

**1 Statistical tests on clustered global earthquake
2 synthetic datasets**

Eric G. Daub,¹ Daniel T. Trugman,^{2,3} and Paul A. Johnson²

Corresponding author: E. G. Daub, Center for Earthquake Research and Information, Memphis, TN 38152, USA (egdaub@memphis.edu)

¹Center for Earthquake Research and Information, University of Memphis, Memphis, TN, USA

²Geophysics Group, Los Alamos National Laboratory, Los Alamos, NM, USA

³Scripps Institution of Oceanography, University of California, San Diego, La Jolla, CA, USA

Abstract. We study the ability of statistical tests to identify nonrandom features of earthquake catalogs, with a focus on the global earthquake record since 1900. We construct four types of synthetic datasets containing varying strengths of clustering, with each dataset containing on average 10000 events over 100 years with magnitudes above $M = 6$. We apply a suite of statistical tests to each synthetic realization in order to evaluate the ability of each test to identify the sequences of events as nonrandom. Our results show that detection ability is dependent on the quantity of data, the nature of the type of clustering, and the specific signal used in the statistical test. Datasets that exhibit a stronger variation in the seismicity rate are generally easier to identify as nonrandom for a given background rate. We also show that we can address this problem in a Bayesian framework, with the clustered datasets as prior distributions. Using this new Bayesian approach, we can quantitatively bound the range of possible clustering strengths that are consistent with the global earthquake data. At $M = 7$, we can estimate 99th percentile confidence bounds on the number of triggered events, with an upper bound of 20% of the catalog for global aftershock sequences, with a stronger upper bound on the fraction of triggered events of 10% for long-term event clusters. At $M = 8$, the bounds are less strict due to the reduced number of events. However, our analysis shows that other types of clustering could be present in the data that we are unable to detect. Our results aid in the interpretation of the results of statistical tests on earthquake catalogs, both worldwide and regionally.

1. Introduction

26 The occurrence of a number of earthquakes above magnitude $M = 8$ in the past decade
27 has led to speculation that large earthquakes cluster in time. The recent occurrence of
28 earthquakes includes three of the six largest earthquakes on record, including the 2004
29 $M_W = 9.2$ Sumatra earthquake, the 2010 $M_W = 8.8$ Maule, Chile, earthquake, and the
30 2011 $M_W = 9.1$ Tohoku earthquake, and a number of additional events above $M = 8$, as
31 can be seen from the earthquake record shown in Fig. 1(a). These large events can have
32 an outsized effect on seismic hazard through their large release of stored elastic energy,
33 causing destructive strong ground motions and tsunamis. If large events do cluster in
34 time, this could change the way that seismic hazard is estimated worldwide.

35 To evaluate this hypothesis, a number of studies have compared the global earthquake
36 record since 1900 to a process that is random in time [*Bufe and Perkins, 2005; Michael,*
37 *2011; Shearer and Stark, 2012; Daub et al., 2012; Parsons and Geist, 2012; Ben-Naim*
38 *et al., 2013*]. A process that is random in time is also known as a time-homogeneous
39 Poisson process, and such a process assumes that the event occurrence times are uncorre-
40 lated. The majority of these studies tend to show that earthquake occurrence worldwide
41 since 1900 shows no deviation from a process that is random in time, other than localized
42 aftershock sequences. This is illustrated in Fig. 1(b) for the *Ben-Naim et al. [2013]* study,
43 showing the likelihood that the catalog is random (through a p -value, calculated using
44 Monte Carlo simulation) for several magnitude thresholds with and without aftershock
45 removal. While some of the statistical tests applied to the catalog appear to show devia-
46 tions from random event occurrence at minimum magnitude levels $M = 8.4 - 8.6$ [*Bufe*

47 *and Perkins, 2005; Ben-Naim et al., 2013*], such as in Fig. 1 at $M \geq 8.4-8.5$, these devia-
48 tions are not strong enough to conclude that the earthquake record is nonrandom. This is
49 because tests using magnitude thresholds selected *a posteriori* underestimate p -values, as
50 shown by *Shearer and Stark [2012]*. The global catalog over this time period is regarded
51 as complete only for $M \geq 7$, thus the catalogs in these studies contain a relatively small
52 number of events, particularly at high magnitude levels.

53 More recent seismic catalogs have been used to examine the ability of large earthquakes
54 to trigger earthquakes above $M = 5$. One study showed that the 2012 $M_w = 8.6$ Indian
55 Ocean event triggered aftershocks above $M = 5$ worldwide [*Pollitz et al., 2012*] followed
56 by a quiescent period [*Pollitz et al., 2014*]. A more comprehensive study looking at many
57 events above $M = 7$ showed that triggering of this nature may not be common [*Parsons*
58 *and Velasco, 2011*], while another more recent study concluded that there is no evidence
59 for elevated seismicity rates from $M = 5.2 - 5.6$ over recent years [*Parsons and Geist,*
60 *2014*]. Further, the *Parsons and Velasco [2011]* study showed that instantaneous triggering
61 (i.e. triggering coincident with surface wave arrivals) of larger events above $M = 5$ is
62 not observed as frequently as observed triggering rates for smaller $M < 5$ earthquakes
63 would predict [*Velasco et al., 2008*]. The observational evidence thus suggests that small
64 events are routinely triggered [*Hill et al., 1993; Gomberg et al., 2004; Freed, 2005*], and
65 there is some evidence large earthquakes can have global effects on moderate-sized events.
66 However, the likelihood that large events trigger other moderately large events on a global
67 scale remains unclear.

68 In this study, we address the ability of statistical tests to identify catalogs as nonrandom,
69 with a particular focus on the global dataset since 1900. We produce a series of simu-

lated datasets that are clustered by construction, and systematically vary the strength
of the clustering to assess how well different statistical tests can identify these datasets
as nonrandom. While this question has been addressed previously in a study by *Dimer
de Oliveira* [2012], here we perform a more systematic study of various dataset types
with clustering of different strengths. In particular, our study aims to bound the range
of clustering strengths that are most likely to be consistent with the global earthquake
catalog. Through these synthetic tests, we can assess the results of the numerous studies
performed on the global earthquake record and interpret the implications for earthquake
interaction and earthquake hazard worldwide.

2. Synthetic Datasets

We generate four types of synthetic datasets that are nonrandom by construction for
analysis using various statistical tests. The synthetic datasets are each designed to be
similar to the global earthquake record since 1900, with a few simplifications. The sim-
ulations are all 100 years in length, and contain an average of 10000 events above a
minimum magnitude threshold of $M = 6$, illustrated in Fig. 2. Event magnitudes are
drawn from a Gutenberg-Richter distribution [*Gutenberg and Richter*, 1954], with a cu-
mulative distribution function $CDF(M) \propto 10^{-bM}$ with $b = 1$, a minimum magnitude of
 $M_{min} = 6$, and a maximum magnitude of $M_{max} = 9.5$, with a few additional simulations
varying these parameters described below. The magnitude frequency statistics of this dis-
tribution are shown in Fig. 3. This distribution shows that the datasets contain ~ 10000
events at $M \geq 6$, ~ 1000 events with $M \geq 7$, and ~ 100 events with $M \geq 8$. All of these
numbers differ slightly from the observed values in the earthquake record since 1900 –
the PAGER catalog [*Allen et al.*, 2009] supplemented with the United States Geological

92 Survey (USGS) Preliminary Determination of Epicenters (PDE) catalog contains ~ 1800
93 events above $M = 7$ over 115 years.

94 The combined PAGER and PDE catalog is estimated to be complete for $M \geq 7$ using
95 both the Maximum Curvature and b -value Stability Methods [*Woessner and Wiemer,*
96 2005]. With this completeness magnitude, $b = 1.26$ in the observed magnitude-frequency
97 distribution using the Maximum Likelihood Method [*Aki, 1965*]. However, an analysis
98 of the more recent International Seismological Centre Global Earthquake Model catalog
99 [*Storchak et al., 2013*] shows that for events in that catalog since 1917, $b \approx 1$ [*Michael,*
100 2014]. For the PAGER global catalog, the number of independent events above $M = 7$
101 varies from $\sim 500 - 1500$, depending on the criteria used in aftershock identification
102 [*Michael, 2011; Shearer and Stark, 2012; Daub et al., 2012*]. Because a different choice of
103 declustering algorithm can always alter the number of events in the catalog, for simplicity
104 we use $b = 1$ with a rate of 100 events/year to produce a round number of events at various
105 magnitude levels for the majority of our simulations. To evaluate if our results are sensitive
106 to the details of the magnitude distribution, we also perform additional simulations with
107 different b -values of 0.8 and 1.2, and a set of simulations with a minimum magnitude of
108 $M_{min} = 5$.

109 For each type of dataset, we create 20 different versions, varying the clustering strength.
110 The background rate ranges from 98 events/year in the least clustered version to 22
111 events/year in the most clustered version. According to each prescription below, we
112 either explicitly add additional events according to certain rules (the case for the ETAS
113 simulations), or we vary the seismicity rate from event to event using certain rules (the
114 case for the other three synthetic datasets). In each case, parameters are selected such

115 that our simulations produce on average 100 events/year over a fixed 100 year duration.
 116 We produce 10000 realizations of each of the 20 different clustering strengths of the four
 117 dataset types. Thus, our results here are based on analyzing a total of 800,000 simulations
 118 (4 clustering types \times 20 clustering strengths \times 10,000 realizations).

2.1. ETAS Simulations

119 Epidemic Time Aftershock Sequence (ETAS) models have been developed over many
 120 years to represent empirically observed features of aftershock sequences [*Ogata, 1998*].
 121 Because of the widespread use of these types of simulations to represent clustering of
 122 seismicity, we generate a series of datasets containing “global aftershock sequences” to
 123 represent one potential type of clustering in our analysis. In the ETAS models, events
 124 are added as aftershocks of background events according to two empirical rules. First,
 125 an aftershock productivity law determines the number of aftershocks produced by a main
 126 shock of magnitude M [*Felzer et al., 2004; Helmstetter et al., 2005*]:

$$127 \quad N_{AS} = C' 10^{\alpha(M-M_{min})}. \quad (1)$$

128 The constant α determines the relative number of aftershocks an earthquake of a given
 129 magnitude produces, and C' , known as the aftershock productivity, determines the overall
 130 number. Observations show that $\alpha \approx 1$, and C' is such that a main shock typically
 131 produces a maximum aftershock magnitude one magnitude unit less than the main shock
 132 (Båth’s law) [*Helmstetter and Sornette, 2003*]. Here, we use $\alpha = 1$, but vary the aftershock
 133 productivity to produce simulations with different levels of clustering.

134 The second empirical rule for the ETAS models is the Omori decay of aftershock rate
 135 with time following the main shock [*Omori*, 1895; *Utsu et al.*, 1995]:

$$136 \quad R(t) = \frac{A}{(c+t)^\beta}. \quad (2)$$

137 In the Omori law, β describes the time decay of aftershock activity, and c is a constant
 138 to make the rate finite at $t = 0$. We use $\beta = 1.07$ and $c = 3 \times 10^{-4}$ years = 0.11 days,
 139 which are similar to ETAS parameters for Japan [*Guo and Ogata*, 1997]. The constant A
 140 is chosen such that Eq. (2) integrates to N_{AS} :

$$141 \quad N_{AS} = \int_0^{t_{max}} R(t) dt. \quad (3)$$

142 We cut off the aftershock decay after $t_{max} = 100$ years since that would exceed the length
 143 of our simulation. This truncation of the aftershock sequence is strictly necessary only if
 144 $\beta \leq 1$.

145 Aftershocks in the ETAS model also produce their own aftershocks, so Eqs. (1) and
 146 (2) are applied recursively to all aftershocks until no further aftershocks are produced.
 147 Note that aftershocks follow the same magnitude-frequency statistics as the main shocks,
 148 which means that the majority of aftershocks are small events. However, occasionally
 149 an event triggers an aftershock whose magnitude exceeds the parent event. In this case,
 150 the initial event is considered to be a foreshock. A consequence of this is that not all
 151 sets of ETAS parameters produce aftershock sequences that terminate [*Helmstetter and*
 152 *Sornette*, 2002], and so care must be taken in selecting parameters.

153 Because the ETAS models are event-based, we set the background rate to the desired
 154 level, and then choose C' such that the synthetic datasets contain a total of 10000 events
 155 in 100 years on average. Specific values of the rate and the aftershock productivity C'

156 are shown in Table 2. Because aftershock sequences can extend over long periods of time
157 for the highest values of aftershock productivity, we ensure that the synthetic datasets
158 are uniform in time by generating 200 years of background events and only selecting the
159 final 100 years of the event sequence for analysis. This ensures that we are not “missing”
160 events whose main shock may have occurred prior to the start of the simulation.

161 ETAS models can also include spatial kernels to simulate clustering of seismicity in space
162 [*Felzer and Brodsky, 2006*]. Because most statistical tests applied to the global catalog
163 neglect spatial information, we do not include this effect in our simulations. However,
164 spatial information is implicitly included in statistical tests on the global catalog through
165 removal of aftershocks, so a more realistic way to treat the ETAS data for our purposes
166 might be to perform a spatial ETAS simulation and then remove aftershocks prior to
167 testing. Due to the large number of simulations considered here and the introduction
168 of additional parameters for both generating and removing aftershocks, we neglect this
169 aspect in our study. Additionally, we note that our simulations approximate this spatial
170 effect, due to the fact that the Omori decay of the aftershock rate with time is assumed
171 to be independent of spatial location. Thus, an ETAS simulation that uses a reduced
172 aftershock productivity can be thought of as a proxy for a spatial simulation where events
173 in the traditional aftershock zone are removed. This leaves a reduced number of global
174 “aftershocks” in the simulated dataset that follow the same time decay of seismicity rate
175 as the traditional aftershocks.

176 A sample ETAS simulation for the strongest level of clustering is shown in Fig. 2(a),
177 illustrating the seismicity rate as a function of time. Because large earthquakes have a
178 pronounced effect on the rate (Eqs. (1)-(2)), a localized spike in the earthquake rate

179 occurs after each large event. The rate quickly decays with time after each event, though
 180 extended aftershock sequences can occur that keep the rate elevated above background
 181 for longer periods of time. Due to the sharp peak in the rate following large events, the
 182 ETAS models exhibit a higher variability in rate when compared to the other simulated
 183 datasets in this study.

2.2. Magnitude-Dependent Simulations

184 The second type of dataset incorporates magnitude-dependent clustering, designed to
 185 be somewhat similar to the ETAS models, yet different in its time dependence. It is based
 186 on the idea that earthquake triggering is related to the strain amplitude of seismic waves,
 187 combined with a model where any earthquake can be a potential precursor to future
 188 seismicity with a rate contribution dependent on magnitude [*Rhoades and Evison, 2004*].
 189 In particular, a study by *van der Elst and Brodsky [2010]* showed that the seismicity rate
 190 increase could be quantitatively tied to wave strain amplitude using a statistical method
 191 applied to a large earthquake catalog. We construct a synthetic dataset based on this
 192 concept, where the seismicity rate at a given time is the sum of the background rate,
 193 plus a variable contribution that depends on the weighted magnitude of the previous 200
 194 events. The seismicity rate λ_i for the time period following event i is given by

$$195 \quad \lambda_i = \lambda_0 + \gamma \left[\sum_{j=i-199}^i 10^{\alpha(M_j - M_{min})} - 200 \right]. \quad (4)$$

196 The background seismicity rate is λ_0 , $\alpha = 1$ determines the relative contribution of earth-
 197 quakes of different magnitude in the same manner as the ETAS models, and γ determines
 198 the overall rate contribution from the triggering effect. Seismic wave amplitudes scale
 199 exponentially with magnitude as $\sim 10^M$ [*Lay and Wallace, 1995*], thus we apply a weight-

200 ing factor that depends exponentially on magnitude to determine their strain amplitude
201 contribution. As with the ETAS models, spatial information is neglected. Finally, sub-
202 tracting the factor of 200 ensures that the rate increase is measured relative to a baseline
203 level where all events have magnitude $M = 6$. Following a large event, the seismicity rate
204 exhibits an approximate step increase, and remains elevated for a longer period of time
205 when compared to the ETAS simulations (Fig. 2(a)-(b)). The additive combination of
206 past events on top of the background seismicity is the same as in the model of *Rhoades*
207 *and Evison* [2004], though with a cutoff after a limited number of events.

208 The duration over which far-field dynamic triggering occurs is not well quantified by ob-
209 servations, as event statistics are usually accumulated over longer time periods to establish
210 a change in seismicity rate [*Freed*, 2005]. Some studies suggest that near-field aftershocks
211 may also be triggered by dynamic stresses [*Gomberg et al.*, 2003], suggesting an Omori
212 time dependence for the triggering effect. Because the ETAS simulations already con-
213 sider the case of clustering following an Omori decay of seismicity rate with time, we use
214 the Magnitude-Dependent simulations to consider an alternative form of a seismicity rate
215 increase that is not as pronounced in time. *van der Elst and Brodsky* [2010] show that
216 a step rate change following the triggering event predicts similar triggering statistics to
217 an Omori decay, and thus we use Eq. (4) to generate this rate pattern in this simulated
218 dataset. The convolution over the previous 200 magnitudes is aimed at making this step
219 change last from 1-2 years, depending on the size of the the rate increase. The choice of
220 200 events (rather than a specific period of time) is also made to ensure that events with
221 $M \geq 8$ influence the seismicity rate over multiple $M \geq 8$ recurrence times (since 1 out of
222 every 100 events will have a magnitude in that range).

223 For the Magnitude-Dependent simulations, we generate 15200 events with magnitudes
224 following the GR distribution, and then perform the sum over event magnitudes to de-
225 termine the event-by-event seismicity rate (Eq. (4)). We then generate recurrence times
226 in an event-by-event fashion by drawing from an exponential distribution (the expected
227 waiting time distribution for a Poisson process) given λ_i . Finally, we sum the recurrence
228 times to get the occurrence times, and remove the first 200 events to ensure uniformity in
229 time. We then select 100 years of events for analysis, and discard events that occur after
230 100 years. We produce 5000 additional events beyond the typical 10000 to ensure that
231 our simulations are never shorter than 100 years in the event of a large number of short
232 interevent times. The coefficient γ is determined for each of the 20 synthetic datasets in
233 order to produce simulations that contain on average 10000 events in 100 years. Parameter
234 values are shown in Table 2.

235 An example Magnitude-Dependent synthetic dataset is shown in Fig. 2(b) for the high-
236 est clustering strength. When compared to the ETAS model, the spikes in the seismicity
237 rate for the Magnitude-Dependent simulations are much reduced in amplitude and are
238 extended in time. This is because the rate increases are much less localized in time when
239 compared to an Omori decay – when a large event occurs, the seismicity rate is elevated,
240 but fairly constant over the ensuing time period.

241 The relative lack of changes in the seismicity rate for the Magnitude-Dependent simula-
242 tions illustrates an important issue in earthquake statistics. Because the time period over
243 which the effects of an earthquake are evident is reasonably long, the true background
244 seismicity rate (22 events/year) is rarely observed in Fig. 2(b). There is always some
245 earthquake that is influencing the seismicity rate, causing the nominal background rate

246 (i.e. the rate inferred by an observer through examination) to not reflect the true back-
 247 ground rate. While this problem exists in all earthquake records (observed and simulated),
 248 its effect is more pronounced here because the true background rate is never observed.
 249 This type of clustering is therefore more difficult to detect, despite the true background
 250 rate being identical to the other datasets, because the tests cannot distinguish the nominal
 251 and true background rates.

2.3. Event Clusters

252 We generate a third type of nonrandom dataset by inserting two randomly placed, non-
 253 overlapping ten-year long clusters with an elevated seismicity rate. In these simulations,
 254 the seismicity rate is elevated from λ_0 to a level $\lambda_0 + \lambda_{clust}$ for a specified number of events
 255 N_{clust} following two random events, chosen such that the two clusters are non-overlapping
 256 and that both clusters fit within the total length of the simulation. The number of events
 257 with an elevated rate N_{clust} increases as the clustering strength increases. The number of
 258 events N_{clust} and the rate change λ_{clust} are chosen so that the duration of each cluster is
 259 on average 10 years, and the dataset contains on average 10000 events over its 100 year
 260 duration. Given a desired background rate λ_0 one can determine N_{clust} and λ_{clust} by:

$$261 N_{clust} = \frac{10000 - \lambda_0 \times 100 \text{ years}}{2} + \lambda_0 \times 10 \text{ years} \quad (5)$$

$$262 \lambda_{clust} = \frac{10000 - \lambda_0 \times 100 \text{ years}}{2 \times 10 \text{ years}}. \quad (6)$$

263 The particular values of λ_{clust} given our values of λ_0 can be found in Table 2. As with
 264 the Magnitude-Dependent rate simulations, the simulations are created in practice by
 265 drawing 15000 recurrence times from an appropriately scaled exponential distribution
 266 given the seismicity rate for a particular event, and then restricting our analysis to 100

267 years. As with the Magnitude-Dependent datasets, we produce far more than 10000
268 events to be certain that all of our simulations are 100 years in duration. Magnitudes are
269 assigned independent of the seismicity rate, following the same distribution as the other
270 simulations.

271 This model is inspired by the “clusters” of magnitude 8.4 and larger events observed
272 in the years 1950-1962 and 2004-2012 [*Bufe and Perkins*, 2005; *Ben-Naim et al.*, 2013].
273 While there is no clear evidence that seismicity at lower levels was higher during these
274 time periods, we use this aspect of the global earthquake catalog as a guide in creating
275 our synthetic datasets to better understand how strong such clusters would need to be
276 in order to be detectable by statistical tests. We also note that the study by *Dimer de*
277 *Oliveira* [2012] used temporal clusters of larger ($M \geq 8$) events in his analysis of simulated
278 clustered data, giving us a point of comparison between our study and other work on this
279 topic. An example of the seismicity rate as a function of time for an Event Clusters
280 simulation is shown in Fig. 2(c).

2.4. Stochastic Rate Simulations

281 The final type of dataset that we consider is one in which the rate varies stochastically
282 in time. As with the Magnitude-Dependent and Event Clusters simulations, the rate λ_i
283 changes from event to event. For the Stochastic Rate datasets, the event rate following
284 event i is $\lambda_i = \lambda_0 + \lambda_{\sigma,i}$, where $\lambda_{\sigma,i}$ is drawn from a one-sided normal distribution. The
285 distribution peaks at zero (i.e. the equivalent two-sided normal distribution has zero
286 mean) and has a standard deviation σ . To set the level of clustering, we vary the value
287 of σ . As with the other simulations, σ is chosen given λ_0 so that the 100 year portion of

288 the synthetic dataset that we use in the analysis contains on average 10000 events. The
289 specific values for each background rate are presented in Table 2.

290 This set of simulations was selected to provide a scenario where the rate does not vary
291 in a predictable fashion with time, and consequently the nonrandom character of the
292 simulation is very difficult to detect, as we will see through our analysis. An example of
293 a Stochastic Rate simulation with the strongest level of clustering is shown in Fig. 2(d).

3. Quantifying Clustering Strength

294 To calibrate the parameter values used in generating our synthetic data and to quantify
295 the clustering in the resulting event sequences, we use two different measures of the
296 clustering strength. The first is the branching ratio, defined as the fraction of events
297 that are in excess of the background seismicity level [*Sornette and Sornette, 1999*]. This
298 quantity was initially developed for ETAS models, as it represents the fraction of events
299 that are aftershocks, but it applies generally to all of the models considered here through
300 our knowledge of the background seismicity rate. For each synthetic dataset, we generate
301 20 different random ensembles with branching ratios varying from 0.02 to 0.78 with a
302 spacing of 0.04. This means that the background rate ranges from 98 events/year to 22
303 events/year, with a difference of 4 events/year between each successive clustering strength.
304 The parameter controlling clustering strength for each simulation type is then changed
305 by trial and error so that the simulations show an overall event rate of 100 events/year.
306 Parameter values for the different models are provided in Table 2.

307 While the branching ratio is only dependent on the background event rate, the actual
308 details of how the event rate varies with time changes across the different synthetic datasets.
309 Thus, we would like to have an independent measure based on fluctuations in the rate

310 that quantifies the strength of clustering. For each type of synthetic dataset, we calculate
311 the coefficient of variation (COV, the standard deviation normalized by the mean) of the
312 rate, assuming the rate is piecewise constant in time. Then we find the mean over all
313 10000 realizations (only 1000 realizations are used in the rate calculations for the ETAS
314 simulations, due to a greater computational cost associated with determining the rate
315 in the ETAS model). The mean rate COV is shown as a function of branching ratio
316 for all four types of synthetic datasets in Fig. 4. For each type, the mean rate COV is
317 linearly dependent on the branching ratio, with a different constant of proportionality for
318 each type. Because the details of the time variation in the rate changes for each type of
319 clustering, we do not expect the mean rate COV to be the same for each type of synthetic
320 data. In particular, we find that the ETAS models have the most strongly varying rate,
321 while the magnitude-dependent rate fluctuates the least in time.

322 In the remainder of this study, we use the branching ratio as our primary indicator of
323 clustering strength due to its simple origin, ease of calculation, and uniformity across the
324 different types of synthetic datasets. However, our results show that mean rate COV is
325 a better indicator of how easily the nonrandom character of a particular dataset can be
326 detected, suggesting that rate variations give a better characterization of the nonrandom
327 behavior for a particular type of synthetic data.

4. Statistical Tests

328 We select eight statistical tests used in the literature to test the synthetic datasets for
329 deviations from random event occurrence. These tests can be divided into two classes:
330 parameter-free tests, and parameter-based tests. We choose four versions of each, and
331 apply each of them to all 800,000 realizations of our synthetic datasets. The parameter-

332 free tests include a test based on the variance of the recurrence times [*Ben-Naim et al.*,
333 2013], a Kolmogorov-Smirnov (KS) test that compares recurrence times to an exponential
334 distribution [*Michael*, 2011], a KS test that compares event occurrence times to a uniform
335 distribution [*Shearer and Stark*, 2012], and an autocorrelation test [*Michael*, 2011; *Parsons*
336 *and Geist*, 2012]. We also choose four tests that require choosing parameter values,
337 including a multinomial chi-squared test [*Gardner and Knopoff*, 1974; *Shearer and Stark*,
338 2012], a Poisson dispersion test [*Shearer and Stark*, 2012; *Daub et al.*, 2012], an alternative
339 chi-squared test [*Brown and Zhao*, 2002; *Luen and Stark*, 2012], and a test that looks for
340 a seismicity rate increase following large events [*Michael*, 2011]. Due to the large number
341 of simulated datasets, we do not vary the parameter values.

342 Most of these tests have been applied to the global earthquake dataset, as well as other
343 types of earthquake catalogs. We summarize the results of application of these tests to
344 the PAGER/PDE dataset through the end of 2014 in Table 1. The tests are applied at
345 minimum magnitudes of 7, 7.5, and 8 both with and without removal of aftershocks using
346 the procedure described in *Daub et al.* [2012]. In general, the catalog does not show a
347 significant deviation from random event occurrence once aftershocks are removed, other
348 than an apparent long-term variation in the seismicity rate at $M = 7$ that has been
349 attributed to differences in event magnitude estimation over time [*Daub et al.*, 2012].
350 The p -values here are in some cases different from values reported in the literature. This
351 is due principally to differences in the declustering methods, and to a lesser extent to
352 differences in the Monte Carlo methods used to estimate the p -values – in particular, for
353 computational reasons we condition on the observed rate in this study, rather than the

354 observed number of events. When these differences are accounted for, our p -values are in
 355 agreement with previously reported values.

356 For all of the results that follow, we will use the criteria that a deviation at the 1%
 357 level constitutes a nonrandom result, meaning that the dataset in question has a value
 358 of the test statistic that is larger than 99% of random realizations. This is smaller than
 359 the frequently used value of 5% often used in hypothesis testing, as well as the 2.3%
 360 value associated with a test statistic that is more than two standard deviations from the
 361 mean for a normal distribution, another common cutoff. While this lower cutoff is chosen
 362 for conservative purposes, we have performed our analysis using different cutoff values,
 363 both higher and lower. We find that the relative frequency of detection between different
 364 tests, clustering strengths, and magnitude levels to be independent of the choice of cutoff
 365 value. The particular value of the detection power for a specific test, clustering level, and
 366 magnitude cutoff will change if a different significance value is selected, but the relative
 367 values and overall trends remain unchanged.

4.1. Variance Test

368 The variance test compares the normalized variance of the recurrence times to the
 369 normalized variance expected if the event times are uncorrelated [*Ben-Naim et al.*, 2013].
 370 Given N events, we compute a sequence of $N - 1$ recurrence (or interevent) times t_r and
 371 compute the normalized variance V

$$372 \quad V = \frac{\langle t_r^2 \rangle - \langle t_r \rangle^2}{\langle t_r \rangle^2} \quad (7)$$

373 where $\langle \cdot \rangle$ indicates the average. The normalized variance is expected to be nearly unity
 374 for a random sequence of events, though for datasets with small numbers of events, the

375 distribution peaks at a smaller value [Ben-Naim et al., 2013]. A dataset is flagged as non-
376 random if the normalized variance is significantly larger than its expected value (i.e. the
377 variance test is one-sided), which is determined by Monte Carlo simulation using 10000
378 random realizations. The random realizations have a background rate of 100 events/year
379 and last for 100 years, and the magnitude distribution is drawn from the same distribution
380 as the simulated datasets. Note that due to the high computational cost of simulating all
381 possible numbers of events, here we condition on the background rate and duration, rather
382 than on the number of events as was done by Ben-Naim et al. [2013]. The variance test
383 compares the entire distribution to the expected distribution for random event occurrence
384 (i.e. all recurrence times are considered in calculating V), weighing the long recurrence
385 times more heavily. If a dataset contains an excess of long recurrence times, the test flags
386 the sequence as nonrandom.

4.2. Kolmogorov-Smirnov Tests

387 The KS methods compare the cumulative distribution function (CDF) determined from
388 the data with the distribution expected for random event occurrence. The KS test com-
389 putes a test statistic based on the largest absolute deviation between the two CDFs, and
390 then assigns a p -value based on the test statistic and the number of observations in the
391 data. For the KS exponential test [Michael, 2011], we look for a deviation between the
392 expected and observed distribution of recurrence times, which is sensitive to short-term
393 clustering in the data but not to long-term changes in the rate. Because the exponential
394 distribution depends on the rate, which must be estimated from the data, we apply an
395 appropriate correction based on Monte Carlo simulations [Lilliefors, 1969]. While both
396 the variance test and the KS exponential test are based on the distribution of recurrence

397 times, the individual events are combined in a different fashion in each test, and thus the
398 tests do not give identical results.

399 The KS uniform test [*Shearer and Stark, 2012*] differs from the KS exponential test in
400 that it uses the occurrence times, rather than the recurrence times. This difference is
401 crucial, because the ordering of the events is important for the KS uniform test. Consider
402 a hypothetical dataset with a set of recurrence times that are exponentially distributed,
403 but ordered from shortest to longest. The KS exponential test would find the sequence
404 consistent with random event occurrence, while the KS uniform test would not. This
405 shows that the KS uniform test is more sensitive to long-term variations in the rate.

4.3. Autocorrelation Test

406 The final parameter-free test looks at the first lag of the autocorrelation of the recurrence
407 times [*Michael, 2011; Parsons and Geist, 2012*]. If there are correlations between the
408 recurrence times of consecutive events (as is expected in an aftershock sequence), this test
409 will identify the data as nonrandom. It is possible to include additional lags beyond the
410 first in the test [*Parsons and Geist, 2012*], though for simplicity we only use the first lag
411 in this study.

4.4. Multinomial Chi-Squared Test

412 This test looks at the detailed distributions of the number of events occurring in a series
413 of K time windows [*Gardner and Knopoff, 1974; Shearer and Stark, 2012*]. From the data,
414 we use the observed number of events per time window λ , and then determine the values
415 of K^- and K^+ . K^- is the smallest integer such that the expected number of windows
416 with no more than K^- events is at least five, and K^+ is the largest integer such that the

417 expected number of windows with at least K^+ events is at least five. Mathematically, this
 418 can be expressed as

$$419 \quad K^- = \min \left\{ k : K \exp(-\lambda) \sum_{i=0}^k \frac{\lambda^i}{i!} \geq 5 \right\}, \quad (8)$$

$$420 \quad K^+ = \max \left\{ k : K \left(1 - \exp(-\lambda) \sum_{i=0}^{k-1} \frac{\lambda^i}{i!} \right) \geq 5 \right\}.$$

421 From the values of K^- and K^+ , the test calculates a test statistic based on the expected
 422 number of events per time bin E_k :

$$423 \quad E_k = \begin{cases} K \exp(-\lambda) \sum_{i=0}^{K^-} \lambda^i / i!, & k = K^-, \\ K \exp(-\lambda) \lambda^k / k!, & K^- < k < K^+, \\ K \left(1 - \exp(-\lambda) \sum_{i=0}^{K^+ - 1} \lambda^i / i! \right), & k = K^+. \end{cases} \quad (10)$$

424 From the data, the test determines the observed number of time bins with fewer than K^-
 425 events X_{K^-} , the number of time bins with k events X_k , and the number of time bins with
 426 at least K^+ events X_{K^+} . These values are used to calculate the Multinomial Chi-Squared
 427 test statistic χ_M^2 :

$$428 \quad \chi_M^2 = \sum_{k=K^-}^{K^+} \frac{(X_k - E_k)^2}{E_k}. \quad (11)$$

429 We choose 1 year for the time window, so the test examines if the distribution of the
 430 number of events in the time windows follow the expected distribution for random event
 431 occurrence. The original study by *Gardner and Knopoff* [1974] compared the distribution
 432 to a chi-squared distribution, though here we follow *Shearer and Stark* [2012] and estimate
 433 the p -values through Monte Carlo simulation using 10000 random realizations. As with
 434 the variance test described above, for computational reasons we condition on the observed
 435 rate and duration, rather than the observed number of events, which are in agreement as
 436 long as the number of events is large.

4.5. Poisson Dispersion Test

437 The Poisson Dispersion test (also referred to as a Conditional Chi-Squared test) divides
 438 the dataset into discrete time bins and examines if the variance of the number of events per
 439 time bin is consistent with random event occurrence. Specifically, given K intervals, each
 440 containing N_k events, and the average number of events per window $\langle N_k \rangle$, the Poisson
 441 Dispersion test calculates a test statistic

$$442 \quad \chi_c^2 = \sum_{k=1}^K \frac{(N_k - \langle N_k \rangle)^2}{\langle N_k \rangle}. \quad (12)$$

443 We use 1 year for the time windows, which was used by both *Shearer and Stark* [2012] and
 444 *Daub et al.* [2012] on the global catalog. *Daub et al.* [2012] also looked at variable time
 445 windows, and found that the results were not strongly dependent on the choice of time
 446 window. We estimate the p -values by random simulation following the procedure of [*Daub*
 447 *et al.*, 2012] and condition on the observed rate and duration, rather than the observed
 448 number of events as was done by *Shearer and Stark* [2012]. Both versions of the test when
 449 applied to the global earthquake catalog gave similar results. The test examines if there
 450 are an anomalous number of windows with a large number of events, which is indicative
 451 of clustering in the sequence of events being tested.

4.6. Brown and Zhao Chi-Squared Test

452 This test is similar to the Poisson Dispersion test, but uses a slightly different test
 453 statistic that converges to a Chi-Squared distribution with sufficient data [*Brown and*
 454 *Zhao*, 2002]. The test calculates the test statistic using the number of observed events in
 455 a series of time windows and the observed seismicity rate. The test divides the dataset
 456 into K windows, with N_k events in the k th window. The Brown and Zhao test statistic

457 χ_{BZ}^2 is then calculated from $Y_k = \sqrt{N_k + 3/8}$ and $\langle Y_k \rangle$ as

$$458 \quad \chi_{BZ}^2 = 4 \sum_{k=1}^K (Y_k - \langle Y_k \rangle)^2. \quad (13)$$

459 We choose 1 year for the time window, and estimate the p -value by Monte Carlo simulation
 460 using 10000 random realizations. As with the other tests, we condition on the rate and
 461 duration when using simulation to quantify the variability the the Brown and Zhao test
 462 statistic. This test is designed to be similar to the Poisson Dispersion test, though we
 463 find in practice that the Poisson Dispersion test tends to be more reliable at detecting the
 464 nonrandomness in our simulated data.

4.7. Big Event Triggering Test

465 The final test looks for a rate increase within a set of time windows following events
 466 above a chosen cutoff magnitude M_{big} [*Michael*, 2011]. We use a time window of 1 year
 467 and $M_{big} = 8.5$ in the test, though *Michael* [2011] examined the various parameter values
 468 more systematically. The test determines the number of events N_w that occur within
 469 the time windows following events with $M \geq M_{big}$, and then uses a binomial test to
 470 determine the probability of N_w events occurring in those windows if the true rate is
 471 always the observed seismicity rate (less the events with $M \geq M_{big}$, which are used to
 472 define the windows). If the number of small events that occur in the windows following
 473 the large events is greater than expected, then the test flags the data as nonrandom. Note
 474 that unlike the other tests, the Big Event Triggering test makes an assumption about
 475 the specific mechanism of triggering (i.e. large events will tend to trigger small events).
 476 This may make the test more sensitive for detecting the nonrandom behavior that occurs
 477 in our simulations where large events trigger small events (ETAS, Magnitude-dependent)

478 while reducing its sensitivity to clustering that does not following the specific triggering
479 model.

5. Test Results

480 We test our synthetic datasets using the eight statistical tests described in Section 4
481 at the 1% level. We threshold each sequence of events in magnitude at several levels, as
482 is frequently done with the global earthquake record [*Michael*, 2011; *Shearer and Stark*,
483 2012; *Daub et al.*, 2012; *Ben-Naim et al.*, 2013], with magnitude levels ranging from $M \geq 6$
484 (the entire dataset) to $M \geq 8$ with increments of 0.1 magnitude units. This allows us
485 determine the likelihood of detection for datasets with between 10000 and 100 events on
486 average, depending on the magnitude level chosen.

487 Each type of synthetic dataset has 20 versions with different clustering strengths, and
488 each of the 20 versions is simulated 10000 times. For each set of magnitude level and
489 clustering strength, we determine the detection power, in other words the probability
490 that the test identifies the synthetic dataset as nonrandom. If the detection power is
491 nearly unity, then the test reliably identifies the nonrandom aspects of the simulated data,
492 while if the detection power is close to zero, then the test cannot distinguish between the
493 simulated events and a random sequence of events.

494 We summarize the results of the statistical tests in Figs. 5-8. Each subplot of the figures
495 shows the magnitude level on the vertical scale and branching ratio on the horizontal scale,
496 with the color scale indicating detection power. Plot (a) shows the detection power for the
497 variance test, (b) illustrates the detection power for the KS exponential test, (c) depicts
498 the detection power for the KS uniform test, (d) shows the results for the autocorrelation
499 test, (e) illustrates the detection power of the Multinomial Chi-Squared test, (f) depicts

500 the results for the Poisson Dispersion test, (g) shows the Brown and Zhao Chi-Squared
501 test, and (h) illustrates the detection power of the Big Event Triggering test. The results
502 for the ETAS simulations are shown in Fig. 5, the results for the Magnitude-Dependent
503 simulations are illustrated in Fig. 6, the results for the Event Clusters can be found in
504 Fig. 7, and the results for the Stochastic Rate simulations are shown in Fig. 8.

505 As expected, all tests show improved detection power as the strength of the cluster-
506 ing increases, and performance decreases as the magnitude threshold increases (and the
507 number of events decreases). We also find the expected result that some tests are better
508 suited for detecting certain types of nonrandom behavior than others. For example, the
509 KS uniform, autocorrelation, and multinomial do not perform as well as the other tests
510 on the ETAS simulations. The KS uniform test is best at detecting long term variations
511 in the seismicity rate, while the aftershock sequences that characterize ETAS simulations
512 are localized in time and do not introduce a change in the seismicity rate over long time
513 periods. Conversely, the KS uniform test performs better than the KS exponential test
514 for the Magnitude-Dependent simulations. This is because the Magnitude-Dependent rate
515 change extends over longer periods of time, which is more difficult to detect when using
516 the interevent time distribution.

517 The nonrandomness of the Event Clusters is easily detected by all of the tests, though
518 the Multinomial Chi-squared test does not perform well at high magnitudes, as it looks
519 at the detailed distribution of events per time bin, details that cannot be discerned for
520 small numbers of events. All tests are successful because this type of clustering intro-
521 duces both long-term variations in the rate, as well as many extra short recurrence times
522 during the clusters. Interestingly, the Big Event Triggering test is still able to flag the

523 Event Clusters datasets as nonrandom, despite making the incorrect assumption that the
524 seismicity rate increases following large events. The test can identify these datasets as
525 nonrandom because it preferentially samples the elevated rate of smaller events during the
526 two clusters, due to the fact that the events with M_{big} are also more likely to occur during
527 the clusters. However, it does require more data than many of the other tests, so it does
528 suffer to some degree for its incorrect assumption. On the other hand, the Stochastic Rate
529 datasets are the most difficult to identify as nonrandom, as the rate changes are variable,
530 being neither localized in time (like the ETAS simulations) nor extended in time (like the
531 Magnitude-Dependent and Event Clusters datasets). Only the tests that use the distribu-
532 tion of the recurrence times (the Variance and KS Exponential tests) have much success
533 with this type of clustering. The other tests fail due to a lack of a long term rate change
534 (KS Uniform), no correlation between successive recurrence times (Autocorrelation), lack
535 of short term clustering over time scales of ~ 1 year (Multinomial, Poisson Dispersion,
536 and Brown and Zhao tests), and no relationship between large events and seismicity rate
537 increases (Big Event Triggering).

5.1. Sensitivity to Magnitude-Frequency Distribution

538 Because of the differences in the magnitude-frequency distribution between the PAGER
539 and GEM catalogs, we perform additional tests using ETAS simulations that are generated
540 using varying magnitude-frequency distributions. In particular, we run two additional
541 sets of simulations maintaining $M_{min} = 6$ but changing the b -value to 0.8 or 1.2, and one
542 additional simulation with $b = 1$ but $M_{min} = 5$. In the case of the variable b -values, the
543 relative number of small versus large events changes, so we can examine if the magnitude
544 distribution within an aftershock sequences affects the ability of the various statistical

545 tests to identify the nonrandom character of that aftershock sequence. Similarly, the
546 $M_{min} = 5$ simulations test whether or not we bias our results by not simulating the
547 smaller events that exist in natural seismicity sequences below the detection threshold in
548 a catalog. The parameter values for these three additional ETAS models are shown in
549 Table 3, and are chosen to maintain the overall rate of 100 events/year in our simulated
550 datasets. Note that by changing the b -value, the number of events at higher magnitudes
551 are different (the number decreases for $b = 1.2$ relative to the $b = 1$ simulations, and the
552 number increases for $b = 0.8$ relative to the $b = 1$ simulations), while the overall number
553 of events at all magnitude levels remains the same for the $M_{min} = 5$ simulations.

554 The results of this analysis are shown in Figs. 9-11. Each figure shows the same set
555 of plots as described for Fig. 5, but with a different magnitude-frequency distribution:
556 Fig. 9 shows the results for $b = 0.8$ and $M_{min} = 6$, Fig. 10 illustrates the detection
557 power for the various statistical tests for $b = 1.2$ and $M_{min} = 6$, and Fig. 11 shows the
558 detection capabilities of the tests for the ETAS simulations with $b = 1$ and $M_{min} = 5$.
559 The detection power changes somewhat for the varying b values, but the differences can
560 be attributed entirely to changes in the number of events. When $b = 0.8$, there are more
561 large magnitude events, and thus we find the statistical tests have greater detection power
562 when compared to the $b = 1$ results, and when $b = 1.2$, the reverse is true. For simulations
563 with $M_{min} = 5$, the results are essentially the same as for $M_{min} = 6$. These additional
564 simulations suggest that our results should be broadly applicable to other earthquake
565 catalogs with different magnitude-frequency distributions, as it is primarily the number
566 of events that controls what clustering can be detected.

6. Analysis of p -values

567 The detection power analysis of the various statistical tests applied to the synthetic
568 datasets provides information on the clustering strengths that are detectable in the global
569 earthquake dataset. However, this analysis requires a choice of the p -value that is con-
570 sidered to be statistically significant, an issue on which there is often some debate. To
571 avoid the problem of choosing a significance level, we instead analyze the p -values that do
572 not give a statistically significant result to see how much information the p -values contain
573 about the clustering level. If p -values turn out to be predictive of clustering strength,
574 then instead of picking a significance level, one could simply test the data and then infer a
575 likely range of clustering strengths based on the p -value. This approach can be thought of
576 as applying a Bayesian framework to the problem, with the clustered datasets producing
577 the prior distribution. Through our analysis, we can use our simulation results to find the
578 posterior distribution of the clustering strength of the earthquake record conditioned on
579 the p -value observed by applying the statistical test to the global earthquake record.

580 To implement this framework, we analyze the p -values that do not constitute a positive
581 test (i.e. $p > 0.01$) for several different magnitude levels, tests, and clustering types. We
582 take all synthetic realizations for which $p > 0.01$ and then bin the p -values specific to each
583 catalog type in five different bins which are logarithmically spaced, centered at $p = 0.015$,
584 0.038 , 0.095 , 0.24 , and 0.602 . Each bin contains simulations that are potentially from any
585 clustering level. We can then determine the cumulative distribution (CDF) as a function
586 of clustering strength for each set of p -values. The CDF for each bin shows the likelihood
587 that the p -values in that particular bin are drawn from any of the clustering strengths
588 above that particular branching ratio, and thus can be used to directly infer upper limits

589 on the branching ratio at particular confidence levels. If the CDF falls rapidly from unity
590 to zero, then there is a strong correlation between p -value and clustering level and p -values
591 have predictive power. On the other hand, if the decrease in the CDF from unity to zero
592 is more gradual, then there is a weaker correlation between p -value and clustering level
593 and little can be learned from the p -value.

594 Example CDF functions calculated using this analysis are shown in Fig. 12. The plots
595 illustrate the CDF as a function of branching ratio for the five bins of p -values for the
596 ETAS simulations. The p -value of each bin decreases from the bottom curve to the top
597 curve. Each plot shows the results for one particular statistical test, and the top set of
598 plots show the CDFs for $M = 7$ and the bottom set of plots are for $M = 8$.

599 The CDF plots illustrate several aspects of the test results. For the tests that perform
600 well on the ETAS simulations, the CDF falls off fairly quickly as the branching ratio
601 increases for all p -values. This indicates that for many statistical tests, the p -value is well
602 correlated with the level of clustering even if the p -value is not low enough to constitute
603 a statistically significant result. At $M \geq 8$ the CDFs decrease more gradually. This is
604 due to the reduced number of events at this magnitude level, and so these results are less
605 useful for constraining the clustering level.

606 These trends are further illustrated for datasets where the clustering is more difficult to
607 detect. Figure 13 shows the same CDF plots as in Fig. 12 for the Magnitude-Dependent
608 synthetic datasets. For the Magnitude-Dependent simulations, the Big Event Triggering
609 test performs best, though a number of other tests also perform well, and this is reflected
610 in the drop-off of the CDF curves. For the Magnitude-Dependent simulations, the tests
611 exhibiting reduced detection power have CDF curves that fall off gradually, and show

612 little difference across the p -value bins. This shows that for tests where detection of the
613 nonrandom character of the synthetic data is less likely, there is also little correlation
614 between p -values and the level of clustering in the dataset. For the Stochastic Rate simu-
615 lations (not shown), nearly all of the CDF curves provide little information on clustering
616 strength.

617 Figure 14 shows the same sets of CDFs for the Event Clusters. Here, all of the tests
618 perform well, and produce CDFs that fall off rapidly with branching ratio. There are
619 differences in precisely how fast the curves decrease across different tests, but in general
620 the p -values of each test provide information on the clustering strength. At $M = 7$, the
621 steep drop in the CDFs with branching ratio indicate that we can use our results to place
622 quantitative upper bounds on the branching ratio. At $M = 8$, the CDFs decrease more
623 slowly, but can still provide upper bounds on the clustering strength in the global catalog.

624 The general trends found in our results help us understand what we can learn from the
625 p -values in the global earthquake catalog. Because all of these tests have been previously
626 applied to the earthquake data since 1900, where the level or type of clustering is unknown,
627 we can use these results to bound the clustering level that is likely to be present in the
628 data, which we discuss in the following section.

7. Discussion

629 The simulations used in this study show that statistical tests cannot always detect that
630 an earthquake dataset is nonrandom. There are two effects responsible for this. First,
631 tests need a certain amount of data in order to distinguish nonrandom event occurrence
632 from the expected fluctuations in a random process. While this result is unsurprising, our
633 results confirm that this is the case for the earthquake catalog at high magnitude levels,

634 and we have quantified how this depends on the strength of the clustering. Second, we
635 find that not all tests perform equally well on a given type of clustering, and that if the
636 wrong test is performed, the test exhibits slower improvement as more data is added than
637 a test that is better suited to identifying the given type of clustering. This suggests that
638 performing the correct test is just as important as having enough data, as it makes a given
639 amount of data more useful.

640 This question has been previously studied by *Dimer de Oliveira* [2012], though only
641 for higher magnitude levels and one type of clustering (similar to our Event Clusters
642 simulations). The clustered datasets of the Dimer de Oliveira study varied the background
643 rate from 0.1 events/year to 0.5 events/year, with a tenfold increase in the seismicity rate
644 during a number of 15 year clusters during a 110 year event sequence, with the number
645 of clusters ranging from 1-5. This best corresponds to our Event Clusters simulations
646 at $M = 8$ at branching ratios of 0.62 or 0.66, with a background rate of 0.38 or 0.34
647 events/year, respectively, and two 10 year clusters where the event rate is 3.4 or 3.6
648 events/year, respectively. Dimer de Oliveira found that for 4 events/decade and two
649 clusters, the detection power at the 5% level was about 40%. In our case, we found that
650 we could almost always detect that these simulated datasets are clustered at the 1% level,
651 though at slightly lower clustering levels our detection power diminishes quickly. These
652 differences in detection power may be due to differences in the exact number of events,
653 or the fact that we generate our data using the event rate at $M = 6$ rather than $M = 8$,
654 resulting in a different amount of variability in the number of large magnitude events
655 across different realizations.

656 Our results have direct relevance for studies that compare the global earthquake record
657 with a process that is random in time. The results of such tests tend to find that the
658 global catalog does not deviate from a random process at p -values that are considered to
659 be significant, and thus our results can bound the range of clustering that could be in the
660 data, yet not be detectable.

661 We can use the CDFs constructed for different p -values (i.e. Figs. 12 and 14) to provide
662 quantitative upper bounds on the branching ratio given the results of tests applied to the
663 earthquake catalog. For instance, *Michael* [2011] reports $p = 0.17$ for the earthquake
664 record for $M \geq 7$ with aftershocks removed. That value falls within our $p = 0.240$ bin
665 (though at the low end of the bin). Based on Fig. 12, this suggests that the percentage of
666 events that are aftershocks is below 20% based on the 99th percentile bound, as the CDF
667 drops below 10^{-2} at a branching ratio of 0.2 for this particular test. Other statistical tests
668 give similar upper bounds when applied to the PAGER/PDE catalog, so this appears to
669 be a robust upper limit based on several statistical tests.

670 We can make a similar estimate for Event Clusters simulations using the results in
671 Fig. 14. The best performing test on this particular type of clustering was the Pois-
672 son Dispersion test. *Shearer and Stark* [2012] were more conservative in their removal
673 of aftershocks, leaving only 509 independent events in their catalog for $M \geq 7$, which
674 corresponds to approximately $M \geq 7.3$ in our simulations. These estimates constrain
675 the branching ratio to be below 0.1 at the 99th percentile if magnitude 7 events occur in
676 two large clusters using the Poisson Dispersion test. In terms of seismicity rate during
677 the clusters, a branching ratio of 0.1 at $M = 7$ indicates a $\sim 50\%$ increase in seismicity
678 rate (from 9 events/year to 14 events/year during the clusters). We find that clustering

679 stronger than this is unlikely to give us the observed p -values. Other tests give similar
680 upper bounds on the maximum branching ratio that is consistent with the PAGER/PDE
681 data.

682 At higher magnitude levels above $M \geq 8$, we cannot constrain the branching ratio
683 based on the p -values of statistical tests to the same degree as $M = 7$. While synthetic
684 datasets that are more clustered tend to have smaller p -values, the correlation is not strong
685 enough to rule out higher branching ratios with high confidence. The study of *Michael*
686 [2011] on the $M \geq 8$ earthquake data reported $p = 0.61$ for the KS exponential test.
687 Our results for the ETAS models for that range of p -values suggest that the branching
688 ratio of the global catalog is below 0.6 at the 99th percentile, a much broader range than
689 we infer for the $M \geq 7$ data. Similarly, *Shearer and Stark* [2012] found $p = 0.898$ for
690 the Poisson Dispersion test at $M \geq 8$, which suggests a branching ratio below 0.6 at
691 the 99th percentile for the ETAS simulations, but an upper bound of 0.35 at the 99th
692 percentile for the Event Clusters datasets. The KS Uniform test applied to the Event
693 Clusters simulations provides a similar estimate of these bounds. In terms of seismicity
694 rate increases for $M = 8$, a branching ratio of 0.35 corresponds to an increase from 0.66
695 events/year to 2.3 events/year during the clusters. Clustering strengths higher than this
696 would be detectable given the amount of data in the PAGER/PDE catalog, but lower
697 values produce sequences of events that are not reliably distinguishable from random
698 event occurrence.

699 We note that each of the clustering studies discussed above used a different method
700 to remove aftershocks, while our simulations are based on the assumption that there
701 are no local aftershocks in the synthetic data. Because aftershock removal is somewhat

702 subjective, there is a good chance that some background events are removed in the process
703 (see *Luen and Stark* [2012] for a discussion of some of the issues raised by the aftershock
704 removal process). Our simulations do not consider the spatial distribution of earthquakes,
705 and thus do not exhibit this artifact of declustering. Further studies, for example using
706 ETAS simulations with spatial kernels, are necessary to better quantify the impact of this
707 effect on the results of statistical tests.

708 One caveat of our analysis is that our synthetic datasets are only four of any number of
709 possible ways that events can cluster in time. We have used various observations and mod-
710 els for clustering to guide our development of our simulations, but this is only a partial,
711 limited set of considerations. Earthquake data may also contain a combination of multiple
712 types of clustering, which may affect the ability of statistical tests to identify nonrandom
713 behavior, and statistical tests cannot reveal the underlying mechanism of clustering, only
714 the likelihood that it is present. Because the rules for aftershock production are based
715 on more robust observations than those for the other types of clustering, the ETAS sim-
716 ulations may be more relevant to seismic hazard estimates. This is especially true since
717 the effect of aftershocks is more localized in time than the other types of clustering, and
718 aftershock forecasts are one of the few types of short-term seismic hazard estimates that
719 can be made with confidence [*Jordan et al.*, 2011]. Thus, such analysis can be extended
720 using the upper bounds on global aftershock production to estimate the effect such global
721 aftershocks would have on seismic hazard. Such an estimate would only serve as an upper
722 bound, but could inform us whether such effects might play a role in hazard estimates.

723 While our focus here is on the global earthquake catalog, our methods can also be ap-
724 plied to regional catalogs at lower magnitude levels. Previous work has mostly focused

725 on earthquake catalogs in Southern California [*Gardner and Knopoff*, 1974; *Luen and*
726 *Stark*, 2012], Japan [*Zhuang et al.*, 2004], and Taiwan [*Wang et al.*, 2014], but as seismic
727 networks continue to grow worldwide, more regional catalogs will become available. For
728 example, our CDF-based methods could be used to provide an assessment of the goodness
729 of fit between ETAS models and regional catalogs and help constrain regional aftershock
730 parameters [*Ogata*, 1992, 1998]. Knowledge of how to analyze such catalogs for seismicity
731 patterns and develop models for earthquake interaction requires quantification of cluster-
732 ing that is consistent with the catalog. The tools outlined here provide a means of doing
733 that, and can help researchers bound the strength of earthquake interactions in a variety
734 of tectonic settings.

735 **Acknowledgments.** We thank Andrew Michael and an anonymous reviewer for con-
736 structive reviews. This research was supported by DOE grant DE-AC52-06NA25396 and
737 institutional (LDRD) funding at Los Alamos. Figures were generated using the Python
738 plotting library Matplotlib [*Hunter*, 2007].

References

- 739 Aki, K. (1965), Maximum likelihood estimate of b in the formula $\log N = a - bM$ and its
740 confidence limits, *Bull. Earthq. Res. Inst., Univ. Tokyo*, *43*, 237-239.
- 741 Allen, T.I., K. Marano, P. S. Earle, and D. J. Wald (2009), PAGER-CAT: A composite
742 earthquake catalog for calibrating global fatality models, *Seism. Res. Lett.*, *80*, 50-56.
- 743 Ben-Naim, E., E. G. Daub, and P. A. Johnson (2013), Recurrence Statistics of Great
744 Earthquakes, *Geophys. Res. Lett.*, *40*, 3021-3025, doi:10.1002/grl.50605.

- 745 Brown, L. D. and Zhao, L. H. (2002), A test for the Poisson distribution, *Sankhyā: Indian*
746 *J. Stat.*, *64*, 611-625.
- 747 Bufe, C. G., and D. M. Perkins (2005), Evidence for a Global Seismic-Moment Release
748 Sequence, *Bull. Seismol. Soc. Am.*, *95*, 833-843.
- 749 Daub, E. G., E. Ben-Naim, R. A. Guyer, and P. A. Johnson (2012), Are megaquakes
750 clustered?, *Geophys. Res. Lett.*, *39*, L06308, doi:10.1029/2012GL051465.
- 751 Dimer de Oliveira, F. (2012), Can we trust earthquake cluster detection tests?, *Geophys.*
752 *Res. Lett.*, *39*, L17305, doi:10.1029/2012GL052130.
- 753 Engdahl, E. R., and A. Villaseñor (2002), Global seismicity: 1900-1999, *International*
754 *Handbook of Earthquake and Engineering Seismology, Volume 81A*, ISBN:0-12-440652-
755 1, 665-690.
- 756 Felzer, K. R., R. E. Abercrombie, and G. Ekstrom (2004), A common origin for after-
757 shocks, foreshocks, and multiplets, *Bull. Seismol. Soc. Am.*, *94*, 88-98.
- 758 Felzer, K. R., and E. E. Brodsky (2006), Decay of aftershock density with distance indi-
759 cates triggering by dynamic stress, *Nature*, *441*, 735-738.
- 760 Freed, A. M. (2005), Earthquake triggering by static dynamic, and
761 postseismic stress transfer, *Ann. Rev. Earth Planet. Sci.* *33*, 335-367,
762 doi:10.1146/annurev.earth.33.092203.122505.
- 763 Gardner, J. K., Knopoff, L. (1974), Is the sequence of earthquakes in Southern California,
764 with aftershocks removed, Poissonian? *Bull. Seismol. Soc. Am.*, *64*, 1363-1367.
- 765 Gomberg, J., P. Bodin, K. Larson, and H. Dragert (2004), Earthquake nucleation by
766 transient deformations caused by the M = 7.9 Denali, Alaska, earthquake, *Nature*, *427*,
767 621-624.

- 768 Gomberg, J., P. Bodin, and P. A. Reasenberg (2003), Observing earthquakes triggered in
769 the near field by dynamic deformations, *Bull. Seismol. Soc. Am.*, *93*, 118-138.
- 770 Guo, Z., and Y. Ogata (1997), Statistical relations between the parameters of af-
771 tershocks in time, space, and magnitude, *J. Geophys. Res.*, *102*(B2), 28572873,
772 doi:10.1029/96JB02946.
- 773 Gutenberg, B., and C. F. Richter (1954), *Seismicity of the Earth and Associated Phenom-*
774 *ena*, 2nd ed., Princeton University Press, Princeton.
- 775 Helmstetter, A., and D. Sornette (2002), Subcritical and supercritical regimes in
776 epidemic models of earthquake aftershocks, *J. Geophys. Res.*, *107*(B10), 2237,
777 doi:10.1029/2001JB001580.
- 778 Helmstetter, A., and D. Sornette (2003), Båth's law derived from the Gutenberg-
779 Richter law and from aftershock properties, *Geophys. Res. Lett.*, *30*, 2069,
780 doi:10.1029/2003GL018186, 20.
- 781 Helmstetter, A., Y. Y. Kagan, and D. D. Jackson (2005), Importance of small earth-
782 quakes for stress transfers and earthquake triggering, *J. Geophys. Res.*, *110*, B05S08,
783 doi:10.1029/2004JB003286.
- 784 Hill, D. P., et al. (1993), Remote seismicity triggered by the M7.5 Landers, California
785 earthquake of June 28, 1992, *Science*, *260*, 1617-1623.
- 786 Hunter, J. D. (2007), Matplotlib: A 2D graphics environment, *Comput. Sci. Eng.*, *9*(3),90-
787 95.
- 788 Jordan, T., Y. Chen, P. Gasparini, R. Madariaga, I. Main, W. Marzocchi, G. Papadopou-
789 los, G. Sobolev, K. Yamaoka, and J. Zschau (2011), Operational Earthquake Fore-
790 casting: State of Knowledge and Guidelines for Utilization, *Ann. Geophys.*, *54*(4),

791 doi:10.4401/ag-5350.

792 Lay, T., and T. C. Wallace (1995), *Modern Global Seismology*, Academic, San Diego,
793 Calif.

794 Lilliefors, H. W. (1969), On the Kolmogorov-Smirnov test for the exponential distribution
795 with mean unknown, *J. Am. Stat. Assoc.*, *64*, 387-389, doi:10.2307/2283748.

796 Luen, B. and Stark, P. B. (2012), Poisson tests of declustered catalogues. *Geophys. J. Int.*,
797 *189*, 691-700, doi: 10.1111/j.1365-246X.2012.05400.x.

798 Michael, A. J. (2011), Random Variability Explains Apparent Global Clustering of Large
799 Earthquakes, *Geophys. Res. Lett.*, *38*, L21301, doi:10.1029/2011GL049443.

800 Michael, A. J. (2014), How Complete is the ISC-GEM Global Earthquake Catalog? *Bull.*
801 *Seismol. Soc. Am.*, *104*(4), 1829-1837, doi:10.1785/0120130227

802 Ogata, Y. (1992), Detection of precursory relative quiescence before great earth-
803 quakes through a statistical model, *J. Geophys. Res.*, *97*(B13), 19845-19871,
804 doi:10.1029/92JB00708.

805 Ogata, Y. (1998), Space-time point-process models for earthquake occurrences, *Ann. Inst.*
806 *Stat. Math.*, *50*(2), 379-402.

807 Omori, F. (1895), On the aftershocks of earthquakes, *J. Coll. Sci., Imp. Univ. Tokyo*, *7*,
808 111-200.

809 Parsons, T., and E. L. Geist (2012), Were Global $M \geq 8.3$ Earthquake Time Intervals
810 Random between 1900 and 2011?, *Bull. Seismol. Soc. Am.*, *102*(4), 1583-1592, doi:
811 10.1785/0120110282.

812 Parsons, T., and E. L. Geist (2014), The 2010-2014.3 global earthquake rate increase,
813 *Geophys. Res. Lett.*, *41*, 4479-4485, doi:10.1002/2014GL060513.

- 814 Parsons, T., and A. A. Velasco (2011), Absence of remotely triggered large earthquakes
815 beyond the mainshock region, *Nat. Geosci.*, *4*, 312-316.
- 816 Pollitz, F. F., R. S. Stein, V. Sevilgen, and R. Bürgmann (2012), The 11 April
817 2012 East Indian Ocean earthquake triggered large aftershocks worldwide, *Nature*,
818 doi:10.1038/nature11504.
- 819 Pollitz, F. F., R. Bürgmann, R. S. Stein, and V. Sevilgen (2014), The Profound Reach
820 of the 11 April 2012 M 8.6 Indian Ocean Earthquake: Short-Term Global Triggering
821 Followed by a Longer-Term Global Shadow, *Bull. Seismol. Soc. Am.*, *104*(2), 972-984,
822 doi: 10.1785/0120130078.
- 823 Rhoades, D. A., and F. F. Evison (2004), Long-range Earthquake Forecasting with Ev-
824 ery Earthquake a Precursor According to Scale, *Pure Appl. Geophys.*, *161*(1), 47-72,
825 doi:10.1007/s00024-003-2434-9.
- 826 Shearer, P. M., and P. B. Stark, (2012), The global risk of big earthquakes has not recently
827 increased, *Proc. Nat. Acad. Sci.*, *109*(3), 717-721.
- 828 A. Sornette, D. Sornette (1999), Renormalization of earthquake aftershocks, *Geophys.*
829 *Res. Lett.*, *26*, 1981-1984, doi:10.1029/1999GL900394.
- 830 Storchak, D. A., D. Di Giacomo, I. Bondär, E. R. Engdahl, J. Harris, W. H. K. Lee,
831 A. Villaseñor and P. Bormann (2013). Public Release of the ISC-GEM Global In-
832 strumental Earthquake Catalogue (1900-2009), *Seism. Res. Lett.*, *84*(5), 810-815,
833 doi:10.1785/0220130034.
- 834 Utsu, T., Y. Ogata, and R. S. Matsu'ura (1995), The centenary of the Omori formula for
835 a decay law of aftershock activity, *J. Phys. Earth*, *43*, 1-33.

- 836 van der Elst, N. J., and E. E. Brodsky (2010), Connecting near-field and far-
837 field earthquake triggering to dynamic strain, *J. Geophys. Res.*, *115*, B07311,
838 doi:10.1029/2009JB006681.
- 839 Velasco, A. A., S. Hernandez, T. Parsons, and K. Pankow (2008), Global ubiquity of
840 dynamic earthquake triggering, *Nature Geosci.*, *1*, 375-379, doi:10.1038/ngeo204.
- 841 Wang, J. P., D. Huang, S.-C. Chang, and Y.-M. Wu (2014), New Evidence and Perspec-
842 tive to the Poisson Process and Earthquake Temporal Distribution from 55,000 Events
843 around Taiwan since 1900, *Nat. Hazards Rev.*, *15*(1), 38-47.
- 844 Woessner, J., and S. Wiemer (2005), Assessing the quality of earthquake catalogs: Esti-
845 mating the magnitude of completeness and its uncertainty, *Bull. Seismol. Soc. Am.*, *95*,
846 doi:10.1785/0120400007.
- 847 Zhuang, J., Y. Ogata, and D. Vere-Jones (2004), Analyzing earthquake cluster-
848 ing features by using stochastic reconstruction, *J. Geophys. Res.*, *109*, B05301,
849 doi:10.1029/2003JB002879.

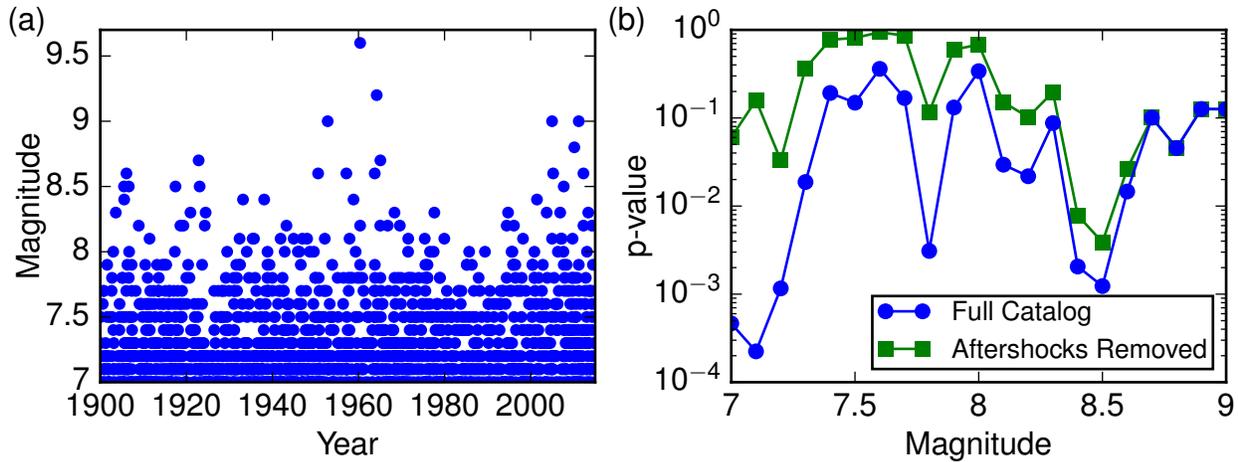


Figure 1. (a) Global earthquake catalog from 1900-2014 [Allen et al., 2009], supplemented by the United States Geological Survey Preliminary Determination of Epicenters catalog through the end of 2014. Occurrence of several earthquakes with $M \geq 8.5$ from 1950-1964 and 2004-2012 led to speculation that the largest earthquakes cluster in time. (b) Statistical analysis using the variance of the recurrence time [Ben-Naim et al., 2013], showing calculated p -values as a function of threshold magnitude, both with and without removal of aftershocks. The p -value is determined by Monte Carlo simulation by calculating the fraction of random catalogs that exhibit a normalized variance that exceeds the value calculated for the PAGER catalog. The results show that in general, the earthquake record does not deviate from a process that is random in time. For magnitude levels $M \geq 8.4$ -8.5, the p -values appear to be low, though they are not low enough to conclude that the earthquake record is nonrandom, as the reported p -values are underestimates due to selection of the magnitude levels *post hoc* [Shearer and Stark, 2012].

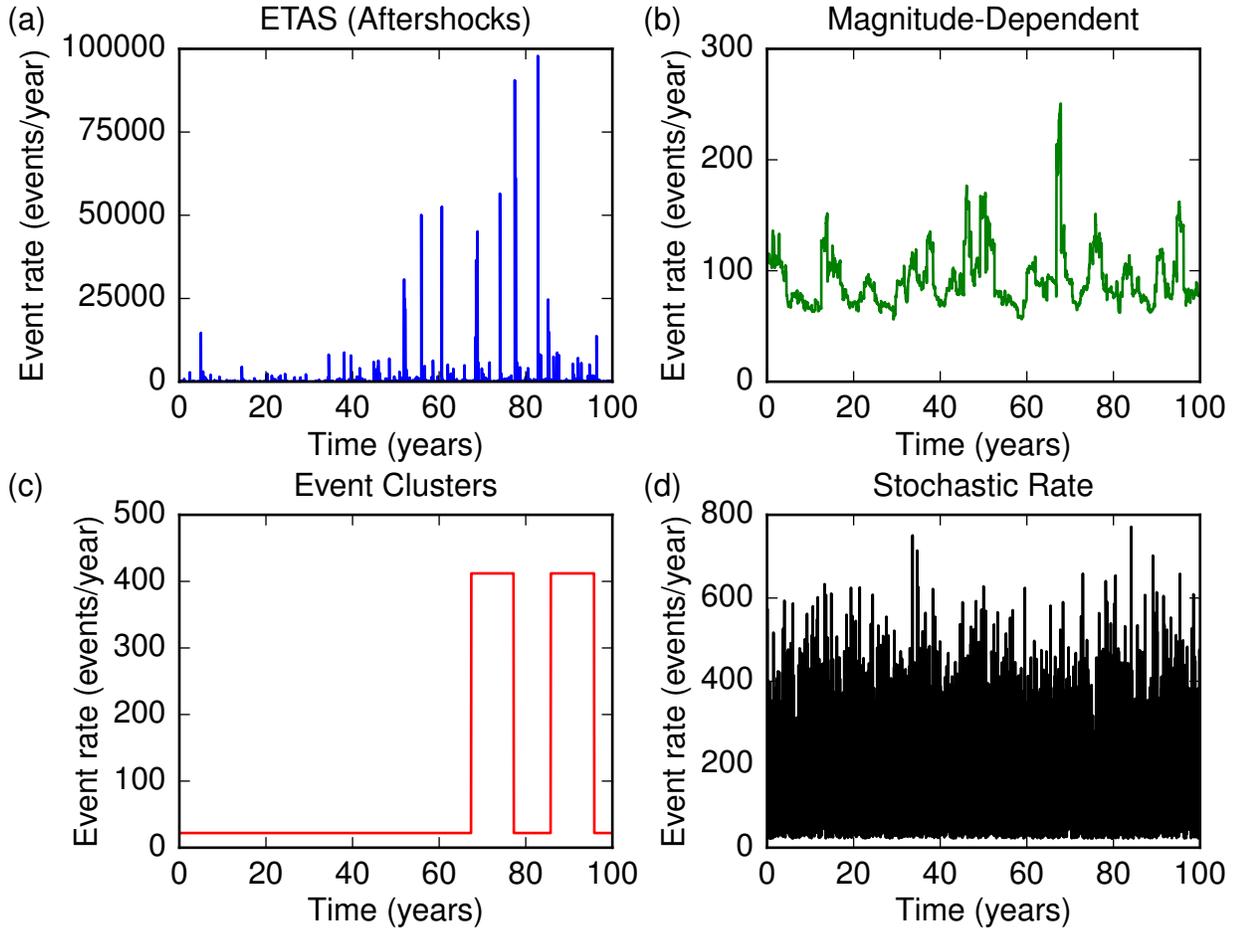


Figure 2. Examples of synthetic clustered datasets. All plots show earthquake event rate as a function of time for a 100 year sequence containing approximately 10,000 events above $M = 6$. The plots shown here illustrate the most strongly clustered example of each simulation type. (a) ETAS model, where additional events are added to the synthetic dataset following empirical aftershock rules. The seismicity rate is sharply peaked following large events, decaying rapidly in time. (b) Magnitude-Dependent rate, where the seismicity rate changes based on the magnitude of the previous 200 events. (c) Event Clusters, where two ten-year clusters of events are placed randomly in time. (d) Stochastic Rate simulations, where the rate varies stochastically in time from event to event, with the variations following a one-sided normal distribution.

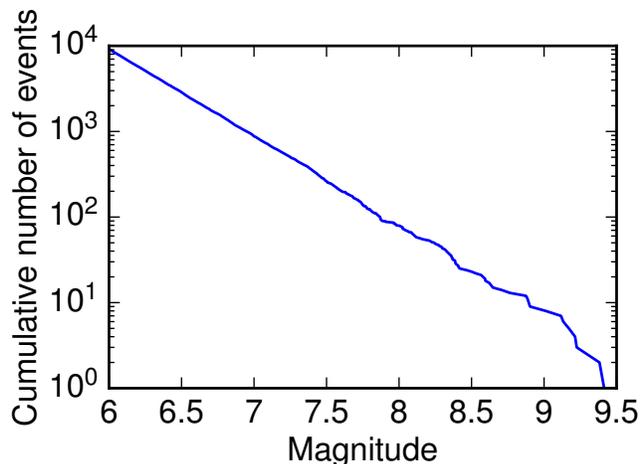


Figure 3. Magnitude-frequency distribution used in the synthetic datasets. Events follow a Gutenberg-Richter distribution with $b = 1$ and a minimum magnitude of $M_{min} = 6$. While this differs from the empirical value of $b = 1.26$ observed for the PAGER/PDE catalog, using $b = 1$ makes determining the number of events at a given magnitude level more straightforward. Alternative magnitude frequency distributions with $b = 0.8$, $b = 1.2$, and $M_{min} = 5$ are also considered for the ETAS simulations.

Table 1. p -values for the statistical tests used in this study applied to the PAGER/PDE catalog through the end of 2014. Tests include: Variance (Var), KS Exponential (KSE), KS Uniform (KSU), Autocorrelation (AC), Multinomial Chi-Squared (MC), Poisson Dispersion (PD), Brown and Zhao Chi-Squared (BZ), and Big Event Triggering, and details on each test are provided in the main text. Tests are applied to the catalog both with and without removal of aftershocks ((AS) denotes removal of aftershocks, using the method described in *Daub et al.* [2012]) at minimum magnitude levels of $M = 7, 7.5$, and 8 .

Catalog	N	Var	KSE	KSU	AC	MC	PD	BZ	Big Event
$M = 7$	1814	6.0×10^{-4}	1.4×10^{-4}	1.6×10^{-5}	0.95	0.0060	3.0×10^{-4}	0.0019	0.0076
$M = 7.5$	462	0.14	0.053	0.11	0.41	0.14	0.17	0.088	0.24
$M = 8$	87	0.34	0.71	0.15	0.29	0.60	0.59	0.51	0.29
$M = 7$ (AS)	1369	0.062	0.40	1.1×10^{-7}	0.58	0.24	0.010	0.024	0.87
$M = 7.5$ (AS)	385	0.81	0.69	0.18	0.21	0.66	0.69	0.60	0.64
$M = 8$ (AS)	77	0.68	0.81	0.31	0.24	0.22	0.91	0.78	0.24

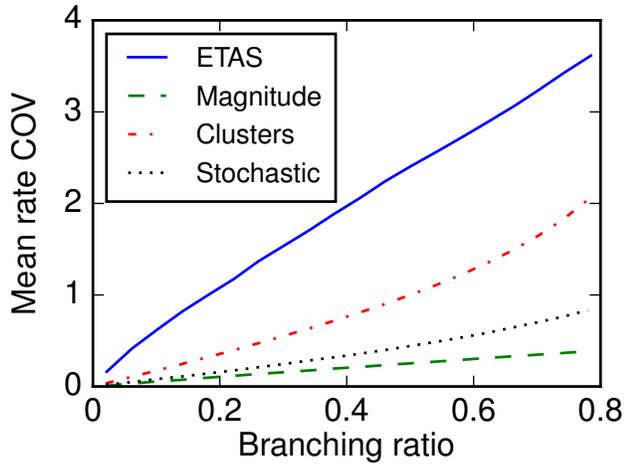


Figure 4. Mean coefficient of variation for the event rate over all synthetic datasets as a function of branching ratio (or fraction of the events that are not background events) for the four synthetic datasets considered in this study. In each case, there is an approximately linear relationship between branching ratio and mean rate COV, and this illustrates that we have a quantitative measure of the clustering strength for each synthetic dataset. Some clustering types exhibit larger rate variations for the same branching ratio when compared to others, and our results show that datasets with stronger rate variations are generally easier to detect as nonrandom.

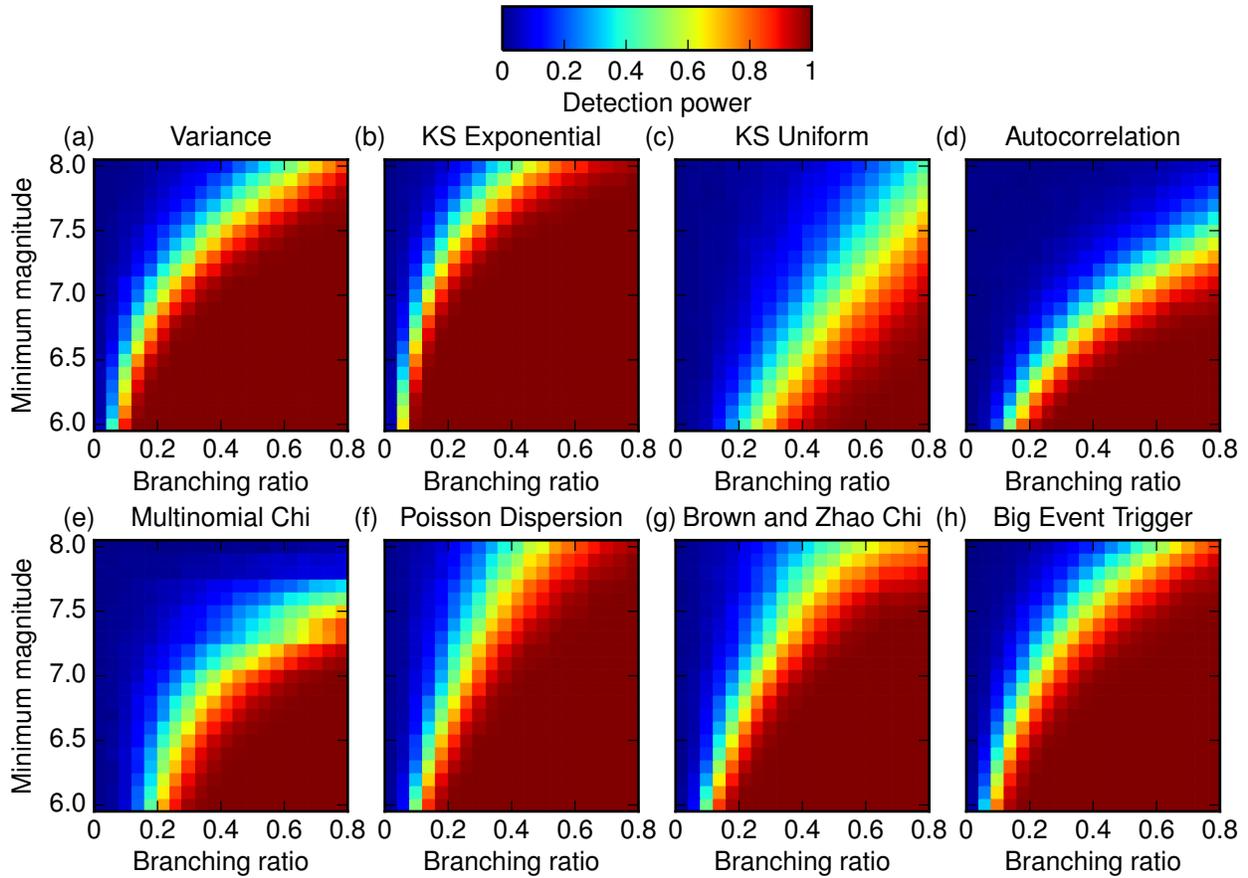


Figure 5. Detection power at the 1% level as a function of branching ratio and minimum magnitude for statistical tests applied to the ETAS models. Vertical axis indicates magnitude level, horizontal axis indicates branching ratio, and color scale indicates detection power, or the probability that the test identifies the event sequence as nonrandom. Results are shown for a set of eight statistical tests: (a) Variance test, (b) KS Exponential test, (c) KS Uniform test (d) Autocorrelation test, (e) Multinomial Chi-Squared test, (f) Poisson Dispersion test, (g) Brown and Zhao Chi-Squared test, and (h) Big Event Triggering test. All of the tests can detect the simulations as nonrandom, though some tests perform slightly better than others. For example, the reduced sensitivity of the KS uniform test is due to the aftershock sequences being very localized in time, while the KS uniform test is more sensitive to long-term variations in seismicity rate. The Autocorrelation and Multinomial Chi-Squared tests do not perform well at high magnitudes, indicating that these tests require more data than the others to detect

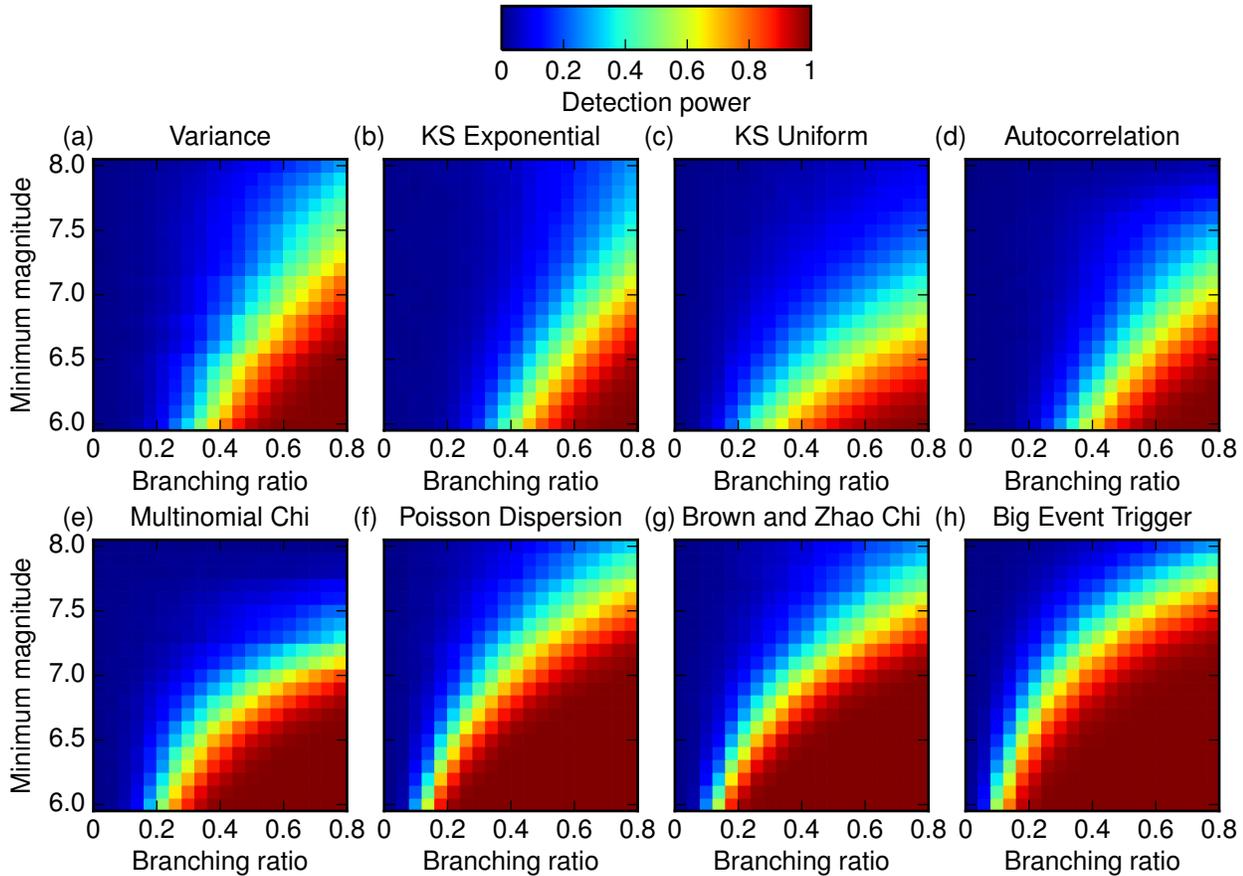


Figure 6. Same as Fig. 5 for the Magnitude-Dependent datasets. Because the mean rate COV is less than in the ETAS simulations, this clustering is more difficult to detect. The Poisson Dispersion and Big Event Triggering tests are the most successful for this type of clustering, as these tests look for variations in rate over longer time periods when compared to tests such as the Variance, KS Exponential, and Autocorrelation, which use interevent times.

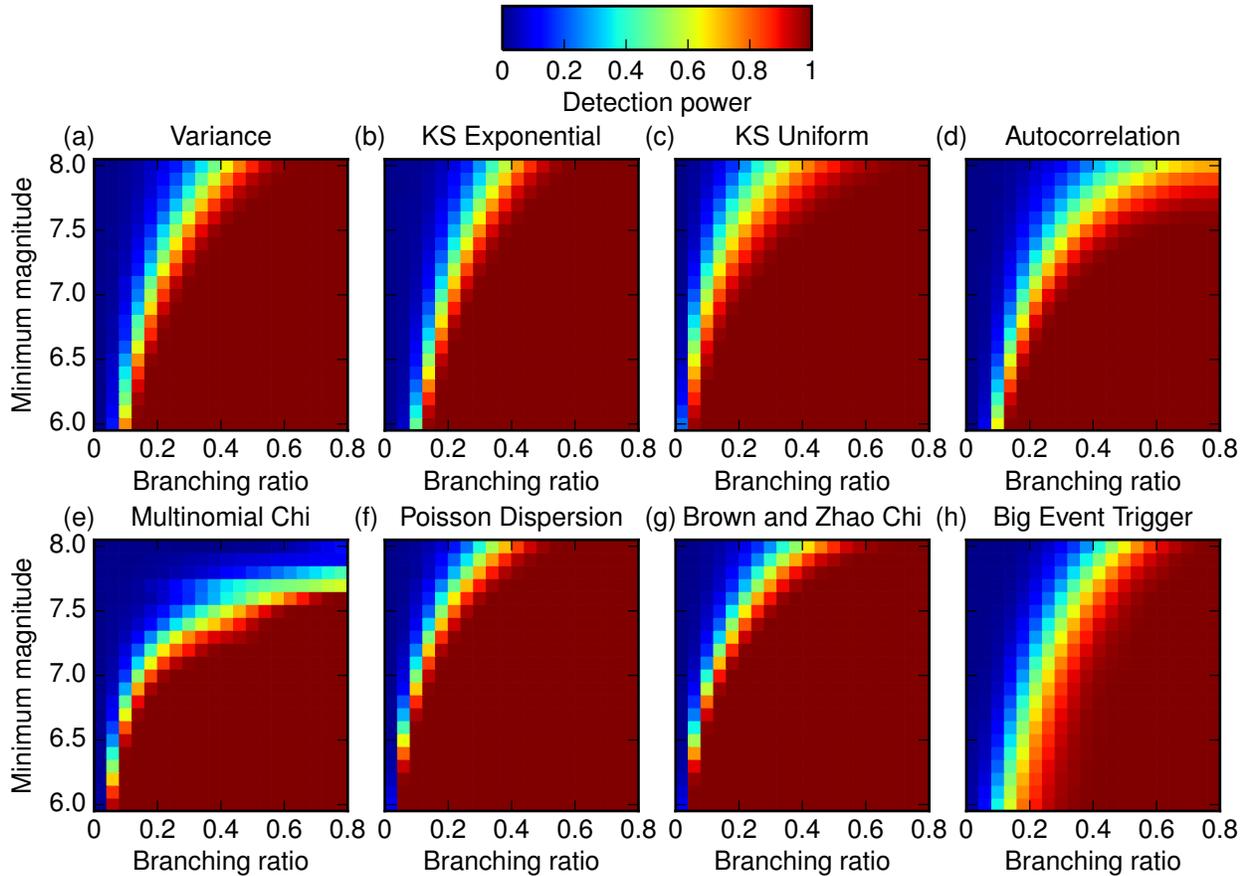


Figure 7. Same as Fig. 5 for the Event Clusters. All tests reliably detect clustering for these synthetic datasets, though the Multinomial Chi-Squared test does not do well at high magnitudes. Additionally, the Big Event Triggering test requires more data than the other tests, due to the fact that the assumption that large events trigger small events does not hold for this particular type of clustering.

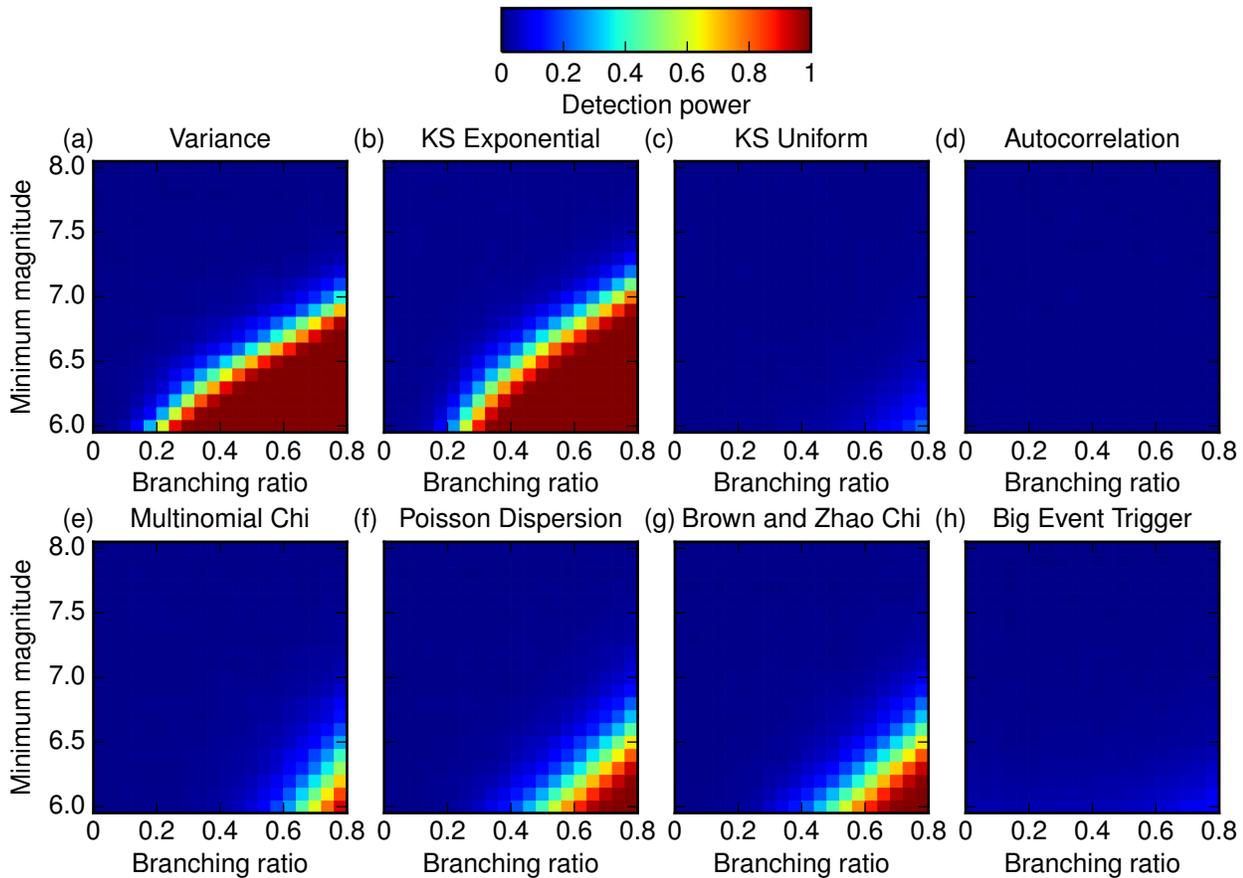


Figure 8. Same as Fig. 5 for the Stochastic Rate simulations. This type of clustering is the hardest to detect, and several tests do not find any nonrandom behavior in the synthetic datasets. The tests that are better able to discern that these sequences are nonrandom use the interevent time distribution (Variance and KS Exponential tests), though the autocorrelation test fails since successive recurrence times vary stochastically and thus are not correlated. The remaining tests perform poorly either due to no long term variation in the rate, or a lack of triggering of small events following large events in the case of the Big Event Triggering test.

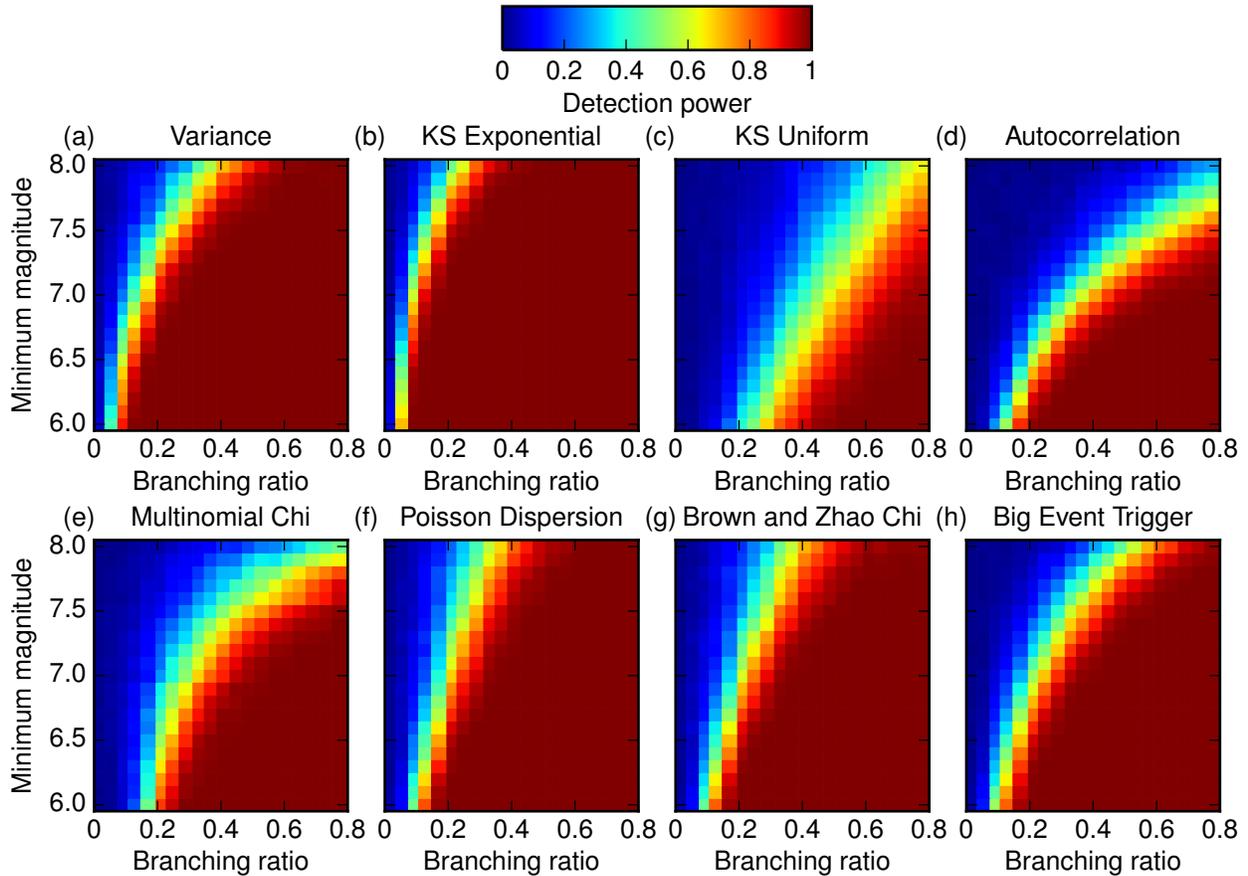


Figure 9. Detection power at the 1% level as a function of branching ratio and minimum magnitude for three statistical tests applied to the ETAS simulations with $b = 0.8$. The tests are better able to distinguish the simulation data as nonrandom for high magnitudes when compared to the $b = 1$ results, due to the increased number of high magnitude events in the $b = 0.8$ simulations.

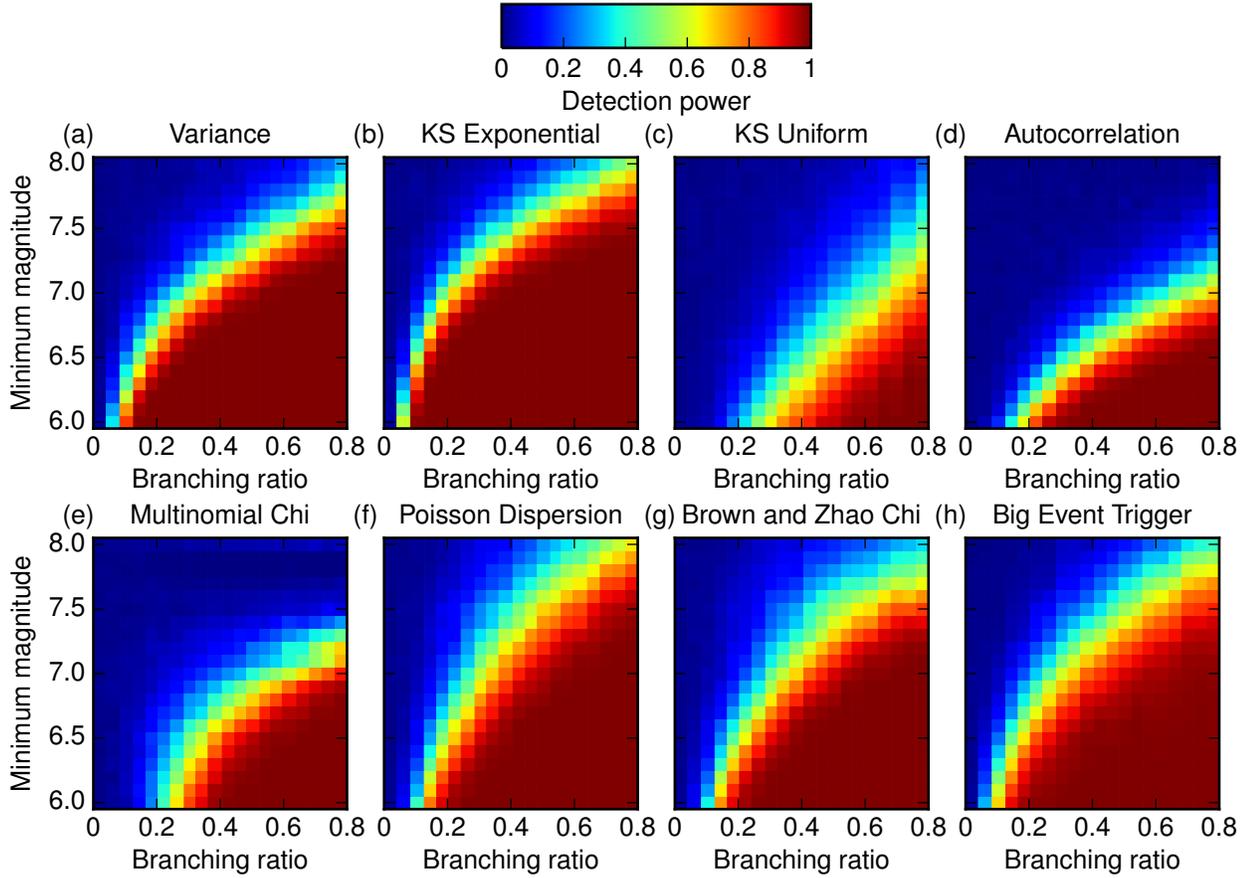


Figure 10. Detection power at the 1% level as a function of branching ratio and minimum magnitude for three statistical tests applied to the ETAS simulations with $b = 1.2$. The tests do not perform as well as the tests on the $b = 1$ results at high magnitudes, as the larger b -value leads to relatively fewer high magnitude events.

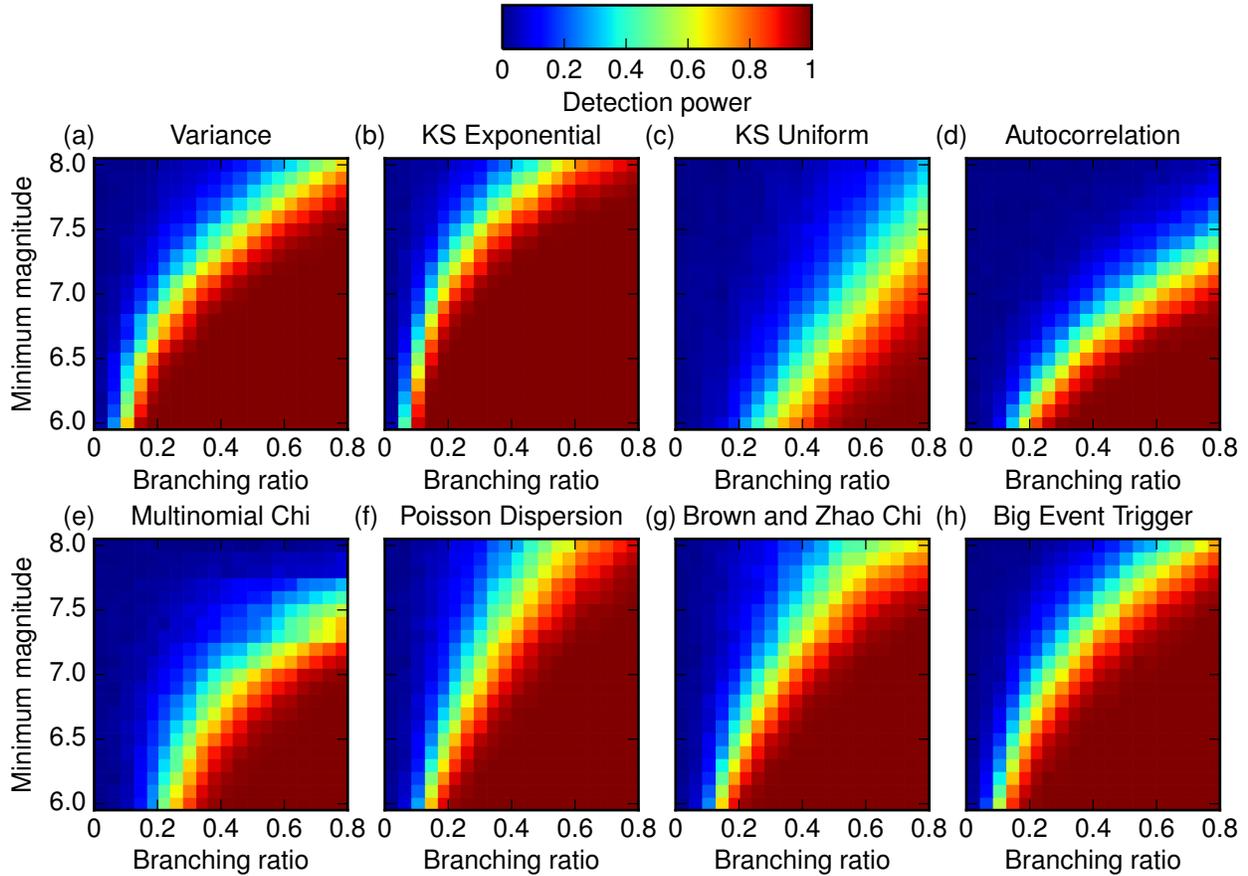


Figure 11. Detection power at the 1% level as a function of branching ratio and minimum magnitude for eight statistical tests applied to the ETAS simulations with $M_{min} = 5$. The results do not show much difference from those in Fig. 5, indicating that the detection power of the statistical tests are not changed by the fact that real earthquake catalogs are missing smaller magnitude events.

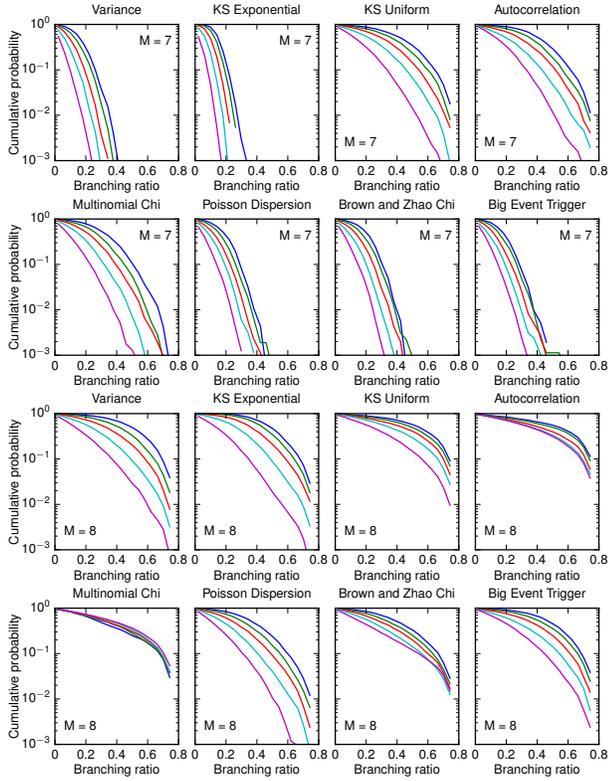


Figure 12. Cumulative distribution functions for ETAS simulation p -values that were not identified as nonrandom by the statistical tests (i.e., $p > 0.01$). The statistical test used in each set of CDF functions is indicated above the plot. In each case, the plots show the CDF as a function of branching ratio for five bins of p -values, centered at the values of 0.015, 0.038, 0.095, 0.240, and 0.602 (p -values increase from top to bottom on each set of curves). The top set of eight plots shows the results for $M = 7$, and the bottom set of eight plots show the results for $M = 8$. Tests that are more likely to detect a synthetic dataset as nonrandom also tend to have p -values that are more predictive of clustering level than tests with lower detection power. The CDF can be easily used to place an upper bound on the level of clustering that is consistent with the p -value observed for the global earthquake record – for instance, the KS exponential test applied to the PAGER/PDE catalog at $M = 7$ gives $p = 0.17$ [Michael, 2011], which is the second curve from the bottom, and we can thus constrain the branching ratio for aftershock-like behavior to be below 0.2 at the 99th percentile. Bounds on the branching ratio at $M = 8$ are higher due to the small number of events in the simulated data.

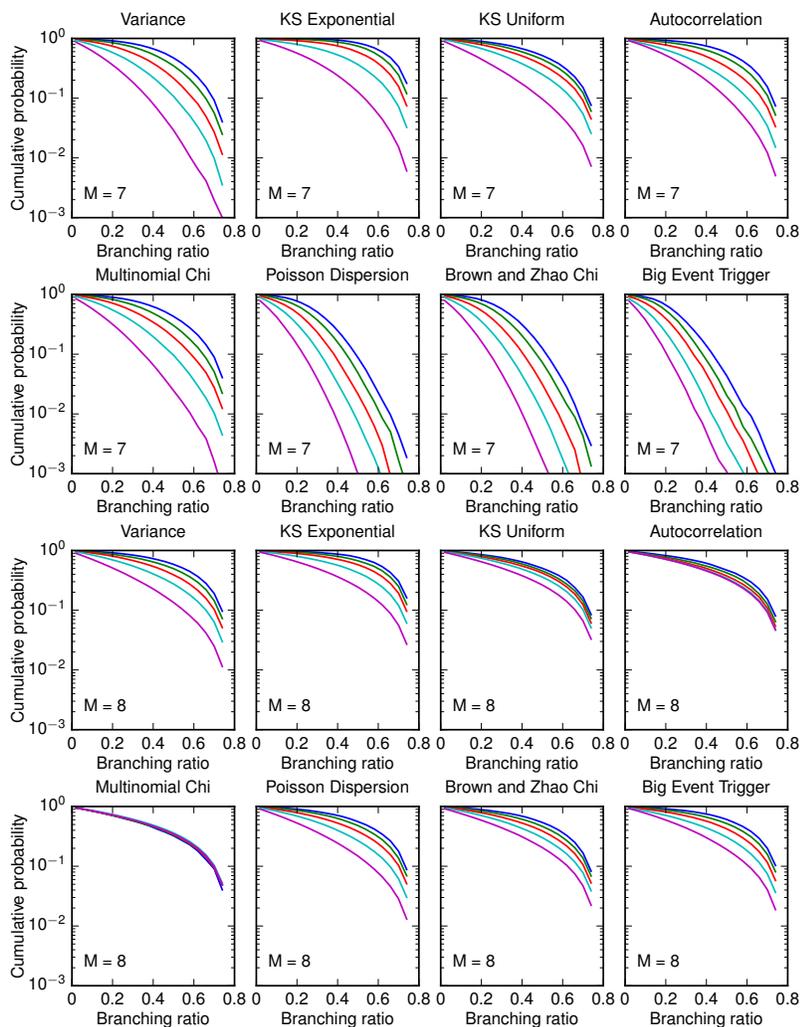


Figure 13. Same as Fig. 12 but for the Magnitude-Dependent simulations. Because this clustering is harder to detect, the CDFs do not fall off as sharply with increasing branching ratio when compared to the ETAS simulations. In several cases, the p -values have very little predictive power of the level of clustering in the data, as the CDF curves fall off slowly with branching ratio, and the curves for different p -values are nearly identical to one another.

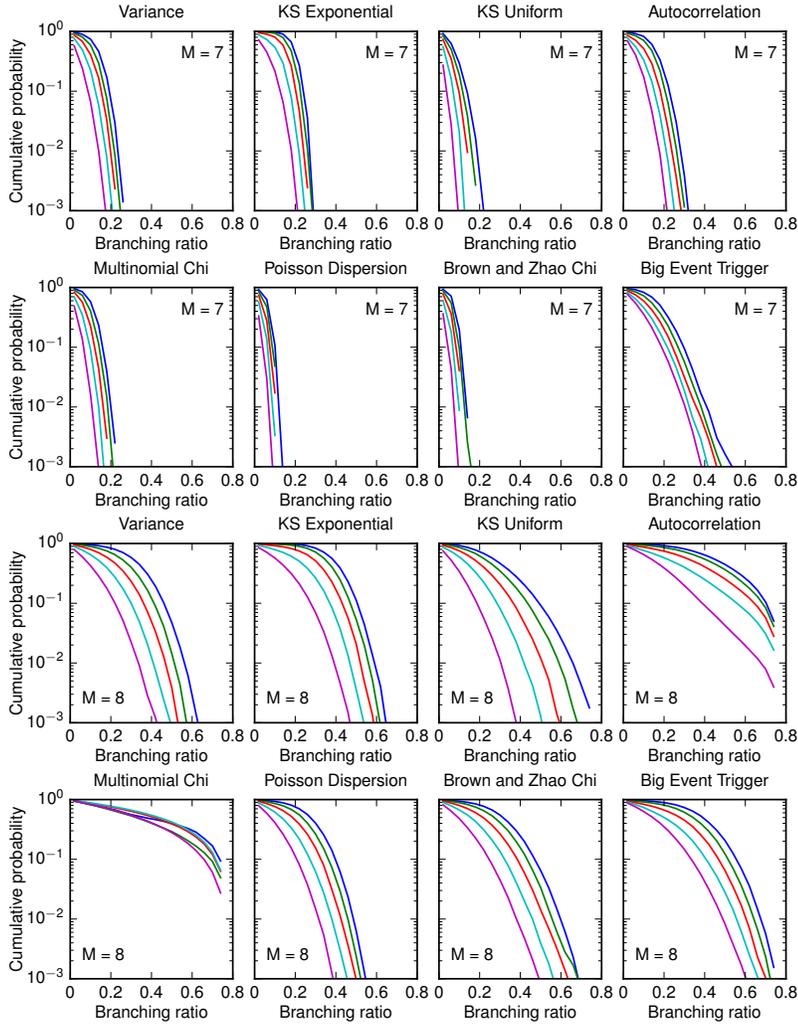


Figure 14. Same as Fig. 12 but for the Event Clusters. With the exception of the Multinomial Chi-squared test at high magnitudes, nearly all tests are able to provide constraints regarding the clustering level, though tests that perform better are better able to constrain an upper bound on the level of clustering in the synthetic dataset. Several of the tests at $M = 7$ suggest that the largest clustering strength consistent with the global earthquake record is 0.1 at the 99th percentile, while at $M = 8$ the upper bound on the branching ratio is below 0.35 at the 99th percentile using the CDFs for the Poisson Dispersion or KS Uniform tests.

Table 2. Parameter values for the synthetic datasets used in the study. Details of all models are described in the main text.

Branching Ratio	Background Rate (events/year)	ETAS C'	Magnitude γ (events/year)	Clusters λ_{clust} (events/year)	Stochastic σ (events/year)
0.02	98	0.0027	0.001	10	3
0.06	94	0.0076	0.004	30	8
0.1	90	0.0125	0.0075	50	14
0.14	86	0.0175	0.01	70	19
0.18	82	0.0225	0.0135	90	25
0.22	78	0.0275	0.0165	110	31
0.26	74	0.0325	0.02	130	38
0.3	70	0.037	0.023	150	45
0.34	66	0.0425	0.026	170	53
0.38	62	0.047	0.03	190	60
0.42	58	0.0525	0.0325	210	68
0.46	54	0.057	0.0365	230	77
0.5	50	0.0623	0.04	250	87
0.54	46	0.06725	0.0435	270	97
0.58	42	0.072	0.047	290	108
0.62	38	0.07725	0.0515	310	120
0.66	34	0.0825	0.0545	330	135
0.7	30	0.08725	0.0585	350	150
0.74	26	0.0925	0.062	370	165
0.78	22	0.0975	0.066	390	185

Table 3. Parameter values for ETAS models varying the magnitude-frequency distribution

Branching Ratio	Background Rate	$b = 0.8$	$b = 1.2$	$b = 1$
		$M_{min} = 6$	$M_{min} = 6$	$M_{min} = 5$
0.02	98	0.0018	0.0035	0.0175
0.06	94	0.004	0.012	0.06
0.1	90	0.0065	0.021	0.1
0.14	86	0.0085	0.03	0.14
0.18	82	0.011	0.038	0.18
0.22	78	0.014	0.046	0.215
0.26	74	0.016	0.054	0.25
0.3	70	0.0185	0.063	0.29
0.34	66	0.02125	0.071	0.33
0.38	62	0.0235	0.079	0.365
0.42	58	0.026	0.088	0.405
0.46	54	0.0285	0.0965	0.44
0.5	50	0.031	0.1045	0.485
0.54	46	0.0335	0.113	0.525
0.58	42	0.03625	0.12	0.56
0.62	38	0.03875	0.128	0.6
0.66	34	0.041	0.138	0.64
0.7	30	0.0435	0.146	0.675
0.74	26	0.04625	0.154	0.715
0.78	22	0.04875	0.163	0.7575