Appendix C: Monte Carlo Method for Determining Earthquake Recurrence Parameters from Short Paleoseismic Catalogs: Example Calculations for California

By Tom Parsons[1]

USGS Open File Report 2007-1437C
CGS Special Report 203C
SCEC Contribution #1138C
Version 1.0

2008

U.S. Department of the Interior
U.S. Geological Survey

California Department of Conservation
California Geological Survey

[1]U.S. Geological Survey, Menlo Park, California

U.S. Department of the Interior
DIRK KEMPTHORNE, Secretary

U.S. Geological Survey
Mark D. Myers, Director

State of California
ARNOLD SCHWARZENEGGER, Governor

The Resources Agency
MIKE CHRISMAN, Secretary for Resources

Department of Conservation
Bridgett Luther, Director

California Geological Survey
John G. Parrish, Ph.D., State Geologist

U.S. Geological Survey, Reston, Virginia 2008

For product and ordering information:
World Wide Web: http://www.usgs.gov/pubprod
Telephone: 1-888-ASK-USGS

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment:
World Wide Web: http://www.usgs.gov
Telephone: 1-888-ASK-USGS

Suggested citation:
Parsons, Tom. 2008. Monte Carlo method for determining earthquake recurrence parameters from short paleoseismic catalogs: example calculations for California, *Appendix C in* The Uniform California Earthquake Rupture Forecast, version 2 (UCERF 2): U.S. Geological Survey Open-File Report 2007-1437C and California Geological Survey Special Report 203C, 29 p.
[http://pubs.usgs.gov/of/2007/1437/c/].

Contents

Monte Carlo Method for Determining Earthquake Recurrence Parameters from Short Paleoseismic Catalogs; Example Calculations for California

Tom Parsons, USGS MS-999, 345 Middlefield Rd. Menlo Park, CA 94025 (650) 329-5074, tparsons@usgs.gov

Abstract

Paleoearthquake observations often lack enough events at a given site to directly define a probability density function (PDF) for earthquake recurrence. Sites with fewer than 10-15 intervals do not provide enough information to reliably determine the shape of the PDF using standard maximum-likelihood techniques [e.g., *Ellsworth et al.*, 1999]. In this paper I present a method that attempts to fit wide ranges of distribution parameters to short paleoseismic series. From repeated Monte Carlo draws, it becomes possible to quantitatively estimate most likely recurrence PDF parameters, and a ranked distribution of parameters is returned that can be used to assess uncertainties in hazard calculations. In tests on short synthetic earthquake series, the method gives results that cluster around the mean of the input distribution, whereas maximum likelihood methods return the sample means [e.g., *NIST/SEMATECH*, 2006]. For short series (fewer than 10 intervals), sample means tend to reflect the median of an asymmetric recurrence distribution, possibly leading to an overestimate of the hazard should they be used in probability calculations. Therefore a Monte Carlo approach may be useful for assessing recurrence from limited paleoearthquake records. Further, the degree of functional dependence among parameters like mean recurrence interval and coefficient of variation can be established. The method is described for use with time-independent and time-dependent PDF's, and results from 19 paleoseismic sequences on strike-slip faults throughout the state of California are given.

1. Introduction

This paper describes a method for estimating most-likely values and resolution on earthquake recurrence interval and coefficient of variation from paleoseismic and historic earthquake records. Even long paleoseismic catalogs cannot generate a complete probability density function (PDF) on recurrence (Figure 1) [*Savage*, 1994]. Further, inconsistent statistical practice between recurrence estimation and earthquake probability calculations can be a concern [e.g., *Savage*, 1991; 1992]. Optimally, we would have enough observations of earthquake intervals to fill out recurrence PDF's; these would eliminate the epistemic uncertainties surrounding recurrence parameters, and define the aleatory uncertainty inherent in earthquake recurrence. Unfortunately, we lack the data to do that. In this paper I show that, by making one decision about the class of recurrence PDF, that PDF can be used to model observed paleoseismic sequences. As will be shown, Monte Carlo fitting tends to be most useful on short sequences and seems primarily sensitive to the histogram of the data. Results reflect epistemic uncertainties by showing the range and uncertainty in distribution parameters that are consistent with observations and their uncertainties. A further issue addressed by this analysis is that of coupled recurrence parameters.

1.1 Example

To highlight some of the issues addressed in this paper, I show an example calculation made from paleoseismology on the south Hayward fault in the San Francisco Bay region of California. *Lienkaemper et al.* [2003] reported the series of events given in Table 1.

Calendar Age		Min Interval	Max Interval	Preferred
Old	Young			
Open to 2006		138	138	138
1868	1868	78	218	148
1650	1790	0	260	130
1530	1740	0	360	180
1380	1590	0	360	180
1230	1410	0	410	205
1000	1270	0	360	180
910	1010	10	260	135
750	900	70	510	290
390	680	0	400	200
280	640	0	450	225
190	550			

Table 1. South Hayward fault paleoseismic catalog

The mean interval is 151 years (calculated by dividing time between the midpoint of the oldest event and the date of the youngest event by the number of intervals) and the distribution of preferred intervals is shown in Figure 1. Translation of a mean interval, like the 151-yr south Hayward calculation into an exponential function, or an asymmetric, time-dependent distribution such as lognormal [e.g., *Nishenko and Buland,* 1987], Weibull [*Hagiwara*, 1974], or Brownian Passage Time [*Kagan and Knopoff*, 1987; *Matthews et al.,* 2002] illustrates some of the difficulty in constraining a recurrence distribution with even a fairly lengthy paleoseismic series (Figure 1).

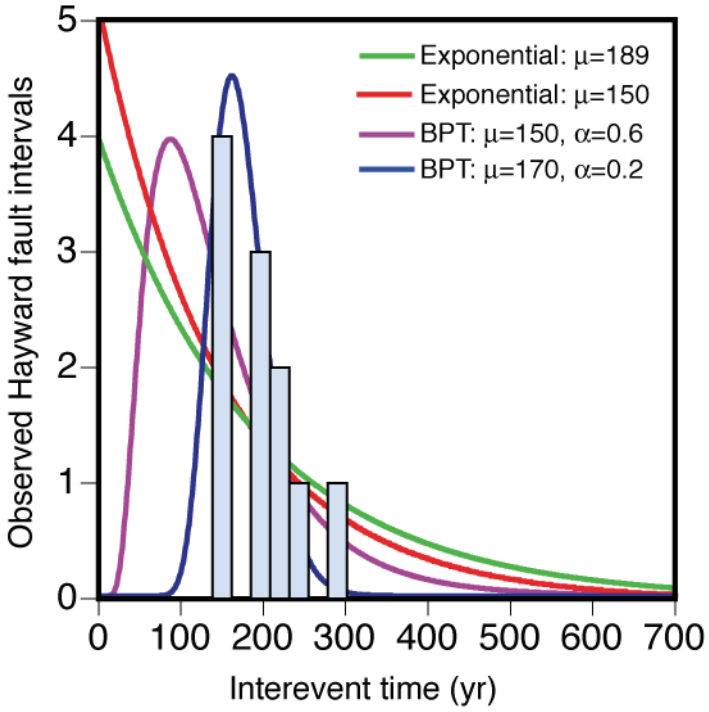

Figure 1: Observed preferred (dating interval centers) earthquake interevent times on the south Hayward fault from Table 1 [*Lienkaemper et al.*, 2003] shown in pale blue; the arithmetic mean of the intervals is ~150 yrs. Purple and blue curves are Brownian Passage Time (BPT) distributions with μ=150 yr, α=0.6 and μ =170 yr, α =0.2 mean-coefficient of variation pairings respectively. Red and green curves show exponential distributions.

Exponential distributions as defined by

$$f(t) = \lambda e^{-\lambda t}, \quad \text{for } t > 0 \tag{1}$$

where t is time, and λ is the mean rate, and Brownian Passage Time distributions as

$$f(t,\mu,\alpha) = \sqrt{\frac{\mu}{2\pi\alpha^2 t^3}} \exp\left(-\frac{(t-\mu)^2}{2\mu\alpha^2 t}\right), \tag{2}$$

where μ is recurrence interval or its proxy and α is coefficient of variation, are also plotted in Figure 1.

From the plot of event mid-times (center of dating interval) it appears that the Hayward fault has an orderly series that does not look like a Poisson process; lacking are observations of short recurrence times that would be expected if the south Hayward fault ruptured randomly in time. However, if all possible event times are bootstrapped from the dating intervals, then the possibility of short-interval events arises (Figure 2). Thus it is probably not possible to distinguish whether earthquake recurrence on the south Hayward fault is distributed according to an exponential or a time-dependent distribution [e.g., *Matthews et al.*, 2002]. However, in other places it may be possible; for example, *Ogata* [1999] found a poor fit to exponential distributions for paleoseismic sites in Japan.

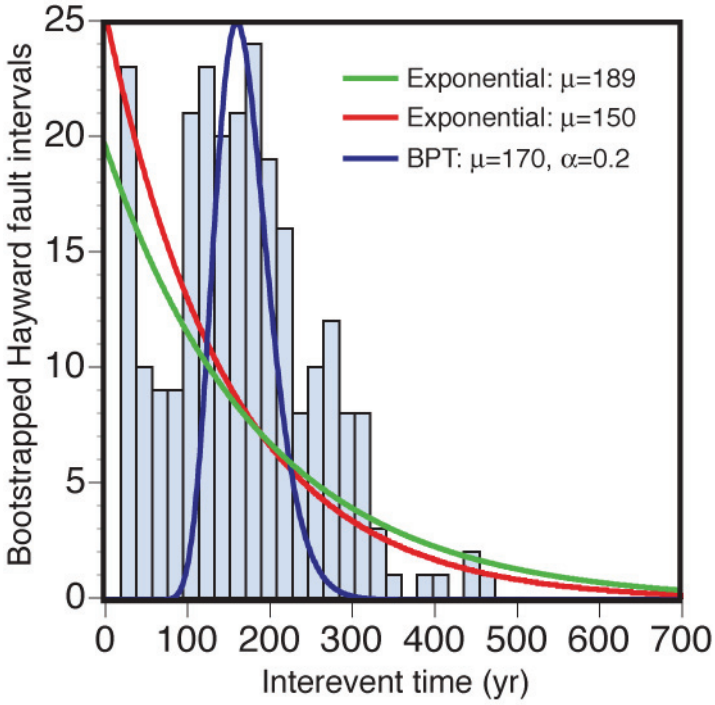

Figure 2: Bootstrapped earthquake interevent times on the south Hayward fault. When all possible interevent times are included, the distribution looks more exponential than the plot of event centers shown in Figure 1. Best-fit exponential and Brownian Passage Time distributions are plotted.

Depending on the application, there may be a desire to use exponential distributions to characterize faults like the Hayward, such as in the National Seismic Hazard Map Program [e.g., *Frankel et al.*, 2002]. Alternatively, the same fault segment may be characterized by time-dependent distributions [e.g., *Working Group on California Earthquake Probabilities (WGCEP)*, 2003]. For consistent statistical practice, recurrence parameters and their uncertainties should be developed specifically using the PDF's from which earthquake probabilities will be calculated. For the most part, paleoseismic data cannot help us choose which PDF to use, and the decision must be made from other criteria. The methods shown in this paper enable any distribution to be fit to paleoseismic observations, and no resolution between distributions is implied. Rather, one can investigate the impact of any given recurrence PDF on probabilistic hazard calculations.

1.2 Current methods

The methods described in this paper differ from other recurrence parameter estimation techniques. Most commonly, variants of maximum-likelihood techniques are applied to observed series to estimate recurrence parameters [e.g., *Nishenko and Buland*, 1987; *Davis et al.,* 1989; *Wu et al.*, 1991; *Ogata*, 1999]. To account for dating uncertainty, *Ellsworth et al.* [1999] developed a process in which carbon-dating-PDF's of paleoseismic intervals were bootstrapped, and then results were used to develop Brownian Passage Time (BPT, also know as the inverse Gaussian distribution) parameters for recurrence interval and coefficient of variation using a maximum likelihood technique. *Biasi et al.* [2002] applied a similar method for lognormal distribution parameters. When there are few events, a large number of short earthquake series are produced. Each short series still suffers the same poor resolution for predicting the shape of recurrence PDF's; for example, *Ellsworth et al.* [1999] showed that a series of 10 events could constrain recurrence coefficient of variation within a range of about 0.2 to 0.8.

As an example, let us assume from Table 1 that the Hayward fault has a ~150-yr average recurrence interval, and for simplicity, that recurrence is distributed as an exponential function. To illustrate resolution issues, I sampled a 150-yr mean exponential distribution, gathering 5,11, and 25 event sequences at random. The distribution of arithmetic means form these randomly sampled series is shown in Figure 3. Only about 10-15% of these series with less than 25 events has a mean within ±10 years of the parent distribution. Thus the arithmetic mean of any given 11-event series has a small probability of representing the actual mean recurrence interval on a fault segment. I used maximum likelihood techniques [*StataCorp.,* 2005] to find the exponential rate parameter on these synthetic series, which returned, as expected, equivalent values to the arithmetic sample means. A maximum-likelihood estimate of BPT distribution parameters for the south Hayward series shown in Figure 1 yielded a mean recurrence interval of μ=155 yr [*W. Ellsworth, written communication*, 2007], also very close to the arithmetic mean for that series (Figure 1).

The methods described in this paper yielded a recurrence interval range of μ=90-340 yr at 95% confidence, and μ=120-230 yr at 67% confidence from the same data. The method tries every reasonable recurrence PDF in a forward sense millions of times, and parameter sets that reproduce observed sequences within age uncertainties significantly more often are considered the most likely PDF's. The method is most effective at filling in gaps posed by very sparse sequences, and/or series with poorly constrained event dates.

In this paper I first show methods for modeling the rate term (λ) for the exponential distribution. Then the discussion is expanded to include time-dependent distribution parameters that include mean recurrence interval and coefficient of variation. Example calculations are made for a compilation of California paleoseismic sites.

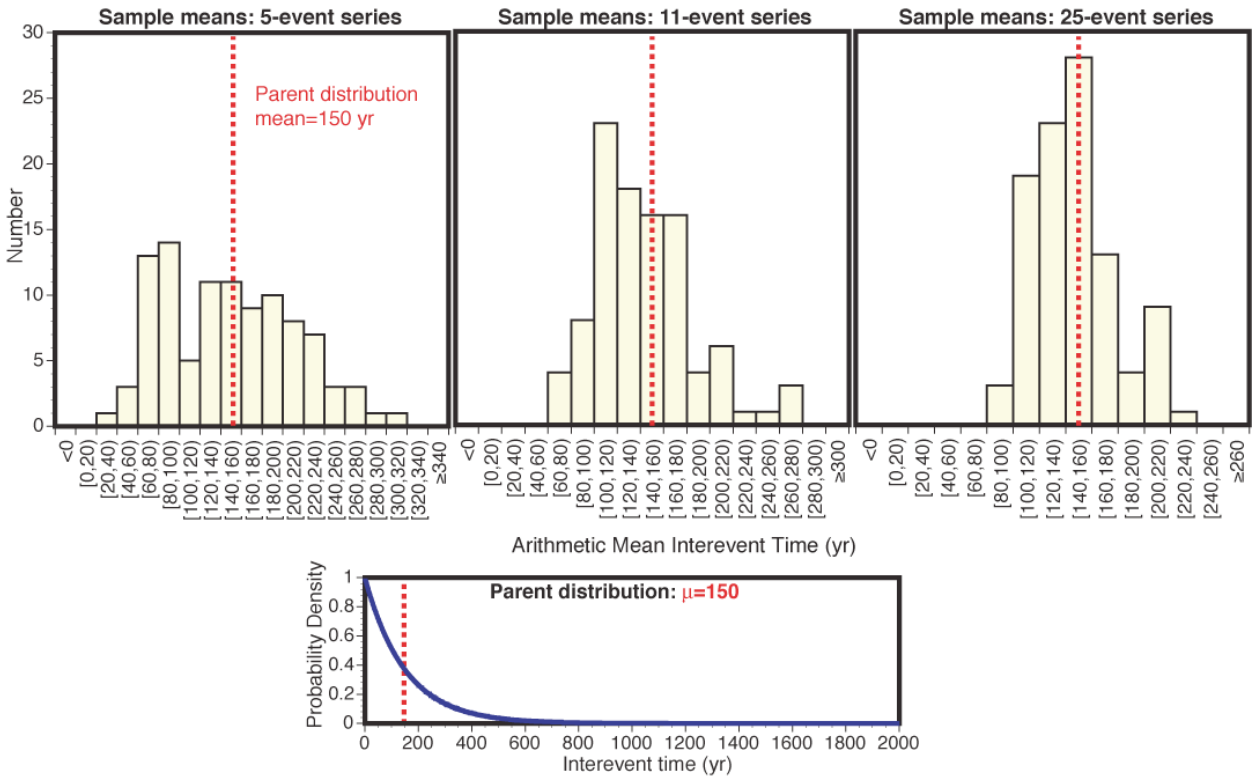

Figure 3: Distribution of arithmetic means of 5-, 11-, and 25-event series randomly sampled from a parent exponential distribution with a 150-yr mean. Only about 10-15% of the series have means within ±10 years of the parent distribution mean for series with fewer than 25 events. Techniques for determining distribution parameters such as maximum likelihood estimators are sensitive to the arithmetic mean, and thus may suffer from poor resolution on short series.

1.3 Paleoseismic data

1.3.1 Data sources. A statewide database of California paleoseismic observations was assembled for the Working Group on California Earthquake Probability (WGCEP) by *T. Dawson* and *R. Weldon* that is available from the WGCEP Paleosites database. The data represent published and unpublished contributions for major strike-slip fault zones in California including the San Andreas fault in southern and central California [*Fumal et al.*, 2002a; 2002b; *Grant and Sieh*, 1994; *Liu et al.*, 2004; *Sims*, 1994; *Biasi et al.*, 2002; *Seitz et al.*, 1996; 2000; *McGill et al.*, 2002; *Yule and Sieh*, 2001; *Sieh*, 1986; *Weldon et al.*, 2004], the Elsinore fault [*T. Rockwell*, unpublished data, 2006; *Vaughan et al.*, 1999], the San Jacinto fault [*Rockwell et al.*, 2006; *Gurrola and Rockwell*, 1996], the Garlock fault [*Dawson et al.*, 2003; *C. H. Madden*, unpublished data, 2006], the San Andreas fault in northern California [*Fumal et al.*, 2003; *Zhang et al.*, 2006], the Hayward fault [*Lienkaemper et al.*, 2003], the San Gregorio fault [*Simpson et al.*, 1997], and the Calaveras fault [*Kelson et al.*, 1996; *Simpson et al.*, 1999].

1.3.2 Uncertainty associated with paleoseismic data.

In the absence of a long term historical earthquake catalog, paleoseismic observations provide perhaps the only non-model-derived earthquake rate estimates that can be used for probabilistic forecasting. Considerable uncertainty is involved in developing a paleoearthquake catalog [e.g., *Grant and Gould*, 2004], much of which results from dating stratigraphic layering and the events that disturb such layers [e.g., *Biasi et al.*, 2002]. A very informative summary of issues surrounding radiocarbon dating, sample collection, and preparation is given by *Fumal et al.* [2002a]. Less quantifiable uncertainty can result from how earthquake features are interpreted within a trench site in terms of which types of disturbances result from large earthquakes. Data and interpretations from the California sites analyzed in this paper have been vetted through multiple reviews for consistency of geologic approach [*Grant and Lettis*, 2002].

Another issue with interpreting paleoseismic data stems from the possibility of missing events in the record. Seismologists examining instrumental earthquake catalogs have suggested that large earthquakes may cluster closely in time on the same fault [*Kagan and Jackson*, 1999]. However, most paleoseismic series are interpreted under a quasi-periodic characteristic earthquake model [*Schwartz and Coppersmith*, 1984], and tend not to indicate clustering (see Tables 1 and 2). At least three explanations are possible: (1) two or more earthquakes striking the same fault in a short time could look very much like a single event, (2) earthquakes might happen that leave no trace at a paleoseismic site, or (3) seismologists interpreting earthquake doublets on the same fault are working at a very different spatial resolution than are paleoseismologists, and could actually be seeing earthquakes rupturing neighboring fault planes. Therefore recurrence parameters derived from paleoseismic observations may carry potential sources of uncertainty beyond those explicitly quantified.

The analysis presented here provides methods for considering dating uncertainty, and open intervals before, during, and after observed paleoseismic events. However the data are not sufficient to determine whether a quasi-periodic or anti-periodic earthquake recurrence model is more appropriate. Thus both types of recurrence distributions are used.

2. Monte Carlo determination of exponential parameters

Here it is assumed that if an exponential distribution is used to calculate earthquake probability, then the best distribution parameters to use are those that most commonly reproduce an observed paleoseismic sequence. The first step is construction of a series of distributions that covers all reasonable rates (10 years to 10 times the sample mean recurrence). Intervals are randomly drawn millions of times from each series and assembled into earthquake sequences (Figure 4).

Those sequences that have one earthquake occurring in order during each observed event windows (range of possible event times as constrained by radio carbon dating), and no earthquakes in the intervals between event windows are tallied. The examples shown in this paper use a uniform distribution for the event time-window defined by dating uncertainty, and an event that happens at any time within the window is considered a match. A refinement to the technique would be weighting of solutions by comparison with the event PDF's defined by radio carbon dating analysis. At the time of this writing, event PDF's were not uniformly available for California sites. It should be noted that an event PDF is defined here as the distribution of possible ages for a single event derived from dating uncertainty, and should not be confused with the recurrence PDF, which is the distribution of possible earthquake recurrence times.

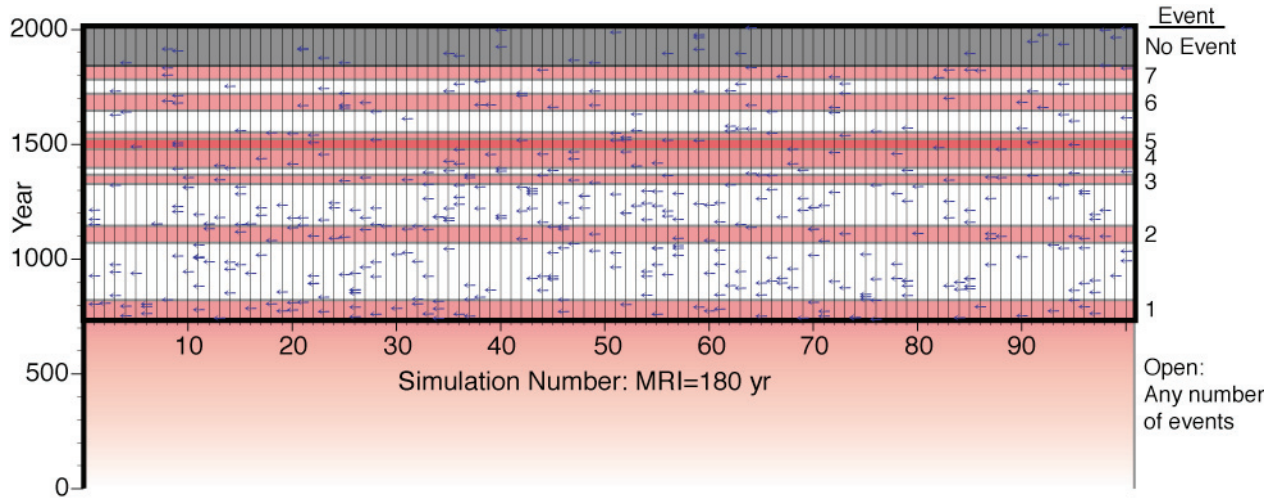

Figure 4: Paleoseismic intervals from the Burro Flats site [*Yule and Sieh*, 2001]; the gray band is the open interval since the last earthquake, the red bands show periods in which earthquakes happened, the white bands are intervals in which they did not occur. The bottom red band is the open interval before the first observed event in which any number of earthquakes may have happened. One hundred randomly drawn earthquake series from an exponential PDF with mean interval of 180 yr are shown by small blue arrows. None of the first 100 attempts fit the record. Clustering behavior inherent with the exponential distributions is evident.

Each exponential distribution (recurrence PDF) for a given rate is randomly sampled 5 million times. Each attempt that matches a paleoseismic catalog is tallied. A distribution of matches to the observed record is produced (Figure 5), and the mode (most frequent value), median, or mean of that distribution could be taken to represent the recurrence parameter (rate, or the inverse of the mean recurrence in the case of the exponential distribution). This approach simultaneously incorporates epistemic uncertainty related to dating intervals, and aleatory uncertainty related to natural interval variation.

In the examples discussed in this paper, the Mote Carlo sequences begin with an event that is given freedom to happen any time prior to the first observed earthquake time window. The extra event contributes nothing other than a starting point for the sampling. This is needed because the first observed time window has some range within which the event might have happened, whereas Monte Carlo simulation must begin at a point in time. It is expected that that an earthquake occurred prior to the first identified event in each paleoseismic series, but we have no knowledge of it other than that. To avoid the starting time having any influence, any number of events are allowed to happen over a long time prior to the first observed window. Conversely, simulations that include earthquakes within the open interval (or any other open interval) between the latest earthquake in the catalog and present time are discarded.

	Modeled Parameters									*Time/Intervals Method*					
	Lat	**Lon**	**Mode**	**Median**	**Mean**	**2.5%**	**97.5%**	**16.5%**	**83.5%**	**T min**	**T max**	**Events**	**RI Min**	**RI Max**	**RI Pref.**
Calaveras fault - North	37.5104	-121.8346	440	640	799	230	2280	370	1160	1861	2381	5	595	372	484
Elsinore - Glen Ivy	33.7701	-117.4909	150	180	232	80	620	120	310	794	947	6	159	189	174
Elsinore Fault - Julian	33.2071	-116.7273	850	1540	1855	400	4570	760	3080	N/A	N/A	2	N/A	N/A	N/A
Elsinore - Temecula	33.4100	-117.0400	1310	1620	1842	460	4350	880	2800	1200	1800	4	400	600	500
Elsinore - Whittier	33.9303	-117.8437	850	1630	1925	410	4610	790	3190	N/A	N/A	2	N/A	N/A	N/A
Garlock - Central	35.4441	-117.6815	1490	1430	1603	520	3810	840	2200	6120	6640	6	1224	1328	1276
Garlock - Western	34.9868	-118.5080	1080	1420	1616	500	3790	830	2370	3920	5350	5	980	1338	1159
Hayward fault - North	37.9306	-122.2977	260	290	343	140	750	200	460	1830	2166	5	261	542	401
Hayward fault - South	37.5563	-121.9739	160	170	189	90	340	120	230	1318	1678	11	132	168	150
N. San Andreas - North Coast	38.0320	-122.7891	N/A	N/A	N/A	N/A	N/A	N/A	N/A	2566	2896	12	233	263	248
SAF - Arano Flat	36.9415	-121.6729	110	120	141	60	270	80	180	796	896	9	100	112	106
N. San Andreas - Fort Ross	38.5200	-123.2400	220	340	456	130	1300	200	670	956	1351	5	239	338	288
San Gregorio - North	37.5207	-122.5135	440	980	1327	220	4160	440	2260	N/A	N/A	2	N/A	N/A	N/A
San Jacinto - Hog Lake	33.6153	-116.7091	N/A	N/A	N/A	N/A	N/A	N/A	N/A	3500	4000	16	233	267	250
San Jacinto - Superstition	32.9975	-115.9436	240	390	552	110	1940	200	830	476	823	3	238	412	325
South San Andreas sites	**Lat**	**Lon**	**Mode**	**Median**	**Mean**	**2.5%**	**97.5%**	**16.5%**	**83.5%**	**Time**		**Events**	**RI Min**	**RI Max**	**RI Pref.**
San Andreas - Burro Flats	33.9730	-116.8170	140	180	226	90	460	130	320	0774-2006		7	85	559	176
SAF - Combined Carrizo Plain	35.1540	-119.7000	190	290	341	120	810	180	470	0598-2006		6	108	640	235
San Andreas - Indio	33.7414	-116.1870	170	250	361	90	1170	140	550	1020-2006		4	96	904	246
San Andreas - Pallett Creek	34.4556	-117.8870	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0645-2006		10	74	283	136
San Andreas - Pitman Canyon	34.2544	-117.4340	130	160	191	80	440	110	250	0931-2006		7	75	382	154
San Andreas - Plunge Creek	34.1158	-117.1370	140	210	349	80	1480	120	480	1499-2006		3	58	820	169
Mission Creek - 1000 Palms	33.8200	-116.3010	160	260	336	100	930	150	500	0824-2006		5	102	728	236
San Andreas - Wrightwood	34.3697	-117.6680	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0533-2006		15	60	175	98

Table 2. Analysis results of 19 paleoseismic sites in California using an exponential PDF. Also given are mean recurrence intervals calculated by dividing the total time by the number of intervals. T min and T max are the minimum and maximum allowable cumulative observation intervals. Event ages calculated by *T. Dawson* [*references given in Section 1.3*], except the southern San Andreas sites, which were calculated by *Biasi et al.* [2002]. Reported confidences are upper and lower one-sided intervals.

Monte Carlo results provide a set of rates that fit observed paleoseismic sequences for use in time-independent earthquake probability calculations. There are a number of possible approaches for using these rates; one can use every value and produce a distribution of probabilities [e.g., *Savage*, 1991; 1992; *Parsons et al.*, 2000, *Parsons*, 2005], or a central value and confidence intervals can be extracted. For a central value from the Hayward example, the mode, median, or mean of of the distribution (160, 170, or 189 yr respectively) might be interpreted as the most likely value; 95% of the frequencies fall in the range between 90 and 340 years, and 67% are found between 120 and 230 years. Example distributions of exponential parameters from a number of other California sites are shown in Figure 5, which cover a broad range in terms earthquake intervals and total duration.

An attempt was made to fit earthquake sequences from 23 California paleoseismic sites using exponential PDF's. Sequences varied from 2 to 16 events spanning a total of ~500 to 6000 years. As a result, resolution on recurrence intervals differed strongly, depending on the site (Table 2). Relative resolution is defined here by the ratio of the mean recurrence interval and width of confidence intervals. The most numerous paleoseismic sequences were difficult to fit to any exponential distribution, and four sites (parameters labeled "N/A" in Table 2) were not fit even after trying 10 million times per frequency. However, broad estimates were obtained for two sites with only two events each which could not be estimated using a time/intervals method (Table 2).

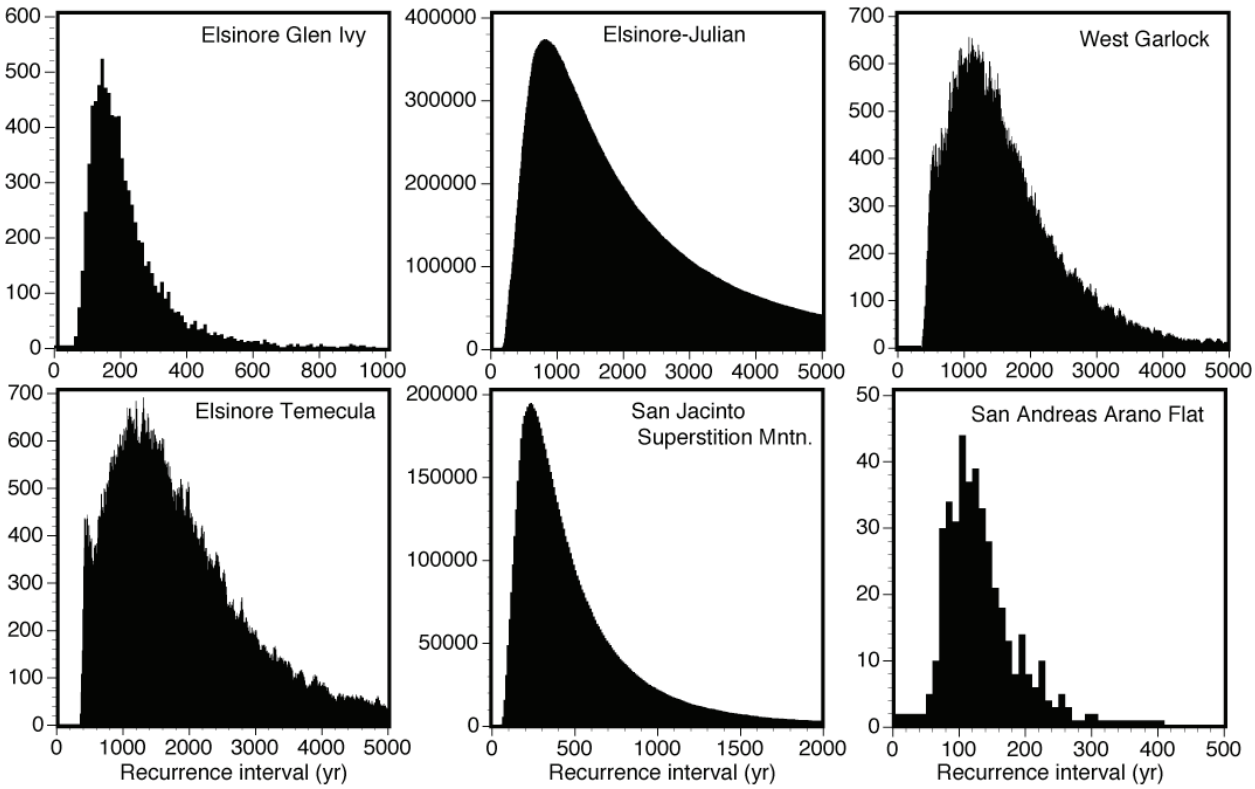

Figure 5. Example distributions of exponential frequencies fit to various California paleoseismic sites. Source: Working Group on California Earthquake Probabilities (WGCEP) paleosites database.

The most likely exponential distributions extrapolated from the paleoseismic record tend to have slightly higher mean intervals than the arithmetic means of observed records on average. The mean recurrence interval for all comparable events was 497 yr using a Monte Carlo method and 380 yr using a time/intervals method [*T. Dawso*n, unpublished data; *Biasi et al.*, 2002]. Different recurrence intervals are expected from the Monte Carlo method because open intervals are accounted for [e.g., *Ogata*, 1999]; allowing a simulated event to occur any time prior to the first observed event and disallowing events within the last open interval most closely represents the state of knowledge, but also has the effect of slightly lengthening recurrence-interval estimates. The impact of open intervals is greatest on sites with the fewest number of events, because they make up a greater proportion of the overall number of intervals.

Perhaps more important than open intervals, there are other factors that influence differences in calculated recurrence intervals between time/intervals methods and the Monte Carlo results, which may be revealed by examining Figure 3. Because of the shape of the parent exponential distribution, randomly sampled synthetic paleoseismic series tended more often to have shorter mean values than the parent distribution. Since most California paleoseismic sites have relatively short earthquake records, and if the actual parent distributions governing earthquake behavior are asymmetric (with medians less than means), odds are that sample means are smaller than the actual means.

To better assess the relative ability of different methods to recover the mean recurrence intervals from a small sample, 100 analyses were performed using Monte Carlo methods and

maximum likelihood. The average California paleoseismic site on which Monte Carlo analysis could be performed had 5 events and an apparent mean recurrence interval of 341 yr. The total average dating uncertainty was about 50% of the recurrence interval. Thus 100 random series were pulled from an exponential distribution with a 341-yr mean and ±75-yr time windows were generated around each event. For 5 events, maximum likelihood is the same as the sample mean (observation period divided by the number of events). Results of the simulations are shown in Figure 6.

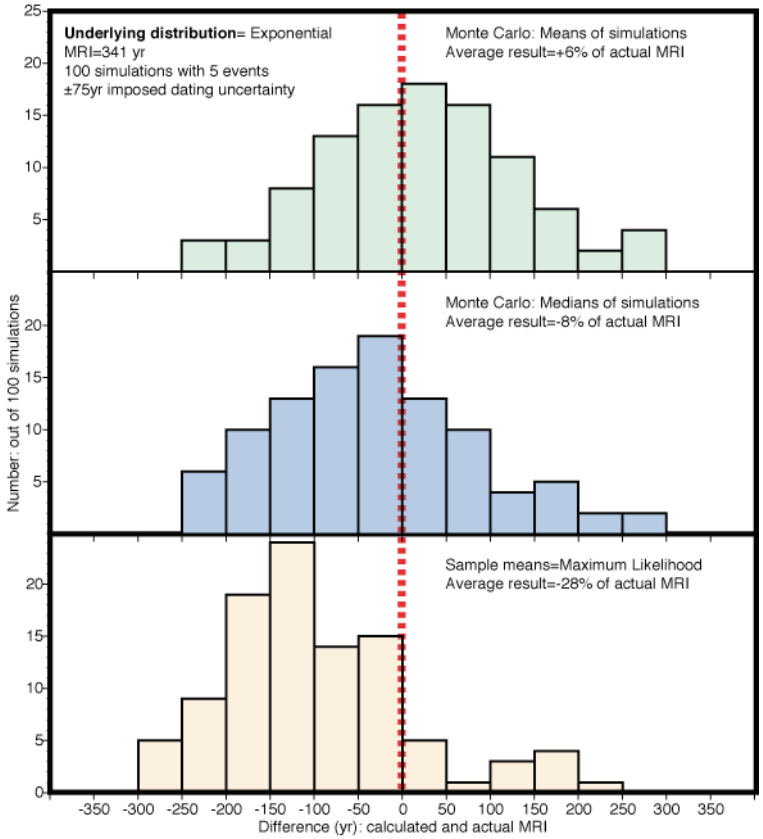

Figure 6. Results of 100 Monte Carlo analyses on 5-event synthetic earthquake series drawn from a known distribution. Histograms show differences between calculated and the correct recurrence value. The asymmetric exponential distribution causes a sampling bias for short series such that the arithmetic mean of the samples is systematically less than the underlying distribution, approaching the distribution median of $0.63/\lambda$ (bottom panel). Monte Carlo techniques are subject to considerable uncertainty (upper panels), but tend to distribute more symmetrically about the mean.

Sample means tended to underestimate recurrence intervals by an average of 97 yr or 28% (see also Figure 3), clustering most closely to the median of the parent distribution ($0.63/\lambda$=215 yr). In the limiting case of a single interval, sample means would cluster exactly at the median because there is a 50% chance of landing on either side of it. With increasing sample size (>25, Figure 3), sample means begin to reflect the mean of the parent distribution. While any given sample mean could be the same as the underlying distribution, odds are that sample means from a collection of paleoseismic sites with 5-10 events each (as is the case in California) will yield a result closer to the underlying medians. The result being a possible overestimate of

the hazard if earthquakes recur according to an asymmetric PDF with medians less than means (e.g., exponential, lognormal, Brownian Passage Time).

Means of Monte Carlo simulations overestimated mean recurrence by an average 20 yr or 6%, while use of the median values underestimated recurrence intervals by 27 yr or 8%. The mode of Monte Carlo analyses tended to underestimate recurrence intervals more, by an average 70 yr (21%); thus use of the median or mean of the simulations appears to be the best choice. If, as is usually assumed [e.g., *WGCEP*, 2003], earthquake recurrence is distributed asymmetrically, then it would appear that Monte Carlo sampling may be a useful technique for analysis of (<5-10 events) short paleoseismic series.

As would be expected, there is an apparent relationship between the number of observed earthquakes in a paleoseismic series and confidence in determining mean recurrence interval. In Figure 7, the number of events in each series is plotted against a normalized expression of the confidence, where means are divided by 95% confidence intervals. Thus larger values imply better-constrained recurrence intervals. A standard-error calculation on expected resolution is shown for comparison.

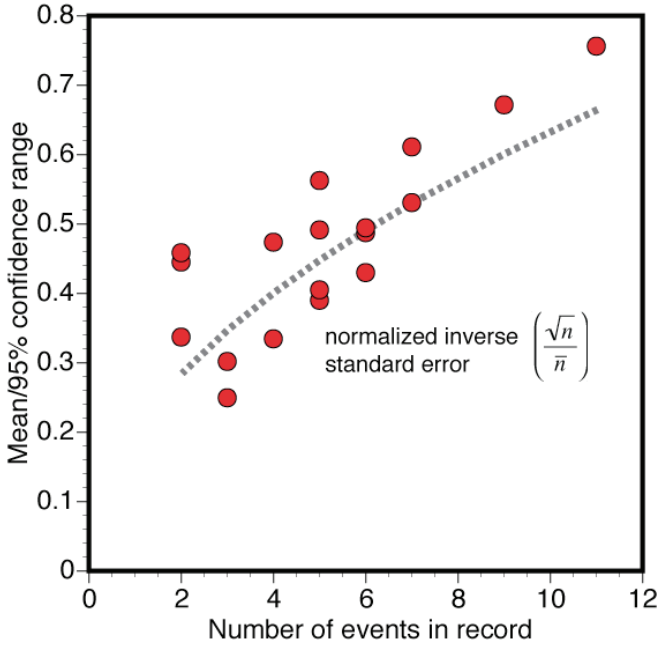

Figure 7. Number of paleoearthquakes at different California sites plotted against normalized expression of confidence on mean recurrence interval (mode divided by 95% confidence interval). Dashed line represents inverse standard error.

To summarize, exponential distribution means were calculated for 19 California paleoseismic sites using a Monte Carlo method in which all reasonable exponential distributions were considered. Relative success rates of different recurrence distribution parameters from the Monte Carlo analysis were used to estimate most-likely interevent times at each site. A test using synthetic paleoseismic series tended to show recovery of the parent distribution mean within about 5-10% of the actual value. The full array of exponential PDF's (e.g., Figure 5) can be retained for use in probability calculations as a way of accounting for uncertainty.

3. Monte Carlo determination of time-dependent recurrence parameters

Time-dependent probability calculations can be fit to mimic the renewal hypothesis of earthquake regeneration such that earthquake likelihood on a fault is lowest just after the last event. As tectonic stress grows, the odds of another earthquake increase. A time-dependent probability calculation sums a PDF *f(t)* as

$$P(t \le T \le t + \Delta t) = \int_{t}^{t+\Delta t} f(t)dt \tag{3}$$

where *f(t)* can be any distribution, such as lognormal [e.g., *Nishenko and Buland,* 1987], Weibull [*Hagiwara*, 1974], or Brownian Passage Time (inverse Gaussian) [*Kagan and Knopoff*, 1987; *Matthews et al.,* 2002]. These functions distribute interevent time or its proxy (μ), and the width of the distributions represents inherent variability (coefficient of variation α) on recurrence. For example, in the case of common practice where α is limited to between 0 and 1, a very narrow distribution implies very regular recurrence.

Two commonly applied probability density functions, the lognormal

$$f(t,\mu,\alpha) = \frac{1}{t\alpha\sqrt{2\pi}} \exp\left(\frac{-(\ln t - \mu)^2}{2\alpha^2}\right) \tag{4}$$

and Brownian Passage Time (inverse Gaussian, Equation 2), have characteristics that qualitatively simulate earthquake renewal when α is limited to be between 0 and 1. The distributions are asymmetric (Figure 1), with less weight at very short recurrence times which, when integrated, translates to very low probability early in the earthquake cycle. They are defined by two parameters, mean interevent time (μ), and a coefficient of variation (α) that govern their shape. The distributions differ in their asymptotic behavior; integration of the lognormal distribution to very long times asymptotes to zero, whereas the Brownian Passage Time distribution asymptotes to a fixed value, behavior that *Matthews at al.* [2002] say favors the Brownian distribution for hazard calculations.

3.1 Recurrence interval and coefficient of variation

The strategy for Monte Carlo determination of time-dependent recurrence parameters is much the same as described for exponential frequencies in Section 2, except the analysis must be expanded to consider a range of coefficient of variation. Thus distributions with means covering the reasonable range of possible recurrence values (10 yr to 10 times the sample mean) are constructed across coefficient of variation values between 0.01 and 1.0. For each PDF described by a paring of coefficient of variation and mean recurrence, event interval sets are drawn at random and assembled into earthquake sequences 5 million times. Those that match observed event windows (range of possible event times as constrained by radio carbon dating assuming a uniform distribution), are tallied. Open intervals and the interval before the first catalog event are treated in the same way as described for the exponential distribution example in Section 2.

Undertaking dual-parameter estimation from sparse paleoseismic data is a difficult problem, and one that generally produces a poorly resolved result. The goal with the Monte Carlo approach is to provide a quantitative basis for limiting parameters, and for assessing most-likely pairings, because time-dependent probability calculations demand these values. Thus at the very least, the results can be used to test the consistency of a chosen time-dependent model with the range of paleoseismic constraints.

As an example, the paleoseismic catalog from the south Hayward site of Table 1 is analyzed. By repeatedly sampling a full range of time-dependent PDF's, the most likely combinations emerge (Figure 8). In this example, Brownian Passage Time distributions [e.g., *Matthews at al.*, 2002] are used. The contour plot of Figure 8 shows the range of possible PDF's that can reproduce the paleoseismic sequence on the south Hayward fault. From model results it is possible to conduct a variety of statistical analyses to aid in selecting appropriate parameters for use probability calculations. The mean of coefficient of variation-recurrence interval combinations yielding PDF's that match observed records most frequently can be taken as most likely following the results of analysis shown in Figure 6. Additionally, the number of matches to observed for a given combination can be divided by the total number of matches, and a relative likelihood expressed as a fraction of the most-frequently successful combination (Table 3).

In Table 3 and Figure 8, results from the south Hayward fault analysis are given with the most likely parameters expressed as a function of coefficient of variation. In that example, a coefficient of variation of 0.2 is favored, similar to the value calculated by *Sykes and Menke* [2006]. The mean of that PDF is 172 yr, and the 95% confidence interval ranges from 140 to 190 yr; confidence intervals were found by counting the hits. Should these parameters be used in time-dependent probability calculations, the relative likelihood of each parameter set could be used to weight the solutions. For example, a coefficient of variation of 0.2 is ~8 times more likely than a value of 0.5 according to the Monte Carlo analysis (Table 3). More examples from a variety of paleoseismic sites are discussed and tabulated in Section 4.

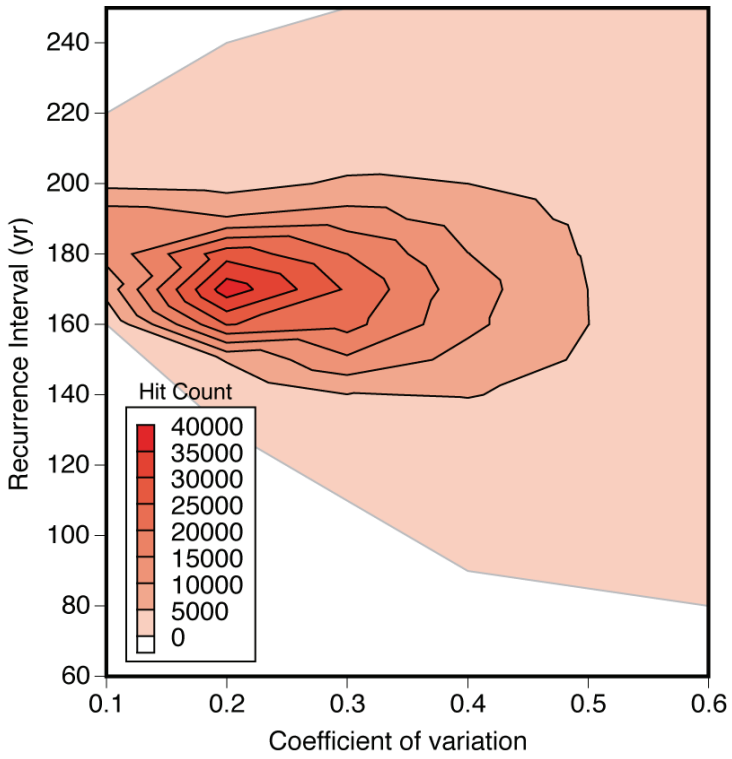

Figure 8. Analysis results for time-dependent parameters for Brownian Passage Time distributions at a site on the south Hayward fault [*Lienkaemper et al.*, 2003]. The number of matches to the observed paleoseismic sequence are contoured vs. recurrence interval and coefficient of variation. The dashed line illustrates dependence between recurrence interval and coefficient of variation.

Prob	COV	Mode	Median	Mean	2.5%	98.5%	16.5%	83.5%
0.0002	0.99	200	153	194	80	340	120	240
0.0004	0.9	190	135	188	80	300	120	230
0.0008	0.8	170	142	184	90	270	130	220
0.0017	0.7	160	148	181	100	260	130	210
0.0039	0.6	170	156	177	100	240	130	200
0.0087	0.5	160	134	173	110	230	130	190
0.0208	0.4	170	137	172	120	210	140	180
0.0420	0.3	170	138	171	130	200	140	180
0.0657	0.2	170	135	172	140	190	150	170
0.0235	0.1	190	143	184	160	190	150	180

Table 3. Probability of coefficient of variation-recurrence interval combinations defining recurrence PDF's from Monte Carlo modeling of the Tule Pond site on the Hayward fault. The most likely mode, median, and mean of the recurrence distributions are given for a range of coefficient of variation values. In addition, 95% and 67% confidence bounds on recurrence intervals are given.

In some cases there can be dependence between the coefficient of variation and recurrence interval. That is, for a given recurrence interval, the most likely range of coefficient of variation differs from that of a different interval. An example of this behavior is shown in Figure 9, where data from a trench on the Elsinore fault at Glen Ivy [*T. Rockwell*, unpublished data] were analyzed. This issue varies in importance depending on the site investigated (e.g., no slope is apparent on Figure 8). Thus model results could be used to identify the most likely combinations on a segment-by-segment basis. Such an analysis for a multi-fault probability forecast reduces the possibility of giving non-zero weight to mutually exclusive recurrence models [*Page and Carlson*, 2006] as compared with applying a single range of coefficient of variation across an entire region.

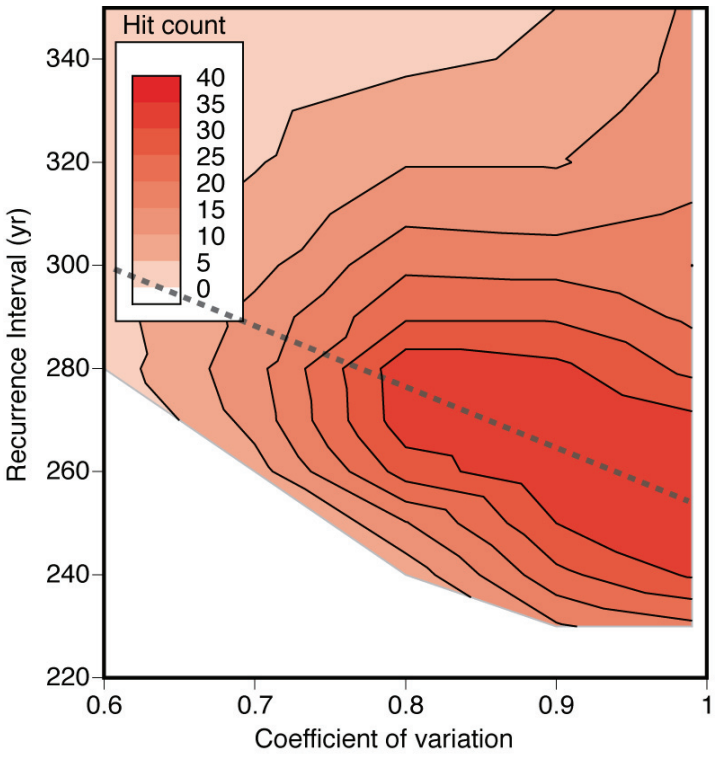

Figure 9. Analysis results for time-dependent parameters for Brownian Passage Time distributions at a site on the Elsinore fault at Glen Ivy (Source: Working Group on California Earthquake Probabilities (WGCEP) paleosites database.) The number of matches to the observed paleoseismic sequence are contoured vs. recurrence interval and coefficient of variation. The dashed line illustrates dependence between recurrence interval and coefficient of variation.

Calculation results yield distributions of coefficient of variation-recurrence interval pairings, which in turn define distributions. Examination of Table 3 shows that recurrence parameter sets are not normally distributed since the mode, mean, and median all differ. Here the mean is taken as the preferred value based on tests shown in Figure 6. However, It is necessary to verify that the distributions are not multi-modal. For example, in Figure 10, a bimodal distribution is evident from analysis of the Whittier site on the Elsinore fault. The site has only 2 reported events that yield one closed and two open intervals, allowing for a broad array of possible PDF's each with comparable likelihood of being correct (Table 4). However, the range of likelihood

does provide a mechanism for weighting parameters in a probability calculation which, given the similarity in weights, would reflect the breadth of the recurrence-parameter uncertainty.

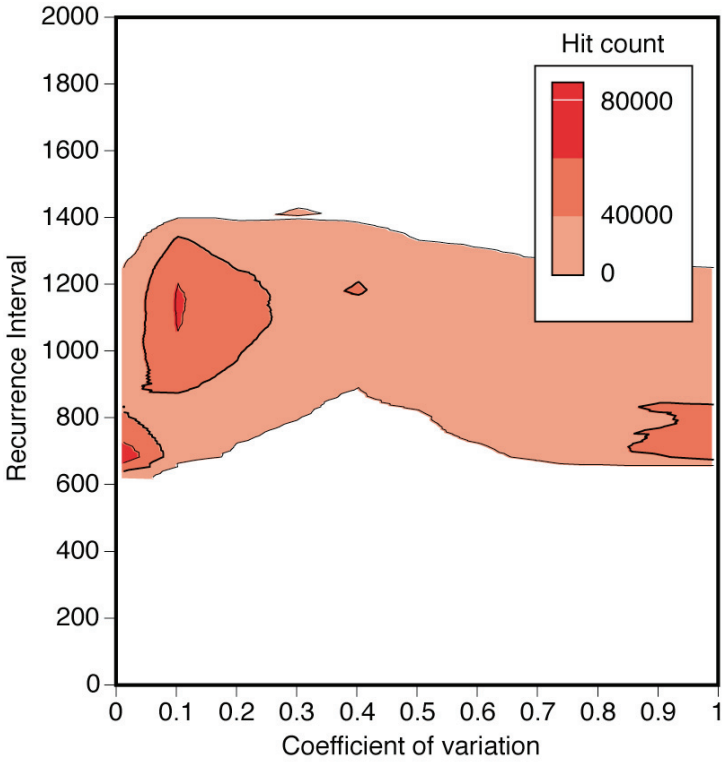

Figure 10. Analysis results for time-dependent parameters for Brownian Passage Time distributions at a site on the Elsinore fault at Whittier (Source: Working Group on California Earthquake Probabilities (WGCEP) paleosites database.) The number of matches to the observed paleoseismic sequence are contoured vs. recurrence interval and coefficient of variation. The plot shows a bimodal distribution of recurrence intervals with modes at about 1100 and 700 years.

Prob	COV	Mode	Median	Mean	-95%	95%	-67%	67%
0.0015	0.99	700	1000	1138.7	670	1950	860	1180
0.0014	0.9	700	1010	1150.5	680	1970	870	1200
0.0012	0.8	750	1030	1167.4	680	2010	890	1230
0.0011	0.7	820	1060	1193.7	690	2040	920	1250
0.0011	0.6	860	1100	1226	720	2080	950	1290
0.0012	0.5	1040	1130	1256.4	740	2080	1000	1310
0.0013	0.4	1180	1190	1286.1	740	2060	1070	1320
0.0011	0.3	1100	1190	1289.7	730	2050	1050	1360
0.0015	0.2	1160	1130	1222.7	720	2010	1010	1250
0.002	0.1	1150	1090	1150.9	700	1690	980	1200

Table 4. Probability of coefficient of variation-recurrence interval combinations defining recurrence PDF's from Monte Carlo analysis from the Whittier site on the Elsinore fault. The analysis was constrained by only 2 intervals (one open), so the relative likelihoods of different combinations are not very different, demonstrating a poorly determined solution.

4. Time dependent recurrence interval estimates from California paleoseismic sites

In this section, the methods outlined in Section 3 are used to analyze a variety of California paleoseismic sites and most-likely values for mean recurrence interval and coefficient of variation are reported. Parameters derived from Monte Carlo analysis are compared with values taken directly from paleoseismic series.

Data from trenches across 19 California strike-slip fault segments were analyzed using Monte Carlo methods. These are the same sites for which exponential parameters were developed in Section 2. A range of coefficient of variation values was used from 0.01 to 0.99 in 0.1 increments, and recurrence intervals from 10 yr to 10 times the sample means were attempted at 10-yr intervals. Each combination was tried 5 million times, resulting in a total of $\sim 1.65 \cdot 10^{10}$ randomly-drawn earthquake series per site that were compared with observed sequences, which represents a maximum reasonable compute time for desktop computers. As in the time-independent calculations, the lengthiest series and ones with most tightly-constrained ages were not reproduced with any combination. Thus, unless higher-powered computer resources are used, the method is best applied to sparse paleoseismic sequences.

Just as recurrence intervals determined from Monte Carlo analysis using exponential functions could be slightly higher than sample arithmetic means depending on whether the modes, means, or medians of the Monte Carlo results were used, so too were those calculated using Brownian Passage Time (BPT) distributions. These effects can be shown graphically by plotting modeled recurrence intervals against the means of observed intervals (Figure 11).

			Modeled Parameters								Time/Intervals Method					
Site	**Lat**	**Lon**	**COV**	**Mode**	**Median**	**Mean**	**2.5%**	**97.5%**	**16.5%**	**83.5%**	**T min**	**T max**	**Events**	**RI Min**	**RI Max**	**RI Pref.**
Calaveras fault - North	37.5104	-121.8346	0.2	440	430	445	320	610	370	500	1861	2381	5	372	595	484
Elsinore - Glen Ivy	33.7701	-117.4909	0.8	270	270	291	240	350	260	280	794	947	6	159	189	174
Elsinore Fault - Julian	33.2071	-116.7273	0.5	900	1060	1193	740	2400	870	1460	N/A	N/A	2	N/A	N/A	N/A
Elsinore - Temecula	33.4100	-117.0400	0.8	680	700	741	560	1090	630	820	1200	1800	4	400	600	500
Elsinore - Whittier	33.9303	-117.8437	0.5	1040	1130	1256	690	2610	880	1570	N/A	N/A	2	N/A	N/A	N/A
Garlock - Central	35.4441	-117.6815	0.6	740	720	729	650	760	670	740	6120	6640	6	1224	1328	1276
Garlock - Western	34.9868	-118.5080	0.7	640	680	711	550	990	610	790	3920	5350	5	980	1338	1159
Hayward fault - North	37.9306	-122.2977	0.7	330	350	363	270	480	310	400	1830	2166	5	261	542	401
Hayward fault - South	37.5563	-121.9739	0.2	170	160	172	140	190	150	170	1318	1678	11	132	168	150
N. San Andreas - North Coast	38.0320	-122.7891	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	2566	2896	12	233	263	248
SAF - Arano Flat	36.9415	-121.6729	0.4	110	100	115	70	140	90	120	796	896	9	100	112	106
N. San Andreas - Fort Ross	38.5200	-123.2400	0.2	360	350	360	270	530	330	390	956	1351	5	239	338	288
San Gregorio - North	37.5207	-122.5135	0.2	440	450	480	320	770	370	560	N/A	N/A	2	N/A	N/A	N/A
San Jacinto - Hog Lake	33.6153	-116.7091	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	3500	4000	16	233	267	250
San Jacinto - Superstition	32.9975	-115.9436	0.8	310	360	402	230	640	280	520	476	823	3	238	412	325
South San Andreas sites	**Lat**	**Lon**	**COV**	**Mode**	**Median**	**Mean**	**2.5%**	**97.5%**	**16.5%**	**83.5%**	**Time**		**Events**	**RI Min**	**RI Max**	**RI Pref.**
San Andreas - Burro Flats	33.9730	-116.8170	0.7	220	220	234	190	340	210	220	0774-2006		7	85	559	176
SAF - Combined Carrizo Plain	35.1540	-119.7000	0.6	210	220	244	160	380	180	280	0598-2006		6	108	640	235
San Andreas - Indio	33.7414	-116.1870	0.7	280	320	358	230	590	270	420	1020-2006		4	96	904	246
San Andreas - Pallett Creek	34.4556	-117.8870	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0645-2006		10	74	283	136
San Andreas - Pitman Canyon	34.2544	-117.4340	0.9	240	240	261	200	370	210	290	0931-2006		7	75	382	154
San Andreas - Plunge Creek	34.1158	-117.1370	0.7	310	330	361	290	600	310	450	1499-2006		3	58	820	169
Mission Creek - 1000 Palms	33.8200	-116.3010	0.6	250	280	303	200	450	240	340	0824-2006		5	102	728	236
San Andreas - Wrightwood	34.3697	-117.6680	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	0533-2006		15	60	175	98

Table 5. Model results from analysis at 19 paleoseismic sites in California using a Brownian Passage Time PDF. Most-frequent combinations of coefficient of variation and recurrence values that reproduced observed event sequences are reported for each site. Also given are mean recurrence intervals calculated by dividing the total time by the number of intervals (same as in Table 2). Event ages calculated by *T. Dawson* [*WGCEP paleosites database*], except the southern San Andreas sites, which were calculated by *Biasi et al.* [2002].

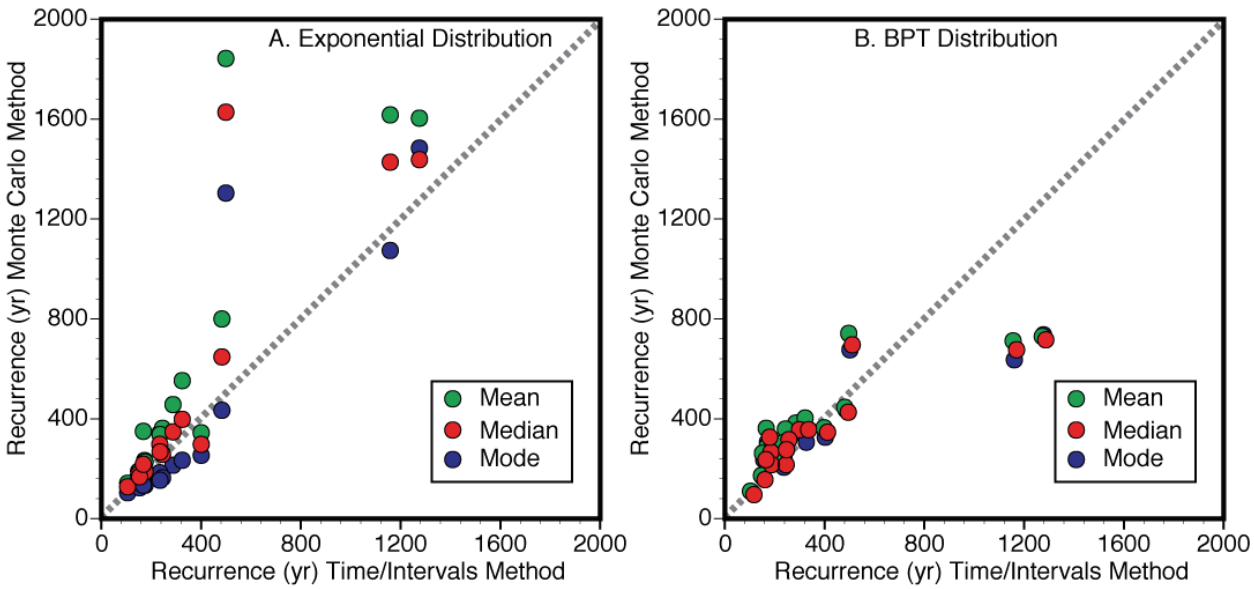

Figure 11. Plot of modeled recurrence-interval means against (a) exponential, and (b) Brownian Passage Time distribution parameters derived directly from observed intervals. The dashed black lines have a slope of 1.0; thus points falling on it would imply both methods yield the same result. The means, medians, and modes of Monte Carlo fits to parameters from Tables 2 and Table 5 are shown by green, red and blue dots respectively.

The Monte Carlo method described in this paper identifies most-likely PDF's that reproduce a paleoseismic sequence. For time-dependent analysis, Brownian Passage Time functions were used, which have long tails (Figure 1). A long-tailed distribution implies that there are low-probability long-interval events expected. The majority of California paleoseismic sites have sequences of fewer than 10 events identified. Thus if earthquakes are actually distributed like a Brownian Passage Time (or lognormal) distribution, infrequent long-interval events are unlikely to be in the record. So, when models impose long-tailed shapes to observed intervals, there is an assumption that long-interval recurrence times will happen in the future. Thus the means of best-fit long-tailed PDF's are expected to be higher than the arithmetic mean of intervals (Figure 11). An exception to that outcome is the Garlock fault, where there are some long intervals in the record (Figure 12). In that instance, Monte Carlo modeling fits the mean recurrence interval at μ=729 yr as compared to the arithmetic mean of μ=1276 yr (Table 5; Figure 12).

Calculations performed here enable an assessment of most-likely values of earthquake coefficient of variation across California. Values range from α=0.2 to α=0.9 (Table 5), implying that choosing a single value for the entire state in probabilistic forecasts might not be the best practice. There appears to be some consistency among fault zones with multiple sites; the four Elsinore fault sites show relatively high coefficient of variation values of α=0.5 to α=0.8, and the two Garlock fault sites are similar at α=0.6 and α=0.7. On the southern San Andreas fault, nearby sites tend to be somewhat consistent: Plunge Creek and Pitman Canyon both have α=0.7-0.9 modeled coefficient of variation, and the Indio and Thousand Palms Oasis sites have values of α=0.6 to 0.7. A counter example is the Hayward fault, with its 2 sites having coefficient of variation values of α=0.2 and α=0.7.

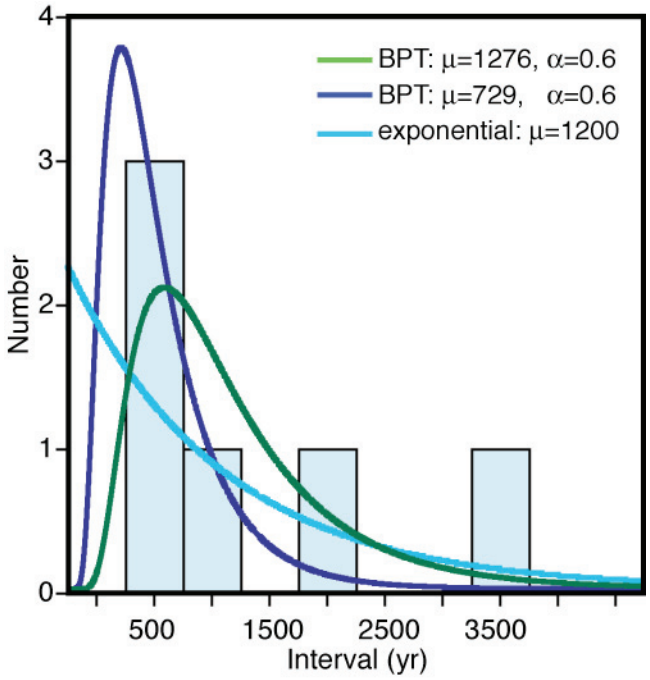

Figure 12: Observed preferred earthquake intervals on the central Garlock fault [*Dawson et al.*, 2003; *C. H. Madden*, unpublished data, 2006] shown in pale blue; the arithmetic mean of the intervals is ~1276 yrs. The green and dark blue curves are BPT distributions with 1276-yr and 729-year means respectively The light blue curve is an exponential distribution corresponding to a 1200-yr mean. In this instance Monte Carlo modeling finds that BPT distributions with lower means than the arithmetic mean can accommodate the long-intervals in the record because of the distribution asymmetry.

As in coefficient of variation, there is a degree of consistency in modeled recurrence interval among faults (Table 5). Three of 4 Elsinore sites are calculated to have recurrence estimates that range from μ=741 to μ=1256 years. The Glen Ivy site shows a higher frequency of μ=291 years. The two Garlock-fault sites are calculated to have consistent recurrence intervals of μ=711 and μ=729 years. Of the southern San Andreas fault sites that could be fit with a Monte Carlo technique there is consistency in modeled recurrence intervals, with values ranging between μ=234 and μ=303 years. However, the southern San Andreas locations not analyzed here, the Wrightwood and Pallet Creek sites, have yielded significantly shorter intervals (μ=105 and μ=135 years respectively) as calculated by *Biasi et al.* [2002]. There is a discrepancy in northern San Andreas earthquake recurrence between the Arano Flat site near Watsonville that shows a relatively short recurrence interval of μ=110 years as compared with the Ft. Ross site north of the San Francisco Bay area that has a 350-year interval.

5. Most-likely-rate and 1-sigma approximation calculated from recurrence interval models

The linear inversion described in Appendix G was informed using rates from the ‘rate' column in Table 6. Uncertainty bounds are meant to approximate 1 standard deviation (1-σ) on rates for potential use in a weighted inversion. Distributions of recurrence intervals were

calculated with three methods, and in most cases were highly asymmetric. I thus had to approximate 1-σ variability by trying to capture a measure of the width of recurrence-interval distributions . Below, methods used are described in detail.

Most-likely rates were calculated in three ways: (1) for sites on the south San Andreas fault, rates were calculated from the reciprocal of the preferred values reported at sites by *Biasi et al.* ([2002], manuscript to be submitted to Bulletin of the Seismological Society of America, and Appendix E of this report). Because of the correlation methods used by *Biasi et al.*, and the relatively large number of events on the south San Andreas fault, these values were deemed better constrained than values that would result from forward modeling with Monte Carlo methods. The *Biasi et al.* approach used time-intervals to calculate Poisson-model means, but used exponential distributions to calculate uncertainties. (2) For most other sites, the forward-modeling method outlined in this appendix (C) was used to find the mode (most frequent value) from fitting exponential distributions (Table 2). (3) For long sequences where the forward-modeling approach did not converge to a solution (north-coast San Andreas fault: Vendanta site, San Jacinto fault: Hog Lake site) the rate and uncertainty were taken from the preferred and min/max rates calculated by Tim Dawson (Appendix B).

In Table 6, the column labeled 'sigma' is an estimated standard deviation, and should only be thought of in the context of the weighted linear inversion where a symmetrical approximation of 1-σ ranges was desired. These values are not reflective of confidence bounds as given in Tables 2 and 5. The method used to approximate a symmetric 1-σ range was to follow

$$\frac{1}{2}\left(\frac{1}{RI_{\min}}-\frac{1}{RI_{\max}}\right).$$

Since the linear least-squares inversion expects normally-distributed errors, this 1-σ range was then treated as if the rate was at the center of a normal recurrence-interval distribution. Since the rates came from three sources, and uncertainty was treated differently by different investigators, some modifications were made through discussions between Weldon and Parsons as to what values should be used for RI_{min} and RI_{max} (minimum and maximum recurrence intervals) to make 1-σ estimates comparable for paleoseismic series of comparable quality and length. The Dawson method produced 1-σ minimum and maximum rates; the *Biasi et al.* [2002] methodology (Appendix E) produced 95% confidence thresholds for minimum and maximum rates (Appendix B). To make comparable estimates of the 1-σ ranges, the factor of ½ in the above equation was replaced with a factor of ¼, under the assumption that 95% confidence is approximately 2σ. For the forward-modeled values, the 95% confidence bounds reported in Table 5 were used, but the span was multiplied by ½ because 67% confidence intervals tended to be very narrow in the asymmetric distributions (Table 2).

Site	Lat	Lon	Poisson rate*	BPT rate*	time/intervals rate	rate sigma (from Poisson)*	rate sigma (from BPT)*	rate sigma (from intervals)	Poisson RI*	BPT RI*	Time/int RI
Calaveras fault - North	37.5104	-121.8346	0.001252	0.00224719	0.002067	0.001955	0.000743	0.000504	799	445	484
Elsinore - Glen Ivy	33.7701	-117.4909	0.004310	0.00343643	0.005744	0.005444	0.000957	0.000499	232	291	174
Elsinore Fault - Julian	33.2071	-116.7273	0.000539	0.00083822	N/A	0.001141	0.000467	N/A	1855	1193	N/A
Elsinore - Temecula	33.4100	-117.0400	0.000543	0.00134953	0.002000	0.000972	0.000434	0.000833	1842	741	500
Elsinore - Whittier	33.9303	-117.8437	0.000519	0.00079618	N/A	0.001111	0.000533	N/A	1925	1256	N/A
Garlock - Central	35.4441	-117.6815	0.000624	0.00137174	0.000784	0.000830	0.000111	0.000032	1603	729	1276
Garlock - Western	34.9868	-118.5080	0.000619	0.00140647	0.000863	0.000868	0.000404	0.000137	1616	711	1159
Hayward fault - North	37.9306	-122.2977	0.002915	0.00275482	0.002491	0.002905	0.000810	0.000993	343	363	401
Hayward fault - South	37.5563	-121.9739	0.005291	0.00581395	0.006676	0.004085	0.000940	0.000812	189	172	150
N. San Andreas - Vendanta	38.0320	-122.7891	0.004237	N/A	0.004028	0.002400	N/A	0.002448	236	N/A	248
SAF - Arano Flat	36.9415	-121.6729	0.007092	0.008696	0.009456	0.006481	0.003571	0.000536	141	115	106
N. San Andreas - Fort Ross	38.5200	-123.2400	0.002193	0.002778	0.003468	0.003462	0.000908	0.000613	456	360	288
San Gregorio - North	37.5207	-122.5135	0.000754	0.002083	N/A	0.002153	0.000913	N/A	1327	480	N/A
San Jacinto - Hog Lake	33.6153	-116.7091	0.004149	N/A	0.004000	0.002316	N/A	0.000273	241	N/A	250
San Jacinto - Superstition	32.9975	-115.9436	0.001812	0.002488	0.003079	0.004288	0.001393	0.000887	552	402	325
Site	**Lat**	**Lon**	**Poisson rate***	**LN rate**	**time/intervals rate**	**rate sigma (from Poisson)**	**rate sigma (from LN)***	**rate sigma (from intervals)**	**Poisson RI***	**LN RI**	**Time/int RI**
San Andreas - Burro Flats	33.9730	-116.8170	0.002381	0.006944	0.005682	0.002494	0.002174	N/A	420	144	176
SAF - Combined Carrizo Plain	35.1540	-119.7000	0.003571	0.005128	0.004255	0.001924	0.001975	N/A	280	195	235
San Andrteas - Indio	33.7414	-116.1870	0.003125	0.004854	0.004065	0.002328	0.001151	N/A	320	206	246
San Andreas - Pallett Creek	34.4556	-117.8870	0.007353	0.009346	0.007353	0.002495	0.002442	N/A	N/A	107	136
San Andreas - Pitman Canyon	34.2544	-117.4340	0.004545	0.008265	0.006494	0.002679	0.002398	N/A	220	121	154
San Andreas - Plunge Creek	34.1158	-117.1370	0.002083	0.002770	0.002941	0.004005	0.000891*	N/A	480	361*	169
Mission Creek - 1000 Palms	33.8200	-116.3010	0.002941	0.005618	0.004237	0.002108	0.000695	N/A	340	178	236
San Andreas - Wrightwood	34.3697	-117.6680	0.010204	0.012500	0.010204	0.005476	0.002174	N/A	N/A	80	98

*from Monte Carlo model

Table 6. Earthquake rates and estimated 1-s uncertainties calculated from paleoseismic series from 23 California sites. "LN" stands for lognormal, and "BPT" stands for Brownian Passage Time. "N/A" denotes cases where a particular method was not used to make calculations. Southern San Andreas site calculations and the Vendanta and Hog Lake sites were made by Biasi et al. [2007]. Methods and values are the same as in Biasi et al. [2002] except for an update to current open intervals.

Site	Lat	Lon	Poisson rate	Est. symmetric σ	16.5%	83.5%	2.5%	97.5%
Calaveras fault - North	37.5104	-121.8346	0.001252	0.001955	0.000862	0.002703	0.000439	0.004348
Elsinore - Glen Ivy	33.7701	-117.4909	0.004310	0.005444	0.003226	0.008333	0.001613	0.012500
Elsinore Fault - Julian	33.2071	-116.7273	0.000539	0.001141	0.000325	0.001316	0.000219	0.002500
Elsinore - Temecula	33.4100	-117.0400	0.000543	0.000972	0.000357	0.001136	0.000230	0.002174
Elsinore - Whittier	33.9303	-117.8437	0.000519	0.001111	0.000313	0.001266	0.000217	0.002439
Garlock - Central	35.4441	-117.6815	0.000624	0.000830	0.000455	0.001190	0.000262	0.001923
Garlock - Western	34.9868	-118.5080	0.000619	0.000868	0.000422	0.001205	0.000264	0.002000
Hayward fault - North	37.9306	-122.2977	0.002915	0.002905	0.002174	0.005000	0.001333	0.007143
Hayward fault - South	37.5563	-121.9739	0.005291	0.004085	0.004348	0.008333	0.002941	0.011111
N. San Andreas - Vendanta	38.0320	-122.7891	0.004237	0.002400	N/A	N/A	0.002193	0.007407
SAF - Arano Flat	36.9415	-121.6729	0.007092	0.006481	0.005556	0.012500	0.003704	0.016667
N. San Andreas - Fort Ross	38.5200	-123.2400	0.002193	0.003462	0.001493	0.005000	0.000769	0.007692
San Gregorio - North	37.5207	-122.5135	0.000754	0.002153	0.000442	0.002273	0.000240	0.004545
San Jacinto - Hog Lake	33.6153	-116.7091	0.004149	0.002316	N/A	N/A	0.002273	0.006993
San Jacinto - Superstition	32.9975	-115.9436	0.001812	0.004288	0.001205	0.005000	0.000515	0.009091
Site	**Lat**	**Lon**	**Poisson rate**	**Est. symmetric σ**	**16.5%**	**83.5%**	**2.5%**	**97.5%**
San Andreas - Burro Flats	33.9730	-116.8170	0.002381	0.002494	0.003125	0.007692	0.002174	0.011111
SAF - Combined Carrizo Plain	35.1540	-119.7000	0.003571	0.001924	0.002128	0.005556	0.001235	0.008333
San Andrteas - Indio	33.7414	-116.1870	0.003125	0.002328	0.001818	0.007143	0.000855	0.011111
San Andreas - Pallett Creek	34.4556	-117.8870	0.007353	0.002495	N/A	N/A	0.003534	0.013514
San Andreas - Pitman Canyon	34.2544	-117.4340	0.004545	0.002679	0.004000	0.009091	0.002273	0.012500
San Andreas - Plunge Creek	34.1158	-117.1370	0.002083	0.004005	0.002083	0.008333	0.000676	0.012500
Mission Creek - 1000 Palms	33.8200	-116.3010	0.002941	0.002108	0.002000	0.006667	0.001075	0.010000
San Andreas - Wrightwood	34.3697	-117.6680	0.010204	0.005476	N/A	N/A	0.005714	0.016667

Table 7. Earthquake rates and confidence intervals developed from Tables 2, 6, and 7. These rates are used to check independent earthquake rate modes for consistency with paleoseismic data.

An additional table (Table 7) is provided here that is used for assessing whether earthquake rates determined from slip-rate-constrained solutions are consistent within 95% confidence intervals with paleoseismic rates. Thus the numbers in Table 7 are not used to constrain a weighted inversion, but are instead used as guidelines to ensure that important data are not violated.

6. Aperiodicity determination for time-dependent probability calculations for Working Group on California Earthquake Probabilities

This section describes methods for suggested segment aperiodicity values from analysis of paleoseismic data. Description is limited to interpretation of Monte Carlo analysis results discussed in previous sections. Output from Monte Carlo analysis generated 3-D distributions of likelihood vs. recurrence interval vs. aperiodicity (Figures 8-10). From distributions like those depicted in Figures 8-10, each aperiodicity-recurrence interval pairing has a relative weight. Ideally, probability calculations could be made for every pair and given an appropriate weight. However, given the number of other logic tree branches under consideration, such an exercise is untenable. Therefore for each of 20 sites that could be analyzed (Figure 13) using the Monte Carlo method, a most-likely aperiodicity value was chosen, as well as 67% and 95% confidence intervals (Table 8).

As described in Section 3, there is some functional dependence between aperiodicity and recurrence interval that may need to be considered. Dependence amongst parameters is varies from site to site, and can be assessed by examining Figure 14, which shows mean recurrence interval vs. aperiodicity.

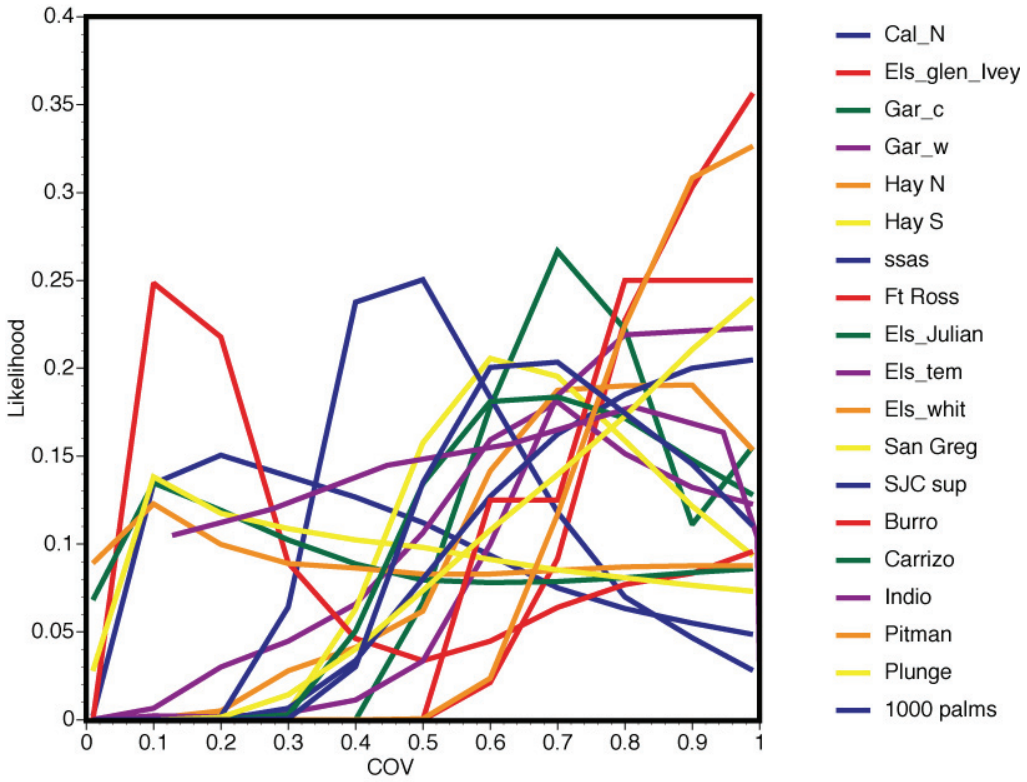

Figure 13. Relative likelihood of aperiodicity for 20 California paleoseismic sites determined from Monte Carlo analysis.

Site	Lat	Lon	COV	2.5%	97.5%	16.5%	83.5%
Calaveras fault - North	37.5104	-121.8346	0.2	0.2	0.9	0.3	0.8
Elsinore - Glen Ivy	33.7701	-117.4909	0.8	0.7	1.0	0.8	1.0
Elsinore Fault - Julian	33.2071	-116.7273	0.5	0.0	1.0	0.2	0.8
Elsinore - Temecula	33.4100	-117.0400	0.8	0.5	1.0	0.7	1.0
Elsinore - Whittier	33.9303	-117.8437	0.5	0.0	1.0	0.2	0.9
Garlock - Central	35.4441	-117.6815	0.6	0.5	1.0	0.6	0.9
Garlock - Western	34.9868	-118.5080	0.7	0.3	1.0	0.5	0.9
Hayward fault - North	37.9306	-122.2977	0.7	0.4	1.0	0.6	1.0
Hayward fault - South	37.5563	-121.9739	0.2	0.0	0.6	0.0	0.5
N. San Andreas - North Coast	38.0320	-122.7891	N/A	N/A	N/A	N/A	N/A
SAF - Arano Flat	36.9415	-121.6729	0.4	0.3	0.9	0.4	0.8
N. San Andreas - Fort Ross	38.5200	-123.2400	0.2	0.1	1.0	0.2	0.9
San Gregorio - North	37.5207	-122.5135	0.2	0.1	1.0	0.2	0.9
San Jacinto - Hog Lake	33.6153	-116.7091	N/A	N/A	N/A	N/A	N/A
San Jacinto - Superstition	32.9975	-115.9436	0.8	0.5	1.0	0.6	1.0
San Andreas - Burro Flats	33.9730	-116.8170	0.7	0.6	1.0	0.7	1.0
SAF - Combined Carrizo Plain	35.1540	-119.7000	0.6	0.4	1.0	0.6	1.0
San Andrteas - Indio	33.7414	-116.1870	0.7	0.5	1.0	0.7	0.8
San Andreas - Pallett Creek	34.4556	-117.8870	N/A	N/A	N/A	N/A	N/A
San Andreas - Pitman Canyon	34.2544	-117.4340	0.9	0.6	1.0	0.8	1.0
San Andreas - Plunge Creek	34.1158	-117.1370	0.7	0.4	1.0	0.6	1.0
Mission Creek - 1000 Palms	33.8200	-116.3010	0.6	0.4	1.0	0.6	1.0
San Andreas - Wrightwood	34.3697	-117.6680	N/A	N/A	N/A	N/A	N/A

Table 8. Estimated aperiodicity values for California paleoseismic sites. The mean of all sites is 0.6. However there is significant variation between sites at the 67% confidence level.

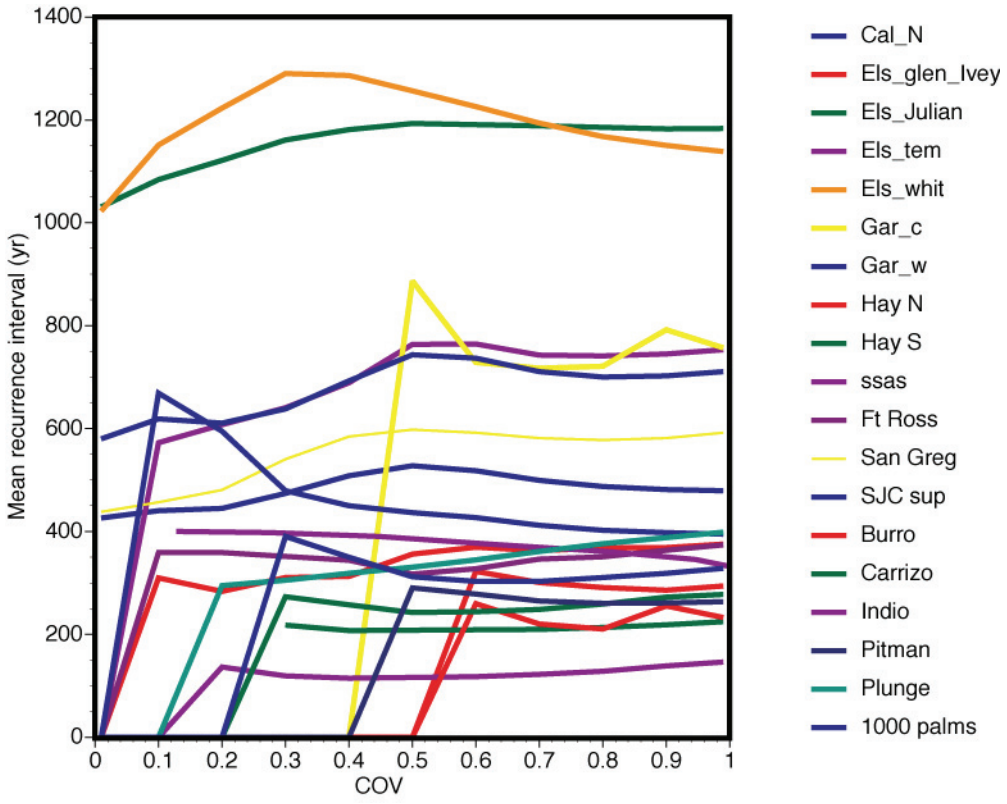

Figure 14. Mean recurrence interval vs. aperiodicity from analysis of 20 California paleoseismic sites. Variable dependence between the two parameters among different sites is evident.

7. Conclusions

A method for estimating most-likely recurrence parameters from short paleoseismic observations is explored. Sample means from short series that are expected to have asymmetric recurrence distributions will tend to reflect the distribution medians rather than the means, and could underestimate recurrence intervals (and thus overestimate hazard). Benchmarking tests on short, synthetic paleoearthquake catalogs indicate that the Monte Carlo methods can give results more reflective of the underlying distribution mean. Monte Carlo draws from every reasonable recurrence PDF of an assumed class are tested for consistency with observed paleoearthquake series. Those models that can reproduce observations within dating uncertainties are tallied, and the mean of PDF's that produce the most fits to observed is taken as most likely. Very long observed sequences with tight age constraints are difficult to reproduce with any specific PDF because of computational limitations; thus the proposed method is most useful in extracting recurrence information from short (•10 events) series, and/or sequences with poor age control. The method was applied to 19 paleoseismic sites across California using time-independent and time-dependent PDF's. Relative weights were calculated for recurrence parameters from even the sparsest catalogs, which are proposed to be used for weighting logic-tree branches in earthquake probability calculations.

Acknowledgments.

This paper benefited from internal USGS and Working Group on California Earthquake Probabilities (WGCEP) reviews by, and discussion with, Allin Cornell, Bill Ellsworth, Eric

Geist, Jeanne Hardebeck, Dave Jackson, and Jim Lienkaemper. Thanks also to JGR reviewers Glenn Biasi and Yan Kagan, and an anonymous AE who all gave very thorough reviews. Special thanks are due to Tim Dawson, who assembled the database of California paleoseismic intervals for WGCEP.

References Cited

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, 2002.

Davis, P. M., D. D. Jackson, and Y. Y. Kagan (1989), The longer it has been since the last earthquake, the longer the expected time till the next?, *Bull. Seismol. Soc. Am. 79,* 1439-1456.

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*, doi:10.1029/2001JB001744.

Ellsworth, W. L., Matthews, M. V., Nadeau, R. M., Nishenko, S. P., Reasenberg, P. A., and Simpson, R. W. (1999) A physically-based earthquake recurrence model for estimation of long-term earthquake probabilities, *U. S. Geol. Surv. Open File Rep. OF99-520,* 22pp.

Frankel, A. D., M. D. Petersen, C. S. Mueller, K. M. Haller, R. L. Wheeler, E.V. Leyendecker, R. L. Wesson, S. C. Harmsen, C. H. Cramer, D. M. Perkins, and K. S. Rukstales (2002), Documentation for the 2002 Update of the National Seismic Hazard Maps, U. S. Geol. Surv. Open File Rep., OFR-02-420, 33pp.

Fumal, T. E., R. J. Weldon II, G. P. Biasi, T. E. Daeson, G. G. Seitz, W. T. Frost, and D. P. Schwartz (2002a) Evidence for large earthquakes on the San Andreas fault at the Wrightwood, California, Paleoseismic site: A.D. 500 to present, *Bull. Seismol. Soc. Am. 92,* 2726-2760.

Fumal, T. E., M. J. Rymer, and G. G. Seitz (2002b), Timing of large earthquakes since A.D. 800 on the Mission Creek strand of the San Andreas fault zone at Thousand Palms Oasis, near Palm Springs, California, *Bull. Seismol. Soc. Am. 92,* 2841-2860.

Fumal, T. E., G. F. Heingartner, T. E. Dawson, R. Flowers, J. C. Hamilton, J. Kessler, L. M. Reidy, L. Samrad, G. G. Seitz, and J. Southon (2003), A 100-Year Average Recurrence Interval for the San Andreas Fault, Southern San Francisco Bay Area, California, EOS Trans, S12B-0388.

Grant, L. B., and M. M. Gould (2004), Assimilation of paleoseismic data for earthquake simulation, *Pure Appl. Geophys. 161*, 2295-2306 doi:10.1007/s00024-003-2564-8.

Grant, L. B., and W. R. Lettis (2002) Introduction to the special issue on paleoseismology of the San Andreas fault system, *Bull. Seismol. Soc. Am. 92,* 2551-2554.

Grant, L. B. and K. Sieh (1994), Paleoseismic Evidence of Clustered Earthquakes on the San Andreas Fault in the Carrizo Plain, California, *J. Geophys. Res., 99*, 6819-6841.

Gurrola, L.D. and T. K. Rockwell (1996), Timing and slip for prehistoric earthquakes on the Superstition Mountain fault, Imperial Valley, southern California, *J. Geophys. Res., 101,* 5,977-5,985.

Hagiwara, Y., Probability of earthquake occurrence as obtained from a Weibull distribution analysis of crustal strain, *Tectonophysics*, *23*, 323-318, 1974.

Kagan, Y. Y., and D. D. Jackson (1999), Worldwide doublets of large shallow earthquakes, *Bull. Seismol. Soc. Am. 89,* 1147-1155.

Kagan, Y. Y., and L. Knopoff, Random stress and earthquake statistics: time dependence, *Geophys. J. R. Astron. Soc. 88*, 723-731, 1987.

Kelson, K. I., G. D. Simpson, W. R. Lettis, and C. C. Haraden (1996), Holocene slip rate and recurrence of the northern Calaveras fault at Leyden Creek, eastern San Francisco Bay region, *J. Geophys. Res., 101*,. 5961-5975.

Lienkaemper, J., P. Williams, T. Dawson, S. Personius, G. Seitz, S. Heller, and D. Schwartz (2003) Logs and data from trenches across the Hayward Fault at Tyson's Lagoon (Tule Pond), Fremont, Alameda County, California: *U.S. Geol. Surv. Open-File Rep. 03-488*, 6 p., 8 plates. http://geopubs.wr.usgs.gov/open-file/of03-488/, 2003

Liu, J., Y. Klinger, K. Sieh, and C. Rubin (2004), Six similar sequential ruptures of the San Andreas fault, Carrizo Plain, California, Geology, 32, 649-652.

Matthews, M. V., W. L. Ellsworth, and P. A. Reasenberg, A Brownian Model for Recurrent Earthquakes, *Bull. Seismol. Soc. Am., 92,* 2233 - 2250, 2002.

McGill, S., S. Dergham, K. Barton, T. Berney-Ficklin, D. Grant, C. Hartling, K. Hobart, R. Minnich, M. Rodriguez, E. Runnerstrom, J. Russell, K. Schmoker, M. Stumfall, J. Townsend and J. Williams (2002), Paleoseismology of the San Andreas Fault at Plunge Creek, near San Bernardino, Southern California, *Bull. Seismol. Soc. Am. 92,* 2803-2840*,* doi:10.1785/0120000607.

Nishenko, S. P., and R. Buland, A generic recurrence interval distribution for earthquake forecasting, *Bull. Seismol. Soc. Am., 77*, 1382-1399, 1987.

NIST/SEMATECH (2006), e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/.

Ogata, Y. (1999), Estimating the hazard of rupture using uncertain occurrence times of paleoearthquakes, *J. Geophys. Res., 104,* 17995-18014.

Page, M. T., and J. M. Carlson (2006), Methodologies for earthquake hazard assessment: Model uncertainty and the WGCEP-2002 forecast, *Bull. Seismol. Soc. Am., 96*, 1624–1633, doi:10.1785/0120050195.

Parsons, T., S. Toda, R. S. Stein, A. Barka, and J. H. Dieterich (2000), Heightened odds of large earthquakes near Istanbul: an interaction-based probability calculation, *Science, 288*, 661-665.

Parsons, T., 2005, Significance of stress transfer in time-dependent earthquake probability calculations, *J. Geophys. Res., 110*, doi:10.1029/2004JB003190.

Savage, J. C., Criticism of some forecasts of the National Earthquake Prediction Evaluation Council, *Bull. Seismol. Soc. Am. 81*, 862-881, 1991.

Savage, J. C., The uncertainty in earthquake conditional probabilities, *Geophjys. Res. Lett., 19,* 709-712, 1992.

Savage, J.C. (1994), Empirical earthquake probabilities from observed recurrence intervals, *Bull. Seismol. Soc. Am. 84,* 219-221.

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.

Sieh, K. E. (1986), Slip rate across the San Andreas fault and prehistoric earthquakes at Indio, California (abstract), EOS, Trans. Amer. Geophys. Union, 67, 1200.

Sieh, K., M. Stuiver, and D. Brillinger, A more precise chronology of earthquakes produced by the San Andreas fault in southern California, *J. Geophys. Res., 94*, 603-623, 1989.

Seitz, G., G. Biasi, and R. Weldon (1996) The Pitman Canyon paleoseismic record suggests a re-evaluation of San Andreas Fault segmentation, Proceedings of the Symposium on Paleoseismology, International Union for Quaternary Research XIV International Congress, Terra Nostra 2/95, eds., Michetti, A.M., and P.L. Hancock, Journal of Geodynamics.

Seitz, G., G. Biasi, and R. Weldon (2000), An improved paleoseismic record of the San Andreas fault at Pitman Canyon, Quaternary Geochronology Methods and Applications, American Geophysical Union Reference Shelf 4, Noller, J. S., et al., eds., p. 563 - 566.

Simpson, G. D., J. N. Baldwin, K. I. Kelson, and W. R. Lettis (1999), Late holocene slip rate and earthquake history for the northern Calaveras fault at Welch Creek, eastern San Francisco Bay area, California, *Bull. Seismol. Soc. Am., 89,* 1250-1263.

Simpson, G. D., S. C. Thompson, J. S. Noller, and W. R. Lettis (1997), The northern San Gregorio fault zone: evidence for the timing of late Holocene earthquakes near Seal Cove, California, *Bull. Seismol. Soc. Am., 87*, 1158-1170.

Sims, J. D. (1994), Stream channel offset and abandonment and a 200-year average recurrence interval of earthquakes on the San Andreas fault at Phelan Creeks, Carrizo Plain, California, in Proceedings of the workshop on paleoseismology, US Geological Survey Open-File Report 94-568, (edited by C. S. Prentice, D. P. Schwartz and R. S. Yeats), 170-172.

StataCorp (2005), *Stata Statistical Software: Release 9*, College Station, TX, StataCorp LP.

Sykes, L. R., and W. Menke (2006), Repeat times of large earthquakes: implications for earthquake mechanics and long-term prediction, *Bull. Seismol. Soc. Am., 96*, 1569–1596, doi:10.1785/0120050083.

Vaughan, P. R., K. M. Thorup, and T. K. Rockwell (1999), Paleoseismology of the Elsinore fault at Agua Tibia Mountain southern California, *Bull. Seismol. Soc. Am., 89,* 1447-1457.

Weldon, R., K. Scharer, T. Fumal, and G. Biasi (2004), Wrightwood and the earthquake cycle: what the long recurrence record tells us about how faults work, GSA Today, 14, 4-10, doi:10.1130/1052-5173(2004)014.

Working Group on California Earthquake Probabilities (2003), Earthquake probabilities in the San Francisco Bay region: 2002 to 2031, *USGS Open-File Report 03-214*.

Wu, S.-C., C. Cornell, and S. R. Winterstein (1991), Estimation of correlations among characteristic earthquakes, Proc. 4th Int. Conf. Seismic Zonation, Stanford, California, 11-18.

Yule, D., and K. Sieh (2001), The paleoseismic record at Burro Flates: Evidence for a 300-year average recurrence of large earthquakes on the San Andreas fault in San Gorgonio Pass, southern California, Cordilleran Section - 97th Annual Meeting, and Pacific Section, American Association of Petroleum Geologists (April 9-11, 2001).

Zhang, H., T. Niemi, and T. Fumal (2006), A 3000-year record of earthquakes on the northern San Andreas fault at the Vedanta marsh site, Olema, California, *Seismological Research Letters, 77*, 176.