

Local Polynomial Based Flood Frequency Estimator for Mixed Population

Somkiat Apipattanavis¹; Balaji Rajagopalan²; and Upmanu Lall³

Abstract: Floods are often generated by more than one physical mechanism, e.g., rainfall and snowmelt. Consequently, traditional flood frequency methods that use a single distribution may not adequately describe the observed flood variability. Mixed distribution models have been proposed but they have two major drawbacks when applied to observed data: (1) determining the appropriate number of components or flood mechanisms and (2) identifying the probability distribution to be used for each component. Further, available flood data are often not sufficient for detecting mixture populations. As a result, mixed-distribution models can be difficult to apply in practice. In this paper we present a nonparametric approach based on local polynomial regression for estimating a flood quantile function that is data driven, flexible, and can capture any arbitrary features present in the data, enabling to alleviate the drawbacks of the traditional methods. We applied the proposed method to a suite of synthetic data from mixture of conventional distributions and to flood records that exhibit mixed population characteristics from Gila River basin of southeast and central Arizona. It is found that the proposed method provides a better fit to both the synthetic and historical data. Although, the proposed method is presented in the context of mixed population flood frequency estimation, the data-driven nature of the method lends itself as a simple, robust, and attractive alternative to traditional flood frequency estimation.

DOI: 10.1061/(ASCE)HE.1943-5584.0000242

CE Database subject headings: Flood frequency; Estimation; Rainfall; Snowmelt.

Author keywords: Flood frequency estimator; Mixed flood population; Local polynomial regression.

Introduction

Flood frequency analysis entails relating the magnitude of annual maximum flood flows to their frequency of occurrence at a gauged site. The typical interest is estimating extreme flood quantiles, i.e., 100-year or 500-year flood, from a small number of observations (~50–90 years) for design of hydraulic structures such as dams, culverts, and bridges. Design flood which is estimated from flood frequency analysis is very important for water resources planning and management, e.g., flood protection, channel improvement, and drainage system.

Traditional parametric methods for design flood computation assume that annual maximum floods are independent and identically distributed and drawn from a single homogeneous population with a known probability density function (PDF). An appropriate PDF is selected from a candidate set or mandated by a regulatory agency for at-site applications. Typical distributions that are prescribed by agencies such as USBR and USGS and widely used in practice are log Pearson Type III (LP3), log nor-

mal (LN), and extreme value Type I (EVI) distributions [see Kite 1977; Interagency Advisory Committee on Water Data (IACWD) 1982; Chow et al. 1988]. There are statistical tests to discriminate between choices of distributions including L-moment methods (see Kite 1977; Vogel 1986; Hosking 1990; Vogel and McMartin 1991). However, it is often difficult to distinguish between candidate models for a given data set, and best fit criteria emphasize the bulk of the distribution rather than its tails. Consequently, there is considerable uncertainty as to the best underlying model for the estimation of the upper flood quantiles.

Increasing evidence has been showing that floods are often generated by two or more distributions and not a single distribution as traditional method assumes, therefore, that this might be a significant reason why none of single population distributions provide an appropriate fit to the flood data. Alila and Mtraoui (2002) summarized that the natural factors that cause the mixed populations are (1) seasonal variations in the flood-producing mechanisms, e.g., hurricanes, thunderstorms, snowmelt flood in spring, and rainfall flood in summer (Waylen and Woo 1982; Jarret and Costa 1988); (2) changes in weather patterns resulting from low-frequency climate shift and/or El Niño/La Niña oscillations (Webb and Betancourt 1992; Alila and Mtraoui 2002, Jain and Lall 2000, 2001); (3) changes in channel routing owing to the dominance of within-channel or floodplain flow (Woltemade and Potter 1994); and (4) changes in antecedent soil moisture and soil cover resulting from basin variability.

Recognizing the mixture of flood populations, more researchers have been using mixed-distribution model methods to fit the heterogeneous flood distributions. Commonly, they first classify annual flood series according to flood-producing mechanisms into subpopulations, then fit each subpopulation with a single population distribution, and finally, combine the distributions to fit the

¹Dept. of Royal Irrigation, Office of Research and Development, Nonthaburi, Thailand (corresponding author).

²Dept. of Civil, Environmental and Architectural Engineering, Univ. of Colorado at Boulder, Boulder, CO; and, Cooperative Institute for Research in Environmental Sciences, Univ. of Colorado, Boulder, CO.

³Dept. of Earth and Environmental Engineering, Columbia Univ., New York, NY.

Note. This manuscript was submitted on October 13, 2008; approved on March 11, 2010; published online on XXXX XX, XXXX. Discussion period open until February 1, 2011; separate discussions must be submitted for individual papers. This paper is part of the *Journal of Hydrologic Engineering*, Vol. 15, No. 9, September 1, 2010. ©ASCE, ISSN 1084-0699/2010/9-1-XXXX/\$25.00.

71 annual floods. For example, Woo and Waylen (1984) combined
 72 two EVI distributions to model annual floods in British Colum-
 73 bia, Canada. Singh (1987) applied two normal distributions to fit
 74 six annual flood series from rivers in Japan. Jarret and Costa
 75 (1988) combined two LP3 distributions to model floods in Colo-
 76 rado. Webb and Betancourt (1992) combined three LP3 distribu-
 77 tions to fit floods in Arizona. Singh et al. (2005) applied two
 78 Pearson Type 3 distributions to model four flood series from riv-
 79 ers in China.

80 Nonparametric methods, on the other hand, do not assume a
 81 distributional form to the data. Rather, the flood magnitude at any
 82 quantile is estimated by locally smoothing the empirical quantile
 83 function of the data or estimating the PDF using a kernel-based
 84 estimator. Because the method is “local,” in which estimates of
 85 the function at a point are based on data points in its neighbor-
 86 hood, this provides the ability to better capture any arbitrary fea-
 87 tures, especially flood feature that exhibits mixture of population
 88 characteristics and furthermore, easily portable across sites.

89 For the estimation of tail quantiles, an extrapolation rather
 90 than interpolation of the empirical quantile function is needed.
 91 The local estimation procedure inherent in nonparametric flood
 92 frequency analysis translates into a model for tail probability es-
 93 timation. Traditional tail probability estimators consider specific
 94 models of tail behavior whose parameters are to be estimated.
 95 Typically, a threshold beyond which the tail probability model
 96 should be applied also needs to be inferred from the data. Moon
 97 and Lall (1994) demonstrated that kernel-based methods often
 98 performed better in practice than some of the tail probability
 99 models that are commonly used. In this paper, we present a higher
 100 order nonparametric estimation scheme, local polynomial regres-
 101 sion (LPR), which improves further on the kernel quantile esti-
 102 mations presented by Moon and Lall (1994). Other extensions of
 103 this sort of approach to the estimation of nonstationary flood fre-
 104 quency distributions are reported by Sankarasubramanian and
 105 Lall (2003). We focus on heterogeneous flood frequency even
 106 though the LPR method can also be effective in homogeneous
 107 cases.

108 Parametric and nonparametric approaches for heterogeneous
 109 flood frequency analysis are next overviewed. The LPR estimator
 110 used for fitting heterogeneous distribution is then illustrated. We
 111 later compare the performance of the proposed estimator with
 112 those of mixed-distribution model estimators as well as that of
 113 traditional frequency estimator on the same synthetic heteroge-
 114 neous data sets, followed by their comparison on four streamflow
 115 data sets in the Gila River basin of southeast and central Arizona
 116 that exhibited mixed of population characteristics.

117 Background

118 Several mixed-distribution model methods have been developed
 119 based on probability rules, resulting in a more appropriate fit to
 120 the mixture of flood populations. Basically, these methods assume
 121 that flood data are a mixture of two or more populations drawn
 122 from different homogeneous distributions according to their
 123 flood-producing processes, e.g., tropical storm, snowmelt in
 124 spring, and rainfall in summer.

125 Assuming that flood populations X are generated by two inde-
 126 pendent processes and characterized by two distinct distributions,
 127 P_1 and P_2 , Waylen and Woo (1982) then applied the multiplica-
 128 tive rule of probability; the composite probability for magnitude x
 129 flood, P_C , can be estimated by

$$P_C(X \leq x) = P_1(X \leq x)P_2(X \leq x) \quad (1) \quad 130$$

where $P_1(X \leq x)$ and $P_2(X \leq x)$ = probabilities of the two popula- 131
 tions. To obtain P_C the following steps are needed: (1) identify the 132
 two subpopulations and their homogeneous frequency distribu- 133
 tions; (2) estimate parameters of each subpopulation; and (3) 134
 apply Eq. (1) to obtain the probability of various magnitude 135
 floods. Using combination of two EVI distributions, this method 136
 was applied to fit annual floods of rivers in British Columbia 137
 (Waylen and Woo 1982) and northern Ontario, Canada (Woo and 138
 Waylen 1984). 139

Similar to the previous method, X is drawn from two distinct 140
 distributions, F_1 and F_2 . Applying to the additive rule of probabil- 141
 ity, the composite exceedance probability F_C is then estimated by 142

$$F_C(X > x) = F_1(X > x) + F_2(X > x) - F_1(X > x)F_2(X > x) \quad (2) \quad 143$$

To obtain F_C the same steps used in the first method are applied. 144
 This method was implemented with combination of two LP3 dis- 145
 tributions to fit annual flood series of front range in Colorado 146
 (Jarret and Costa 1988) and with combination of three LP3 dis- 147
 tributions to model annual floods in Arizona (Webb and Betan- 148
 court 1992). 149

Fundamentally, Eqs. (1) and (2) are the same and share the 150
 same probability characteristic. It might be proved by substituting 151
 $P(X \leq x) = 1 - F(X > x)$ into Eq. (1). It turns into Eq. (2). Singh 152
 et al. (2005) commented that these methods might overestimate 153
 the frequency distribution. 154

Singh (1968) proposed a curve fitting method by considering 155
 that the annual maximum floods belong to a number of popula- 156
 tions with distinct homogeneous distributions. The composite 157
 probability, P_C , in case of two populations, is estimated by 158

$$P_C(X \leq x) = \alpha P_1(X \leq x) + (1 - \alpha)P_2(X \leq x) \quad (3) \quad 159$$

where α = weight factor relating contribution of each population. 160
 The two distributions, P_1 and P_2 , have means μ_1 and μ_2 and 161
 standard deviations σ_1 and σ_2 , respectively. This method does not 162
 require a priori separation of flood data; however, the curve fitting 163
 requires a priori assumed frequency distributions. Therefore, it 164
 requires the estimation of five parameters, i.e., α , μ_1 , μ_2 , σ_1 , and 165
 σ_2 . Singh (1987) combined two normal distributions to model six 166
 annual flood series in Japan, USSR, Poland, Czechoslovakia, 167
 Italy, and U.S.A. 168

Similar to Eq. (3) the exceedance probability, F_C , can be esti- 169
 mated by 170

$$F_C(X \geq x) = \alpha F_1(X \geq x) + (1 - \alpha)F_2(X \geq x) \quad (4) \quad 171$$

Alila and Mtraoui (2002) used Eq. (4) for combining two LN 172
 distributions to fit the observed flood series in Gila River basin. 173
 They estimated the five parameters (α , μ_1 , μ_2 , σ_1 , and σ_2) by 174
 minimizing objective function $\Sigma(\Delta z)^2$ subject to the constraints 175
 suggested by Cohen (1967) and using a nonlinear optimization 176
 algorithm of Singh and Nakashima (1981). Note that Δz equals 177
 the difference between the observed probability of annual floods 178
 and theoretical probability estimated from mixture of assumed 179
 distributions. The observed probability was obtained from the 180
 Cunnane plotting position formula $[F_T = (n + 0.2)/(m - 0.4)]$, 181
 where T is the return period in years, n is the number of obser- 182
 vations, and m is the rank from the smallest ($m = 1$) to the largest 183
 ($m = n$) observation. 184

Singh et al. (2005) presented a mixed-distribution model 185
 method using conditional probability and multiplication rule. The 186
 composite exceedance probability (F_C) can be estimated by 187

$$F_C(X \geq x) = \sum_{i=1}^s F(A_i)F(X \geq x|A_i) \quad (5)$$

where $F(A_i)$ =probability of annual maximum flood that occurs in seasons i ($i=1, 2, \dots, s$) and s =number of season. $P(x|A_i)$ is conditional frequency probability of flood magnitude x given that annual maximum flood occurs in season i . The method requires a priori both separation and assumed distribution. This method was applied by combining two LP3 distributions to fit four flood data sets from three basins in the United States (Idaho, Louisiana, Arizona) and one basin in China.

While providing more appropriate fit to heterogeneous flood compared to traditional method, the mixed-distribution model methods are complicated as more number of parameters have to be estimated. From the above discussion it can be seen that the mixed model methods have two main drawbacks when applied to observed data: (1) how to determine the number of flood populations (mechanisms) and (2) how to identify homogeneous distribution for each population. In practice, available flood data are short, thus making it difficult to identify the population mixture and also the form of each distribution. Furthermore, no clear relationships between flood mechanisms and their frequency distributions are available for selecting distributions suitably. Therefore, in practice, it is difficult to apply or often not feasible for modeling mixed population floods with the mixed-distribution model methods; consequently, practitioners settle for a suboptimal solution by selecting a single PDF.

213 Nonparametric Approach

Nonparametric flood frequency estimators were developed and studied by Schuster and Yakowitz (1985), Adamowski (1985, 1989), Adamowski and Feluch (1990), and Bardsley (1988, 1989) and subsequently by Lall et al. (1993), Moon et al. (1993), and Moon and Lall (1994) who also compared their performance with several alternatives available at the time. The nonparametric methods are more advantageous than the parametric methods in flood frequency analysis for both annual maximum and partial duration flood series (Adamowski et al. 1998) in that they do not require a priori assumption of the underlying PDF and the estimation is local and data driven, which enables them to capture any distributional features (homogeneous or heterogeneous) exhibited by the data. This will be described in the context of our proposed method in the following section.

Lall et al. (1993) developed a kernel-based quantile estimator, in which a kernel density estimator is used to estimate the probability distribution function and consequently, the quantiles of interest. They also showed that parametric estimates based on the cumulative distribution function are more appropriate than those based on density estimates in the flood frequency context. Estimators based on kernel density are easy to implement; however, they suffer from (1) loss of efficiency of estimation with respect to the true distribution; (2) an uncertain and likely negligible ability to extrapolate beyond the data (Lall et al. 1993); and (3) oversmooth the distribution function. Adamowski (1989) suggested a variable bandwidth kernel density estimator that addresses the extrapolation problem. Later, Moon and Lall (1994) developed a nonparametric kernel-based regression estimator for quantiles. Here, the empirical quantile function is smoothed using a kernel regression estimator. They found that both the density and regression based estimators are competitive compared to other estimators. However, both these methods suffer from

Table 1. Details of Selected Stream Gauges in the Gila River Basin

Number	Station name	USGS number	Data period	Sample size
1	San Francisco River at Clifton	94445000	1891–2005	98
2	San Pedro River near Redington	94720000	1926–1998	67
3	Santa Cruz River at Tucson	94825000	1915–2004	87
4	Salt River near Roosevelt	94985000	1916–2005	81

Note: Missing data exist in record.

boundary problems, i.e., the tail quantiles are biased (Lall et al. 1993; Moon and Lall 1994). Here, we present a LPR (Loader 1999) based estimator that improves on the kernel-based methods in this respect.

LPR Estimator

The LPR estimator is based on fitting of the observed quantile function of the flood data. The quantile function is prescribed through a standard “plotting position formula:” $X_i=i/(n+1)$, where n is the sample size and $i=1, 2, \dots, n$. Makkonen (2008) recently recommended that “the plotting position should be considered not as an estimate but to be equal to $i/(n+1)$.” He also added that the result from this plotting position is unique and independent of the parent distribution.

Given an n -year historical record of annual maximum floods, we can define the observed quantile function through the following set of ordered pairs: (X_i, Y_i) , $i=1, 2, \dots, n$, where $X_i=i/(n+1)$ and Y_i =ranked annual maximum flood data. The X_i are the so-called plotting positions.

Then, we consider a general model for the quantile function as

$$Y_i = \mu(X_i) + \varepsilon_i \quad (6)$$

where $\mu(\cdot)$ =nonlinear function; ε_i =assumed to be identically distributed errors with mean 0 and finite variance; and $X_i \in [0, 1]$. In this context if we consider the estimation of the T year flood, then we are interested in an estimate $\mu(X_T)$ such that $X_T=1-1/T$. The specific proposal here is that $\mu(X_T)$ be estimated using LPR, where we assume that $\mu(X_T)$ is a general function that is continuous and has $(p-1)$ derivatives. Hence, it is reasonable to approximate $\mu(X_T)$ using a local polynomial of order p , following Taylor series arguments. Local here refers to an approximation in the neighborhood of X_T . The size of the neighborhood depends on the smoothness of the target regression function and on the nature of the residual process that generates.

More details of local regression technique are provided by Loader (1999). The estimation algorithm is summarized below:

- For any point of estimate, X_T , nearest neighbors (i.e., nearest data points), $k=(\alpha n)$, are identified, where α varies from 0 to 1 (when $\alpha=1$ then all the data points are neighbors to X_T). The bandwidth $h(X_T)$ of this window of k neighbors around X_T is the distance to the k th neighbor. For tail quantiles, this translates into the number of upper order statistics that are used to fit a polynomial tail quantile model.
- Each of the k data pairs used is then weighted according to the distance to X_T via a weight function (e.g., bisquare,

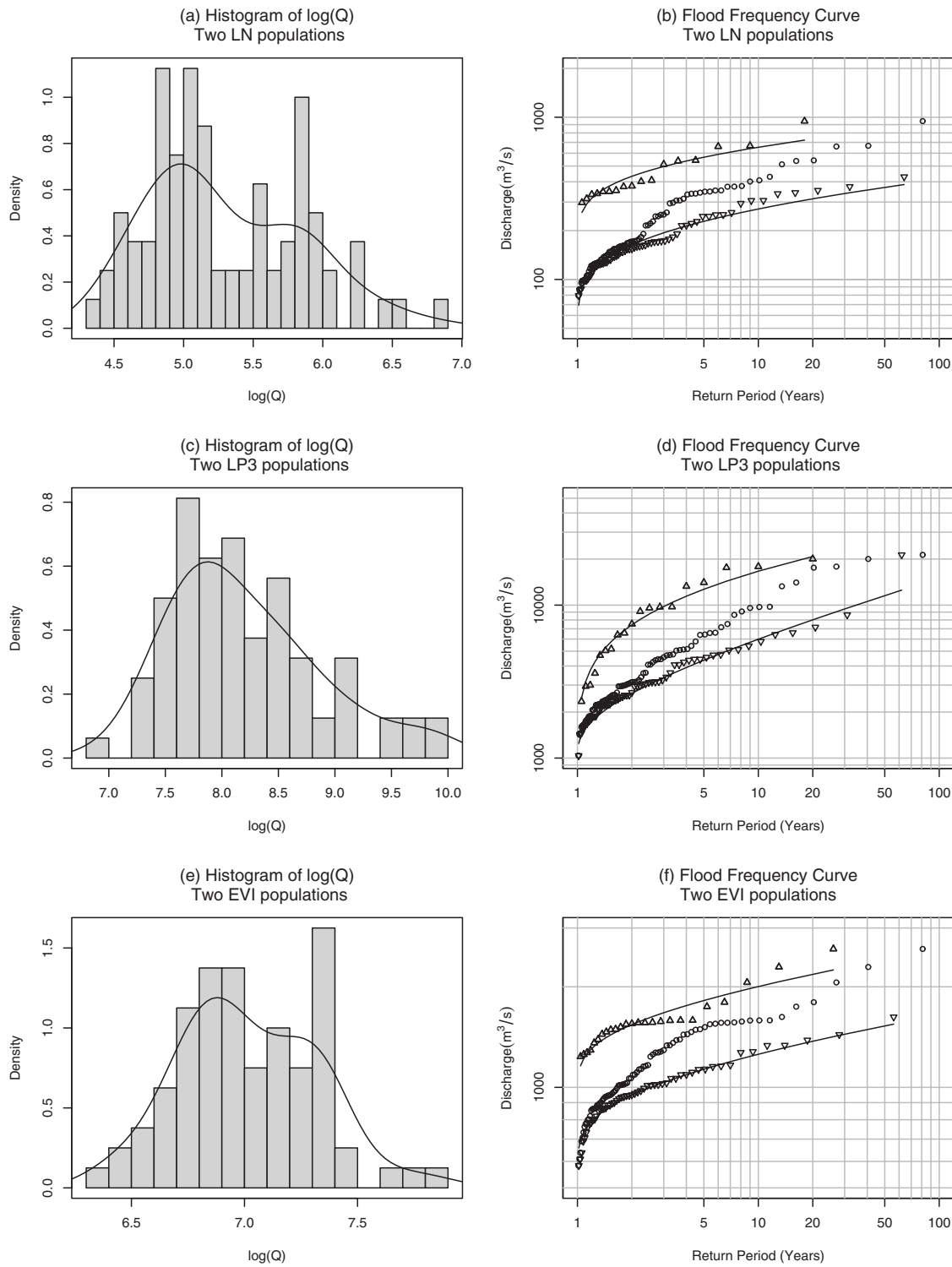


Fig. 1. Synthetic flood data (Q) from a mixture of two distributions; (a) histogram of $\log(Q)$ drawn from two LN distributions and the Kernel density estimated PDF (line) and (b) flood frequency curves of two LN populations (point-up triangles represent flood data drawn from the first LN distribution, point-down triangles from the other, and circles represent flood data drawn from mixture of two LN distributions). Similarly, (c) and (d) are the same as (a) and (b) but for two LP3 while (e) and (f) for two EVI.

290 tricubic, etc.). The bisquare weight is given as $W(u_i)$
 291 $= (15/16)(1-u_i^2)^2$, where $u=(X_i-X_T)/h(X_T)$ and $|u| \leq 1$.
 292 3. Within the smoothing window (i.e., with the k neighbors),
 293 $\mu(X)$ is approximated by a polynomial order p . For example,
 294 a local quadratic model would be

$$\mu(X) = a_0 + a_1(X) + a_2(X)^2 \tag{7} \quad 295$$

The coefficients of the polynomial a_0 , a_1 , and a_2 are obtained 296
 by minimizing the weighted least-squares function 297

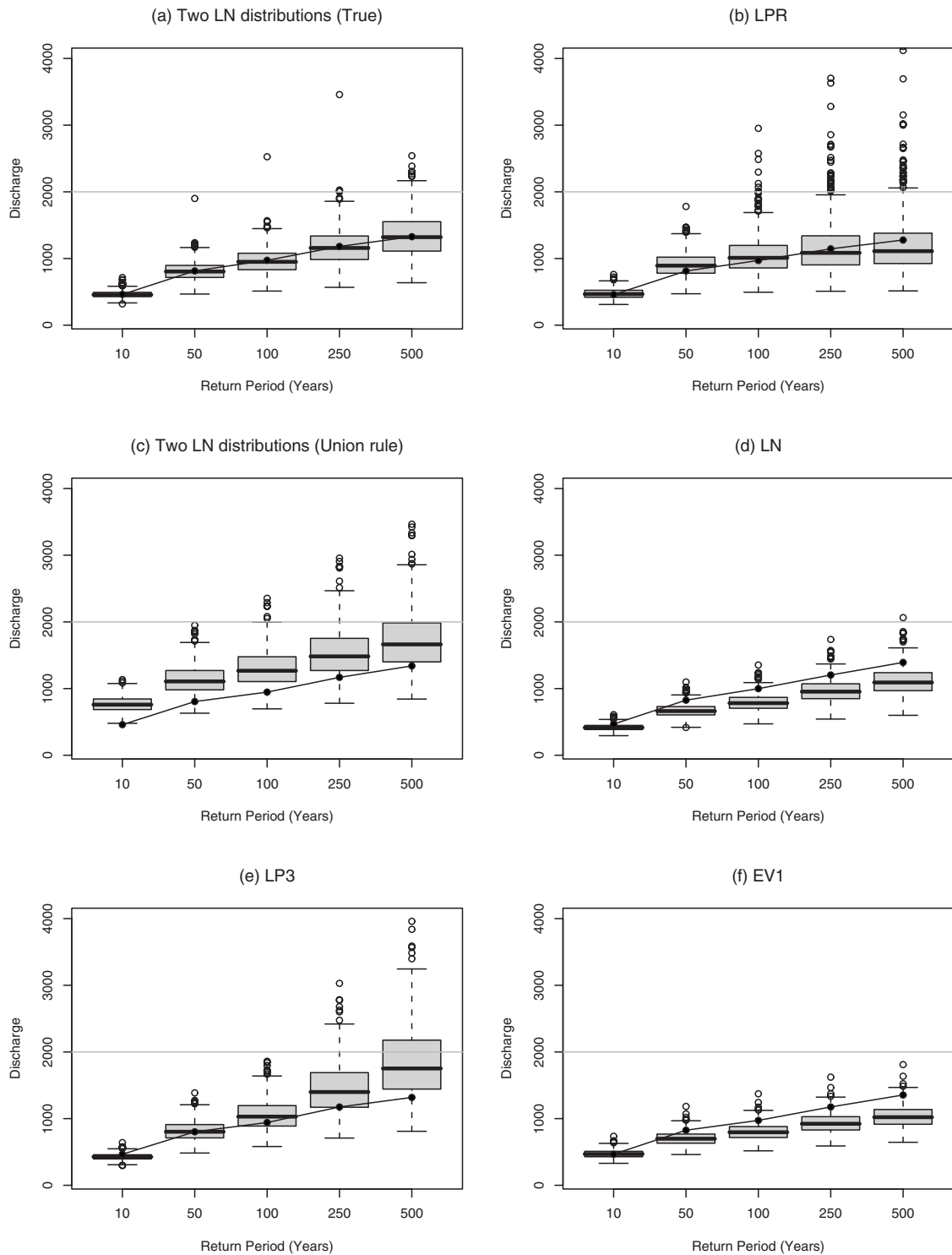


Fig. 2. Boxplots of estimates of 10-, 50-, 100-, 250-, and 500-year return periods of mixture of two LN distributions estimated from (a) true distribution; (b) LPR; (c) mixed-distribution model; (d) LN; (e) LP3; and (f) EVI. The points connected by a solid line in boxplots are the true values. Note that the heights of boxplots represent sampling errors and boxes away from the solid points represent modeling errors.

298

$$\sum_{i=1}^k W_i(X_T)(Y_i - \mu(X_i))^2 \quad (8)$$

299 These steps are repeated for each estimation point.
 300 The key parameters are the optimal number of neighbors k and
 301 the order of polynomial p . These are obtained via minimization of
 302 a generalized cross-validation (GCV) function described below. If

$h(X)$ is too small, insufficient data fall within the smoothing win- 303
 dow; the estimated quantile value will have a very high variance. 304
 On the other hand, if $h(X)$ is too large, the quantile estimate may 305
 have a large bias. Therefore, the bandwidth must be chosen to 306
 compromise this bias-variance trade-off. Similar to the band- 307
 width, the degree of the local polynomial p also affects the bias- 308
 variance trade-off. 309

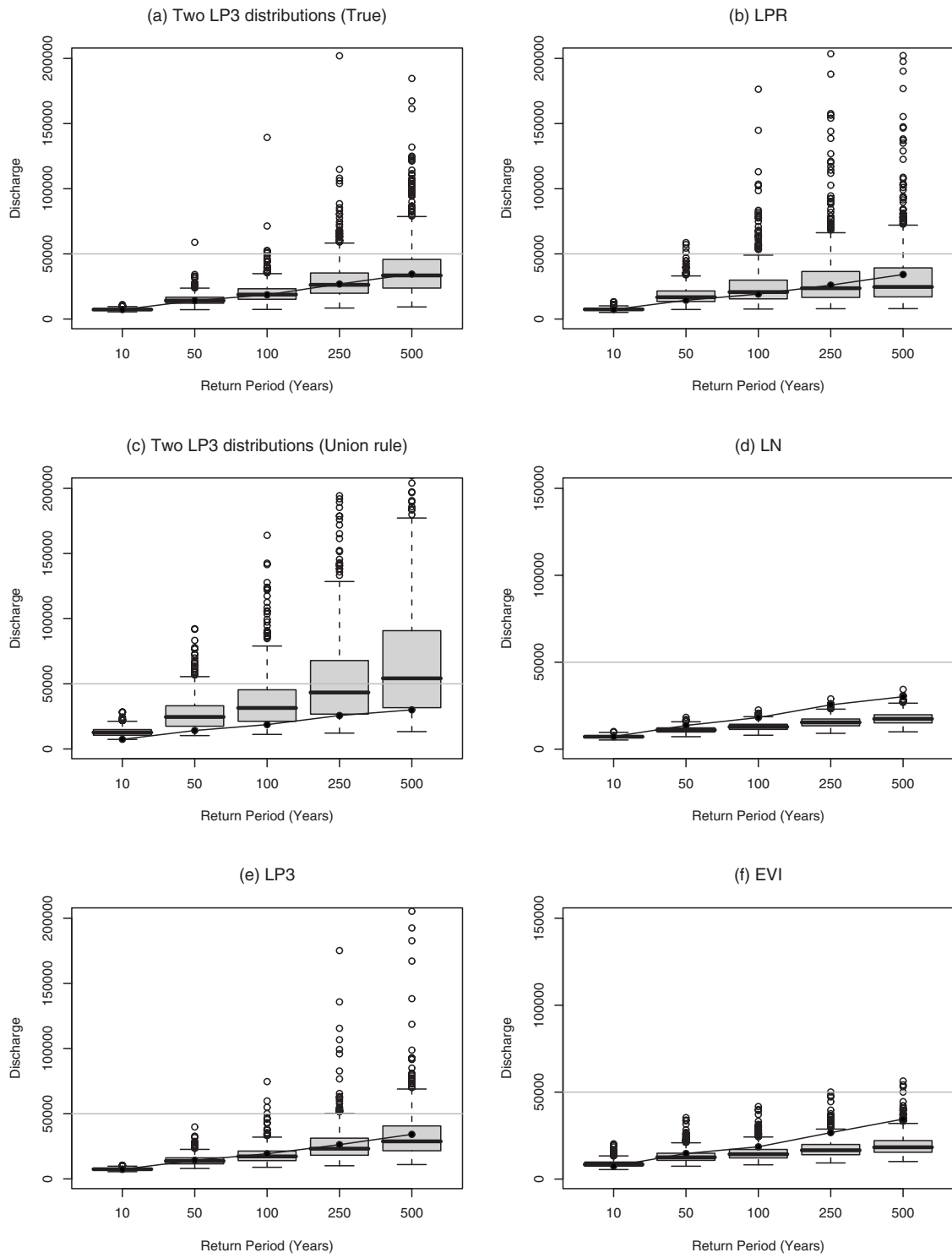


Fig. 3. Same as Fig. 2 but synthetic data from a mixture of two LP3 distributions

310 Loader (1999) suggested that it often suffices to choose a low
 311 order polynomial and concentrate on choosing the bandwidth to
 312 obtain a satisfactory fit. Typically, in parametric regression, mean
 313 squared error is used to assess the performance of the fit. How-
 314 ever, this is a poor indicator of future performance of the model
 315 (i.e., predictive error). Craven and Wahba (1978) developed the
 316 GCV, similar to Akaike information criteria and Bayesian infor-
 317 mation criteria, which approximates predictive risk

$$GCV(\alpha, p) = n \sum_{i=1}^n (Y_i - \hat{\mu}(X_i))^2 / \left(1 - \sum_{i=1}^n h_{ii} \right)^2 \quad (9) \quad 318$$

where n =sample size; $Y_i - \hat{\mu}(X_i)$ =residual; and h_{ii} =diagonal 319
 terms of the hat matrix H . The hat matrix can be estimated using 320
 standard linear regression procedures. For fairly small data sets 321

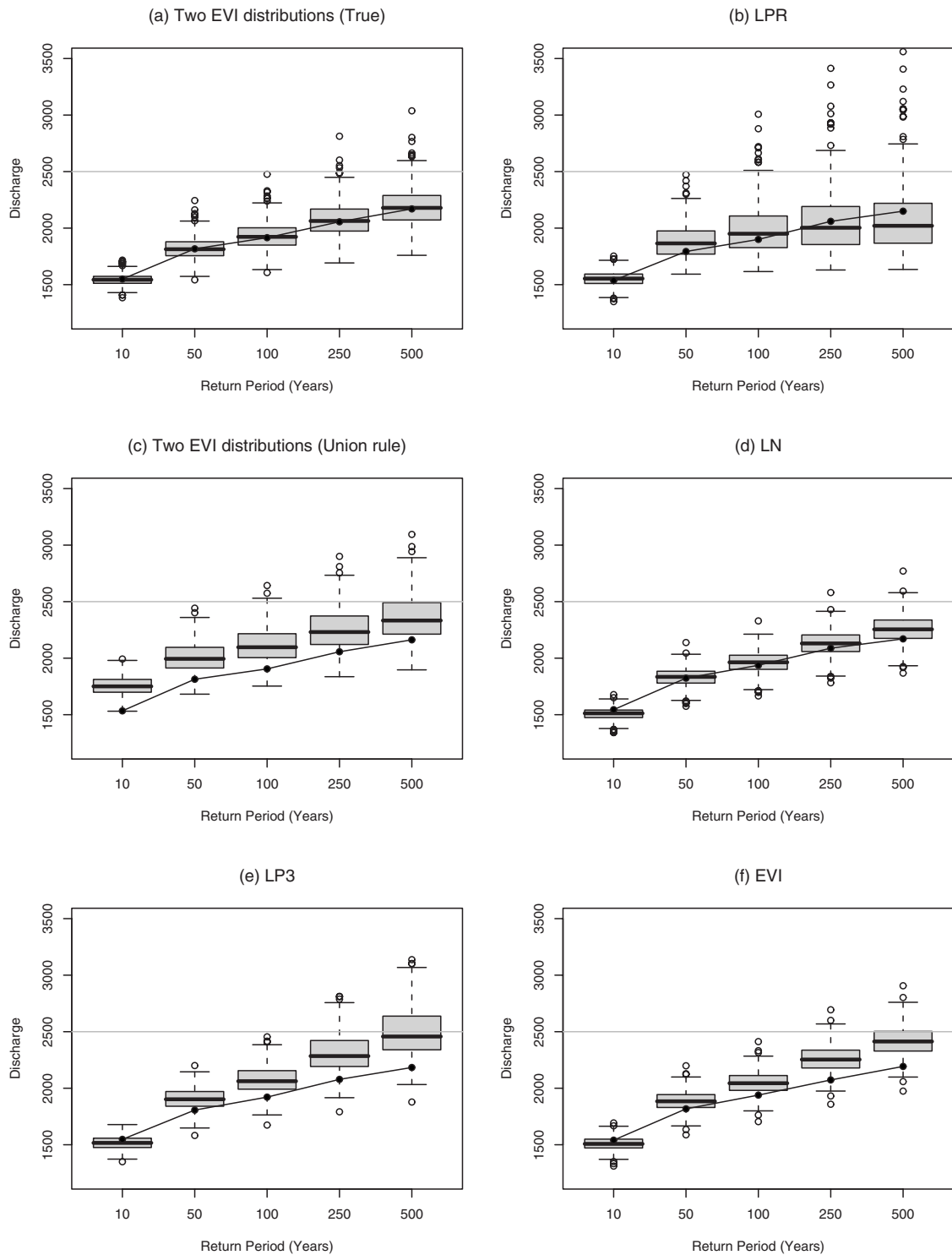


Fig. 4. Same as Fig. 2 but synthetic data from a mixture of two EVI distributions

322 Loader (1999) suggested the use of the leave-one-out cross-
 323 validation (CV) function

$$CV(\alpha, p) = \frac{1}{n} \sum_{i=1}^n (Y_i - \hat{\mu}_{-i}(X_i))^2 \quad (10)$$

324
 325 where $\hat{\mu}_{-i}(X_i)$ denotes the leave- X_i -out estimate of $\hat{\mu}(X_i)$. That is,
 326 each X_i is removed from the data set in turn and the local regres-
 327 sion estimate computed from the remaining $n-1$ data points.

Applications

The performance of LPR quantile estimator was tested on a suite 329
 of synthetic heterogeneous data sets that composed of a mixture 330
 of two conventional distribution populations and also compared 331
 with those of mixed model methods and traditional models. We 332
 then applied the LPR estimator on four streamflow data sets that 333
 exhibit mixed population characteristics. 334

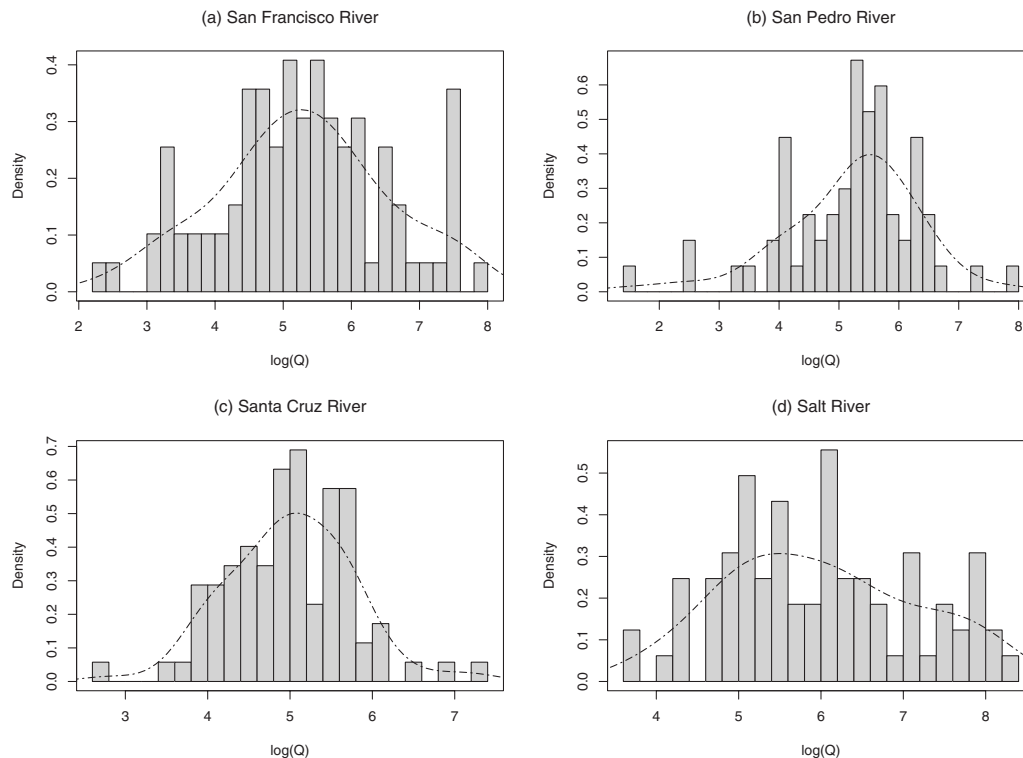


Fig. 5. Histogram of logarithm of annual peak flow data [$\log(Q)$] for (a) San Francisco River; (b) San Pedro River; (c) Santa Cruz River; and (d) Salt River. The dashed lines are the PDFs estimated using kernel density estimators.

335 **Synthetic Heterogeneous Flood Distribution** 336 **Experiments**

337 To simulate synthetic heterogeneous flood data sets, three sets of
338 combined probability distribution models as “parents” for at-site
339 flood generation processes were considered. Consequently,
340 we generated 500 samples of size 80 each from Eq. (3) using
341 different combined probability distributions. The combinations
342 were (1) two LN distributions with mean $\mu_{y1}=5.0$, standard de-
343 viation $\sigma_{y1}=0.4$, and $\mu_{y2}=6.0$, $\sigma_{y2}=0.5$; (2) two LP3 distributions
344 with $\mu_{y1}=3.5$, $\sigma_{y1}=0.2$, coefficient of skewness, $\gamma_{y1}=0.4$,
345 and $\mu_{y2}=3.7$, $\sigma_{y2}=0.3$, and $\gamma_{y2}=0.5$; and (3) two EVI distribu-
346 tions with $\mu_{x1}=1,500$, $\sigma_{x1}=200$ and $\mu_{x2}=1,000$, $\sigma_{x2}=200$, where
347 $y=\log(x)$. All combinations used weight factor, $\alpha=0.25$. The pa-
348 rameters of the distribution and the weight factor of combination
349 were selected to provide distributions that are well discriminated
350 and a clear mixture distribution. The appropriate mixed-
351 distribution model estimators were used to estimate five quantiles
352 (10-year, 50-, 100-, 250-, and 500-year return periods) across the
353 suite of parents. The LPR estimator was applied to each of the
354 generated samples and five quantiles were estimated. Three con-
355 ventional PDFs—LN, LP3, and EVI—were also applied for esti-
356 mating the quantiles. The quantile estimates are displayed as
357 boxplots along with the true values from the parent.
358 Our hypothesis is that the proposed nonparametric method will
359 be competitive against the parametric alternatives. The proposed
360 method as described in the previous section requires neither the
361 flood-producing mechanism information in basin nor a priori dis-
362 tribution for each flood mechanism. In contrast, the mixed model
363 method requires data of both flood mechanisms and flood distri-
364 butions, which often are deficient. Furthermore, the traditional
365 single PDF estimators cannot recognize the mixture of two distri-
366 bution populations in the synthetic flood data resulting in inap-
367 propriate fits. We test our hypothesis by examining the model

error (bias) and sampling error (variance) of each estimator from 368
the true quantiles of the parent combined distributions. 369

370 **Observed Data**

371 The LPR estimator was subsequently applied to annual maximum 371
flood series from four stream gauges in the Gila River basin of 372
central and southern Arizona. Several researchers have described 373
various causes of heterogeneity in the flood data of this basin 374
(Webb and Betancourt 1992; Yarnal and Diaz 1986; Reyes and 375
Cadet 1988; Hjalmarson 1990). Lately, Alila and Mtraoui (2002) 376
concluded that heterogeneity in the flood data of this basin results 377
from three potential causes (1) different types of storms, i.e., 378
monsoonal storm, frontal storm, and tropical cyclone, (2) El Niño 379
southern oscillation (ENSO) conditions, during ENSO years 380
precipitation and its variability are enhanced, and/or (3) decadal 381
climatic fluctuations, shifts in the climate of the southwestern 382
U.S.A. around 1,930 and 1,960 were driven by decadal scale vari- 383
ability. 384

385 The four stream gauges are on four tributaries: (1) San Fran- 385
cisco River; (2) San Pedro River; (3) Santa Cruz River; and (4) 386
Salt River. The details of each gauge such as station name, USGS 387
station number, data period, and sample size are shown in Table 1. 388
Lacking information of flood-producing mechanisms for separat- 389
ing the flood data into subpopulations, thus, applying the mixed- 390
distribution model to these observed flood data was not possible. 391

392 Bootstrap based confidence intervals were also computed for 392
each quantile estimates. In this, we generated 1,000 bootstrap 393
samples and the quantiles are estimated for each sample using the 394
LPR method—the 5th and 95th percentiles of the estimates from 395
the bootstrap samples provide the 90% confidence intervals for 396
the quantile estimates. The bootstrap approach is nonparametric 397
in that no distributional assumption of the estimator is required to 398

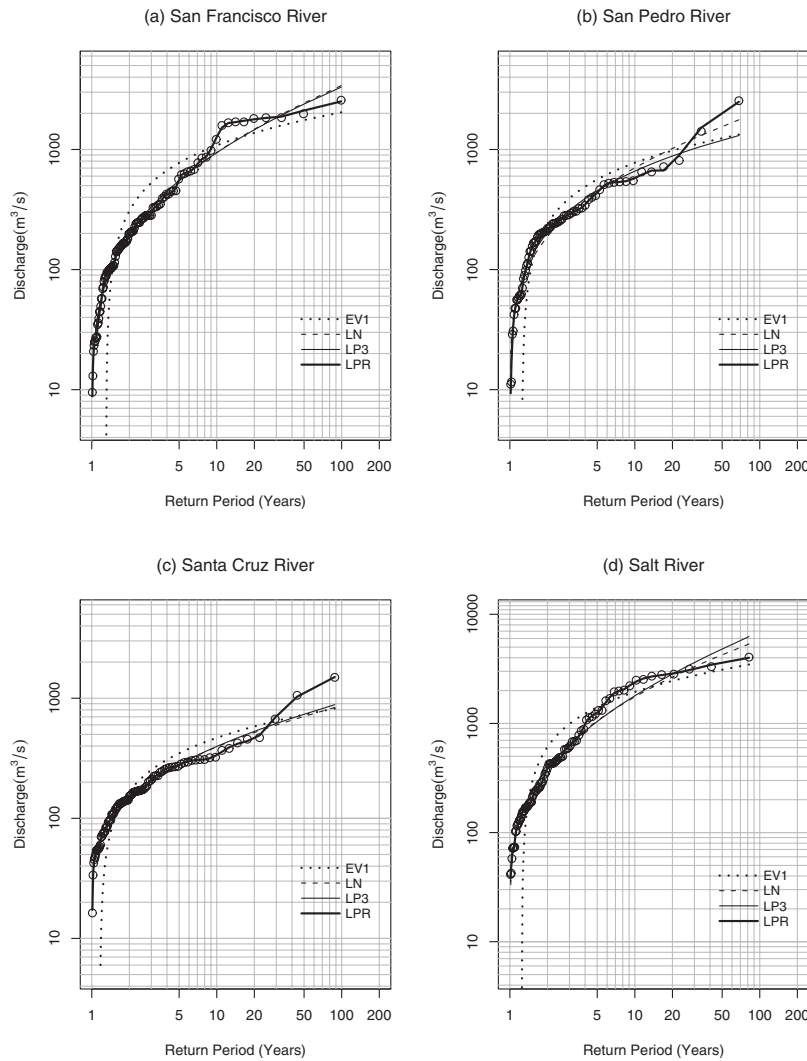


Fig. 6. Quantile estimates from EV1, LN, LP3, and LPR estimators for (a) San Francisco River; (b) San Pedro River; (c) Santa Cruz River; and (d) Salt River. Circles represent empirical quantiles and lines represent estimated quantiles.

399 generate surrogate independent and identically distributed
400 samples of the same length as the original data by sampling with
401 replacement.

402 Finally, the predictive capability of the LPR method was tested
403 in a leave-one-out CV mode. In this, a flood observation is
404 dropped from the data set and the LPR estimators on the re-
405 maining data are used to estimate the quantile of the dropped
406 observation; this is repeated for all the observations. The results
407 from the synthetic and observed data are described in the follow-
408 ing section.

409 Results

410 Synthetic Data

411 The synthetic mixed data series from all parent combined distri-
412 butions show their mixed population characteristics by their bi-
413 modal distributions and their dog-leg shaped flood frequency
414 curves (Fig. 1 shows one of the 500 samples). Potter (1958) sug-
415 gested that the dog-leg shape of flood frequency curve indicates
416 the mixture of two populations.

417 As mentioned earlier, the LPR estimator and the parametric

estimators are applied to each mixed synthetic data and we esti- 418
mate the 10-, 50-, 100-, 250-, and 500-year return period magni- 419
tudes. These estimates from the simulations are shown as 420
boxplots along with the true values as solid points connected with 421
a solid line (Figs. 2–4). Box sizes provide the variance of the 422
estimates resulting from sampling error, while departures of true 423
values from median of the estimates (horizontal line in box) pro- 424
vide the bias of the estimates resulting from modeling error. The 425
bias is acceptable if the true value falls within the box where 50% 426
of estimate values (between 25th and 75th quartile) are contained, 427
which can be considered as 50% confidence interval. 428

The LPR estimator exhibits good performance for all the par- 429
ent distributions [Figs. 2(b), 3(b), and 4(b)]. The variance of the 430
estimates from LPR increases (bigger boxes) as the return period 431
increases—more so for return periods of 250 and 500 years. This 432
is to be expected from standard regression theory, as LPR ex- 433
trapolates beyond the range of the data at higher return periods 434
and, hence, has larger variance and larger bias. However, the true 435
values are still within the boxes. This shows the ability of the 436
LPR estimator to recognize the mixed population characteristic. 437

Of course, the mixed-distribution model [Eq. 3 of Singh 438
(1968)] with true distributions performs better than LPR estimator 439

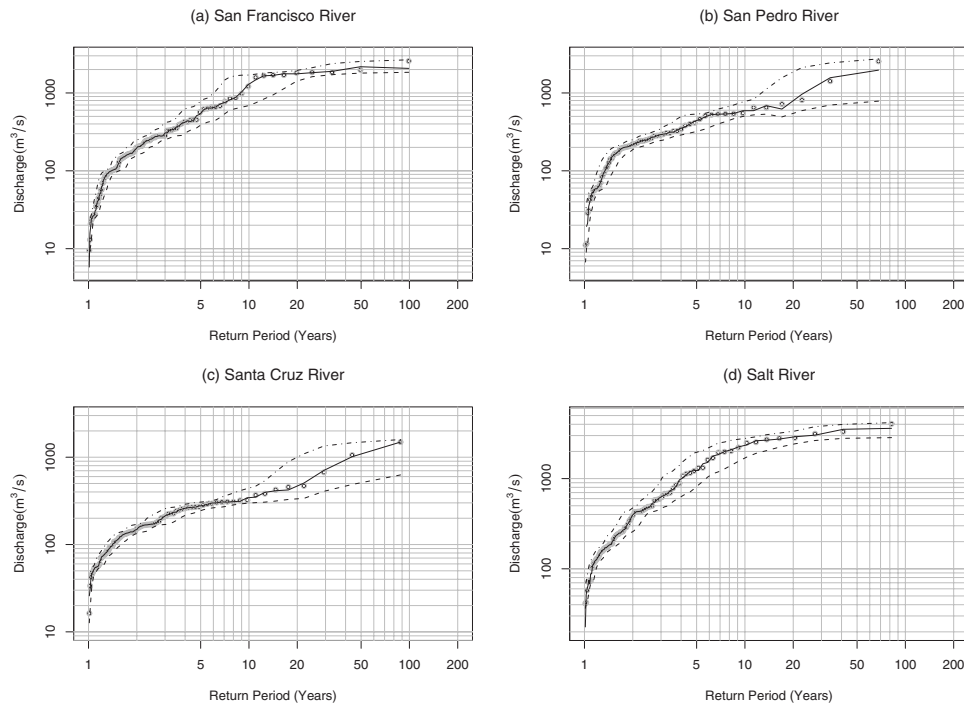


Fig. 7. Leave-one-out quantile estimates (solid lines) and 90% confidence intervals (dashed and dotted-dashed lines) for (a) San Francisco River; (b) San Pedro River; (c) Santa Cruz River; and (d) Salt River

440 does. As expected, the true mixed model estimates show no bias
 441 [Figs. 2(a), 3(a), and 4(a)] while LPR estimates show some bias
 442 for 250- and 500-year return period magnitudes, but their perfor-
 443 mances on variance are comparable.

444 None of the traditional parametric homogeneous distributions
 445 i.e., LN, LP3, and EVI, perform well on the parent distributions
 446 due to their inability in recognizing the mixed populations. They
 447 either underestimate or overestimate the upper tail quantiles, i.e.,
 448 beyond 100-year return periods and have unacceptable bias (true
 449 value falls outside box) for 250- and 500-year return periods
 450 [Figs. 2(d), 3(d), and 4(d) for LN, Figs. 2(e), 3(e), and 4(e) for
 451 LP3, and Figs. 2(f), 3(f), and 4(f) for EVI]. Surprisingly, LP3
 452 performs well for the two LP3 mixed distribution [Fig. 3(e)] and
 453 LN performs well for the two EVI mixed distribution [Fig. 4(d)].

454 The performance of the LPR estimator is quite competitive
 455 with the true underlying model estimates. Similar results were
 456 seen with synthetic data generated from homogeneous distribu-
 457 tions (Apipattanavis et al. 2003). Given that in practice it is often
 458 not feasible to obtain the true mixture distributions and mixed-
 459 distribution models such as Eq. 1 of Waylen and Woo (1982) tend
 460 to overestimate for all parent distributions (Singh et al. 2005), the
 461 LPR method with its data-driven feature is very attractive.

462 Observed Data

463 Unlike the synthetic data, annual flood series from the four
 464 gauges in the Gila River basin do not show clearly their multi-
 465 modal characteristics (Fig. 5) due to the short record length re-
 466 lated to their three flood-producing mechanisms (Alila and
 467 Mtiraoui 2002). However, their flood frequency curves show the
 468 dog-leg shape indicating their mixed populations (see their em-
 469 pirical quantiles shown as circles in Fig. 6).

470 The LPR estimators closely follow and smooth the empirical
 471 quantiles of annual flood data of all four rivers for all return
 472 periods (shown as solid lines in Fig. 6). The traditional methods

(LN, LP3, and EVI estimators), on the other hand, do not capture
 the quantile features present in the data. The LPR estimator pro-
 vides an appropriate fit to mixed flood data obviating the tough
 task of identifying and separating the data into homogeneous
 flood populations.

The leave-one-out cross-validated quantile estimates (Fig. 7)
 appear to be within the 90% confidence interval (obtained from
 the bootstrap approach) for all four sites. Residuals from the
 cross-validated estimates were found to be normally distributed
 with no significant autocorrelation (figures are not shown) indi-
 cating the goodness of the LPR model. The quantile estimates for
 250- and 500-year return periods along with the 90% confidence
 intervals from the bootstrap approach, at all sites are shown in
 Fig. 8. It can be seen that the confidence intervals are asymmetric,
 unlike the symmetric intervals, from traditional approaches.

Summary and Conclusions

A nonparametric flood frequency estimator based on LPR was
 developed and applied on synthetic and historical data sets, then
 analyzed the flood quantiles of return periods up to 500 years. The
 method performs a local regression on the empirical quantiles to
 smooth them and also to extrapolate in the tails. The local aspect
 of the estimation provides the ability to capture any arbitrary
 features that might present in the data. Thus, it is particularly
 suited for estimating quantiles from mixture of flood populations.
 Traditional parametric mixed-distribution-model methods require
 the knowledge of the generating mechanisms and frequency distri-
 bution for each mechanism. In practice, it is often not feasible
 to identify the distributions from short data sets. Unlike the para-
 metric counterparts, the LPR estimator requires no prior assump-
 tion of the underlying distribution, which makes it portable across
 sites. This also improves upon kernel-based nonparametric esti-

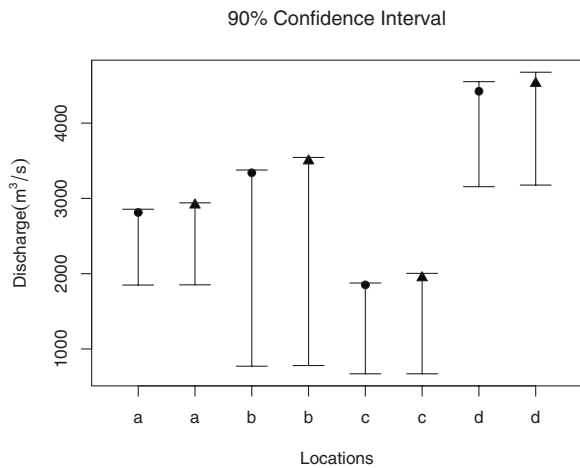


Fig. 8. Ninety percent confidence intervals for 250- and 500-year return periods (shown as circles with whiskers and point-up triangles with whiskers, respectively) for (a) San Francisco River; (b) San Pedro River; (c) Santa Cruz River; and (d) Salt River

504 mators developed in the past in that it is easy to alleviate the
 505 boundary problems that plague the kernel estimators. The LPR
 506 estimator showed good performance on a variety of synthetic data
 507 sets from mixed population characteristics and also on observed
 508 flood data. The method offers an attractive alternative. In situation
 509 in which traditional models can be well identified and are well
 510 suited, we suggest using them, otherwise, the LPR estimator is a
 511 good alternative. Our aim in this research is to offer a simple and
 512 flexible framework that can supplement and compliment the tra-
 513 ditional methods. There is emerging research on estimating flood
 514 frequency conditioned on large-scale ocean-atmospheric informa-
 515 tion using semiparametric approaches (Sankarasubramanian and
 516 Lall 2003) and Bayesian methods (Lima and Lall 2010). These
 517 estimations can provide flood frequency information for each year
 518 that will be of immense help for water resources planning. Gen-
 519 eralized extreme value distributions have also been used for esti-
 520 mating flood quantiles using covariates (Katz et al. 2002).
 521 Furthermore, these approaches are useful in estimating extreme
 522 events under changing climate (Towler et al. 2010). The LPR
 523 estimator proposed here can be easily adapted for the same pur-
 524 pose (Apipattanavis 2007) and it could also be extended for re-
 525 gional flood frequency analysis.

526 Acknowledgments

527 The first writer thanks Royal Irrigation Department, Thailand, for
 528 their support. Partial support for the second and third writers from
 529 National Science Foundation through Grant No. EAR 9973125 is
 530 thankfully acknowledged. We thank the reviewers and the associ-
 531 ate editor for insightful comments that greatly improved the
 532 paper.

533 References

534 Adamowski, K. (1985). "Nonparametric kernels estimation of flood fre-
 535 quencies." *Water Resour. Res.*, 21(11), 1585–1590.
 536 Adamowski, K. (1989). "A Monte Carlo comparison of parametric and
 537 nonparametric estimation of flood frequencies." *J. Hydrol.*, 108, 295–
 538 308.

Adamowski, K., and Feluch, W. (1990). "Nonparametric flood-frequency
 analysis with historical information." *J. Hydraul. Eng.*, 116(8), 1035–
 1047. 539
 540
 Adamowski, K., Liang, G., and Patry, G. G. (1998). "Annual maxima and
 partial duration flood series analysis by parametric and non-
 parametric methods." *Hydrolog. Process.*, 12, 1685–1699. 542
 543
 Alila, Y., and Mtiraoui, A. (2002). "Implications of heterogeneous flood-
 frequency distributions on traditional stream-discharge prediction
 techniques." *Hydrolog. Process.*, 16, 1065–1084. 545
 546
 Apipattanavis, S. (2007). "Stochastic nonparametric methods for multi-
 site weather generation and flood frequency estimation applications to
 construction delay, hydrology and agricultural modeling." Ph.D. the-
 sis, Civil, Environmental, and Architectural Engineering, Univ. of
 Colorado, Boulder, Colo. 548
 549
 Apipattanavis, S., Rajagopalan, B., and Lall, U. (2003). "Local regression
 quantile estimate for flood frequency analysis." *Hydrology days*,
 American Geophysical Union, Fort Collins, Colo. 553
 554
 Bardsley, W. E. (1988). "Toward a general procedure for analysis of
 extreme random events in the Earth sciences." *Math. Geol.*, 20(5),
 513–528. 557
 558
 Bardsley, W. E. (1989). "Using historical data in nonparametric flood
 estimation." *J. Hydrol.*, 108, 249–255. 559
 560
 Chow, V. T., Maidment, D. R., and Mays, L. W. (1988). *Applied hydrology*,
 Vol. 572, McGraw-Hill, New York. 561
 562
 Cohen, A. C. (1967). "Estimation in mixture of two normal distribution."
Technometrics, 9(1), 15–28. 563
 564
 Craven, P., and Wahba, G. (1978). "Smoothing noisy data with spline
 functions." *Numer. Math.*, 31, 377–403. 565
 566
 Hjalmanson, H. W. (1990). "Flood of October 1983 and history of flood-
 ing along the San Francisco River, Clifton, Arizona." *USGS Water
 Resources Rep. No. 85-4225-B*. 567
 568
 Hosking, J. R. M. (1990). "L-moment: Analysis and estimation of distri-
 butions using linear combinations of order statistics." *J. R. Stat. Soc.
 Ser. B (Methodol.)*, 52(2), 105–124. 570
 571
 Interagency Advisory Committee on Water Data (IACWD). (1982).
*Guidelines for determining flood flow frequency: Bulletin 17B of the
 Hydrology Subcommittee*, USGS, Reston, Va. 574
 575
 Jain, S., and Lall, U. (2000). "The magnitude and timing of annual maxi-
 mum floods: Trends and large-scale climatic associations for the
 Blacksmith Fork River, Utah." *Water Resour. Res.*, 36(12), 3641–
 3651. 576
 577
 Jain, S., and Lall, U. (2001). "Floods in a changing climate: Does the past
 represent the future?" *Water Resour. Res.*, 37(12), 3193–3205. 580
 581
 Jarret, R. D., and Costa, J. E. (1988). "Evaluation of the flood hydrology
 in the Colorado front range using precipitation, streamflow, and paleo-
 flood data for the Big Thompson River basin." *USGS Water Re-
 sources Investigations Rep. No. 87-4177*. 582
 583
 Katz, R. W., Parlange, M. B., and Naveau, P. (2002). "Statistics of ex-
 tremes in hydrology." *Adv. Water Resour.*, 25, 1287–1304. 584
 585
 Kite, G. W. (1977). *Frequency and risk analyses in hydrology*, Water
 Resources Publications, Fort Collins, Colo. 588
 589
 Lall, U., Moon, Y.-I., and Bosworth, K. (1993). "Kernel flood frequency
 estimators: Bandwidth selection and kernel choice." *Water Resour.
 Res.*, 29(4), 1003–1015. 590
 591
 Lima, C. H. R., and Lall, U. (2010). "Spatial scaling in a changing
 climate: A hierarchical Bayesian model for nonstationary multi-site
 annual maximum and monthly streamflow." *J. Hydrol.*, 383, 307–
 318. 593
 594
 Loader, C. (1999). *Local regression and likelihood*, Vol. 290, Springer,
 New York. 597
 598
 Makkonen, L. (2008). "Extreme value analysis and order statistics, bring-
 ing closure to the plotting position controversy." *Commun. Stat:
 Theory Meth.*, 37, 460–467. 599
 600
 Moon, Y.-I., and Lall, U. (1994). "Kernel function estimator for flood
 frequency analysis." *Water Resour. Res.*, 30(11), 3095–3103. 602
 603
 Moon, Y.-I., Lall, U., and Bosworth, K. (1993). "A comparison of tail
 probability estimators." *J. Hydrol.*, 151, 343–363. 604
 605
 Potter, W. D. (1958). "Upper and lower frequency curves for peak rates
 606

- 607 of runoff." *Trans., Am. Geophys. Union*, 39(1), 100–105.
- 608 Reyes, S., and Cadet, D. L. (1988). "The southwest branch of the North
609 American monsoon during 1979." *Mon. Weather Rev.*, 116, 1175–
610 1187.
- 611 Sankarasubramanian, A., and Lall, U. (2003). "Flood quantiles in a
612 changing climate: Seasonal forecasts and causal relations." *Water Re-
613 sour. Res.*, 39(5), 1134.
- 614 Schuster, E., and Yakowitz, S. (1985). "Parametric/nonparametric mix-
615 ture density estimation with application to flood-frequency analysis." *Water
616 Resources Research Bulletin*, 21(5), 797–803.
- 617 Singh, K. P. (1968). "Hydrologic distributions resulting from mixed
618 populations and their computer simulation." *International Association
619 of Scientific Hydrology Publications*, 81, 671–681.
- 620 Singh, K. P. (1987). "A versatile flood frequency methodology." *Water
621 Int.*, 12(3), 139–145.
- 622 Singh, K. P., and Nakashima, M. (1981). "A new methodology for flood
623 frequency analysis with objective detection and modification of
624 outliers/inliers." *Rep. No. 272*, Illinois State Water Survey, Cham-
625 paign, Ill.
- 626 Singh, V. P., Wang, S. X., and Zhang, L. (2005). "Frequency analysis of
627 nonidentically distributed hydrologic flood data." *J. Hydrol.*, 307,
628 175–195.
- Towler, E., Rajagopalan, B., Gilleland, E., Summers, R., Yates, D., and
629 Katz, R. (2010). "Modeling hydrologic and water quality extremes in
630 a changing climate." *Water Resour. Res.*, ■, ■–■, in press. 631 AQ:
#10
- Vogel, R. M. (1986). "The probability plot correlation coefficient test for
632 the normal, lognormal, and Gumbel distribution hypothesis." *Water
633 Resour. Res.*, 22(4), 587–590. 634
- Vogel, R. M., and McMartin, D. E. (1991). "Probability plot goodness-
635 of-fit and skewness estimation procedures for the Pearson type 3 dis-
636 tribution." *Water Resour. Res.*, 27(12), 3149–3158. 637
- Waylen, P. R., and Woo, M. K. (1982). "Prediction of annual floods
638 generated by mixed process." *Water Resour. Res.*, 18(4), 1283–1286. 639
- Webb, R. H., and Betancourt, J. L. (1992). "Climatic variability and flood
640 frequency of the Santa Cruz River, Pima County, Arizona." *U.S. Geol.
641 Surv. Water Supply Pap.*, ■, 2379–■. 642 AQ:
#11
- Woltemade, C. J., and Potter, K. W. (1994). "A watershed modeling
643 analysis of fluvial geomorphologic influences on flood peak attenua-
644 tion." *Water Resour. Res.*, 30(6), 1933–1942. 645
- Woo, M. K., and Waylen, P. R. (1984). "Areal prediction of annual floods
646 generated by two distinct processes." *Hydrol. Sci. J.*, 29(1), 75–88. 647
- Yarnal, B., and Diaz, H. F. (1986). "Relationships between extremes of
648 the southern oscillation and the winter climate of the Anglo-American
649 Pacific Coast." *J. Climatol.*, 6, 197–219. 650

AUTHOR QUERIES — 005009QHE

- #1 Au: Please check that ASCE Membership affiliations are provided for all authors that are members.
- #2 Au: Please check affiliation 1 and supply authors' title/role in affiliations 1-3.
- #3 Au: Please check setting of “i” as subscript throughout.
- #4 Au: Please check changes in Apipattanavis et al. (2003).
- #5 Au: Please verify author in Bardsley (1988, 1989).
- #6 Au: Please verify year in Craven and Wahba (1978).
- #7 Au: Please supply location in Hjalmarson (1990) and Jarret and Costa (1988).
- #8 Au: Please check change in Interagency Advisory Committee on Water Data (IACWD) (1982).
- #9 Au: Please check our update of Lima and Lall (2010).
- #10 Au: Please update Towler et al. (2010), if possible.
- #11 Au: Please supply volume and last page in Webb and Betancourt (1992).