**Natural Hazards
and Earth
System Sciences**

The Gumbel hypothesis test for left censored observations using regional earthquake records as an example

E. M. Thompson[1], **J. B. Hewlett**[2], **L. G. Baise**[1]**, and R. M. Vogel**[1]

[1]Tufts University, Department of Civil and Environmental Engineering, Medford, MA, USA
[2]GZA GeoEnvironmental, Inc. Norwood, MA, USA

Abstract. Annual maximum (AM) time series are incomplete (i.e., censored) when no events are included above the assumed censoring threshold (i.e., magnitude of completeness). We introduce a distrtibutional hypothesis test for left-censored Gumbel observations based on the probability plot correlation coefficient (PPCC). Critical values of the PPCC hypothesis test statistic are computed from Monte-Carlo simulations and are a function of sample size, censoring level, and significance level. When applied to a global catalog of earthquake observations, the left-censored Gumbel PPCC tests are unable to reject the Gumbel hypothesis for 45 of 46 seismic regions. We apply four different field significance tests for combining individual tests into a collective hypothesis test. None of the field significance tests are able to reject the global hypothesis that AM earthquake magnitudes arise from a Gumbel distribution. Because the field significance levels are not conclusive, we also compute the likelihood that these field significance tests are unable to reject the Gumbel model when the samples arise from a more complex distributional alternative. A power study documents that the censored Gumbel PPCC test is unable to reject some important and viable Generalized Extreme Value (GEV) alternatives. Thus, we cannot rule out the possibility that the global AM earthquake time series could arise from a GEV distribution with a finite upper bound, also known as a reverse Weibull distribution. Our power study also indicates that the binomial and uniform field significance tests are substantially more powerful than the more commonly used Bonferonni and false discovery rate multiple comparison procedures.

1 Introduction

A wide variety of hypothesis tests are available for evaluating distributional alternatives, including the Kolmogorov-Smirnov and the Chi-square tests. Research has shown that the probability plot correlation coefficient (PPCC) hypothesis test is more powerful (power is the complement of the false negative rate) than either of these tests for a number of distributional alternatives (Stedinger et al., 1993; Chowdhury et al., 1991; Heo et al., 2008). For example, the PPCC test of normality compared favorably with seven other commonly used hypothesis tests of normality on the basis of empirical power studies performed by Filliben (1975) and Looney and Gulledge (1985). Subsequently, the PPCC test has been extended to many other distributions.

The PPCC test is based on a probability plot, which leads to two advantageous properties: (1) it is easily extended to any probability distribution with a known quantile function (the inverse of the cumulative distribution function), and (2) interpretation of results is intuitive because the test is based on a widely used graphical aid. Thus, the PPCC test is often combined with a graphical display of goodness-of-fit using a probability plot. PPCC test statistics are now available in the form of regression equations for numerous distributions (Heo et al., 2008) and are widely used in the field of statistics, as evidenced by its inclusion in most standard statistical computing packages such as MINITAB, ChemStat, and S-PLUS (Millard and Neerchal, 2001).

A censored observation is an observation in which the exact value is unknown. Although many types of censoring are possible, this paper is concerned with left-censoring, in which all the censored observations are below a detection threshold. Thus, we know that the censored observations are below the threshold, but we do now know their exact values. Censored observations are additionally categorized as either type I censoring, where the measurement threshold is fixed

and the number of censored data points varies, or as type II censoring, where the number of censored data points is fixed and the implicit threshold varies (David and Nagaraja, 2003). The purpose of this study is to extend the Gumbel PPCC hypothesis test developed by Vogel (1986) and Heo et al. (2008) for the case of type I left-censored observations.

Since an annual maximum (AM) series of earthquake magnitudes is an example of an AM series that is usually censored (censoring threshold is the magnitude of completeness), we apply the new hypothesis test to a global earthquake catalog. In environmental statistics, research on censoring has concentrated on comparisons of various estimators of the mean, standard deviation, median, and other statistics of censored data sets. Helsel and Hirsch (2002) and Berthouex and Brown (1994) summarized estimation methods for use with left censored data. Little attention has been given to the development of distributional hypothesis tests for censored observations.

One important application of the censored Gumbel hypothesis test is regional frequency analysis of earthquake magnitudes. Extreme value theory (Gumbel, 1958) continues to be applied to the AM series of earthquake records to assess seismic hazard (e.g., Burton et al., 2004; Öztürk et al., 2008; Shanker et al., 2007). Burton et al. (2004) discuss the difficulties that arise when applying the AM series approach to earthquake catalogs. They refer to censored years as "dummy observations" and suggest that the sampling interval be adjusted to minimize these occurrences. Adjusting the sampling interval is reasonable for the series of earthquake maxima because earthquake records do not exhibit a strong seasonal dependence. However, adjusting the sampling interval does not always completely remove censored values.

The most famous earthquake frequency model is the Gutenberg-Richter (Gutenberg and Richter, 1954) model of earthquake magnitudes. It is equivalent to an exponential probability distribution function (PDF) for the peaks over threshold (POT) series of earthquake magnitudes (Kijko and Graham, 1998; Utsu, 1999). The POT series consists of observations above a threshold, termed the "magnitude of completeness" (m_c). As this term implies, the POT series is assumed to include all occurrences above a specific threshold. Stedinger et al. (1993) and others have shown that a Gutenberg-Richter (i.e., exponential) model of a POT series is equivalent to a Gumbel model of the corresponding AM series of earthquake magnitudes, assuming a Poisson distribution of earthquake arrival times. See Thompson et al. (2007) for further discussion of the relationship between probability distributions of AM and POT series of earthquake magnitudes.

One advantage of the POT series over the AM series is that the number of samples in the POT series is always larger than for the AM series, often by a substantial margin. However, an important advantage of the AM series over the POT series is that the assumption of independence of the observations is more reasonable for the AM series than the POT series. Serial correlation or persistence of the POT series is expected for many natural processes, including earthquakes and such persistence leads to violation of the single most important assumption of such frequency analysis: independence of observations. We are not advocating the use of the AM series over the POT series for estimation of probabilities of exceedance or return periods. However, we hope that researchers and practitioners will consider the wealth of research and discussions on this subject published in the field of hydrology, which has led to use of the AM series for estimation of hydrologic design events (see Stedinger et al., 1993, Sect. 18.6). Additional research, analogous to the type of research performed in the field of hydrology, is needed in the field of earthquake engineering to determine whether an AM or POT type of analysis is preferred in a particular situation.

2 Type I left censored annual maximum observations

When a dataset is missing observations above or below a threshold, such data are said to be "censored". Left censoring arises when some fraction of a dataset are below the detection limit of the available sensing equipment. Left-censored environmental data typically follow type I censoring because the censoring threshold is fixed by the measurement technology. Values for data below a single measurement threshold are generally reported as "less than the detection limit", and data sets containing such points are referred to as type I left singly censored data. This study concentrates on type I left censoring, the type of censoring present in AM series of earthquake magnitudes. We also assume a single magnitude of completeness m_c (single censoring), though extensions are possible to multiple censoring levels (see Helsel and Hirsch, 2002). When working with censored samples, one cannot replace the censored values with either the measurement threshold or some other fixed value because research has shown that such approaches can lead to enormous bias in derived statistics (Helsel and Hirsch, 2002).

For seismic studies on a global scale, one encounters the problem of widely varying methods and equipment used to collect and process seismic data. This issue is present both spatially, where regions with more seismic stations will provide more reliable records, and temporally, where network sensitivity and analytical methods change over time (Rydelek and Sacks, 1989). To avoid such observational bias, care must be taken in determining the appropriate m_c for a particular catalog of earthquakes. Initially, the value of m_c may appear to be a trivial issue in the analysis, but if it is too small it can severely bias estimates of the shape of the distribution. Woessner and Wiemer (2005) showed that small errors in m_c can result in substantial bias in computed seismicity rates. Over time, m_c typically decreases as the extent and quality of seismic instrumentation improves (Wiemer and Wyss,

2000). Wiemer and Wyss (2000) proposed two methods for determining m_c based on the assumption of self-similarity; Woessner and Wiemer (2005) present a modification of the method introduced by Ogata and Katsura (1993) to estimate m_c that uses a Monte Carlo approximation of the bootstrap method to evaluate the precision of estimates of m_c.

As the number of AM observations increases, the precision of the estimates of extreme events also increases. One attempts to minimize m_c, which in turn maximizes the number of complete samples. One also strives to create the longest possible AM series by including older data. As older data are included, however, m_c necessarily increases, which in turn increases the number of events that are censored. The final number of observations included in the analysis is a function of these two competing factors.

3 Data

The AM series used in this study is the same series used by Thompson et al. (2007) except that data before 1977 are now excluded because we have since learned that data may not be complete. Thompson et al. (2007) computed the moment magnitudes from the moment tensor solutions in the global CMT catalog with records from 1976–2005 (Ekström et al., 2005). The data from 1977–2005 were derived from data collected from the Global Digital Seismograph Network and the data before 1977 were derived from the High-Gain Long-Period network in operation at that time (Ekström and Nettles, 1997).

The earthquakes were classified into the 50 Flinn-Engdahl geographic regions (Young et al., 1996), which are loosely based on tectonic setting. Thompson et al. (2007) showed that one could not reject the hypothesis of temporal independence of the AM samples using this regional grouping scheme and a 5% level test based on the lag-one serial correlation coefficient. Kagan (1997) showed that a magnitude of completeness of $m_c = 5.8$ is appropriate for the CMT catalog from 1977–1995. Censoring occurred in 34 of the 50 regions. Thus, the AM series for these regions are type I left censored samples. Indeed, this problem is so severe that every year is censored for $m_c = 5.8$ in Eastern South America (Region 35), Northwestern Europe (Region 36), Northern Eurasia (Region 49), and Antarctica (Region 50). Only 16 regions are complete for the entire extent of the catalog.

4 Probability plots and the probability plot correlation coefficient

The probability plot is a widely used graphical tool for assessing the goodness-of-fit of various probability distributions to data and for illustrating the cumulative distribution of a sample. If a sample arises from a hypothesized distribution, a probability plot of the ordered observations versus their expected values under that hypothesized distribution will be approximately linear. The PPCC is a measure of the linearity of the probability plot and offers a quantitative goodness-of-fit metric to accompany the graphical plot. A probability plot is constructed as follows:

1. Rank the n observations $y_i = y_1, \ldots, y_n$ from a sample in ascending order that yields the ordered observations: $y_{(i)} = y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$.

2. Estimate the non-exceedance probability associated with each ranked observation using a suitable unbiased plotting position p_i, which depends on the sample size n and rank i.

3. Compute estimates of the ordered observations from the hypothesized distribution at p_i using the distribution's inverse cumulative distribution function (also known as its quantile function).

4. Plot the n ordered observations against their estimated values based on the hypothesized distribution and compare with the line with a slope of unity that passes through the origin.

The above approach for constructing a probability plot is termed a quantile-quantile (Q-Q) plot. One could also develop a P-P plot of the percentiles of the observations against their expected percentiles. Other types of probability plots are possible, and their advantages and disadvantages have been discussed for both complete (Wilk and Gnanadesikan, 1968) and censored (Waller and Turnbull, 1992) samples.

4.1 Probability plot for complete Gumbel samples

The type I extreme value distribution is also known as the Gumbel distribution since Gumbel (1941, 1958) first applied it to flood frequency analysis. Gumbel (1958) showed that if samples of a random variable are independent and identically distributed, then the cumulative distribution function (CDF) of the largest observation in a random sample asymptotically approaches the Gumbel distribution

$$F_Y(y) = p = \exp\left[-\exp\left(-\frac{y - \xi}{\beta}\right)\right], \tag{1}$$

where p is the non-exceedance probability associated with y, and β and ξ are model parameters which may be estimated from the first two ordinary product moments using

$$\hat{\beta} = \frac{s_y \sqrt{6}}{\pi} \tag{2}$$

$$\hat{\xi} = \bar{y} - \gamma \hat{\beta}, \tag{3}$$

where $\gamma \approx 0.5772$ is known as Euler's constant, and $\bar{y}$ and s_y are estimates of the mean and standard deviation of y, respectively. Note that the "b-parameter" of the Gutenberg-Richter model is related to Eq. (1) by $\beta = \log_{10}(e)/b$, where
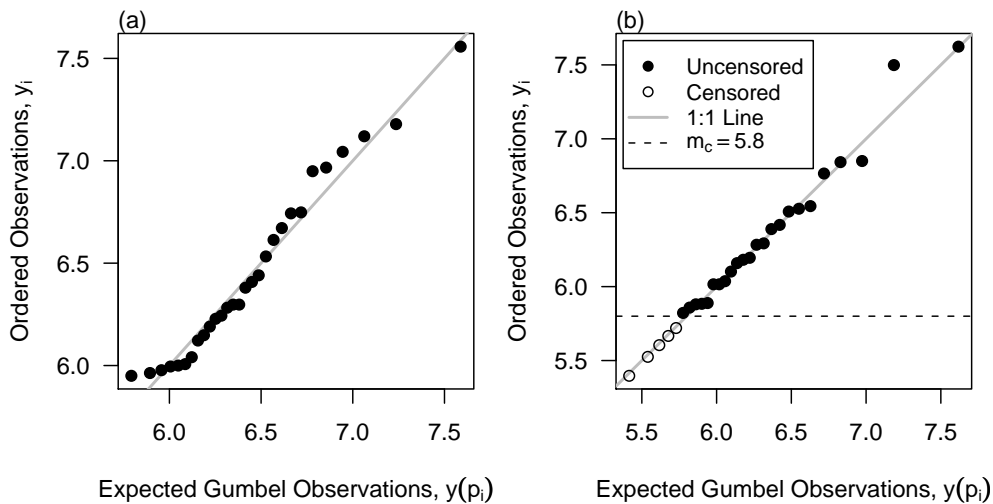
Fig. 1. (a) Gumbel probability plot using a complete series of $n = 29$ random Gumbel variates (Region 10) and **(b)** a censored Gumbel probability plot using the $n = 29$ Gumbel variates, with $n - m = 24$ uncensored observations and $m = 5$ left-censored values (Region 26).

e is the base of the natural logarithm. The inverse of the Gumbel CDF in Eq. (1) is termed its quantile function

$$y(p) = \xi - \beta \ln[-\ln(p)]. \tag{4}$$

A probability plot illustrates the ordered observations $y_{(i)}$ versus estimates of the ordered observations using the quantile function in Eq. (4) with a suitable estimate of p_i. A suitable plotting position for the Gumbel distribution that reproduces the expected values of the ordered observations is the Gringorten (1963) plotting position

$$p_i = \frac{i - 0.44}{n + 0.12}. \tag{5}$$

Figure 1a illustrates a Gumbel probability plot for a complete sample of earthquake magnitudes, using the $n = 29$ observations from the Southern Antilles (Region 10). The uncensored probability plot is constructed by plotting the ordered observations $y_{(i)}$ versus their expectation $y(p_i)$ from the Gumbel quantile function given by Eq. (4) and p_i from Eq. (5). Qualitatively, one expects the observations to fall near the line shown which passes through the origin and has a slope of unity, if the sample originates from a Gumbel distribution.

4.2 Plotting positions for censored Gumbel observations

Probability plots are a convenient tool for handling censored datasets (see Waller and Turnbull, 1992; and Chapter 13 in Helsel and Hirsch, 2002). Here, we describe how to construct a probability plot for censored Gumbel samples. Consider a type I left censored sample with m censored values $y_{(i)}^{c} = y_{(1)}, y_{(2)}, \ldots, y_{(m)}$ followed by the $n - m$ uncensored values

$y_{(i)}^{u} = y_{(m+1)}, \ldots, y_{(n)}$. Adapting the Gringorten (1963) plotting position in Eq. (5) gives

$$p_i^{c} = \frac{i - 0.44}{n + 0.12} \quad \text{for} \quad i = 1, \ldots, m \tag{6}$$

$$p_i^{u} = \frac{i - 0.44}{n + 0.12} \quad \text{for} \quad i = m + 1, \ldots, n \tag{7}$$

where p_i^{c} and p_i^{u} correspond to the plotting positions for censored and uncensored observations, respectively.

4.3 Probability plot regression for censored Gumbel samples

Gupta (1952), Helsel and Gilliom (1986), and Gilliom and Helsel (1986) introduced the idea of probability plot regression (PPR) for constructing probability plots for censored samples. The PPR method for censored observations can also be used for estimating sample statistics because it provides unbiased estimates of the censored observations (Helsel and Hirsch, 2002). This approach has also been referred to as the "regression on order statistics" approach by Shumway et al. (2002). Helsel and Gilliom (1986), Shumway et al. (2002) and others have shown that such an approach is competitive, in terms of mean square error of the estimated statistics, with a variety of alternative estimation methods for censored data, especially for small samples.

The PPR approach fits a regression between the ordered uncensored observations and their expected values. The regression equation can then be used to estimate distributional parameters or the missing (censored) observations. Helsel and Gilliom (1986), Helsel and Hirsch (2002), and Shumway et al. (2002) document that the PPR method can be expected to perform about as well as an MLE for censored samples for estimation of a distribution's parameters and for estimation

of a variety of statistics, particularly for small samples. The PPR method yields unbiased estimates of the missing observations under the assumption that the data arise from a Gumbel distribution. However, this does not reproduce the variance of the original observations, for which a full Bayesian analysis would be more accurate.

Here we extend the PPR method to left censored Gumbel observations. The first step is to construct a probability plot from the uncensored observations, and the second step is to estimate the censored observations using their expected values. The probability plot for the uncensored observations is constructed by plotting the ordered uncensored observations $y_{(i)}^{u}$ versus their expected values using the quantile function $y(p_i^{u})$ in Eq. (4) with p_i^{u} given in Eq. (7).

To estimate the parameters ξ and β in Eq. (4) from the uncensored observations, we note that the quantile function is simply a linear model between the dependent variable $y(p)$ and the transformed variable $\eta = \ln[-\ln(p)]$. Thus, the ordinary least squares (OLS) regression of $y_{(i)}^{u} = a - b\eta$ provides estimates of the Gumbel parameters for the complete sample $\hat{\xi} = a$ and $\hat{\beta} = b$. Instead of the OLS estimates of the Gumbel model parameters, one could also use the maximum likelihood estimators (MLE) for the parameters of a Gumbel distribution for the case of type I left censoring introduced by Leese (1973). It remains an open question as to how the Leese (1973) MLE compares with PPR for estimation of the parameters of a Gumbel distribution under type I left censoring.

Figure 1b illustrates the Gumbel PPR method for censored datasets applied to the $n - m = 24$ uncensored observations of Region 26 (India – Xizand – Sichuan – Yunnan) with $m = 5$ estimates of the censored observations obtained from the PPR method. This approach yields an estimate of the complete sample, which can then be used to estimate any desired statistic.

PPR can also be used to estimate the magnitude of events with larger average return periods than the sample size n. For example, one is often interested in the event that has a 2% or 10% probability of exceedance in 50 years, which corresponds to return periods of 2475 and 475, respectively, assuming a Poisson process. Thus, the magnitude of the 2475- and 475-year earthquake can be estimated using $y_p = a - b\ln[-\ln(p)]$, where $p = 1 - (\text{return period})^{-1}$. The PPR approach could also be combined with the index earthquake method introduced by Thompson et al. (2007).

5 The probability plot correlation coefficient hypothesis test for censored Gumbel observations

The PPCC is a goodness-of-fit measure that describes the degree of linearity of the probability plot for the hypothesized distribution. Following Vogel (1986) and others, we define the PPCC as the Pearson product moment correlation coeffi-

Table 1. Critical values of the PPCC test statistic for $\lambda = 0$. Reported values are $1000(1 - r)$.

n			α			
	0.05	0.10	0.25	0.75	0.90	0.95
10	91.25	73.30	49.95	19.97	12.73	9.688
20	61.50	48.74	33.12	14.25	9.736	7.754
30	48.11	37.70	25.52	11.22	7.832	6.627
100	22.03	16.78	11.15	5.022	3.623	3.001
500	6.667	5.043	3.323	1.649	1.116	0.935
1000	3.791	2.876	1.918	0.956	0.652	0.548
5000	0.994	0.761	0.507	0.239	0.176	0.149

cient between the ordered uncensored observations and their expectation under the hypothesized Gumbel model.

The PPCC goodness-of-fit statistic can be used as a test statistic for hypothesis testing. Vogel (1986) provides tables of the critical values of the PPCC for complete Gumbel samples and Heo et al. (2008) provides regression equations that relate the critical values of the Gumbel PPCC to sample size and significance level. Our goal is to extend those results to left-censored Gumbel samples, where the fraction of censored data, is termed the censoring level: $\lambda = m/n$.

We use the Monte-Carlo method to generate $M = 50\,000\,000\,n$ complete Gumbel samples with sample sizes from $n = 10$ to 5000. The reason for generating $M = 50\,000\,000\,n$ samples is that the PPCC statistic is more stable for larger sample sizes. Thus, a smaller number of replicate Monte-Carlo experiments are needed for large sample sizes than for small sample sizes. Critical values of the PPCC (r_α) are computed from each set of M experiments so that the α percentile of the PPCC is the value of r that is exceeded $100(1-\alpha)\%$ of the time. Assumed values of α include 0.01, 0.05, 0.10, 0.25, 0.5, 0.75, 0.90, 0.95 and 0.99. We then censor each complete Gumbel sample such that the lowest $(100\lambda)\%$ of the observations were removed for $\lambda = (0.0, 0.1, \ldots, 0.8)$. We implement these simulations within the open-source statistical software R (R Development Core Team, 2010).

Tables 1 through 3 provide the critical values of the PPCC for selected values of sample size n, censoring level λ, and significance level α. Tables 1 through 3 report values of $1000(1 - r)$ because many of the values of r are nearly indistinguishable from unity. All of the values reported in Tables 1 through 3 have 95% confidence intervals that fall within $\pm 2\%$ of the values reported in the table.

Most previous efforts to develop PPCC hypothesis tests have reported critical values of the test statistics in tables analogous to Tables 1 through 3. For example, Vogel and MacMartin (1991) developed regression equations that relate the PPCC values to n, α, and the skew coefficient for complete samples drawn from a Log Pearson type III

Table 2. Critical values of the PPCC test statistic for $\lambda = 0.3$. Reported values are $1000(1-r)$.

n	α					
	0.05	0.10	0.25	0.75	0.90	0.95
10	114.8	92.24	61.24	20.65	11.85	8.376
20	79.77	63.67	42.86	16.75	10.64	8.120
30	63.99	50.52	33.87	13.70	8.987	7.002
100	31.56	24.07	15.77	6.635	4.589	3.697
500	10.34	7.744	5.020	2.166	1.539	1.265
1000	6.082	4.567	2.960	1.292	0.9227	0.7618
5000	1.608	1.222	0.8123	0.3641	0.2646	0.2204

Table 3. Critical values of the PPCC test statistic for $\lambda = 0.6$. Reported values are $1000(1-r)$.

n	α					
	0.05	0.10	0.25	0.75	0.90	0.95
10	151.2	121.5	73.83	16.15	6.750	3.428
20	110.3	88.72	58.83	19.85	11.46	8.172
30	89.74	71.89	48.22	17.83	10.92	8.149
100	46.86	36.32	24.00	9.786	6.553	5.176
500	16.94	12.66	8.183	3.449	2.415	1.964
1000	10.21	7.625	4.926	2.104	1.488	1.219
5000	2.833	2.148	1.411	0.6208	0.4453	0.3697

distribution. Similarly, Heo et al. (2008) developed regression based PPCC tests for complete samples from numerous other distributions including the normal, Gumbel, Gamma, GEV, and Weibull distributions.

We develop the following relationship between values of $r = f(n, \alpha, \lambda)$ using multivariate regression methods:

$$\hat{r} = 1 - \frac{1}{\exp([A + B\ln(n)]^2)}, \quad \text{where:}$$

$$A = C + D\lambda$$

$$B = \frac{1}{(\exp[E + F\lambda])^2}$$

$$C = 0.5257 + 1.037 \left[\ln\left(\frac{1}{1-\alpha}\right) \right]^{1/6}$$

$$D = -0.2930 - 1.815\alpha + 18.83\alpha^2 - 73.42\alpha^3$$
$$\quad + 140.8\alpha^4 - 129.6\alpha^5 + 45.88\alpha^6$$

$$E = 0.8937 \cdot 1.080^\alpha \cdot \alpha^{0.00927}$$

$$F = 0.05572 - 1.167\alpha + 12.67\alpha^2 - 55.64\alpha^3$$
$$\quad + 117.4\alpha^4 - 116.1\alpha^5 + 43.34\alpha^6. \tag{8}$$

where n, α, λ are the sample size, significance level, and censoring level. Equation (8) was developed by noting that a linear relationship exists between the logit function
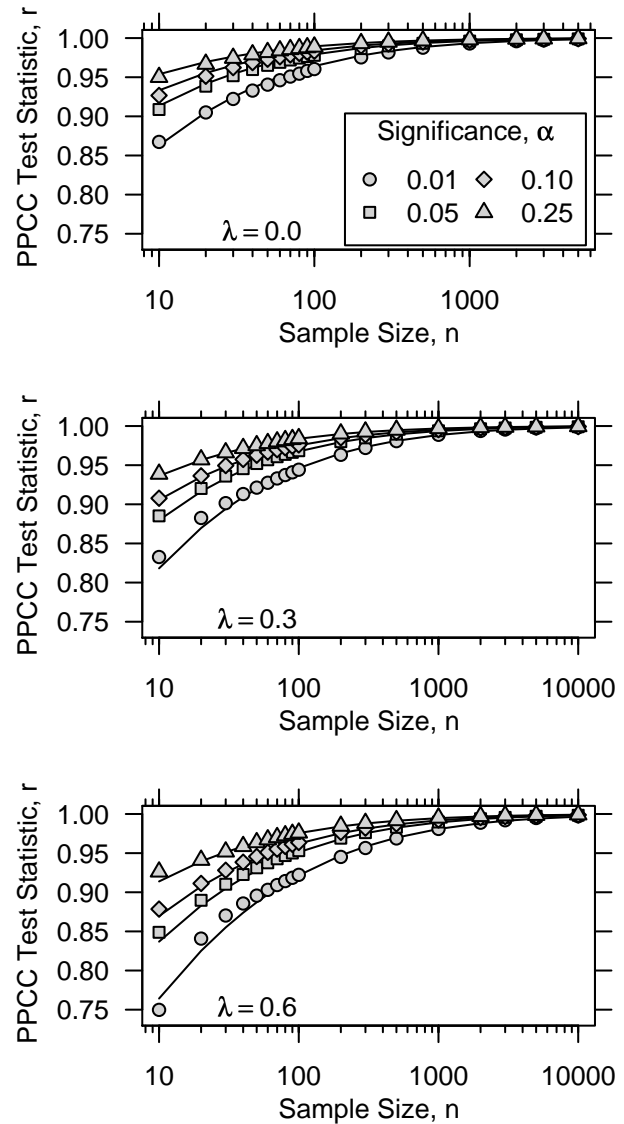

Fig. 2. Comparison of regression (solid line) and simulated values of the censored Gumbel PPCC $= r$ values for various values of significance level α, censoring level λ, and sample size n.

$\ln[1/(1-r)]$ and $\ln(n)$ for a given value of λ. Figure 2 compares the regression estimates (curves) with the simulated values of the PPCC $= r$ for selected values of n, λ, and α. Figure 2 documents a satisfactory fit between the regression in Eq. (8) and the results of the Monte-Carlo experiments, with the regression performing best as both n and α increase.

6 Applications of censored Gumbel PPCC test to earthquake magnitudes

Thompson et al. (2007) explored various distributional hypotheses for observations of AM earthquake magnitudes in 50 seismic regions across the globe, based on the

Flinn-Engdahl regionalization scheme (Young et al., 1996). Based on the Gumbel PPCC test developed by Vogel (1986) for complete samples, Thompson et al. (2007) only rejected the Gumbel hypothesis for 3 of the 46 tested regions using a 5% significance level. When one applies a hypothesis test 46 times, using a 5% level significance level, one expects $0.05(46) = 2.3$ rejections. Since Thompson et al. (2007) obtained 3 rejections, they could not reject the overall null hypothesis that AM earthquake magnitudes for all regions of the globe follow a Gumbel distribution. However, the Gumbel PPCC test statistic developed by Vogel (1986) and used by Thompson et al. (2007) does not consider the impact of censoring, so their conclusions are open to question.

The censored Gumbel hypothesis test presented in this study overcomes this limitation. Here we replicate the Gumbel PPCC test performed by Thompson et al. (2007) but also account for the important fact that no earthquakes were observed above the assumed $m_c = 5.8$ in many years. Table 4 reports the record lengths of the censored m and uncensored $n - m$ observations along with the censoring level λ for each of the 46 earthquake regions. Here, censoring levels λ range from 0 to 0.86 (note that we only analyze regions with $n - m \geq 4$) with a median of 0.12. See Sect. 2 for further discussion of the earthquake data.

Note that Table 4 also reports the computed significance level $\hat{\alpha}$ for each region computed from Eq. (8). The estimated significance level $\hat{\alpha}$ is the nonexceedance probability of obtaining the PPCC test statistic under the assumption that the Gumbel null hypothesis is true. This value is often referred to as the "p-value" in the context of hypothesis testing, however, we have already defined the variable "p" as the plotting position so we use the symbol $\hat{\alpha}$ to avoid ambiguity.

Assuming a significance level of 5%, we also use Eq. (8) to estimate the critical values of the Gumbel PPCC for each region. Table 4 documents that the 5% critical value of the PPCC is greater than the computed PPCC for only one region (Region 45: Macquarie Loop). Region 45 is the only region where $\hat{\alpha} < 5\%$. Therefore, we could only reject the null hypothesis that AM earthquake magnitudes follow a Gumbel distribution in one of the 46 regions using a 5% significance level. The suite of 46 individual 5% level tests in Table 4 is difficult to interpret, so in the following section we performed more rigorous tests termed "field significance tests" or "multiple comparison procedures". These tests evaluate the behavior of the entire ensemble of 46 independent hypothesis test results into a single collective hypothesis test.

6.1 Probability plot regression for censored Gumbel samples

Table 4 summarizes the results of a large number of individual hypothesis tests. A set of multiple hypothesis tests such as this is difficult to interpret because the null hypothesis may be rejected by chance even if the null hypothesis is in fact true. Put differently, the likelihood of a type I error

Table 4. Summary statistics of the global CMT catalog earthquake data.

Region	m	$n - m$	l	PPCC	Crit. PPCC	$\hat{\alpha}$
1	0	29	0.00	0.9609	0.9525	0.0865
2	15	14	0.52	0.9791	0.9105	0.6520
3	10	19	0.34	0.9386	0.9269	0.0853
4	17	12	0.59	0.9655	0.9032	0.4222
5	0	29	0.00	0.9816	0.9525	0.4131
6	0	29	0.00	0.9772	0.9525	0.2879
7	5	24	0.17	0.9866	0.9409	0.7148
8	0	29	0.00	0.9845	0.9525	0.5242
9	14	15	0.48	0.9486	0.9140	0.1912
10	0	29	0.00	0.9872	0.9525	0.6472
11	5	24	0.17	0.9753	0.9409	0.3552
12	0	29	0.00	0.9783	0.9525	0.3151
13	0	29	0.00	0.9843	0.9525	0.5150
14	0	29	0.00	0.9730	0.9525	0.2068
15	0	29	0.00	0.9694	0.9525	0.1575
16	0	29	0.00	0.9896	0.9525	0.7679
17	14	15	0.48	0.9841	0.9140	0.7534
18	0	29	0.00	0.9847	0.9525	0.5304
19	0	29	0.00	0.9778	0.9525	0.3007
20	4	25	0.14	0.9674	0.9434	0.2021
21	2	27	0.07	0.9869	0.9481	0.6744
22	0	29	0.00	0.9719	0.9525	0.1902
23	0	29	0.00	0.9689	0.9525	0.1519
24	0	29	0.00	0.9765	0.9525	0.2715
25	16	13	0.55	0.9865	0.9069	0.8253
26	5	24	0.17	0.9880	0.9409	0.7677
27	10	19	0.34	0.9290	0.9269	0.0549
28	22	7	0.76	0.9806	0.8827	0.7561
29	6	23	0.21	0.9588	0.9383	0.1485
30	3	26	0.10	0.9854	0.9458	0.6327
31	14	15	0.48	0.9641	0.9140	0.3430
32	1	28	0.03	0.9507	0.9504	0.0510
33	0	29	0.00	0.9810	0.9525	0.3918
34	22	7	0.76	0.8960	0.8827	0.0759
37	12	17	0.41	0.9833	0.9207	0.7112
38	25	4	0.86	0.9876	0.8690	0.8868
39	20	9	0.69	0.9486	0.8913	0.2822
40	21	8	0.72	0.9604	0.8871	0.4240
41	6	23	0.21	0.9544	0.9383	0.1176
42	24	5	0.83	0.9484	0.8737	0.3449
43	1	28	0.03	0.9973	0.9504	0.9988
44	14	15	0.48	0.9797	0.9140	0.6512
45	1	28	0.03	0.8972	0.9504	0.0022
46	3	26	0.10	0.9492	0.9458	0.0610
47	21	8	0.72	0.9438	0.8871	0.2610
48	2	27	0.07	0.9741	0.9481	0.2658

for the group of tests as a whole is not equal to the likelihood of a type I error chosen for each individual test. Numerous multiple comparison procedures (MCP) have been developed in the statistics literature to account for this phenomenon. The most widely used approaches in the statistics literature are based on the Bonferroni equality (Simes, 1986) or various modifications of that test (e.g., Rice, 1989;

Benjamini and Hochberg, 1995). Other MCP tests have been proposed by Livezey and Chen (1983), Douglas et al. (2000), and Vogel and Kroll (1989). The field significance (α_f) is the collective significance of a group of individual hypothesis tests. Following Ventura et al. (2004) and Vogel et al. (2008), we use the methods of Simes (1986), Benjamini and Hochberg (1995), Livezey and Chen (1982), and Vogel and Kroll (1989) to evaluate the overall, or joint, significance level associated with the group of individual tests reported in Table 4.

Here we define the field significance α_f as the probability that a suite of N individual hypothesis tests will reject the null hypothesis when it is true. It should be thought of as the overall collective significance level of the group of hypothesis tests. We describe four different MCPs below. In each, the criterion for rejection is defined differently.

6.2 Bonferroni-type multiple comparison procedure

Suppose that we want to evaluate an ensemble of N independent hypothesis tests and that the probability of rejecting any one or more of the individual hypotheses is α_f, given that the hypothesis is true. Also assume that the N individual hypothesis tests are independent of one another, and that each has type I error probability α. A Bonferroni type test (Simes, 1986) states that

$$1 - \alpha_f = (1 - \alpha)^N. \tag{9}$$

Thus, we can compute the α required for each individual test to achieve a chosen field significance α_f with

$$\alpha = 1 - \sqrt[N]{1 - \alpha_f}. \tag{10}$$

Equation (10) is often approximated by $\alpha = \alpha_f/N$. This approximation is only accurate to about 2.5% of the exact value for $\alpha_f = 0.05$, however, and is not justified because of the ease with which Eq. (10) can be computed for any value of N.

Using $\alpha_f = 5\%$ for $N = 46$ tests in Table 4, Eq. (10) leads to an individual hypothesis test significance level of $\alpha \approx 0.00111$. Since $\hat{\alpha} > 0.00111$ for all regions, there is no evidence to reject the Gumbel null hypothesis using a 5% level Bonferroni-type test. The overall Gumbel null hypothesis would be rejected for an $\alpha_f = 0.05$ level Bonferroni-type MCP if one or more of the computed significance levels in Table 4 were less than 0.00111.

6.3 False discovery rate multiple comparison procedure

Benjamini and Hochberg (1995) introduced an improvement over the Bonferroni-type test that attempts to control for what they term the false discovery rate (FDR). FDR is the number of false rejections of the null hypothesis (see Rice, 1989). Benjamini and Hochberg (1995) also found that their FDR procedure led to considerable gains in statistical power over the traditional Bonferroni-type test. The traditional

Bonferroni-type test rejects the null hypothesis for all regions $i = 1, \ldots, N$ if any $\hat{\alpha}$ is less than the value of α computed from Eq. (10). In contrast, the rejection threshold for the FDR procedure of Benjamini and Hochberg (1995) is variable. Let $\hat{\alpha}_{(i)}$ be the N values of $\hat{\alpha}$ ranked in ascending order. The FDR procedure rejects the null hypothesis for all regions $i = 1, \ldots, k$, where k is the largest value of i for which

$$\hat{\alpha}_{(i)} \le \frac{i}{N} \alpha_f. \tag{11}$$

After ranking the computed significance levels in Table 4, we find that Eq. (11) holds for all N regions for $\alpha_f = 0.05$. Thus, we cannot reject the Gumbel null hypothesis for all 46 regions, using $\alpha_f = 5\%$ with the FDR MCP. The null hypothesis would be rejected for an $\alpha_f = 0.05$ level FDR MCP if Eq. (11) was violated for one or more individual samples.

6.4 Binomial multiple comparison procedure

If each hypothesis test is independent, then each test is a Bernoulli trial with probability $\alpha = 0.05$ of rejecting the Gumbel null hypothesis, given that the data are sampled from a Gumbel distribution. Thus, the probability mass function of the number of rejections X in a series of N independent tests follows a binomial distribution with parameters N and α. It follows that α_f is a function of the number of rejections and the binomial distribution parameters

$$\alpha_f = P[X \ge x] = 1 - P[X < x]$$
$$= 1 - \sum_{k=0}^{x-1} \frac{N!}{k!(N-k)!} \alpha^k (1-\alpha)^{N-k}. \tag{12}$$

Figure 3 plots α_f as a function of the number of rejections for $\alpha = 0.01$, 0.05 and 0.10 assuming $N = 46$. Further, Eq. (12) gives $\alpha_f = 0.906$ for an $\alpha = 0.05$ level test and $x = 1$ rejection, as reported in Table 4. The probability that these results are not due to chance is 90.6% assuming the AM earthquake observations are Gumbel. Note that $\alpha_f = 0.079$ for $x = 4$ rejections, and $\alpha_f = 0.026$ for $x = 5$ rejections. Thus, we would reject the null hypothesis for an $\alpha_f = 0.05$ level binomial MCP with $N = 46$ only if five or more individual tests were rejected.

6.5 Uniform PPCC multiple comparison procedure

The previous tests only considered the results of N individual hypothesis tests, each with fixed significance level α. If each test is independent, then the N values of $\hat{\alpha}$ in Table 4 are random samples from a uniform distribution over the interval [0, 1] (Casella and Berger, 1990). Thus, another MCP evaluates the null hypothesis that the $\hat{\alpha}$ values in Table 4 follow a uniform distribution.

Vogel and Kroll (1989) developed a PPCC hypothesis test for the uniform distribution with critical values of the uniform PPCC reported for sample sizes in the range [10, 1000] and for significance levels in the range [0.01, 0.99]. To test

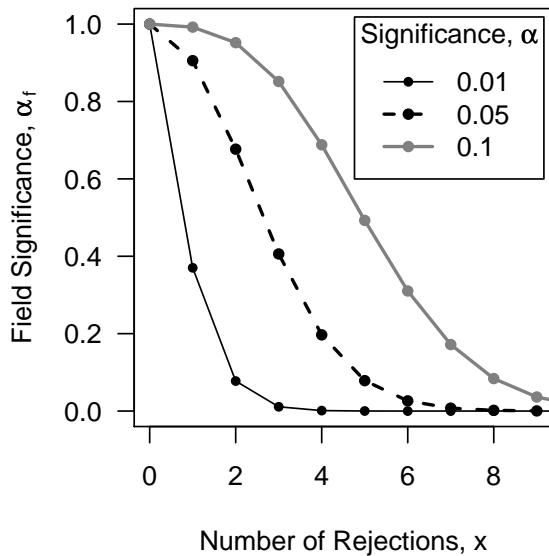**Fig. 3.** Plot of the field significance level α_f associated with $n = 46$ hypothesis tests, each with an individual significance levels of $\alpha = 0.01$, 0.05 and 0.10 based on Eq. (10).
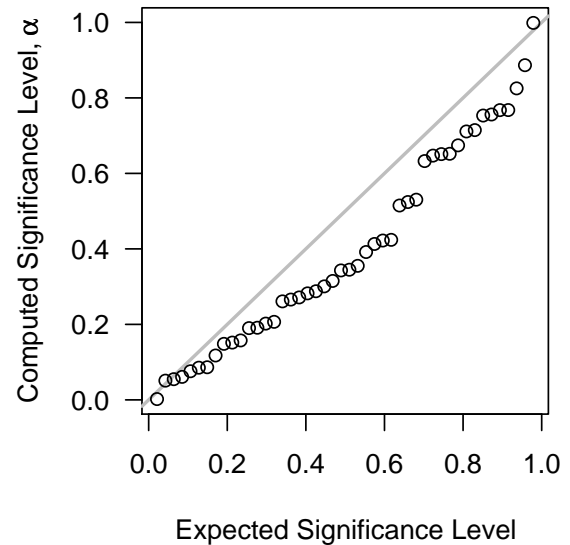

Fig. 4. Uniform probability plot of local significance levels α reported in Table 2. Here the uniform PPCC $r = 0.9852$ and the 5% critical value is $r_{0.05} = 0.9801$.

the null hypothesis that $\hat{\alpha}$ follows a uniform distribution over the interval [0, 1], we construct a uniform probability plot in Fig. 4. If the ordered $\hat{\alpha}$ values $\hat{\alpha}_{(i)}$ are considered independent and are ranked in ascending order, then they follow a beta distribution (David and Nagaraja, 2003; Loucks et al., 1981) with expectation

$$E[\alpha_{(i)}] = u_i = \frac{i}{N+1}. \tag{13}$$

Note that u_i is known as the Weibull plotting position and here it provides an unbiased estimate of the expectation of the ordered uniform random variables, analogous to the way the Gringorten plotting position in Eq. (5) provides unbiased estimates of the ordered values from a Gumbel distribution.

Figure 4 displays the uniform probability plot of the computed significance levels, $\hat{\alpha}_{(i)}$ versus the Weibull plotting position u_i. If the earthquake magnitudes follow a Gumbel distribution, then the points in Fig. 4 should be located near the line of equality. The more linear the uniform probability plot in Fig. 4 is, the less evidence there is for rejecting the Gumbel null hypothesis. The data in Table 4 give $r = 0.9852$. From Table 1 in Vogel and Kroll (1989), the critical value of the uniform PPCC test statistic at a 5% significance level is $r_{0.05} = 0.9801$. Since the computed uniform PPCC of 0.9852 is slightly greater than the critical value for a 5% level hypothesis test, we fail to reject the overall Gumbel null hypothesis for the 46 seismic regions at a field significance level of $\alpha_f = 5\%$. However, the computed PPCC only slightly exceeds the critical value. It follows that if we were to slightly increase our overall field significance level α_f, we would reject the null hypothesis, and thus our test results are by no means conclusive.

6.6 Power for the Generalized Extreme Value (GEV) alternative

The 5% level censored Gumbel hypothesis test combined with the various MCPs are designed so that the overall test will only reject 5% of samples drawn from a Gumbel distribution, which is termed the type I error. Since the MCP tests in the preceding sections all fail to reject the Gumbel null hypothesis, in this section we compute the likelihood that the test would reject the Gumbel null hypothesis when the samples actually arise from a different distribution (termed the type II error). The "power" of the hypothesis test is defined as the complement of the probability of a type II error and reflects the ability, or power, of the test to detect departures from the null hypothesis.

The GEV distribution is a generalization of Gumbel's type I, II, and III distributions, introduced to hydrology by Jenkinson (1969) and to seismology by Makjanic (1980). It is a more flexible (three parameter) distribution than the Gumbel (two parameter) distribution. Importantly, the GEV can exhibit upper and lower bounds depending on the value of its shape parameter κ, which overcomes the physically unrealistic unboundedness of the Gumbel distribution. When the magnitudes associated with the POT series follows a Generalized Pareto distribution and the number of arrivals of earthquakes are assumed to follow a Poisson process, then the AM series is GEV. The Generalized Pareto distribution has recently been applied to earthquake POT series by Pisarenko and Sornette (2003).

The Gumbel distribution is a special case of the GEV distribution when the shape parameter $\kappa = 0$. For $\kappa = 0$, the GEV reduces to a Gumbel distribution and exhibits no upper bound. However, for $\kappa > 0$ the GEV distribution exhibits an upper bound and is termed the reverse Weibull distribution (Simiu and Heckert, 1996). Thus, the reverse Weibull is an important alternative distribution to consider since physical constraints require that earthquake magnitudes exhibit a finite upper bound. In the following power study, we document that a global Gumbel hypothesis test often fails to reject samples that arise from a reverse Weibull distribution. Thus, the reverse Weibull distribution is also a viable choice for modeling earthquakes.

We simulate the earthquake catalog by generating $N = 46$ samples with $n = 29$ complete record lengths from the reverse Weibull distribution for a given value of $\kappa \geq 0$. Thompson et al. (2007, Eq. A31) provides the GEV quantile function in terms of the AM lower bound ξ, scale parameter β^*, and shape parameter κ. For a fixed κ, we use the method of moments to estimate ξ and β^* from Eqs. (B7) and (B8) in Thompson et al. (2007). Note that we estimate the sample mean and sample standard deviation using the PPR method for censored samples. We then censor these samples to match the λ of the corresponding observation in the earthquake catalog. For each $\kappa = 0, 0.05, \ldots, 0.5$ we generate $M = 100\,000$ synthetic earthquake catalogs and compute $\hat{\alpha}$ for the censored Gumbel hypothesis test. We then apply each of the previously described MCP tests and count the number of times the MCP rejects the null hypothesis N_r. For all cases where $\kappa > 0$, we estimate the power of the censored Gumbel hypothesis test against the reverse Weibull alternative distribution, where Power $= 1 - (N_r/M)$. Figure 5 plots the estimate of the probability of a type II error (N_r/M) of the censored Gumbel test against reverse Weibull alternatives as a function of κ for each MCP.

Figure 5 illustrates that the censored Gumbel PPCC hypothesis test is generally unable to detect slight departures from the Gumbel distribution but can begin to discriminate between the GEV and Gumbel models for larger values of κ ($\kappa > 0.2$). These results indicate that, although we cannot reject the Gumbel model, it is possible that that the samples arises from a reverse Weibull model with a small positive value of the shape parameter ($0 < \kappa < 0.2$).

As expected from previous research (Benjamini and Hochberg, 1995), the FDR test exhibits slightly higher power than the Bonferroni test. Also, the uniform test shows slightly higher power than the binomial test for detecting departures from a Gumbel PDF for the sampling characteristics of this earthquake catalog. More importantly, the binomial and uniform MCPs exhibit substantially higher power than either the Bonferroni and FDR MCPs. This result is specific to the sample sizes and distributions that we analyze in this paper, so further investigation is warranted to determine if these findings can be generalized.
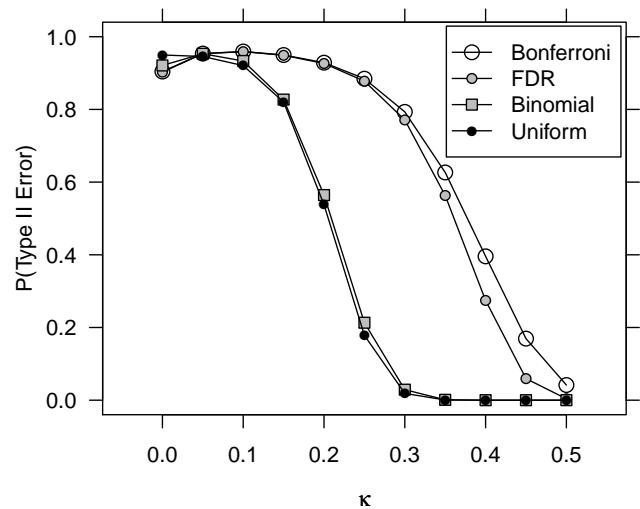

Fig. 5. Power of a type I left censored Gumbel PPCC hypothesis test combined with the four field significance level tests discussed in this paper against the reverse Weibull alternative. Power is a function of the reverse Weibull shape parameter $\kappa > 0$. As κ increases, the distribution becomes less similar to the Gumbel distribution, so the differences are easier to detect.

7 Summary and conclusions

Our primary goal is to present a rigorous hypothesis test for evaluating whether or not left-censored observations arise from a Gumbel distribution. To accomplish this, we extend the PPCC hypothesis test for complete Gumbel samples to left-censored observations. The censored Gumbel PPCC hypothesis test can be easily extended to other types of censoring, as well as other probability distributions and the tests outlined here are applicable to any problem where type I left censoring arises, such as in the analysis of water quality, and other environmental data (Helsel and Hirsch, 2002).

We compute critical values for the left censored Gumbel PPCC test statistic from Monte Carlo simulations for a variety of sample sizes n, censoring levels λ, and significance levels α. The results are summarized in the form of a regression equation (see Tables 1 through 3, Eq. (8), and Fig. 2). We illustrate the application of the hypothesis test with 46 time series of AM earthquakes. Each of the 46 hypothesis tests have significance levels greater than 0.05, except for one region. To help interpret the results of the suite of 46 individual hypothesis tests, we employ four field significance level hypothesis tests (termed multiple comparison procedures or MCPs) that have been previously developed in the statistical, climate, and hydrology literature. These MCPs failed to reject the overall hypothesis that censored AM earthquake magnitudes arise from a Gumbel distribution using an overall 5% significance level. Since the results were inconclusive a power study was performed.

A power study documents that the various MCPs could not detect small but reasonable departures from the null hypothesis (GEV with $0 < \kappa < 0.2$). Thus, although the global earthquake catalog is consistent with the Gumbel hypothesis, we could not rule out the possibility that the earthquake observations arise from a reverse Weibull model (a generalized extreme value (GEV) distribution which exhibits a finite upperbound). Furthermore, the power study indicates that the binomial and uniform MCPs are substantially more powerful than the Bonferonni and FDR tests. These results warrant further exploration to determine whether or not these findings can be generalized beyond the special cases considered here.

Edited by: M. E. Contadakis
Reviewed by: D. Shanker and another anonymous referee

References

Benjamini, Y. and Hochberg, Y.: Controlling the false discovery rate: A practical and powerful approach to multiple testing, J. R. Stat. Soc., 57, 289–300, 1995.

Berthouex, P. M. and Brown, L. C.: Statistics for Environmental Engineers, Lewis, Boca Raton, FL, 335 pp., 1994.

Burton, P. W., Qin, C. Y., Tselentis, G. A., and Sokos, E.: Extreme earthquake and earthquake perceptibility study in Greece and its surrounding area, Nat. Hazards, 32, 227–312, 2004.

Casella, G. and Berger, R: Statistical Inference, Wadsworth and Brooks/Cole, Pacific Grove, CA, 650 pp., 1990.

Chowdhury, J., Stedinger, J. R., and Lu, L.: Goodness-of-fit tests for regional generalized extreme value flood distributions, Water Resour. Res., 27, 1765–1776, 1991.

David, H. A. and Nagaraja, H. N.: Order Statistics, 3rd Edn., Wiley, Hoboken, NJ, 458 pp., 2003.

Douglas, E. M., Vogel, R. M., and Kroll, C. N.: Trends in flood and low flows in the United States, J. Hydrol., 240, 90–105, 2000.

Ekström, G. and Nettles, M.: Calibration of the HGLP seismograph network and centroid-moment tensor analysis of significant earthquakes of 1976, Phys. Earth Planet In., 101, 219–243, 1997.

Ekström, G., Dziewonski, A. M., Maternovskaya, N. N., and Nettles, M.: Global seismicity of 2003: Centroid-moment-tensor solution for 1087 earthquakes, Phys. Earth Planet In., 148, 327–351, 2005.

Filliben, J. J.: The probability plot correlation coefficient test for normality, Technometrics, 17, 111–117, 1975.

Gilliom, R. J. and Helsel, D. R.: Estimation of distributional parameters for censored trace level water quality data, 1. Estimation techniques, Water Resour. Res., 22, 135–146, 1986.

Gringorten, I. I.: A plotting rule for extreme probability paper, J. Geophys. Res., 68, 813–814, 1963.

Gumbel, E. J.: The return period of flood flows, Ann. Math. Stat., 12, 163–190, 1941.

Gumbel, E. J.: Statistics of Extremes, Columbia University Press, New York, NY, 375 pp., 1958.

Gupta, A. K.: Estimation of the mean and standard deviation of a normal population from a censored sample, Biometrika, 39, 260–273, 1952.

Gutenberg, B. and Richter, C. F.: Seismicity of the Earth, Princton University Press, Princeton, NJ, 310 pp., 1954.

Helsel, D. R. and Gilliom, R. J.: Estimation of distributional parameters for censored trace level water quality data, 2. Verification and applications, Water Resour. Res., 22, 147–155, 1986.

Helsel, D. R. and Hirsch, R. M.: Statistical Methods in Water Resources, Techniques of Water Resources Investigations, Book 4, Chapter A3, US Geological Survey, available at: http://pubs.usgs.gov/twri/twri4a3/, 2002.

Heo, J.-H., Kho, Y. W., Shin, H., Kim, S., and Kim, T.: Regression equations of probability plot correlation coefficient test statistics from several probability distributions, J. Hydrol., 335, 1–15, 2008.

Jenkinson, A. F.: Statistics of Extremes, in: Estimation of Maximum Floods, WMO No. 233, TP 126, (Technical note no. 98), 183–228, 1969.

Kagan, Y.: Seismic moment-frequency relation for shallow earthquakes: Regional comparison, J. Geophys. Res., 102, 2835–2852, 1997.

Kijko, A. and Graham, G.: Parametric-historic procedure for probabilistic hazard analysis: Part I: Estimation of maximum regional magnitude m(max), Pure Appl. Geophys., 152, 413–442, 1998.

Leese, M. N.: Use of censored data in the estimation of Gumbel distribution parameters for annual maximum flood series, Water Resour. Res., 9, 1534–1542, 1973.

Livezey, R. E. and Chen, W. Y.: Statistical field significance and its determination by Monte Carlo techniques, Mon. Weather Rev., 111, 46–59, 1983.

Looney, S. W. and Gulledge Jr., T. R.: Use of the correlation coefficient with normal probability plots, Am. Stat., 39, 75–79, 1985.

Loucks, D. P., Stedinger, J. R., and Haith, D. A.: Water Resources Systems Planning and Analysis, Prentice Hall, Inc., New Jersey, 559 pp., 1981.

Makjanic, B.: On the frequency distribution of earthquake magnitudeand intensity, B. Seismol. Soc. Am., 70(6), 2253–2260, 1980.

Millard, S. P. and Neerchal, N. K.: Environmental Statistics with S-Plus, CRC Press, Boca Raton, NIST/SEMATECH, e-Handbook of Statistical Methods, available at: http://www.itl.nist.gov/div898/handbook/, 2001.

Ogata, Y. and Katsura, K.: Analysis of temporal and spatial heterogeneity of magnitude frequency distribution inferred from earthquake catalogs, Geophys. J. Int., 113, 727–738, 1993.

Öztürk, S., Bayrak, Y., Cinar, H., Koravos, G. C., and Tsapanos, T. M.: A quantitative appraisal of earthquake hazard parameters computed from Gumbel I method for different regions in and around Turkey, Nat. Hazards, 47, 471–495, 2008.

Pisarenko, V. K. and Sornette, D.: Characterization of the frequency of extreme earthquake events by the generalized Pareto distribution, Pure Appl. Geophys., 160, 2343–2364, 2003.

R Development Core Team R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, 2010.

Rice, W. R.: Analyzing tables of statistical tests, Evolution, 43,

223–225, 1989.

Rydelek, P. A. and Sacks, I. S.: Testing the completeness of earthquake catalogs and the hypothesis of self-similarity, Nature, 337, 251–253, 1989.

Shanker, D., Yadav, R. B. S., and Singh, H. N.: On the seismic risk in the Hindukush-Pamir-Himalaya and their vicinity, Current Science, 92, 1625–1630, 2007.

Shumway, R. H., Azari, R. S., and Hayhanian, M.: Statistical approaches to estimating mean water quality concentrations with detection limits, Environ. Sci. Technol., 36, 3345–3353, 2002.

Simes, R. J.: An improved Bonferroni procedure for multiple tests of significance, Biometrika, 73, 751–754, 1986.

Simiu, E. and Heckert, N. A.: Extreme wind distribution tails: a "peaks over threshold" approach, J. Struct. Eng., 122, 539–547, 1996.

Stedinger, J. R., Vogel, R. M., and Foufoula-Georgiou, E.: Frequency Analysis of Extreme Events, Chapter 18, Handbook of Hydrology, McGraw-Hill Book Company, New York, NY, 1993.

Thompson, E. M., Baise, L. G., and Vogel, R. M.: A global index earthquake approach to probabilistic assessment of extremes, J. Geophys. Res., 112, B06314, doi:10.1029/2006JB004543, 2007.

Utsu, T.: Representation and analysis of the earthquake size distribution: A historical review and some new approaches, Pure Appl. Geophys., 155, 509–535, 1999.

Ventura, V., Paciorek, C. J., and Risby, J. S.: Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatalogical data, J. Climate, 17, 4343–4356, 2004.

Vogel, R. M.: The probability plot correlation coefficient test for normal, lognormal, and Gumbel distributional hypotheses, Water Resour. Res., 22, 587–590 (see correction in 23, p. 2013), 1986.

Vogel, R. M., Hosking, J. R. M., Elphick, C. S., Roberts, D. L., and Reed, J. M.: Goodness-of-fit of Probability Distributions for Sightings as Species Approach Extinction, B. Math. Biol., 71, 701–719, 2008.

Wiemer, S. and Wyss, M.: Minimum Magnitude of Completeness in Earthquake Catalogs: Examples from Alaska, the western United States and Japan, B. Seismol. Soc. Am., 90, 859–869, 2000.

Wilk, M. B. and Gnanadesikan, R.: Probability plotting methods for analysis of data, Biometrika, 55, 1–17, 1968.

Woessner, J. and Wiemer, S.: Assessing the Quality of Earthquake Catalogues: Estimating the Magnitude of Completeness and Its Uncertainty, B. Seismol. Soc. Am., 95, 684–698, 2005.