Geophys. J. Int. (1998) **133,** 57–64

A systematic test of time-to-failure analysis

Susanna Gross and John Rundle*

CIRES, Campus Box 216, *University of Colorado at Boulder, Boulder,* CO 80309, *USA. E-mail: sjg@quake.colorado.edu*

Accepted 1997 October 2. Received 1997 August 18; in original form 1997 May 16

SUMMARY

Time-to-failure analysis is a technique for predicting earthquakes in which a failure function is fit to a time-series of accumulated Benioff strain. Benioff strain is computed from regional seismicity in areas that may produce a large earthquake. We have tested the technique by fitting two functions, a power law proposed by Bufe & Varnes (1993) and a log-periodic function proposed by Sornette & Sammis (1995). We compared predictions from the two time-to-failure models to observed activity and to predicted levels of activity based upon the Poisson model. Likelihood ratios show that the most successful model is Poisson, with the simple Poisson model four times as likely to be correct as the best time-to-failure model. The best time-failure model is a blend of 90 per cent Poisson and 10 per cent log-periodic predictions. We tested the accuracy of the error estimates produced by the standard least-squares fitter and found greater accuracy for fits of the simple power law than for fits of the more complicated log-periodic function. The least-squares fitter underestimates the true error in time-to-failure functions because the error estimates are based upon linearized versions of the functions being fitted.

Key words: earthquake prediction, fitting, seismicity, statistics.

INTRODUCTION

Seismologists have been seeking a reliable basis on which to make forecasts for earthquakes for at least 40 years, and have explored a large variety of observable quantities. Researchers have studied seismic phase velocities (Kazi 1990), seismic anisotropy (Crampin, Booth & Evans 1990) and seismic quiescence (Wyss & Weimer 1997; Kisslinger 1988). Fluctuations in the water table (Sato *et al.* 1995), electromagnetic signals (Varotsos, Alexopoulos & Lazaridou 1993), and chemical tracers (Wakita 1996) have also been studied. Advocates of these methods have reported changes in many of these potential predictors before large earthquakes, but have not proven statistically significant patterns for any of them.

Minster & Williams (1995) have carried out careful statistical analyses of the proposed earthquake prediction method M8. Kagan & Jackson (1991) have evaluated the seismic gap hypothesis. Wyss *et al.* (1996) and Kagan *et al.* (1996) have made assessments of the significance of electromagnetic signals. Aster *et al.* (1996) have evaluated changes in shear-wave splitting. Generally, advocates of the method have criticized statistical assessments of earthquake prediction. Advocates dislike the requirement that prediction methods be rigidly defined and applied when tested. Tests have seldom detected any predictive skill in the proposed methods. Careful statistical analysis of earthquake prediction methods is an essential first step before the application of any earthquake prediction method, and also the best way for the method to gain acceptance in the wider scientific community.

Time-to-failure analysis is different from most approaches to earthquake forecasting because it is based upon widely available seismicity data, and the technique has a well-defined methodology. Time-to-failure analysis is much simpler than the pattern-recognition techniques applied to similar data (Keilis-Borok, Knopoff & Kossbokov 1990). Accumulated Benioff strain from a seismically active region is fitted to an increasing function. One such function is a power law, a time-inverted version of the modified Omori relation (Utsu 1961), the classic model of aftershocks. Bufe & Varnes (1993) developed the time-to-failure method in an application to the Loma Prieta earthquake. Sornette & Sammis (1995) proposed an extension of the method, and found that they could get a better fit to the time of occurrence of the Loma Prieta earthquake by fitting a function that included log-periodic fluctuations in the seismicity. Both groups have applied the technique to a variety of other cases, finding increasing Benioff strain before earthquakes in the Virgin Islands (Varnes & Bufe 1996) and the Aleutians (Bufe, Nishenko & Varnes 1994). These earthquakes might have been predictable using the technique.

* Also at: Geological Sciences and Physics Departments, University of Colorado at Boulder, Boulder, CO 80309, USA.

PLAN OF THE WORK

The approach we have taken in our study is to estimate all the factors that quantify the usefulness of the time-to-failure earthquake prediction method in a statistical sense. We have fitted the functions when there was no earthquake to predict, to quantify false alarms. We have counted cases in which no prediction was made, but an earthquake occurred. When false alarms or unpredicted events occur, they reduce the probability that the models are correct. We have also systematically collected cases in which the techniques do successfully predict events. Our statistical measure of how well the models fit the data is a number called the likelihood. Our result is a table of likelihoods that provides estimates of the relative chances that the time-to-failure models of seismicity are correct, given all the observations. For a null hypothesis, we have chosen the Poisson model. We computed likelihoods for combinations of the Poisson model with both kinds of time-to-failure models. These combinations of time-to-failure models with the Poisson model should be sensitive to predictive skill in the time-to-failure models without requiring that they always function. We have explored alternative test catalogues, spatial subsets, magnitude uncertainties and definitions of a prediction. We have studied the fitting method using synthetic data, so that we can find out how dependable the earthquake prediction is without relying entirely upon estimates produced by the fitting algorithm.

THE TIME-TO-FAILURE FUNCTIONS

The power-law fit of Bufe & Varnes (1993) is

$$E(t) = A + B(t - t_f)^m, \tag{1}$$

where $E(t)$ is the cumulative Benioff strain as a function of time t, predicting an earthquake at time t_f, A, B and m are arbitrary constants. Benioff strain is proportional to $10^{0.75m_b}$, where m_b is body-wave magnitude. Small-magnitude events dominate sums of Benioff strain, but less than they dominate sums of earthquake numbers.

Sornette & Sammis (1995) have fitted a log-periodic function,

$$E(t) = A + B(t_f - t)^m \left[1 + C \cos\left(2\pi \frac{\log(t_f - t)}{\log \lambda} + \Psi \right) \right], \tag{2}$$

to similar data, where A, B, C, m, λ and Ψ are arbitrary constants.

We have fitted both the power-law and the log-periodic functions to the original Bufe & Varnes data set, to synthetic sequences and to global data for the systematic test.

THE NON-LINEAR LEAST-SQUARES FITTING METHOD

The models we present were fitted using a Levenberg–Marquardt fitter that we modified from the one presented by Press *et al.* (1992). This algorithm finds the closest fit of a non-linear function to a set of data, as defined by the smallest chi-square. The Levenberg–Marquardt method is a combination of steepest decent optimization, used far from the best fit, and the inverse Hessian method, used near the best fit. We chose the Levenberg–Marquardt method because it was used by Sammis, Sornette & Saleur (1996) and because it formalizes the minimization of chi-squares that Bufe & Varnes (1993) used in their original paper. We have also tried applying the maximum-likelihood technique used in Gross & Kisslinger (1994) for fitting aftershock decay models. The maximum-likelihood fitter could not reproduce the results of either Bufe & Varnes (1993) or Sornette & Sammis (1995), so we have concentrated instead on the chi-square technique.

We ran the basic Levenberg–Marquardt fitting algorithm several hundred times using different starting parameters to ensure that the best fit was found for each model. We selected a widely spaced grid of initial parameters and minimized the chi-square of the difference between observed and modelled Benioff strain until the fitting algorithm had converged to a set of best-fitting model parameters. A new grid of initial parameters was then chosen, more tightly spaced about the best chi-square found thus far. Every fit of the power-law function included 15 nested grids of initial parameters with 375 minimization procedures. The fits of the log-periodic function involved an additional 264 minimizations arranged in a set of nine nested grids of initial parameters, centred on the best-fitting power-law parameters. Our extended fitting algorithm ran without any human intervention, and effectively found the best-fitting model as defined by the chi-square. Our first test of the fitting technique was on the sequence preceding Loma Prieta.

THE LOMA PRIETA CASE

The sequence of activity several decades before Loma Prieta is the classic case for which time-to-failure analysis seems to 'predict' the main shock. Bufe & Varnes (1993) started studying the build-up of activity before the main shock occurred. The data set is useful for comparisons between fitting methods as well. Fig. 1 shows the original Loma Prieta data set, derived from the Ellsworth *et al.* (1981) catalogue used by Bufe & Varnes, and two fits to the cumulative Benioff strain we computed using the modified Levenberg–Marquardt method. Both fits accurately predict the time of the Loma Prieta main shock, but the log-periodic function is slightly more accurate in its prediction of the failure time than the power-law function is. The reduced chi-squares for the fits in Fig. 1 are significantly smaller than one, which means that the fits are much better than expected based on the magnitude uncertainties, assumed to be 0.3 units. We would have to assume an unrealistic magnitude uncertainty of 0.03 to bring the goodness of fit into agreement with the error estimate. The number of events in the sequence (31) is significantly greater than the numbers of free parameters in the models (four or seven), so an under-constrained model seems unlikely. One possible explanation for the anomalously low reduced chi-squares in the time-to-failure fits arises from an incompatibility of the chi-square measure with the statistical properties of Benioff strain data. Accumulated Benioff strains are not statistically independent of one another, and so produce a smoother curve than independent measures of Benioff strain would. A smooth curve is much more readily fitted with a smooth function, and so the scatter of the points about the fit is much less than we expect based upon reasonable estimates of their uncertainty. The chi-square can still be used to fit Benioff strain data, but the scatter of the fit is a poor measure of the uncertainty. We have not used error estimates based upon the chi-square in any of our tests; instead we derived the uncertainty from the

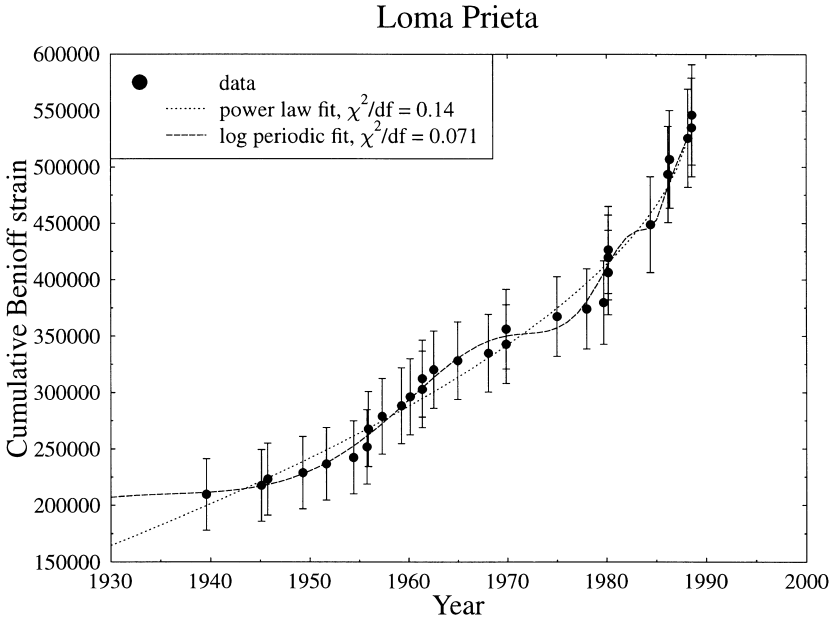

Figure 1. A plot of accumulated Benioff strain as a function to time before the Loma Prieta earthquake. The two smooth curves represent the two types of time-to-failure models. Error bars are assigned on the assumption that magnitudes are accurate to 0.3 units.

assumption that magnitudes can be reliably measured to within 0.3 units.

The Loma Prieta case is a useful example with which to explore the effects of magnitude uncertainties upon the time-to-failure fitting methods. To quantify the effects of magnitude uncertainties, we added Gaussian perturbations to the magnitudes of the earthquakes in the Loma Prieta data set. We then fitted the data with the log-periodic function and observed the scatter in the best-fitting failure times as we varied the perturbations. The failure times did not show any systematic bias, remaining similar to the true failure time in late 1989. The scatter in the fits increased smoothly, ranging from nine months to 6.6 years as the perturbations increased from 0.1 to 0.4 units. With magnitude perturbations of 0.3 units, the Benioff strain typically varies by 68 per cent, but the failure time remains stable within five years. These simulations suggest that the technique is robust enough to apply to catalogues with magnitude uncertainties of several tenths of a unit.

A TEST WITH SYNTHETIC DATA

We tested the accuracy and uncertainties of the failure times produced by the Levenberg–Marquardt fitting method by applying it to synthetic data. We fit 200 synthetic sequences of 100 events each with known log-periodic parameters similar to the Loma Prieta model, and a failure time of 1990. The times of the synthetic earthquakes were evenly spaced and the magnitudes were derived from the log-periodic function. The magnitudes included additional Gaussian errors of known width σ_E, set to either 1 or 10 per cent of the total Benioff strain so we could assess the accuracy of the error estimates. An error in Benioff strain of 1 per cent corresponds to a magnitude uncertainty of 0.006 units, and a 10 per cent error in E is equivalent to a magnitude uncertainty of 0.06 units. Because the errors in the synthetic data sets were independent Gaussian distributions, the sequences were not as smooth as real data, and fits commonly resulted in reduced chi-squares close to one.

Table 1 summarizes the results of our tests of the fitting algorithm, with both choices of scatter in the synthetic data and both possible functions fitted. The table shows that the fitting method works, because the average failure time estimated from the fits is close to the true failure time of 1990. However, the true misfit in failure time estimated from the variation in fitted failure times (last column in Table 1) is greater than the estimates of misfit obtained from the algorithm. Error estimates from the log-periodic fits (19 years) are smaller than the estimates from fits of the power-law (58 years) for the case with 10 per cent error. The true scatter of log-periodic fitted failure times (69 years) is not much less than the corresponding scatter of the power-law fits (73 years).

One explanation for the inaccuracy of the error estimates that come from the Levenberg–Marquardt fitter is visible in Figs 2 and 3. The Levenberg–Marquardt fitter assumes that errors in the fit will be dominated by linear variations in chi-square close to the solution. Models with significantly different failure times but remarkably similar chi-squares appear as

Table 1. Fits of synthetic data.

σ_E	Model	Error est.	Average time
1% E	power-law	2.0 ± 0.9	1988.9 ± 4.6
1% E	log-periodic	2.4 ± 1.8	1989.5 ± 4.3
10% E	power-law	58 ± 166	2006 ± 73
10% E	log-periodic	19 ± 65	2001 ± 69

Error estimates and fitted failure times for sets of 100 synthetic sequences fitted by either the log-periodic function of the power law. The sequences have Gaussian errors equal to 1 or 10 per cent of the accumulated Benioff strain, E, and all have a true failure time in 1990. The reduced chi-squares are not shown and were very close to one, as expected for good fits.

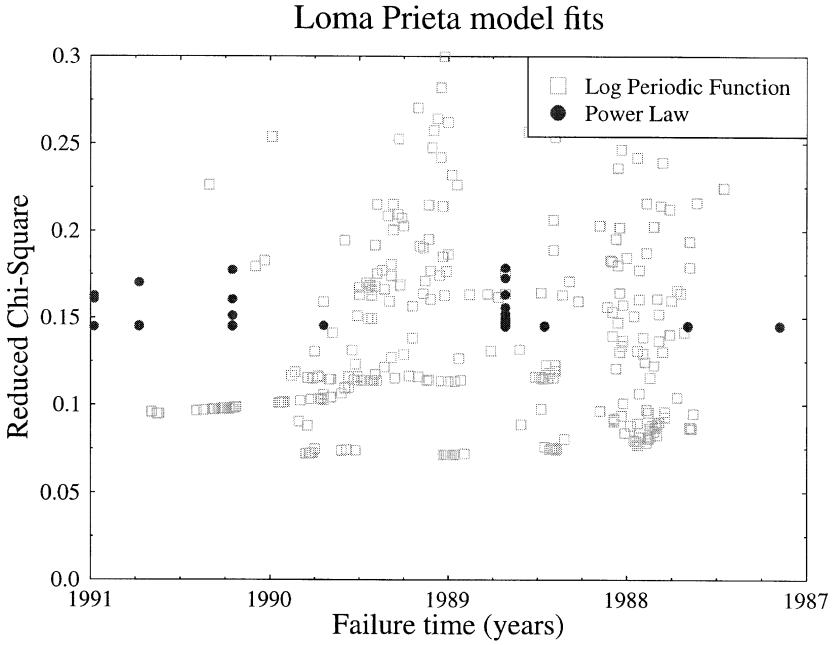

Figure 2. A plot of reduced chi-square versus failure time for a wide variety of models of the Loma Prieta sequence. Clumping of the points around several low values shows how nearly equally good solutions are distributed in failure time. Values of the reduced chi-square much less than one indicate an unexpectedly good fit.

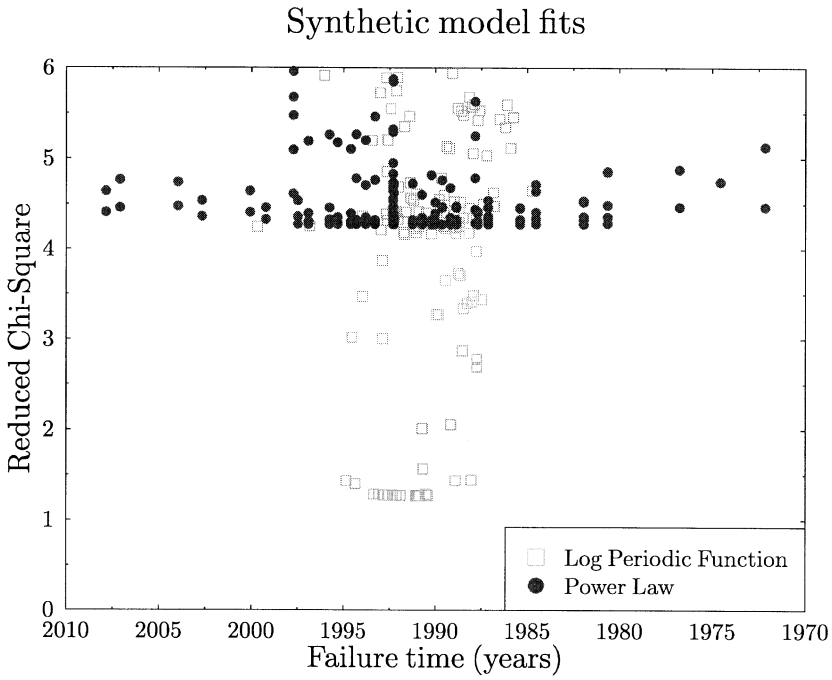

Figure 3. A plot of reduced chi-square versus failure time for a wide variety of models of synthetic data for which the error is 1 per cent of the accumulated Benioff strain. Because this case has a broad minimum in chi-square, the width of the minimum is a good measure of the fitting uncertainty. Values of the reduced chi-square approximately equal to one are expected for a good fit.

well-separated clusters of points in Fig. 2. Changes in chi-square near the solution are not a good measure of the true uncertainty. The case from the synthetic data shown in Fig. 3 has less scatter and a much broader minimum, whose width is a better measure of the true uncertainty. These tests of the Levenberg–Marquardt fitter show that non-linearity of the failure functions results in underestimates of the misfit, even when the statistical properties of the data being fit are consistent with the assumptions of the fitting algorithm.

CONSTRUCTING A LIKELIHOOD RATIO

The statistical measure we use to summarize the effectiveness of models fitted to many sequences of seismicity is the

likelihood, L. The likelihood L of a set of N observed earthquakes occurring at times t_i can be computed from the probability $P(t_i)$ of an event occurring,

$$L = \prod_{i=1}^{N} P(t_i), \tag{3}$$

if the earthquakes are statistically independent. We use the likelihood to decide which model best represents a set of data. The model with the greatest likelihood has the greatest probability of being correct, given that the earthquakes occurred. The models are combinations of the best-fitting power-law or log-periodic model with the Poisson model. The Poisson model assumes that the rate of earthquakes λ_i depends on location only. In the following expression, r is the weight given to the Poisson model, and $1-r$ is the weight given to the power-law or log-periodic model:

$$P(t_i) = Fr\lambda_i + \frac{G(1-r)}{\sqrt{\pi}\sigma_f} \exp\left(-\frac{(t_i - t_f)^2}{\sigma_f^2}\right). \tag{4}$$

The normalization factors F and G can be found by integrating the probabilities of the models over time periods T_j including both data and models. The Poisson model probability density is integrated using the expression

$$F \equiv \frac{N}{\sum_j \lambda_j T_j}, \tag{5}$$

and the power-law or log-periodic fit to the failure time t_f is integrated with

$$G \equiv N \Big/ \int_{T_j} \frac{1}{\sqrt{\pi}\sigma_f} \exp\left(-\frac{(t-t_f)^2}{\sigma_f^2}\right) dt. \tag{6}$$

The normalization ensures that the probability density of each class of model integrates to the total number N of earthquakes observed in all the intervals of length T_j for which the fitting converged. The predicted failure time t_j from the time-to-failure fit is assumed to have a Gaussian distribution with a standard deviation computed from the error estimated by the Levenberg–Marquardt method. Because the tests with synthetic data showed that the fitted error underestimates the true error, we assumed the true error was twice as great as the error reported by the Levenberg–Marquardt method. Likelihoods are very small numbers, so we report the log of the likelihood. The log-likelihood ratio is the difference between log likelihoods. Likelihoods should be interpreted only by comparing models fitted to exactly the same data.

In summary, evaluating a likelihood requires fits to the seismicity giving failure times t_f and their uncertainties, σ_f. Rates of background activity determine the λ_js for all periods for which the fits converged, and then the models may be normalized by summing the probabilities over the test intervals, and the likelihoods computed from eq. 3.

DATA FOR THE SYSTEMATIC TEST

We used the global DNAG catalogue (Engdahl & Rinehart 1991) as a test set for estimating failure times. The test is constructed as if a set of predictions had been made in 1985 when the DNAG catalogue ended, and then the predictions were tested for the following 10 years. We divided events from the DNAG catalogue of $m_b \geq 5$ and above occurring after 1960 into spatial subsets (Fig. 4) and fitted those sets with at least 20 events with both the log-periodic and the power-law functions. Fig. 5 shows an example of a fit from the systematic test, in which an increasing rate of seismicity results in predicted failure times early in the test interval, although no earthquakes occurred. We used two different catalogues, the NEIC and the Harvard catalogues, to define the 'truth' or observed set. We computed likelihoods for alternative definitions of a prediction and the magnitude uncertainty.

The definition of arbitrary spatial and temporal limits for the testing data is one respect in which our test differs from prior studies of time-to-failure analysis. We have applied the test to arbitrary spatial volumes because we wish to learn if time-to-failure analysis shows promise as an earthquake prediction method. Consequently, we cannot use the events we wish to predict in any way when making the predictions. Bufe & Varnes (1993) have introduced a regional classification scheme because some regions exhibit decelerating seismicity before large events instead of accelerating seismicity. They suggest that the time-to-failure prediction method is only applicable to those regions that exhibit accelerating seismicity before main shocks. Their reasonable suggestion presents a practical difficulty. The regional classification is based upon application of time-to-failure analysis to previous seismic cycles in each region. Our data set is too short to allow us to classify

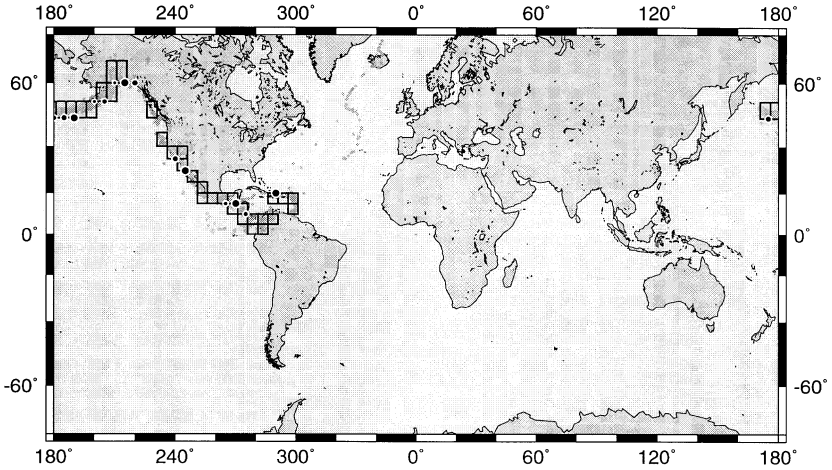

Figure 4. The map shows both the circles and the non-overlapping boxes which contained enough seismicity to allow both models to converge.

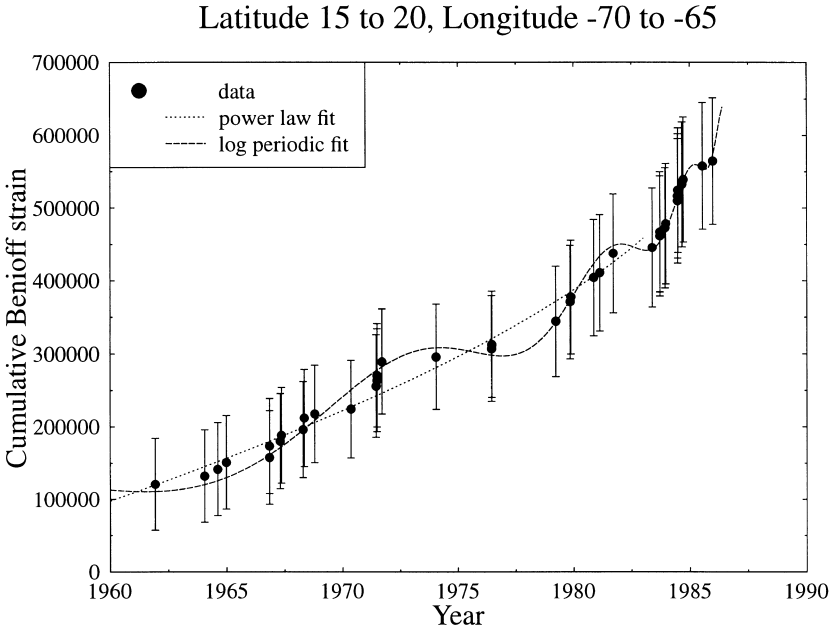

Figure 5. This example of time-to-failure fits drawn from the systematic test illustrates a 'false alarm' because an earthquake was predicted in 1986 by the power-law fit, and in 1987 by the log-periodic fit, but no event occurred during the test period, 1985–1995.

the regions without using events we are trying to predict in the classification. If data past the end of the test set contributed to the regional classification scheme, we could not fail to enhance the apparent effectiveness of the prediction method. We would be eliminating exactly those cases for which the prediction failed, and would completely invalidate our test. Although we apply the technique to regions for which it may not be appropriate, our test should still be quite sensitive to the effectiveness of time-to-failure analysis in predicting some earthquakes. We only assume that the time-to-failure fits do not converge to answers that are systematically worse than chance in the cases for which the analysis does not apply.

The models sometimes failed to converge; we decided these were unusable cases. If the models converged to a date well separated from the test interval, either before 1970 or after 2015, the fit might imply a prediction of no earthquake or no prediction at all. If we count the cases with predictions far from the test interval as no prediction, then the likelihoods should be computed without any contribution from cases with predictions far from the test interval. The sixth column of Table 2, labelled 'near fits' is an example of likelihoods computed without including a contribution from fits that converged far from the test interval. It included 427 observed earthquakes in 302 intervals. Defining the cases with failure times far outside the test interval as predictions of no earthquake made more data available (946 observed earthquakes in 953 intervals for column seven), so predictions far from the test interval are included for all the other cases in Table 2.

The first three columns of results in Table 2 list log likelihoods for suites of models computed using the observed occurrence of $m_b \geq 6.5$ earthquakes in the NEIC catalogue as a truth set. Events with $M_S \geq 6.5$ from the Harvard catalogue

Table 2. Log likelihoods.

	NEIC truth set			Harvard truth set		
Model	One grid	25 grids	circles	all fits	near fits	M ± 1.0
100% Poisson	−44.33	**−1650**	**−26.96**	**−2238**	**−866**	**−2238**
10% power-law	−42.33	−1692	−27.69	−2304	−901	−2279
30% power-law	**−41.05**	−1799	−29.45	−2477	−988	−2408
50% power-law	−41.84	−1948	−31.81	−2717	−1105	−2603
70% power-law	−44.75	−2173	−35.38	−3081	−1282	−2907
90% power-law	−52.90	−2643	−43.07	−3832	−1647	−3524
99% power-law	−69.67	−3537	−59.19	−5221	−2356	−4520
100% power-law	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$
10% log-periodic	−44.42	−1680	−22.92	−2281	−894	−2283
30% log-periodic	−45.40	−1767	−24.67	−2414	−965	−2412
50% log-periodic	−47.53	−1893	−27.01	−2611	−1066	−2600
70% log-periodic	−51.56	−2087	−30.54	−2919	−1220	−2885
90% log-periodic	−61.22	−2497	−38.04	−3562	−1541	−3450
99% log-periodic	−82.00	−3274	−52.96	−4732	−2165	−4317
100% log-periodic	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$	$-\infty$

provided the truth set for the last three columns. The percentage of time-to-failure model from the first column of Table 2 is equal to $1-r$ from eq. 4, and the table shows how well the various mixtures of Poisson with log-periodic or power-law models fit the data. Three different choices of the spatial domains over which the test was carried out are represented by the first three columns of results in Table 2. The second column of Table 2 lists likelihoods for spatial subsets drawn from a non-overlapping grid of $5^\circ \times 5^\circ$ boxes covering the whole Earth and having a vertex at 0° latitude, 0° longitude. The single-grid spatial subsets included 21 events in the 34 intervals for which both models converged to some failure time, so the second column lists results from a fairly modest number of events in independent data sets. Column three shows combined results from 25 such grids, offset in latitude and longitude by whole degrees ranging from zero to four. The tests constructed from overlapping grids are not fully independent, but they do provide larger numbers for statistical robustness, with 566 observed quakes in 953 intervals. The tests using the Harvard catalogue are also based upon data from 25 overlapping grids.

The fourth column lists results for a set of circles whose radii were optimized to make the cumulative number of events least linear. Sammis (personal communication, 1997) proposed the optimization as a way of finding the spatial subsets which show the strongest changes in Benioff strain rate. Since the optimization does not use any information about the seismicity being predicted, the optimization is permissible in a test of earthquake prediction. Circles ranged in radius from 50 to 190 km were centred on the points at which the non-overlapping $5^\circ \times 5^\circ$ grid lines intersected. Only a single circle radius was selected for each centre point, so the results in the fourth column are statistically independent. The circular spatial cuts have the smallest numbers, so we relaxed the requirement that both models converge, giving up the ability to compare log-periodic and power-law fits. There were seven observed quakes in 14 circles for which the log-periodic model converged, and seven quakes in 19 circles for which the power-law converged.

The least negative numbers, shown in bold type in Table 2, represent the best models in each set. The log-periodic likelihoods can be directly compared with the power-law likelihoods within a column for all the cases except the circles, because the sets of data for which predictions were made are the same. The best-fitting model for the independent grids in the second column of Table 2 is the case with 30 per cent power-law and 70 per cent Poisson. Of the end-member models, the Poisson model is clearly the best, because it never reports a zero probability of an event, while the other models can. The models which combine the peaked distributions of the time-to-failure fits with a uniform background from the Poisson model are designed to detect any predictive skill in the time-to-failure models, while not requiring that they function well in all cases. Combining the Poisson model with fits from the power-law function does not improve the Poisson model's ability to predict the seismicity in any cases except for that mentioned above. The log-periodic model does not improve the ability of the Poisson model to predict seismicity for any of the cases studied, but it produces higher likelihoods than the power-law model for most cases.

As a final step in interpretation of the results in Table 2, we can convert the differences in log likelihoods, log L, into ratios of probabilities that two models are correct given the data. The formula simply relates probabilities and likelihoods for models designated A and B,

$$\frac{P_A}{P_B} = \exp(\log L_A - \log L_B). \tag{7}$$

The formula must be applied cautiously, to fully independent data and to models that have equal *a priori* probability of being correct. The requirement that the data sets be independent is violated by the overlapping data sets in Table 2. Results from the two independent data sets in the table are contradictory, with the grids suggesting the 30 per cent power-law model is 26 times as good a model as the simple Poisson. However, the results from the optimized circles suggest the Poisson model is 12 times more likely to be correct than the model with 30 per cent power-law. The Poisson model has fewer parameters than the power-law, so its *a priori* probability of being correct may be greater. The particular choice of spatial grid used to construct the test in column two of the table is only one of many possibilities. Those possibilities are summarized in the cases computed using overlapping grids. Log likelihoods from cases with overlapping grids equal the sum of log likelihoods for 25 independent grid cases. The overlapping grid cases can be used to compute the difference in log likelihoods averaged over all 25 choices of the grid origin. Likelihoods computed using alternative choices of catalogue, magnitude uncertainty, and definition of a prediction are all valid estimators of the likelihood, and are equivalent to 100 different cases. An average of the 100 cases gives a log likelihood for the pure Poisson model of -69.92. Similarly, the average log likelihood for the best hybrid model is -71.38, giving a probability ratio of 4.3. The probability ratio implies that the simple Poisson model is four times more likely to be correct than the best hybrid model. The best hybrid model is composed of a blend of 90 per cent Poisson predictions with 10 per cent log-periodic predictions. Greater advantages of Poisson over time-to-failure models results from cases including a greater proportion of time-to-failure predictions.

CONCLUSIONS

Our systematic test has found that generally the time-to-failure models are less effective at predicting earthquakes than the pure chance Poisson model is. Combinations of the time-to-failure models with the Poisson model are much better predictors of future activity than the pure time-to-failure models, but only in one case were they actually better than the pure Poisson model. A combination of 30 per cent power-law model with 70 per cent Poisson model appears to be 26 times as likely to be correct than the simple Poisson model when a single grid of $5^\circ \times 5^\circ$ spatial subsets is considered. The apparent skill of the power-law model disappears when different choices of the grid, catalogue, prediction definition, and magnitude uncertainty are explored. In aggregate, the most successful time-to-failure earthquake model we evaluated is a blend of 90 per cent Poisson and 10 per cent log-periodic model. The 100 per cent Poisson model still has a four times greater probabilty of generating the observed seismicity than the most successful time-to-failure model.

Both the power-law and log-periodic functions do a good job of 'predicting' the time of the Loma Prieta earthquake. We also found that the time-to-failure functions fit the accumulated

Benioff strain much better than typical magnitude uncertainties would lead us to expect, and so the fits have a reduced chi-square much less than one. The low chi-square is a result of the smoothness of the accumulated Benioff strain curve, and imples that scatter of the data about the fit should not be used as a measure of the uncertainty.

Tests with synthetic seismicity show that estimates of the error in the time-to-failure from the Levenberg–Marquardt algorithm are generally less than the true errors. Underestimates of the fitting error were worst for fits of the log-periodic function in cases with significant scatter in the data. Inaccuracies in error estimates result from non-linearities of the functions. The Levenberg–Marquardt fitting method assumes that the function being fit may be linearized when estimating errors in the parameters.

ACKNOWLEDGMENTS

Paul Reasenberg and an anonymous reviewer provided insightful reviews. Charles Bufe kindly gave us the Loma Prieta data set and a thoughtful review. Dr Steven Ward insisted on more active and simplified writing style. This work was supported by DOE grant DE-FG03-95ER14499 and NSF grant EAR-9526814.

REFERENCES

Aster, R., Slad, G., Henton, J. & Antolik, M., 1996. Differential analysis of coda Q using similar micro earthquakes in seismic gaps; Part 1, Techniques and application to seismograms recorded in the Anza seismic gap, *Bull. seism. Soc. Am.,* **86,** 868–889.

Bufe, C.G. & Varnes, D.J., 1993. Predictive modeling of the seismic cycle of the greater San Francisco Bay region, *J. geophys. Res.,* **98,** 9871–9883.

Bufe, C.G., Nishenko, S.P. & Varnes, D.J., 1994. Seismicity trends and potential for large earthquakes in the Alaska–Aleutian region, *Pure appl. Geophys.,* **142,** 83–99.

Crampin, S., Booth, D.C. & Evans, R., 1990. Changes in shear wave splitting at Anza near the time of the North Palm Springs earthquake, *J. geophys. Res.,* **95,** 11 197–11 212.

Ellsworth, W.L., Lindh, A.G., Prescott, W.H. & Herd, D.G., 1981. The 1906 San Francisco earthquake and the seismic cycle, in *Earthquake Prediction: An International Review, Maurice Ewing Ser.,* vol. 4, pp. 126–140, eds Simpson D.W. & Richards, P.G., AGU, Washington, DC.

Engdahl, E.R. & Rinehart, W.A., 1991. Seismicity map of North America project, in *Neotectonics of North America,* pp. 21–27, GSA, Boulder, CO.

Gross, S.J. & Kisslinger, C., 1994. Tests of models of aftershock rate decay, *Bull. seism. Soc. Am.,* **84,** 1571–1579.

Kagan, Y.Y. & Jackson, D.D., 1991. Seismic gap hypothesis; ten years after, *J. geophys. Res.,* **96,** 21 419–21 431.

Kagan, Y.Y., Jackson, D.D., Varotsos, P. & Lazaridou, M., 1996. Statistical test of VAN earthquake predictions; comment and reflections [discussion and reply], *Geophys. Res. Lett.,* **23,** 1433–1440.

Kazi, W.M., 1990. Velocity changes in the central Aleutian Islands determined from the analysis of earthquake doublets using cross-spectral analysis method, *PhD thesis,* University of Colorado, Boulder, CO.

Keilis-Borok, V.I., Knopoff, L. & Kossbokov, V., 1990. Intermediate-term prediction in advance of the Loma Prieta earthquake, *Geophys. Res. Lett.,* **17,** 1461–1464.

Kisslinger, C., 1988. An experiment in earthquake prediction and the 7 May 1986 Andreanof Islands earthquake, *Bull. seism. Soc. Am.,* **78,** 218–229.

Minster, J.B. & Williams, N.P., 1995. Worldwide performance of a seismicity-based intermediate-term prediction algorithm; the M8 algorithm, $M > 7.5$, 1985–94, *EOS, Trans. Am. geophys. Un.,* **76,** 358–358.

Press, W.H., Flannery, B.P., Teukolsky, S.A. & Vetterling, W.T., 1992. *Numerical Recipes in C, The Art of Scientific Computing,* 2nd edn, Cambridge University Press, Cambridge.

Sammis, C.G., Sornette, D. & Saleur, H., 1996. Complexity and earthquake forecasting, in *Reduction and Predictability of Natural Disasters,* pp. 143–156, eds Rundle, J., Turcotte, D. & Klein, W., Santa Fe Institute Proceedings, **25.**

Sato, T., Matsumoto, N., Takahashi, M. & Tsukuda, E., 1995. Ground water changes related to the 1995 Kobe earthquake *EOS, Trans. Am. geophys. Un.,* **76,** 378.

Sornette, D. & Sammis, C.G., 1995. Complex critical exponents from re-normalization group theory of earthquakes: implications for earthquake predictions, *J. Phys. I France,* **5,** 607–619.

Utsu, T., 1961. A statistical study on the occurrence of aftershocks, *Geophys. Mag.*, **30,** 521–605.

Varnes, D.J. & Bufe, C.G., 1996. The cyclic and fractal seismic series proceeding an m_b 4.8 earthquake on 1980 February 14 near the Virgin Islands, *Geophys. J. Int.,* **124,** 149–158.

Varotsos, P., Alexopoulos, K. & Lazaridou, M., 1993. Latest aspects of earthquake prediction in Greece based on seismic electric signals, II, *Tectonophysics,* **224,** 1–37.

Wakita, H., 1996. Geochemical challenge to earthquake prediction, *Proc. Nat. Acad. Sci.,* **93,** 3781–3786.

Wyss, M. & Wiemer, S., 1997. Two current seismic quiescences within 40 km of Tokyo, *Geophys. J. Int.,* **128,** 459–473.

Wyss, M., Allmann, A., Varotsos, P., Eftaxias, K., Lazaridou, M., Dologlou, E. & Hadjicontis, V., 1996. Probability of chance correlations of earthquakes with predictions in areas of heterogeneous seismicity rate; the VAN case [discussion and reply], *Geophys. Res. Lett.,* **23,** 1302–1314.