Geophys. J. Int. (2008) **172,** 715–724 doi: 10.1111/j.1365-246X.2007.03676.x

Testing alarm-based earthquake predictions

J. Douglas Zechar and Thomas H. Jordan

Department of Earth Sciences, University of Southern California, Los Angeles, CA 90089*, USA. E-mail: zechar@usc.edu*

Accepted 2007 October 26. Received 2007 September 24; in original form 2007 June 29

SUMMARY

Motivated by a recent resurgence in earthquake predictability research, we present a method for testing alarm-based earthquake predictions. The testing method is based on the Molchan diagram—a plot of miss rate and fraction of space–time occupied by alarm—and is applicable to a wide class of predictions, including probabilistic earthquake forecasts varying in space, time, and magnitude. A single alarm can be simply tested using the cumulative binomial distribution. Here we consider the more interesting case of a function from which a continuum of well-ordered alarms can be derived. For such an 'alarm function' we construct a cumulative performance measure, the area skill score, based on the normalized area above a trajectory on the Molchan diagram. A score of unity indicates perfect skill, a score of zero indicates perfect non-skill, and the expected score for a random alarm function is $^1/_2$. The area skill score quantifies the performance of an arbitrary alarm function relative to a reference model. To illustrate the testing method, we consider the 10-yr experiment by J. Rundle and others to predict $M \geq 5$ earthquakes in California. We test forecasts from three models: relative intensity (RI), a simple spatial clustering model constructed using only smoothed historical seismicity; pattern informatics (PI), a model that aims to capture seismicity dynamics by pattern recognition; and the U. S. Geological Survey National Seismic Hazard Map (NSHM), a model that comprises smoothed historical seismicity, zones of 'background' seismicity, and explicit fault information. Results show that neither PI nor NSHM provide significant performance gain relative to the RI reference model. We suggest that our testing method can be used to evaluate future experiments in the Collaboratory for the Study of Earthquake Predictability and to iteratively improve reference models for earthquake prediction hypothesis testing.

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

INTRODUCTION

Despite the notable lack of success in reliably predicting destructive earthquakes, there has been a resurgence of research on earthquake predictability motivated by better monitoring networks and data on past events, new knowledge of the physics of earthquake ruptures, and a more comprehensive understanding of stress evolution and transfer. However, the study of earthquake predictability has been hampered by the lack of an adequate infrastructure for conducting prospective prediction experiments under rigorous, controlled conditions and evaluating them using accepted criteria specified in advance. To address this problem, the working group on Regional Earthquake Likelihood Models (RELM), supported by the Southern California Earthquake Center (SCEC) and U. S. Geological Survey (USGS), has recently established a facility for prospective testing of scientific earthquake predictions in California, and a number of experiments are now underway (Field 2007, and references therein).

The RELM project conforms to the requirements for well-posed prediction experiments (e.g. Rhoades & Evison 1989; Jackson 1996) through a strict set of registration and testing standards. For a 5 yr experiment, models are constructed to predict earthquakes in California above magnitude 4.95 during 2006–2010 by specifying time-invariant earthquake rates in prescribed latitude–longitude–magnitude bins. Three tests based on likelihood measures will be used to evaluate the forecasts (Schorlemmer *et al*. 2007).

The interest in the RELM project shown by earthquake scientists has motivated an international partnership to develop a Collaboratory for the Study of Earthquake Predictability (CSEP). CSEP is being designed to support a global program of research on earthquake predictability (Jordan 2006), and one of its goals is to extend the testing methodology to include alarm-based predictions. In this paper, we outline such a methodology and apply it to the retrospective testing of three prediction models for California.

ALARM-BASED PREDICTION

Earthquake alarms are a natural construct when we consider the problem of predicting the locations and origin times of earthquakes above some minimum magnitude—target earthquakes. A common

approach to this problem is to search for precursory signals that indicate an impending target earthquake in a given space–time window (Keilis-Borok 2002, 2003 and references therein; Kossobokov & Shebalin 2003). These signals can be represented by precursory functions, the values of which are computed and analysed in moving time windows. For example, at the present time t we consider a region R where we wish to predict target earthquakes using precursory function f, which is based on information available up to time t. If $f(t)$ exceeds some threshold value (typically optimized by retrospective testing), an alarm is declared, indicating that one or more target earthquakes are expected in R during the period $(t, t + \Delta t)$, a 'time of increased probability' (TIP) (Keilis-Borok & Kossobokov 1990). In practice, pattern recognition algorithms often combine several precursory functions. For example, the Reverse Tracing of Precursors (RTP) algorithm employs eight intermediate-term precursory patterns and yields alarms with fixed duration but highly variable spatial extent (Keilis-Borok *et al.* 2004; Shebalin *et al.* 2006), which makes testing difficult.

To place alarm-based testing in the RELM context, we consider spatially varying but time-invariant prediction models. At the beginning of the experiment, we assume there exists some (unknown) probability P_k that the next target earthquake in the RELM testing region R will occur in r_k, the kth subregion of R. We further assume that P_k is identical for every target earthquake in the testing interval; that is, the conditional probability that the nth earthquake locates in r_k after $n - 1$ earthquakes have already occurred also equals P_k. We suppose that, prior to the experiment, some reference model of this time-invariant distribution, $\tilde{P}_k$, is available. For example, the prior distribution might be the next-event probability calculated from the smoothed, average historical rate of earthquake occurrence in r_k (Kagan & Jackson 2000; Kafka 2002; Kossobokov 2004; Rhoades & Evison 2004; Helmstetter *et al.* 2007). In this paper, we will follow Tiampo *et al*. (2002) by calling this the relative intensity (RI) forecasting strategy. By definition, the summation of P_k or $\tilde{P}_k$ over all subregions in R is unity.

An alarm-based prediction uses fresh information or insights to identify a 'domain of increased probability,' $A \subseteq R$, the alarm region, where the true probability is hypothesized to exceed the reference value:

$$P_A \equiv \sum_{r_k \in A} P_k > \sum_{r_k \in A} \tilde{P}_k \equiv \tilde{P}_A. \tag{1}$$

At the end of the testing interval, we observe that N target earthquakes have occurred and that h of these are located in the alarm region A. Under the null hypothesis $H_0 : P_A = \tilde{P}_A$, the probability of h 'hits' in A follows a binomial distribution,

$$B(h|N, \tilde{P}_A) = \binom{N}{h} (\tilde{P}_A)^h (1 - \tilde{P}_A)^{N-h}. \tag{2}$$

H_0 can be rejected in favour of $H_1 : P_A > \tilde{P}_A$ if

$$\sum_{n=h}^{N} B(n|N, \tilde{P}_A) \leq \alpha, \tag{3}$$

for some critical significance level α; that is, if the probability of obtaining h or more hits by chance is less than or equal to α. Rejection of H_0 in favour of H_1 at a high confidence level ($\alpha \ll 1$) is evidence that the alarm-based prediction has significant skill relative to a prediction based on the reference model $\tilde{P}_k$.

Following Molchan (1990, 1991) and Molchan & Kagan (1992), we consider how the miss rate, $\nu = (N - h)/N$, varies with the probability-weighted area of alarm region A, $\tau = \tilde{P}_A$. Plots of (τ, ν) where $\tau, \nu \in [0, 1]$ are called Molchan diagrams ('error diagrams' in Molchan's terminology). At the end of the testing period, the total number of target events, N, is known, so the value of ν is restricted to the discrete set $\{n/N : n = 0, 1, \ldots, N\}$. The boundary conditions are fixed: if no alarm is declared ($\tau = 0$; the optimist's strategy), all events are missed ($\nu = 1$), whereas if an alarm is declared over the entire testing region R ($\tau = 1$; the pessimist's strategy), no events are missed ($\nu = 0$).

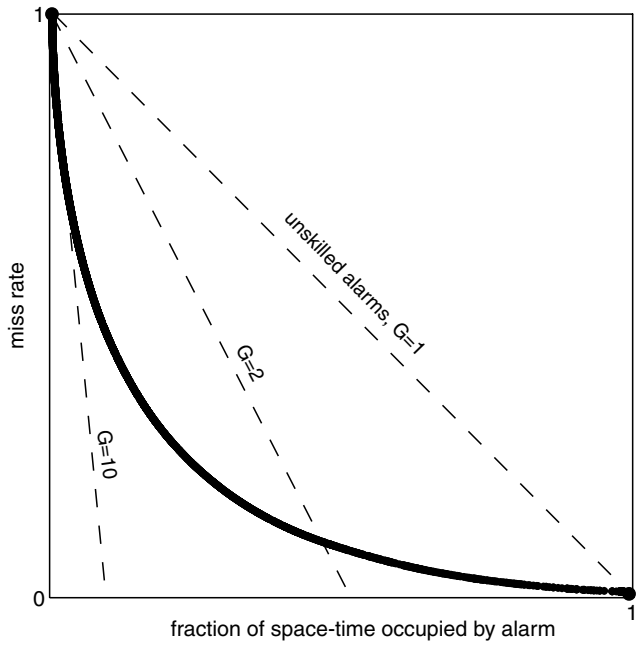

Figure 1. Molchan diagram—a plot of miss rate versus fraction of space–time occupied by alarm—with dashed isolines of probability gain. The descending diagonal, corresponding to unit probability gain, represents the expected performance when $P_A = \tilde{P}_A$. The dark line represents a hypothetical optimal trajectory which depends on the unknown 'true' distribution.

Under H_0, the distribution of ν is given by (2), and its expected value, $\langle \nu \rangle$, lies on the descending diagonal of the Molchan diagram, $\langle \nu \rangle = 1 - \tau$ (Fig. 1). More generally, $1 - \langle \nu \rangle$ measures the long-run probability of a subregion being in the alarm region conditional on it containing an event, $P(A|E)$, while τ is the prior probability of the alarm region, $P(A) = \tilde{P}_A$. The Bayes identity requires

$$P(E|A) = \left[\frac{P(A|E)}{P(A)} \right] P(E), \tag{4}$$

where the quantity in brackets is called the probability gain (Aki 1981; Molchan 1991; McGuire *et al.* 2005):

$$G \equiv \frac{P(E|A)}{P(E)} = \frac{1 - \langle \nu \rangle}{\tau}. \tag{5}$$

On the Molchan diagram, the sample value of G is the slope of the line connecting (0,1) to (τ, ν) and (3) provides a test of the null hypothesis $H_0 : G = 1$ against the alternative $H_1 : G > 1$.

For the grid-based RELM models, the values of τ are also discrete, given by summations over the cell values $\tilde{P}_k$. In the continuum limit where the cell size shrinks to zero, $\tilde{P}_k$ becomes a probability density function (p.d.f.) $\tilde{p}(\mathbf{x})$ of an event at a geographic location $\mathbf{x} \in R$. The analysis is also simplified by representing target earthquakes as discrete points at their geographic epicentres $\{\mathbf{x}_n : n = 1, 2, \ldots, N\}$. In this point-process limit, τ is a continuous variable on [0,1], and all realizable values of ν are stepwise constant functions of τ. Moreover, as cell size shrinks to zero, τ becomes the measure of false positives (false alarms) and $(1 - \tau)$ becomes the measure of

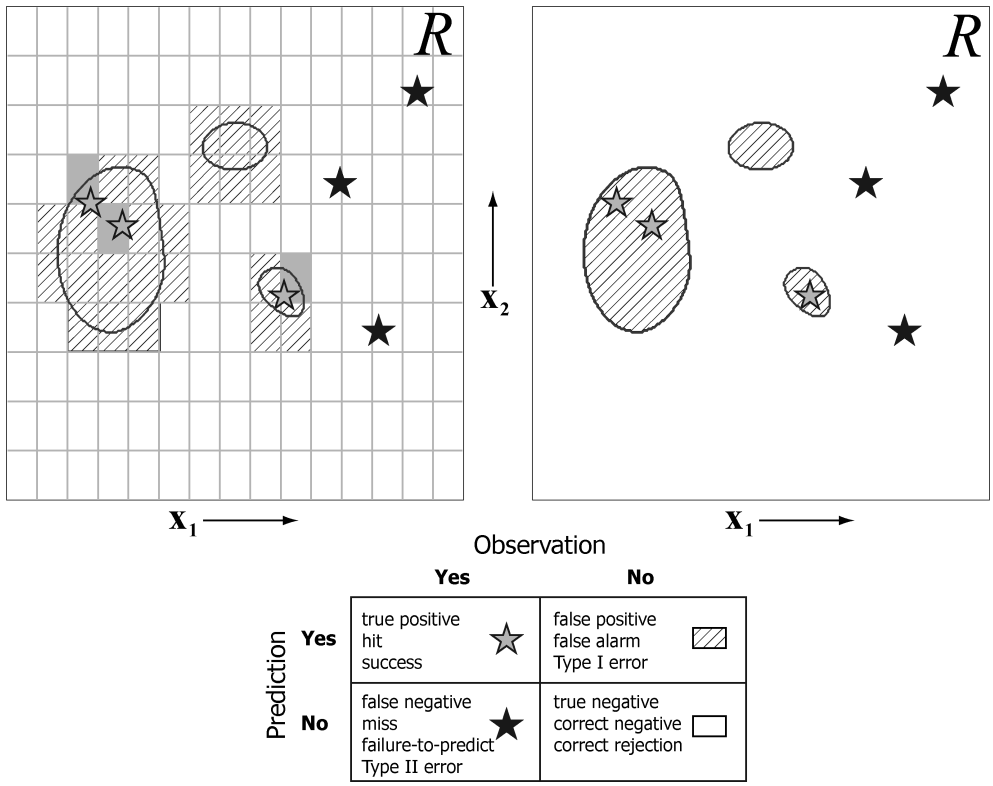

Figure 2. Example alarms in the case of coarse spatial discretization (left-hand panel) and in the continuum limit, where cell size approaches zero (right-hand panel). A full contingency table (with alternate names for each contingency) provides a legend. In the left-hand panel, the shaded boxes are hit regions, which are not explicitly represented in the contingency table because hits and misses describe events rather than cells. In the right-hand panel, the hit regions are infinitesimally small, so all alarm regions become false alarms.

correct negatives. The Molchan diagram then describes the complete contingency table (Fig. 2) and is equivalent to the receiver operating characteristic (ROC) diagram, a plot of hit rate versus false alarm rate that has been employed in weather forecast verification, medical testing and signal analysis (Mason 2003 and references therein). In the continuum limit, we can contour constant values of α on the Molchan diagram by finding the minimum value of τ that solves the equality in (3) for each discrete value of ν. These stepwise confidence intervals are illustrated in Fig. 3.

OPTIMAL MOLCHAN TRAJECTORIES

Minimizing $\langle\nu\rangle$ for a fixed value of τ yields an optimal alarm region A^*. We consider the special case where two conditions apply:

1. the prior distribution is uniform over R; that is, the values of the p.d.f. $\tilde{p}(\mathbf{x})$ are everywhere equal, and τ measures the normalized geographic area covered by an alarm; and
2. the true distribution has no flat spots; that is, in general, the contours $\{\mathbf{x}(\lambda) : p(\mathbf{x}) = \lambda\}$ are sets of measure zero (lines, points, or empty) for all contour levels λ.

The optimization problem is then solved by the 'water-level principle', which states that a region on a map above a contour level λ has the highest average elevation of any region with the same area (Fig. 4). In the case we consider here, the optimal alarm is the domain of R where the topography represented by $p(\mathbf{x})$ rises above a water level λ; this alarm can be expressed as $A^*(\lambda) = \{\mathbf{x} \in R : H(p(\mathbf{x}) - \lambda) = 1\}$, where H is the Heaviside step function. Dropping the water level from the maximum value of $p(\mathbf{x})$ to zero traces out an optimal trajectory,

$$\tau^*(\lambda) = \int_R H[p(\mathbf{x}) - \lambda]\,\mathrm{d}\mathbf{x}, \tag{6a}$$

$$\nu^*(\lambda) = \int_R [1 - p(\mathbf{x})]\,H[p(\mathbf{x}) - \lambda]\mathrm{d}\mathbf{x}. \tag{6b}$$

The optimal trajectory lies on or below the descending diagonal of the Molchan diagram (Fig. 1).

We can relax condition (a) by considering an arbitrary prior p.d.f. satisfying $\tilde{p}(\mathbf{x}) > 0$ for all $\mathbf{x}$ in R. In this case, the optimal alarm is given by

$$A^*(\lambda) = \{\mathbf{x} \in R : H[g^*(\mathbf{x}) - \lambda] = 1\}, \tag{7}$$

where $g^*(\mathbf{x}) = \frac{p(\mathbf{x})}{\tilde{p}(\mathbf{x})}$ is the optimal local probability gain at $\mathbf{x}$. The optimal trajectory becomes

$$\tau^*(\lambda) = \int_R \tilde{p}(\mathbf{x})\; H[g^*(\mathbf{x}) - \lambda]\,\mathrm{d}\mathbf{x} = \tilde{P}_{A^*}, \tag{8a}$$

$$\nu^*(\lambda) = \int_R [1 - p(\mathbf{x})]\,H(g^*(\mathbf{x}) - \lambda)\mathrm{d}\mathbf{x} = 1 - P_{A^*}. \tag{8b}$$

We note that the probability gain of the optimal trajectory can be written as the weighted average of the local gain over the optimal alarm region,

$$G^*(\lambda) = \frac{\int_{A^*(\lambda)} g^*(\mathbf{x})\;\tilde{p}(\mathbf{x})\;\mathrm{d}\mathbf{x}}{\int_{A^*(\lambda)} \tilde{p}(\mathbf{x})\;\mathrm{d}\mathbf{x}}. \tag{9}$$

We can relax condition (b) by considering an arbitrary optimal local gain function $g^*(\mathbf{x})$ that may contain flat spots. We consider a flat

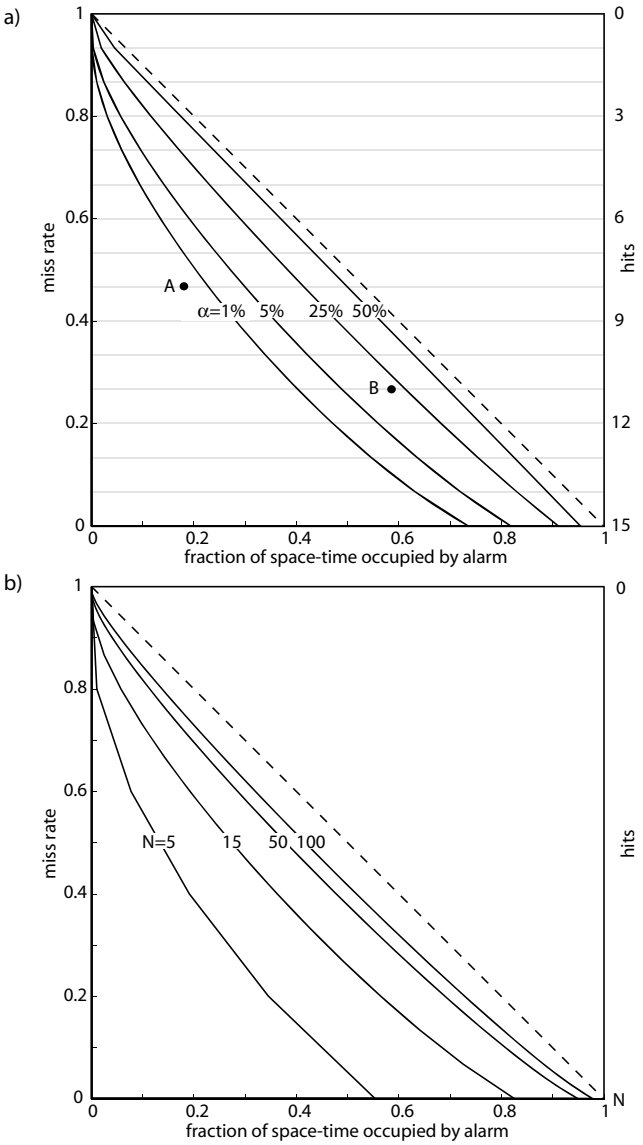

Figure 3. Molchan diagram confidence bounds computed by solving Equation 3 for (a) fixed N and varying α and (b) fixed α and varying N. In (a), $N = 15$ and the curves are contours for $\alpha = \{1, 5, 25, 50$ per cent$\}$. Here, point A represents an alarm region that has obtained 8 hits and indicates that the null hypothesis $H_0 : P_A = \tilde{P}_A$ can be rejected at a confidence level greater than 99 per cent while the point B (11 hits) supports rejection at just above 75 per cent confidence. In (b), $\alpha = 5$ per cent and the curves are contours for $N = \{5, 15, 50, 100\}$. As N increases, the contours approach the descending diagonal.

spot domain $D \subset R$ where $p(\mathbf{x}) = \lambda_D\, \tilde{p}(\mathbf{x})$ for all $\mathbf{x}$ in D and let

$$A^{*\pm}(\lambda_D) = \lim_{\varepsilon \to 0} \{\mathbf{x} \in R : H(g^*(\mathbf{x}) - \lambda_D \pm \varepsilon) = 1\}, \tag{10}$$

such that $A^{*+} = A^{*-} \cup D$. Then, the optimal trajectory 'jumps' from (τ^{*-}, ν^{*-}) to (τ^{*+}, ν^{*+}) at λ_D. Sampling any two subsets of the same size from D yields the same Molchan trajectory point, and therefore, relaxing condition (b) can lead to non-unique optimal alarms. We note, however, that such a sampling can only yield points on the line connecting (τ^{*-}, ν^{*-}) to (τ^{*+}, ν^{*+}), and therefore, the Molchan trajectory remains unique and no alarm region can achieve a lower value of $\langle \nu \rangle$.

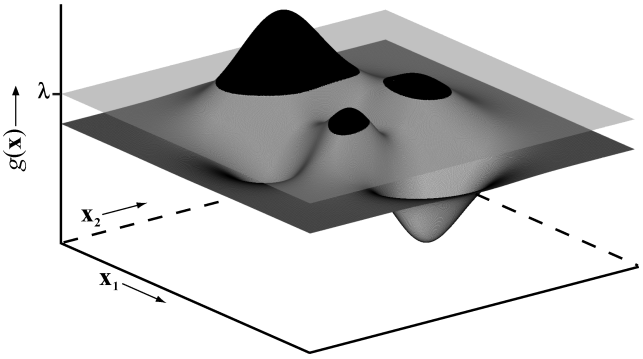

Figure 4. Illustration of alarm optimization technique via water-level threshold procedure. Here, the curved surface on R represents a gain function g ($\mathbf{x}$); the plane intersects the surface at a height of λ. The resulting region above this threshold is the optimal alarm region A^*, or the region of this fixed area with the highest average elevation. A map view of this alarm region is shown in Fig. 2.

In statistical hypothesis testing, the power of a test is the probability that a false null hypothesis is rejected—in other words, that a Type II error is not committed—and is equal to $1 - \beta$ where β is the Type II error rate (Lehman & Romano 2005). In our problem, where the Type I error rate is measured by τ and the Type II error rate by $\nu(\tau)$, an appropriate measure of the power of an alarm is $1 - \nu(\tau)$. In these terms, $A^*(\lambda)$ is the most powerful alarm of size $\tau(\lambda)$. As the reference model approaches the true distribution, $\tilde{P}_{A^*} \to P_{A^*}$, the power of the optimal alarm approaches the average power of a random alarm, $1 - \nu^*(\tau) \to \tau(\lambda)$, and a larger number of events N is needed to discriminate H_1 from H_0.

When $\tilde{p}(\mathbf{x}) = p(\mathbf{x})$ for all $\mathbf{x}$ in R, $g^*(\mathbf{x})$ is flat—and in particular, equal to unity—throughout R, and the optimal trajectory coincides with the descending diagonal of the Molchan diagram. In this case, no alarm-based strategy can reject H_0, and the time-invariant prediction problem for the simple RELM set-up is solved.

ALARM FUNCTIONS AND THE AREA SKILL SCORE

We consider alarm sets $\{A(\lambda) : \lambda \geq 0\}$ that are ordered by $\tau(\lambda)$ such that

$$\tau(A(\lambda)) < \tau(A(\lambda')) \Rightarrow A(\lambda) \subset A(\lambda'); \tag{11}$$

and complete on $0 \leq \tau \leq 1$; that is, sufficient to generate complete Molchan trajectories $\{\nu(\tau) : \tau \in [0, 1]\}$. A complete, ordered alarm set can be represented as an unscaled contour map on R (Fig. 5). Such a set can be generated from a continuous, positive-semi idefinite 'alarm function' $g(\mathbf{x})$ by water-level contouring,

$$A_g(\lambda) = \{\mathbf{x} \in R : H(g(\mathbf{x}) - \lambda) = 1\}. \tag{12}$$

An example of an alarm function is the optimal gain function, $g^*(\mathbf{x})$. In fact, any p.d.f. constitutes an alarm function; not all alarm functions, however, specify a p.d.f.

Alarm functions that have Molchan trajectories with the same values as $g(\mathbf{x})$ form an equivalence class indexed by g:

$$C_g = \{f(\mathbf{x}) : \nu_f(\tau) = \nu_g(\tau), \forall\, \tau \in [0, 1]\}. \tag{13}$$

An infinite number of alarm functions yield the same alarms as $g(\mathbf{x})$. For example, consider any order-preserving functional; that is,

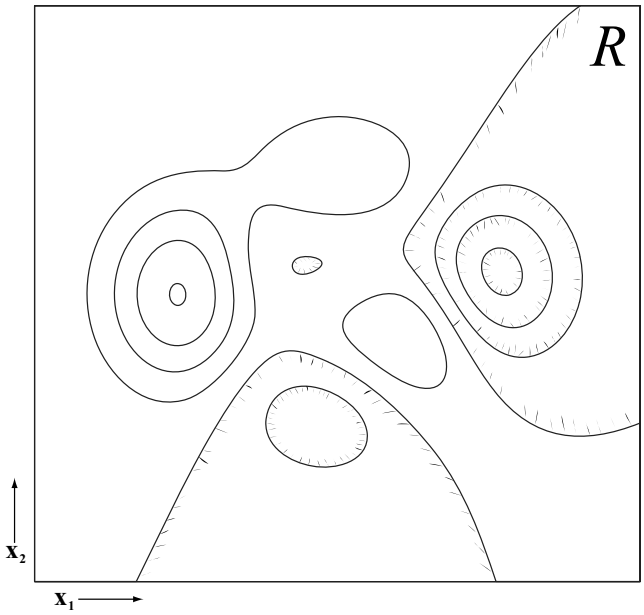

Figure 5. Map view of an example alarm function; here, the contour map corresponds to the spatial alarm function shown in Fig. 4. Hashed contours indicate descending values.

one that satisfies $h(z) < h(z') \Leftrightarrow z < z'$. Then $f(\mathbf{x}) = h(g(\mathbf{x}))$ is also in the equivalence class C_g. Such a functional only rescales the contour map, so that $A_f = A_g$, and the trajectory $\nu_f(\tau)$ is identical to $\nu_g(\tau)$.

All members of a given equivalence class C_g provide equal probability gain relative to the prior $\tilde{p}(\mathbf{x})$. The interesting extremes are C_1, comprising functions equivalent to $g(\mathbf{x}) = 1$ and C^*, comprising functions equivalent to g^*. Alarm functions belonging to C_1 yield trajectories with expected values lying on the descending diagonal and thus provide no gain in the long run, whereas alarm functions belonging to C^* yield the optimal trajectory. For any alarm function $f(\mathbf{x})$, we seek to test the null hypothesis $H_0 : f \in C_1$, against the alternative that f performs 'better' than alarm functions belonging to C_1. (An alarm function with an expected trajectory above the descending diagonal is 'worse' than C_1; however, as noted by Molchan & Kagan (1992), we can use it to generate a set of 'anti-alarms' whose complements in R have probability gains greater than unity. We therefore consider only one-sided tests.)

The performance of an alarm function $f(\mathbf{x})$ can be measured by the area above its Molchan trajectory evaluated at a given τ, a statistic we call the 'area skill score':

$$a_f(\tau) = \frac{1}{\tau} \int_0^\tau [1 - \nu_f(t)]\, \mathrm{d}t. \tag{14}$$

This statistic is normalized such that its value is between 0 and 1 and under H_0 its expectation is

$$\langle a_f(\tau) \rangle = \frac{\tau}{2}. \tag{15}$$

We can use this statistic to assess the skill of $f(\mathbf{x})$ relative to $\tilde{p}(\mathbf{x})$ by testing the null hypothesis $H_0 : a_f(\tau) = \tau/2$ against the alternative H_1: $a_f(\tau) > \tau/2$. In the limit of infinitesimal discretization in τ, the area skill score is equivalent to the area under curve (AUC) measure used in ROC analyses (Mason 2003).

In order to use the area skill score for hypothesis testing, we have explored the score distribution of unskilled alarm functions with an arbitrary prior (Zechar & Jordan in prep). We have an analytic approach for generating moments of the distribution and find that this distribution is related to the distribution of cross-sectional wedge 'area' of an N-dimensional hypercube along its principal diagonal. An application of the Central Limit Theorem shows that, in the case of continuous alarm functions, the area skill score distribution at $\tau = 1$ is asymptotically Gaussian with a mean of $^1/_2$ and a variance of $1/(12N)$. Furthermore, the distribution's kurtosis excess—a factor dependent on the second and fourth central moments and an indicator of deviation from the Gaussian distribution—is equal to $-5/(4N)$. For N on the order of a dozen or more, the Gaussian approximation provides an excellent estimate of confidence bounds.

We can estimate the area skill score distribution by simulation for any number of target events N at any value of τ. It can be shown that the power of the area skill score, while dependent on the prior, tends to increase with increasing τ, and therefore it is best to use $a_f(\tau = 1)$ for the hypothesis test. In the illustrative experiment described below, we consider discrete alarm functions and the observed seismicity yields multiple earthquakes in a single forecast cell. In this case, the Molchan trajectory and area skill score confidence bounds are most easily estimated by simulation of unskilled alarm functions (i.e. those belonging to C_1).

MODELS AND DATA

To demonstrate the area skill score testing procedure, we consider three models of spatial predictability—RI, pattern informatics (PI), and the United States Geological Survey National Seismic Hazard Map (NSHM)—in a quasi-prospective prediction experiment. For visual comparison, the alarm function values for each model are shown in Figs 6–8. These models make for an interesting set of examples because they represent distinct hypotheses about the spatial distribution of earthquakes.

RI suggests that future earthquakes are most likely to occur where historical seismicity rates are highest (Tiampo *et al*. 2002). RI uses a particularly simple measure of seismicity—the rate of past earthquakes occurring in each spatial cell—and belongs to a general class of smoothed seismicity models. Proximity to Past Earthquakes (Rhoades & Evison 2004), Cellular Seismology (Kafka 2002) and others (e.g. Kagan & Jackson 2000; Kossobokov 2004; Helmstetter *et al.* 2007) are members of this class that offer slightly different representations of the same basic hypothesis; each has been recommended as a reference model.

PI suggests that the locations of future earthquakes are indicated by anomalous changes in seismic activity. Regions undergoing seismic activation or seismic quiescence are found by computing the short-term seismicity rate in a given spatial cell—say, for the previous 10 yr—and comparing with the long-term seismicity rate in this cell—say, for the previous 50 yr. If the short-term rate is anomalously low/high, the cell is considered to be undergoing seismic quiescence/activation in preparation for a target earthquake in the near future (Tiampo *et al.* 2002). As shown in Figs 6 and 7, the PI and RI alarm functions are highly correlated—in particular, many regions with a high PI index also have a high RI index.

NSHM suggests that future earthquakes will occur where past earthquakes have occurred, with the qualification that moderate to large earthquakes are likely to occur near mapped faults and some earthquakes will be surprises. Therefore, the NSHM earthquake rate model combines smoothed historical seismicity, fault information, and 'background' zones where a spatially uniform seismicity rate is assumed. According to Frankel *et al*. (1996, 2002), this combination represents the best knowledge of faults and spatial distribution of

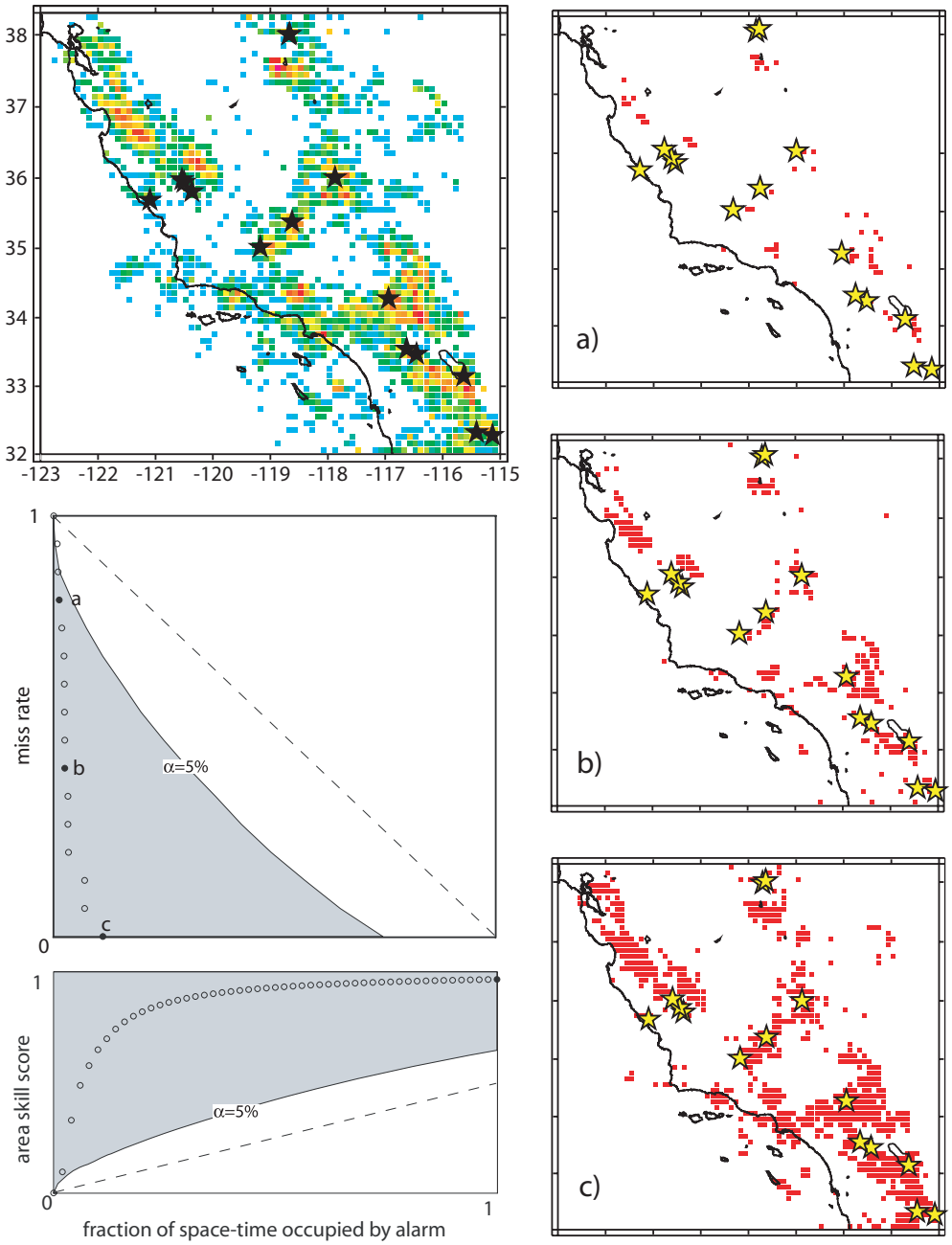

Figure 6. Illustration of testing Relative Intensity (RI) alarm function relative to a uniform spatial prior. Top left frame shows map of RI alarm function values and observed target earthquakes (stars), created with Generic Mapping Tools software (Wessel & Smith 1998). Panels (a), (b) and (c) show alarm regions for three decreasing threshold values. The corresponding Molchan diagram points are labelled in the plot on the left-hand side of the second row. The left-hand panel in the third row shows the corresponding area skill score trajectory. The shaded areas on the plots are the $\alpha = 5$ per cent critical region.

earthquakes. The NSHM alarm function values in Fig. 8 reveal a forecast that is much smoother than that of RI and PI.

Rundle *et al.* (2003) issued 10-yr 'hotspot' maps based on RI and PI . The hotspots are alarms that last the duration of the experiment and are derived from underlying alarm functions. The alarm function of the PI model is not a p.d.f. and does not provide explicit forecasts of earthquake rate; PI simply provides a ranking of cells and, therefore, the RELM likelihood testing procedures cannot be applied in a straightforward way. We were provided the PI hotspot map values by J. Holliday (personal communication 2005). The RI alarm function constitutes a next-event spatial p.d.f. and, upon assumption of a magnitude distribution and regional seismicity rate, can be tested using the RELM methods (e.g. Zechar *et al*. 2007). For this experiment, we computed the RI values using the parameters suggested by Rundle *et al.* (2002). The alarm function of the NSHM model is a p.d.f. of space and magnitude and yields a forecast of expected seismicity rates; a previous version of the model is currently being tested by RELM (Petersen *et al.* 2007). We computed the 2002 NSHM values using the OpenSHA platform (Field *et al.* 2005). These three models also make for an interesting set of examples as they demonstrate the potential to compare heterogeneous forecasts.

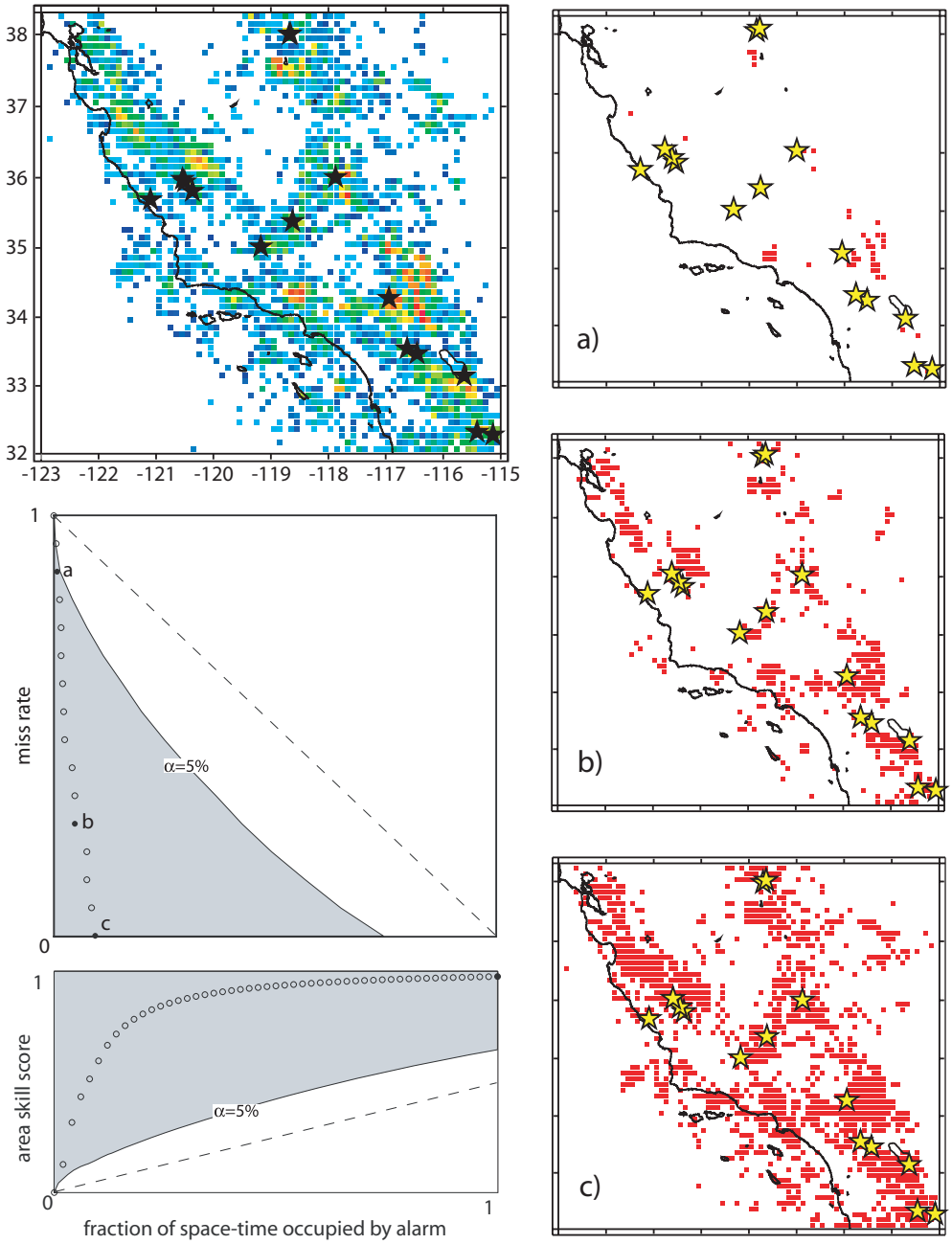

Figure 7. Same as Fig. 6 for the Pattern Informatics (PI) alarm function.

We consider the experiment specified by Rundle *et al.* (2003): to forecast the epicentral locations of $M \geq 5$ earthquakes during the 10 yr period starting 1 January 2000 in the gridded region with latitude ranging from 32° to 38.3°, longitude ranging from −123° to −115°, and a spatial discretization of 0.1°. We consider this a quasi-prospective experiment because the PI and RI forecasts were issued in 2002; none of the forecasts, however, use data collected after the beginning of the experiment. Although the magnitude scale and earthquake catalogue to be used for verification were not stated in the original experiment specification, we followed the RELM project in taking the ANSS composite catalogue to be the authoritative data source for this natural laboratory. We selected all tectonic earthquakes in this region since 2000 that had ANSS reported magnitudes greater than or equal to 5.0, regardless of the reported magnitude scale. This selection process yielded the 15 target earthquakes listed in Table 1; the corresponding earthquake catalogue is available in the Supplementary Material online.

RESULTS

Earthquakes cluster in space and time, and therefore, any forecast that captures this clustering behaviour should outperform a uniform reference model (Kagan 1996; Stark 1996, 1997; Michael 1997). Figs 6–8 confirm the expectation that RI, PI and NSHM provide significant gain relative to a spatially uniform prior distribution. From the area skill score trajectories, and in particular, the points at $\tau = 1$, it can be seen that, at greater than 95 per cent confidence, each forecast obtains an area skill score that is greater than $^1/_2$.

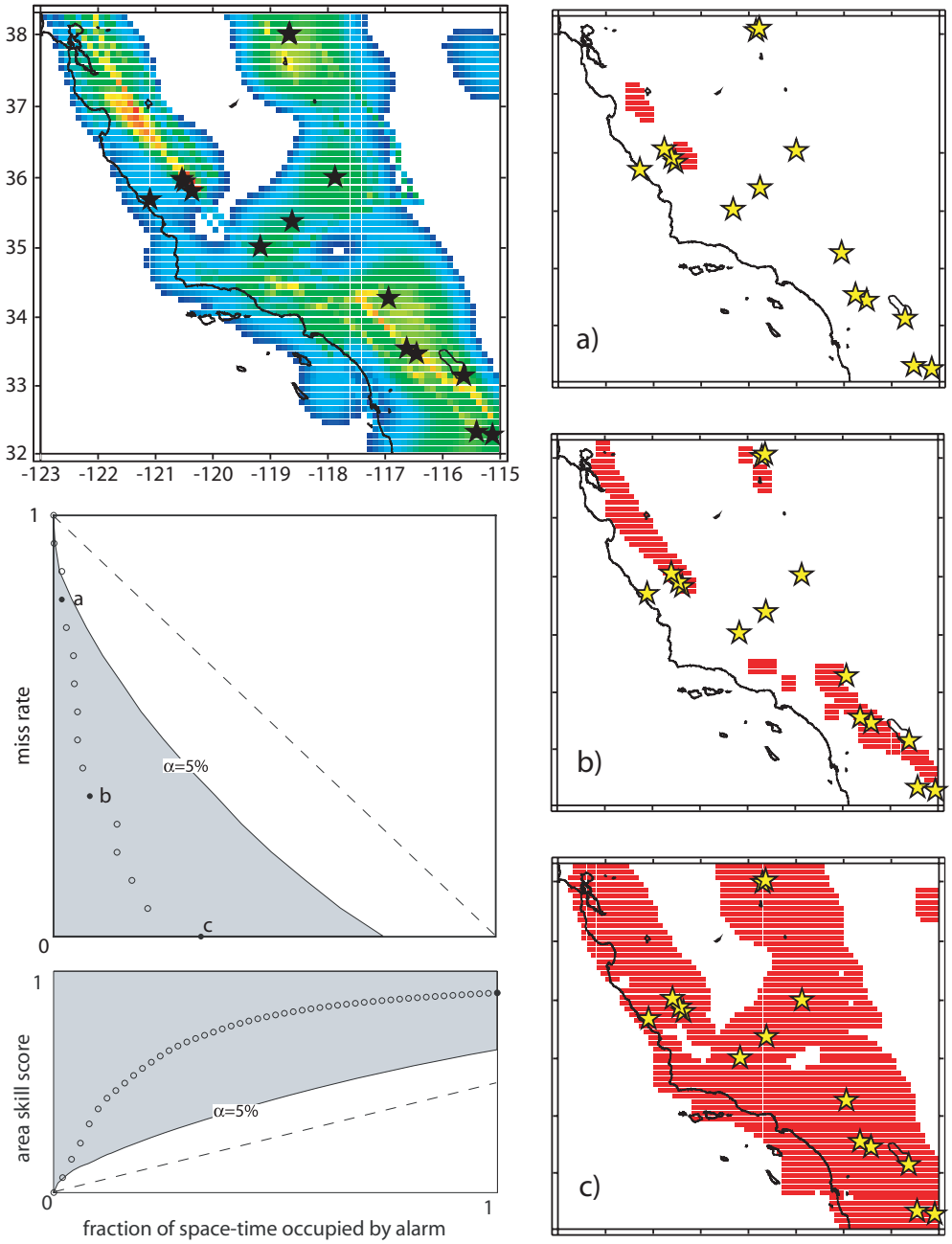

Figure 8. Same as Fig. 6 for the National Seismic Hazard Map (NSHM) alarm function.

To include time-invariant spatial clustering in the reference model, we use the RI alarm function values as the prior distribution—in other words, the RI index defines the measure of space for τ—and compute the Molchan trajectory and corresponding area skill score curve for the PI and NSHM forecasts. Computationally, this means that the 'cost' of declaring an alarm in a given cell is proportional to the RI alarm function value of this cell; in the case of a uniform prior distribution, the cost is everywhere equal. Fig. 9 shows the result of testing for the 15 target earthquakes since 1 January 2000. In the calculation of τ and ν, we include the margin of error suggested by Rundle *et al.* (2003); namely, if a target earthquake occurs in an alarm cell or in one of the alarm cell's immediate neighbours (Moore neighbourhood), it is considered a hit. Accordingly, all cells in the Moore neighbourhood of alarm regions are counted as alarms when computing τ. We note that our method for generating alarms from an alarm function is exact and efficient; we use as the alarm thresholds all of the unique values of the alarm function, rather than iterating the thresholds by some constant. The codes for generating the Molchan and area skill score trajectories are provided in the Supplementary Material online.

With RI as the reference model, the Molchan trajectories for PI and NSHM are closer to the descending diagonal, indicating much smaller probability gains than in the case of a uniform reference model. The NSHM forecast, however, yields three exceptional trajectory points at low values of τ. These points arise from the fact that three of the target earthquakes—numbers 3, 8 and 13 in Table 1—occurred in cells where NSHM had very high alarm function values and RI had low values. These three hits also manifest themselves in the area skill score trajectory, where NSHM obtains a few exceptional points at small τ. Because the statistical power

Table 1. Fifteen target earthquakes occurring in the testing region with latitude ranging from 32° to 38.3°, longitude ranging from −123° to −115°, during the interval 2000 January 1–2007 June 30.

Number	Origin time	Magnitude	Latitude (°)	Longitude (°)
1	2001/02/10 21:05	5.13 ML	34.2895	−116.9458
2	2001/07/17 12:07	5.17 Mw	36.0163	−117.8743
3	2001/10/31 07:56	5.09 ML	33.5083	−116.5143
4	2002/02/22 19:32	5.70 Mw	32.3188	−115.3215
5	2003/12/22 19:15	6.50 Mw	35.7002	−121.0973
6	2004/09/18 23:02	5.55 Mw	38.0095	−118.6785
7	2004/09/18 23:43	5.40 Mw	38.0187	−118.6625
8	2004/09/28 17:15	5.96 Mw	35.8182	−120.3660
9	2004/09/29 17:10	5.00 Mw	35.9537	−120.5022
10	2004/09/29 22:54	5.03 Mw	35.3898	−118.6235
11	2004/09/30 18:54	5.00 Mw	35.9890	−120.5378
12	2005/04/16 19:18	5.15 ML	35.0272	−119.1783
13	2005/06/12 15:41	5.20 Mw	33.5288	−116.5727
14	2005/09/02 01:27	5.10 Mw	33.1598	−115.6370
15	2006/05/24 04:20	5.37 Mw	32.3067	−115.2278

of the area skill score increases with increasing τ, however, we test the area skill score value at $\tau = 1$; here, neither PI nor NSHM obtain a score that is significantly greater than $^1/_2$. Given these results and because we test at standard significance values $\alpha = 1$, 5 and 10 per cent, we cannot reject at the 90 per cent confidence level the null hypothesis that PI and NSHM belong to C_1. In other words, the observed set of 15 target earthquakes during this experiment is consistent with the spatial distribution forecast by RI and neither PI nor NSHM provide significant gain relative to this simple model of smoothed seismicity.

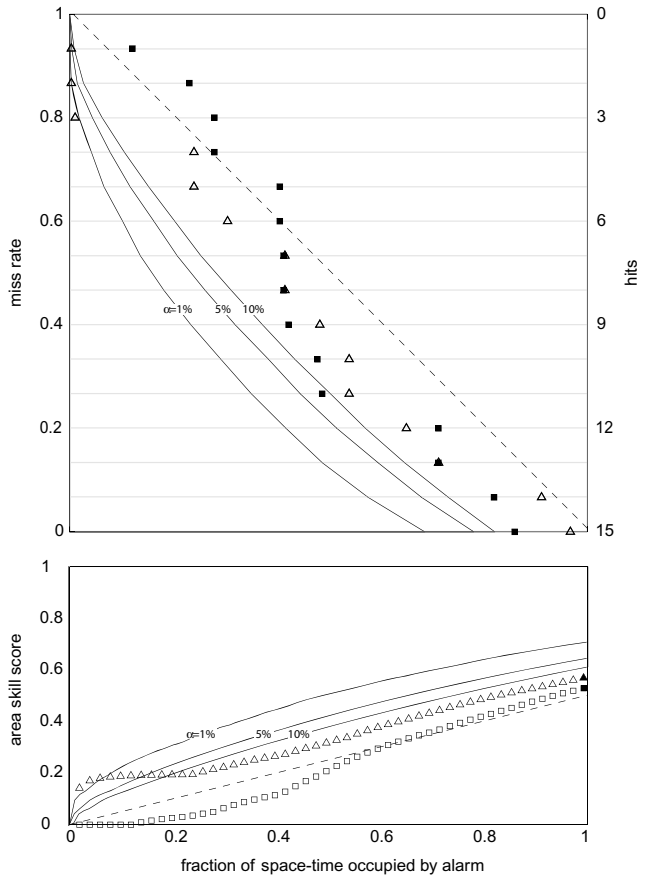

Figure 9. Results of Molchan trajectory/area skill score analysis for PI (squares) and NSHM (triangles) relative to the RI reference model. Top panel shows complete Molchan trajectories for both predictions and bottom panel shows corresponding area skill score curve. Each plot also shows the $\alpha = 1$, 5 and 10 per cent critical boundaries. In the Molchan trajectory plot, points below these boundaries reject the alarm region null hypothesis; in the area skill score trajectory plot, points above the boundaries reject the alarm function null hypothesis. We test the area skill score points at unit τ—the filled points on the bottom panel—and find that neither PI nor NSHM provides significant gain relative to RI.

CONCLUSIONS

In an illustration of an alarm-based earthquake prediction evaluation technique, we have shown that, contradictory to the retrospective testing of Rundle *et al.* (2002, 2003), the PI forecast model does not yield statistically significant performance in a quasi-prospective earthquake forecast. In particular, at the 90 per cent confidence level, we are unable to reject the null hypothesis that PI and NSHM provide no gain relative to RI.

With respect to NSHM, we note that this model was constructed to forecast large earthquakes in the long term, and we have tested it for a period of only 7.5 yr during which only one damaging earthquake occurred. By increasing either the duration of the experiment or the minimum magnitude of target earthquakes, the fault information included in the NSHM forecast might provide better spatial resolution and accuracy than purely statistical methods. Both of these increases require more time to collect a meaningful number of events but may offer insight into the spatial predictability of a region's largest earthquakes. Fault-based experiments and testing thereof will be investigated further by CSEP researchers.

While the alarm functions considered here focus on forecasting the geographic location of future earthquakes above a minimum magnitude, our method can be applied to more complex forecasts, including time-varying, magnitude-varying and fault-based alarm functions. In the experiment considered here, we have disregarded catalogue errors. Because the forecasts are time-invariant, timing errors are irrelevant to the test. The spatial discretization of the forecasts is of such a scale that location errors are probably negligible. Because this experiment concerns earthquakes above a minimum magnitude without magnitude discretization, magnitude errors are only relevant for earthquakes close to the minimum target magnitude and are also unlikely to change the result of the hypothesis test. In general, however, it is important to consider catalogue errors when testing earthquake forecasts (e.g. Werner & Sornette submitted ms), and our testing method can account for such errors by using simulations comparable to those planned for the RELM experiments (Schorlemmer *et al.* 2007).

The framework for evaluating multilevel alarms has been described by Molchan and Kagan (1992); applying these principles would allow further disaggregation of testing results. For example, magnitude discretization can reveal that one model accurately predicts small earthquakes and another is better at predicting intermediate size earthquakes. A bootstrap approach where these models are combined may be an effective way to proceed with earthquake prediction research. Hypothesis testing using the area skill score can be used as a guide in this process.

The consistent failure to find reliable earthquake precursors leads us to believe that a more effective way to advance earthquake prediction is a 'brick-by-brick' approach that synthesizes hypotheses, models, and data across space- and timescales (Jordan 2006). Rigorous testing methods like the one described here are vital in identifying the most robust characteristics of seismicity and improving ref-

erence models. Such testing may provide a better means of communicating earthquake forecast performance and progress to the public.

ACKNOWLEDGMENTS

This research has been supported by the National Science Foundation via a Graduate Research Fellowship (JDZ) and grant CMG 0621119, and by the Southern California Earthquake Center (SCEC). SCEC is funded by NSF Cooperative Agreement EAR-0106924 and USGS Cooperative Agreement 02HQAG0008. We thank Yan Kagan and an anonymous reviewer for insightful comments on the manuscript. We thank Nitin Gupta for assistance in computing the NSHM forecast. The SCEC contribution number for this paper is 1099.

REFERENCES

Aki, K., 1981. A probabilistic synthesis of precursory phenomena, in *Earthquake Prediction: An International Review,* pp. 566–574, eds. Simpson, D. & Richards, P., Am. Geophys. Union, Washington, DC.

ANSS Earthquake Catalog. Produced by Advanced National Seismic System (ANSS) and hosted by the Northern California Data Center (NCEDC), 1932–2007. http://quake.geo.berkeley.edu/anss.

Field, E.H., 2007. Overview of the working group for the development of regional earthquake likelihood models (RELM), *Seismol. Res. Lett.,* **78**(1), 7–16.

Field, E.H., Gupta, N., Gupta, V., Blanpied, M.L., Maechling, P.J. & Jordan, T.H., 2005. Hazard calculations for the WGCEP-2002 forecast using OpenSHA and distributed object technologies, *Seismol. Res. Lett*., **76,** 161–167.

Frankel, A., Mueller, C., Barnhard, T., Perkins, D., Leyendecker, E., Dickman, N., Hanson, S. & Hopper, M., 1996. National seismic-hazard maps: Documentation June 1996 *U.S. Geol. Surv. Open-file report 96–532*, 41 pp.

Frankel, A. *et al.*, 2002. Documentation for the 2002 update of the national seismic hazard maps, *U.S. Geol. Surv. Open-file report 02–420*, 33 pp.

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

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

Jordan, T.H., 2006. Earthquake predictability, brick by brick, *Seismol. Res. Lett.,* **77**(1), 3–6.

Kafka, A.L., 2002. Statistical analysis of the hypothesis that seismicity delineates areas where future large earthquakes are likely to occur in the Central and Eastern United States, *Seismol. Res. Lett.,* **73,** 990–1001.

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

Kagan, Y.Y. & Jackson, D.D., 2000. Probabilistic forecasting of earthquakes, *Geophys. J. Int.,* **143,** 438–453.

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., Springer-Verlag, Berlin.

Keilis-Borok, V.I. & Kossobokov, V., 1990. Premonitory activation of earthquake flow: algorithm M8, *Phys. Earth Planet. Inter.,* **61**(1/2), 73–83.

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.

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

Kossobokov, V. & Shebalin, P.N., 2003. Earthquake prediction, in *Nonlinear Dynamics of the Lithosphere and Earthquake Prediction,* pp. 141–205, eds Keilis-Borok, V.I. & Soloviev, A., Springer-Verlag, Berlin.

Lehman, E.L. & Romano, J.P., 2005. *Testing Statistical Hypotheses,* 3rd edn, pp. 784, Springer, New York.

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

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

Michael, A.J., 1997. Testing prediction methods: earthquake clustering versus the Poisson model, *Geophys. Res. Lett.,* **24**(15), 1891–1894.

Molchan, G.M., 1990. Strategies in strong earthquake prediction, *Phys. Earth Planet. Inter.,* **61,** 84–98.

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

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

Petersen, M., Cao, T., Campbell, K. & Frankel, A., 2007. Time-independent and time-dependent seismic hazard assessment for the state of California: uniform California earthquake rupture forecast model 1.0, *Seismol. Res. Lett.,* **78**(1), 99–109.

Rhoades, D.A. & Evison, F.F., 1989. Time-variable factors in earthquake hazard, *Tectonophysics,* **167,** 201–210.

Rhoades, D.A. & Evison, F.F., 2004. Long-range earthquake forecasting with every earthquake a precursor according to scale, *Pure Appl. Geophys.,* **161,** 47–72.

Rundle, J.B., Tiampo, K., Klein, W. & Sa Martins, J., 2002. Self-organization in leaky threshold systems: the influence of near-mean field dynamics and its implications for earthquakes, neurobiology, and forecasting, *Proc. Natl. Aca. Sci. USA,* **99,** 2514–2521.

Rundle, J. B., Turcotte, D.L., Shcherbakov, R., Klein, W. & Sammis, C.G., 2003. Statistical physics approach to understanding the multiscale dynamics of earthquake fault systems, *Rev. Geophys.,* **41**(4), 5.1–5.30.

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.

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

Stark, P.B., 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,** 495–499.

Tiampo, K. F., Rundle, J.B., McGinnis, S., Gross, S.J. & Klein, W., 2002. Mean-field threshold systems and phase dynamics: an application to earthquake fault systems, *Europhys. Lett.,* **60**(3), 481–488.

Werner, M.J. & D. Sornette. submitted ms. Magnitude uncertainties impact seismic rate estimates, forecasts and predictability. Submitted to *J. Geophys. Res.*

Wessel, P. & W. Smith, 1998. New, improved version of Generic Mapping Tools released, *Eos Trans. AGU,* **79**(47), 579.

Zechar, J.D. & Jordan, T.H., 2007. The area skill score statistic for evaluating earthquake predictability experiments, in preparation.

Zechar, J.D., Jordan, T.H., Schorlemmer, D. & Liukis, M., 2007. Comparison of two earthquake predictability evaluation approaches, Molchan error trajectory and likelihood, *Seism. Res. Lett.,* **78**(2), 250.

SUPPLEMENTARY MATERIAL

The following supplementary material is available for this article:

Supplement S1. The catalog data and code used to compute RI alarm function values and create the Molchan diagram and area skill score trajectory plots for Figs 3, 6–9.

This material is available as part of the online article from: http://www.blackwell-synergy.com/doi/abs/10.1111/j.1365-246X.2007.03676.x (This link will take you to the article abstract.)

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