Geophys. J. Int. (2012) **188,** 336–342 doi: 10.1111/j.1365-246X.2011.05274.x

Testing earthquake forecasts using reliability diagrams

J. R. Holliday,[1] W. J. Gaggioli[2] and L. E. Knox[1]

[1]*Department of Physics, University of California, Davis, California, USA. E-mail: jrholliday@ucdavis.edu*
[2]*XeraSys Incorporated,* 231 *G St, Davis, CA* 95616*, USA*

Accepted 2011 October 18. Received 2011 September 28; in original form 2011 January 3

SUMMARY

Reliability diagrams are a standard forecast evaluation tool for comparing forecasted probabilities of binary events with their observed frequencies. To make this comparison for rare events, forecasts of similar probability are grouped together. Here, we demonstrate the utility of reliability diagrams for earthquake forecast evaluation with an application to a modified version of the Holliday *et al.* time-dependent earthquake forecasting method. We quantify reliability with a χ^2 statistic and find that the forecasting method is consistent with reliability; that is, forecast probabilities match rates of occurrence to within the expected uncertainties.

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

1 INTRODUCTION

In this paper, we introduce the reliability diagram (Wilks 1995; Mason 2003) and show how it can be used to quickly assess the quality of an earthquake forecast. The reliability diagram is a very straightforward way of assessing how well forecasts of binary events agree with the actual rates of occurrence. Forecasts of similar probability are grouped together and the expectation value for the number of events in that group is directly compared with the actual number of observed events. Because of this simplicity, reliability diagrams are used in many fields, including meteorology, statistics, hydrology and pharmacology. Here we show how they can also be used to better understand and evaluate earthquake forecasts.

We advocate the use of reliability diagrams, not to exclude other testing methods, but to bring out additional information. For example, one motivated and popular way to compare forecasts is through a calculation of the likelihood (Bevington & Robinson 1992; Gross & Rundle 1998; Kagan & Jackson 2000; Tiampo *et al.* 2002a; Schorlemmer *et al.* 2007). The likelihood is proportional to the probability of the model, but the proportionality constant is unknown. Reliability diagrams can help quantize this measure of proportionality.

As a case study, we consider a forecasting methodology based on the work of Holliday *et al.* (2006a,b, herein referred to as H06). We use reliability diagrams in several different ways to probe the properties of the forecast, and its performance in comparison with observed rates of occurrence. Among our findings from this investigation are that the H06-modified forecasts are reliable (observed rates of occurrence match forecast probabilities to within the uncertainties) but not demonstrably superior to a much simpler static model based on a spatially varying Gutenberg–Richter distribution.

2 TWO FORECASTING METHODS

We begin by introducing the forecasting methods we are going to test. The first is an extension of the ensemble intensity difference (EID) method of H06, which we call in this paper EID2. The second is a reference forecast method which we find to be valuable for comparison of its performance with EID2's performance. We call it the static GR forecast because its basic premise is that earthquakes are a time-independent stochastic phenomenon with a frequency-magnitude distribution that follows the Gutenberg–Richter power law. This is the basic premise of what is also called the relative intensity (RI) method of Tiampo *et al.* (2002b) and Holliday *et al.* (2005). We first describe EID2, then the static GR forecast in more detail and then describe how the parameters of these models are calibrated.

2.1 EID2

The EID2 method produces time and space-dependent probabilities of future earthquakes with $M \geq 6$. The probability is factored into a probability that there will be an earthquake in a given time period in a given region, and a probability of the location of the earthquake within the region, given that there is an earthquake within the region:

$$\begin{aligned} &P(\text{earthquake in } x_\alpha \text{ during } i) \\ &\quad = P(\text{earthquake in } x_\alpha | \text{ earthquake during } i) \\ &\qquad \times P(\text{earthquake during } i \text{ in region } R), \end{aligned} \tag{1}$$

where x_α is a $0.1^\circ \times 0.1^\circ$ spatial bin within region R, and i is an interval of time (temporal bin). The vertical bar | is standard probabilistic notation for 'given'. We refer to these separate factors as the spatial factor and the temporal factor.

2.1.1 The temporal factor

The output of the EID method is a function of time called a risk classifier (RC) curve (see Fig. 1). H06 refer to times when the RC curve is positive as 'high-risk' periods and times when the RC curve is negative as 'low-risk' periods. For their northern California and southern California RC curves, H06 show a strong concentration of earthquakes during the high-risk periods. We consider the significance of this clustering in Section 4.

To obtain the temporal factor of the EID2 method, we extend the EID method in two ways. First, we have a different method for calibrating the $F_{\max}$ parameter that affects the EID RC curve. Second, we introduce a method for assigning quantitative earthquake probabilities to the high- and low-risk periods.

Our different method for calibrating $F_{\max}$ addresses a weakness with the original EID method that was obscured by an error in the text. In the original analysis, $F_{\max}$ was described as the value which maximizes the marginal utility of the forecast (Holliday *et al.* 2006b). In practice, however, the parameter was chosen to maximize the performance of the method on retrospective testing. Our modified calibration does not use test (or future) data, yet it produces similar results and thereby reduces the possibility of future data influencing the forecast.

Our fundamental assumption for the distribution of probability in time is that the waiting time between 'events' follows a Weibull distribution. We define an event as either a transition between risk states, or an $M \geq 6$ earthquake. Thus, the probability of an event occurring between the time of some event, t_e and some later time t_s (for a guide to our temporal notation, see Table 1) is given by

$$P(i_{es}) = 1 - \exp[-(\Delta t_{es}/\tau)^b], \tag{2}$$

where $\Delta t_{es} = t_s - t_e$. Note that the Poisson distribution is a special case of the Weibull distribution with $b = 1$. Of course, we are interested in the probability of an earthquake, not the probability of an event. In practice, we simply equate the probability of an earthquake with the probability of an event.

For forecasts over intervals beginning at some time t_f that is not coincident with an event, we have

$$P(i_{fs}) = \left[P(i_{es}) - P(i_{ef})\right]/N, \tag{3}$$

where t_e is the time of the prior event closest to t_f and N is a normalization constant equal to $1 - P(i_{ef})$, chosen so that $P(i_{fs}) = 1$ in the limit that $\Delta t \to \infty$. With some algebra, we obtain

$$P(i_{fs}) = 1 - \exp\left[\left(\frac{\Delta t_{ef}}{\tau}\right)^b - \left(\frac{\Delta t_{es}}{\tau}\right)^b\right]. \tag{4}$$

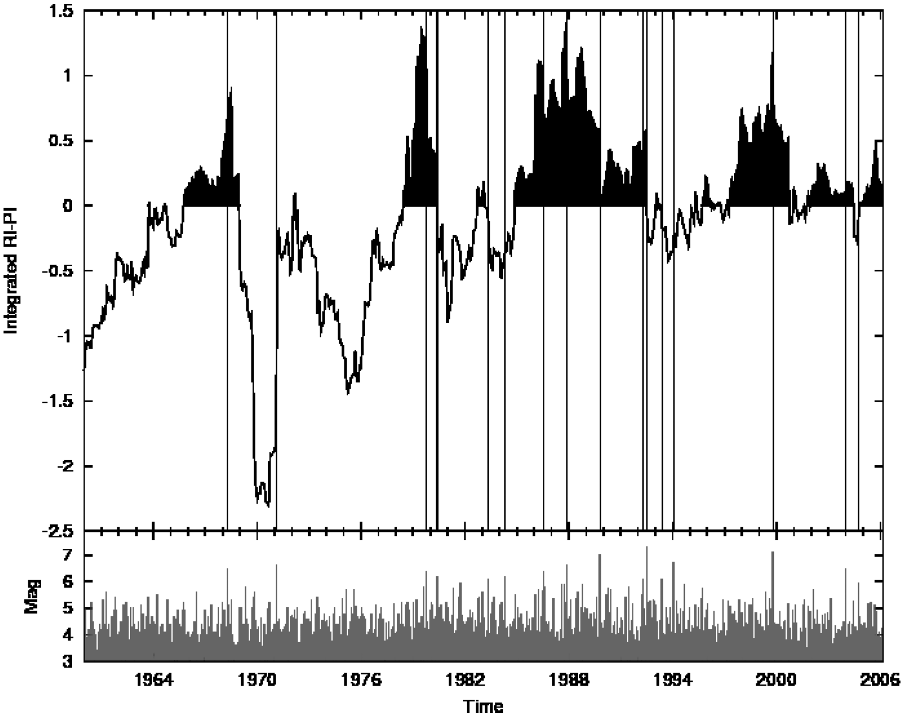

Figure 1. Sample risk classifier (RC) curve (top panel) and earthquake magnitude (bottom panel) as a function of time for events occurring in California. Vertical black lines represent times of major earthquakes having $M \geq 6.0$. Reproduced from Holliday *et al.* (2006a).

Table 1. Reference table for the notation we use to distinguish times, durations and intervals.

Symbol	Meaning	Example
t	A particular point in time	Midnight on 2006 January 30
t_e	Time of an event (earthquake or risk-state transition)	
t_f	Time at which forecast is made	
t_b	Beginning time of a forecast interval	We always set $t_b = t_f$
t_s	Stop time of a forecast time interval	
Δt	A temporal duration	$\Delta t_{ef} = t_e - t_f$
i	An interval of time	i_{fs} is the interval from t_f to t_s

For t_f in high risk, the Weibull parameters τ and b are calibrated with historic data (Gelman *et al.* 2003). For t_f in low risk, we assume a Poisson distribution; that is, we set $b = 1$ and calibrate τ with historic data. We treat the high- and low-risk τ values as independent and call them τ_h and τ_l, respectively.

2.1.2 *Algorithm*

In cookbook format, the modified EID2 method as applied in this paper is composed of the following steps.

(1) The seismically active region is binned into boxes of some characteristic size, and all events having $M \geq M_c$ are used. These boxes are labelled x_i. For the western United States and Cascadia regions, we typically use boxes with edge length equal to 0.1° and a threshold magnitude M_c equal to 3.0. In Mexico, we typically use boxes with edge length equal to 0.2° and a threshold magnitude M_c equal to 4.0 due to degraded catalogue completeness.

(2) The seismicity obtained from the regional catalogue for each day in each box is considered to be uniformly spread over that box. The resulting intensities for each box forms a time-series.

(3) Three time parameters are determined: t_0, t_1 and t_2. t_0 is chosen to be the base time. For California, we typically take $t_0 =$ 1932 January 1. For other regions, we typically take t_0 to be in the 1960s, depending on data availability. t_2 is chosen such that the number of $M > M_T$ events during the time period $t_2 \rightarrow t$ is equal to some value specified by the regional b-value. M_T is allowed to sweep from M_c to $M_c + 2$. t_1 is chosen such that $t_2 - t_1 = 13$ yr.

(4) RI, $I(x_i;\, t_0,\, t_2)$, and PI, $\Delta I(x_i;\, t_0,\, t_1,\, t_2)$, maps are created for the region.

(5) These maps are converted to binary forecasts, and ROC diagrams are constructed for the snapshot window $t_2 \rightarrow t$. $\Delta A(t)$ is calculated by integrating $A_{\mathrm{RI}}(t) - A_{\mathrm{PI}}(t)$ over $F \in [0, F_{\max}]$.

(6) $\Delta A(t)$ is averaged (or, for simplicity, summed) over a range of M_T values yielding $\Delta\mathcal{A}(t)$. As stated above, M_T is allowed to sweep from M_c to $M_c + 2$. If $\Delta\mathcal{A}(t) \geq 0$, the region is assumed to be in a high-risk state. If $\Delta\mathcal{A}(t) < 0$, the region is assumed to be in a low-risk state.

(7) Finally, the time-dependent probability was allocated over the seismic region by multiplying it against a suitable spatial density function. For this analysis, we created a density function by stacking a modified PI map on top of a 'background' RI map and normalizing to have unit total area. For the modified PI map, we followed the procedure outlined by Chen *et al.* (2005) and used only the top 15 per cent of active sites.

2.2 Static GR

Many statistical tests for the performance of weather forecasts include a comparison forecast called 'climatology'. The climatological forecast for rain on 2005 December 21 in Davis, CA is that it will occur with a probability equal to the observed frequency of occurrence of rain on previous December 21 days in Davis, CA. This is the forecast you get in the absence of any meteorological modelling assumptions and extra data beyond the historical record of occurrence of rain. It is a useful forecast for comparison, to test if your modelling is allowing you to use extra data to make more accurate forecasts; that is, it helps to answer the question, 'is my modelling actually improving the forecasts?'

The static GR forecast is a seismicity analogue of the climatological weather forecast. We assume a Gutenberg–Richter distribution of frequency with earthquake magnitudes such that the rate of earthquakes with magnitude greater than or equal to M is given by

$$r(\geq M) = a 10^{-bM}, \tag{5}$$

where a and b are constants. We find the least squares fit to a and b using the differential version

$$\frac{\mathrm{d}r}{\mathrm{d}M} = a \ln(10) b\, 10^{-bM} \tag{6}$$

and assuming standard deviations given by the square root of the number of earthquakes in each magnitude bin of width 0.1, uncorrelated from bin to bin, starting from some minimum magnitude bin. The choice of minimum magnitude depends on the quality of the data. Our use of the number of earthquakes in each bin to determine the errors leads to a bias in our a and b estimate because downward fluctuations in number (from average) are treated as having smaller uncertainty and the opposite for upward fluctuations. We have determined that this bias is negligibly small. Although the parameters of the model are assumed to be time-independent, forecast probabilities do change with time as our estimates of the parameters change with more available data. The a parameter in particular is strongly influenced by aftershock sequences. With a and b fit for a region, the final step is to distribute that rate across the testing region.

2.2.1 *Algorithm*

In cookbook format, the Static GR method as applied in this paper is composed of the following steps:

(1) We bin the number of events having $M \geq M_c$ into map bins labelled x_α. For the western United States and Cascadia regions, we typically use bins with edge length equal to 0.1° and a threshold magnitude M_c equal to 3.0. In Mexico, we typically use bins with edge length equal to 0.2° and a threshold magnitude M_c equal to 4.0 due to degraded catalogue completeness.

(2) The seismicity obtained from the regional catalogue for each day in each bin is considered to be uniformly spread over that bin. The resulting intensities for each bin forms a time-series.

(3) By integrating each time-series from t_0 to the current time and dividing by the length of the time-series, the historic rate of earthquake occurrences is estimated for each bin. For California, we typically take $t_0 = 1932$ January 1. For other regions, we typically take t_0 to be in the 1960s, depending on data availability. These rates are then normalized by requiring the sum over all bins is unity. The resulting map is then smoothed by convolution with a Gaussian with $\sigma = 20$ km. The resulting map is the spatial factor, $P(\text{earthquake in } x_\alpha \,|\, \text{earthquake during } i)$ of eq. (1).

(4) To get the temporal factor, we calculate the expected rate of $m \geq 6.0$ earthquakes in the seismic region. We do so by finding the best-fitting Gutenberg–Richter a and b parameters of the seismic region. We assume that earthquake probabilities are constant in time and uncorrelated from one day to the next; that is, the number of earthquakes in any given time interval (larger than a day) follows a Poisson distribution.

3 TESTING METHODOLOGY

We now describe the production of reliability diagrams, and some of the associated terminology. We also describe the reduction of reliability diagrams to a χ^2 statistic, that can be used for a quantitative measure of the performance.

3.1 Reliability diagrams

A reliability diagram can be used to see how reliably the forecasted probabilities are matched with the observed rates of occurrence. Typically, each single forecast for a given bin (a particular geographical area over a particular stretch of time) has a probability of earthquake occurrence of much less than 1. To reduce the large sample variance that would plague the direct comparison of such low-forecasted probabilities with actual occurrences, we group bins together with similar forecasted probabilities. For each group of bins, we determine the average expectation value of occurrence and plot that number on the *x*-axis, with the actual rate of occurrence per bin on the *y*-axis.

Examples of reliability diagrams for six hypothetical forecasts are given in Fig. 2. Reliability is indicated by the proximity of the plotted points to the diagonal. If the curve lies below the line, this indicates overforecasting (probabilities too high); points above the line indicate underforecasting (probabilities too low). These panels include inset histograms indicating the total number of forecast bins in each probability grouping.

One property of forecasts illustrated by a reliability diagram is known as 'resolution'. A forecast that does not discriminate from one bin to another (as in the top left-hand panel) has zero resolution. A high-resolution forecast is one that has a large dynamic range of probabilities. The width of the range is evident in the dynamic range from the lowest probability grouping of bins to the highest probability grouping of bins. High resolution is desirable if achieving, it does not ruin reliability. The higher the resolution, the more impressive is reliability.

The reliability diagram is conditioned on the forecasts; that is, it answers the question 'given that *X* was predicted, what was the outcome?' It is, thus, complementary to the ROC diagram, which is conditioned on the observations.

The technique for constructing the reliability diagram is similar to that for calculating the ROC score, only instead of plotting the hit rate against the false alarm rate, the hit rate is calculated only from the sets of forecasts for each probability separately. It is then plotted against the corresponding forecast probabilities.

The hit rate (H_n) for each probability group n is defined as

$$H_n = \frac{O_n}{O_n + N_n}, \tag{7}$$

where O is the number of observed instances and N is the number of non-observed instances.

Frequency histograms (often shown as insets) are constructed from the same contingency tables as those used to produce the reliability diagrams. Frequency histograms show the frequency of forecasts as a function of the probability group. The frequency of forecasts (F_n) for probability group n is defined as

$$F_n = \frac{O_n + N_n}{T}, \tag{8}$$

where T is the total number of forecasts.

3.2 Probability of exceeding χ^2

The reliability diagrams are a qualitative tool for judging the success of a forecast, but quantitative measures are useful as well. To quantify the reliability, and to help us understand the implications for the viability of the model, we calculate a χ^2 statistic as well as its distribution under different hypotheses.

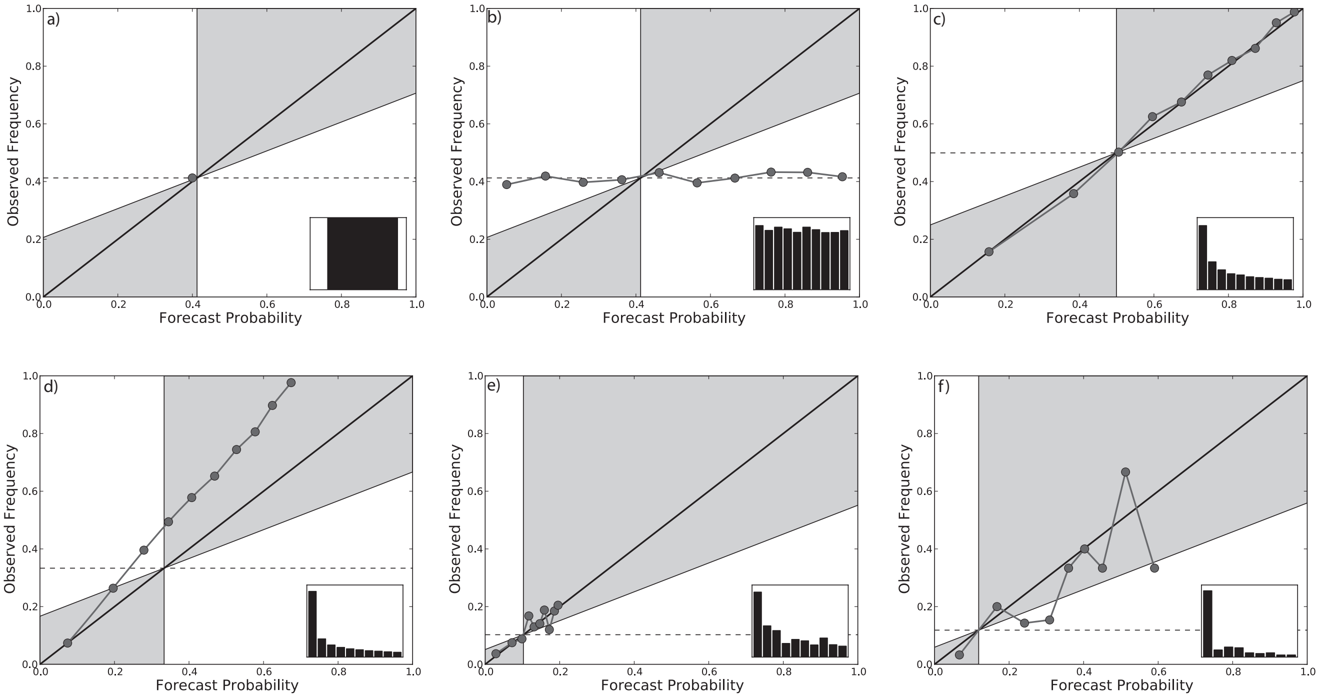

Figure 2. Hypothetical reliability diagrams showing observed relative frequency as a function of forecast probabilities for (a) forecasts with no resolution, (b) forecasts exhibiting small but non-zero resolution, (c) forecasts showing good resolution and good reliability, (d) forecasts showing an underforecasting bias, (e) reliable forecasts of a rare event and (f) verification data set limited by small sample size. Inset boxes indicate number of events forecasted for each probability group. Note that diagrams are ordered left-to-right-hand side, top-to-bottom panels.

To quantify the departure of the forecast from perfect reliability, compared to the level of departure expected due to sample variance, we calculate a χ^2 statistic as follows:

$$\chi^2 = \sum_{ij} (p_i - o_i) C_{ij}^{-1} (p_j - o_j), \tag{9}$$

where p_i is the average probability of earthquake occurrence for bins in group i (as determined from Monte Carlo simulations), o_i is the observed rate of occurrence for bins in group i and C_{ij} is the covariance of the simulated observed rate of occurrence. We estimate the covariance matrix via

$$C_{ij} = \sum_{\alpha} \left(p_{i\alpha} - p_i\right)\left(p_{j\alpha} - p_j\right), \tag{10}$$

where $p_{i\alpha}$ is average frequency of earthquake occurrence for bins in group i for Monte Carlo simulation α.

This χ^2 statistic is a measure of how well the forecasted probabilities agree with the observed rate of occurrence. We calculate the expected distribution of χ^2 from the Monte Carlo simulations and see where the observed χ^2 falls in the distribution. If it is way out on the tails of the distribution, this is an indicator of a problem with the model. One of the statistics we use is the 'χ^2 exceedance', $p(> \chi^2|\text{model})$ which is the probability of exceeding the observed χ^2 value, given the forecasting model. We take this as a measure of how unusual the observed χ^2 value is, given the assumed model. The more unusual ($p(> \chi^2|\text{model})$ being close to one or zero) the more improbable the model.

If the occurrence rates have a normal distribution, then the expectation value for χ^2 is equal to the number of groups. If χ^2 is very low—much lower than the great majority of the Monte Carlo simulations—this is an indicator that the data being used to test the forecasts has also been used to adjust the model being tested. If this is the case, then the present appearance of reliability may be spurious and not indicative of reliability with future data or data from other regions.

4 APPLICATION

In this section, we begin with the use of a reliability diagram to analyse the consistency of the EID2 forecast for northern California. We then move on to looking at how reliability diagrams can be used to examine the consistency of this model in more depth, and to compare its consistency with our climatology analogue, the static GR model.

4.1 EID2 forecasts for northern California

For our first example application, we examine forecasts of the EID2 model for spatio-temporal bins in northern California from 1985 to 2008. Each bin has a duration of 6 months and a spatial extent of 11 km × 11 km. Forecasts were updated every month. The reliability diagram for this set of forecasts and distribution of χ^2 values are shown in Fig. 3.

To produce the reliability diagram, we began by sorting the forecast bins into groups. First, we sorted the bins from highest probability to lowest. Then, starting with the lowest probability bin and continuing upward, we kept adding bins to the first group until the sum of probability in these bins was one quarter of the total. Note that in this application, data sparsity resulted in the need to use only four groups rather than, say, 10. The second group contained all remaining bins below some probability threshold chosen so that the sum of second group bin probabilities was equal to another quarter of the total sum. Similarly, the remaining two groups each contained a quarter of the probability.

The next step was to determine the X- and Y-axes locations for each of these groups. The Y-axis location was determined by taking the number of earthquakes that actually occurred in a group's bins and dividing by the group's total number of bins; this was the observed frequency of occurrence. The X-axis location can be determined analytically in some cases, but we chose to determine it by Monte Carlo simulation. In the Monte Carlo procedure, we created synthetic catalogues by sampling from the earthquake probability distribution of the model in question. These catalogues were then

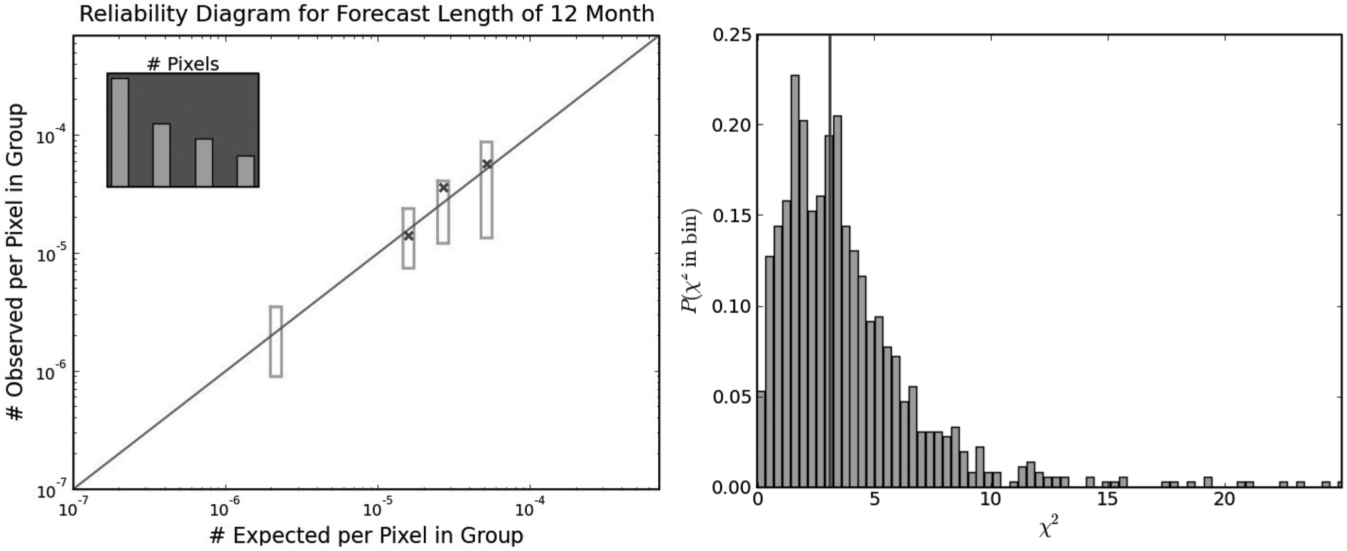

Figure 3. Example Reliability Diagram (left-hand panel, see text) and Associated χ^2 versus $P(\chi^2)$ (right-hand panel, see text). The model here is the EID2 model evaluated from 1985 to 2008 in northern California. The missing data point is below the lower limit of the Y-axis due to zero earthquakes observed for the lowest probability group of bins. The height of a bar in the right-hand panel is given by the fraction of Monte Carlo catalogues with χ^2 values in the χ^2 range for that bar.

fed through the same procedure that reduces the real catalogue data to rates of occurrence in groups ('observed' frequencies). In this implementation, any errors in the model distribution cascaded into the Monte Carlo sampling. We thus generated a distribution of synthetic frequencies and set the X-axis value for the group in question at the mean.

The plot not only show how the predicted mean observed frequency lines up with the actually observed frequency, but also shows the expected fluctuations in the observed frequency, given that the model used to generate the Monte Carlo catalogues is correct. The lower (upper) end of the error bars is set so that 16 per cent of the Monte Carlo catalogues have a lower (higher) observed frequency.

Temporal overlap of predictions complicates the procedure. When an earthquake occurs, there are multiple bins it is assigned to. For example, if we are testing forecasts of 6 month duration, then our bins will be 6 months long in the time direction. If the forecasts are updated monthly, a single earthquake will belong to six different space–time bins. To avoid overcounting, in this case we assign 1/6 of an earthquake to each bin.

From our Monte Carlo simulations, we also calculate the expected covariance matrix for the observed frequencies and use this to calculate χ^2. Further, for each Monte Carlo simulation, we can calculate the χ^2 value and thus the distribution of χ^2 values under the assumption that the forecasts are correct. We see that there is nothing unusual about the observed χ^2 value as it lies in a range with high probability.

In this case study, we find that the EID2 forecast is indeed reliable. The departures from perfect reliability are consistent with those expected from sample variance. Sample variance leads to a spread in the expected number of events that give us error bars that extend up and down factors of 2–3; thus we have tested the probabilities to this level of precision. The model possesses a resolution such that forecasts for the bins in its most probable quartile of bins have 25 times the probability of the forecasts for the bins in the least probable quartile.

4.2 EID2 examined in depth and compared with static GR model

We now use reliability diagrams to investigate the model in more depth. We study separately the spatial forecasts, temporal forecasts and spatio-temporal forecasts. We also produce reliability diagrams for the static GR model for comparison purposes. Our analysis is summarized in Fig. 4.

For the left-hand panel of Fig. 4, we are testing the spatial component of the forecasts only. The spatial component is the probability of the location of the earthquake within the region during a given time interval *given* that there is an earthquake in the region during that time interval. See eq. (1). Thus, we include (and only include) in the analysis those space–time bins for which there is an earthquake in the region to which the bin belongs.

We see that the static GR and EID2 spatial forecasts have very similar reliability diagrams. This is to be expected because the spatial forecasts are very similar. Their departures from perfect reliability are consistent with those expected due to sample variance; that is, the observed frequencies are consistent with the probabilities given by both models.

We note that our conclusions regarding these spatial forecasts are consistent with those of Zechar & Jordan (2008). They calculated Molchan diagrams to test three forecasts for California over the period 2000 January 1 to 2006 December 31, and found that the PI forecast and the RI forecast had very similar performances. They refer to these tests as 'quasi-prospective' because the forecasts were made with some of the testing interval existing in the past: the forecast was created in 2002 but covers the testing interval 2000–2010. The EID2 spatial forecast is a combination of the PI and RI forecasts. And RI itself is a version of static GR.

For the central panel of Fig. 4, we are testing only the temporal component of the forecasts. We use only two groups in this test, with temporal bins grouped according to whether the EID2 forecast is for high risk or low risk. For ease of comparison, the groups are the same for both forecasts. For static GR, the forecasts are nearly time-independent and that is why the forecasted probabilities are so close on the graph. The small amount of variation in probability with time is due to the change in calibration of the GR parameters as more data become available.

The central panel shows that both forecast models are consistent with the data. The EID2 forecast is bolder (has higher resolution) and its reliability diagram has a lower χ^2 value. However, neither χ^2 value is highly unusual. 81 percent of the Monte Carlo catalogues produced under the EID2 hypothesis have larger χ^2 values than the EID2 χ^2 value for the real catalogue. The corresponding percentage for the static GR model is 27 per cent. Statistically, the χ^2 test does not rule out either the static GR model or the EID2 model.

Finally, in the right-hand panel, we see the reliability diagrams for the complete spatio-temporal forecasts of our two methods, this time for northern and southern California combined. Again, both models reinforce the data.

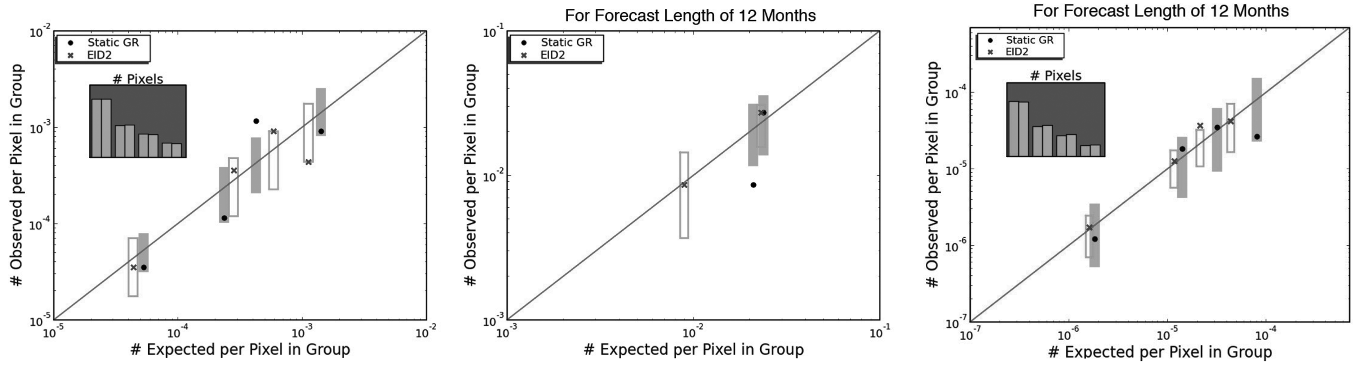

Figure 4. Reliability diagrams for California spatial forecasts (left-hand panel), temporal forecasts (centre panel) and complete spatio-temporal forecast (right-hand panel) for both the EID2 forecasts (open rectangles) and the static GR forecasts (filled rectangles) from 1980 January 1 to 2007 December 31. Forecasts are for a duration of 12 months and are updated monthly.

5 SUMMARY AND CONCLUSIONS

The chief virtue of the reliability diagram is that it provides a straightforward way of comparing the forecasted frequencies of occurrence with observed rates of occurrence. As a demonstration of its utility, we analysed a modified version of the Holliday *et al.* (2006b) temporal earthquake forecasting method (H06). We also analysed a simple reference model based on static Gutenberg–Richter scaling.

To determine the significance of departures from exact reliability, we produced a suite of Monte Carlo catalogues. We used the χ^2 value as a summary statistic for the reliability diagrams and used the synthetic catalogues to study its expected distribution under the two differing hypotheses.

We found that, as expected, the static GR and EID2 spatial forecasts have very similar reliability diagrams and that their departures from perfect reliability are consistent with those expected due to simple sample variance. Our results are consistent with the conclusions published by Zechar & Jordan (2008). Their analysis of PI and RI forecasts for California found the two models performed similarly well at forecasting future earthquakes. The EID2 forecast in our analysis is a combination of PI and RI forecasts, and the RI forecast by itself is a version of static GR.

We also found that the distinction between high- and low-risk states is compatible with the observed rates of occurrence, although the same is true for a static model in which there is no such distinction.

In conclusion, we find that reliability diagrams are useful tools for investigating the consistency of earthquake probability forecasting methods.

ACKNOWLEDGMENTS

The authors gratefully thank Dr. J. B. Rundle and Dr. D. L. Turcotte for many useful conversations. In addition, we thank the rest of the XeraSys staff (M. E. Clayton, D. Hirst, B. Johnson, T. Knutstad, C. L. Sershen, A. J. Smith and B. A. Wilson) for useful conversations and camaraderie. JRH was supported by NASA Grants #NNX08AF69G and #NNX07AQ93H.

REFERENCES

Bevington, P.R. & Robinson, D.K., 1992. *Data Reduction and Error Analysis for the Physical Sciences,* McGraw-Hill, New York.

Chen, C.C., Rundle, J.B., Holliday, J.R., Nanjo, K.Z., Turcotte, D.L., Li, S.C. & Tiampo, K.F., 2005. The 1999 Chi-Chi, Taiwan, earthquake as a typical example of seismic activation and quiescence, *Geophys. Res. Lett.,* **32,** L22315, doi:10.1029/2005GL023991.

Gelman, A., Carlin, J.B., Stern, H.S. & Rubin, D.B., 2003. *Bayesian Data Analysis,* 2nd edn, Chapman & Hall/CRC, New York, NY.

Gross, S. & Rundle, J.B., 1998. A systematic test of time-to-failure analysis, *Geophys. J. Int.,* **133,** 57–64.

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

Holliday, J.R., Rundle, J.B., Tiampo, K.F. & Turcotte, D.L., 2006a. Using earthquake intensities to forecast earthquake occurrence times, *Nonlin. Process. Geophys.,* **13,** 585–593.

Holliday, J.R., Rundle, J.B., Turcotte, D.L., Klein, W., Tiampo, K.F. & Donnellan, A., 2006b. Space-time clustering and correlations of major earthquakes, *Phys. Rev. Lett.,* **97,** 238501, doi:10.1103/PhysRevLett.97.23850.

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

Mason, I.B., 2003. Binary events, in *Forecast Verification,* pp. 37–76, John Wiley, Chichester.

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

Tiampo, K.F., Rundle, J.B., McGinnis, S., Gross, S.J. & Klein, W., 2002a. Eigenpatterns in southern California seismicity, *J. geophys. Res.,* **107**(B12), 2354, doi:10.1029/2001JB000562.

Tiampo, K.F., Rundle, J.B., McGinnis, S., Gross, S.J. & Klein, W., 2002b. Mean field threshold systems and earthquakes: an application to earthquake fault systems, *Europhys. Lett.,* **60**(3), 481–487.

Wilks, D.S., 1995. *Statistical Methods in the Atmospheric Sciences: An Introduction,* Academic Press, London.

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