

# Paleoseismic interevent times interpreted for an unsegmented earthquake rupture forecast

Tom Parsons<sup>1</sup>

Received 8 May 2012; revised 8 June 2012; accepted 8 June 2012; published 10 July 2012.

[1] Forecasters want to consider an increasingly rich variety of earthquake ruptures. Past occurrence is captured in part by paleoseismic observations, which necessarily see three-dimensional ruptures only at a point. This has not been a problem before, because forecasts have assumed that faults are segmented, and that repeated ruptures occur uniformly along them. A technique is now required to calculate paleo-earthquake rates at points that may be affected by multiple recurrence processes, and that is consistent with an all-possible-ruptures forecast. Dating uncertainties are addressed by bootstrapping across event time windows, and the resulting distributions are transformed into log space as  $f(\ln(T))$  where  $T$  is interevent time. This takes advantage of a property of time-dependent recurrence distributions in which their logarithms are normally distributed. Paleoseismic series necessarily have a finite number of observations such that the true long-term mean interevent time ( $\mu$ ) is hard to estimate. However the mode (most frequent value) is easier to identify. Since the mode is equal to the mean of a normal distribution,  $\mu$  can thus be found at the mode ( $m$ ) of  $f(\ln(T))$  as  $\mu = e^m$ . The point  $\mu - \sigma$  occurs where 32% of a folded (half) normal distribution is found in the interval between  $\ln(T) = 0$  and  $m$ . The  $\mu + \sigma$  value is identified by symmetry, which overcomes the difficulty of absent long intervals in the record. Tests are conducted with complex synthetic interevent distributions, and applications to real data from the Hayward and Garlock faults in California are shown. **Citation:** Parsons, T. (2012), Paleoseismic interevent times interpreted for an unsegmented earthquake rupture forecast, *Geophys. Res. Lett.*, 39, L13302, doi:10.1029/2012GL052275.

## 1. Introduction

[2] Paleoseismic records of earthquake sequences arise only at some points along faults because they require specific geological conditions such as a continuous and datable sedimentary record that is thick enough to capture multiple earthquake disturbances, and that is deposited near an active fault. Because of this, many faults have limited or irregular spatial paleoseismic coverage. Thus paleo-sites are essentially a point process, and like earthquake epicenters, often cannot reveal much information about rupture dimensions or variability. However, they provide essential empirical mean earthquake rates that are crucial to seismic hazard assessment.

[3] If a fault is assumed to behave according to a characteristic earthquake model [e.g., *Schwartz and Coppersmith*,

1984; *Wesnowsky*, 1994], where fault segments repeatedly rupture segments in a similar fashion, then one paleoseismic site along a fault segment is representative of that segment's earthquake recurrence distribution. However, if one hypothesizes a broader range of possible ruptures [e.g., *Field and Page*, 2011] that overlap, branch, or that have variable magnitudes [*Weldon et al.*, 2004] with different recurrence distributions, then interpreting paleoseismic information at one point on a segment can become more complicated. Indeed, sites where overlapping ruptures are thought to occur, like Wrightwood and Pallet Creek on the San Andreas fault [*Biasi and Weldon*, 2009], have paleoseismic series that cannot be reproduced by any one recurrence distribution even after  $50 \cdot 10^6$  attempts [*Parsons*, 2008a], signaling a more complex process. Additionally, earthquake sequences may change character, branching into long-term cycles of increased or diminished activity owing to fault interactions [e.g., *Marzocchi and Lombardi*, 2008] that obey different recurrence distributions.

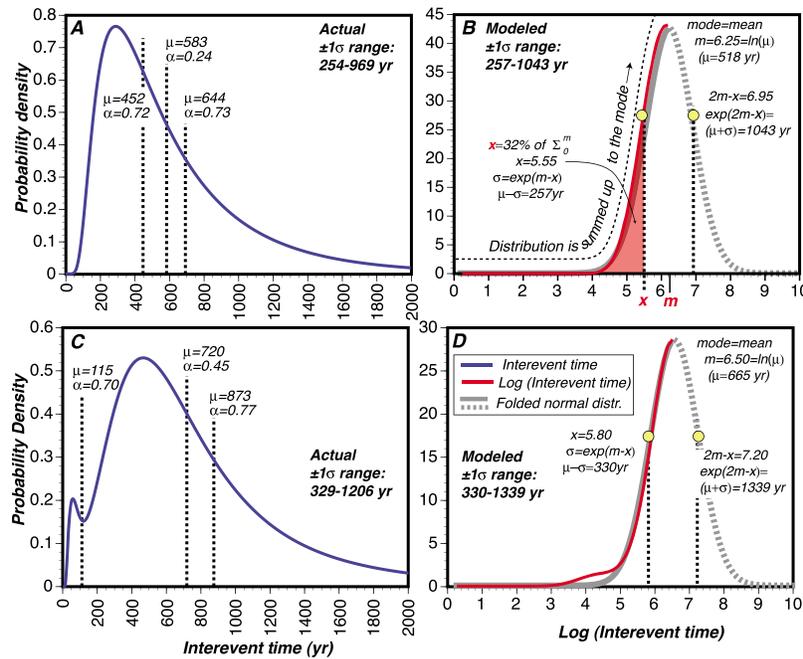
[4] Paleoseismic observations reveal a number of earthquakes above an observable surface slip threshold (assumed proportional to magnitude) in a period. This empirical information is extremely valuable to earthquake forecasters who need a rate to make probability calculations. The variability in earthquake recurrence intervals due to inconsistency in the rupture process, which gives rise to aleatory uncertainty in hazard modeling, is perhaps best represented by paleoseismic observations. Two sources of epistemic uncertainty must be accounted for before a paleoseismic rate constrains a probability calculation: (1) dating uncertainty (usually radiocarbon dating), and (2) the effects of undersampling that can cause a time-limited historical or paleoseismic record to preferentially reflect the shortest intervals and miss the longest ones [*Stein and Newman*, 2004]. Dating uncertainty can be addressed by bootstrapping across the possible event time ranges (sampling a uniform PDF determined by the reported uncertainties) [e.g., *Ellsworth et al.*, 1999; *Biasi et al.*, 2002]. Undersampling has been accounted for by Monte Carlo sampling from long-tailed recurrence distributions [e.g., *Console et al.*, 2008; *Parsons*, 2008a]; this has been necessary because the arithmetic mean of observed interevent times is not likely to represent the true average recurrence because the means of distributions thought to represent earthquake occurrence are all skewed to the right of their modes, and it requires many samples to capture that.

[5] In this paper I present a method to estimate the long-term mean and confidence bounds on the earthquake rate at a point when a segmented and/or characteristic earthquake rupture concept is not assumed. This application is to be explored for the Uniform California Earthquake Rupture Forecast version 3 (UCERF3); prior California forecasts have segmented faults by characteristic ruptures [e.g., *Field*

<sup>1</sup>U.S. Geological Survey, Menlo Park, California, USA.

Corresponding author: T. Parsons, U.S. Geological Survey, Menlo Park, CA 94025, USA. (tparsons@usgs.gov)

This paper is not subject to U.S. copyright. Published in 2012 by the American Geophysical Union.



**Figure 1.** (a) Theoretical example of three different time-dependent recurrence processes affecting a point on a fault displayed as a combined probability density function. Relative amplitudes are governed by coefficients of variation. (b) The natural logarithms of recurrence times  $\ln(T)$  are binned, which are distributed approximately normally, indicating that the sum of lognormal distributions can be interpreted as lognormal. Real observations are expected to be sparse for long interevent times, so the log distribution is summed up to the mode of a folded (half) normal distribution. The  $-1\sigma$  bound is the point  $x$  where 32% of the density of the folded distribution occurs. If the half-normal distribution is reflected across the mode ( $m$ ), then the  $+1\sigma$  bound can be identified by symmetry. (c, d) The same process is shown but for a more complex example with very different recurrence means. More uncertainty is introduced, but reasonable values for the  $\pm 1\sigma$  bounds are found.

*et al.*, 2009]. Results are given as interevent times ( $T$ ) rather than earthquake rates ( $1/T$ ), because  $T$  is more intuitive and more often reported in the paleoseismic literature. The primary result of interest for UCERF3 is the range of allowable interevent times rather than the mean.

## 2. Method

[6] In accordance with UCERF, I assume that earthquake recurrence is time dependent, meaning that a Poisson process is not applicable. Probability density functions commonly thought to represent time dependent earthquake recurrence like the lognormal [e.g., *Nishenko and Buland*, 1987], Brownian Passage Time (inverse Gaussian) [*Kagan and Knopoff*, 1987; *Matthews et al.*, 2002] or Weibull [*Hagiwara*, 1974] are skewed such that the logarithms of their probability density are approximately normally distributed [e.g., *Sachs*, 1984] (Figure 1).

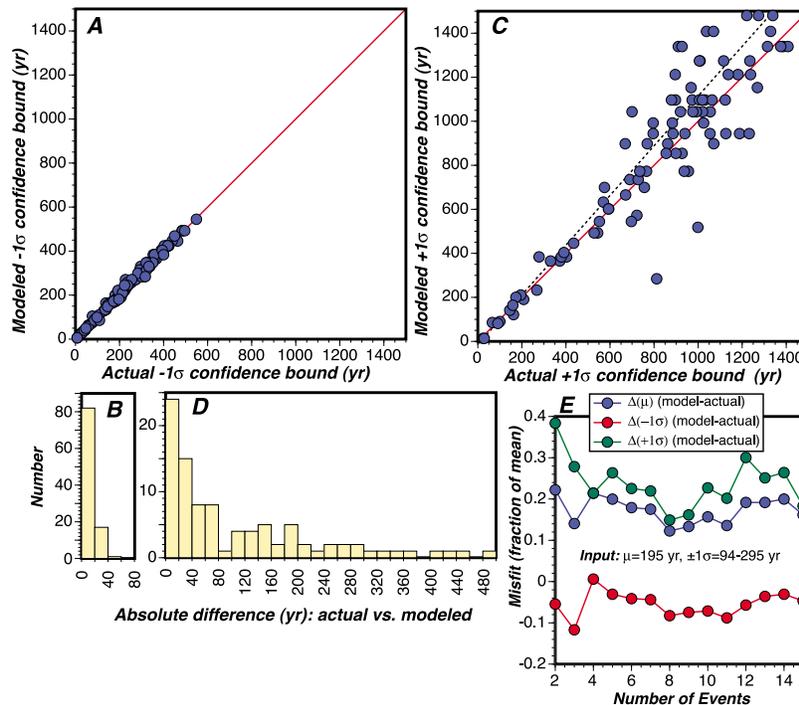
[7] The point ( $m$ ) where the mode of the normal distribution  $f(\ln(T))$  occurs is also its mean (Figure 1). Thus  $e^m$  yields the “true” mean ( $\mu$ ) of the interevent time distribution. The left half of the  $\ln(T)$  distribution can be thought of as a folded, or half normal distribution. Solving for the point  $x$  from

$$\frac{\sum_{\ln(T)=0}^{\ln(T)=x} f(\ln(T))}{\sum_{\ln(T)=0}^{\ln(T)=m} f(\ln(T))} = 0.32 \quad (1)$$

where 32% of the summed value of the folded distribution is located, is also the point where one standard deviation ( $\sigma$ ) on the mean of the complete distribution can be found. This is because 32% of the folded distribution is 16% of the complete distribution, which in turn marks the lower bound of where 68% of the density lies. The value of  $e^{m-x}$  thus represents the  $\mu - \sigma$  bound on interevent time. The variable  $\sigma$  is used here to denote 68% confidence bounds on the mean of the skewed interevent distribution because it is also the standard deviation of the normal  $f(\ln(T))$  distribution. Operationally,  $f(\ln(T))$  is expressed as a histogram, which is summed numerically. Maximum likelihood estimators exist for folded normal distributions, but their calculation and error estimation are “troublesome” according to *Johnson* [1962]. The histogram mode is subject to binning uncertainties, so I make 10 bootstrap calculations for every series to ensure that it is stable.

[8] If there were total sampling of the underlying time dependent distribution, then  $\ln(T)$  would be a complete normal distribution, fully symmetric about its mean. This is however unlikely for most paleoseismic sites unless a very long record is present. An advantage of transforming the data into log space is that symmetry applies. Therefore, if the more complete, left side of the normal distribution  $f(\ln(T))$  is reflected across its mode ( $m$ ), then the  $+1\sigma$  bound on the recurrence interval estimate can be extrapolated to  $e^{(2m-x)}$  (Figure 1).

[9] The primary uncertainty associated with this approach is that it approximates an unknown interevent time distribution,



**Figure 2.** (a) Actual vs. modeled ( $\mu - \sigma$ ) values on interevent time are plotted against each other; the red line shows a slope = 1.0 for reference. “Actual” refers to the variability of the tested distributions. (b) The absolute misfits are shown as a histogram. The misfits are relatively small on the lower ( $\mu - \sigma$ ) bounds. (c) The actual and modeled ( $\mu + \sigma$ ) values are plotted. The dashed black line shows linear fit with a slope of 1.12, suggesting that the method skews a little high. (d) Misfits are greater on the upper ( $\mu + \sigma$ ) bound because of logarithmic binning. (e) Misfits are given as a function of the number of events used to model mean interevent times (from 100 realizations).

or combination of distributions, as generally lognormal. For example, a normal distribution fit to a Weibull or Brownian Passage Time distribution on a log axis could be biased because the Brownian Passage Time has a slightly heavier right tail and the Weibull a slightly heavier left tail. Combinations of different interevent behavior at a single site would also affect the resolution of this technique. I next examine the impacts of these effects.

### 3. Testing With Known Distributions

[10] I generate 100 recurrence distributions at random that are meant to simulate point process behavior in an unsegmented fault system. These are combinations of 1 to 5 Brownian Passage Time distributions, which can have means ranging from 0.1–1000 yr, and coefficients of variation from 0.1 to 1.0. Each is given a normalized random weighting ranging from 0.1 to 1.0 when combined; Figure 1 shows two examples. I choose to combine 5 or fewer distributions because I am most concerned about strongly multimodal distributions. If there are more distributions combined, then their modes tend to get smoothed out.

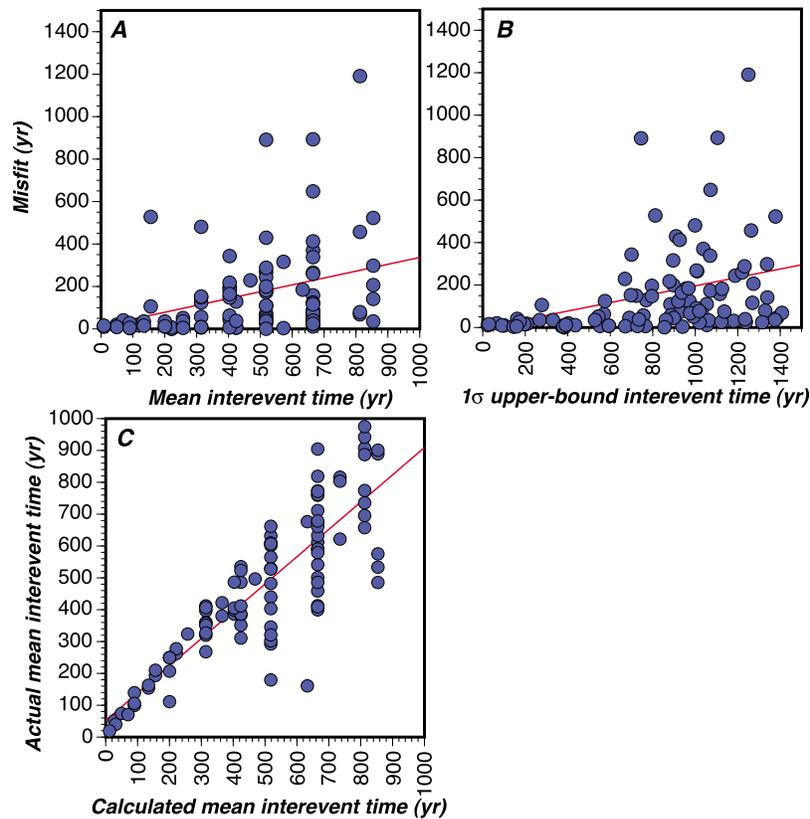
[11] The folded  $\ln(T)$  method as described above is applied to each of the 100 synthetic distributions, and their  $\pm 1\sigma$  ranges are found by counting where 68% of their density occurs. These numbers are compared with calculated values from the proposed method as a way to assess it (Figure 2). The  $\mu - \sigma$  bounds on interevent times are fit fairly well, with maximum magnitude of misfits all less than 50 yr, and the majority less than 20 yr (Figures 2a and 2b).

This isn’t surprising since the method is tuned to fit the  $\mu - \sigma$  bound, and the logarithmic bins of the  $\ln(T)$  histogram are smallest at low  $T$ . The tradeoff from being able to apply Gaussian statistics to  $\ln(T)$  is that each bin has increasing width with increasing  $T$ . When the folded normal distributions of  $\ln(T)$  are reflected across the modes/means, and the  $\mu + \sigma$  bounds on interevent times are extrapolated, the misfits are thus proportionately larger (Figures 2c and 2d), increasing with greater mean interevent times ( $\mu$ ) (Figure 3). However, while  $\mu + \sigma$  misfits are greater, the majority are less than 20 yr.

[12] Real paleoseismic observations are subject to radiocarbon dating uncertainties, and have varying numbers of events. A second test is applied where intervals are drawn at random from an example known distribution (parameters also generated at random:  $\mu = 195$ , coefficient of variation  $\alpha = 0.6$ ). Dating uncertainty bounds of  $\pm 50$  yr are added to each interval, and interval distributions are created by bootstrapping across time windows for each event. The number of included events is systematically reduced from 15 to 2, with misfits being stable down to  $\sim 4$  events (Figure 2e).

### 4. Example Application to Observations: South Hayward and Central Garlock Faults

[13] When the method is applied at California sites, it returns reasonable earthquake rate estimates based on comparison with results from prior methods ( $\pm 1\sigma$  ranges



**Figure 3.** (a) Mean interevent times ( $\mu$ ) are plotted against  $+1\sigma$  misfits, and (b),  $\mu + \sigma$  values are plotted against their misfits. Red lines show linear fits that indicate misfit is a function of increasing interevent time. (c) Calculated  $\mu$  values are plotted against the actual means. If the method worked perfectly all the points would fall on the red line, which would have a slope of 1.0, but instead there is scatter and the slope is 0.9.

encompass past results: see Table S1 in Text S1 of the auxiliary material).<sup>1</sup>

[14] However, in the interest of full disclosure, I show applications to two California paleoseismic series that cover the range from closest to greatest mismatch from past studies. The first series comes from Tyson’s Lagoon, which lies on the south Hayward fault in the San Francisco Bay region of California, and preserves a  $\sim 1900$ -yr record of 11 paleoearthquakes and one historic event in 1868 [Lienkaemper *et al.*, 2010]. Possible intervals are bootstrapped across reported time windows for each event from radiocarbon dating uncertainty as shown in Figure 4a. In this example, 1000 series of 12 events are drawn from within the uncertainty bounds, and their intervals calculated.

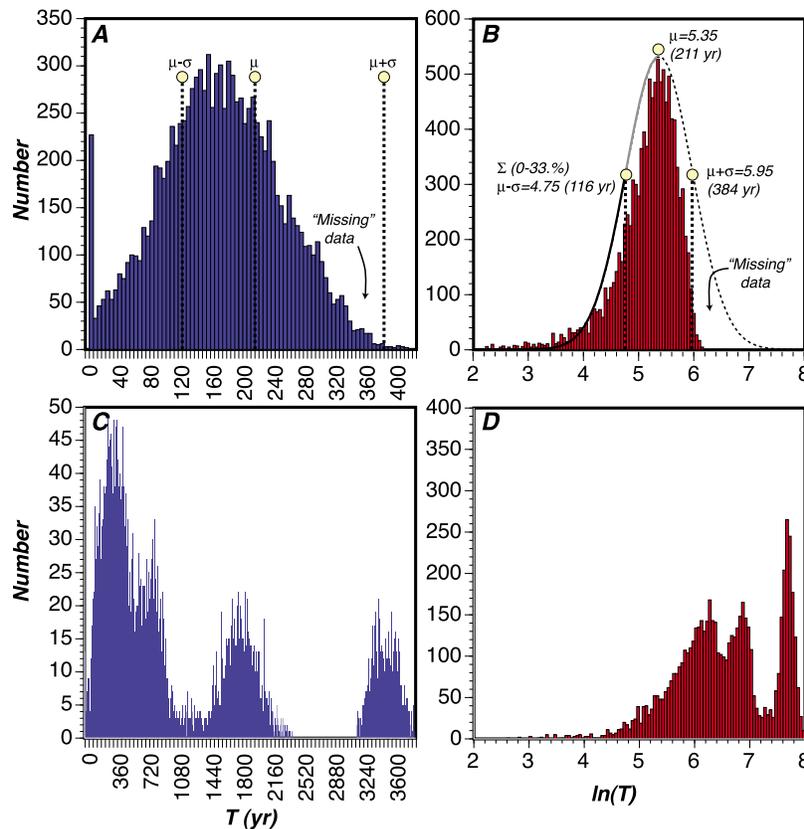
[15] The same procedure was used on the real data as described in Sections 2 and 3. Observations were rebinned by  $\ln(T)$ , and the mode identified (Figure 4b). Equation (1) is solved to find the  $\mu - \sigma$  value. The folded normal distribution in log space is then reflected across the mode, and the  $\mu + \sigma$  value is found through symmetry. The exponentials of these values yield a mean interevent time on the south Hayward fault of 211 yr, and a  $\pm 1\sigma$  range from 116 to 384 yr. Rate values are the reciprocals, yielding 0.0026–0.0086 events/yr, with a mean of 0.0047 events/yr. The mean recurrence interval value of 211 yr is essentially the

same as the 210-yr value calculated using a Monte Carlo technique by Parsons [2008b].

[16] The concern about absent longest intervals seems to be present with the Hayward fault record, as the normal distribution of  $\ln(T)$  is truncated (“missing” data area shown in Figure 4b). This issue is important in cases like UCERF, where most of the calculated earthquake rates come from long-term fault slip rate measurements inferred from offsets of geologic features that have occurred over  $10^3$ – $10^6$  yr periods. Since California paleoseismic constraints mostly represent  $10^2$ – $10^3$  yr scales, their rates need to be consistent with longer-term, fault-slip-rate solutions that may have involved many more earthquake cycles. Further, UCERF has applied time-dependent Brownian Passage Time functions (with their long tail assumptions) for probability calculations; thus the underlying earthquake rate values also need to be consistent with that assumption.

[17] The  $\sim 250$ -km long left-lateral Garlock fault trends roughly east–west across southern California. The El Paso Peaks paleoseismic site lies on the central segment of the fault, and is located in an extensional step over that has been filled by an ephemeral stream. Dawson *et al.* [2003] report on six well-resolved earthquakes that happened during the past 7000 yr. They also note that earthquake occurrence has been very irregular (intervals range from 215 to 3300 yr), which can be seen in Figure 4c. Bootstrapping of the central Garlock intervals results in a distinctly multimodal distribution. Re-binning by natural logarithm puts the mode of the

<sup>1</sup>Auxiliary materials are available in the HTML. doi:10.1029/2012GL052275.



**Figure 4.** Best and worst case results: (a) observed Hayward fault paleoseismic event intervals from *Lienkaemper et al.* [2010] are bootstrapped across reported radiocarbon dating uncertainties. (b) The same information is displayed except binning is by natural logarithm of interevent time. The mode of the folded normal distribution is 5.35 ( $T = 211$ ). This value is interpreted as the mean of the complete distribution, and the  $\pm 1\sigma$  range for Hayward fault interevent times lies between 116–384 years. “Missing” data refers to the longest interevent times that are inferred to exist based on the long-tailed time-dependent recurrence distributions used to make earthquake probability calculations. (c, d) The same method is applied on a strongly multimodal series from the central Garlock fault. Distinct clustering behavior makes it more difficult to interpret this series as having one mean interevent time.

distribution on the very high end, which returns a very high mean interevent time of  $\mu = 3362$  yr. In cases like this, it may be best to consider different modes individually, because of potential rupture mode switching [e.g., *Zöller et al.*, 2007; *Hillers et al.*, 2009], or double branching behavior [*Marzocchi and Lombardi*, 2008] in which the Garlock fault may only be stressed by slip events on the San Andreas fault, rather than directly by plate motions [*Parsons*, 2006]. As a comparative measure, I calculated the fit of the observed intervals to a lognormal distribution using a Kolmogorov-Smirnov (KS) test (see auxiliary material for details). While neither the Hayward nor Garlock series can be confirmed as lognormally distributed at high significance (if they could this paper would not be necessary), the significance level of the Hayward series (55% confidence) far exceeds the central Garlock (1%).

[18] Lastly, a general consequence of dating uncertainty is that some event windows overlap, which when bootstrapped, leads to many interevent times that are close to zero (Figure 4a). These are not part of the transformed normal distribution shown in Figure 4b. They instead result in an isolated spike in negative log space. This occurrence can be interpreted as short-term clustering behavior (like

aftershocks), and therefore can be added as a static shift to the interpreted rate if desired.

## 5. Conclusions

[19] A simple method is identified to calculate long-term earthquake rates from point process observations of paleoearthquakes. A feature of time-dependent earthquake recurrence distributions is that binning them by their natural logarithms results in a normal distribution. This allows identification of the  $\mu - \sigma$  bound from the most complete part of the observed record, while the  $\mu + \sigma$  bound is extrapolated by symmetry. Tests using complicated, multi-modal synthetic distributions show that the method works. This method can estimate earthquake rates at sites known to have multiple recurrence processes operating, and where Monte Carlo methods have failed such as at the Pallet Creek and Wrightwood sites. An application using real data from the south Hayward fault returns virtually the same mean as that found with computationally intensive Monte Carlo sampling. However, very irregular sequences with long interevent times such as those on the Garlock fault remain difficult to interpret.

[20] **Acknowledgments.** My thanks to Eric Geist and David Schwartz for their constructive review comments on an initial draft and to Seth Stein, Max Werner, and Editor Andrew Newman for their helpful GRL reviews.

[21] The Editor thanks Maximilian Werner and Seth Stein for their assistance in evaluating this paper.

## References

- Biasi, G. P., and R. J. Weldon II (2009), San Andreas Fault rupture scenarios from multiple paleoseismic records: Stringing pearls, *Bull. Seismol. Soc. Am.*, *99*, 471–498, doi:10.1785/0120080287.
- Biasi, G. P., R. J. Weldon II, T. E. Fumal, and G. G. Seitz (2002), Paleoseismic event dating and the conditional probability of large earthquakes on the southern San Andreas fault, California, *Bull. Seismol. Soc. Am.*, *92*, 2761–2781, doi:10.1785/0120000605.
- Console, R., M. Murru, G. Falcone, and F. Catali (2008), Stress interaction effect on the occurrence probability of characteristic earthquakes in central Apennines, *J. Geophys. Res.*, *113*, B08313, doi:10.1029/2007JB005418.
- Dawson, T. E., S. F. McGill, and T. K. Rockwell (2003), Irregular recurrence of paleoearthquakes along the central Garlock fault near El Paso Peaks, California, *J. Geophys. Res.*, *108*(B7), 2356, doi:10.1029/2001JB001744.
- Ellsworth, W. L., M. V. Matthews, R. M. Nadeau, S. P. Nishenko, P. A. Reasenberg, and R. W. Simpson (1999), A physically-based earthquake recurrence model for estimation of long-term earthquake probabilities, *U.S. Geol. Surv. Open File Rep. OF99-520*, 22 pp.
- Field, E. H., and M. T. Page (2011), Estimating earthquake-rupture rates on a fault or fault system, *Bull. Seismol. Soc. Am.*, *101*, 79–92, doi:10.1785/0120100004.
- Field, E. H., et al. (2009), The uniform California earthquake rupture forecast, version 2 (UCERF 2), *Bull. Seismol. Soc. Am.*, *99*, 2053–2107, doi:10.1785/0120080049.
- Hagiwara, Y. (1974), Probability of earthquake occurrence as obtained from a Weibull distribution analysis of crustal strain, *Tectonophysics*, *23*, 313–318, doi:10.1016/0040-1951(74)90030-4.
- Hillers, G., J. M. Carlson, and R. J. Archuleta (2009), Seismicity in a model governed by competing frictional weakening and healing mechanisms, *Geophys. J. Int.*, *178*, 1363–1383, doi:10.1111/j.1365-246X.2009.04217.x.
- Johnson, N. L. (1962), The folded normal distribution: accuracy of estimation by maximum likelihood, *Technometrics*, *4*, 249–256.
- Kagan, Y. Y., and L. Knopoff (1987), Random stress and earthquake statistics: Time dependence, *Geophys. J. R. Astron. Soc.*, *88*, 723–731, doi:10.1111/j.1365-246X.1987.tb01653.x.
- Lienkaemper, J. J., P. L. Williams, and T. P. Guilderson (2010), Evidence for a twelfth large earthquake on the Southern Hayward Fault in the past 1900 years, *Bull. Seismol. Soc. Am.*, *100*, 2024–2034, doi:10.1785/0120090129.
- Marzocchi, W., and A. M. Lombardi (2008), A double branching model for earthquake occurrence, *J. Geophys. Res.*, *113*, B08317, doi:10.1029/2007JB005472.
- Matthews, M. V., W. L. Ellsworth, and P. A. Reasenberg (2002), A Brownian model for recurrent earthquakes, *Bull. Seismol. Soc. Am.*, *92*, 2233–2250, doi:10.1785/0120010267.
- Nishenko, S. P., and R. Buland (1987), A generic recurrence interval distribution for earthquake forecasting, *Bull. Seismol. Soc. Am.*, *77*, 1382–1399.
- Parsons, T. (2006), Tectonic stressing in California modeled from GPS observations, *J. Geophys. Res.*, *111*, B03407, doi:10.1029/2005JB003946.
- Parsons, T. (2008a), Monte Carlo method for determining earthquake recurrence parameters from short paleoseismic catalogs: Example calculations for California, *J. Geophys. Res.*, *113*, B03302, doi:10.1029/2007JB004998.
- Parsons, T. (2008b), Earthquake recurrence on the south Hayward fault is most consistent with a time dependent, renewal process, *Geophys. Res. Lett.*, *36*, L21301, doi:10.1029/2008GL035887.
- Sachs, L. (1984), *Applied Statistics*, 707 pp., Springer, New York.
- Schwartz, D. P., and K. J. Coppersmith (1984), Fault behavior and characteristic earthquakes: Examples from the Wasatch and San Andreas fault zones, *J. Geophys. Res.*, *89*, 5681–5698, doi:10.1029/JB089iB07p05681.
- Stein, S., and A. Newman (2004), Characteristic and uncharacteristic earthquakes as possible artifacts: Applications to the New Madrid and Wabash seismic zones, *Seismol. Res. Lett.*, *75*, 173–187, doi:10.1785/gssrl.75.2.173.
- Weldon, R., K. Scharer, T. Fumal, and G. Biasi (2004), Wrightwood and the earthquake cycle: what a long recurrence record tells us about how faults work, *GSA Today*, *14*, 4–10, doi:10.1130/1052-5173(2004)014<4:WATECW>2.0.CO;2.
- Wesnowsky, S. G. (1994), The Gutenberg-Richter or characteristic earthquake distribution, which is it?, *Bull. Seismol. Soc. Am.*, *84*, 1940–1959.
- Zöller, G., Y. Ben-Zion, M. Holschneider, and S. Hainzl (2007), Estimating recurrence times and seismic hazard of large earthquakes on an individual fault, *Geophys. J. Int.*, *170*, 1300–1310, doi:10.1111/j.1365-246X.2007.03480.x.