Estimating trends in data from the Weibull and a generalized extreme value distribution

Robin T. Clarke

Instituto de Pesquisas Hidráulicas, Porto Alegre, Brazil

Received 3 April 2001; revised 17 December 2001; accepted 17 December 2001; published 26 June 2002.

[1] Where changes in hydrologic regime occur, whether as a result of change in land use or climate, statistical procedures are needed to test for the existence of trend in hydrological data, particularly those expected to follow extreme value distributions such as annual peak discharges, annual minimum flows, and annual maximum rainfall intensities. Furthermore, where trend is detected, its magnitude must also be estimated. A later paper [*Clarke*, 2002] will consider the estimation of trends in Gumbel data; the present paper gives results on tests for the significance of trends in annual and minimum discharges, where these can be assumed to follow a Weibull distribution. The statistical procedures, already fully established in the statistical analysis of survival data, convert the problem into one in which a generalized linear model is fitted to a power-transformed variable having Poisson distribution and calculates the trend coefficients (as well as the parameter in the power transform) by maximum likelihood. The methods are used to test for trend in annual minimum flows over a 19-year period in the River Paraguay at Cáceres, Brazil, and in monthly flows at the same site. Extension of the procedure to testing for trend in data following a generalized extreme value distribution is also discussed. Although a test for time trend in Weibull-distributed hydrologic data is the motivation for this paper, the same approach can be applied in the analysis of data sequences that can be regarded as stationary in time, for which the objective is to explore relationships between a Weibull variate and other variables (covariates) that explain its behavior. *INDEX TERMS:* 1860 Hydrology: Runoff and streamflow; 1803 Hydrology: Anthropogenic effects; 1812 Hydrology: Drought; *KEYWORDS:* trends, Weibull, GEV, GLM

1. Introduction

[2] The *Intergovernmental Panel for Climate Change* (*IPCC*) [2001] summarized its conclusions concerning the likely increases in globally averaged surface temperatures (1.4°–5.8°C by 2100, relative to 1990) and in globally averaged sea level (0.09–0.88 m by 2100). They conclude that warming will vary regionally and be accompanied by increases and decreases in precipitation. "In addition," says *IPCC* [2001], "there would be changes in the variability of climate, and changes in the frequency and intensity of some climate phenomena." Such forecasts, now being made with ever-increasing confidence, imply that the statistical stationarity necessary for many hydrologic analyses can no longer be safely assumed, and the spatial and temporal availability of water resources must be expected to change as and when regional climate changes. This is of crucial importance to a developing country like Brazil, in which >90% of its energy is provided by hydropower.

[3] This paper is concerned with some aspects of the detection and estimation of "changes in the frequency and intensity of some climate phenomena" referred to in the above quote from *IPCC* [2001]. An later paper [*Clarke*, 2002] deals with the detection and estimation of trends in annual maximum river discharges following a Gumbel distribution, although the methods that it presents for detecting and estimating trends over time are equally applicable to other Gumbel-distributed hydrologic and climate variables, such as annual maximum rainfall intensities for various durations, and wind velocities. This later paper also suggests that there is a logical inconsistency involved in estimating floods with *T*-year return period by fitting a Gumbel distribution when time trends are absent, while abandoning the Gumbel hypothesis when time trends may be present in data. Where it is believed that time trends may be present, two common procedures [*Salas*, 1993] are to base significance tests on (1) linear regression, which assumes that data follow a normal distribution with time-varying mean or (2) nonparametric methods, such as Mann-Kendall, either procedure having very little to do with extreme value theory by which use of the Gumbel is often justified. There seems to be an inconsistency in the logic by which, when time trends in hydrologic data are believed to be absent, data are modeled by an appropriate extreme value distribution and inferences drawn about the frequency of occurrence of extremes, while if there is a possibility that a time trend exists in the data, the working hypothesis of extreme value distribution is immediately abandoned in favor of the normal distribution (in the case of linear regression) or no distribution at all (in the case of Mann-Kendall). To remove this inconsistency, the approach explored by *Clarke* [2002] is to allow the location parameter in the Gumbel distribution to be a function of time, such as a polynomial, whose coefficients could be tested for

0043-1397/02/2001WR000575$09.00

significance; if none of the polynomial parameters except the constant term are statistically significant, the procedure reduces to fitting a Gumbel distribution to a stationary data sequence by maximum likelihood (ML). If one or more of the polynomial coefficients is found to be statistically significant, the data are taken to follow a Gumbel distribution with time-varying mean. While the emphasis of *Clarke* [2002] is the detection and estimation of trends over time, the method would be equally applicable where time trends are shown to be absent for exploring relationships between Gumbel-distributed data and other variables that might explain the behavior of the variable of primary interest. For example, a sequence of annual maximum rainfall intensities for duration D minutes might be related to velocity and direction of wind, even where the intensities themselves show no evidence of changing over the course of years.

[4] For the Gumbel distribution of *Clarke* [2002] an efficient ML estimation procedure was found by (1) transforming to a new variable, related to the variable of interest by a transformation involving a single unknown parameter, and (2) using iteratively weighted least squares (IWLS) [*McCullagh and Nelder*, 1989, pp. 40–43] to obtain ML estimates of the trend coefficients and of the transformation parameter. IWLS is a standard computational procedure which effectively converts ML estimation into a weighted least squares calculation in which both the dependent variable and the weights given to it depend on the fitted values, for which only current estimates are available. The problem was formulated in terms of a generalized linear model (GLM), for which departure from a hypothetical model structure was measured by the deviance statistic. The present paper extends the approach given by *Clarke* [2002] to the detection and estimation of time trends in data following a Weibull distribution. If W_i are the minimum stream flows in different days of the year, then the annual minimum is the smallest of the W_i, each of which is bounded by zero. In this case the random variable $X_i = \min(W_i)$ may be well described by the Weibull (or extreme value type 3) distribution [*Stedinger et al.*, 1993] given in their notation as

$$f_Y(y) = (k/\sigma)(y/\sigma)^{k-1}\exp\left[-(y/\sigma)^k\right] \quad y > 0,\ \sigma,\ k > 0. \tag{1}$$

The procedure is illustrated using records of annual minimum discharge (or, more exactly, annual minimum mean daily flow, estimated from a calibration or "stage-discharge" curve by twice-daily stage readings) from the gauging station at Cáceres on the River Paraguay; this river has shown long-term changes in flow which have been related [*Collischonn et al.*, 2001] to gradual changes in regional rainfall.

[5] The procedure described in section 2 is not original (although its use for detecting time trends in hydrologic data is believed to be) since it was developed over 20 years ago for the statistical analysis of survival data [*Aitkin and Clayton*, 1980; *Aitkin et al.*, 1989]. The notation used by these authors also differs from that in equation (1), and it is simpler here to retain their notation in which the Weibull is written as

$$f_Y(y) = \alpha\lambda y^{\alpha-1}\exp[-\lambda y^{\alpha}], \tag{2}$$

with cumulative distribution

$$F_Y(y) = 1 - \exp(-\lambda y^{\alpha}), \tag{3}$$

so that the parameters α and λ in equation (2) are related to σ and k in equation (1) by $\alpha = k$ and $\lambda = 1/\sigma^k$. Using the notation of equation (2), the mean and variance of the Weibull distribution are $\Gamma(1 + \alpha^{-1})/\lambda^{1/\alpha}$ and $[\Gamma(1 + 2\alpha^{-1}) - \Gamma(1 + \alpha^{-1})^2]/\lambda^{2/\alpha}$, respectively.

2. ML Estimation of Weibull Parameters in the Presence of Time Trends

[6] We assume a sequence of hydrologic data, such as annual minimum flows in each of N years, denoted by $\{y_1, y_2 \ldots y_t \ldots y_N\}$, for which the possibility of time trend must be explored. A trend may be described by $\boldsymbol{\beta}^{\mathbf{T}}\mathbf{F}$, where $\boldsymbol{\beta}^{\mathbf{T}}$ and $\mathbf{F}$ are the vectors $[\beta_0\beta_1 \ldots \beta_{p-1}]$ and $[f_0(t)\, f_1(t) \ldots f_{p-1}(t)]^T$, with the $f_i(\,)$ functions of time; if a polynomial trend is used, $f_i(t) = t^i$. (The parameter $f_i(t)$, here used as known function of time, should be distinguished from the parameter $f_Y(y)$, used above, for a probability density function). We modify the Weibull distribution equation (2) to allow its mean and, in particular, the parameter λ, which enters both the mean and variance given above, to vary in time according to this functional form, so that for the ith data value y_i the parameter now has a suffix and is written λ_t to indicate its time dependence. Specifically, we write

$$\lambda_t = \exp\left(\boldsymbol{\beta}^{\mathbf{T}}\mathbf{F}\right),$$

so that the Weibull becomes

$$f_Y(y) = \alpha y^{\alpha-1}\exp\left[\boldsymbol{\beta}^{\mathbf{T}}\mathbf{F} - y^{\alpha}\exp\left(\boldsymbol{\beta}^{\mathbf{T}}\mathbf{F}\right)\right] \quad y \geq 0,\ \alpha > 0 \tag{4}$$

with mean $E[y] = \Gamma(1 + \alpha^{-1})\exp(-\boldsymbol{\beta}^{\mathbf{T}}\mathbf{F})$ having index parameter α [*Cox and Oakes*, 1984, p. 19] and trend parameters $\boldsymbol{\beta}$ are to be estimated by ML.

[7] Following *Aitkin et al.* [1989], we define $h(y) = f_Y(y)/[1 - F_Y(y)]$, $S(y) = 1 - F_Y(y)$, and $H(y) = \int_0^y h(u)du$. Thus $h(y) = f(y)/S(y)$, $S(y) = \exp[-H(y)]$, and $f_Y(y) = h(y)\exp[-H(y)]$. Because of its basis in survival analysis (in which the functions $h(y)$, $S(y)$, and $H(y)$ are termed the hazard function, survivor function, and integrated hazard function, respectively), Aitkin et al.'s method includes the possibility that data are censored on the right (i.e., it is known only that the value of the random variable Y is greater than some value y^*). Thus a variable w_t ($t = 1, 2 \ldots N$) is introduced, having the value 1 if a data value is uncensored and 0 otherwise. In the context of annual minimum flows in which an uncensored value exists for each year of record, all of the w_t will be 1, but if the record of daily flow in a year is incomplete, the annual minimum flow will be less than or equal to the minimum in the partial record: a case of censoring on the left. Data censored on the left cannot be dealt with by the method being described for fitting the Weibull distribution (although at least one statistical package, S-plus, fits the model given in equation (4) when data are censored on the left (J. R. M. Hosking, personal communication, 2001)). Censoring on the left, as for incomplete years of daily flow, can be dealt with by

Table 1. Annual Minimum Flows During the Period 1966–1984 for the River Paraguay at Cáceres

Year	Q_{min}, $m^3\ s^{-1}$
1	148
2	138
3	155
4	140
5	152
6	145
7	144
8	138
9	189
10	191
11	225
12	237
13	237
14	274
15	281
16	255
17	324
18	280
19	283

another adaptation of *Aitkin et al.* [1989] method but requires a distribution other than the Weibull. However, it would, of course, be possible to include left-censored data when maximizing the Weibull-derived log likelihood function, but the IWLS algorithm would need to be replaced by some other procedure such as the *E-M* algorithm [*Dempster et al.*, 1977].

[8] Now consider the likelihood function of N years of low-flow data, possibly censored, where the elements of the vector $[w_1, w_2, \ldots w_N]$ are equal to 1 for full years of data and 0 for right-censored data. By definition, the likelihood is

$$L = \prod_{t=1}^{N} f_Y(y_t)^{w_t} S(y_i)^{1-w_t},$$

which, since $h(y) = f(y)/S(y)$, may be written formally as

$$L = \prod_{t=1}^{N} h(y_t)^{w_t} S(y_t), \tag{5}$$

recalling that the w_t are either 1 or 0 (and will be all unity for the case of interest, noncensored data).

[9] For the Weibull, $h(y_t) = \alpha\lambda_t y_t^{\alpha-1}$, $S(y_t) = \exp(-\lambda_t y_t^{\alpha})$, and $H(y_t) = \lambda_t y_t^{\alpha} = \theta_t$, say. With the parameter λ_t expressed in terms of the trend $\boldsymbol{\beta}^{\mathbf{T}}\mathbf{F}$, the likelihood function in equation (5) can be written

$$L(\boldsymbol{\beta}, \alpha) = \prod_{t=1}^{N} (\alpha\theta_t/y_t)^{w_t} \exp(-\theta_t) = \alpha^{\sum w_t} \prod_t \theta_t^{w_t} \exp(-\theta_t) / \prod_t y_t^{w_t}. \tag{6}$$

The denominator of the expression on the right-hand side of equation (6) can be ignored because it does not involve the parameters $\boldsymbol{\beta}$ and α, so that when derivatives of $\log_e L$ are taken with respect to these parameters, the denominator disappears. The term

$$\prod_{t=1}^{N} \theta_t^{w_t} \exp(-\theta_t)$$

is the likelihood function of N independent "Poisson variates" w_t with means θ_t, although the full likelihood function contains the additional factor

$$\alpha^{\Sigma w_t} = \alpha^N$$

in the absence of censoring. The log likelihood then becomes

$$\log_e L = N \log_e \alpha + \sum_{t=1}^{N} \{\log_e \theta_t - \theta_t), \tag{7}$$

recalling that $\theta_t = \lambda_t y_t^{\alpha} = y_t^{\alpha} \exp(\boldsymbol{\beta}^{\mathbf{T}}\mathbf{F})$. Taking derivatives with respect to β_j, α gives the equations

$$\partial \log_e L / \partial \beta_j = \sum_t (1 - \theta_t) f_j(t_t) = 0 \tag{8a}$$

$$\partial \log_e L / \partial \alpha = N/\alpha + \sum_t (1 - \theta_t) \log_e y_t = 0, \tag{8b}$$

so that the ML estimate of α satisfies

$$\hat{\alpha} = N \Big/ \left[\sum_t \left(\hat{\theta}_t - 1 \right) \log_e y_t \right]. \tag{9}$$

The parameters α and $\boldsymbol{\beta}$ can be efficiently estimated using the IWLS algorithm, given, for example, by *McCullagh and Nelder* [1989]. The estimation proceeds by casting the problem in the form of a GLM; the work by McCullagh and Nelder is the standard text on these models, which generalize techniques of multiple regression to conditions where (1) data do not have to be normally distributed but may have binomial, Poisson, gamma, or inverse gamma distributions (or, more generally, may have distributions belonging to a general family for which all the data may be summarized in terms of a small number of sufficient statistics) and (2) the expected value μ_t of the random variable giving the tth observation y_t is expressed in terms of explanatory variables $\mathbf{x}$ not necessarily just of the linear form $\mu_t = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x}_t$ but may be more generally $g(\mu_t) = \boldsymbol{\beta}^{\mathbf{T}}\mathbf{x}_t$, where $g(\,)$ is any known nondecreasing function.

[10] If the parameter α were known, the variables w_t, despite all being equal to 1 in this application, could be modeled by a GLM since (1) the w_t are Poisson distributed and (2) $E[w_t] = \theta_t$ is related to the explanatory variables $\mathbf{F}$ by $\log_e \theta_t = \alpha \log_e y_t + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{F}_t$, in which the "known" term

Table 2a. Annual Minimum Flows at Cáceres: Analysis of Deviance and Estimated Parameters From Fitting the Poisson Model for *wi* With Offset ($\log_e yi$, $\boldsymbol{\beta} = [\beta_0\ \beta_1]^T$, and $\mathbf{F} = [1\ t\ ]^T$

	Degrees of Freedom	Deviance	Mean Deviance
Regression	1	101.85	101.852
Residual	17	21.97	1.292
Total	18	123.82	6.879
	Estimate	Standard Error	Ratio
$\hat{\beta}_0$	−54.909	0.442	−124.28
$\hat{\beta}_1$	−0.5579	0.0378	−14.77

Table 2b. Annual Minimum Flows at Cáceres: Analysis of Deviance and Estimated Parameters From Fitting the Poisson Model for *wi* With Offset $\alpha \log_e yi$, $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2\ ]^T$, and $\mathbf{F} = [1t\ t^2]^T$

	Degrees of Freedom	Deviance	Mean Deviance
Regression	2	113.97	56.987
Residual	16	22.92	1.432
Total	18	136.89	7.605
Change	−1	−1.90	1.899

$\alpha \log_e y_t$ is termed an offset. In ordinary multiple regression an offset can be simply absorbed by subtracting it directly from the dependent variable; in the case of GLMs a different but equally straightforward method must be used [*McCullagh and Nelder*, 1989]. However, the approach defined by points 1 and 2 is complicated by the fact that α is not known and must be estimated from equation (9), so that a two-stage iterative procedure is required with the following steps.

1. In step 1, set $\alpha = 1$. Fit the Poisson model to the w_t ($w_t = 1$ for $t = 1, 2,\ldots N$) using $\alpha \log_e y_t$ as an offset. This step therefore yields estimates of the trend parameters $\boldsymbol{\beta}$ in $\log_e\theta_t = \alpha \log_e\ y_t + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{F}_t$.

2. In step 2, with this estimate of the vector β, use equation (9) to reestimate α; use the new estimate of α, return to step 1, and repeat until current estimates of both $\boldsymbol{\beta}$ and α have converged.

[11] *Aitkin et al.* [1989] found that the solution oscillates around the ML solution but that convergence could be accelerated by averaging each two successive estimates of α. Thus if $\alpha^{(k-1)}$ and $\alpha^{(k)}$ are estimates of α at iterations $k-1$ and k, then the smoothed estimate $\alpha^* = (\alpha^{(k-1)} + \alpha^{(k)})/2$ is taken. Once the estimate of α has converged, the "Poisson model" immediately provides estimates of the parameters β; then either by comparing the estimates $\hat{\beta}$ with their standard errors (obtained in the usual way from the matrix of second derivatives at the point where the likelihood function has its maximum) or by using an analysis of deviance [*McCullagh and Nelder*, 1989], the significance of the trend can be tested.

3. Fitting a Time Trend to Annual Minimum Flows on the River Paraguay at Cáceres

[12] We illustrate the procedure using a 19-year record of annual minimum flows recorded at Cáceres, on the River Paraguay during the period 1966–1984. These data are shown in Table 1. Water levels and flows in this river, which flows through the Pantanal, the world's largest wetland, have shown remarkable fluctuations [*Collischonn et al.*, 2001; *Tucci and Clarke*, 1998] which led, during one period of about a decade, to settlement on land previously flooded. Subsequently, the river rose again, displacing ranchers and their cattle; analysis by Collischonn et al. detected variations in regional rainfall over the period concerned, but land use change may also have been a contributing factor.

[13] Tables 2a and 2b shows the result of fitting the Poisson model. The estimate of α was 11.33, with convergence after 26 iterations. The estimates of β_0 and β_1 are $\hat{\beta}_0 = -54.909 \pm 0.442$ and $\hat{\beta}_1 = -0.5579 \pm 0.0378$. Since the estimate $\hat{\beta}_1$ is $\sim$15 times its standard error, there is strong evidence of trend. The fitted values are the estimates of $E[y_t]$, namely, $\Gamma(1 + 11.33^{-1})$ exp $(54.909 + 0.5579\ t)$,

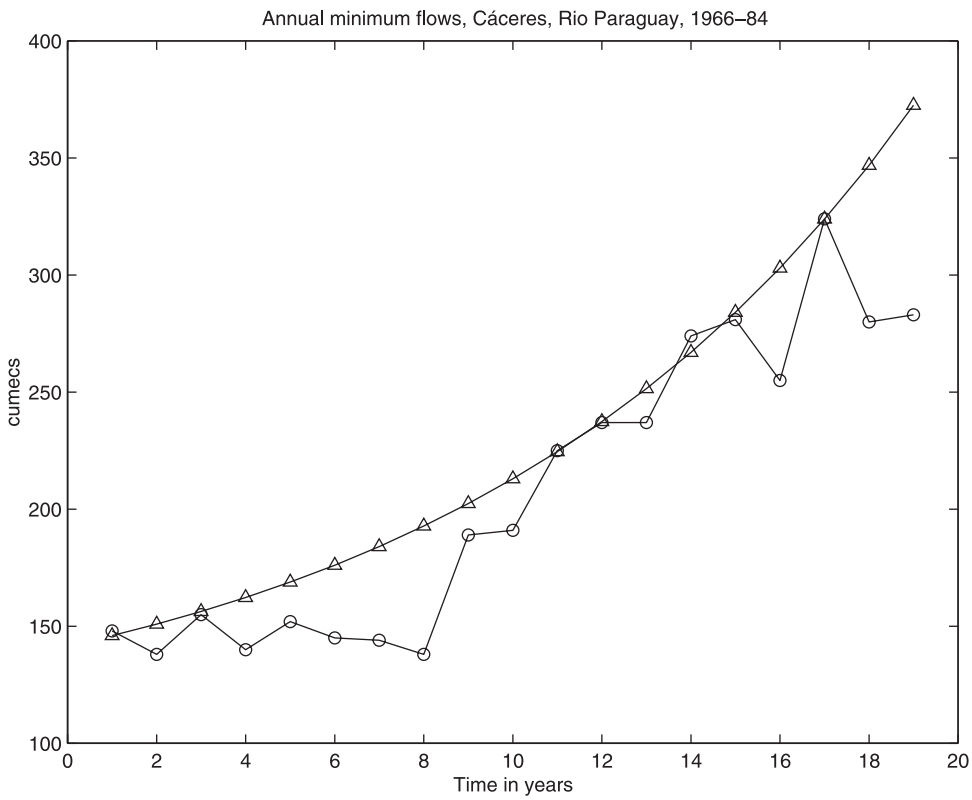

Figure 1. Annual minimum flows 1966–1984 at Cáceres, River Paraguay. Circles denote observed annual minima; triangles are fitted values estimated from the Poisson model described in the text.

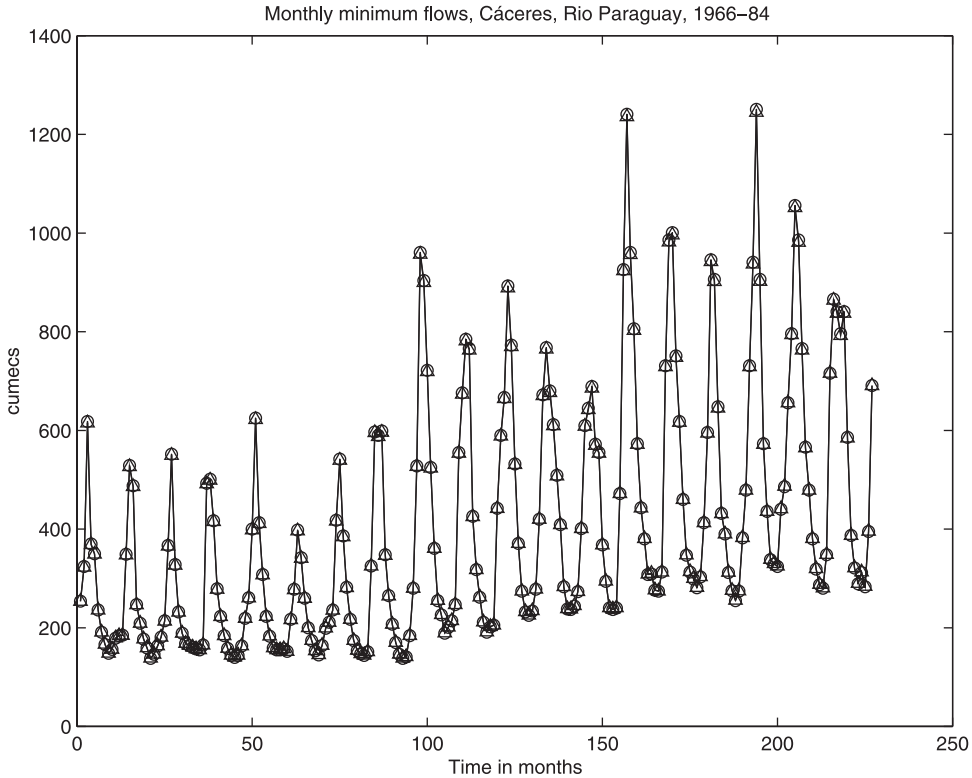

Figure 2. Monthly minimum flows 1966–84 at Cáceres, River Paraguay. Circles denote observed annual minima; triangles are fitted values estimated from the Poisson model described in the text.

where t is year number, and these are shown in Figure 1 together with the observed annual minimum flows. Just as in polynomial fitting by multiple regression, additional powers of the explanatory variable t can be added; inclusion of a variable t^2 so that $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2]^T$, $\mathbf{F} = [1\ t\ t^2]^T$ changes $\hat{\alpha}$ to 12.16 and gives an estimate $\hat{\beta}_2 = -0.0136 \pm 0.0101$. The ratio of the estimate to its standard error is 1.35, suggesting that the inclusion in the model of a term in t^2 does not improve fit. The modification to the analysis of deviance is shown in Table 2b; the total deviance is different for the two models because the estimate of α has changed.

[14] As a further illustration, the Poisson model was also fitted to the monthly minimum flows. The validity of this calculation is open to question since the assumption of serial independence in the data, implicit in the method, is likely to be invalidated in the case of monthly minimum flows. Since one monthly minimum was missing, there were $(12 \times 19) - 1 = 227$ data values shown plotted in Figure 2, which as expected, gives evidence of an annual cycle superimposed upon the general upward trend in minimum values. The vectors $\boldsymbol{\beta}$ and $\mathbf{F}$ were initially $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2]^T$ and $\boldsymbol{F} = [1\ \cos(2\pi t/12)\ \sin(2\pi t/12)]^T$; subsequently, an additional term t was added, giving $\boldsymbol{\beta} = [\beta_0\ \beta_1\ \beta_2\ \beta_3]^T$, $\mathbf{F} = [1t\ \cos(2\pi t/12)\ \sin(2\pi t/12)]^T$. The consequences of including this additional term are discussed in section 5. Tables 3a

Table 3a. Monthly Minimum Flows at Cáceres: Analysis of Deviance and Estimated Parameters From Fitting the Poisson Model for the *wi* With Offset $\alpha\ \log_e y_t$, $\boldsymbol{\beta} = [\beta_0\beta_1\ \beta_2\beta_3]^T$, and $\boldsymbol{F} = [1\ t\ \cos(2\pi t/12)\sin(2\pi t/12)]^T$

	DF	Deviance	Mean Deviance
Regression	3	238.60	79.5333
Residual	223	35.32	0.1584
Total	226	273.92	1.2120
	Estimate	Standard Error	Ratio
$\hat{\beta}_0$	−10.258	0.133	−76.96
$\hat{\beta}_1$	−0.00904	0.00101	−8.94
$\hat{\beta}_2$	−0.0944	0.0988	−0.96
$\hat{\beta}_3$	−1.2313	0.0902	−13.66

Table 3b. Monthly Minimum Flows at Cáceres: Analysis of Deviance and Estimated Parameters From Fitting the Poisson Model for the w_i With Offset $\alpha\ \log_e y_t$, $\boldsymbol{\beta} = [\beta_0\beta_1\beta_2\ \beta_3\ \beta_4\beta_5]^T$, and $\mathbf{F} = [1\ \mathrm{t}\ \cos(2\pi t/12)\ \sin(2\pi t/12)\ \cos(4\pi t/12)\ \sin(4\pi t/12)]^T$

	DF	Deviance	Mean Deviance
Regression	5	243.66	48.7315
Residual	221	30.26	0.1369
Total	226	273.92	1.2120
Change	−2	−5.06	2.5290
	Estimate	Standard Error	Ratio
$\hat{\beta}_0$	−10.232	0.133	−76.70
$\hat{\beta}_1$	−0.00916	0.00101	−9.05
$\hat{\beta}_2$	−0.0907	0.0943	−0.96
$\hat{\beta}_3$	−1.2316	0.0938	−13.13
$\hat{\beta}_4$	0.1742	0.0935	1.86
$\hat{\beta}_5$	−0.1225	0.0945	−1.30

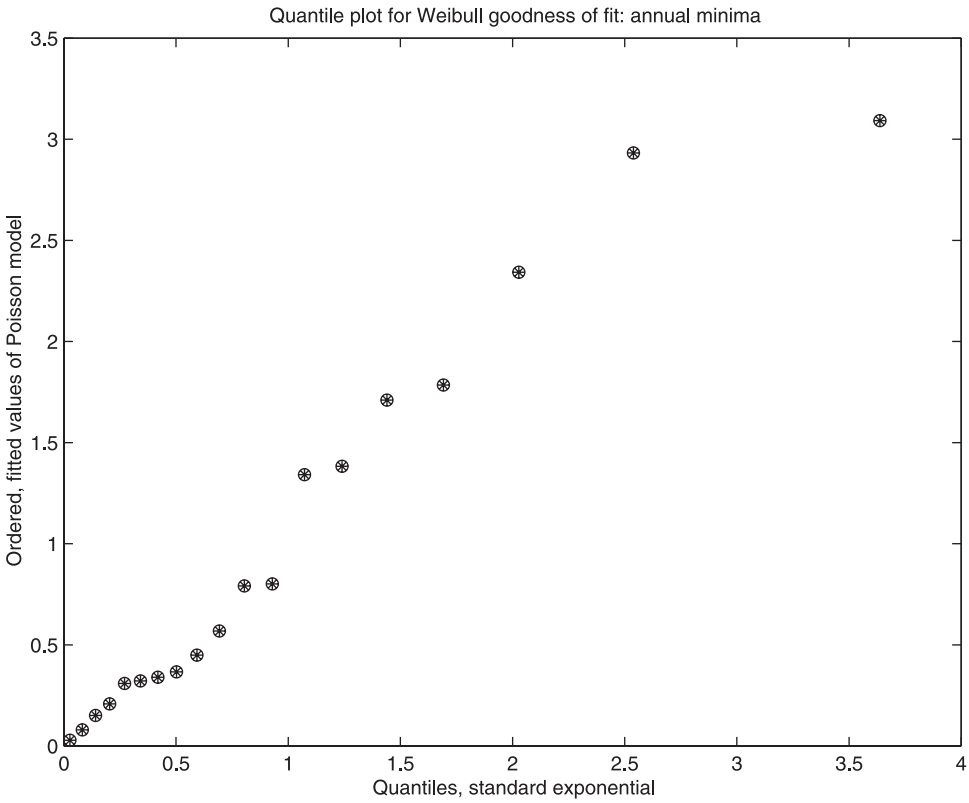

Figure 3a. Goodness of fit for annual minimum flows at Cáceres 1966–1984: quantile-quantile plot in which linearity denotes conformity with the Weibull model.

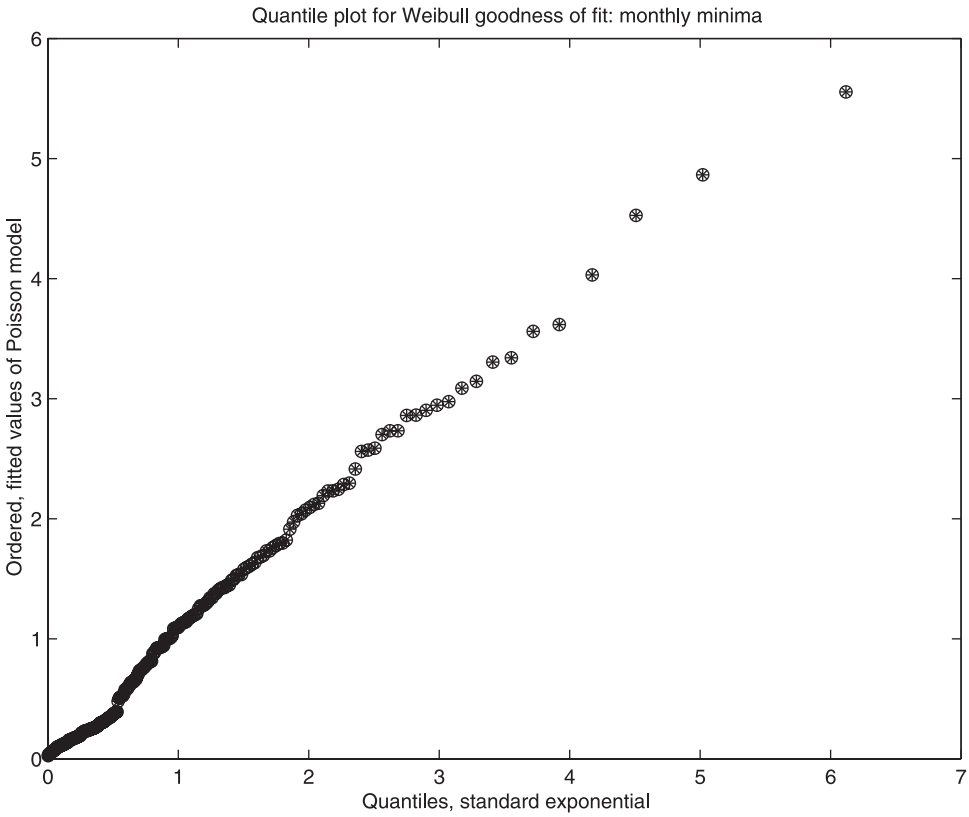

Figure 3b. Goodness of fit for monthly minimum flows at Cáceres 1966–1984: quantile-quantile plot in which linearity denotes conformity with the Weibull model.

and 3b show the analysis of deviance and the estimated coefficients $\boldsymbol{\beta}$ together with the changes resulting for the inclusion of additional terms $\cos(4\pi t/12)$, $\sin(4\pi t/12)$ in $\mathbf{F}$. The fitted value of α was $\hat{\alpha} = 1.925$. Apart from the constant β_0 (estimated as -10.258 ± 0.133), the trend coefficients that are large relative to their standard errors are the coefficients of linear trend (-0.009004 ± 0.00101) and of the sine component (-1.2313 ± 0.0902). Observed and fitted monthly minima are plotted together in Figure 2. Figure 2 shows that the model fitted well during months when minimum flows were low but that there was considerable discrepancy between observed and fitted flows in months when minimum flows were larger.

[15] Goodness of fit of the Weibull distribution can be shown by a plot analogous to that used to check normality of residuals in multiple regression. The residuals in this case are the fitted values in the Poisson model, and since these have exponential distributions [*Aitkin and Clayton*, 1980] when the Weibull distribution is appropriate, they may be put in ascending order and plotted against the quantiles of the standard exponential distribution $\exp(-t)$, corresponding to cumulative probabilities $p = (i - 0.5)/N$. As in the normal case a linear plot suggests conformity with the Weibull hypothesis. Figures 3a and 3b show the plots obtained for residuals from the analysis of annual and monthly minima, respectively; no strictly formal significance test is possible because the residuals are correlated. There is some departure from linearity at the upper end of the annual plot Figure 3a, but for the monthly data (Figure 3b), there is no strong evidence that the Weibull distribution is inappropriate. In Figure 3a, $\mathbf{F} = [1t]^T$; in Figure 3b, $\mathbf{F} = [1\cos(2\pi t/12)\ \sin(2\pi t/12)]^T$; we refer to them again in the discussion below.

4. Extension to Other Extreme Value Distributions

[16] *Aitkin and Clayton* [1980] show how the Poisson model may be used to fit a generalized extreme value (GEV) distribution. However, their GEV distribution has a different appearance from the GEV distribution well known to hydrologists and given (for example) by *Stedinger et al.* [1993]. The GEV of Aitkin and Clayton is written as

$$f_Y(y) = \alpha\delta y^{\delta-1} e^{\alpha y^\delta} \exp\left(-e^{\alpha y^\delta}\right) \quad -\infty < y < \infty,\ \alpha > 0. \quad (10)$$

A notable difference between this GEV distribution and the GEV widely used in hydrology, namely,

$$f_Y(y) = (1/\sigma)[1 - k(y-u)/\sigma]^{(1/k)-1} \exp\left\{-[1 - k(y-u)/\sigma]^{1/k}\right\}, \quad (11)$$

is that the latter distribution (equation (11)) is bounded (when $k > 0$, $y < (u + \sigma/k)$ and so is bounded above; when $k \langle 0, y \rangle (u + \sigma/k)$ and so is bounded below), while the GEV in equation (10) is unbounded, $-\infty < y < \infty$. It is not obvious whether or how one distribution may be transformed into the other (except insofar that any univariate continuous distribution can be transformed into any other by mapping it on to a uniform distribution). When $\delta = 1$, the distribution equation (10) is equal to the reversed Gumbel distribution (with y replaced by $-y$, giving a mirror image of the Gumbel, with negative skew), while the distribution in equation (11) reduces to the Gumbel (with positive skew) when $k \to 0$. Also, as $\delta \to 0$ in equation (10), the Weibull distribution is obtained.

[17] To fit equation (10), the Poisson model is fitted with offset αy_t^δ, for which α and δ must be calculated iteratively. *Aitkin and Clayton* [1980] recommend that initially, δ should be set equal to 1 and an initial estimate of α obtained by fitting a reversed Gumbel distribution using their Poisson model. However, an attempt to fit the reversed Gumbel to the Cáceres data, using the Poisson model, failed to converge, so a moment estimator was used to give the initial value of α. The IWLS procedure then becomes as follows: given the initial estimates of α and δ, the coefficients $\boldsymbol{\beta}$ ($= [\beta_0\ \beta_1 \ldots]^T$) are estimated, together with the fitted values $\hat{\mu}_t$ given by $\log_e \hat{\mu}_t = \alpha\, y_t^\delta + \boldsymbol{\beta}^{\mathbf{T}} \boldsymbol{F}_t$; then α and δ are reestimated from the maximum likelihood equations $\partial \log_e L/\partial\alpha = 0$ and $\partial \log_e L/\partial\delta = 0$, the process being repeated until convergence. Explicit forms for these two equations are given in Appendix A.

[18] When this procedure was used to fit equation (10) to the Cáceres data, convergence was extremely slow and showed no sign of nearing termination by 500 iterations (taking perhaps 15 s a Pentium PC). Iterating still further, the value of δ continued to approach 0, the value at which the GEV distribution (10) becomes a Weibull distribution. A possible explanation is that the Weibull distribution is a more appropriate distribution than the GEV distribution (10) for the annual minimum flow record at Cáceres.

[19] There is, however, some question (J. R. M. Hosking, personal communication, 2001) as to whether the Aitkin-Clayton GEV distribution, given in equation (10), is a valid probability density function since values of y^δ are undefined for $y < 0$ (except for δ integer); this may or may not have a bearing on the failure to converge for the Cáceres data. Aitkin and Clayton deduced the form of their GEV distribution by writing the Weibull density in the form

$$f_Y(y) = \alpha\, e^{\alpha \log y} \exp\left(-e^{\alpha \log y}\right)/y \qquad -\infty < \log y < \infty$$

and then noting that $\log y$ is the limiting case, as $\delta \to 0$, of the Box-Cox power transform $(y^\delta - 1)/\delta$, defined for $y > 0$, which led to the form shown in equation (10) for nonzero δ. Further research is needed to explore whether the difficulty in defining the Aitkin-Clayton GEV for negative y is of practical importance and whether it can be avoided by modifying the form given in equation (10) to $f_Y^*(y) = f_Y(y)/\int_0^\infty f_Y(y)dy$ for $y \geq 0$, although it is uncertain whether the procedure given in this paper for estimating the parameters $\boldsymbol{\beta}$, α, and δ could still be used. J. R. M. Hosking's (personal communication, 2001) recommendation is to treat the three extreme value families (Gumbel, Weibull, and Frechet) separately; the first of the three cases has been dealt with [*Clarke*, 2002], and the second is dealt with in the present paper. The Frechet family, with cumulative distribution function $F_Y(y) = \exp[-\lambda y^{-\alpha}]$, can then be transformed to the Gumbel case by setting $y = \log x$ and using the methods already given.

5. Discussion and Conclusions

[20] It is clear that generalized linear models, fitted to the Weibull and to the GEV distribution in equation (10) by

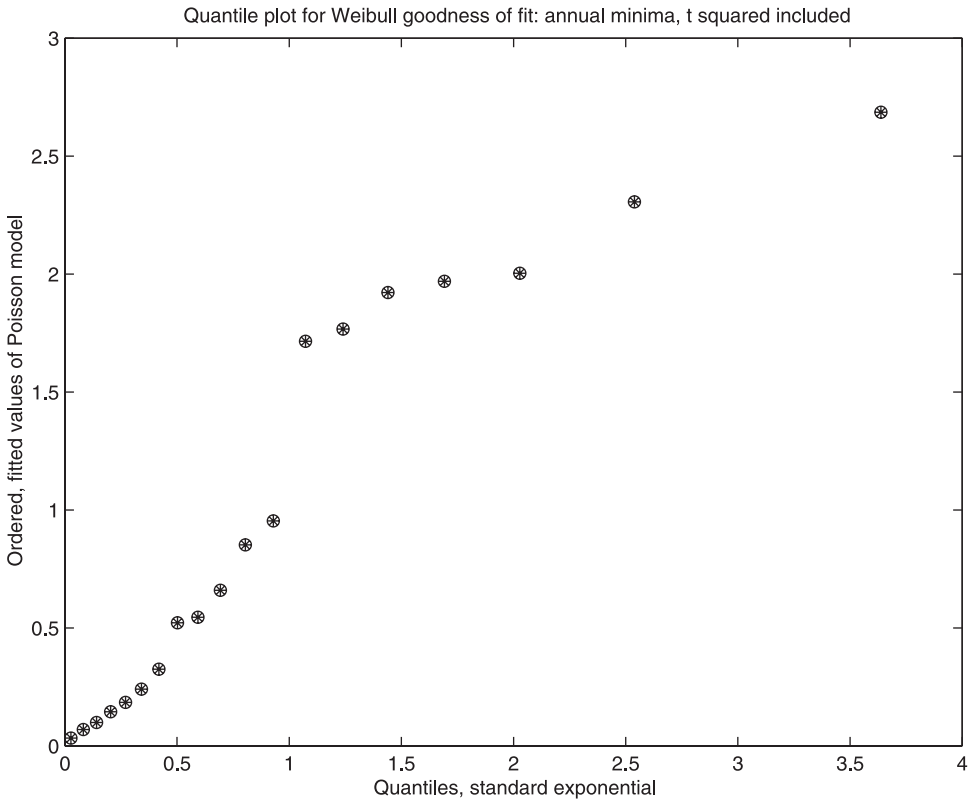

Figure 4a. Goodness of fit for annual minimum flows at Cáceres 1966–1984: effect on quantile-quantile plot of including a quadratic term t^2 in addition to a constant and a term in t.

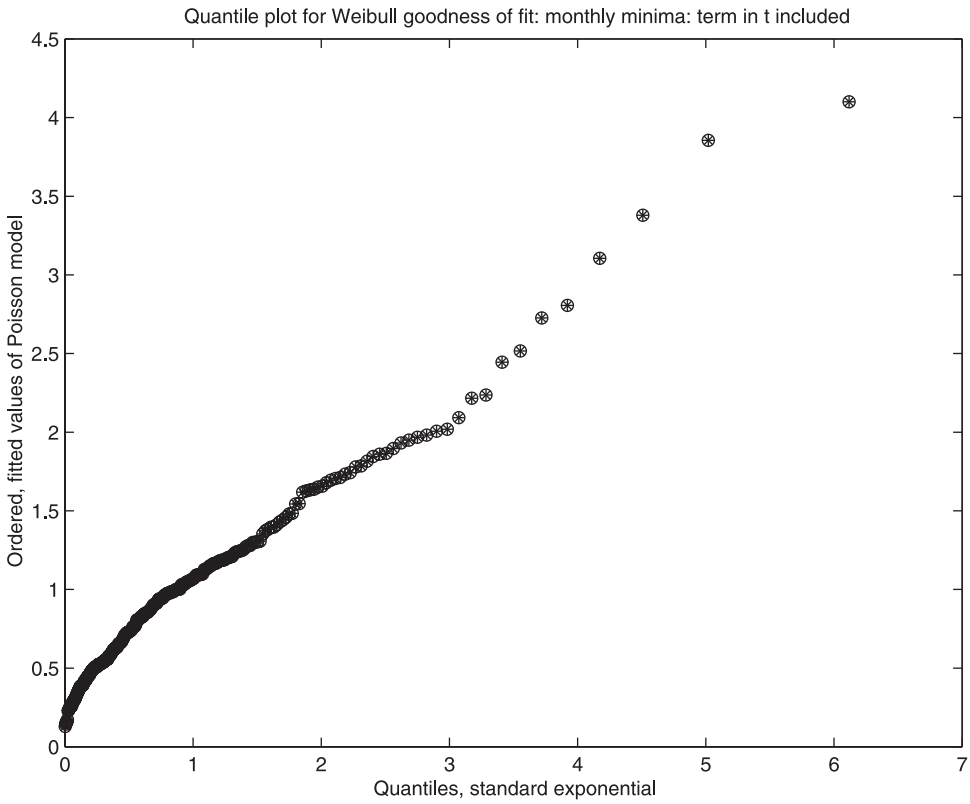

Figure 4b. Goodness of fit for monthly minimum flows at Cáceres 1966–1984: effect on quantile-quantile plot of including a term t in addition to a constant and a terms in cos $(2\pi t/12)$ and sin $(2\pi t/12)$.

casting the estimation procedure in the form of a Poisson model and using IWLS, are a powerful tool both for the analysis of time trends and for study of relationships between annual and/or monthly minimum flows and explanatory variables ***F***. Using good statistical software, the calculation needed to fit them is extremely rapid.

[21] However, it appears that convergence may be uncertain unless (1) the log link is used in the Poisson model and (2) the Weibull distribution is appropriate for the data. *Aitkin and Clayton* [1980] show how the Weibull may be fitted using identity or reciprocal links instead of the log link function, but they also point out that negative fitted values may occur, resulting in an undefined deviance. Using the Cáceres low-flow data, convergence was no problem using the Weibull and log link functions, but there were computational difficulties with the identity link which led to some negative fitted values and an undefined deviance statistic. In practice, this did not matter because the trend in the Cáceres data was well fitted using the log link. This was true for both the annual minimum flows and the monthly minimum flows.

[22] It is clear also that in principle, the Poisson model can be used to fit the particular GEV distribution given by *Aitkin and Clayton* [1980], although this distribution is unbounded for the random variable *Y*, in contrast to the GEV distribution used by hydrologists, which is bounded above or below, depending on whether the parameter k in equation (11) is positive or negative. It is not clear, however, how far the two GEV distributions (10) and (11) are equivalent, with either capable of being transformed into the other and, if they are not equivalent, whether the Aitkin-Clayton GEV has merits that justify its hydrological use for the types of data to which it can be fitted satisfactorily. As shown by its use with the Cáceres data, convergence of the IWLS algorithm cannot be guaranteed, possibly because it was less appropriate for annual and monthly minimum flows than the Weibull distribution, obtained when the parameter δ in equation (10) tends to zero and toward which the iterated values of δ were tending for these data. Failure of the ML estimates to converge was not simply a consequence of using the IWLS procedure; attempts to use a modified Newton method [*Genstat 5 Committee*, 1993] to maximize $\log_e L$ as a function of four parameters α, δ, β_0, and β_1, where α and δ are as in equation (10) and β_0 and β_1 define a trend term $\beta_0 + \beta_1 t$, also failed to converge before 30 iterations had been completed. A further attempt to find a solution, using the simplex procedure to maximize $\log_e L$, also failed to converge after 2000 iterations, at which the estimate of δ continued to decrease very slowly. Thus it seems likely that failure to converge indicates model unsuitability, rather than any fault or limitation in the Poisson model as applied to annual minimum flows.

[23] The number of terms included in the vector $\mathbf{F}$ of explanatory variables and therefore the number of parameters in the vector $\boldsymbol{\beta}$ of coefficients have an influence on how well the goodness of model fit appears in plots of fitted values against percentiles of the exponential distribution, of the type shown in Figures 3a and 3b for annual and monthly minima respectively. Figure 3a shows the quantile plot (in which linearity confirms consistence with the Weibull model) with $\mathbf{F} = [1\ t]^T$, and it can be seen that with the exception of perhaps one point, the plot is fairly close to linearity; Figure 4a shows the corresponding quantile plot when an additional power of t was included, giving $\mathbf{F} = [1\ t\ t^2]^T$, and the quantile plot now shows a considerably greater departure from linearity. The coefficient β_2 of t^2 was small relative to its standard error in this case, so it was reasonable to omit it. In the analysis of monthly minima, however, the picture was rather different: Figure 3b shows the quantile plot with $\mathbf{F} = [1\cos(2\pi t/12)\ \sin(2\pi t/12)]^T$, again with a fair degree of linearity, but Figure 4b shows the quantile plot when $\mathbf{F} = [1\ t\ \cos(2\pi t/12)\ \sin(2\pi t/12)]^T$, this time with the inclusion of t in addition to the harmonic terms. The coefficient β_t of t was large relative to its standard error and so should be included in the model, yet the quantile plot now shows considerable nonlinearity.

[24] In conclusion, this paper has shown how the existence of time trends in Weibull-distributed data can be tested by converting the problem into one in which a generalized linear model (GLM) is fitted to a power-transformed variable having a Poisson distribution. The trend coefficients (as well as the parameter in the power transform) are estimated by maximum likelihood, and their significance can be tested using properties of the deviance statistic. The method is used to test for trend in annual minimum flows over a 19-year period in the River Paraguay at Cáceres, Brazil, and in monthly flows at the same site. Extension of the procedure to testing for trend in data following a generalized extreme value (GEV) distribution is also discussed. Although a test for time trend in Weibull-distributed hydrologic data is the motivation for this paper, the same approach can be applied in the analysis of data sequences that can be regarded as stationary in time, for which the objective is to explore relationships between a Weibull variate and other variables (covariates) that explain its behavior.

Appendix A

[25] In the procedure described in the text for fitting the GEV distribution (10) of *Aitkin and Clayton* [1980], the two ML equations for iterative calculation of the parameters α and δ are as follows. For α,

$$\hat{\alpha} = \left(\sum_t (\mu_t - 1) y_t^{\delta} / N \right)^{-1}$$

$$\hat{\delta} = \left(\sum_t (\mu_t - 1) y_t^{\delta} \log_e y_t \Big/ \sum_t (\hat{\mu}_t - 1) y_t^{\delta} - \sum_t \log_e y_t / N \right)^{-1},$$

where $\mu_t = \exp(\alpha y_t^{\delta} + \boldsymbol{\beta}^{\mathbf{T}}\mathbf{F})$.

[26] **Acknowledgments.** The author wishes to thank J. R. M. Hosking, IBM Thomas J. Watson Research Center, for valuable advice and two anonymous authors for helpful comments.

References

Aitkin, M., and D. Clayton, The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM, *Appl. Stat.*, *29*, 156–163, 1980.

Aitkin, M., D. Anderson, B. Francis, and J. Hinde, *Statistical Modelling in GLIM*, Clarendon, Oxford, UK, 1989.

Clarke, R. T., Estimating time trends in Gumbel-distributed data by means of generalized linear models, *Water Resour. Res.,38*, 10.1029/2001WR000917, in press, 2002.

Collischonn, W., C. E. M. Tucci, and R. T. Clarke, Further evidence of changes in hydrological regime of the River Paraguay: Part of a wider phenomenon of climate change?, *J. Hydrol.*, *245*, 218–238, 2001.

Cox, D. R., and D. Oakes, *Analysis of Survival Data*, Chapman and Hall, New York, 1984.

Dempster, A. P., N. M. Laird, and D. B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J. R. Stat. Soc., Ser. B*, *39*, 1–38, 1977.

Genstat 5 Committee, *Genstat 5 Release 3 Reference Manual*, Clarendon, Oxford, UK, 1993.

Intergovernmental Panel for Climate Change (IPCC), *Climate Change 2001: Impacts, Adaptation and Vulnerability*, Cambridge Univ. Press, New York, 2001.

McCullagh, P., and J. A. Nelder, *Generalized Linear Models*, 2nd ed., Chapman and Hall, New York, 1989.

Salas, J. D., Analysis and modeling of hydrologic time series, in *Handbook of Hydrology*, edited by E. R. Maidment, chap. 19, pp. 19.1–19.72, McGraw-Hill, New York, 1993.

Stedinger, J. R., R. M Vogel, and E. Foufoula-Georgiou, Frequency analysis of extreme events, in *Handbook of Hydrology*, editor by E. R. Maidment, chap. 18, pp. 18.1–18.66, McGraw-Hill, New York, 1993.

Tucci, C. E. M., and R. T. Clarke, Environmental issues in the la Plata basin, *Water Resour. Dev.*, *14*(2), 157–173, 1998.

R. T. Clarke, Instituto de Pesquisas Hidráulicas, Caixa Postal 15029, Avenida Bonto Goncalves 9500, Porto Alegre, CEP 91501-970, Brazil. (clarke@if.ufrgs.br)