ORIGINAL ARTICLE

Goodness of Fit of Probability Distributions for Sightings as Species Approach Extinction

Richard M. Vogel[a,*], Jonathan R.M. Hosking[b], Chris S. Elphick[c],
David L. Roberts[d,e], J. Michael Reed[f]

[a] *Department of Civil and Environmental Engineering, Tufts University, Medford, MA 02155, USA*
[b] *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598, USA*
[c] *Department of Ecology and Evolutionary Biology, and Center for Conservation and Biodiversity, University of Connecticut, 75 North Eagleville Rd. U-3043, Storrs, CT 06269, USA*
[d] *Museum of Comparative Zoology, Harvard University, 26 Oxford St., Cambridge, MA 02138, USA*
[e] *Royal Botanic Gardens, Kew, Richmond, Surrey, TW9 3AB, UK*
[f] *Department of Biology, Tufts University, Medford, MA 02155, USA*

Abstract Estimating the probability that a species is extinct and the timing of extinctions is useful in biological fields ranging from paleoecology to conservation biology. Various statistical methods have been introduced to infer the time of extinction and extinction probability from a series of individual sightings. There is little evidence, however, as to which of these models provide adequate fit to actual sighting records. We use L-moment diagrams and probability plot correlation coefficient (PPCC) hypothesis tests to evaluate the goodness of fit of various probabilistic models to sighting data collected for a set of North American and Hawaiian bird populations that have either gone extinct, or are suspected of having gone extinct, during the past 150 years. For our data, the uniform, truncated exponential, and generalized Pareto models performed moderately well, but the Weibull model performed poorly. Of the acceptable models, the uniform distribution performed best based on PPCC goodness of fit comparisons and sequential Bonferroni-type tests. Further analyses using field significance tests suggest that although the uniform distribution is the best of those considered, additional work remains to evaluate the truncated exponential model more fully. The methods we present here provide a framework for evaluating subsequent models.

Keywords L-moments · Extinct birds · Field significance test · Biological records · Extirpation

*Corresponding author.
E-mail address:* richard.vogel@tufts.edu (Richard M. Vogel).

1. Introduction

A variety of statistical methods have been introduced for inferring the time of extinction as well as the probability of extinction associated with a sighting record (defined here as a sequence of observations across years) (Solow, 2005). Estimating the probability that a species is extinct and the timing of extinctions is useful in various biological fields, ranging from paleoecology to conservation biology. For instance, it is highly unlikely that the fossil record will capture the exact timing of extinction events. Consequently, statistics that provide estimates of extinction timing can refine interpretations of past extinction patterns in light of potential drivers of extinction (e.g., Marshall, 1995; Marshall and Ward, 1996; Wang and Marshall, 2004). In applied settings, the difficulty of confirming that a population is truly extinct (Solow, 1993a; Reed, 1996) necessitates the use of repeatable methods that can reliably assess when conservation efforts should end, so as to balance the risk of giving up when a species can still be saved against the risk of misdirecting limited resources (Collar, 1998). Though classical statistical methods introduced here may be used to estimate the date of extinction, estimating the probability of extinction is a more complex statistical problem requiring Bayesian methods (Solow, 1993a) and is not covered here.

Solow (2005) reviewed both parametric and nonparametric methods for estimating extinction parameters from sighting records. Although numerous probabilistic models have been introduced and applied, there is little evidence from previous work as to which of these models provide adequate fit to actual sighting records. Recent innovations in other fields relating to goodness of fit evaluations of probabilistic models has critically contributed to and benefited from advancement of these methods in other fields including statistics (Hosking, 1990) and hydrology (Hosking and Wallis, 1997; Stedinger et al., 1993). These statistical innovations provide an opportunity to evaluate the goodness of fit of the various probabilistic models introduced by Solow (2005) and others for modeling extinction times and their associated probabilities. In some sense, this work is analogous to recent work by Thompson et al. (2007) who used innovations in hydrology and statistics to improve our understanding of the statistical behavior of earthquakes. Our primary goal here is to assess the goodness of fit of various statistical models for sighting records, using data collected for a set of North American and Hawaiian bird populations that have either gone extinct, or are suspected of having gone extinct, over the last 150 years.

2. The theory of L-moments

There are a variety of methods for assessing the goodness of fit of a particular probability distribution to either an individual sample or a set of samples. Any such goodness of fit method applied to an individual sample will generally lack the statistical power (i.e., the ability to discriminate among alternative hypotheses) unless either the sample size is large, or one has many samples all arising from the same population so that the number of samples multiplied by the average sample size is large. An increasingly popular approach for assessing the goodness of fit of a particular probability distribution to observations involves the construction of L-moment ratio diagrams, introduced by Hosking (1990). Such a diagram provides a graphical evaluation of the goodness of fit of a particular probability distribution to a number of samples all drawn from the same population.

L-moments were introduced by Hosking (1990). Like the ordinary moments (mean, variance, etc.), they are measures of the location, scale, and shape of probability distributions. Denote by $X_{k:n}$ the kth smallest observation from a sample of size n, so that the ordered sample is $X_{1:n} \leq X_{2:n} \leq \cdots \leq X_{n:n}$. The first four L-moments of a probability distribution are defined by

$$\lambda_1 = E(X_{1:1}),$$

$$\lambda_2 = \frac{1}{2} E(X_{2:2} - X_{1:2}),$$

$$\lambda_3 = \frac{1}{3} E(X_{3:3} - 2X_{2:3} + X_{1:3}),$$

$$\lambda_4 = \frac{1}{4} E(X_{4:4} - 3X_{3:4} + 3X_{2:4} - X_{1:4}).$$

The first L-moment λ_1 is the mean of the distribution; λ_2 is a scale or dispersion measure. The L-moment ratios $\tau_r = \lambda_r/\lambda_2, r = 3, 4, \ldots$, are dimensionless measures of the shape of the distribution; they all take values between -1 and $+1$. Analogously to the ordinary moment ratios, the L-skewness τ_3 is a measure of skewness and the L-kurtosis τ_4 is a measure of kurtosis. The L-CV λ_2/λ_1 is the L-moment analogue of the conventional coefficient of variation C_v.

From a sample of data, the L-moments of the distribution from which the sample was drawn can be estimated by "sample L-moments," which are linear combinations of the ordered data values. Corresponding sample L-moment ratios can also be defined. Details are given by Hosking (1990, Section 3.1).

L-moment ratio diagrams are now used routinely in hydrology and meteorology, and increasingly in other fields. L-moment ratio diagrams are constructed by plotting sample L-moment ratios versus each other: for example, L-CV versus L-skewness. On the same plot, the theoretical relationships for various probability density functions are compared to the observations. Vogel and Fennessey (1993) have shown that L-moment ratio diagrams are nearly always an improvement over ordinary moment ratio diagrams because L-moment ratios are approximately unbiased, whereas ordinary moment ratios can exhibit enormous downward bias, particularly for skewed samples, even with extremely large samples. Hosking (1990), Stedinger et al. (1993), Hosking and Wallis (1997), and others have summarized the theory of L-moments, thus we do not reproduce the theory here.

3. Parametric approaches for estimation of extinction times T_E

We review the parametric approaches introduced by Solow (2005) and provide extensions to his results using recent innovations relating to the theory of L-moments. Consider the series of n sighting times at which a species is sighted $t_1 < t_2, < \cdots < t_n$ during a particular observation period $[0, T]$, ordered from earliest to latest sighting. Here, the upper end of the range $[0, T]$ is considered to be 2007. Further consider that the species of interest becomes extinct at an unknown time T_E.

3.1. Stationary Poisson process

When the preextinction sighting record follows a stationary Poisson process, then Solow (1993a, 2005) showed that the sighting times will follow a uniform probability distribution (UN)

$$f(t) = \frac{1}{T_E} \quad \text{for } 0 \leq t \leq T_E. \tag{1}$$

Note that this distribution is the simplest of the parametric models considered by Solow (2005) because it only has a single parameter T_E. In this case the cumulative distribution function (cdf) is given by

$$F(t) = p = \frac{t}{T_E} \tag{2}$$

and the quantile function is given by

$$t(p) = pT_E, \tag{3}$$

where $p = F(t)$ is the probability that the sighting time is less than a particular value t and $t(p)$ is the pth percentile of the sighting time. In this case, the first four L-moments are given by $\lambda_1 = T_E/2$, $\lambda_2 = T_E/6$, $\lambda_3 = 0$ and $\lambda_4 = 0$ (see Table 1 in Hosking, 1990). The first two L-moment ratios are then given by $\mathrm{L-CV} = \lambda_2/\lambda_1 = 1/3$ and L-skewness $= \lambda_3/\lambda_2 = 0$ (see triangle in Fig. 1).

A maximum likelihood estimator of the extinction time T_E is simply the largest observed sighting time t_n; however, this estimator of T_E is biased (see David and Nagaraja, 2003, p. 174; Solow, 2005). A preferred uniformly minimum-variance unbiased estimator (MVUE) of the extinction time is

$$\hat{T}_E = \frac{n+1}{n} t_n. \tag{4}$$

Since (4) is the MVUE, it would be impossible to find a better estimator for large samples if the stationary Poisson model is correct. However, in practice, sighting records are often short and we are not certain that the stationary Poisson is the correct model; hence, there is little assurance that the estimator in (4) is best among all possible estimators. Future research is needed to evaluate the small sample properties and robustness of alternative estimators of T_E.

3.2. Nonstationary Poisson process

If the sighting record rate prior to extinction declines, then the above approach will be incorrect, leading to extinction time estimates that occur too early. Solow (1993b, 2005) considered the case where the preextinction sighting rate declined exponentially at an unknown rate β, in which case the resulting probability distribution of the sighting times is given by the truncated exponential distribution (TEX)

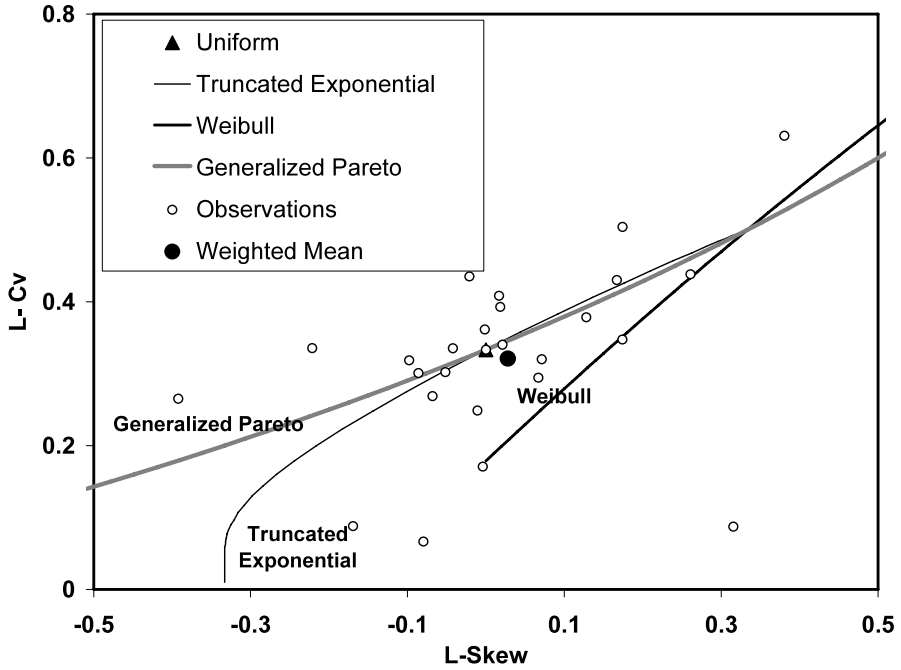$$f(t) = \frac{\beta \exp(-\beta t)}{1 - \exp(-\beta T_E)} \quad \text{for } 0 \leq t \leq T_E. \tag{5}$$

Fig. 1 L-moment ratio diagram of bird sighting data

In this case, the cdf is given by

$$F(t) = p = \frac{1 - \exp(-\beta t)}{1 - \exp(-\beta T_E)} \quad \text{for } 0 \le t \le T_E, \tag{6}$$

which can be inverted to yield the quantile function

$$t(p) = \frac{\ln[1 - p(1 - e^{-\beta T_E})]}{-\beta}. \tag{7}$$

Solow (2005) provides the MVUE for this case, which was originally derived by Beg (1982). Again, there is no assurance that the MVUE is the best choice among estimators for small samples or in situations when the model assumption is incorrect. For example, Hosking and Wallis (1987) showed that both method of moments and L-moment estimators of the parameters of a generalized exponential model (termed the generalized Pareto model) are preferred to the maximum likelihood estimators (MLE) for sample sizes below 500. MLEs are only MVUE for large samples; hence, there is no assurance that the estimator provided by Solow (2005) has lower mean square error than either an L-moment or moment estimator for the small sample sizes considered here. Furthermore, the moment and L-moment estimators are often much simpler to implement than either the MVUE or the MLE. Appendix A introduces L-moment estimators for the parameters of the truncated exponential distribution.

4. Testing whether extinction has occurred

Suppose the sighting times t_i have a cdf given by $F(t)$. If the n observed sighting times are assumed to be independent, then the distribution of the most recent sighting $T_n \max(t_1, t_2, \ldots, t_n)$ is given by

$$P[T_n \leq t_n] = p = \left[F(t_n) \right]^n. \tag{8}$$

The result in (8) is a well-known general result (David and Nagaraja, 2003) which also applies to the largest observed flood on record (see derivation in Douglas and Vogel, 2006).

The probability p in (8), when evaluated using the form of F(t) for $T_E = T$, can be regarded as the p-value associated with a test of the null hypothesis that extinction has not occurred by time T (Solow, 1993a). The interpretation of this probability in classical (frequentist) statistical inference is that when the last sighting occurs at time t_n one may conclude that either the species became extinct before time T or that an event (species is extant but not seen) of probability at most p has occurred. Direct inference about the probability of extinction requires additional assumptions; for example, Bayesian inference (as in, e.g., Solow, 1993a) requires the specification of a prior distribution of the time of extinction.

4.1. Stationary Poisson process

The stationary Poisson process assumes that the preextinction population size, sighting probabilities, and sighting efforts were stable over the observation period (Solow, 1993a). Under the assumption of a stationary Poisson process, the sighting times t_i are uniformly distributed, and the p-value of a test that extinction has occurred is obtained by substitution of (2), with $T_E = T$, into (8), which yields

$$p_{\text{UN}} = \left[\frac{t_n}{T} \right]^n, \tag{9}$$

where the subscript "UN" denotes the assumption that sighting times follow a uniform distribution.

4.2. Nonstationary Poisson process

Similarly, for a nonstationary Poisson process, the p-value of a test that extinction has occurred is obtained by substitution of (6), with $T_E = T$, into (8), and yields

$$p_{\text{TEX}} = \left[\frac{1 - \exp(-\beta t_n)}{1 - \exp(-\beta T)} \right]^n, \tag{10}$$

where the subscript "TEX" denotes the assumption that sighting times follow a truncated exponential distribution. Note that this probability is always greater than (9), so assuming that the Poisson process is nonstationary reduces the probability that the hypothesis of nonextinction will be rejected.

4.3. Weibull distribution

Solow (2005) shows that when the number of sightings n is large, the joint distribution of the k most recent sightings $t_{n-k+1} < t_{n-k+2} < \cdots < t_n$ will follow a Weibull distribution regardless of the parent probability distribution of the sighting times. Sighting records are often short, however, so the Weibull model which is an asymptotic extreme value distribution may not offer a good approximation. Solow (2005) used an estimate of the shape parameter of the Weibull distribution, $\hat{v}$,

$$\hat{v} = \frac{1}{k-1} \sum_{i=1}^{k-2} \ln\left[\frac{t_n - t_{n-k+1}}{t_n - t_{i+1}}\right] \tag{11}$$

to produce an approximate p value for testing for extinction.

$$p_{\text{WEI}} = \exp\left(-k\left(\frac{T - t_n}{T - t_{n-k+1}}\right)^{1/\hat{v}}\right). \tag{12}$$

This method had been used previously to estimate the extinction statistics for the Dodo (*Raphus cucullatus*) (Roberts and Solow, 2003).

5. Sighting data

To test how well actual sighting records fit the distributions underlying each of the analytical methods, we used a recently compiled set of sighting records for bird populations that have putatively gone extinct in North America and Hawaii during the last 150 years (Elphick et al., unpublished data). Forty-one bird species, subspecies, and distinct island populations that are either considered extinct or for which there are no unambiguous recent sightings were included in the original data set. For each population, we identified years in which there were undisputed records of the species. Of these sighting records, we eliminated those with fewer than 5 observations, leaving 27 sighting series for analysis.

6. Results

6.1. Graphical goodness of fit

L-moment ratios were obtained for the bird sighting data using unbiased L-moment estimators introduced by Hosking (1990) and reported elsewhere (Stedinger et al., 1993; Hosking and Wallis, 1997). The resulting L-moment diagram for the data series is shown in Fig. 1. The open circles are the estimated L-moment ratios for the series of bird sightings. The solid circle depicts the record-length weighted average value of the sample estimates of L-CV and L-skewness which represents the overall average L-moment ratio for all 27 samples, accounting for the fact that each sample has a different length (i.e., number of sightings in the series). For comparison with those empirical or sample L-moment ratios, Fig. 1 also illustrates the theoretical relationships between L-CV and L-skewness

for a variety of probability distributions that have previously been considered for modeling the sighting times, including the truncated exponential (TEX), Weibull and uniform (UN) models. Also shown is the generalized Pareto (GP) model, which is a generalized exponential model that may prove useful in future studies. See Hosking and Wallis (1987) for further background on the GP model. The UN model plots as a point in Fig. 1, because the UN model always has the same L-CV and L-skewness, regardless of the values of its model parameters.

Since estimated L-moment ratios are known to be approximately unbiased (see Hosking, 1990), one expects the theoretical relationships in Fig. 1 to pass through the center of the observed L-moments, which is roughly the case for the TEX, UN and GP models. We conclude from the qualitative assessment in Fig. 1 that the TEX, UN and GP models all provide moderately good representations of the probability distribution of bird sightings, but that the Weibull model does not provide an adequate representation.

Note that we have only evaluated one- and two-parameter probability distributions in Fig. 1. For an evaluation of the goodness of fit of three-parameter distributions, an L-moment diagram illustrating the relationship between L-kurtosis and L-skewness would need to be constructed.

6.2. *Probability plot correlation coefficient goodness of fit comparisons*

A probability plot is a useful graphical tool for illustrating the cdf of a set of observations and for evaluating the goodness of fit of alternative probability distributions to those observations. A probability plot compares the ordered observations versus estimates of the ordered observations under an assumed probability distribution. For example, Solow et al. (2006) illustrate the use of probability plots for evaluating the goodness of fit of a probabilistic model to the observed radiocarbon ages of the 25 most recent remains of Alaskan horses and mammoths.

Consider the following example of a probability plot. Suppose one wishes to illustrate and evaluate the goodness of fit of a uniform distribution to the sighting record for the Eskimo Curlew which began in 1945 with sightings in 1946, 47, 48, 50, 55, 56, 59, 60, 61, 62, 63, 64, 68, 70, 72, 73, 74, 76, 77, 80, 81, 82, 83, and 1985. A total of $n = 24$ sightings were made during this period with sighting times equal to $t_i = 1, 2, 3, 5, 10, 11, 14, 15, 16, 17, 18, 19, 23, 25, 27, 28, 29, 31, 32, 35, 36, 37, 38, 40$ years since the initiation of sighting in 1945. Figure 2 illustrates a uniform probability plot for the Eskimo Curlew sighting times, which is simply a plot of the ordered sighting times t_i versus their cumulative (nonexceedance) probability which is computed using the unbiased plotting position $P_i = i/(n + 1)$ for the uniform distribution, where i is the rank of each observation (see Section 6.8 for further information on uniform probability plots). See Section 6.8 for further information on construction of a uniform probability plot. The MVUE for the upper bound of the uniform distribution given in (4) is $\hat{T}_E = t_n((n + 1)/n) = 40(24 + 1)/24 = 41.66$ which is used to plot the fitted uniform distribution as a solid line in Fig. 2 using the quantile function in (3).

The probability plot correlation coefficient (PPCC) measures the linearity of the plot under an assumed distribution and provides a quantitative measure for comparing the relative goodness of fit of a fitted distribution. The PPCC, termed r, for the Eskimo Curlew sighting times is simply the ordinary product moment correlation coefficient between the sighting times t_i and their respective cumulative probabilities P_i. As described in the
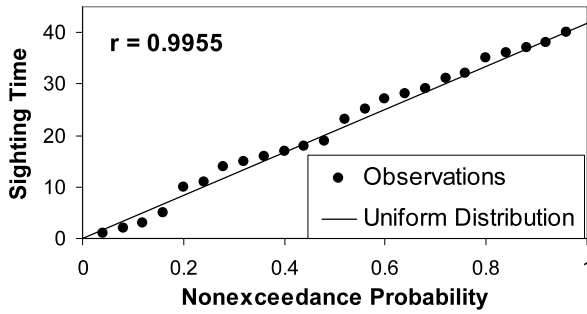
Fig. 2 A uniform probability plot of the sighting times for the Eskimo Curlew (*Numenius borealis*) illustrating the relationship between the ordered sighting times (in years since 1945) versus an estimate of their cumulative nonexceedance probability using the Weibull plotting position. The solid line illustrates the fitted uniform distribution using the MVUE estimator in (4).

Table 1 Comparison of sample size weighted average value of the probability plot correlation coefficients (PPCC) for the four distributions considered

Distribution	$\bar{r}$
Uniform (UN)	0.9706
Truncated Exponential (TEX)	0.9648
Generalized Pareto (GP)	0.9580
Weibull	0.9383

next section, the PPCC statistic also can serve as a powerful hypothesis test statistic. In Fig. 3, we use the PPCC to compare the relative goodness of fit between each of the three theoretical models TEX, GPA, and Weibull and the simpler one-parameter uniform (UN) probability distribution. From Fig. 3, one observes that the TEX and GPA models yield values of PPCC that are very similar to the UN model, whereas the PPCC values for the Weibull distribution are generally lower than those for the UN model.

Comparisons of the PPCC values in Fig. 3 can be misleading, because among the $j = 1, \ldots m$ ($m = 27$) samples, each sample j has a different sample size n_j so that the value of the PPCC for each sighting time series j, denoted $r_j = \text{PPCC}_j$ is based on a different sample size. A more meaningful comparison would correct for the differences among sample sizes. For example, the sample size weighted average value of the PPCC is given by

$$\bar{r} = \frac{\sum_{j=1}^{m} n_j r_j}{\sum_{j=1}^{m} n_j}. \tag{13}$$

Table 1 compares the sample size weighted estimates of the PPCC computed from (13) for the four distributions considered and shows that the Uniform distribution yields the best overall goodness of fit although the TEX distribution, fit using the method of L-moments, yields very similar results.
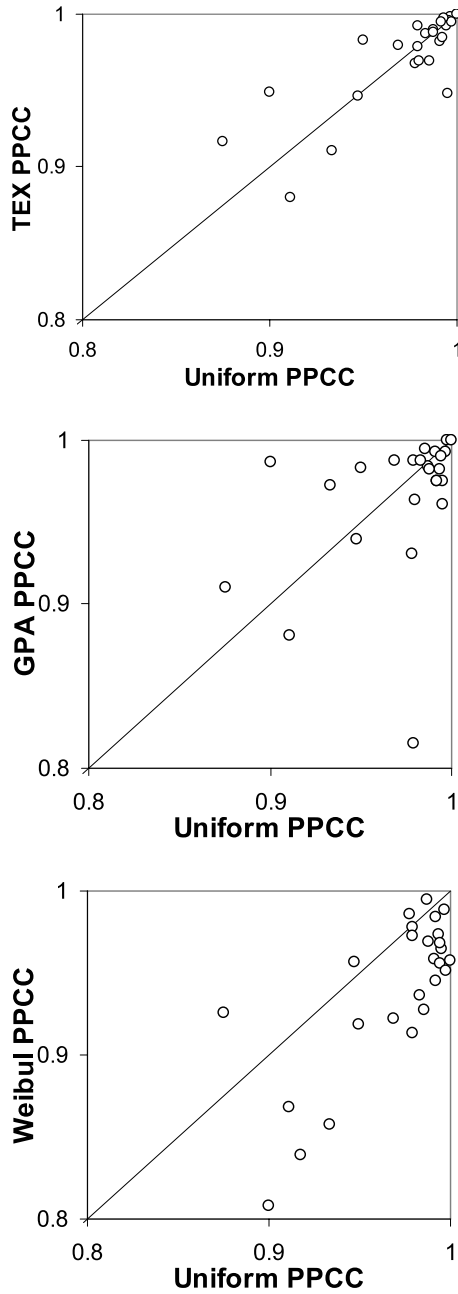
Fig. 3 Comparison of probability plot correlation coefficients (PPCC) for the truncated exponential (TEX), generalized Pareto (GP), and Weibull distributions versus the uniform model.

6.3. Hypothesis tests for distributional fit

A variety of quantitative hypothesis tests are available for evaluating whether or not a particular distributional hypothesis is consistent with observed data. Given the small sample sizes associated with typical sighting records, hypothesis tests based on individual sighting records can have very low statistical power. To address this issue, we apply distributional hypothesis tests repeatedly to many individual samples, thus increasing the statistical power of the resulting suite of tests.

PPCC hypothesis tests are attractive because they combine a commonly used graphical tool (probability plots) with a quantitative goodness of fit statistic (correlation coefficient), and they exhibit power that is comparable with other recommended hypothesis tests. The PPCC test was first introduced by Filliben (1975) for the normal hypothesis and is now widely used (e.g., Stedinger et al., 1993; Kottegoda and Rosso, 1997; Beirlant et al., 2005; Heo et al., 2008). While numerous hypothesis tests are available for testing alternative distributional hypotheses, the PPCC test is more powerful than many alternatives (Chowdhury et al., 1991; Stedinger et al., 1993; Heo et al., 2008), and compares favorably with seven other tests of normality on the basis of empirical power studies (Filliben, 1975).

Of the distributional hypotheses considered in Fig. 1, PPCC distributional hypothesis tests are only available for the Weibull and Uniform distributions (Vogel and Kroll, 1989). Since our previous results reported above favor the Uniform distribution, we only consider that hypothesis. A PPCC test for evaluating the fit of the uniform distribution to a set of ordered data values $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(m)}$ is a measure of the linearity of a plot of the $x_{(i)}$ values against their expected values $p_i = i/(m+1)$ and is computed using

$$r = \frac{\sum_{i=1}^{m}(x_{(i)} - \bar{x})(p_i - \bar{p}_i)}{\sqrt{\sum_{i=1}^{m}(x_{(i)} - \bar{x})^2 \sum_{i=1}^{m}(p_i - \bar{p}_i)^2}}, \tag{14}$$

where $\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_{(i)}$ and $\bar{p}_i = \frac{1}{m}\sum_{i=1}^{m} \frac{i}{m+1} = \frac{1}{2}$.

Lower values of r indicate greater evidence against the uniform null hypothesis. Critical values of the test are reported by Vogel and Kroll (1989, their Table 1) for sample sizes in the range [10, 1,000] and for significance levels in the range [0.01, 0.99].

For the Eskimo Curlew sighting record reported in Fig. 2, a PPCC test yields $r = 0.996$ and Vogel and Kroll (1989 show that in 95% of samples of length $n = 24$, the values of r would exceed 0.96, hence there is little evidence in support of the rejection of the uniform hypothesis for that sighting record. Given the short length of record ($n = 24$), such a hypothesis test lacks statistical power, hence we emphasize that in spite of the fact that there is little evidence to support rejection of the uniform hypothesis, we cannot with any certainty, accept the uniform hypothesis, because it is entirely possible that the data arise from another distribution. To improve our ability to discern whether or not we can reject the uniform hypothesis for sighting times, we now examine the results of 27 independent hypothesis tests.

The values of the r_j of the PPCC and their corresponding p-values p_j for testing the uniform hypothesis are summarized in Table 2 for each of the 27 sighting records. The p-values represent the probability that the PPCC statistic would take a value as small as the one actually observed, if the null hypothesis that the bird sighting times are uniformly

Table 2 Values of sample i, of number of records n, probability plot correlation coefficient (PPCC) r_j, and p-values p_j arising from the PPCC test of the hypothesis that the sighting times follow a uniform distribution. Species that deviate significantly from the uniform distribution using, a 5% level test, are highlighted in bold

i	Species	n	r_j	p_j
1	Labrador Duck *Camptorhynchus labradorius*	6	0.984	0.921
2	Heath Hen *Tympanuchus c. cupido*	8	0.994	0.999
3	Laysan Rail *Porzana palmeri*	17	0.986	0.651
4	Hawaiian Rail *P. sandwichensis*	8	0.917	0.080
5	Eskimo Curlew *Numenius borealis*	24	0.996	0.905
6	Great Auk *Pinguinus impennis*	15	0.979	0.432
7	Passenger Pigeon *Ectopistes migratorius*	28	1.000	1.000
8	Carolina Parakeet *Conuropsis carolinensis*	9	0.991	0.967
9	Ivory-billed Woodpecker *Campephilus p. principalis*	13	0.992	0.934
10	Kauai Oo *Moho braccatus*	19	0.993	0.893
11	Hawaii Oo *M. Nobilis*	7	0.988	0.951
12	Laysan Millerbird *Acrocephalus familiaris*	11	0.98	0.683
13	Kamao *Myadestes myadestinus*	19	0.997	0.996
14	Dusky Seaside Sparrow *Ammodramus maritimus nigrescens*	10	0.996	0.999
15	Ou *Psittirostra psittacea* (Kauai)	18	0.969	0.154
16	Ou *P. psittacea* (Hawaii)	11	0.988	0.863
17	Ou *P. psittacea* (Lana'i)	7	0.947	0.287
18	Greater Koa-finch *Rhodacanthis palmeri*	6	0.95	0.383
19	Lesser Akialoa *Hemignathus obscurus*	11	0.995	0.999
20	**Greater Akialoa *H. ellisianus***	**11**	**0.900**	**0.017**
21	**O'ahu Alauahio *Paroreomyza maculata***	**9**	**0.749**	**0.001**
22	**Maui Alauahio *P. montana* (Lana'i)**	**9**	**0.875**	**0.015**
23	Akepa *Loxops coccineus ochraceus*	6	0.911	0.128
24	Kakawahie *Paroreomyza flammea*	10	0.933	0.086
25	Hawaii Mamo *Drepanis pacifica*	8	0.979	0.787
26	Laysan Honeycreeper [Apapane] *Himateone sanguinea freethii*	12	0.978	0.581
27	Po'ouli *Melamprosops phaeosoma*	26	0.992	0.745

distributed were true. Using a test with significance level 5% (Type I error probability = 0.05) for each sample (species/populations), we reject the null hypothesis that the sighting times are uniformly distributed for 3 of the 27 bird populations (highlighted in bold in Table 2). When one applies a hypothesis test 27 times using a 5% level significance level for each sample, one expects $0.05(27) = 1.35$ rejections and we obtained 3 rejections, thus there is little evidence to support a rejection of the uniform hypothesis. Using a 5% level test for each sample is potentially misleading, as described in the section below, and so we also performed the more rigorous "field significance tests" to evaluate the behavior of the 27 independent hypothesis test results.

6.4. Field significance tests for the uniform distribution

In Table 2, we summarize the results of 27 individual uniform PPCC hypothesis tests. Such a set of multiple hypothesis tests may be difficult to interpret because, even if the null hypothesis is in fact true, after performing enough tests one will inevitably encounter a test result that supports rejection of the null hypothesis. One must beware of attaching too much importance to an individual test result among a group of multiple tests, because it may have resulted by chance alone, and since they are based on such small samples,

each individual hypothesis test contains very little information. To account for this phenomenon numerous multiple comparison procedures (MCP) have been advanced in the statistics literature. The most widely used approaches are based on the Bonferroni equality (Simes, 1986) or modifications of that test (e.g., Rice, 1989; Benjamini and Hochberg, 1995). Other MCPs used in the fields of climate and hydrology include the field significance tests introduced by Livezey and Chen (1983), Vogel and Kroll (1989), and Douglas et al. (2000). Field significance is the collective significance of a group of hypothesis tests. Analogously to the work of Ventura et al. (2004), in the following sections, we use the ideas of Livezey and Chen (1983), Simes (1986), Vogel and Kroll (1989), and Benjamini and Hochberg (1995) to evaluate the overall field significance associated with the group of individual tests reported in Table 2.

6.5. Bonferroni-type field significance test

Suppose that we want to test the overall hypothesis that each of m individual hypotheses is true, and that we would like the probability of rejection of the overall hypothesis, if it is in fact true, to be a specified value α_F, the field significance level. Suppose further that the m individual hypothesis tests are statistically independent, and that each has type I error probability α. A Bonferroni-type test (Simes, 1986) would relate the field significance level α_F associated with the overall test to the significance level α of each individual test using

$$1 - \alpha_F = (1 - \alpha)^m. \tag{15}$$

When α is small, as is usually the case, $(1 - \alpha)^m \cong 1 - m\alpha$, which yields

$$\alpha = \alpha_F/m. \tag{16}$$

This procedure thus determines the significance level that should be used for the individual tests in order to achieve a specified field significance level for the overall test.

Using a field significance level of 5% ($\alpha_F = 0.05$) for $m = 27$ tests in Table 2, Eq. (16) leads to an individual hypothesis test significance level of $\alpha = 0.05/27 = 0.00185$. In Table 2, we observe only a single bird sighting record with a p-value less than 0.00185; hence there is little evidence to reject the uniform null hypothesis using a Bonferroni-type test.

6.6. False discovery rate procedure

Benjamini and Hochberg (1995) introduced an improvement over the Bonferroni-type test which attempts to control for what they term the false discovery rate (FDR), which is the number of false rejections of the null hypothesis, H_0 (see also 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 H_0 for all samples $i = 1, \ldots, m$, if any of the p-values of the individual tests is less than α in (16); so that the same threshold α is applied to every bird population. In contrast, the rejection threshold for the FDR procedure of Benjamini and Hochberg (1995) differs as follows. Consider the $i = 1, \ldots, m$ values of the p-values in Table 2

ranked in ascending order which we denote by $p_{(i)}$. The FDR procedure rejects the null hypothesis for tests $1, \ldots, k$, where k is the largest value of i for which

$$p_{(i)} \leq \frac{i}{m} \alpha_F. \tag{17}$$

After ranking the p-values in Table 2, we find that the largest value of i for which (17) holds, when $\alpha_F = 0.05$, is 1; hence the null hypothesis can only be rejected for one of the $m = 27$ populations.

6.7. Binomial field significance test

Of the samples tested using a 5% level uniform PPCC test, Table 2 revealed that three sighting records were rejected whereas on average one expects $0.05(27) = 1.35$ rejections. If each hypothesis test is considered to be independent of the others, the individual tests are independent Bernoulli trials each having probability $\alpha = 0.05$ of rejecting the uniform null hypothesis if it is in fact true. The probability distribution of the number of rejections, X, in a series of m independent tests will follow a binomial distribution with parameters m and α. If the criterion for rejecting the overall null hypothesis is that at least x of the individual tests are significant at level α, then the field significance level α_F is given by

$$\alpha_F = P[X \geq x] = 1 - P[X < x] = 1 - \sum_{k=0}^{x-1} \frac{m!}{k!(m-k)!} \alpha^k (1-\alpha)^{m-k}. \tag{18}$$

Figure 4 illustrates the field significance level α_F as a function of the number of rejections using (18), for $\alpha = 0.01, 0.05$ and 0.1 assuming $m = 27$. Table 2 indicates that the number of rejections corresponding to hypothesis tests with significance levels of $\alpha = 0.01, 0.05$ and 0.1 are 1, 3, and 5, respectively. Using Fig. 4, these results correspond to field significance levels of 0.24, 0.15, and 0.13 based on the individual hypothesis test significance levels of $\alpha = 0.01, 0.05$, and 0.1, respectively.

Using this binomial field significance test, the smallest field significance level at which the null hypothesis would be rejected ranges between 0.13 and 0.24. This provides little evidence against the overall null hypothesis that sighting times are uniformly distributed for all of the bird populations in Table 2. Below we consider a more general field significance test, which considers the behavior of all of the p-values associated with all of the individual hypothesis tests.

6.8. Uniform PPCC field significance test

The previous tests only considered the results of m individual hypothesis tests each with fixed significance level. Of interest is whether the widely varying p-values of the 27 tests reported in Table 2 behave in the way that would be expected under the uniform null hypothesis of bird sightings. Again, consider a collection of m hypothesis tests that yield p-values $p_i, i = 1, \ldots, m$. If each test is independent, the p-values arising from the individual hypothesis tests should constitute a random sample from a uniform distribution over the interval [0, 1] (Casella and Berger, 1990; Springer, 1990). Note that this can be
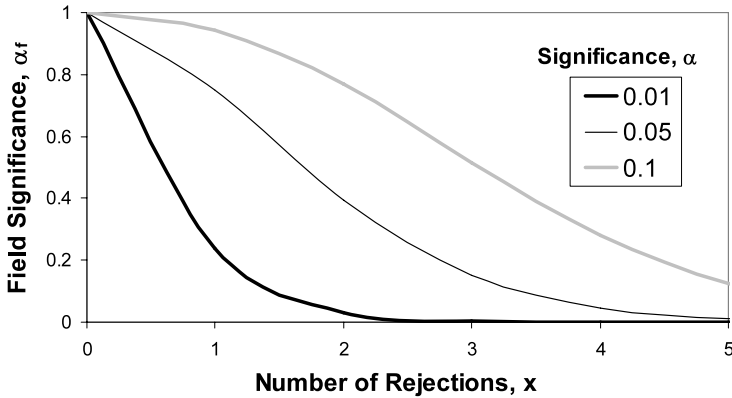
Fig. 4 Relationship between field significance and number of rejections for various assumed significance levels α, of individual tests.

confusing because our null hypothesis is that the bird sightings follow a uniform distribution, which implies that the collection of p-values associated with the multiple tests of that hypothesis also follow a uniform distribution. Thus, another field significance test would evaluate the null hypothesis that the p-values reported in Table 2 follow a uniform distribution.

Analogous to Springer (1990) who used a Kolmorogov–Smirnov (KS) test to evaluate whether independent significance levels from a suite of hypothesis tests follow a uniform distribution we use the PPCC test of uniformity derived by Vogel and Kroll (1989). We expect from the results of Heo et al. (2008) that the PPCC test of uniformity should have statistical power at least as high as the KS test of uniformity. The PPCC test of uniformity described in Section 6.3, is based on a uniform probability plot and is constructed as follows. If the m p-values p_i are considered independent and are ranked in ascending order, which we denote $p_{(i)}$, then the ranked values follow a beta distribution (David and Nagaraja, 2003) with expectation equal to

$$E[p_{(i)}] = u_i = \frac{i}{m+1}. \tag{19}$$

Note that (19) is often termed the Weibull plotting position.

A uniform probability plot of the p-values associated with each of the $m = 27$ hypothesis tests is illustrated in Fig. 5. It was obtained by plotting the ordered p-values $p_{(i)}$ versus the Weibull plotting position u_i in (19) based on their ranks i, with $m = 27$. To compare the fitted uniform distribution to its theoretical expectation, we show a line passing through the origin with a slope of unity. If the bird sighting data follow a uniform distribution, then the p-values associated with each individual hypothesis test should also be uniformly distributed, and the open circles in Fig. 5 should lie close to a straight line. The more nearly linear the probability plot in Fig. 5 is, the less evidence is there for rejecting the null hypothesis.

For the 27 p-values in Table 2, we obtain $r = 0.9597$ from (14). From Table 1 in Vogel and Kroll (1989), the critical value of the uniform PPCC test statistic at a 2.2% field significance level is $r_{0.05} = 0.9597$. Hence, the p-value associated with this test is only 2.2%;
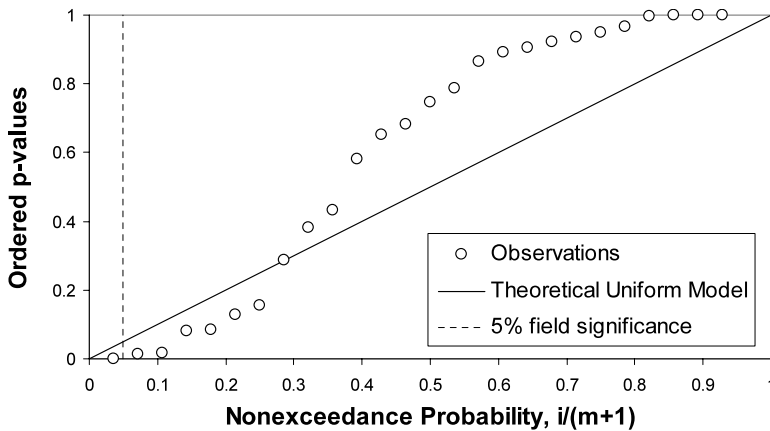
Fig. 5 Uniform probability plot of estimated *p*-values associated with the *m* = 27 individual PPCC hypothesis tests reported in Table 2.

this is also the smallest significance level at which the null hypothesis would be rejected. This low *p*-value occurs not because the uniform model for sighting times yields a poor fit, but because it often yields a fit that is better than could be expected under the null hypothesis. This low overall *p*-value seems to occur because many of the individual *p*-values in Table 2 are close to 1, so the fit of a uniform distribution to the sightings data is in many cases "too good to be true." A typical reason why this occurs is not that the underlying distribution is not uniform, but that it is not independent and identically distributed. Possible reasons why this could occur for our example data include: (1) the sighting records report only whether a sighting occurred in a given time period, rather than the number of sightings in each time period; (2) the intensity of observations, and hence the probability of making a sighting given that the species is not extinct, is not uniform but changes after the species has not been sighted for some length of time; or (3) the sighting records are correlated across species because surveys are done for multiple species simultaneously. These issues highlight the need for careful consideration of the assumptions underlying these methods, and the value of conducting simulations to experimentally test how robust the methods are to assumption violations. For example, these statistical methods may not be appropriate for evaluating eradiation programs as the sightings (e.g., trap captures) are not independent of the decline (see Solow et al., 2008 for an alternative method). While others (Burgman et al., 1995; McCarthy, 1998) have attempted to resolve questions over sighting effort, more work is needed in this area.

7. Conclusions

L-moment diagrams and probability plot correlation coefficient hypothesis tests were performed to examine the goodness of fit of various probability distribution functions (pdfs) that have previously been used to model the likelihood of bird sighting data. Our goodness of fit analyses reveal that among the four hypothetical pdfs, the uniform, TEX, and GP models all perform well, whereas the Weibull model performs poorly. Of the three

pdfs that perform well, the uniform distribution appears to perform best based on PPCC goodness of fit comparisons.

Hypothesis tests were then performed for the uniform null hypothesis of bird sighting times, to evaluate whether or not the results of $m = 27$ independent hypothesis tests would be expected to behave the way that our tests performed when the null hypothesis is correct. Using traditional Bonferroni-type hypothesis tests and sequential Bonferroni-type tests that control for the number of false rejections of the null hypothesis, our results reveal little evidence to support the rejection of the uniform null hypothesis. Still, we are unable to accept the uniform null hypothesis because there is always the possibility that the observations arise from another distribution. Further analyses using field significance tests, which attempt to evaluate the overall behavior of all $m = 27$ hypothesis tests as an ensemble, show that although the uniform null hypothesis was the most strongly supported, the fit to this distribution is better than one would expect. A plausible reason for this result is that the bird sighting records may not be independent and identically distributed. Future work is needed to develop powerful hypothesis tests for the TEX and GPA hypotheses so that those alternatives may be more fully evaluated. The methods we present here provide a framework for evaluating future models.

Acknowledgements

Appendix A: Theory of L-moments for the truncated exponential distribution

Consider the truncated exponential probability density function given in (5) with scale parameter β, truncated at T_E.

L-moments: Let $\eta = e^{-\beta T_E}$. Then the first four L-moments are given by

$$\lambda_1 = \frac{1}{\beta} - \frac{T_E \eta}{1 - \eta}, \tag{A.1}$$

$$\lambda_2 = \frac{1}{1 - \eta}\left[\frac{1 + \eta}{2\beta} - \frac{T_E \eta}{1 - \eta}\right], \tag{A.2}$$

$$\lambda_3 = \frac{1}{(1 - \eta)^2}\left[\frac{1 + 10\eta + \eta^2}{6\beta} - \frac{T_E \eta(1 + \eta)}{1 - \eta}\right], \tag{A.3}$$

$$\lambda_4 = \frac{1}{(1 - \eta)^3}\left[\frac{1 + 29\eta + 29\eta^2 + \eta^3}{12\beta} - \frac{T_E \eta(1 + 3\eta + \eta^2)}{1 - \eta}\right]. \tag{A.4}$$

Parameter estimation based on the first two L-moments: Rewriting (A.1) and (A.2) as

$$\lambda_1 = \frac{1 - \eta + \eta \ln(\eta)}{\beta(1 - \eta)}, \tag{A.5}$$

$$\lambda_2 = \frac{1 + 2\eta \ln(\eta) - \eta^2}{2\beta(1-\eta)^2}. \tag{A.6}$$

The L-CV is then given by

$$\tau = \frac{\lambda_2}{\lambda_1} = \frac{1 + 2\eta \log \eta - \eta^2}{2(1-\eta)(1-\eta+\eta\log\eta)}, \tag{A.7}$$

and is a monotonic function of η decreasing from $\tau = \frac{1}{2}$ at $\eta = 0$ to $\tau = \frac{1}{3}$ at $\eta = 1$.

So, given a τ value in the range $\frac{1}{3}$ to $\frac{1}{2}$, we can solve for η, e.g., by the bisection method with starting interval $(0, 1)$. Then L-moment estimates of the parameters are given by

$$\beta = \frac{1 - \eta + \eta \ln(\eta)}{(1-\eta)\lambda_1} \quad \text{and} \quad T_E = \frac{-\ln(\eta)}{\beta}. \tag{A.8}$$

References

Beg, M.A., 1982. Optimal tests and estimators for truncated exponential families. Metrika 29, 103–113.

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

Beirlant, J., Goegebeur, Y., Segars, J., Teugels, J., 2005. Statistics of Extremes: Theory and Applications. Wiley, Hoboken.

Burgman, M.A., Grimson, R.C., Ferson, S., 1995. Inferring threat from scientific collections. Conserv. Biol. 9, 923–928.

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

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

Collar, N.J., 1998. Extinction by assumption; or the Romeo error on Cebu. Oryx 32, 239–244.

David, H.A., Nagaraja, H.N., 2003. Order Statistics, 3rd edn. Wiley, Hoboken.

Douglas, E.M., Vogel, R.M., 2006. The probabilistic behavior of the flood of record in the United States. J. Hydrol. Eng. 11, 482–488.

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

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

Heo, J.-H., Kho, Y.W., Shin, H., Kim, S., Kim, T., 2008. Regression equations of probability plot correlation coefficient test statistics from several probability distributions. J. Hydrol. 335(1–4), 1–15.

Hosking, J.R.M., 1990. L-moments: Analysis and estimation of distributions using linear combinations or order statistics. J. R. Stat. Soc. B Met. 52, 105–124.

Hosking, J.R.M., Wallis, J.R., 1987. Parameter and quantile estimation for the generalized Pareto distribution. Technometrics 9, 339–349.

Hosking, J.R.M., Wallis, J.R., 1997. Regional Frequency Analysis: An Approach Based on L-moments. Cambridge Univ. Press, New York.

Kottegoda, N.T., Rosso, R., 1997. Statistics, Probability, and Environmental Engineers. McGraw–Hill, New York.

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

Marshall, C.R., 1995. Distinguishing between sudden and gradual extinctions in the fossil record: Predicting the position of the Cretaceous-Tertiary iridium anomaly using the ammonite fossil record on Seymour Island, Antarctica. Geology. 23, 731–734.

Marshall, C.R., Ward, P.D., 1996. Sudden and gradual molluscan extinctions in the latest Cretaceous of western European Tethys. Science 274, 1360–1363

McCarthy, M.A., 1998. Identifying declining and threatened species with museum data. Biol. Conserv. 83, 9–17.

Reed, J.M., 1996. Using statistical probability to increase confidence of inferring species extinction. Conserv. Biol. 10, 1283–1285.

Rice, W.R., 1989. Analyzing tables of statistical tests. Evolution 43, 223–225.

Roberts, D.L., Solow, A.R., 2003. When did the Dodo become extinct? Nature 426, 245.

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

Solow, A.R., 1993a. Inferring extinction from sighting data. Ecology 74, 962–964.

Solow, A.R., 1993b. Inferring extinction in a declining population. J. Math. Biol. 32, 79–82.

Solow, A.R., 2005. Inferring extinction from a sighting record. Math. Biosci. 195, 47–55.

Solow, A.R., Roberts, D.L., Robbirt, K.M., 2006. On the Pleistocene extinctions of Alaskan mammoths and horses. Proc. Natl. Acad. Sci. USA 103, 7351–7353.

Solow, A., Seymour, A., Beet, A., Harris, S., 2008. The untamed shrew: On the termination of an eradication programme for an introduced species. J. Appl. Ecol. 45, 424–427.

Springer, M.S., 1990. The effect of random range truncations on patterns of evolution in the fossil record. Paleobiology 16, 512–520.

Stedinger, J.R., Vogel, R.M., Foufoula-Georgiou, E., 1993. Frequency analysis of extreme events. In: Maidment, D.R. (Ed.), Handbook of Hydrology, pp. 18.1–18.68. McGraw–Hill, New York.

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

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

Vogel, R.M., Fennessey, N.M., 1993. L-moment diagrams should replace product-moment diagrams. Water Resour. Res. 29, 1745–1752.

Vogel, R.M., Kroll, C.N., 1989. Low-flow frequency analysis using probability plot correlation coefficients. J. Water Resour. Plan. Manag. 115, 338–357.

Wang, S.C., Marshall, C.R., 2004. Improved confidence intervals for estimating the position of a mass extinction boundary. Paleobiology 30, 5–18.