Geophysical Journal International

Geophys. J. Int. (2010) **182,** 1319–1326 doi: 10.1111/j.1365-246X.2010.04666.x

Risk and return: evaluating Reverse Tracing of Precursors earthquake predictions

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

[1] *Swiss Seismological Service, Institute of Geophysics, ETH Zurich, Sonneggstrasse* 5, 8092 *Zurich, Switzerland. E-mail: jeremy.zechar@sed.ethz.ch*
[2] *Lamont-Doherty Earth Observatory, Columbia University, NY, USA*
[3] *Institute of Statistical Mathematics,* 10-3 *Midori-Cho, Tachikawa-Shi, Tokyo* 190-8562, *Japan*

Accepted 2010 May 17. Received 2010 May 3; in original form 2010 February 25

SUMMARY

In 2003, the Reverse Tracing of Precursors (RTP) algorithm attracted the attention of seismologists and international news agencies when researchers claimed two successful predictions of large earthquakes. These researchers had begun applying RTP to seismicity in Japan, California, the eastern Mediterranean and Italy; they have since applied it to seismicity in the northern Pacific, Oregon and Nevada. RTP is a pattern recognition algorithm that uses earthquake catalogue data to declare alarms, and these alarms indicate that RTP expects a moderate to large earthquake in the following months. The spatial extent of alarms is highly variable and each alarm typically lasts 9 months, although the algorithm may extend alarms in time and space. We examined the record of alarms and outcomes since the prospective application of RTP began, and in this paper we report on the performance of RTP to date. To analyse these predictions, we used a recently developed approach based on a gambling score, and we used a simple reference model to estimate the prior probability of target earthquakes for each alarm. Formally, we believe that RTP investigators did not rigorously specify the first two 'successful' predictions in advance of the relevant earthquakes; because this issue is contentious, we consider analyses with and without these alarms. When we included contentious alarms, RTP predictions demonstrate statistically significant skill. Under a stricter interpretation, the predictions are marginally unsuccessful.

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

1 INTRODUCTION

In this paper, we present an analysis of earthquake predictions that were generated by Reverse Tracing of Precursors (RTP), a pattern recognition algorithm intended to predict large earthquakes in a time window of at least 9 months (Keilis-Borok *et al*. 2004; Shebalin *et al*. 2004, 2006). Below, we provide a conceptual overview of RTP with an emphasis on the details relevant to this study, and we refer the interested reader to Section 2 and Table 3 of Shebalin *et al.* (2004) for a comprehensive description of the algorithm.

RTP searches for seismicity precursors in two distinct steps: short-term recognition and intermediate-term confirmation. In the first step, the algorithm decomposes a declustered regional earthquake catalogue into chains, which are composed of neighbours (Shebalin 2006); two earthquakes are neighbours if they occur sufficiently close in space and time, where the definition of 'sufficiently close' depends on the magnitudes of the two earthquakes. After grouping all earthquakes into chains, RTP keeps only the chains that have a large spatial extent and comprise several events. The algorithm interprets chains that satisfy these criteria as signals of a short-term precursor to a target earthquake. A chain's spatial domain is delimited by the union of circles of radius r that are centred on each epicentre in the chain and the area connecting these circles. See Fig. 1 for an example of a chain's spatial extent with varying r.

The second step of RTP seeks to confirm (or disconfirm) the short-term precursor by searching for intermediate-term patterns within each precursory chain's spatial domain. In this step, values of eight precursory functions (described in Table 2 of Shebalin *et al*. 2004) are computed. These functions capture four types of supposed precursory behaviour: increased seismic activity, increased spatial clustering, increased correlation length and a transformed relationship between earthquake frequency and magnitude. Researchers have used such precursory behaviours in previous prediction studies with some limited success (see reviews by Keilis-Borok 2002, 2003).

The algorithm computes the values of the intermediate-term precursory functions in sliding time windows, and it defines threshold values for each combination of precursory function and windowing value, yielding 64 threshold values. If many of these threshold values are exceeded, RTP declares an alarm that lasts for 9 months

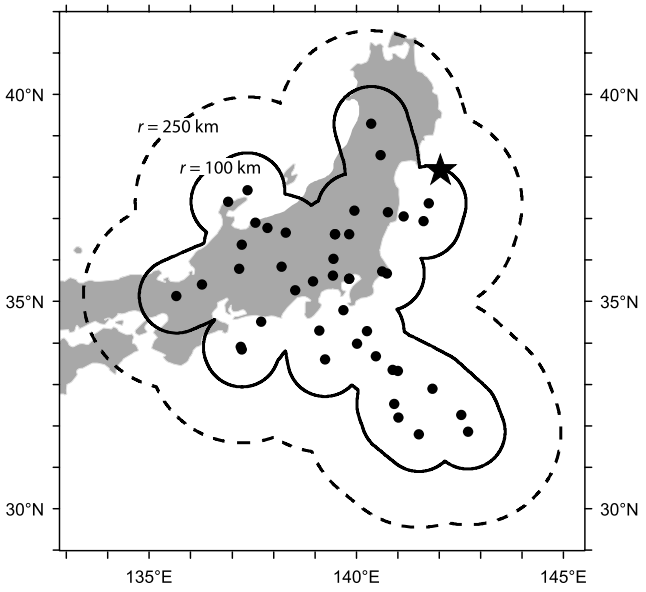

Figure 1. Example RTP alarm (modified from Shebalin, private communication, 2005): Alarm 10 from Table 2. Solid and dashed lines indicate the boundaries of the spatial extent of this alarm with $r = 100$ and 250 km, respectively. The circles represent the epicentres of the earthquakes forming the short-term chain, and the star shows the location of the contentious Earthquake 6.

from the end of the chain's formation—the point in time at which RTP recognized the chain—and covers the chain's spatial domain. RTP extends alarms if a chain continues to grow in space and time. Declaration of an alarm indicates that RTP expects at least one earthquake above a specified minimum magnitude within the space–time domain of the alarm.

The intermediate-term patterns used in RTP find fluctuations in seismicity, but the patterns are not based on a specific physical mechanism. On the other hand, the short-term component of RTP is based on a seismicity pattern that identifies a rapid increase of earthquake correlation range (Shebalin 2006). Researchers have found this pattern in seismicity models (Gabrielov *et al*. 1999, 2000) and in regional earthquake observations (Shebalin *et al*. 2000; Zaliapin *et al*. 2002). One physical explanation of this pattern is that distant fault network elements begin to interact as they approach a critical state; when this state is reached, several fault elements rupture simultaneously and produce a large earthquake.

Since late 2003, RTP investigators have been conducting a prospective prediction experiment. Beginning in 2004, they announced RTP alarms via emails to a group of colleagues, and details of the alarms are archived at http://www.rtptest.org. In this paper, we describe and present an up-to-date independent evaluation of the RTP experiment. We emphasize that we evaluated the RTP alarm statements; that is, in this study we did not attempt to recreate the RTP algorithm or reproduce the resulting alarms. The prospective prediction experiment includes study regions in Japan, California–Oregon–Nevada, Italy, the eastern Mediterranean and the northern Pacific. The polygon coordinates which enclose each study region are listed in Table 1. So far, RTP has produced 29 alarms; the characteristics of each are reported in Table 2. During this experiment, 14 target earthquakes have been observed, and these are listed in Table 3. Some of the outcomes analysed in this study are contentious, and therefore we considered evaluations of RTP with and without these alarms and earthquakes; a detailed explanation follows in the next section.

2 CONTENTIOUS ALARMS AND EARTHQUAKES

For several alarms that are listed in Table 2, determining the outcome of the alarm is not straightforward. For example, Shebalin *et al*. (2004) claimed that Alarm 1 successfully predicted the September 2003 Tokachi-oki earthquake (Earthquake 1 in Table 3). But Alarm 1 was initially presented in July 2003 as a successful retrospective prediction of the 2003 May 26 M 7.0 earthquake, not as a

Table 1. List of study regions and corresponding earthquake catalogues. For Japan, the JMA catalogue was used for Alarm 1, because this was the catalogue used to determine that alarm (P. Shebalin, private communication, 2009). For subsequent alarms in the Japanese region, the Global Centroid Moment Tensor (GCMT) catalogue (Dziewonski *et al*. 1981; Ekström *et al*. 2005) was used.

Name	Catalogue of interest[a]	Polygon enclosing study region
Japan	JMA 1923–2003 June 1[b] CMT 1977–2003 June 1	(30,140), (38,136), (49,136), (49,153), (46,156), (31,144)
California	ANSS 1932–2003 June 1	(31.5,−114), (31.5,−120), (39,−124.75), (39,−130), (44,−130), (44, −120), (41, −120), (41, −116), (35, −116), (35, −114)
Eastern Mediterranean	CMT 1977–2003 June 1	(28,32), (36,32), (36,38), (28,38)
Italy	PDE 1973–2003 June 1	(41,18), (41,10), (43.84,10), (43,5), (47,5), (47,17), (45,17), (44.5,14)
North Pacific	CMT 1977–2006 November 1	(45, −175), (53, −140), (40, −130), (21, −115), (25, −105), (45, −120), (65, −140), (65, −150), (55,180), (60,164), (41,134), (32,134), (32,148), (36,148), (50,165)

Notes: [a]Catalogue references: JMA—produced by the Japanese Meteorological Agency (JMA), 1921–2007, http://www.hinet.bosai.go.jp. CMT—CMT Earthquake Catalogue, produced by the Global Centroid Moment Tensor (GCMT) group, 1976–2007, http://www.globalcmt.org. ANSS—ANSS Earthquake Catalogue, produced by Advanced National Seismic System (ANSS) and hosted by the Northern California Data Center (NCEDC), 1932–2007, http://quake.geo.berkeley.edu/anss.
[b]Used for Alarm 1 only.

Table 2. List of all announced RTP alarms, including the study region, time range of interest, target magnitude range and alarm radius. Alarms are listed in chronological order according to their start date. The epicentres that define the spatial extent of each alarm region are listed in Supporting Information.

Alarm no., i	Region	Alarm start	Alarm end	Magnitude	Radius, r (km)
1[a]	Japan	2003 March 27	2003 November 27	$M_{\rm JMA} \geq 7.0$	75
2[a]	California	2003 May 5	2004 February 27	$M_{\rm ANSS} \geq 6.4$	75
3	California	2003 November 13	2004 September 5	$M_{\rm ANSS} \geq 6.4$	50
4	Japan	2004 February 8	2004 November 8	$M_{\rm w} \geq 7.2$	100
5[a]	Italy	2004 February 29	2004 November 29	$M_{\rm w} \geq 5.5$	50
6	California	2004 November 14	2005 August 14	$M_{\rm ANSS} \geq 6.4$	50
7	California	2004 November 16	2005 August 16	$M_{\rm ANSS} \geq 6.4$	50
8	Italy	2004 December 31	2005 October 1	$M_{\rm PDE} \geq 5.5$	50
9	Italy	2005 May 6	2006 February 6	$M_{\rm PDE} \geq 5.5$	50
10[a]	Japan	2005 June 2	2006 March 2	$M_{\rm w} \geq 7.2$	100
11	California	2005 June 17	2006 March 17	$M_{\rm ANSS} \geq 6.4$	50
12	California	2006 March 18	2006 September 18	$M_{\rm ANSS} \geq 6.4$	50
13	California	2006 March 24	2006 December 24	$M_{\rm ANSS} \geq 6.4$	50
14	California	2006 December 25	2007 May 2	$M_{\rm ANSS} \geq 6.4$	50
15	Italy	2006 May 2	2007 February 3	$M_{\rm PDE} \geq 5.5$	100
16	Japan	2006 May 11	2007 February 11	$M_{\rm w} \geq 7.2$	100
17	California	2006 September 23	2007 June 23	$M_{\rm ANSS} \geq 6.4$	50
18	Japan	2006 September 30	2007 June 30	$M_{\rm w} \geq 7.2$	100
19	North Pacific	2006 October 28	2007 July 28	$M_{\rm w} \geq 7.2$	100
20	California	2007 January 17	2007 October 17	$M_{\rm ANSS} \geq 6.4$	50
21	California	2007 May 3	2008 January 28	$M_{\rm ANSS} \geq 6.4$	50
22	California	2007 October 18	2008 January 14	$M_{\rm ANSS} \geq 6.4$	50
23	North Pacific	2007 July 29	2008 January 28	$M_{\rm w} \geq 7.2$	100
24	North Pacific	2007 August 24	2008 May 24	$M_{\rm w} \geq 7.2$	100
25	California	2008 January 29	2008 September 26	$M_{\rm ANSS} \geq 6.4$	50
26	Italy	2008 April 7	2009 January 7	$M_{\rm PDE} \geq 5.5$	50
27	California	2008 April 14	2009 January 14	$M_{\rm ANSS} \geq 6.4$	50
28	North Pacific	2008 July 17	2009 April 17	$M_{\rm w} \geq 7.2$	100
29	California	2009 January 29	2009 October 29	$M_{\rm ANSS} \geq 6.4$	50

Note: [a]Contentious alarm; see Section 2 for details.

Table 3. Earthquakes with magnitudes that equal or exceed the typical target magnitude for the RTP study region in which they occurred, having occurred since testing began in the respective region. Events with magnitudes listed as $M_{\rm w}$ are listed as reported from the GCMT catalogue.

Earthquake no.	Region	Origin date	Magnitude	Latitude (°)	Longitude (°)	Inside alarm?
1[a]	Japan	2003 September 25	$M_{\rm JMA} = 8.0$	42.21	143.84	Maybe
2[a]	Japan	2003 September 25	$M_{\rm w} = 7.3$	41.75	143.62	Maybe
3[a]	California	2003 December 22	$M_{\rm ANSS} = 6.5$	35.70	−121.10	Maybe
4	California	2005 June 15	$M_{\rm ANSS} = 7.2$	41.29	−125.95	No
5	California	2005 June 17	$M_{\rm ANSS} = 6.6$	40.77	−126.57	No
6[a]	Japan	2005 August 16	$M_{\rm w} = 7.2$	38.24	142.05	Maybe
7	Japan	2006 November 15	$M_{\rm w} = 8.3$	46.71	154.33	Yes
8	Japan	2007 January 13	$M_{\rm w} = 8.1$	46.17	154.80	Yes
9	North Pacific	2007 December 19	$M_{\rm w} = 7.2$	51.02	−179.27	Yes
10	North Pacific	2008 July 5	$M_{\rm w} = 7.7$	54.12	153.37	No
11	North Pacific	2008 November 24	$M_{\rm w} = 7.3$	54.27	154.71	No
12	North Pacific	2009 January 15	$M_{\rm w} = 7.4$	46.97	155.39	No
13	Italy	2009 April 6	$M_{\rm w} = 6.3$	42.33	13.33	No
14	Italy	2009 April 7	$M_{\rm w} = 5.5$	42.28	13.46	No

Note: [a]Contentious earthquake; see Section 2 for details.

prospective prediction (Shebalin *et al*. 2003). In addition, Shebalin *et al*. (2003) did not include the magnitude range of interest or the alarm duration in the alarm description. One could argue that the magnitude range could be inferred from the fact that the 2003 May 26 event was considered a target earthquake, but the alarm duration was not specified. Therefore, it is contentious to interpret Alarm 1 as a successful prediction of Earthquakes 1 and 2.

Similarly, RTP investigators declared Alarm 2 in a memo sent to colleagues, and the memo included neither an explicit statement of the alarm time window nor a definition of the target earthquake; rather, the authors used vague phrases such as 'preparing for a major earthquake' (K. Aki *et al.*, private communication, 2003). Therefore, it is suspect to interpret Alarm 2 as a successful prediction of the 2003 San Simeon earthquake (Earthquake 3).

For Alarm 5 in Italy, RTP investigators announced the target magnitude range as $M_{\rm w} \geq 5.5$, although they had developed and optimized RTP using the 'official magnitude' of the PDE catalogue (P. Shebalin, private communication, 2009). If the PDE catalogue

contains more than one estimate of magnitude for an earthquake, the official magnitude is the maximum of all estimates. An event with M_L 5.7 occurred within the space–time extent of Alarm 5, but, because the reported moment magnitude was M_w 5.2, Alarm 5 formally cannot be considered successful.

Because of a delay in the earthquake catalogue data used to search for chains in the Japan testing region, RTP investigators did not declare Alarm 10 until after Earthquake 6 occurred (P. Shebalin private communication, 2009). If they had declared this alarm in advance of the earthquake, it would be a clear success; yet given the circumstances, the success of Alarm 10 as a prospective prediction of Earthquake 6 is questionable.

Because these alarms are contentious, we present two evaluations: one that includes these alarms and the corresponding target earthquakes, and another that excludes them. We refer to these evaluations as the loose interpretation and the strict interpretation, respectively.

3 EVALUATION METHODS

In the context of deterministic prediction of binary events, there are four possible outcomes: If one predicts that an event will happen (a positive prediction) and it happens, this is a hit. If one predicts that an event will happen but it does not happen, this is a false alarm. If one predicts that an event will not happen (a negative prediction) but it does happen, this is a miss. If one predicts that an event will not happen and it does not happen, this is a correct negative.

To measure the skill of a set of predictions, one can organize the outcomes in a contingency table, which denotes the frequency of each outcome, and choose some related metric to quantify skill (Mason 2003). In the specific context of deterministic earthquake prediction, one such metric that has been used is the *R*-score (also known as the Hanssen–Kuiper skill score) (Shi *et al*. 2001; Harte & Vere-Jones 2005). The *R*-score is defined as the difference between the hit rate and the false alarm rate

$$R = \frac{a}{a+c} - \frac{b}{b+d}, \tag{1}$$

where a is the number of hits, b is the number of false alarms, c is the number of misses and d is the number of correct negatives. The *R*-score of RTP is uninformative because RTP only produces positive predictions; therefore, c and d are zero and, regardless of the alarms and outcomes, R is always zero. Wherever there is no RTP alarm in space and time, we interpreted this as a statement of no prediction rather than inferring a negative prediction; we discuss the alternative interpretation in Section 6. We note that most contingency table metrics, including the *R*-score, implicitly assume that the probability of a hit is the same for each prediction; this is not the case for RTP alarms (details in Section 4), and it is rarely the case for earthquake predictions in general.

Another diagnostic that is commonly used to evaluate the performance of alarm-based earthquake predictions is the Molchan diagram (Molchan & Keilis-Borok 2008; Zechar & Jordan 2008), which compares the fraction of space–time–magnitude covered by alarms with the miss rate, $\nu = c/(a+c)$. Unfortunately, because RTP does not issue negative alarms, the miss rate of RTP is always zero, and therefore the Molchan diagram approach is also uninformative.

Because the usual contingency table measures are not informative for RTP, we used the gambling score of Zhuang (2010). This scoring approach requires the explicit choice of a reference model for earthquake probability; the Poisson process is a reasonable reference model for 'independent' events (Gardner & Knopoff 1974), and the Omori–Utsu distribution (Utsu *et al*. 1995) is appropriate for evaluating 'aftershock' forecasts. For evaluation of RTP alarms, the Poisson reference model yields prior probabilities for the success of each alarm.

To understand the gambling score method, suppose that the reference model indicates a probability p that at least one target earthquake will occur in a given space–time–magnitude window. Think of the forecaster as a gambler and a prediction as being a bet of one credit of professional reputation. RTP does not yield negative predictions, so one only needs to consider positive predictions. (We refer the reader to Zhuang (2010) for details on negative predictions and an extension to probabilistic forecasts.) The RTP forecaster can bet one reputation credit on 'Yes', or he may abstain from betting if no RTP prediction is available. If no target earthquake satisfies the alarm, the forecaster loses and his reputation is diminished by one credit. If at least one target earthquake satisfies the alarm, he correctly bet on 'Yes' and he gains credits according to the return ratio for 'Yes' bets, $r_{\mathrm{YES}} = (1-p)/p$. This ratio is designed such that the expected change in reputation, ΔR, for this bet is zero if the reference model is an exact representation of the system (i.e. it is the 'true' model):

$$\mathrm{E}[\Delta R] = r_{\mathrm{YES}} p - (1-p) = 0, \tag{2}$$

where $E[x]$ denotes the expectation of x. Zhuang (2010) proved that if a set of predictions obtains a positive ΔR, the prediction model is superior to the reference model; in particular, a gain in reputation indicates that the correlation between the prediction model and the (unknown) true model is greater than the correlation between the reference model and the true model.

4 REFERENCE PROBABILITY ANALYSIS

The gambling score method requires the explicit choice of a reference model with which to compare a forecaster's bets. Considering alarm-based predictions, the reference model yields an estimate of the probability that an alarm will be successful or, equivalently, the probability that at least one earthquake will occur in the space–time–magnitude domain of an alarm. Assuming that target earthquake rates are well described by a Poisson process, the reference model needs to estimate only the probability of at least one earthquake occurring in the space–*duration*–magnitude domain of an alarm, without regard for the specific time period of the alarm. In other words, if the alarm duration is 9 months, the Poisson reference probability is relevant to any 9-month period covering the same space–magnitude domain, not only the 9-month period during which the alarm is active. Although there is ample evidence that the Poisson distribution is less than ideal for describing earthquake number variation (Kagan 2010; Schorlemmer *et al*. 2010), we nevertheless used a Poisson reference model because of its uniform applicability to all alarms and because of its simplicity—namely, it includes few assumptions and it is easily interpreted. To estimate the reference probability for the ith alarm A_i, we used the relevant catalogue (Table 1) and computed the rate of earthquakes, $N(m_i)$, in A_i's space–magnitude domain. When estimating this rate, we included all events that occurred before the alarm was declared. Under the Poisson assumption, the probability of observing at least one earthquake in the space–magnitude domain of A_i in a time window of duration t_i days is

$$p_i = 1 - \exp(-N(m_i)t_i). \tag{3}$$

Table 4. Duration t_i, minimum magnitude m_i, historical daily rate, $N(m_i)$ or $\tilde{N}(m_i)$, and corresponding Poisson reference probability p_i (from eq. 3) for each alarm. When the historical rate $N(m_i)$ was zero, we used the rate of smaller events $\tilde{N}(m_i) \sim N(m_i - \Delta m)$ to estimate the alarm's success probability. If $\tilde{N}(m_i)$ was zero, we arbitrarily set the rate to one-half earthquake per the duration of the catalogue; this is marked with an asterisk.

Alarm no., i	t_i (days)	m_i	$N(m_i)$	$\tilde{N}(m_i)$	p_i (%)
1	246	7.2	1.40E−4	–	29.12
2	299	6.4	7.68E−5	–	2.27
3	308	6.4	3.05E−4	–	8.96
4	275	7.2	1.01E−4	–	2.74
5	275	5.5	7.03E−4	–	17.58
6	274	6.4	1.50E−4	–	4.03
7	274	6.4	–	1.01E−4	2.74
8	275	5.5	5.13E−4	–	13.17
9	277	5.5	7.62E−4	–	19.03
10	274	7.2	9.63E−5	–	2.61
11	274	6.4	2.61E−4	–	6.90
12	185	6.4	1.11E−4	–	2.03
13	276	6.4	3.32E−4	–	8.76
14	129	6.4	2.92E−4	–	3.70
15	278	5.5	8.21E−4	–	20.42
16	277	7.2	9.33E−4	–	22.77
17	274	6.4	–	7.33E−5	1.99
18	274	7.2	6.44E−4	–	16.18
19	274	7.2	3.67E−4	–	9.57
20	274	6.4	1.09E−4	–	2.95
21	271	6.4	2.54E−4	–	6.66
22	89	6.4	1.08E−4	–	0.96
23	184	7.2	3.58E−4	–	6.38
24	275	7.2	–	4.47E−5	1.22
25	242	6.4	2.52E−4	–	5.91
26	276	5.5	1.55E−4	–	4.20
27	276	6.4	–	1.04E−4	2.83
28	275	7.2	–	3.58E−6*	0.98
29	274	6.4	7.10E−5	–	1.93

For any alarm where $N(m_i) = 0$—that is, the catalogue contained no earthquakes with magnitude at least m_i in the alarm's spatial domain—we estimated the rate of target earthquakes using $N(m_i - \Delta m)$, a rate that includes smaller events. We did this because earthquake catalogues are finite and a zero rate/probability seems unphysical. In estimating this rate of target earthquakes, we further assumed that the magnitude–frequency distribution follows the Gutenberg–Richter relation with a b-value of unity:

$$\tilde{N}(m_i) = \frac{N(m_i - \Delta m)}{10^{\Delta m}}. \tag{4}$$

For alarms where $N(m_i) = 0$, we chose $\Delta m = 1$ and substituted the estimate $\tilde{\mathrm{N}}(m_i)$ of eq. 4 for $N(m_i)$ in eq. 3. For Alarm 28, even this estimation yielded a zero rate. For this special situation, we arbitrarily set the rate to one earthquake per twice the duration of the catalogue. This choice corresponds to a posterior estimate in which the rate has a prior density that is proportional to the likelihood.

For each alarm, we report in Table 4 the minimum magnitude of interest, duration, estimated daily rate of earthquakes and corresponding prior probability. The probabilities represent the chance of each alarm's success under the Poisson reference model, and they vary widely from alarm to alarm. As mentioned in the previous section, this variation makes the application of most contingency table measures invalid.

We emphasize that the reference probabilities for false alarms are not used for the RTP gambling score calculation and we report them here to be consistent and to facilitate comparison of alarms. Therefore, our most uncertain estimates (for Alarm 28 and all others based on eq. 4) do not affect the gambling score results.

5 RESULTS

In Table 5, we report the results of testing the RTP alarms with the gambling score and a Poisson reference model. Because some of the outcomes are contentious, we present results from two end-member interpretations. The loose interpretation counts the contentious alarms as successful predictions, whereas the strict interpretation ignores the contentious alarms or, for Alarm 5, counts them as false alarms. The total return from the 29 RTP alarms is 84.4 credits under the loose interpretation and −4.15 credits under the strict interpretation. Because the former value is positive, under the loose interpretation the gambling score deems RTP alarms superior to the Poisson reference model. On the other hand, the strict interpretation result indicates performance that is marginally worse than the Poisson model. The biggest difference between the results is caused by the reputation credits associated with Alarm 2 and the 2003 San Simeon earthquake (Earthquake 3), which occurred in a region of relatively low historical seismic activity. Neglecting this earthquake and the corresponding alarm (as we did under the strict interpretation), the total reputation gain is negative. The other major contribution to the reputation gain is Alarm 23 (Earthquake 9), which resulted in a 14.7 credit increase.

The gambling score approach allows a straightforward hypothesis test regarding the RTP alarms. In the context of this study, the null hypothesis is that the performance of the RTP alarms, quantified by the net gambling score return, is no better than what could be expected from the Poisson reference model. In other words, we are interested in the question: is the RTP reputation change significantly larger than what one would obtain if bets were placed according to the reference probabilities? To address this question, we used a simple simulation method. For each RTP alarm, a number was chosen randomly from the uniform distribution on [0, 1). If the selected number was smaller than the reference probability for the RTP alarm, an identical, positive alarm was declared; if the number was larger, a negative alarm covering the same space–time–magnitude volume was declared. After this process was repeated for all RTP alarms, a gambling score was computed for this set of simulated alarms $\{X_i\}$:

$$\Delta R = \sum_i \left[X_i Y_i \left(\frac{1 - p_i}{p_i} \right) + (1 - X_i)\, Y_i\, (-1) + X_i\,(1 - Y_i)(-1) + (1 - X_i)(1 - Y_i) \left(\frac{p_i}{1 - p_i} \right) \right]. \tag{5}$$

Here, $X_i = 1$ if the alarm was positive and 0 otherwise; $Y_i = 1$ if at least one earthquake occurred within the space–time domain of the ith alarm and 0 otherwise; p_i is the reference probability for the ith alarm, and the sum is performed over all alarms. The four summands in eq. 5 correspond to the reputation gain from hits, misses, false alarms and correct negatives, respectively. This simulation procedure also has a gambling analogy: Each iteration represents an additional gambler at the betting table. These gamblers wait until RTP bets, at which point the reference model tells them the house odds, and each gambler bets with or against RTP. After all the alarms have been scored, how do the earnings of RTP compare to those of the other gamblers?

Using this procedure, we simulated one million gambling score returns, using both the strict and loose interpretations. In Fig. 2, we

Table 5. Success probability, loose and strict interpretations of the outcome, and corresponding changes in reputation for each alarm. For those outcomes interpreted as hits, we also note the earthquake number (from Table 3) that the alarm has been interpreted to predict.

Alarm no., i	Success probability, p_i (%)	Outcome (loose)	Outcome (strict)	ΔR_i (loose)	ΔR_i (strict)
1	29.12	Hit[a] (Eqk. 1, 2)	–	2.43	–
2	2.27	Hit[b] (Eqk. 3)	–	43.07	–
3	8.96	False alarm	False alarm	−1.00	−1.00
4	2.74	False alarm	False alarm	−1.00	−1.00
5	17.58	Hit[c]	False alarm	4.69	−1.00
6	4.03	False alarm	False alarm	−1.00	−1.00
7	2.74	False alarm	False alarm	−1.00	−1.00
8	13.17	False alarm	False alarm	−1.00	−1.00
9	19.03	False alarm	False alarm	−1.00	−1.00
10	2.61	Hit[d] (Eqk. 6)	–	37.38	–
11	6.90	False alarm	False alarm	−1.00	−1.00
12	2.03	False alarm	False alarm	−1.00	−1.00
13	8.76	False alarm	False alarm	−1.00	−1.00
14	3.70	False alarm	False alarm	−1.00	−1.00
15	20.42	False alarm	False alarm	−1.00	−1.00
16	22.77	False alarm	False alarm	−1.00	−1.00
17	1.99	False alarm	False alarm	−1.00	−1.00
18	16.18	Hit (Eqk. 7, 8)	Hit (Eqk. 7, 8)	5.18	5.18
19	9.57	False alarm	False alarm	−1.00	−1.00
20	2.95	False alarm	False alarm	−1.00	−1.00
21	6.66	False alarm	False alarm	−1.00	−1.00
22	0.96	False alarm	False alarm	−1.00	−1.00
23	6.38	Hit (Eqk. 9)	Hit (Eqk. 9)	14.67	14.67
24	1.22	False alarm	False alarm	−1.00	−1.00
25	5.91	False alarm	False alarm	−1.00	−1.00
26	4.20	False alarm	False alarm	−1.00	−1.00
27	2.83	False alarm	False alarm	−1.00	−1.00
28	0.98	False alarm	False alarm	−1.00	−1.00
29	1.93	False alarm	False alarm	−1.00	−1.00
Total reputation change				84.42	−4.15

Notes: [a]The magnitude range was not specified in advance of the target earthquake; see Section 2 for details.
[b]The magnitude range was not specified in advance of the target earthquake; see Section 2 for details.
[c]Magnitude ambiguously specified in original alarm statement; see Section 2 for details.
[d]Due to delay of catalogue data, the alarm was declared after a satisfactory target earthquake; see Section 2 for details.

show the resulting empirical cumulative distribution functions for each set of simulated gambling scores. Under the loose interpretation, the null hypothesis is rejected with a high level of confidence: less than 0.009% of all simulations resulted in a reputation gain of more than 84.4 credits. But under the strict interpretation, the result is decidedly less compelling, and the null hypothesis cannot be rejected with much confidence: nearly 93.70% of the simulations had reputation gains larger than the 4.15 credits lost by RTP.

6 DISCUSSION AND CONCLUSIONS

Using a Poisson reference model and the gambling score method of Zhuang (2010), we evaluated the 29 prospective earthquake prediction alarms produced by RTP investigators. Because the details of several of the alarms and the corresponding outcomes are debatable, we considered a very formal, strict interpretation of the alarms using the exact statements of the RTP investigators. To be fair, we also considered a very loose interpretation that favoured RTP in any doubtful situation. When using the loose interpretation, we found that RTP was significantly better than a simple Poisson reference model, mostly due to the prediction involving the 2003 San Simeon earthquake. When using the strict interpretation, because the 2003 San Simeon earthquake and the related alarm were disregarded, the null hypothesis that RTP is no better than a Poisson model could not be rejected with much confidence; indeed, RTP fared marginally worse than the Poisson model.

At least one aspect of this study may be discomfiting to the reader: because RTP declared only positive alarms, it is not penalized for the occurrence of those large earthquakes inside RTP study regions and outside any alarm (Earthquakes 4, 5, 10–14 in Table 3)—outcomes that normally might be considered misses. Nevertheless, RTP did not issue negative alarms and therefore we treated these events as happening in regions of space–time–magnitude where no RTP bet was placed; we contend that this interpretation is the most judicious.

Alternatively, what if one treated any lack of an alarm as a negative prediction, rather than as a statement of 'no available prediction'? Because this is a dubious interpretation—if not an outright misinterpretation—of a predictive non-statement, any evaluation that resulted from this approach would be far from defensible. Beyond this important semantic issue, several technical challenges would remain: How would one estimate the reference probability for a missed earthquake (as opposed to the reference probability for a well-defined space–time–magnitude volume, as we used in this study)? Could one fairly and reasonably define an appropriate volume *after* a large earthquake happened? More generally, how

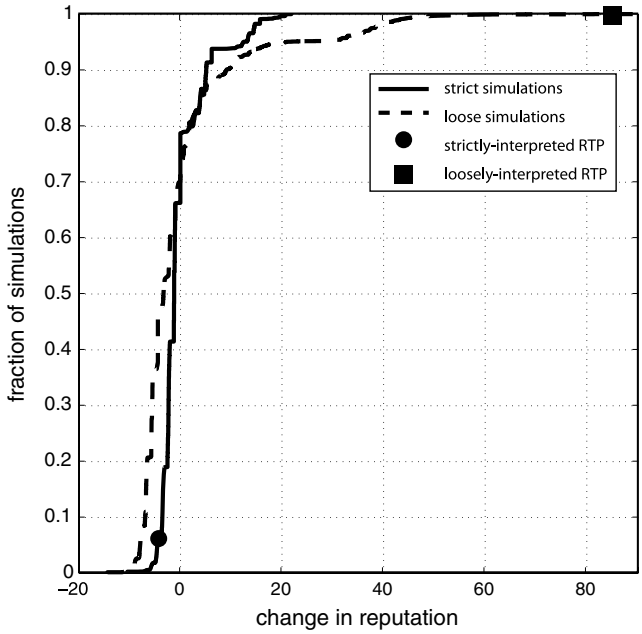

Figure 2. Empirical cumulative distribution functions for simulated gambling returns under the strict interpretation (solid line) and loose interpretation (dashed line). RTP reputation changes are indicated by a circle (strict interpretation) and a square (loose interpretation). Because almost no simulations under the loose interpretation yield a return that is greater than RTP's gain, the null hypothesis is rejected with great confidence—one can say that the RTP alarms are superior to the reference model. On the other hand, under the strict interpretation, because a substantial percentage of simulations have a larger reputation increase than RTP, the null hypothesis cannot be rejected with great confidence.

would the volume of each inferred negative alarm be defined, so as to also count correct negatives? If one discretized the total available space–time–magnitude volume in any or all dimensions, the choice of discretization size would affect the scoring results. Because there are so many pitfalls with the alternatives, we argue that the interpretations used and the evaluation presented in this study involve the fewest assumptions and yield the most robust results, despite effectively ignoring some of the large earthquakes listed in Table 3.

As mentioned in the Introduction, one of the RTP parameters is the radius, r, of circles centred on each event in the chain; this parameter is used to define the spatial extent of each alarm. In this study, we only considered the alarms defined by the smaller r-value for each alarm, consistent with the value specified for the early RTP alarms. Later in the experiment, RTP investigators introduced alarms based on a larger r; they announced these alarms simultaneously with the smaller alarms (P. Shebalin private communication, 2009); we chose to test only the smaller alarms. Nevertheless, for the alarms analysed in this study, the larger alarm variants would fare no better than their smaller counterparts. In particular, the reference probabilities would be at least as large as those reported in Table 4, and no target earthquake occurred in the larger alarms that did not fall within the smaller alarms; therefore, the total changes in reputation necessarily would be no greater than those reported in Table 5. Along these same lines, we emphasize that we tested the RTP alarm statements, not the entire algorithm as it has been described in various publications. In particular, we did not decluster the catalogues when determining target earthquakes and we did not disregard deep events.

There are some general lessons to be learned from the contentious alarms and earthquakes discussed in Section 2. When making earthquake forecast statements, it is critically important that every prediction is well defined and unambiguously falsifiable. In the context of RTP, the fact that the San Simeon alarm (Alarm 2) is controversial obscures the distinction between a result that demonstrates compelling predictive skill (under the loose interpretation) and one that does not (under the strict interpretation). As earthquake modellers increasingly rely on computer codes to analyse data and to produce forecasts, an open-source, hands-off approach is beneficial to the forecaster's credibility. If the RTP codes were fully automated, or at least made available and documented well enough to allow reproducibility, the integrity of the forecasts would increase.

Testing the RTP alarms is not straightforward: owing to the irregularity with which they are announced and because there are no explicit negative alarms, many commonly used evaluation techniques are not informative. More generally, we are unaware of any other models making predictions in the same format, manner and study regions that RTP does, and this precludes comparisons with other earthquake prediction algorithms. The simultaneous, multiple-investigator experiments occurring within the Collaboratory for the Study of Earthquake Predictability (CSEP) testing centres (Jordan 2006; Zechar *et al.* 2009) offer a unique opportunity to advance earthquake forecasting research, and including the RTP algorithm in CSEP experiments might be enlightening. The investigators of RTP have attempted to share and archive their prospective earthquake alarms. Nevertheless, it remains difficult to reproduce the algorithm by following only the publications that describe the RTP approach and the outcomes of its alarms. It is also difficult to judge if the RTP algorithm or any of the myriad model parameter values have changed, or are changing, as the experiment progresses—such changes would make RTP a moving target for evaluation. Moreover, in this study we applied a testing procedure that was designed after the experiment began, and certainly it is not the only evaluation possible.

CSEP is designed to address these specific problems: testing centres curate and execute model codes in an automated, reproducible environment where the testing metrics are defined in advance (e.g. Schorlemmer *et al*. 2007; Zechar *et al.* 2010); and during prospective prediction experiments, codes and model parameter values cannot be changed (Schorlemmer & Gerstenberger 2007; Zechar *et al*. 2009). Integrating RTP into a CSEP testing centre would reduce controversy and ambiguity related to alarm declaration and evaluation. The chains, alarms and the results of any test would be formally reproducible. This would obviate the need for multiple analyses of the same set of alarms and further clarify whether RTP demonstrates genuine predictive skill.

ACKNOWLEDGEMENTS

The authors thank Peter Shebalin for his cooperation and lively exchanges. The authors thank Christine Smyth and Patricia S. Elzie for carefully reading, and thoughtfully responding to, an early draft of the paper. The authors thank two anonymous reviewers for their constructive comments.

REFERENCES

Dziewonski, A.M., Chou, T.-A. & Woodhouse, J.H., 1981. Determination of earthquake source parameters from waveform data for studies of global and regional seismicity, *J. geophys. Res.,* **86,** 2825–2852.

Ekström, G., Dziewonski, A.M., Maternovskaya, N.N. & Nettles, M., 2005. Global seismicity of 2003: centroid-moment-tensor solutions for 1087 earthquakes, *Phys. Earth planet. Inter.,* **148,** 327–351.

Gabrielov, A., Newman, W.I. & Turcotte, D.L., 1999. An exactly soluble hierarchical clustering model: inverse cascades, self-similarity and scaling, *Phys. Rev. E,* **60,** 5293–5300.

Gabrielov, A., Keilis-Borok, V., Zaliapin, I. & Newman, W.I., 2000. Critical transitions in colliding cascades, *Phys. Rev. E,* **62,** 237–249.

Gardner, J.K. & Knopoff, L., 1974. Is the sequence of earthquakes in southern California, with aftershocks removed, Poissonian?, *Bull. seism. Soc. Am.,* **64**(5), 1363–1367.

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

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

Kagan, Y.Y., 2010. Statistical distributions of earthquake numbers: consequence of branching process, *Geophys. J. Int.,* **180,** 1313–1328.

Keilis-Borok, V.I., 2002. Earthquake prediction, state-of-the-art and emerging possibilities, *Annu. Rev. Earth planet. Sci.,* **30,** 1–33.

Keilis-Borok, V.I., 2003. Fundamentals of earthquake prediction: four paradigms, in *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction,* pp. 1–36, eds Keilis-Borok, V.I. & Soloviev, A.A., Springer-Verlag, Berlin.

Keilis-Borok, V.I., Shebalin, P.N., Gabrielov, A. & Turcotte, D.L., 2004. Reverse tracing of short-term earthquake precursors, *Phys. Earth planet. Inter.*, **145**, 75–85.

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

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

Schorlemmer, D. & Gerstenberger, M.C., 2007. RELM testing centre, *Seism. Res. Lett.,* **78,** 30–36.

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

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

Shebalin, P., 2006. Increased correlation range of seismicity before large events manifested by earthquake chains, *Tectonophysics,* **424,** 335–349.

Shebalin, P.N., Zaliapin, I. & Keilis-Borok, V.I., 2000. Premonitory rise of the earthquakes' correlation range: Lesser Antilles, *Phys. Earth Planet. Inter.,* **122,** 241–249.

Shebalin, P.N., Keilis-Borok, V.I., Zaliapin, I., Uyeda, S., Nagao, T. & Tsybin, N., 2003. *Short-Term Premonitory Rise of the Earthquake Correlation Range*. IUGG Abstracts, Sapporo, Japan.

Shebalin, P.N., Keilis-Borok, V.I., Zaliapin, I., Uyeda, S., Nagao, T. & Tsybin, N., 2004. Advance short-term prediction of the large Tokachi-oki earthquake, September 25, 2003, M = 8.1: A case history, *Earth planet. Space,* **56,** 715–724.

Shebalin, P., Keilis-Borok, V.I., Gabrielov, A., Zaliapin, I. & Turcotte, D., 2006. Short-term earthquake prediction by reverse analysis of lithosphere dynamics, *Tectonophysics,* **413,** 63–75.

Shi, Y., Liu, J. & Zhang, G., 2001. An evaluation of Chinese earthquake prediction, 1990–1998. *J. Appl. Probab.,* **38,** 222–231.

Utsu, T., Ogata, Y. & Matsu'ura, R.S., 1995. The centenary of the Omori formula for a decay law of aftershock activity, *J. Phys. Earth,* **43,** 1–33.

Zaliapin, I., Keilis-Borok, V.I. & Axen, G., 2002. Premonitory spreading of seismicity over the faults' network in southern California: precursor accord, *J. geophys. Res.,* **107,** 2221, doi: 10.1029/2000JB000034.

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

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

Zechar, J.D., Gerstenberger, M.C., & Rhoades, D.A., 2010. Likelihood-based tests for evaluating space-rate-magnitude earthquake forecasts, *Bull. seism. Soc. Am.,* **100,** doi: 10.1785/0120090192.

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

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article:

Supplement. We list the epicentral latitudes and longitudes for events that define the spatial extent of each alarm.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.