UC Santa Barbara
UC Santa Barbara Previously Published Works

Title

Fitting Minima of Flows Via Maximum Likelihood

Permalink

https://escholarship.org/uc/item/5xz238ck

Journal

Journal of Water Resources Planning and Management, 114(1)

ISSN

0733-9496

Authors

Loaiciga, Hugo A
Marino, Miguel A

Publication Date

1988

DOI

10.1061/(asce)0733-9496(1988)114:1(78)

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

Fitting Minima of Flows via Maximum Likelihood

By Hugo A. Loaiciga,[1] Associate Member, ASCE, and Miguel A. Mariño,[2] Member, ASCE

Abstract: A statistical method for deriving frequency distribution functions of minima of streamflows is presented. An innovative feature of the proposed methodology is that it does not require the specification of a parent distribution for streamflows, i.e., it is distribution free. The only assumption necessary is that the realizations of streamflows be independent, identically distributed random variables. The validity of this assumption is established with a nonparametric test. The main use of the methodology developed herein is in estimating small quantiles of the flow distribution for water supply planning and low-flow investigations. An example is included to illustrate the applicability of the approach, using a record of annual flows.

Introduction

The method developed in this study is aimed at providing water resource analysts with a statistically consistent method to fit the lower tail of the cumulative distribution function (CDF) of streamflows. The potential uses of the proposed approach are in water supply stability studies, effluent design, wildlife habitat and fishery resources management, estimation of hydroelectric power potential, and assessment of environmental quality (Task Committee on Low-Flow Evaluation, Methods and Needs 1980).

The knowledge of streamflow characteristics is important for water resource planning purposes. Due to the annual, seasonal, and daily variability of streamflows, it is necessary to characterize their statistical properties for suitable allocation among competing uses. In this regard one must mention the work of Riggs (1972) dealing with the application of (statistical) parametric methods and regional analysis for low-flow investigations. MacMahon (1976) presented a survey of computational procedures on low-flow analysis. In this study, a new method, which is classified as distribution free, is developed to estimate the lower tail of the cumulative distribution function of streamflows. In contrast to the parametric methods based on some parent distribution (e.g., lognormal, gamma, or Gumbel), the distribution-free approach does not require a parent distribution model. It will be shown herein that to estimate low-flow quantiles, it is enough to approximate the shape of the lower tail of the cumulative distribution function of streamflows by a parsimonious model involving only location, scale, and shape parameters. Subsequently, the maximum

[1] Asst. Prof., Dept. of Geological Sci., Wright State Univ., Dayton, OH 45435.

[2] Prof., Depts. of Land, Air, and Water Resour. and Civ. Engrg., Univ. of California, Davis, CA 95616.

Note. Discussion open until June 1, 1988. To extend the closing date one month, a written request must be filed with the ASCE Manager of Journals. The manuscript for this paper was submitted for review and possible publication on July 9, 1986. This paper is part of the *Journal of Water Resources Planning and Management*, Vol. 114, No. 1, January, 1988. ©ASCE, ISSN 0733-9496/88/0001-0078/$1.00 + $.15 per page. Paper No. 22143.

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

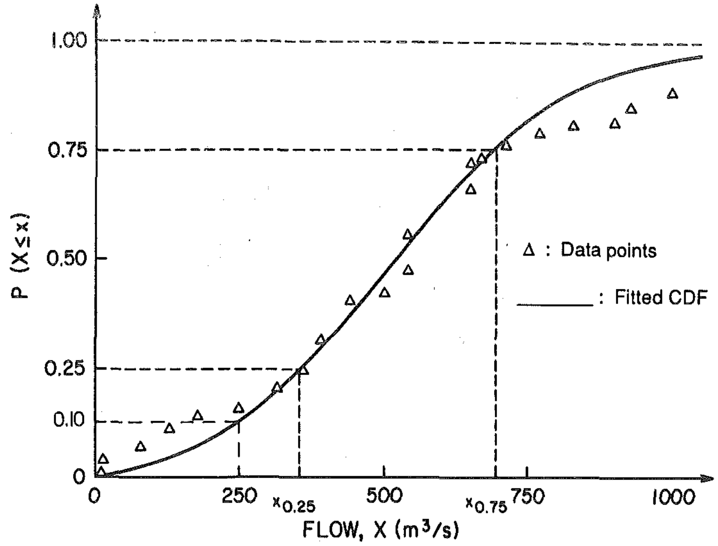

FIG. 1. Fitted CDF (three-parameter lognormal) to Streamflow Sample from Smith River, California

likelihood (ML) method is used to obtain consistent estimators of those parameters, from which low-flow quantiles are readily derived.

Problem Statement

Suppose a sample of streamflows X_1 , X_2 , . . . , X_N is available. The objective is to fit a CDF to the observed sample, with the ultimate goal being to estimate lower quantiles. It is documented in the hydrologic literature that some of the popular density functions (e.g., lognormal or those from the gamma family) usually do not approximate the upper or lower tails of streamflow records well. The main difficulty in fitting a theoretical CDF to a streamflow sample is shown in Fig. 1. It is clear that the theoretical CDF (in this case, a three-parameter lognormal) provides a good approximation of the observed flows in the interquartile range $x_{0.75}$–$x_{0.25}$, where $P(X \leq x_{0.75}) = 0.75$ and $P(X \leq x_{0.25}) = 0.25$. It is observed that fitted quantiles using a theoretical density function tend to underestimate (overestimate) observed flows in the upper (lower) tail of the CDF of streamflows.

In either the upper or lower tail cases, the use of theoretically estimated quantiles usually leads to risky decisions in planning studies, in the sense that the computed quantiles give an overly optimistic picture of the streamflow distribution. For example, $x_{0.10}$ is 250 m^3/s from the theoretical curve, a value that most likely overestimates the actual 10th quantile.

Distribution-Free Approach

Suppose that the flow variates, i.e., X_1 , X_2 , . . . X_N , are ordered from smallest to largest, and the sorted record is $X_{(1)} < X_{(2)} < . . . < X_{(N)}$. The

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

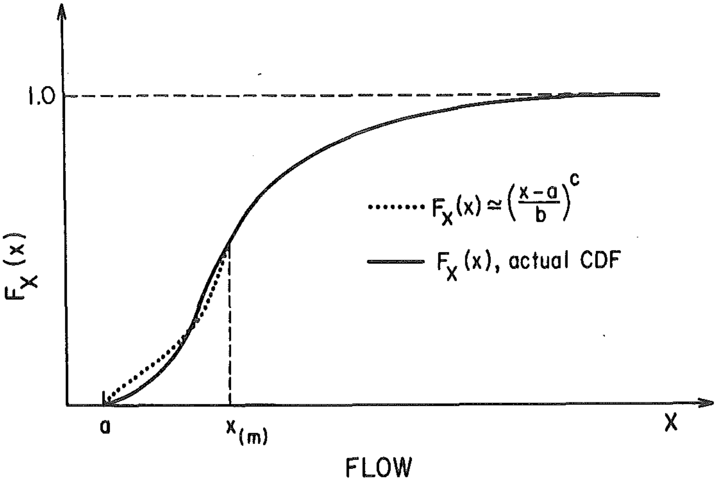

FIG. 2. Approximate and Actual CDFs

next step in the distribution-free approach is to censor out the variates $X_{(m+1)}, \ldots, X_{(N)}$, and the trimmed subsample $X_{(1)}, X_{(2)}, \ldots, X_{(m)}$ is used to fit the lower tail of the CDF. The purpose of censoring out $X_{(m+1)}$, $X_{(m+2)}, \ldots, X_{(N)}$ is to derive an approximation to the lower tail of the CDF that is not adversely affected by the flow values above a threshold flow $X_{(m)} = x_{(m)}$. Suppose that the CDF from which observations are taken satisfies the following condition (Rockette et al. 1974; David 1981)

$$P(X \leq x) = F_X(x) \simeq \left(\frac{x-a}{b}\right)^c \qquad (1)$$

in which a = the lower bound to X; and b and c are positive constants. Eq. 1 implies that the lower tail of the CDF of flows can be approximated by a power function that is defined by a location parameter a, and scale and shape parameters b and c, respectively. It is assumed that Eq. 1 is valid only as X approaches the lower bound a from above. Notice that Eq. 1 does not correspond to any of the well-known CDFs, e.g., lognormal, gamma, etc. It is simply a parsimonious approximation (i.e., in terms of location, scale, and shape parameters only) to the lower tail of a theoretical CDF. Models such as that implied by Eq. 1 but involving only location and scale parameters have been used in the analysis of order statistics. The reader is referred to David (1981) for a survey of the subject matter. Fig. 2 shows the idea behind the approximation given in Eq. 1. The actual CDF of X is given by the solid line, which is an unknown probabilistic model. The CDF is approximated by the dotted line, which corresponds to Eq. 1 as x tends to a from above. The selection of the number of observations in the trimmed sample, m, is discussed in the model application section later.

The objective is to derive consistent estimates $\hat{a}$, $\hat{b}$, and $\hat{c}$ of a, b, and c, respectively, and this is done by means of the ML method.

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

Likelihood Function

From Eq. 1, differentiation with respect to x yields the approximate density function

$$f_X(x) = cb^{-1}\left(\frac{x-a}{b}\right)^{c-1} \quad \text{(2)}$$

Since the random variables $X_1, X_2, \ldots, X_N$ are assumed to be independent and identically distributed, it can be shown that the joint distribution of the trimmed sample $X_{(1)}, X_{(2)}, \ldots, X_{(m)}$ is given by (David 1981)

$$f_{X_{(1)},\ldots,X_{(m)}}(x_{(1)},\ldots,x_{(m)}) = \frac{N!}{(N-m)!}\left[\prod_{i=1}^{m} f_X(x_{(i)})\right][1-F_X(x_{(m)})]^{N-m} \quad \text{(3)}$$

from which the log-likelihood function corresponding to Eq. 3 is

$$L = \ln N! - \ln (N-m)! + \sum_{i=1}^{m} \ln f_X(x_{(i)}) + (N-m)\ln[1-F_X(x_{(m)})] \quad \text{(4)}$$

Substitution of Eqs. 1 and 2 into Eq. 4 yields

$$L = C + m\ln c - m\ln b + (c-1)\sum_{i=1}^{m}\ln\left[\frac{x_{(i)}-a}{b}\right]$$

$$+ (N-m)\ln\left\{1-\left[\frac{x_{(m)}-a}{b}\right]\right\}^{c} \quad \text{(5)}$$

in which $C = \ln N! - \ln(N-m)!$. By letting

$$z = \left[\frac{x_{(m)}-a}{b}\right]^{c} \quad \text{(6)}$$

a first-order Taylor series approximation to the function e^{-z} yields

$$e^{-z} \simeq 1 - z \quad \text{(7)}$$

Replacing $1 - z$ by e^{-z} in the last term of Eq. 5 results in the final expression for the log-likelihood function

$$L = m\ln c - m\ln b + (c-1)\sum_{i=1}^{m}\ln\left[\frac{x_{(i)}-a}{b}\right]$$

$$-(N-m)\left[\frac{x_{(m)}-a}{b}\right]^{c} \quad \text{(8)}$$

where constant terms are omitted. The ML estimators of a, b, and c maximize L as given in Eq. 8.

Maximum Likelihood Estimators

Differentiating Eq. 8 with respect to b, setting the resulting expression equal to zero, and solving for b yields

$$\hat{b} = \left(\frac{N-m}{m}\right)^{1/c}[x_{(m)} - a] \quad \text{(9)}$$

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

If $\hat{b}$, as given in Eq. 9, replaces b in Eq. 8, one obtains the following expression

$$L(a, c) = m \ln c + (c - 1) \sum_{i=1}^{m} \ln [x_{(i)} - a] - m \ln (N - m)$$

$$+ m \ln m - cm \ln [x_{(m)} - a] - m \quad \ldots\ldots (10)$$

Differentiating Eq. 10 with respect to c, equating to zero, and solving for c, results in the following expression

$$\hat{c} = \frac{m}{\sum_{i=1}^{m} \ln \left[\frac{x_{(m)} - a}{x_{(i)} - a}\right]} \quad \ldots\ldots (11)$$

Replacing c in Eq. 10 by $\hat{c}$ as expressed in Eq. 11 leads to the concentrated log-likelihood function in terms of a only, i.e.,

$$L(a) = -m \ln \left\{ \sum_{i=1}^{m} \ln \left[\frac{x_{(m)} - a}{x_{(i)} - a}\right]\right\} - \sum_{i=1}^{m} \ln[x_{(i)} - a] \quad \ldots\ldots (12)$$

where constant terms are omitted. The function $L(a)$, where $0 \leq a < x_{(1)}$, can be maximized by any univariate search technique (e.g., Fibonacci search) to find the ML estimator $\hat{a}$. If one lets $\hat{a} \to x_{(1)}$, then $L(\hat{a})$ becomes unbounded, i.e., $L(\hat{a}) \to +\infty$, implying that $x_{(1)}$ is an inconsistent estimator of $\hat{a}$. This follows from Eq. 11, since for $\hat{a} \to x_{(1)}$, $\hat{c} \to 0$, and from Eq. 9, $\hat{b} \to +\infty$, which clearly are inconsistent estimators of c and b. Therefore, one must search for a local maximizer of $L(a)$ other than $x_{(1)}$. As it is shown in the example given later, it is fairly straightforward to detect the location of such local maximizers by first plotting $L(a)$ in the interval $0 \leq a < x_{(1)}$.

The steps to solve for the ML estimators $\hat{a}$, $\hat{b}$, and $\hat{c}$ are: (1) Plot $L(a)$ in the interval $0 \leq a < x_{(1)}$ to approximate $\hat{a}$; (2) use an univariate search technique to locate $\hat{a}$ precisely; and (3) find $\hat{b}$ from Eq. 9 and $\hat{c}$ from Eq. 11. Having $\hat{a}$, $\hat{b}$, and $\hat{c}$, the ML estimator of the pth quantile is obtained from Eq. 1, i.e.,

$$\hat{x}_p = \hat{a} + \hat{b}(p)^{1/\hat{c}} \quad \ldots\ldots (13)$$

in which $p = P(X \leq x)$ is close to zero.

Testing for Independent Identically-Distributed Flows

Test of Time Stationary Distribution

The approach presented in this study hinges on the assumption that the X_is are independent and drawn from the same (but unknown) distribution function. Two plausible examples are annual flows and seven-day lowest mean flows. Intuitively, such flows are due to the cumulative effect of a large number of hydroclimatic factors, and the resulting observations will have the character of random, independent variables. If this is indeed the case, one must verify that the flow realizations over different periods of time have the same distribution. A suitable test of the null hypothesis

$$H: F_1 = F_2 = \ldots = F_N \quad \ldots\ldots (14)$$

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

i.e., that the distribution functions are the same, is a variation of the Mann-Whitney form of the Wilcoxon test, as suggested in Lehmann (1975).

Suppose that a series of flow values $x_1, x_2, \ldots, x_N$ observed at times $1, 2, \ldots, N$ are available. If the observations are ordered from smallest to largest, each observation will have a rank corresponding to its location in the sorted sample, e.g., if x_N is the second smallest, its rank is $T_N = 2$, and so forth. Define the statistic

$$D = \sum_{t=1}^{N} (T_t - t)^2 \qquad (15)$$

If the flows have an upward trend (over time), this means that the distribution functions change over time, violating Eq. 14. Thus, if indeed the flows have an upward trend, then large values of the ranks T_t will tend to occur for large values of t, and small values of T_t for small values of t, and the statistic D in Eq. 15 will be small. Therefore, when testing the null hypothesis in Eq. 14 against the alternative of an upward trend, the null hypothesis is rejected for small values of D. A similar argument indicates that the null hypothesis should be rejected for large values of D when testing against the alternative that flows show a downward trend. For sufficiently large N, it can be shown that under the null hypothesis, D is approximately normal, with expected value

$$E_D = \frac{N(N^2 - 1)}{6} \qquad (16)$$

and variance $$\sigma_D^2 = \frac{N^2(N + 1)^2(N - 1)}{36} \qquad (17)$$

If the null hypothesis is tested against an upward trend, the rejection criterion is that the null hypothesis should be rejected if

$$P(D \le \bar{D}) = P\left(Z \le \frac{\bar{D} - E_D}{\sigma_D}\right) \qquad (18)$$

is small (i.e., less than 0.10), in which $\bar{D}$ = the computed statistic D from the actual data (see Eq. 15); and Z = a standardized normal variate. When testing against a downward trend, the null hypothesis (Eq. 14) should be rejected whenever

$$P(D \ge \bar{D}) = 1 - P\left(Z \le \frac{\bar{D} - E_D}{\sigma_D}\right) \qquad (19)$$

is small (i.e., less than 0.10). The rank test is illustrated later in the application example.

Testing for Independent Flows

Eqs. 18 and 19 permit the testing for a time stationary distribution of flow variates. In addition, one must test the assumption that flows are independent. Flow series such as annual runoff and seven-day lowest mean flows are dependent on a large number of climatic and hydrogeologic factors so that consecutive flow variates can be treated as independent. If successive flow values are dependent, then there would be a tendency toward

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

clustering so that high or low values tend to occur together. In view of this, one plausible test for independence is to consider runs of like elements in a flow series, and to reject the independence hypothesis when the number of runs is too small. Suppose that the runs are defined by letting Z_t be equal to zero or one, depending on whether X_t is below or above the median of the X_ts. Based on this criterion, there will be a total number of runs, say R, of zeroes and ones. The distribution of the statistic R, under the null hypothesis that successive flows are independent, has been derived by Wald and Wolfowitz (1940), and is given by

$$P(R = 2k) = \frac{2\binom{n-1}{k-1}\binom{n-1}{k-1}}{\binom{2n}{n}} \qquad (20)$$

and $$P(R = 2k + 1) = \frac{2\binom{n-1}{k}\binom{n-1}{k-1}}{\binom{2n}{n}} \qquad (21)$$

in which n = the number of ones and the parentheses in the right-hand side of Eqs. 20 and 21 indicate binominal coefficients. The null hypothesis of independent random variables should be rejected whenever

$$P(R \leq \bar{R}) \qquad (22)$$

is small (e.g., less than 0.10), in which $\bar{R}$ = the observed number of runs in the data set. The computation of the probability in Eq. 22 can be significantly simplified for large values of the total sample size N by using the fact that the null distribution of R is approximately normal (Hogg and Craig 1978), with mean

$$E_R = n + 1 \qquad (23)$$

and variance $$\sigma_R^2 = \frac{n(n-1)}{2n-1} \qquad (24)$$

so that the required probability in Eq. 22 becomes

$$P(R \leq \bar{R}) = P\left(Z \leq \frac{\bar{R} - E_R}{\sigma_R}\right) \qquad (25)$$

and is readily available from standard normal tables.

Model Application

The methods developed earlier are illustrated with a series of annual runoff volumes in the American River, upstream of Folsom Lake, which is located in the foothills of the Sierra Nevada, northeast of Sacramento in northern California. The contributing drainage area at the gage station is approximately 4,980 km^2 . Table 1 contains the 76-year long record. In this

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

TABLE 1. Annual Runoff Volumes of American River Upstream of Folsom Lake, California

Water year[a] (1)	Annual total[b] (k acre-ft) (2)	Water year[a] (3)	Annual total[b] (k acre-ft) (4)	Water year[a] (5)	Annual total[b] (k acre-ft) (6)
1904–05	2,024.5	1930–31	654.8	1956–57	2,296.6
1905–06	4,761.7	1931–32	2,574.1	1957–58	4,205.4
1906–07	5,710.4	1932–33	1,325.1	1958–59	1,315.4
1907–08	1,453.6	1933–34	1,128.8	1959–60	1,760.7
1908–09	4,544.6	1934–35	2,572.1	1960–61	1,180.5
1909–10	3,647.2	1935–36	3,414.6	1961–62	2,171.0
1910–11	5,477.7	1936–37	2,400.7	1962–63	3,386.6
1911–12	1,264.7	1937–38	4,522.0	1963–64	1,914.3
1912–13	1,433.7	1938–39	1,086.0	1964–65	4,421.3
1913–14	3,949.6	1939–40	3,442.1	1965–66	1,516.5
1914–15	3,061.3	1940–41	3,212.5	1966–67	3,987.0
1915–16	3,848.4	1941–42	3,990.7	1967–68	1,844.6
1916–17	2,831.7	1942–43	3,931.0	1968–69	4,548.8
1917–18	1,419.5	1943–44	1,537.0	1969–70	3,380.0
1918–19	2,155.0	1944–45	2,564.0	1970–71	3,040.4
1919–20	1,391.2	1945–46	2,857.7	1971–72	2,067.9
1920–21	3,221.5	1946–47	1,419.2	1972–73	3,093.1
1921–22	3,349.3	1947–48	2,262.5	1973–74	4,407.8
1922–23	2,750.2	1948–49	1,906.0	1974–75	2,785.7
1923–24	530.4	1949–50	2,704.9	1975–76	1,142.3
1924–25	2,759.0	1950–51	4,667.5	1976–77	356.0
1925–26	1,374.0	1951–52	5,030.2	1977–78	2,963.0
1926–27	3,627.9	1952–53	2,706.5	1978–79	2,346.5
1927–28	2,527.2	1953–54	2,067.9	1979–80	3,971.8
1928–29	1,156.3	1954–55	1,685.7		
1929–30	1,578.6	1955–56	4,781.3		

[a] Water year spans from October 1 to September 30.
[b] Units are in k acre-ft; 1 k acre-ft = 1.23×10^6 m^3.

section, the annual flow series is first tested for independence and stationary distribution, and subsequently, low quantiles are estimated.

Test of Hypotheses

In testing for independence according to Eq. 25, the number of runs was found to be $R = 41$ (i.e., $k = 20$) and the number of ones was $n = 38$ (i.e., an equal number of flows are above and below the median). The probability $P(R \leq 41)$ [$= P(R = 2) + P(R = 3) + \ldots + P(R = 40)$] is approximately 68%, and exceeds 0.10, and therefore the null hypothesis should not be rejected, i.e., the null hypothesis of independence of successive flows is supported by the available data. It was calculated that the hypothesis of independence should be rejected if the number of runs is less than 33.

The test for stationary distribution, as stated in Eq. 14, was conducted by first testing against the alternative of an upward trend on the flows. The test statistic was obtained as $\overline{D} = 73{,}844$, and $P(D \leq \overline{D})$ (see Eq. 18) was found to be approximately 53%; this exceeds 0.10, so the null hypothesis of stationary distribution is not rejected. The null hypothesis should be rejected only if $\overline{D} < 61{,}451$. The test of the null hypothesis against the alternative of an upward trend (see Eq. 19) indicated that the null hypothesis should not be rejected. In this case, rejection would occur if $\overline{D} > 84{,}849$.

In conclusion, the available flow data support the basic assumption of independent identically distributed flows, and hence, the proposed methodology for maximum likelihood estimates of quantiles is well founded.

Results of ML Estimation

The first step in the implementation of the estimation approach consisted of choosing an adequate value of m, the size of the trimmed sample. In this study, a semianalytical approach was used in which Akaike's (1974) information criterion (AIC) yielded a preliminary value of m and subsequently several additional values of m were tried in the estimation of low-flow quantiles to contrast their performance against observed data. The final choice of the trimmed sample size was based on the accuracy of the approximation to the observed data by the estimated quantiles for a given size of the trimmed sample. As will be seen later, different values of m appear to perform better, depending on whether the quantiles fall in the ranges $p < 1/100$ or $p \geq 1/100$ [recall that $p = P\ (X \leq x)$]. It must be kept in mind that the model Eq. 1 proposed for the approximation of the CDFs lower tail is theoretically valid only as quantiles approach the lower bound a of flows. In theory, this means that the procedure should be restricted to $p < 1/100$, but it is shown later that in practice it can be successfully applied to quantiles corresponding to p-values as large as 1/10.

The preliminary scanning of the trimmed sample size according to the AIC indicated that $m = 30$. Nevertheless, a wide range of m-values from $m = 5$ to $m = 40$ was tested to acquire a clear understanding of the performance of the method. Fig. 3 shows a plot in logarithm scales of the flow quantiles 356.0, 530.4, 654.8, 1,086.0, 1,128.8, 1.142.3, 1,156.3, and 1,180.5 (all units are in k acre-ft; 1 k acre-ft = $1.23 \times 10^6\ \mathrm{m}^3$) corresponding to $p = P(X \leq x) = 1/77, 2/77, 3/77, 4/77, 5/77, 6/77, 7/77$, and $8/77$, respectively. It is clearly seen that lower quantiles corresponding to $p =$

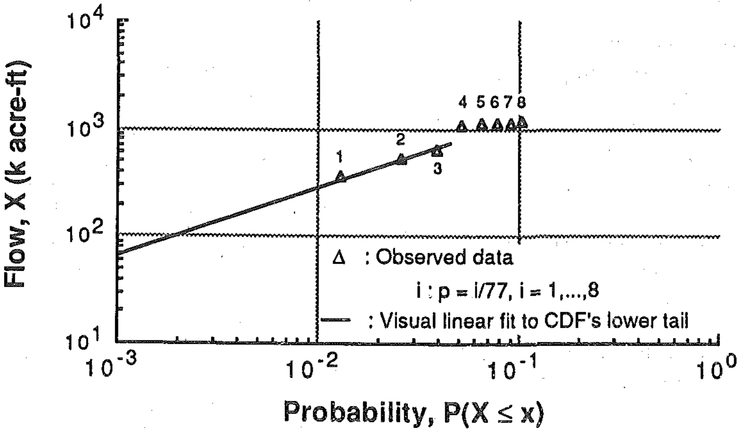

FIG. 3. Observed and Extrapolated Behavior of Lower Tail of CDF of Annual Runoff

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

TABLE 2. Results of Maximum Likelihood Estimation of Flow Quantiles Based on Annual Runoff Data

	Parameters			Quantiles, $\hat{x}_p$[b]							
m[a] (1)	$\hat{a}$ (2)	$\hat{b}$ (3)	$\hat{c}$ (4)	1/77 (5)	2/77 (6)	3/77 (7)	4/77 (8)	5/77 (9)	6/77 (10)	7/77 (11)	8/77 (12)
20	0	2,194.3	3.1262	547	682	777	852	915	970	1,019	1,063
25	0	2,583.0	2.3796	416	557	660	745	818	884	943	997
30	0	2,643.7	2.1698	357	492	592	677	749	815	876	931
35	0	2,750.3	1.8701	270	390	485	566	637	703	763	819

[a] m defines the value $x_{(m)}$ at which the sample is trimmed.

[b] The quantiles are defined in terms of the Weibull plotting positions ($i/N+1 = i/77$). The lowest and largest plotting positions considered are 1/77 = 1.3% and 8/77 = 10.4%, respectively. Units are in k acre-ft; 1 k acre-ft = 1.23 × 10^6 m^3.

[c] The observed values of the quantiles from the flow record in Table 1 are 356.0; 530.4; 654.8; 1,086.0; 1,128.8; 1,142.3; 1,156.3; and 1,180.5.

1/77 and 2/77 fall in a fairly straight line that has been extrapolated to the quantile corresponding to $p = 0.001$. There is a clear transition from the quantile corresponding to $p = 3/77$ to that corresponding to $p = 4/77$. Evidently, the CDF shows a distinctive behavior for $p \leq 2/77$ and for $p \geq 4/77$. Notice that the extrapolated linear behavior of the CDFs lower tail in the logarithmic scales is in agreement with the power law structure in model Eq. 1, which was proposed as an approximation to the CDF for small values of p. A word of caution is warranted regarding the extrapolation for $p < 1/77$, for it represents only a guess of actual behavior in the absence of further data points.

Table 2 summarizes the results of ML estimation of the location (a), scale (b), and shape (c) parameters, and of low-flow quantiles. The results are given for several values of m. It was found that for values of $m \leq 6$, the log-likelihood function was monotonically increasing in the interval $[0, x_{(1)}]$, with a global, yet inconsistent, maximum at $x_{(1)}$. Notice that even though there were no observed data in the interval $p < 0.01$ the graphical plot in Fig. 3 suggests that the lower bound is not equal to $x_{(1)} = 356$ k acre-ft but to some lower value. It is also emphasized that the linear behavior of the CDFs lower tail for $p < 1/77$ represents only a hypothetical guess as to its actual behavior in that range.

Several other values of m larger than six were tried, and those shown in Table 2 give a good summary of the dependence of estimates on m. For $m = 20$, 25, 30, and 35, the ML of the lower bound a was found to be equal to zero. Although on physical grounds it may be argued that zero runoff is not plausible, the estimate of a at a zero level is statistically consistent and maximizes the joint probability of having observed the runoff data. Certainly, Fig. 3 indicates an extrapolated behavior representative of a monotonic linear rate of decrease (in the logarithmic scales) of the distribution tail. Had we had a longer data set that included observed quantiles in the range $p \leq 1/100$, it could have been quite plausible to obtain nonzero estimates of the lower bound. For the observed data, and given the model proposed in Eq. 1, the nonzero estimates for the lower bound are statistically consistent and they lead to good approximations to observed quantiles, which is the ultimate goal of the estimation approach. Notice that in Table 2, quantiles associated with Weibull plotting positions ranging from 1/77 = 1.3% to 8/77 = 10.4% were presented, although in theory, the model Eq. 1 of the CDFs lower tail is strictly valid as X approaches the location parameter a from above. From a theoretical point of view, this means that the approximation is valid, say, for $p \leq 0.01$. However, higher quantiles were included to explore the usefulness of Eq. 1 away from the location parameter a. It is evident from Table 2 that for $m = 20$, the quantile estimates exceed observed quantities for $p \leq 3/77$, but gave the better approximation for $p \geq 4/77$ than the other estimates corresponding to $m = 25$, 30, and 35. On the other hand, it is seen that the excellent approximation is obtained of the lower quantiles ($p \leq 2/77$) when $m = 30$. In fact, due to the validity of Eq. 1 as an approximation of the CDF of minima as $X \rightarrow a$, there is a strong indication that the proper value of m should be 30. Parametric models based on common distribution models usually fail to provide good estimates of low quantiles, and in this application it is apparent the suitability of Eq. 1 as a model for the lower tail of CDFs of streamflows. Even though the flow record is not long enough and the value corresponding to $p = 1/100$ is not available, its extrapolated estimate using $m = 30$ yields that $\hat{x}_{0.01} = 316$ k acre-ft ($= 390 \times 10^6$ m^3), which is to be contrasted with the graphical value of 300 k acre-ft ($= 370 \times 10^6$ m^3) obtained from Fig. 3. Therefore, the flow value, which is likely to be observed in the long run once every 100 years is 316 k acre-ft. For $m = 25$, estimates are reasonably accurate, with an overestimation of observed values for $p \leq 3/77$, and an underestimation of the actual quantiles for $p \geq 4/77$. When $m = 35$, an accentuated bias to underestimate observed quantiles is clear for the range of considered plotting positions.

The estimation results suggest that quantile estimates for $p < 0.02$ should be computed using $m = 30$, whereas estimates for $0.05 \leq p \leq 0.10$ should be based on $m = 20$. For $0.02 \leq p < 0.05$, $m = 25$ appears to yield the better estimates.

Summary and Conclusions

An approach for estimating low-flow quantiles was developed in this study. Its innovative feature is that it is a distribution-free methodology based on a parsimonious approximation of the CDFs lower tail, involving location, scale, and shape parameters only, and does not depend on any particular probabilistic model (e.g., lognormal or gamma-type densities). The approach is suitable for estimating low-flow quantiles, i.e., return periods larger than or equal to 10, in arbitrary time units, and hinges on the assumption of independent identically distributed flows. Two nonparametric tests based on ranks were used to test for independence and stationary distribution function, showing the suitability of the assumption on a record of annual flows.

The estimation of location, scale, and shape parameters, as well as low-flow quantiles, was done via maximum likelihood using a trimmed sample that included the first m-order statistics, i.e., $X_{(1)}, X_{(2)}, \ldots, X_{(m)}$. The computational example analyzed the behavior of the estimates for a broad range of values of m. It was found that different values of the trimmed sample size m were associated with the better low-flow quantile estimation in the ranges $p < 0.02$, $0.02 \leq p < 0.05$, and $0.05 \leq p \leq 0.10$.

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.

Even though the proposed model Eq. 1 to the CDFs lower tail is valid as the lower bound is approached, its application to cumulative probabilities $p \leq 10$ yielded quite accurate estimates. Evidence in the research literature points to the severe difficulty in accurately estimating low quantiles, and the example described herein indicates that such flow minima are well approximated using the distribution-free method advanced herein.

Acknowledgment

The research leading to this report was supported by the University of California, Water Resources Center, as part of Water Resources Center Project UCAL-WRC-W-634.

Appendix I. References

Akaike, H. (1974). "A new look at the statistical model identification." *IEEE Trans. Automatic Contr.*, AC-19(6), 716–723.

David, H. A. (1981). *Order statistics*, 2nd Ed. John Wiley and Sons, Inc., New York, N.Y.

Hogg, R. V., and Craig, A. T. (1978). *Introduction to mathematical statistics*, 4th Ed. MacMillan Publishing Co., Inc., New York, N.Y.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks.* Holden-Day, Inc., San Francisco, Calif.

MacMahon, T. A. (1976). "Low flow analyses of streams: Details of computational procedures and annotated bibliography." *Research Report No.* 5, Department of Civil Engineering, Monash University, Clayton, Australia.

Riggs, H. C. (1972). "Low-flow investigations." *U.S. Geological Survey techniques for water resources investigations*, Book 4, Chapter B1, U.S. Geological Survey, Washington, D.C.

Rockette, H., Antle, C., and Klimbo, L. A. (1974). "Maximum estimation with the Weibull model." *J. Amer. Statist. Assoc.*, 69(345), 246–249.

Task Committee on Low-Flow Evaluation, Methods and Needs, (1980). "Characteristics of low flows." *J. Hydr. Div.*, ASCE, 106(5), 717–731.

Wald, A., and Wolfowitz, J. (1940). "On a test whether two samples are from the same population." *Ann. Math. Statist.*, 11, 147–162.

Appendix II. Notation

The following symbols are used in this paper:

a = location parameter;
$\hat{a}$ = estimate of location parameter;
b = scale parameter;
$\hat{b}$ = estimate of scale parameter;
CDF = cumulative distribution function;
c = shape parameter;
$\hat{c}$ = estimate of shape parameter;
D = test statistic in test for stationary distribution;
$\overline{D}$ = observed or computed value of D;
E = expected value;
F_X = cumulative distribution function of X;
f_x = probability density function of X;
L = log-likelihood function of trimmed sample;

ML = maximum likelihood;
m = number of observations in trimmed sample;
N = total number of flow observations;
p = value between 0 and 1 defining quantiles, $p = P(X \leq x)$;
R = number of runs in test for independence;
$\overline{R}$ = observed value of R;
T_t = rank of flow value at time t;
t = time ($t = 1, 2, \ldots, N$);
X = flow variate;
$X_{(i)}$ = ith-order statistic of X;
x = realization of flow variate;
$x_{(i)}$ = observed value of ith-order statistic;
Z = standard normal variate; and
σ^2 = variance.

Downloaded from ascelibrary.org by University of California, Santa Barbara on 09/30/24. Copyright ASCE. For personal use only; all rights reserved.