Local Polynomial Technique for Flood Frequency Analysis

Somkiat Apipattanavis[1,2], Balaji Rajagopalan[1,3] and Upmanu Lall[4]

[1] Department of Civil, Environmental and Architectural Engineering, University of Colorado at Boulder, Boulder, USA
[2] Royal Irrigation Department, Bangkok, Thailand
[3] Co-operative Institute for Research in Environmental Sciences, University of Colorado, Boulder, USA
[4] Department of Earth and Env. Engg., Columbia University, New York, NY

Abstract

Nonparametric estimators of at-site flood frequency using annual maximum flow data present an alternative to methods that a priori assume a specific probability distribution function. They approximate a wide class of distribution functions. Past work in this direction using kernel density and quantile estimators is extended here using a higher order approximation technique, locally weighted polynomial regression, for estimating the quantile function. An empirical investigation of the performance of this method relative to selected alternatives and for selected target distributions is presented here.

Introduction

Flood frequency analysis entails relating the magnitude of annual maximum flood events to their frequency of occurrence at a gauged site. The typical interest in estimating extreme flood quantiles i.e. - 100-year or 500-year flood, from a small number of observations (~50 to 100 years) for the magnitude of the design of hydraulic structures such as dams, culverts and bridges. Traditional parametric methods for this problem assume that the annual maximum floods are independent and identically distributed and drawn from a 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, USGS and widely used in practice are Log-Pearson Type III, Log-Normal and Extreme Value distributions (see Kite, 1977, Chow et al., 1988). There are statistical tests to discriminate between choices of distributions including L-moments based (see Kite, 1977; Hosking, 1990; Vogel and McMartin, 1991; Vogel 1986). However, often, it is difficult to discriminate between candidate models for a given data set, and the 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.

Nonparametric methods, on the other hand, do not assume a distributional form to the data. Rather, the flood magnitude at any quantile is estimated by locally smoothing the empirical quantile function of the data or estimating the pdf using a kernel based estimator. Because the method is "local", in that estimates of the function at a point are based on data points in its neighborhood – this provides the ability to better capture an arbitrary features exhibited by the data and furthermore, easily portable across sites. For the estimation of tail quantiles, an extrapolation rather than interpolation of the empirical quantile function is needed. The "local" estimation procedure inherent in nonparametric flood frequency analysis translates into a model for tail probability estimation. Traditional tail probability estimators consider specific models of tail behavior whose parameters are to be estimated. Typically, a threshold beyond which the tail probability model should be applied also needs to be inferred from the data. Moon et al (1994) demonstrated that kernel based methods often performed better in practice than some of the tail probability models that are commonly used. In this paper, we present a higher order nonparametric estimation scheme that improves further on the kernel quantile estimations presented by Moon et al (1994).

Nonparametric flood frequency estimators were developed and studied by Schuster and Yakowitz (1985), Adamowski (1985, 1989), Adamowski and Feluch (1990), Bardley (1988, 1989) and more recently Lall et al. (1993), Moon et al. (1993) and Moon and Lall (1994). Lall et al. (1993) developed a kernel based quantile estimator, where in, 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 distribution function are more appropriate than those based on density estimates in the flood frequency context. Kernel density based estimators while easy to implement, 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 find that both, the density and regression based estimators are competitive to other estimators. However, both these nonparametric quantile estimators suffer from boundary problems, i.e., the tail quantiles are biased (Lall et al., 1993; Moon and Lall, 1994). Here, we present a local polynomial (Loader, 1999) based estimator that improves upon the kernel regression estimator.

The local polynomial estimator is first described. We then compare the performance of this estimator with traditional parametric estimators on a suite of synthetic data set, followed by their comparison on two streamflow data sets.

Local Polynomial Estimator

Given an n-year historical record of annual maximum floods, we can define the empirical quantile function through the following set of ordered pairs: (X_i, Y_i), $i = 1,2,\ldots,n$ where $X_i = (i-0.25)/(n+0.50)$, Y_i = ranked annual maximum flood data (in this study, we use log-transformed Y_i). The X_i are the so-called plotting positions, and one can use any other formula of interest for the purpose. Here Adamowski's (1981) formula is used.

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

$$Y_i = \mu(X_i) + \varepsilon_i \tag{1}$$

where $\mu(.)$ is a nonlinear function, ε_i are assumed to be identically distributed errors with mean 0 and finite variance, and it is understood that $X_i \in [0,1]$.

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 advanced here is that $\mu(X_T)$ be estimated using locally weighted polynomial regression, 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, ε_i.

For details as to the specific local polynomial estimation method (LOCFIT) used here see Loader (1999). The estimation algorithm is summarized below:

1. For any point of estimate, $X_T, k(=\alpha n)$, nearest neighbors (i.e. nearest data points) are identified, where a 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.
2. Each of the *k* data pairs used is then weighted according to the distance to *T* via a weight function (e.g. Bisquare, Tricubic etc.). The Bisquare weight is given as $W(u_i) = 15/16(1-u_i^2)^2$, where $u = (X_i - X_T)/h(X_T)$, and $|u| \leq 1$.
3. Within the smoothing window (i.e. with the *k* neighbors), $\mu(X)$ is approximated by a polynomial order *p*. For example, a local quadratic model would be

$$\mu(X) = a_0 + a_1(X) + a_2(X)^2 \qquad (2)$$

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

$$\sum_{i=1}^{k} W_i(X_T)(Y_i - \mu(X_i))^2 \qquad (3)$$

These steps are repeated for each estimation point.

The key parameters identify are the optimal number of neighbors *k* and the order of polynomial *p*. These are obtained via minimization of a Generalized Cross Validation (GCV) function described below. If *h(X)* is too small, insufficient data fall within the smoothing window, the local polynomial fits the data very noisy. The resulting regression will have large variance. On the other hand, if *h(X)* is too large, the local polynomial may not fit the data well within the smoothing window, and important features of the mean function $\mu(X)$ may be distorted i.e., the model will have a large bias. Therefore, the bandwidth must be chosen to compromise this bias-variance trade-off. Similar to the bandwidth, the degree of the local polynomial *p*, affects the bias-variance trade-off. A higher polynomial degree may provide a better approximation to the target function $\mu(X)$ than a low polynomial degree. Thus, fitting a high degree polynomial will usually lead to an estimate $\mu(X)$ with less bias. But high order polynomial has large numbers of coefficients to estimate, and the result is higher variability of the estimate.

It often suffices to choose a low order polynomial and concentrate on choosing the bandwidth to obtain a satisfactory fit. Typically, in parametric regression, mean squared error is used to assess the performance of the fit. However, this is a poor indicator of future performance of the model (i.e. predictive error). Craven and Wahba (1979) developed a measure called the GCV (similar to the AIC or BIC) that approximates predictive risk.

$$GCV(\alpha,p)=n\frac{\sum_{i=1}^{n}\left(Y_i-\hat{\mu}(X_i)\right)^2}{\left(1-\sum_{i=1}^{n}h_{ii}\right)^2} \tag{4}$$

where *n* is the sample size, $Y_i-\hat{\mu}(X_i)$ is the residual and h_{ii} are the diagonal terms of the hat matrix H. The hat matrix can be estimated using standard linear regression procedures –typically, it is $X\left(X^T X\right)^{-1} X^T$. For fairly small datasets Loader (1999) suggests the use of the cross validation (CV) function:

$$CV(\alpha, p) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\mu}_{-i}(X_i)\right)^2 \tag{5}$$

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

Loader (1999) also developed an approximate confidence interval for the estimates from the local polynomial, assuming the residuals to be normally distributed locally (within the neighborhood of *k* data points). Approximate confidence interval for the true mean is

$$I(X) = \left(\hat{\mu}(X) - c\hat{\sigma}\|l(X)\|, \hat{\mu}(X) + c\hat{\sigma}\|l(X)\|\right) \tag{6}$$

where $\hat{\mu}(X)$ is an unbiased estimate of $\mu(X)$, *c* is the appropriate quantile of the standard normal distribution, $\hat{\sigma}$ is an estimate of the residual standard deviation, and $\|l(X)\|$ is the variance reducing factor that measures the reduction in variance due to local regression. Usually, the reduction factor decreases as the bandwidth increases. Thus, a prediction interval has limits

$$\hat{\mu}(X_{new}) \pm c\hat{\sigma}\left(1 + \|l(X)\|^2\right)^{1/2} \tag{7}$$

Note that prediction intervals assume normality: If X_{new} is not normally distributed, the prediction interval will not be correct, even asymptotically.

Applications

We tested the LOCFIT quantile estimator on a suite of synthetic data sets and two streamflow data sets. We also compared with traditional parametric estimators.

Synthetic experiments

To simulate the “choice” of models that a practitioner may face, we considered a set of probability distribution models as “parents” for the at-site flood generation process, and similarly for the estimation of quantiles. The LOCFIT procedure is considered as an alternative for estimation across the suite of “parents”. The setting is of interest where a public regulatory agency may have mandated as a “best practice”, the use of a specific distributional model across all enterprises. This has been the case in the U.S., since the Bulletin 17, USWRC procedures (USWRC, 1981) were adopted. Our hypothesis is that the nonparametric procedure will be competitive against parametric alternatives, where a mix of parent populations may be appropriate across the country or region. We test this hypothesis by examining the success or lack thereof versus the proper specification and mis-specification of the parametric model. Consequently, we generated 500 samples of size 75 each from the following parent populations:

1) Log-Normal: $f(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp\left[-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right]$, $x > 0$ where $y = \log x$,

   $\mu_y = \bar{y},\ \sigma_y = s_y$

2) Log-Pearson type III: $f(x) = \frac{\lambda^{\beta}(x\text{-}\varepsilon)^{\beta\text{-}1}e^{\text{-}\lambda(\mathrm{x}\text{-}\varepsilon)}}{\mathrm{x}\Gamma(\beta)}$, $\log x \geq \varepsilon$ where

   $y = \log x$, $\lambda = s_y/\sqrt{\beta}$, $\beta = [2/C_s(y)]$, $\varepsilon = \bar{y} - s_y\sqrt{\beta}$ (assuming $C_s(y)$ is positive)

3) Extreme Value Type I: $f(x) = \frac{1}{\alpha}\exp\left[-\frac{x-u}{\alpha} - \exp\left(-\frac{x-u}{\alpha}\right)\right]$,

   $-\infty < x < \infty$ where $\alpha = \sqrt{6}s_x/\pi$, $u = \bar{x} - 0.5772\alpha$

4) Mixture of Normal: $f(x) = c_1 N(\mu_1, \sigma_1) + c_2 N(\mu_2, \sigma_2)$, $-\infty < x < \infty$ where c_1 and c_2 are weighted constant and c_1+ c_2=1.

In each case we compare the performance of LOCFIT for selected tail quantiles relative to properly and improperly specified parametric models. The above distributions were chosen for the synthetic experiments because they from the exponential family-which is widely considered in practice.

Real Data

We also applied the LOCFIT estimator to two streamflow data sets (i) Annual maximum flow on the Santa Cruz at Tucson, Arizona for the period 1915-2000 and (ii) 3-day annual maximum flow on the American river at Folsom dam, CA for the period 1905-2001. The LOCFIT estimate at these sites are compared to Log-Normal (LN), Log-Pearson type III (LPIII), Extreme Value I distribution (EVI). Confidence intervals for the traditional methods were also computed (Chow et al., 1988) for estimates at selected return periods.

Results

Synthetic Data

The LOCFIT estimator and the parametric estimators are applied to each synthetic data and we estimated the 10-, 50-, 100-, 250- and 500-year return period magnitudes. These estimates from the simulations are shown as boxplots along with the true values as a solid line. Box sizes provide the variance of the estimates.

It can be seen from Figures 1-4 that the LOCFIT estimator exhibits good performance for all the tested distributions. The variance of the estimates from LOCFIT increases (bigger boxes) as the return period increases – more so for return periods 250 and 500 years. This is to be expected from standard regression theory, as

LOCFIT extrapolates beyond the range of the data at higher return periods and hence, increased variance. It performs well especially with the mixture distribution (Figure 4) when compared with other estimators. LPIII estimator is also good for all the distributions with relatively smaller variance in comparison to LOCFIT. This is because, LPIII performs very well for exponential family of distributions, as is the case with the distributions tested here. However, for the mixture distribution (Figure 4) the variances from LPIII are higher than those from LOCFIT. LN estimator works well for simulations from EVI and LN distributions (Figure 1 and 2), but performs badly in all other cases. The EVI estimator exhibits poor performance in all the distributions except of course, the simulations from EVI.

Real Data

The LOCFIT estimator and the traditional methods are applied to the Santa Cruz annual maximum flow data. As shown in Figure 5, the LOCFIT estimates closely follow and smooth the empirical quantiles (shown as circles) in the tails. The parametric methods fit well at lower return periods, but for higher return periods they grossly underestimate relative to the observed quantiles, with EVI performing especially poorly. The cross-validated quantile estimates (Figure 5b) appear to be within the 90% confidence interval (obtained from equation 7). Residuals from cross-validated estimates are normally distributed and show no significant lag-1 autocorrelation (Figures 5c and d) indicating the goodness of the LOCFIT model. Furthermore, the LOCFIT estimator has a smaller confidence interval compared to the other methods (Table 1). These findings are consistent with those seen in Moon and Lall (1994) on the same data set for the period 1915-1986.

In the case of American River as well, the LOCFIT estimator follows the empirical quantiles quite well, while parametric methods perform poorly in the tails (Figure 6a). Also the Log-Normal estimator appears to overestimate the higher return period flows and EVI seems to underestimate. Here too the confidence intervals from the LOCFIT estimator are tighter than those from the parametric models (Table 2). The residuals from the cross-validated estimates show a Gaussian structure and significant autocorrelations (Figure 6c and d). This indicates that an iterated estimation may be required (Loader, 1999). Similar results are found when applying these methods to the data from the 1905-1945 (Figures 7a and b) and 1946-2001 (Figures 7c and d) sub periods (Tables 3 and 4). These sub periods were chosen because the Folsom dam was built in 1945. The Folsom dam was a case study of a recent National Research Council report (Potter et al., 1999) - where they suggest climate variability as being a reason for increased annual maximum flows in the latter sub period.

Summary

Locally weighted polynomial regression technique, a nonparametric approach, is applied to flood frequency estimation. The estimation is "local" and therefore, has the ability to capture any smooth distribution that generated the data. Unlike its parametric counterparts, no prior assumption of the underlying distribution is required, which makes it portable across sites. This also improves upon kernel-based nonparametric estimators developed in the past. Good performance on a variety of synthetic and real data sets is observed. Multivariate extensions of this approach to forecasting regional flood quantiles conditioned on large-scale ocean-atmospheric information are underway.

Acknowledgements

The first author thanks Thai Government for his graduate student fellowship and also to Royal Irrigation Department, Thailand for their support. Partial support for the second and third authors from National Science Foundation through grant EAR 9973125 is thankfully acknowledged.

References

Adamowski, K., Plotting formula for flood frequency, *Water Resources Research Bulletin, 17*(2), 197-202, 1981.

Adamowski, K., Nonparametric kernels estimation of flood frequencies, *Water Resources Research, 21*(11), 1585-1590, 1985.

Adamowski, K., A Monte Carlo comparison of parametric and nonparametric estimation of flood frequencies, *Journal of Hydrology, 108*, 295-308, 1989.

Adamowski, K., and W. Feluch, Nonparametric flood-frequency analysis with historical information, *Journal of Hydraulic Engineering, 116*(8), 1035-1047, 1990.

Adamowski, K., and C. Labaliuk, Estimation of flood frequencies by a nonparametric density procedure, in Hydrologic Frequency Modeling, edited by V.P. Singh, p.p. 97-106, D. Reidel, Norwell, Mass., 1987.

Bardsley, W. E., Using historical data in nonparametric flood estimation, Journal of Hydrology, 108, 249-255, 1989.

Chow, V. T., D. R. Maidment and L. W. Mays, *Applied Hydrology*, 572 p.p., McGraw Hill, New York, 1988.

Lall, U. and J. Niu, Variable bandwidth kernel density estimation (abstract), *Eos Trans. AGU, 70*, 324, 1989.

Lall, U., Y.-I. Moon and K. Bosworth, Kernel Flood Frequency Estimators: Bandwidth Selection and Kernel Choice, *Water Resources Research*, 29(4), 1003-1015, 1993.

Loader, C., *Local Regression and Likelihood*, 290 p.p., Springer, New York 1999.

Moon, Y.-I., U. Lall, and K. Bosworth, A comparison of tail probability estimators, *Journal of Hydrology, 151*, 343-363, 1993.

Moon, Y.-I., and U. Lall, Kernel function estimator for flood frequency analysis, *Water Resources Research*, 30(11), 3095-3103, 1994.

Potter, K. et al., Committee on American River Flood Frequencies, *Improving American River Flood Frequency Analysis*, Water Science and Technology Board, Commission on Geosciences, Environment, and Resources National Research Council, National Academy Press, 120 p.p., Washington, D.C., 1999.

Schuster, E., and S. Yakowitz, Parametric/nonparametric mixture density estimation with application to flood-frequency analysis, *Water Resources Research Bulletin, 21*(5), 797-803, 1985.

U.S. Water Resources Council, Guidelines for determining flood flow frequency, *Bulletin 17*, Wasington, D.C.: U.S. Government Printinig Office.

Webb R. H. and J. L. Betancourt, Climatic variability and flood frequency of the Santa Cruz River, Pima County, Arizona, U.S. Geological Survey, *Water Supply Paper; 2379*, 1992.

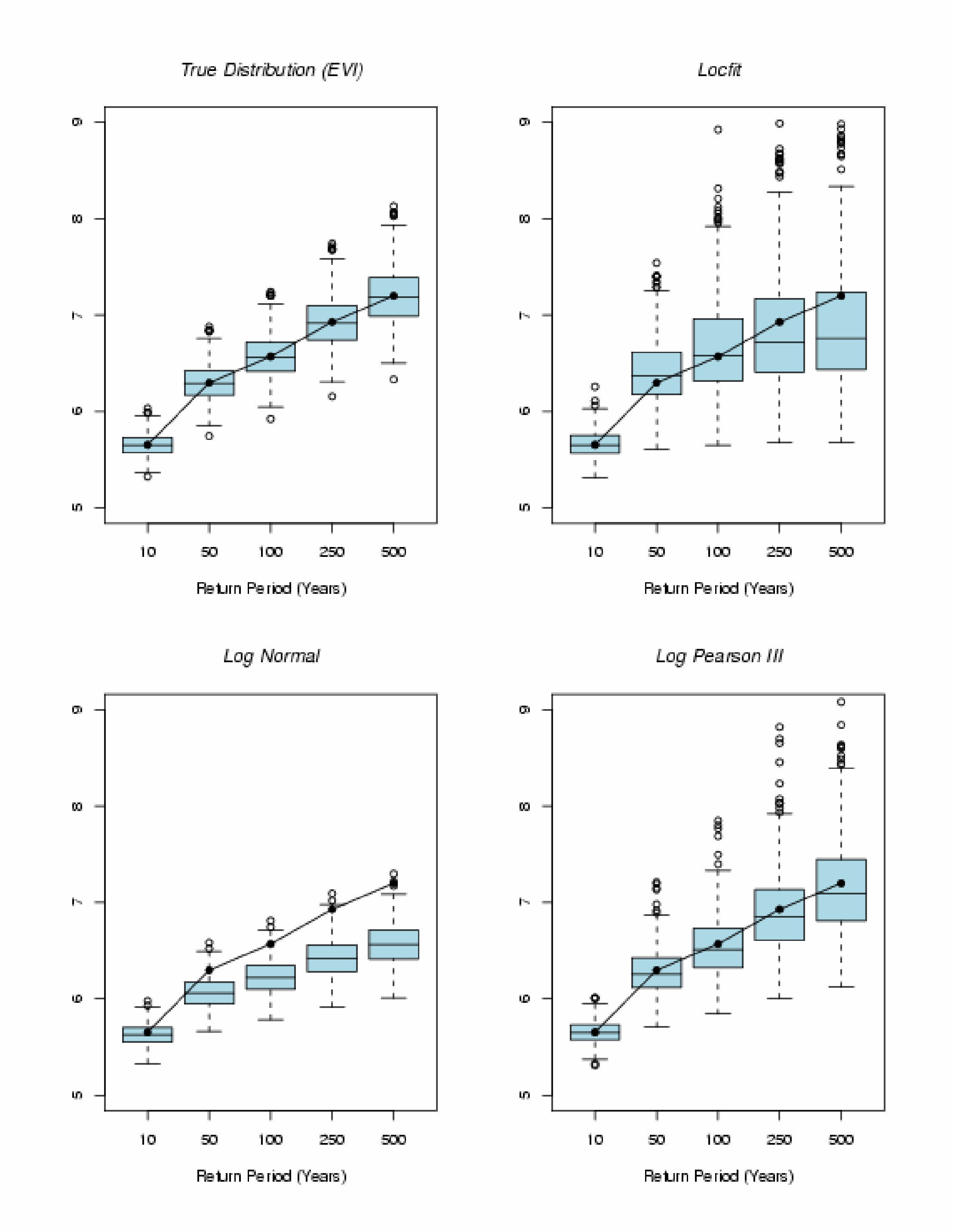

Figure 1: Boxplots of estimates of 10, 50, 100, 250 and 500-year return period of data from EVI distribution. The estimates are from (i) True distribution (EVI), (ii) locfit (iii) Log Normal,and (iv) Log Pearson III estimators. The solid lines in all the figures join the true value.

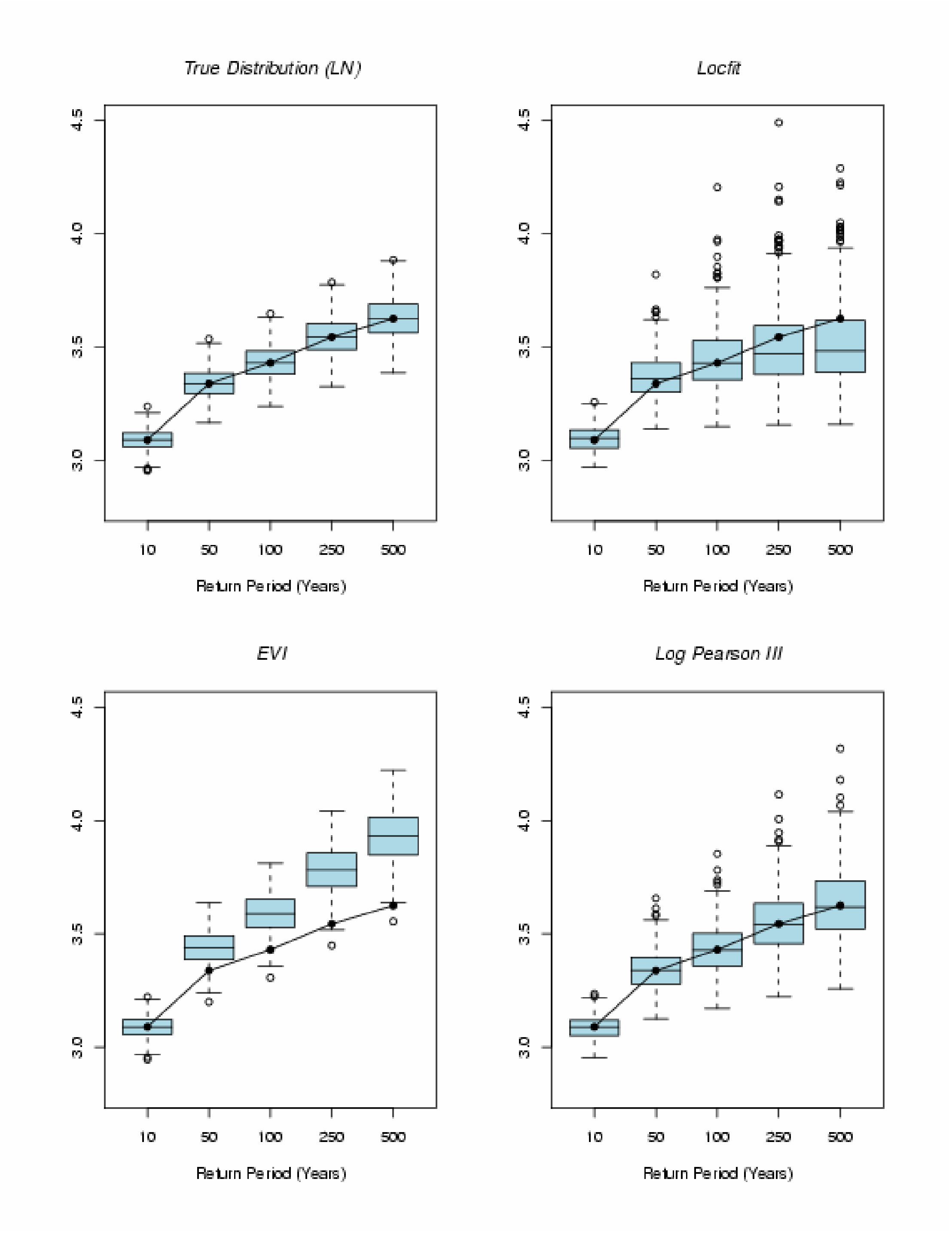

Figure 2: Same as Figure 1, but for data from Log Normal distribution.

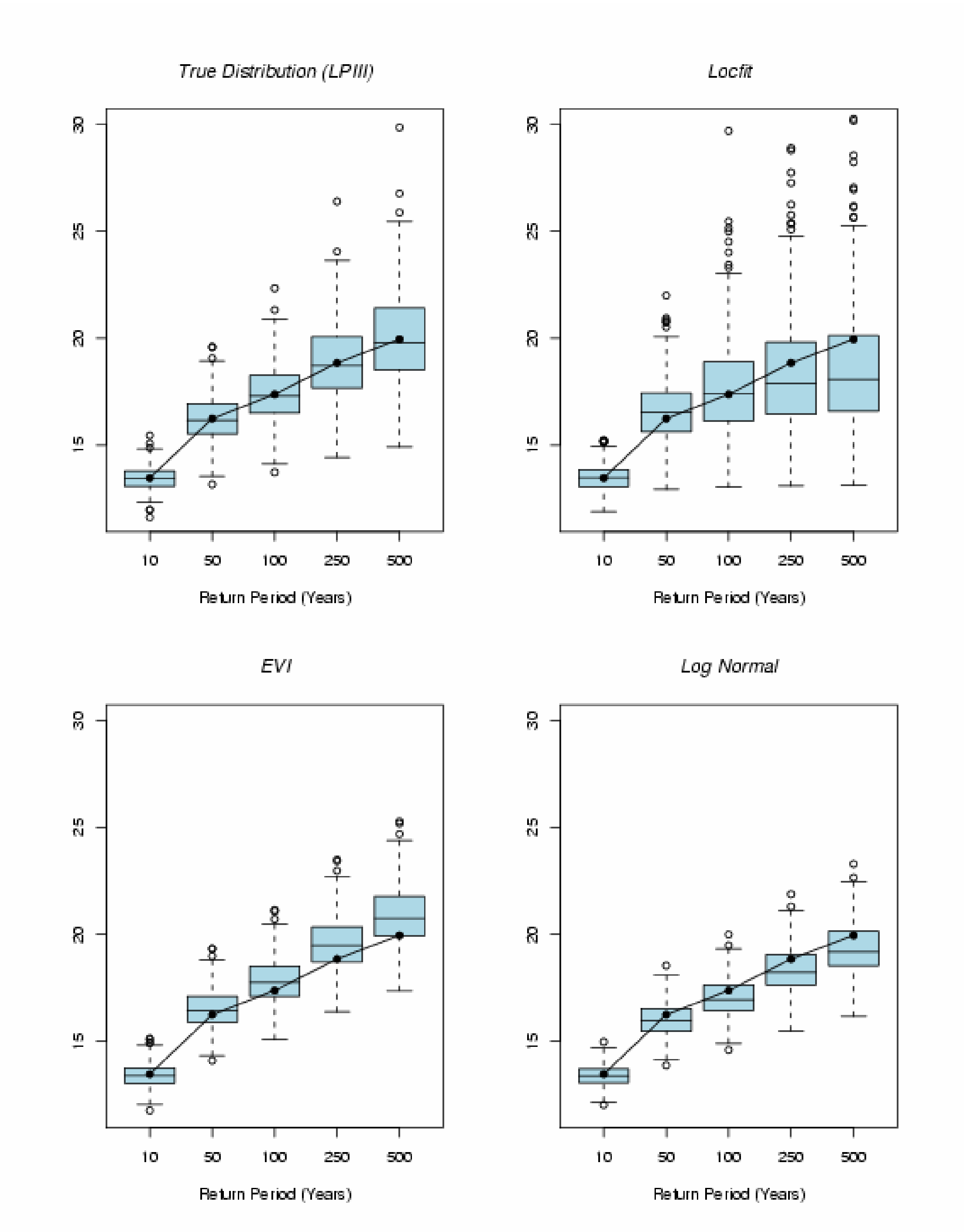

Figure 3: Same as Figure 1, but for data from Log Pearson III distribution

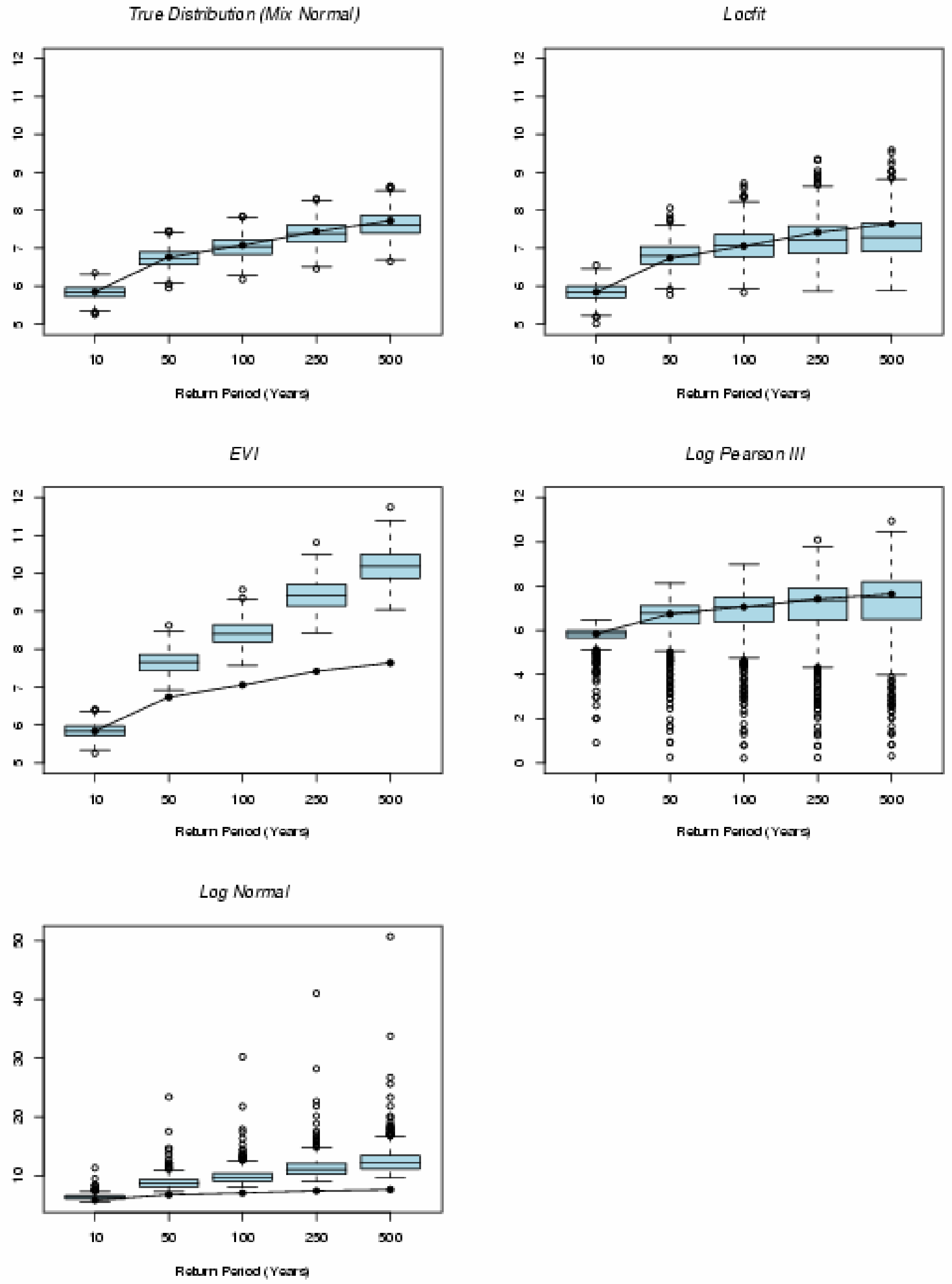

Figure 4: Same as Figure 1, but for data from Mixture Normal distribution

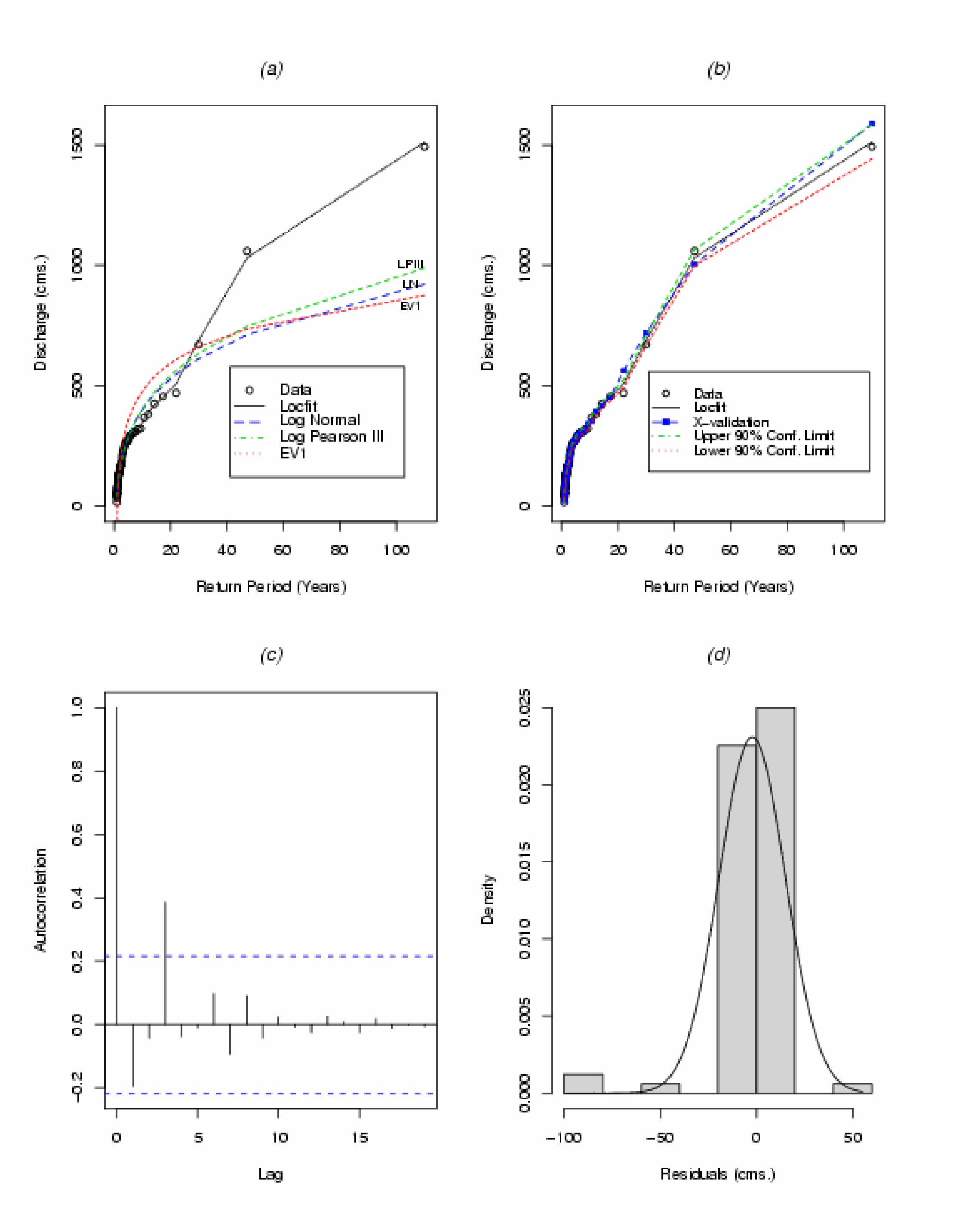

Figure 5: Quantile function estimates and x-validation for Santa Cruz, AZ (1915-2000).

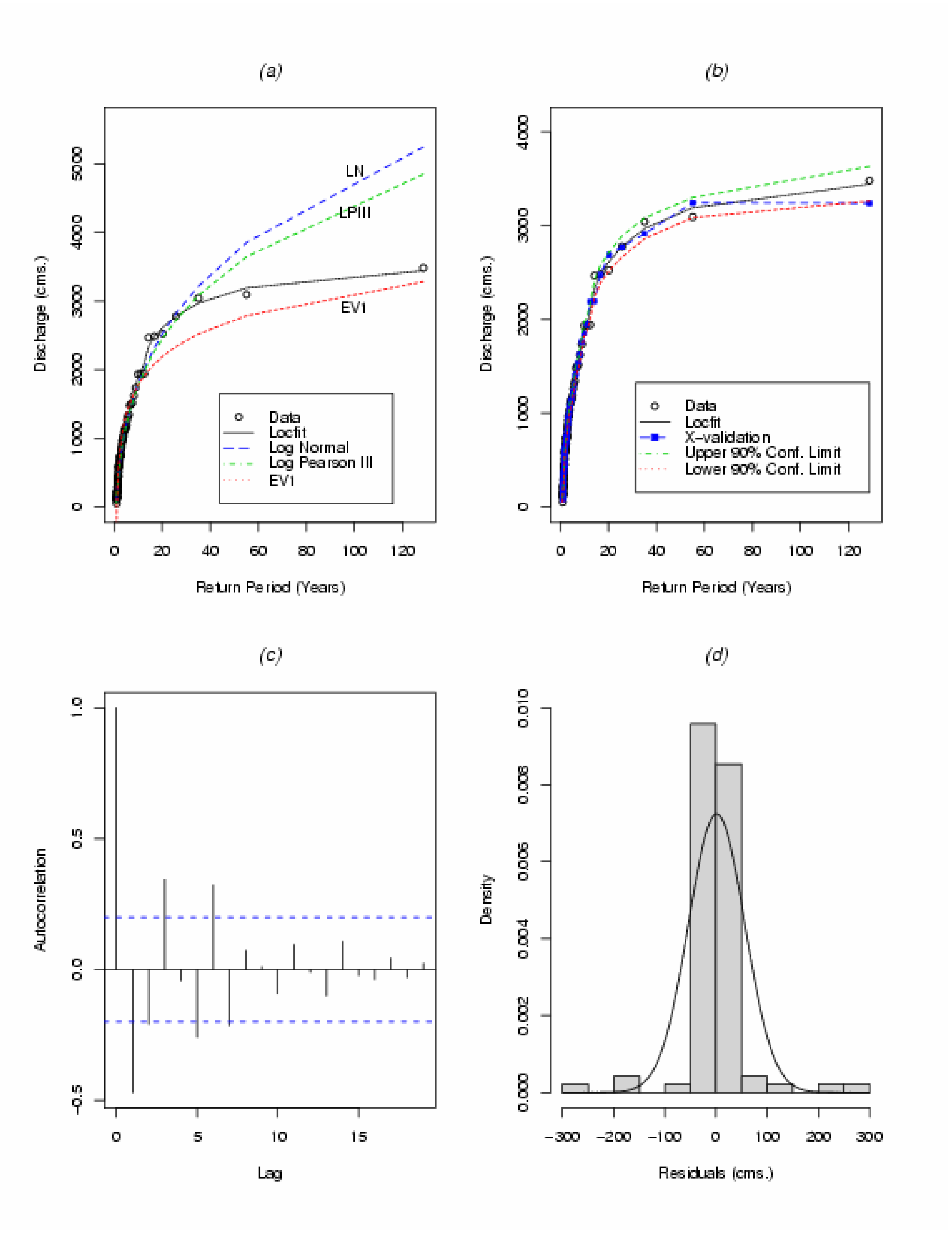

Figure 6: Quantile function estimates and x-validation for American River 3-day maximum flood data (1905-2000).

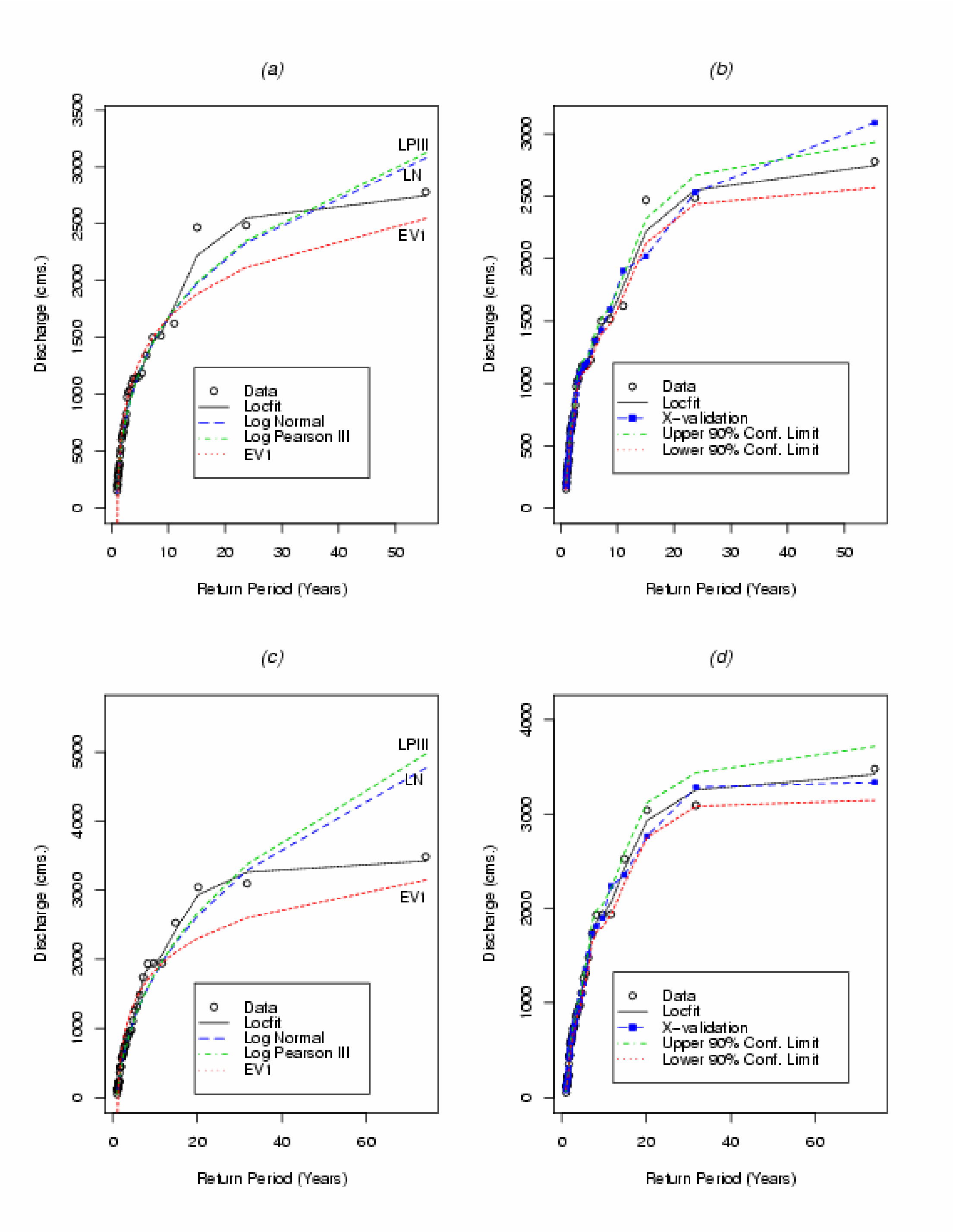

Figure 7: Quantile function estimates for American River 3-day maximum flood data (1905-1945) and (1946-2000).

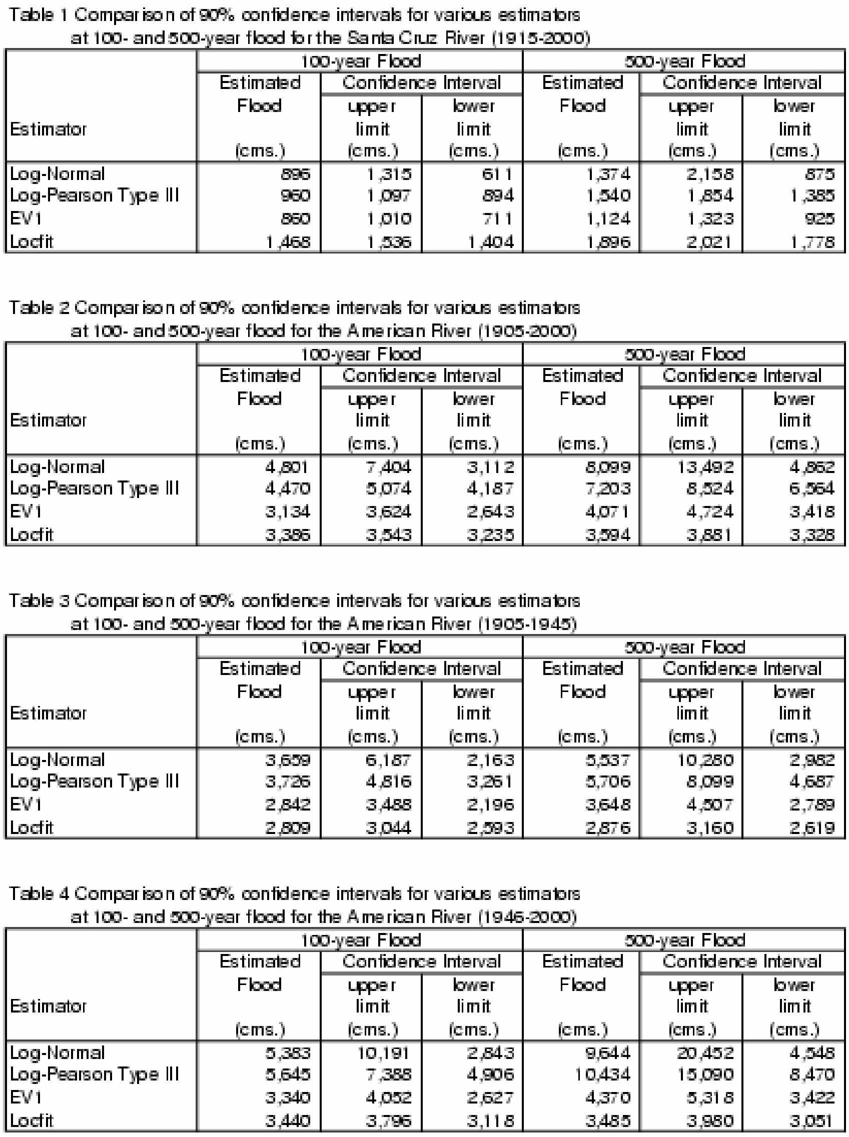

Table 1 Comparison of 90% confidence intervals for various estimators
at 100- and 500-year flood for the Santa Cruz River (1915-2000)

	100-year Flood			500-year Flood		
	Estimated	Confidence Interval		Estimated	Confidence Interval	
	Flood	upper	lower	Flood	upper	lower
Estimator		limit	limit		limit	limit
	(cms.)	(cms.)	(cms.)	(cms.)	(cms.)	(cms.)
Log-Normal	896	1,315	611	1,374	2,158	875
Log-Pearson Type III	960	1,097	894	1,540	1,854	1,385
EV1	860	1,010	711	1,124	1,323	925
Locfit	1,468	1,536	1,404	1,896	2,021	1,778

Table 2 Comparison of 90% confidence intervals for various estimators
at 100- and 500-year flood for the American River (1905-2000)

	100-year Flood			500-year Flood		
	Estimated	Confidence Interval		Estimated	Confidence Interval	
	Flood	upper	lower	Flood	upper	lower
Estimator		limit	limit		limit	limit
	(cms.)	(cms.)	(cms.)	(cms.)	(cms.)	(cms.)
Log-Normal	4,801	7,404	3,112	8,099	13,492	4,862
Log-Pearson Type III	4,470	5,074	4,187	7,203	8,524	6,564
EV1	3,134	3,624	2,643	4,071	4,724	3,418
Locfit	3,386	3,543	3,235	3,594	3,881	3,328

Table 3 Comparison of 90% confidence intervals for various estimators
at 100- and 500-year flood for the American River (1905-1945)

	100-year Flood			500-year Flood		
	Estimated	Confidence Interval		Estimated	Confidence Interval	
	Flood	upper	lower	Flood	upper	lower
Estimator		limit	limit		limit	limit
	(cms.)	(cms.)	(cms.)	(cms.)	(cms.)	(cms.)
Log-Normal	3,659	6,187	2,163	5,537	10,280	2,982
Log-Pearson Type III	3,726	4,816	3,261	5,706	8,099	4,687
EV1	2,842	3,488	2,196	3,648	4,507	2,789
Locfit	2,809	3,044	2,593	2,876	3,160	2,619

Table 4 Comparison of 90% confidence intervals for various estimators
at 100- and 500-year flood for the American River (1946-2000)

	100-year Flood			500-year Flood		
	Estimated	Confidence Interval		Estimated	Confidence Interval	
	Flood	upper	lower	Flood	upper	lower
Estimator		limit	limit		limit	limit
	(cms.)	(cms.)	(cms.)	(cms.)	(cms.)	(cms.)
Log-Normal	5,383	10,191	2,843	9,644	20,452	4,548
Log-Pearson Type III	5,645	7,388	4,906	10,434	15,090	8,470
EV1	3,340	4,052	2,627	4,370	5,318	3,422
Locfit	3,440	3,796	3,118	3,485	3,980	3,051