30	2010 July 7, 23:53	33.42	−116.49	5.43
31	*2010 September 14, 10:52*	*32.05*	*−115.20*	*4.96*

Kagan *et al.* map (Fig. 4). The trace of the San Andreas Fault is delineated by blue cells, suggesting that the Kagan *et al.* forecast obtained a positive gambling return in these cells by deemphasizing faults relative to the other forecasts. Recall that the ideal forecast has high probabilities in bins where earthquakes occur and low probabilities elsewhere. Again, we note that maps of likelihood could not be used to identify such features (see Figs S2–S5), and maps of information gain are limited to pairwise comparisons.

5 DISCUSSION

As demonstrated with the RELM experiment, the parimutuel gambling score can be used to compare probabilistic seismicity forecasts. In many applications across the statistical sciences, likelihood ratios (or related information criteria, or Bayes factors) are used for pairwise model comparison. (For examples in the context of seismicity forecasting, we refer you to the articles by Schorlemmer *et al.* 2007; Harte & Vere-Jones 2005; Marzocchi *et al.* 2012.) The parimutuel gambling score can also be applied in pairwise fashion, and this permits a direct comparison of the resulting 'head-to-head' metric—a special case of parimutuel gambling with only two bettors—with the log-likelihood ratio. Consider two forecasts for a single observation with corresponding likelihoods of p_1 and p_2, respectively. The log-likelihood ratio of these forecasts is

$$LLR(p_1, p_2) = \log \frac{p_1}{p_2}. \tag{7}$$

The difference in winnings of the two forecasts using the head-to-head score is

$$\Delta R_1 - \Delta R_2 = G(p_1, p_2) = \frac{2p_1}{p_1 + p_2} - 1. \tag{8}$$

We can express the log-likelihood ratio in terms of the gambling score difference

$$LLR(p_1, p_2) = \log \left[\frac{1 + G(p_1, p_2)}{1 - G(p_1, p_2)} \right]. \tag{9}$$

The log-likelihood ratio is a monotone function of the head-to-head gambling score difference, but their relationship is not perfectly linear, and the range of the head-to-head gambling difference is $[-1, 1]$ while the log-likelihood ratio is unbounded. In particular, these measures diverge at low probabilities, with the possibility that the log-likelihood is dominated by a single earthquake in an experiment; the gambling approach avoids such an instability.

The head-to-head gambling score is also closely related to the rate-corrected information gain per earthquake suggested by Rhoades *et al.* (2011). However, as with the log-likelihood ratio, information gains are unbounded; moreover, the analysis suggested by Rhoades *et al.* (2011) groups all bins where no target earthquakes occurred, while parimutuel gambling considers each bin individually.

Model optimization can be thought of as a special case of model comparison, and therefore researchers can use parimutuel gambling to optimize their models. For example, consider the TripleS (Simple Smoothed Seismicity) model of Zechar & Jordan (2010a). This model has a single adjustable parameter: the size of the smoothing kernel to be applied to past epicentres. As is common practice, the optimal value of this parameter is estimated using a retrospective forecast experiment in which the most recent target earthquakes are forecast. Several candidate values of the parameter are used to generate forecasts, and these forecasts are compared. In the implementation for seismicity in Italy, the area skill score (Zechar & Jordan 2010b) was used for comparison; for an illustration using seismicity in China, spatial likelihood was used (Mignan *et al.* 2013). For the reasons we suggested in Section 1 (e.g. stability with respect to low-probability earthquakes), the parimutuel gambling score could be useful in future applications of this model to other regions. Of course, we mention TripleS only as a representative example of a seismicity model with adjustable parameters; you could optimize arbitrarily complex models using the same technique.

The parimutuel gambling analysis in this study is an example of inference based on multiple comparisons, a common research topic in medical studies. In that context, researchers seek to measure the differences between several treatments on many subjects and thereby find the best treatment. The analogy here is forecasts as treatments, bins as patients and gambling returns as patient responses. Hsu (1996) describes a number of inferential methods that apply when making multiple comparisons. For example, perhaps the most important question we can ask is whether the forecast obtaining the largest average gambling return—that is, the 'sample best' forecast—is truly the best. Hsu (1996) shows that to answer this question of 'multiple comparisons with the sample best', you can simply perform a two-sample Student's *t*-test on the sample best and the sample second best. (This assumes normality and equal variances of the responses, which a non-parametric approach could be employed to relax.) For the mainshock group, this corresponds to a paired *t*-test with the samples being the returns in each bin for the Helmstetter *et al.* forecast and the Wiemer–Schorlemmer

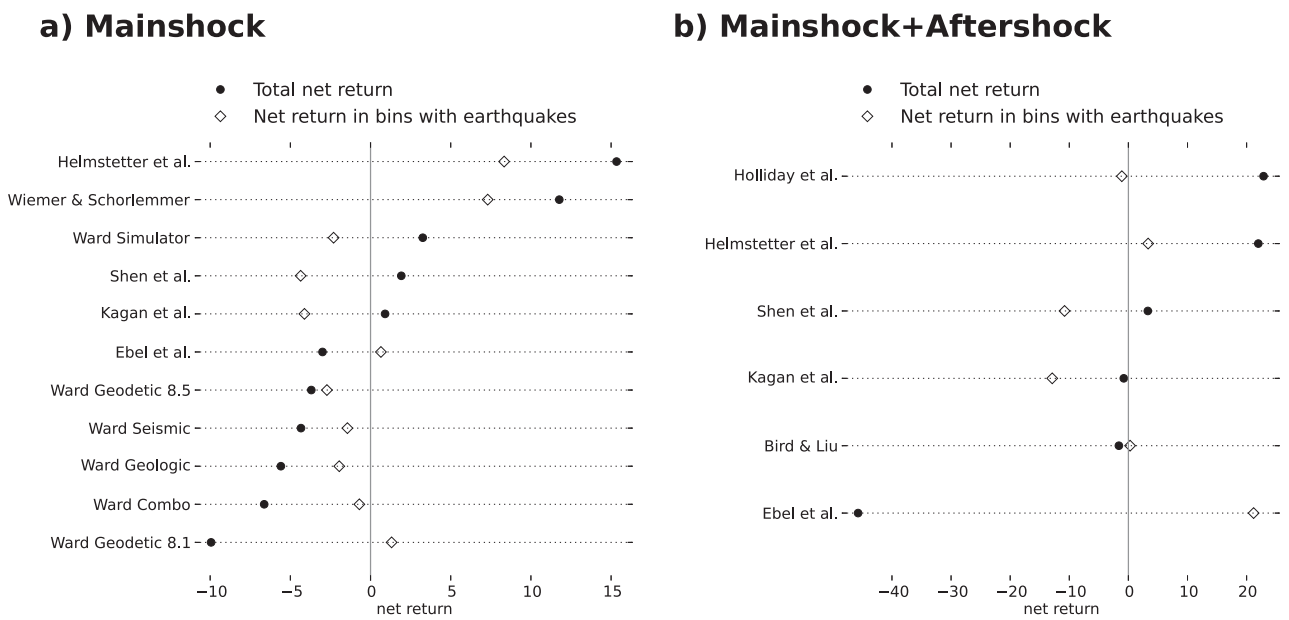

Figure 2. Parimutuel gambling returns from round table analysis of RELM (a) mainshock forecasts and (b) mainshock+aftershock forecasts. Hollow diamonds show the returns based on the bins where target earthquakes occurred; filled circles show total return from all bins. Forecasts are ordered by total return.

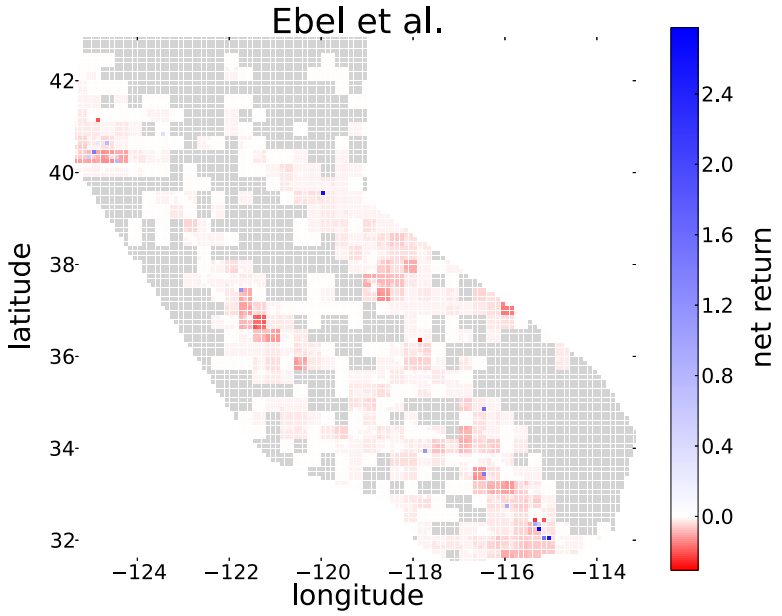

Figure 3. Map showing the spatial distribution of Ebel *et al.*'s gambling returns from round table analysis of RELM mainshock+aftershock forecasts. Light grey indicate cells that were masked (see text for details), white indicates returns near zero, blue indicate positive returns and red negative.

forecast: the resulting p-value is 0.449, suggesting that the performances of the two models are not significantly different. For the mainshock+aftershock group, a p-value of 0.914 suggests that the performances of the Holliday *et al.* (the sample best) and Helmstetter *et al.* forecasts are not significantly different. We note that these results are nearly identical to those reported by Zechar *et al.* (2013), the only difference being that a paired t-test of information gain suggested that the Helmstetter *et al.* forecast was better than the Holliday *et al.* forecast (although, again, without statistical significance).

But we do not want to overemphasize questions of statistical significance: one can imagine employing various Monte Carlo methods to answer such questions, but these methods would likely include questionable assumptions. For example, to simulate additional catalogues based on the observations would be to make the mistake of putting 'the randomness in an intractable place (the Earth)' (Stark 1996) and/or would likely require you to assume that seismicity exhibits some form of stationarity (see Wang *et al.* 2010, for evidence that this assumption does not hold in California). Simulating catalogues based on each model would almost surely indicate that none of the models is the data-generating model for seismicity (i.e. in this situation every model would obtain larger returns with simulated catalogues than its return based on the observed catalogue), but we know this without resorting to simulations. In

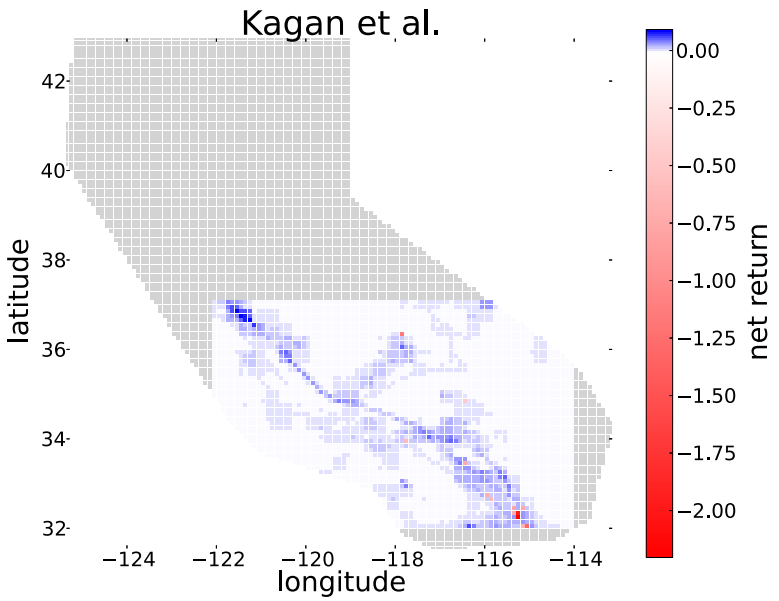

Figure 4. Same as Fig. 3 for Kagan *et al.* mainshock+aftershock forecast.

general, we are not as interested in obtaining a strict statistically significant ranking of models (e.g. Smyth *et al.* 2012) as we are in getting an idea of how models differ from each other. And perhaps more importantly, we ask: in what ways do the models not fit the observations well? Answers to this question can guide model improvement, which is the ultimate goal of this enterprise. Following this line of thinking, we suggest that parimutuel gambling returns could be used as weights to build ensemble forecasts; Taroni *et al.* (2014) did this for a global forecast experiment and found that an ensemble using weights derived from parimutuel gambling returns outperformed individual models as well as several alternative linear combinations. (For a more general treatment of model combination, see Cesa-Bianchi & Lugosi 2006.) And parimutuel gambling analysis could be applied to different model dimensions separately: you could potentially find one model has a superior spatial forecast, another has a superior magnitude forecast, another the best overall rate and use parimutuel gambling weights to combine them.

6 CONCLUSIONS

In this paper, we described a parimutuel gambling method that can be used to compare seismicity forecasts, and we illustrated the method using the 5-yr RELM experiment in California. This method is different from previous techniques because it does not require the choice or construction of a reference forecast model and it can be used to simultaneously compare multiple models. Moreover, this method is intuitive because of the simple analogues in gambling: each model is a bettor, every earthquake forecast bin is a game, you can bet on every possible outcome or even abstain from betting, and so on. And, although we only demonstrated its application to one class of space–rate–magnitude forecasts, it applies generally: you could make similar analyses of probabilistic models in weather, climate, finance, etc.

While scientific forecast experiments similar to RELM have flourished (Gerstenberger & Rhoades 2010; Schorlemmer *et al.* 2010a; Nanjo *et al.* 2011; Eberhard *et al.* 2012; Mignan *et al.* 2013; Taroni *et al.* 2014), there is an increased interest in short-term, time-varying seismicity models and the seismic hazard estimates they inform; such efforts are now referred to as operational earthquake forecasting (OEF; Jordan & Jones 2010; Jordan *et al.* 2011). Other than the shortened forecast horizon, many OEF models are similar to those considered here, and parimutuel gambling can provide guidance for model combination and insight into model performance.

ACKNOWLEDGEMENTS

JDZ was supported in part by NSF grant EAR-0944202 and USGS grant G11AP20038 and JZ was partially supported by JSPS grant-in-aid (C) 25330052. This research was partially carried out in the framework of the REAKT project funded by FP7 of the European Commission, contract number 282862. We thank Editor Duncan Agnew and Yan Kagan for insightful comments, and we thank an anonymous referee for referring us to the multiple comparisons literature. We thank Jonathan Larroquette and Seth Romatelli for pointing us to the term ‘parimutuel’.

REFERENCES

Bird, P. & Liu, Z., 2007. Seismic hazard inferred from tectonics: California, *Seismol. Res. Lett.,* **78**(1), 37–48.

Cesa-Bianchi, N. & Lugosi, G., 2006. *Prediction, Learning, and Games,* 1st edn, Cambridge Univ. Press.

Clements, R.A., Schoenberg, F.P. & Schorlemmer, D., 2011. Residual analysis methods for space-time point processes with applications to earthquake forecast models in California, *Ann. appl. Stat.,* **5**(4), 2549–2571.

Ebel, J.E., Chambers, D.W., Kafka, A.L. & Baglivo, J.A., 2007. Non-Poissonian earthquake clustering and the hidden Markov model as bases for earthquake forecasting in California, *Seismol. Res. Lett.,* **78**(1), 57–65.

Eberhard, D.A.J., Zechar, J.D. & Wiemer, S., 2012. A prospective earthquake forecast experiment in the western Pacific, *Geophys. J. Int.,* **190**(3), 1579–1592.

Evison, F.F. & Rhoades, D.A., 1993. The precursory earthquake swarm in New Zealand: hypothesis tests, *N.Z. J. geol. Geophys.,* **36**(1), 51–60.

Geller, R.J., Jackson, D.D., Kagan, Y.Y. & Mulargia, F., 1997. Earthquakes cannot be predicted, *Science,* **275**(5306), 1616.

Gerstenberger, M.C. & Rhoades, D.A., 2010. New Zealand earthquake forecast testing centre, *Pure appl. Geophys.,* **167**(8–9), 877–892.

Harte, D. & Vere-Jones, D., 2005. The entropy score and its uses in earthquake forecasting, *Pure appl. Geophys.,* **162**(6–7), 1229–1253.

Helmstetter, A., Kagan, Y.Y. & Jackson, D.D., 2007. High-resolution time-independent grid-based forecast for M ≥ 5 earthquakes in California, *Seismol. Res. Lett.,* **78**(1), 78–86.

Holliday, J.R., Nanjo, K.Z., Tiampo, K.F., Rundle, J.B. & Turcotte, D.L., 2005. Earthquake forecasting and its verification, *Nonlinear Process. Geophys.,* **12**(6), 965–977.

Holliday, J.R., Chen, C.-C., Tiampo, K.F., Rundle, J.B., Turcotte, D.L. & Donnellan, A., 2007. A RELM earthquake forecast based on pattern informatics, *Seismol. Res. Lett.,* **78**(1), 87–93.

Hsu, J., 1996. *Multiple Comparisons: Theory and Methods,* CRC Press.

Jordan, T.H. & Jones, L.M., 2010. Operational earthquake forecasting: some thoughts on why and how, *Seismol. Res. Lett.,* **81**(4), 571–574.

Jordan, T.H. *et al.*, 2011. Operational earthquake forecasting: state of knowledge and guidelines for implementation, *Ann. Geophys.,* **54**(4), 316–391.

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

Kagan, Y. & Knopoff, L., 1987. Statistical short-term earthquake prediction, *Science,* **236**(4808), 1563–1567.

Kagan, Y.Y., 2007. On earthquake predictability measurement: information score and error diagram, *Pure appl. Geophys.,* **164**(10), 1947–1962.

Kagan, Y.Y. & Jackson, D.D., 1994. Long-term probabilistic forecasting of earthquakes, *J. geophys. Res.,* **99**(B7), 13685–13700.

Kagan, Y.Y. & Knopoff, L., 1977. Earthquake risk prediction as a stochastic process, *Phys. Earth planet. Inter.,* **14**(2), 97–108.

Kagan, Y.Y., Jackson, D.D. & Rong, Y., 2007. A testable five-year forecast of moderate and large earthquakes in southern California based on smoothed seismicity, *Seismol. Res. Lett.,* **78**(1), 94–98.

Keilis-Borok, V.I. & Shebalin, P.N., 2004. Reverse tracing of short-term earthquake precursors, *Phys. Earth planet. Inter.,* **145**(1–4), 75–85.

Kossobokov, V.G., 2006. Testing earthquake prediction methods: 'The West Pacific short-term forecast of earthquakes with magnitude MwHRV ≥ 5.8', *Tectonophysics,* **413**(1–2), 25–31.

Kossobokov, V.G., 2004. Earthquake prediction: basics, achievements, perspectives, *Acta Geod. Geophys. Hung.,* **39**(2), 205–221.

Kossobokov, V.G., Romashkova, L.L. & Healy, J.H., 1999. Testing earthquake prediction algorithms: statistically significant advance prediction of the largest earthquakes in the Circum-Pacific, 1992- 1997, *Phys. Earth planet. Inter.,* **111,** 187–196.

Luen, B. & Stark, P.B., 2008. Testing earthquake predictions, in *Probability and Statistics: Essays in Honor of David A. Freedman,* pp. 302–315, Institute of Mathematical Statistics.

Marzocchi, W. & Zechar, J.D., 2011. Earthquake forecasting and earthquake prediction: different approaches for obtaining the best model, *Seismol. Res. Lett.,* **82**(3), 442–448.

Marzocchi, W., Sandri, L. & Boschi, E., 2003. On the validation of earthquake-forecasting models: the case of pattern recognition algorithms, *Bull. seism. Soc. Am.,* **93**(5), 1994–2004.

Marzocchi, W., Zechar, J.D. & Jordan, T.H., 2012. Bayesian forecast evaluation and ensemble earthquake forecasting, *Bull. seism. Soc. Am.,* **102**(6), 2574–2584.

Mignan, A., Jiang, C., Zechar, J.D., Wiemer, S., Wu, Z. & Huang, Z., 2013. Completeness of the mainland China earthquake catalog and implications for the setup of the China earthquake forecast testing center, *Bull. seism. Soc. Am.,* **103**(2A), 845–859.

Molchan, G.M., 1991. Structure of optimal strategies in earthquake prediction, *Tectonophysics,* **193**(4), 267–276.

Molchan, G.M., 1997. Earthquake prediction as a decision-making problem, *Pure appl. Geophys.,* **149**(1), 233–247.

Molchan, G.M., 2010. Space–time earthquake prediction: the error diagrams, *Pure appl. Geophys.,* **167**(8–9), 907–917.

Molchan, G.M., 2011. On the testing of seismicity models, *Acta Geophys.,* **60**(3), 624–637.

Molchan, G.M. & Kagan, Y.Y., 1992. Earthquake prediction and its optimization, *J. geophys. Res.,* **97**(B4), 4823–4838.

Molchan, G.M. & Keilis-Borok, V.I., 2008. Earthquake prediction: probabilistic aspect, *Geophys. J. Int.,* **173**(3), 1012–1017.

Molchan, G.M. & Romashkova, L., 2010. Earthquake prediction analysis based on empirical seismic rate: the M8 algorithm, *Geophys. J. Int.,* **183**(3), 1525–1537.

Molchan, G.M. & Romashkova, L., 2011. Gambling score in earthquake prediction analysis, *Geophys. J. Int.,* **184**(3), 1445–1454.

Nanjo, K.Z., Tsuruoka, H., Hirata, N. & Jordan, T.H., 2011. Overview of the first earthquake forecast testing experiment in Japan, *Earth Planets Space,* **63,** 159–169.

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

Rhoades, D.A., Schorlemmer, D., Gerstenberger, M.C., Christophersen, A., Zechar, J.D. & Imoto, M., 2011. Efficient testing of earthquake forecasting models, *Acta Geophys.,* **59**(4), 728–747.

Schorlemmer, D., Gerstenberger, M.C., Wiemer, S., Jackson, D.D. & Rhoades, D.A., 2007. Earthquake likelihood model testing, *Seismol. Res. Lett.,* **78**(1), 17–29.

Schorlemmer, D., Christophersen, A., Rovida, A., Mele, F., Stucchi, M. & Marzocchi, W., 2010a. Setting up an earthquake forecast experiment in Italy, *Ann. Geophys.,* **53**(3), 1–9.

Schorlemmer, D., Zechar, J.D., Werner, M.J., Field, E.H., Jackson, D.D. & Jordan, T.H., 2010b. First results of the Regional Earthquake Likelihood Models experiment, *Pure appl. Geophys.,* **167**(8–9), 859–876.

Shen, Z.-K., Jackson, D.D. & Kagan, Y.Y., 2007. Implications of geodetic strain rate for future earthquakes, with a five-year forecast of M5 earthquakes in southern California, *Seismol. Res. Lett.,* **78**(1), 116–120.

Smyth, C., Yamada, M. & Mori, J., 2012. Earthquake forecast enrichment scores, *Res. Geophys.,* **2**(1), 7–12.

Stark, P., 1996. A few considerations for ascribing statistical significance to earthquake predictions, *Geophys. Res. Lett.,* **23**(11), 1399–1402.

Stark, P.B., 1997. Earthquake prediction: the null hypothesis, *Geophys. J. Int.,* **131**(3), 495–499.

Taroni, M., Zechar, J.D. & Marzocchi, W., 2014. Assessing annual global M6+ seismicity forecasts, *Geophys. J. Int.,* **196**(1), 422–431.

Wang, Q., Jackson, D. & Zhuang, J., 2010. Are spontaneous earthquakes stationary in California? *J. Geophys. Res.,* **115**(B8), B08310, doi:10.1029/2009JB007031.

Ward, S.N., 2007. Methods for evaluating earthquake potential and likelihood in and around California, *Seismol. Res. Lett.,* **78**(1), 121–133.

Wiemer, S. & Schorlemmer, D., 2007. ALM: an asperity-based likelihood model for California, *Seismol. Res. Lett.,* **78**(1), 134–140.

Zechar, J.D. & Jordan, T.H., 2008. Testing alarm-based earthquake predictions, *Geophys. J. Int.,* **172**(2), 715–724.

Zechar, J.D. & Jordan, T.H., 2010a. Simple smoothed seismicity earthquake forecasts for Italy, *Ann. Geophys.,* **53**(3), 99–105.

Zechar, J.D. & Jordan, T.H., 2010b. The area skill score statistic for evaluating earthquake predictability experiments, *Pure appl. Geophys.,* **167** (8–9), 893–906.

Zechar, J.D. & Zhuang, J., 2010. Risk and return: evaluating Reverse Tracing of Precursors earthquake predictions, *Geophys. J. Int.,* **182**(3), 1319–1326.

Zechar, J.D., Schorlemmer, D., Werner, M.J., Gerstenberger, M.C., Rhoades, D.A. & Jordan, T.H., 2013. Regional Earthquake Likelihood Models I: first-order results, *Bull. seism. Soc. Am.,* **103**(2A), 787–798.

Zhuang, J., 2010. Gambling scores for earthquake predictions and forecasts, *Geophys. J. Int.,* **181**(1), 382–390.

Zhuang, J. & Jiang, C., 2012. Scoring annual earthquake predictions in China, *Tectonophysics,* **524,** 155–164.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article: For the sake of transparency and reproducibility, all codes and data to perform analyses, generate figures, generate supplementary figures and generate supplementary tables are included in the src directory, you can do all these things with relm_gambling.py, have a look there. Then run the .sh scripts in figures/used to create figures/ to produce the final figures.

Figure S1. From the round table analysis of the RELM experiment, these are the parimutuel net returns in the bins where target earthquakes occurred. The earthquake numbers on the horizontal axis

refer to the earthquakes in Table 2. In the legend, for each forecast the net return for all bins where target earthquakes occurred is shown. (These values are the same as the hollow diamonds in Fig. 2.) Part a) shows results for all mainshock forecasts, part b) shows results for mainshock+aftershock forecasts. The total return for a forecast is usually not dominated by a single earthquake (as it may be in fixed-odds gambling), and no model is superior for all earthquakes.

Figure S2. Maps of the spatial distribution of gambling returns from round table analysis of RELM mainshock forecasts. The colour scale varies from map to map, but pure white is always zero, blue is always positive, and red is always negative. Each pixel is the net return for all magnitude bins at the plotted location. Forecast cells that were designated by the model developer as masked are excluded from analysis and therefore not plotted here.

Figure S3. Same as Fig. S2 but showing spatial distribution of log-likelihood rather than gambling returns.

Figure S4. Same as Fig. S2 for mainshock+aftershock forecasts.

Figure S5. Same as Fig. S3 for mainshock+aftershock forecasts.

Table S1. Values from Fig. S1(a), parimutuel net returns in the bins where target earthquakes occurred for mainshock forecasts.

Table S2. Values from Fig. S1(b), parimutuel net returns in the bins where target earthquakes occurred for mainshock+aftershock forecasts (http://gji.oxfordjournals.org/lookup/suppl/doi:10.1093/gji/ggu137/-/DC1).