January 1993

A Kernel Quantile Function Estimator For Flood Frequency Analysis

Young-Il Moon

Upmanu Lall

Recommended Citation

Moon, Young-Il and Lall, Upmanu, "A Kernel Quantile Function Estimator For Flood Frequency Analysis" (1993). *Reports.* Paper 194.
https://digitalcommons.usu.edu/water_rep/194

A KERNEL QUANTILE FUNCTION ESTIMATOR FOR FLOOD FREQUENCY ANALYSIS

Young-Il Moon and Upmanu Lall
Utah Water Research Laboratory
Utah State University, Logan, UT 84322-8200

ABSTRACT

A kernel estimator (KQ) of the quantile function is presented here. Boundary kernels are used for extrapolation of tail quantiles. The bandwidth of the estimator is chosen using an automatic, "plug-in" method. Confidence intervals for the estimated quantile are estimated by bootstrapping. Comparisons of the estimator with selected tail probability estimators are offered. The KQ estimator presented here is shown to be competitive with other estimators.

INTRODUCTION

An objective of flood frequency analysis is to obtain an estimator of flood quantile magnitude (Q_T) for one or more locations on a river system. Correspondingly, a flood magnitude may be specified and an estimate of its return period (T) desired. In this paper, our objective was to estimate the flood quantile relationship using data from a gaged site. Traditionally an annual maximum frequency model f(x;θ) is proposed and calibrated from the N-year record of annual maximum flood peaks at a site. The quantile Q_T is then estimated as

9792

$$\hat{Q}_T = \hat{\theta}_1 + \hat{\theta}_2 \, y_T(\hat{\theta}_3) \qquad (1)$$

where $\hat{\theta}_1$, $\hat{\theta}_2$, and $\hat{\theta}_3$ are estimates of location, scale, and shape parameters of a selected distributional form $f(x;\theta)$, and $y_T(\theta_3)$ is a standardized variate value of return period T from $f(x;\theta)$.

Past and current research into methods of flood quantile estimation at a gaged site has concentrated mostly on the statistical aspects of the problem, based on the assumption that the sample of flood observations comes from a population with a known probability density function (pdf). However, no unique pdf or procedure is always best. Classical parametric estimation procedures are also most heavily weighted towards fitting the main body of the assumed probability density, and accord a negligible weight to the estimation of the tail of the distribution. Even if the parametric pdf fits well, considerable uncertainties for the magnitude of the floods at the frequencies of interest exist. Often, discriminating between different parametric probability models for the sample sizes available using standard tests such as the Chi-square and the Kolmogrov-Smirnov (e.g., Kite 1977) is difficult. Such tests are rather insensitive to tail behavior. This is an onerous mismatch in objectives.

Annual maximum flows at a site may be due to different causes (e.g., snowmelt, rainfall runoff, cyclonic activity). This leads to statistically heterogeneous populations or mixture distributions. The identification of finite mixtures of arbitrary or unknown populations from short records (typically n= 20-70) is not an attractive proposition and is not usually pursued. Webb and Betancourt (1992) tried to develop storm type classifications, separate events on that basis, fit a parametric pdf to each storm type, and finally combine the estimates. While this is a good demonstration that floods may arise from a mixture of processes, such a procedure is not easily implemented by a field engineer. The tail behavior of a mixture is often dictated by the tail behavior corresponding to the distribution in the mixture having the heaviest tail and by the relative proportion of events that correspond to each component. Methods that are robust in such situations (mixtures), are parsimonious and can give reasonable answers for a limited extrapolation of the data (e.g., 100-year flood), are

THE KERNEL QUANTILE ESTIMATOR

The Kernel Quantile estimator (KQ) is based on a kernel smoothing of the empirical quantile function of the data. The empirical quantile function is prescribed through a standard "plotting position formula". Let y_i, i=1...n, be the observed sequence of n annual maximum flows, arranged in ascending order. Let p_i, i=1...n, be the corresponding plotting positions estimated using a standard formula (e.g., the Weibull, Beard or Adamowski formula). Here we use Adamowski's (1981) formula :

$$p_i = \frac{i - 0.25}{n + 0.5} \qquad (2)$$

The empirical quantile function $x(p_i)$ is defined by the sample values y_i corresponding to each p_i. The quantile function x(p) to be estimated is defined as the event magnitude corresponding to the p^{th} quantile.

For flood frequency analysis, we are interested in the upper quantiles, i.e., p between 0.5 and 1, and in particular, for $0.9 \le p < 1$. Typical sample sizes for flood frequency analysis range from 20 to 100. An extrapolation of the data to $p > p_n$ is consequently needed.

The KQ estimator is based on the Gasser-Müller (1984) kernel regression estimator. It considers a convolution of the empirical quantile function with a kernel or weight function, as illustrated in Figure 1.

$$\hat{x}(p) = \sum_{i=1}^{n} \frac{1}{h} \int_{s_{i-1}}^{s_i} y_i \, K\left(\frac{p-u}{h}\right) du = \sum_{i=1}^{n} \frac{1}{h}\, y_i \int_{s_{i-1}}^{s_i} K\left(\frac{p-u}{h}\right) du = \sum_{i=1}^{n} y_i \, w_i \qquad (3)$$

where s_i is an interpolating sequence of the p_i, given as $s_i=(p_i+p_{i+1})/2$, i=1...n-1, $s_0=0$, $s_n=1$; h is a bandwidth associated with the point p; and K(.) is a kernel or weight function, and $p \in [0,1]$.

The kernel function K(.) is usually taken to satisfy the requirements, $\int K(t)dt = 1$, K(t) =

- K(t), and $\int K^2(t)dt < a$, i.e., it is a symmetric probability density with finite variance; where $t = (p-u)/h$. Müller(1988) points out that while different kernels perform similarly in terms of Mean Square Error (MSE), kernels of higher order lead to estimated functions with a higher degree of differentiability. A kernel of order p has finite moments up to order p, and vanishing moments of order higher than p.

It is preferred that K(.) have compact support to minimize the effect of the bounded domain ($0<p<1$) on the nonparametric estimate of the quantile function. A specialized boundary kernel that corresponds to the kernel used in the interior is needed within a bandwidth of the boundary to take care of the bias in the weighted convolution in the boundary region. The interior kernels provide a weight sequence that is suitable for interpolating the observed data, while the boundary kernels provide extrapolation. Müller (1991) develops boundary kernels corresponding to specific interior kernels. Here we use the Epanechnikov kernel in the interior (that is MSE optimal for order 2) and the corresponding Müller boundary kernel. These kernels are:

Epanechnikov kernel: $K(t) = 0.75(1 - t^2)$ (4)

Boundary kernel corresponding to the Epanechnikov kernel:

$$K_x(q,t) = 6(1+t)(q-t)\frac{1}{(1+q)^3}\left\{1 + 5\left(\frac{1-q}{1+q}\right)^2 + 10\frac{1-q}{(1+q)^2}t\right\} \tag{5}$$

where $t=(p-p_i)/h$; $K_{-}\{(1-p)/h,t\}$ is the boundary kernel used for the right boundary, i.e., if $p\in[1-h,1]$, and $q=(1-p)/h$; and $K_{+}(p/h,t)$ is the boundary kernel used for the left boundary, i.e. if $p\in[0,h]$, and $q=p/h$; $0<q<1$ and $-1\leq t\leq 1$.

Note that for q=1, the boundary kernel (5) is identical to the Epanechnikov kernel (4). The kernel quantile estimation process is illustrated through the following example. With y(i) given by the 71 year record of annual maximum floods for the Santa Cruz River from Webb and Betancourt (1992), (see Table 1), and p(i) given by equation 2, consider the estimation of a quantile at p = 0.8 using the Epanechnikov kernel function K(t) with the bandwidth h=0.064. The values of y(i) for $0.736\leq p\leq 0.864$ will contribute to the estimate. The kernel estimate for x(0.8) is not in the boundary area in Figure 1. The curve in Figure 1 is the Epanechnikov kernel function, and the weight w_i is the area under this kernel around the i^{th} point (i.e., from s_{i-1} to s_i). The quantile estimate is:

x(0.8)=0.144 y(53) + 0.071 y(54) + 0.116 y(55) + 0.146 y(56) + 0.161 y(57) + 0.160 y(58) + 0.144 y(59) + 0.112 y(60) + 0.065 y (61) + 0.010 y(62)=282 m3/s. (6)

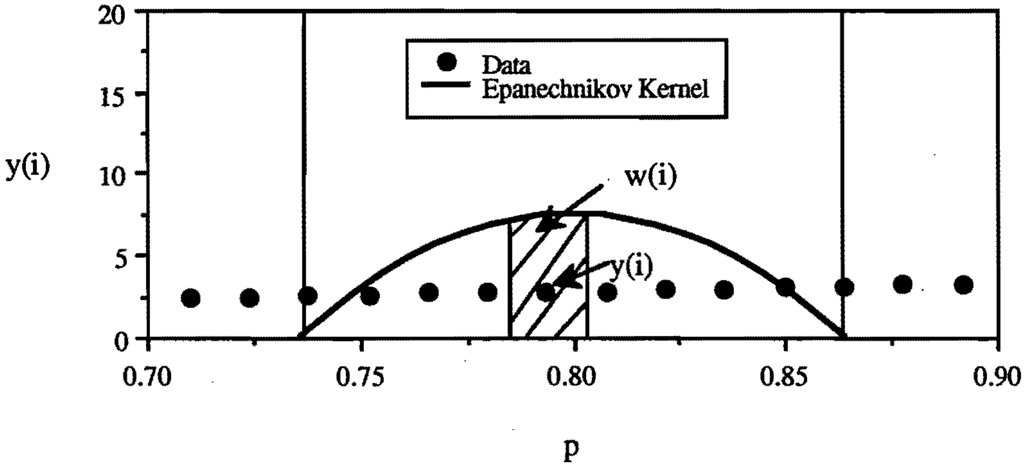


Figure 1. Kernel estimate for x(0.8), y(i) in hundreds of cubic meter per second.

However, if we are interested in x(0.99) then we are in the boundary region (0.936 <p<1) , since the point of estimate lies within a bandwidth (h=0.064) of the right boundary. In this case we are extrapolating the empirical quantile function, and the weight sequence or kernel used has to be modified. The interior (4) and the boundary kernel (5) are shown for q=0.15625 in Figure 2, corresponding to an h of 0.064 and p of 0.99. The quantile x(0.99) for Santa Cruz's River data with 71 data points without considering the boundary effect is:

x(0.99)=0.0001y(66) + 0.037y(67) + 0.092y(68) + 0.130y(69) + 0.154y(70) + 0.202y(71) = 524 m3/s. (7)

This is a biased estimate since the kernel centered at p=0.99 extends past 1.0, outside the domain of interest, and the data values (i.e. p(i)) are not symmetrically distributed around the point of estimate. This situation is remedied by the boundary kernel, which is defined over the domain of interest, and also accounts for the asymmetric data distribution relative to p. The resulting estimate is:

x(0.99) = -0.001y(66) - 0.177y(67) - 0.105y(68) + 0.245y(69) + 0.545y(70) + 0.494y(71) = 1094 m3/s. (8)

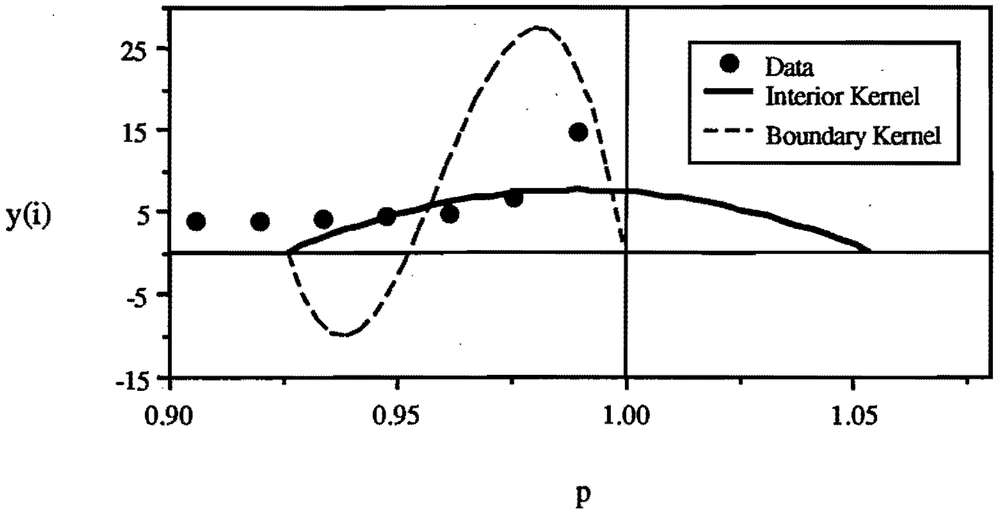


Figure 2. Kernel estimate for x(0.99), y(i) in hundreds of cubic meter per second.

When using the interior kernel, the estimate x(0.99) is formed using a weighted moving average of the empirical quantile function, with a symmetric weighting scheme about the point of estimate. In the situation where extrapolation is needed, this yields an estimate that is effectively centered somewhere in the span of the observations, and not at p=0.99. Consequently, the estimate of x(0.99) is lower than the empirical quantile corresponding to p(i)=0.976. However, when the asymmetry is accounted for using the boundary kernel, a much more reasonable estimate is obtained.

The kernel quantile estimate is sensitive to the choice of the kernel and the choice of the bandwidth. Examination of an expansion of the MSE of the kernel estimator, in terms of a

Taylor series, suggests (see Härdle (1991)) that sensitivity to the bandwidth is perhaps an order of magnitude more important than kernel choice. We observe from the example above that bandwidth variation has the effect of admitting a different number of upper order statistics into the KQ estimate. Note also that the KQ estimator differs from traditional tail probability estimators in that a sliding neighborhood around the desired point of estimate is used, rather than a preset number of upper order statistics. In the next section we discuss how an MSE optimal bandwidth can be estimated, once a kernel function has been specified.

Bandwidth Estimation

The bandwidth or smoothing parameter h determines the roughness or smoothness of the estimated function. Smaller bandwidths result in fewer data points contributing to the estimate at any point, and hence a rougher on more bumpy estimator. Larger bandwidths however allow averaging over a larger data space resulting in a smoother estimator. As bandwidth increases, bias increases and variance decreases. For pointwise consistency of the estimate, the bandwidth must get smaller as the sample size increases. Consider the estimation problem at the data points as :

$$\hat{x}_i(p_i) = x(p_i) + \varepsilon_i \qquad (9)$$

where ε_i is a residual term.

The asymptotic mean squared error (up to the leading terms in the Taylor series expansion) of KQ is seen to be (Müller, 1991):

$$\mathrm{MSE}\left(\hat{x}(p)\right) = E\left[\hat{x}(p) - x(p)\right]^2$$

$$-\frac{\sigma^2}{nh}\int_{-1}^{q}(K_x(q,t))^2\,dt + \frac{1}{4}h^4\left\{x''(p)\right\}^2\left\{\int_{-1}^{q}K_x(q,t)\,t^2\,dt\right\}^2 \qquad (10)$$

where $K_x(q,t) = K_+(1,t)$ for interior, $h \le p \le 1-h$; and is given by equation 5 in the boundary regions, $x''(p)$ is the second derivative of $x(p)$; $\sigma^2 = \mathrm{var}(\varepsilon_i)$.

The first term in equation 10 provides an estimate of the estimation variance, while the second term corresponds to the bias squared. Some methods to find an optimal bandwidth that balance bias and variance include the Generalised Cross Validation (GCV) method proposed by Craven and Wahba (1979) and the Plug-In method by Gasser et al. (1991) as well as local least absolute deviation and least squares cross validation aimed at minimizing the mean square error of x(p). We found that Gasser et al.'s Plug-In method with an Epanechnikov kernel worked better than the others in our Monte Carlo tests.

An optimal global bandwidth (Gasser et al., 1991) that minimizes the Average Integrated MSE over the domain (0<p<1), is given by:

$$h = \left\{ \frac{1.5}{n} \, \frac{c_1}{c_2} \, \frac{\sigma^2}{\int_0^1 \{x''(p)\}^2 \, dt} \right\}^{0.2} \qquad (11)$$

where $c_1 = 2\int_{-1}^{1} K_x(q,t)^2 dt$ and $c_2 = 4\int_{-1}^{1} K_x(q,t)t^2 \, dt$

The Gasser et al. plug in method seeks to recursively estimate h through kernel estimates of the a priori unknown term $\int_0^1 \{x''(p)\}^2 dp$. Such an estimator $\hat{r}_2(p;\, h_2)$ for $x''(p)$ is

$$\hat{r}_2(p;h_2) = \frac{1}{h_2^3} \sum_{i=1}^{n} \int_{s_{i-1}}^{s_i} D_x\left\{q, \frac{p-u}{h_2}\right\} du \; y_i \qquad (12)$$

where $D_x\{q,(p-u)/h_2\}$ is an optimal fourth order kernel suitable for estimating the second derivative of the target function (see Müller (1991)), and h_2 is a bandwidth appropriate for estimating the second derivative of the target function.

Using asymptotic arguments, Gasser et al. (1991) specify the bandwidth $h_2 = h\, n^{-1/10}$. They show that this leads to convergence rates of the order of $n^{-1/2}$. The residual variance σ^2 is also unknown a priori. However, a number of nonparametric estimators for σ^2 are available. We used the following estimator given by Gasser et al. (1986):

$$\hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=2}^{n-1} c_i^2 \tilde{\varepsilon}_i^2 \qquad (13)$$

where

$\tilde{\varepsilon}_i = a_i x(p_{i-1}) + b_i x(p_{i+1}) - x(p_i)$

and $a_i = (p_{i+1}-p_i)/(p_{i+1}-p_{i-1})$; $b_i=(p_i - p_{i-1})/ (p_{i+1}-p_{i-1})$; $c_i^2 = (a_i^2+b_i^2+1)^{-1}$

The following procedure is followed to estimate the bandwidths h:

i) Set $h_1 = 1/n$.

ii) Iterate i = 2, 3, ... until i=11.

$$\hat{h}_i = \hat{G}(\hat{h}_{i-1}\, n^{1/10}) = \left\{ \frac{1.5}{n} \frac{c_1}{c_2} \frac{\hat{\sigma}^2}{\int_0^1 \hat{r}_2(p; \hat{h}_{i-1}\, n^{1/10})^2\, dt} \right\}^{0.2} \qquad (14)$$

iii) Set $\hat{h} = \hat{h}_{11}$

Confidence Intervals for KQ Estimates

A strategy for the estimation of pointwise confidence intervals for KQ estimates is presented in this subsection. A difficulty with the construction of direct confidence intervals for KQ is the presence of bias in the estimates. Eliminating the bias is not possible, but on average the variance dominates the MSE. Two main ideas have been considered for constructing confidence intervals of kernel regression estimators. These are the use of an asymptotic distribution (typically Gaussian) for the residuals and bootstrap approximations of KQ. The asymptotic distribution of kernel regression estimates has been considered by Wahba (1983), Nychka (1988, 1990), and Hall and Titterington (1988). Wahba (1983) and Nychka (1988, 1990) considered confidence intervals based on Bayesian considerations and smoothing spline estimators. Hall and Titterington (1988) described the construction of confidence bands based on interpolation formula in numerical differentiation. Bootstrap confidence bands based on kernel estimators have been studied by Hardle and Bowman (1988) and Hardle and Marron (1991). The method of construction of confidence bands in this paper is based on the bootstrap. The bootstrap (Efron (1979)) is a technique for resampling the data with replacement. The bootstrap resample is taken from the empirical quantile function. The resampling can be done from the data pairs $\{(p_i,y_i), i=1,..,n\}$ according to the following algorithm.

i) Given a sample $\{(p_i, y_i), i=1,..,n\}$

ii) Generate $\{\delta_{j;}\ j=1,..,n\}$ from a uniform distribution.

iii) Construct a new sample Y_j, $j=1...n$, where $Y_j=y_i$ such that $s_{i-1} < \delta_j < s_i$.

iv) Find x(p) using the KQ estimator and the new data set

v) Repeat (ii)-(iv) M times (e.g. M=1000)

vi) From the estimates $\hat{x}_m(p)$, $m=1...M$, identify the β and $(1-\beta)$ confidence limits for $\hat{x}(p)$.

A bootstrap estimate of the sampling distribution of $\hat{x}(p)$ is likewise obtained. Such estimates are presented for our running example in the applications section. Note that the bootstrap cannot address estimation bias, i.e. if a biased estimator is used, the bootstrap confidence intervals or sampling density will be likewise biased. This limits the utility of the bootstrap to compare results across methods with markedly different amounts of bias. The resulting bands reflect only the estimation variance.

TAIL ESTIMATORS

Many of the proposed estimators of tail probabilities (Hill 1975, and Breiman et al. 1981) assume that a distribution function F(x) is in the domain of attraction of a known distribution function G(x) for all values greater than some predetermined value x_{p_0}. Hill (1975) and Hosking and Wallis (1987) developed tail probability estimators by forming an estimate of an extreme right tail quantile under the assumption that the behavior in the upper tail follows the Pareto distribution. Another method proposed by Breiman and Stone (1985) assumes that the tail of the distribution is exponential or approximately linear in log(p). They also considered a quadratic tail method in which x_p is assumed to be a quadratic function of log(p).

Since only the upper part of the data is used for estimating upper tail probabilities or upper tail quantiles, tail estimators do not care whether or not the lower data values follow the distribution. A disadvantage is that tail estimators still need to specify parametric family behavior and the place at which tail starts. Pickands (1975), Hill (1975), Hall (1982), and Hall and Welsh (1985) examined the problem of estimating the number of extreme values or the cutoff point for the tail to achieve optimal performance, and showed that, in general, this number depends on unknown properties of the tail. Therefore, the size of the extreme subsample used to construct the estimators must also be estimated from the sample. These methods are based on asymptotics and one must consider whether or not asymptotics can be invoked for the small sample sizes available in practice. In our Monte Carlo simulations from known parent populations, using samples of size 20 and 100, we found a simple strategy of specifying 5 and 10 upper order statistics, respectively, outperformed the sophisticated asymptotic strategies presented by these authors. We suspect that this is due to the high variance associated with these methods for such small samples.

For the sake of brevity, the reader is referred to Moon et al. (1993) for algorithms of selected tail probability estimators used in the comparisons that follow. Hill's method (PT1) is presented for historical reasons, and for comparison with a recent Pareto model due to Hosking and Wallis (1987) (PT2). The Exponential and Quadratic tail methods (ET, QT) due to Breiman and Stone are also presented. The Type I Extreme Value distribution (EV1) is also considered because EV1 is often used as a model for annual maximum floods, and can be considered a tail estimation method.

APPLICATIONS

We conducted a Monte Carlo experiment similar to those reported in Lall et al. (1993) and Moon et al. (1993) to compare the performance of KQ with PT1, PT2, ET, QT, and EV1 where the underlying population was assumed to be Normal (0,1), Pearson III with parameters (0,1,1), and a Normal location mixture (0.5N(0,1) + 0.5N(3,1)). One thousand samples of size 20 and 100 were generated in each case.The performance of the methods for the two sample sizes, and with the Normal and with the Pearson III data was qualitatively similar. The performance of PT1 degraded substantially for the smaller sample size. Results for Bias and Root MSE (RMSE) of $\hat{x}_p$ (p=0.9,0.95,0.98,.0.99,0.995), for samples of size 20 and 100, for the normal, Pearson III, and mixture data are shown in figures 3 through 14.

A perusal of these figures suggest that KQ and QT are the best methods in these situations, with consistent performance in terms of bias and rmse. There are cases in which one of the other estimators may do better, but typically the same estimator performs rather poorly in other situations. QT had a very high rmse for n=20, with Pearson III data, while the performance of KQ was stable. In terms of bias, QT typically performed marginally better across the simulations. However, KQ was marginally superior in terms of rmse. The rmse performance of KQ is also somewhat superior to that of QT as the degree of extrapolation (i.e., (1/np)) increases. Both of these methods can be recommended on the basis of our Monte Carlo simulations.

Santa Cruz River Annual Maximum Floods

A comparison between KQ, tail estimators (PT1, PT2, ET, QT, EV1), and a kernel distribution function estimator (VK-C-AC) of the quantile function for the Santa Cruz River data is shown in Figure 15. Note that the largest recorded flood (1493 m^3/s) is more than double the magnitude of the second largest flood (671 m^3/s). Reported parametric estimates (see Table 2) of the 100-year flood range from 572 to 2,780 m^3/s. Of interest is the first estimate computed by Webb and Betancourt (1992). They separated floods above base discharge (48 m^3/s) by storm type into three categories: monsoonal storms (56 data points), frontal systems (18 data points), and dissipating tropical cyclones (19 data points). A log-Pearson type III distribution using maximum-likelihood analysis was fit to each partition. The 100-year flood was then estimated by combining the three estimates as 1,050 m^3/s. Note from figure 15 that the KQ (1094 m^3/s) and QT (1102 m^3/s) estimates of the 100 year flood are quite comparable to this estimate and are in the middle of the range of the parametric estimates. In both cases, the methods was applied to the full data set, and parameters were chosen automatically.

Note that KQ effectively interpolates the empirical quantile function for this data set for values of p up to approximately 0.92, and smooths it thereafter. Recall that the boundary region for the kernel estimator is $0.936<p<1$. The behavior of the other estimators is also of interest. PT2 clearly appears inconsistent with the empirical quantile function. VK-C-AC interpolates the empirical quantile function all the way to p_n. This leads one to suspect that it would be rather sensitive to extreme values in the data set, and also to the plotting position formula selected. The agreement of KQ and QT for $p=0.99$ appears fortuitous. QT's tail behavior appears to be closer to the empirical quantile function, than that of KQ. We also see that the other methods (EV1, ET and PT1) are more strongly influenced by the main body of the data, rather than the tails.

Bootstrap confidence intervals with $\beta=0.05$, and standard errors of estimate of the 100 year flood for KQ, QT, PT1, and EV1, are reported in Table 4. Of the nonparametric methods QT has the smallest standard error, and the tightest confidence interval. The standard error and confidence interval for EV1 are considerably smaller, reflecting the reduced variance of estimation in using a parametric method. However, the bias issue remains unresolved. Bootstrap confidence intervals for KQ are also reported for a range of p values, in Table 4 and in Figure 16. The large width of the confidence intervals as p increases reflects the growing uncertainty in the estimate of the rarer events. Note that the confidence intervals obtained cover virtually all the methods of estimation considered with this data set at most values of p. This is partly because of the high uncertainty in the tail, and the local nature of KQ. It reflects also on the usual dilemma of choosing between models for tail behavior.

Bootstrap estimates of the sampling densities of estimates at $p=0.9$ and 0.99, for KQ and QT are presented in Figure 17 and 18 respectively. There is little difference between the methods for $p=0.9$. For $p =0.99$, one can see three peaks in the distribution for KQ, but only two in that for QT. The bandwidth (and hence the number of upper order statistics) used by KQ varies by sample, and the peaks seem to reflect sensitivity to the inclusion (possibly repeated) or deletion of specific observations. On the other hand, QT is always fit using the upper 7 order statistics, and seems to be less sensitive to the inclusion or deletion of the largest observation. Note also that the QT estimate was closer to the empirical quantile function in the neighborhood of the largest sample value. These observations would suggest that QT may be preferred as the estimator in this situation. Recall, however, that the bootstrap density of x(0.99) does not account for the different bias that KQ and QT estimates are likely to have.

Sensitivity to Plotting Position Formula

The kernel quantile estimator needs prior estimates of the empirical quantile function based on plotting position formula. Many plotting position formula are special cases of the general formula:

$$p_i = \frac{i - \alpha}{n + 1 - 2\alpha} \tag{15}$$

where α is a constant that depends on the underlying distribution. For example, $\alpha = 0$ for the uniform (Weibull's formula), 0.25 for Adamowski's formula, 0.44 for EV1 and the exponential distribution, and 0.5 for Hazen's formula. We chose the extreme members of this set, i.e, Weibull's and Hazen's formulae to investigate the sensitivity of KQ to the choice of plotting position formulae. The results of this evaluation are presented in Table 4 and graphically in Figure 16. The 100 year flood estimate would now range from 947 to 1244 m^3/s, a range that is still substantially smaller than the variation in the parametric estimates reported for this data set. Recall that the KQ bootstrap confidence intervals at β=0.05, using the Adamowski formula for p_i, range from 466 to 1658 m^3/s.

CONCLUSIONS

KQ and QT performed similarly. They appeared relatively robust with respect to the other methods for the variety of situations tested. Results from KQ applied to log transformed data (not reported here) were similar. Both KQ and QT consider weighted linear combinations of order statistics to form the quantile function. The weight sequence and the number of order statistics used by the two methods differ. The analysis of the Santa Cruz

River data revealed that these methods can give reasonable results with data from mixed populations. However, they also illustrated the futility inherent in flood frequency estimation - it is easy to design innumerable schemes that are equally plausible within the range of the data and quite different under extrapolation.

There is no shortage of methods for the frequency analysis of annual maximum flood data. A number of hydrologists have been concerned with the search for the "best" distributional model, and the best parameter estimation scheme for such models. Clearly, this philosophy can extend to a search for the best model for tail extrapolation, once a recognition sets in that the estimation of tail behavior may be a fundamentally different problem than that of estimating a suitable density function for the main body of the data. This is exacerbated where the data represents a finite mixture of generating mechanisms. Parsimonious models are important in any estimation situation. Parametric approaches that attempt to model mixtures, or allow for more flexible curves (e.g. Wakeby) suffer from a lack of parsimony (and a corresponding increased estimation variance), and may still be inappropriate in a given situation. On the other hand, a simple parametric model may be quite inappropriate for the tail of the data distribution, even if it provides an adequate fit elsewhere and "wins" in terms of having lower variance.

We feel that the search for the "best" tail distributional model is just as futile as the search for the best p.d.f., perhaps more so given the uncertainty induced by the small samples and complex mechanisms (is the process really stationary and statistically homogeneous ?). The comments above apply to at site as well as regional flood frequency estimation. Given these comments, we are comfortable recommending adaptive tail extrapolation methods such as KQ and QT, together with an understanding of the relatively large associated uncertainty of such estimates, as indicated by the large bootstrap standard errors. The primary assumptions here are differentiability of the quantile function, and the estimation of a weight sequence, that depends on one parameter for KQ, and on a fixed number of upper order statistics for QT. Tail behavior is assumed for QT while KQ is more adaptable. Both approaches sacrifice variance for reduced model bias relative to parametric methods. However, note that the variance (across site) of a procedure that includes the selection of an appropriate parametric model at each site may be no better.

Robustness in performance across different situations is a desirable attribute. We feel kernel quantile estimators can be developed that are superior to KQ. Such developments may require further theoretical analysis of tail properties to determine the bandwidth and the appropriate kernel functions.

ACKNOWLEDGEMENTS

The work reported here was supported in part by the U.S. Geological Survey through their funding of the second author's 1992-3 sabbatical leave, when he worked with the Branch of Systems Analysis, WRD, USGS, National Center, Reston VA.

REFERENCES

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

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

Adamowski, K. and Labatiuk, C. (1987) “Estimation of flood frequencies by a nonparametric density procedure”, Hydrologic Frequency Modeling, 97-106.

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

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

Bardsley, W. E. (1988) “Toward a general procedure for analysis of extreme random events in the earth sciences”, Mathematical Geology 20(5), 513-528.

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

Breiman, L., Stone, C. J., and Ginns, J. (1981) “Further developments of new methods for estimating tail probabilities and extreme value distributions”, Technical Report #TSD-PD-A243-1, Tech. Service Corp., Santa Monica, California 90405.

Breiman, L., and Stone, C. J. (1985) “Broad spectrum estimates and confidence intervals for tail quantiles", Technical Report 46, Univ. of California, Berkeley, California.

Craven, P. and Wahba, G. (1979) "Smoothing noise data with spline functions", Numerische Mathematik 31, 377-403.

Efron, B. (1979). "Bootstrap method: another look at the jackknife", Annals of Statistics 7,1-26.

Gasser, T. and Müller, H. G. (1984) "Estimating regression functions and their derivatives by the kernel method", Scandinavian Journal of Statistics 11, 171-185.

Gasser, T., Sroka, L., and Jennen-Steinmetz, C. (1986) "Residual variance and residual pattern in nonlinear regression", Biometrika 73(3), 625-633.

Gasser, T., Kneip, A., and Kohler,W. (1991) "A flexible and fast method for automatic smoothing", Journal of the American Statistical Association 86(514), 643-652.

Guo, S. L. (1991) "Nonparametric variable kernel estimation with historical floods and paleoflood information", Water Resources Research 27(1), 91-98.

Hall, P. (1982) "On some simple estimates of an exponent of regular variation", J. R. Statist. Soc. B 44(1), 37-42.

Hall, P. and Welsh, A. H. (1985) "Adaptive estimates of parameters of regular variation", Annals of Statistics 13, 331-341.

Hall, P. and Titterington, D. M. (1988) "On confidence bands in nonparametric density estimation and regression", J. Multivariate Analysis 27, 228-254.

Hardle, W. and Bowman, A. W. (1988) "Bootstrapping in nonparametric regression:Local adaptive smoothing and confidence bands" J. Amer. Statist. Assoc. 83, 102-110.

Hardle, W. and Marron, S. (1991) "Bootstrap simultaneous error bars for nonparametric regression", Annals of Statistics 19, 778-796.

Hill, B. M. (1975) "A simple general approach to inference about the tail of a distribution", Ann. Math. Statist. 3, 1163-1174.

Hosking, J. R. M. and Wallis, J. R. (1987) "Parameter and quantile estimation for the generalized pareto distribution", Technometrics 29(3), 339-349.

Kite, G. W. (1977) Frequency and Risk Analyses in Hydrology, Water Resources Publications, Fort Collins, Colorado.

Lall, U. and Niu, J. (1989) "Variable bandwidth kernel density estimation", AGU Spring Meeting, Baltimore, Maryland, May 7-12, 1989.

Lall, U., Moon, Y.-I., and Bosworth, K. (1993) "Kernel flood frequency estimators: Bandwidth selection and kernel choice", Water Resources Research 29 (4), 1003-1015.

Moon, Y.-I., Lall, U., and Bosworth, K. (1993) "A comparison of tail probability estimators", Journal of Hydrology, in press.

Müller, H.-G. (1988) "Lecture notes in statistics", Springer-Verlag, Berlin, Germany.

Müller, H.-G. (1991) "Smooth optimum kernel estimators near endpoints", Biometrika 78(3), 521-530.

Nychka, D. (1988) "Bayesian confidence intervals for smoothing splines", Journal of the American Statistical Association 83, 1134-1143.

Nychka, D. (1990) "The average posterior variance of a smoothing spline and a consistent estimate of the average squared error", Annals of Statistics 18, 415-428.

Parzen, E. (1979) "Density quantile estimation approach to statistical data modelling" Smoothing Techniques for Curve Estimation 757, 155-180.

Pickands, J. (1975) "Statistical inference using extreme order statistics", Ann. Stat. 3(1), 119-131.

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

Sheather, S. J. and Marron, J. S. (1990) "Kernel quantile estimators", Journal of the American Association 85(410), 410-416.

Takeuchi, H. and Yokoyama, S. (1991) "A note on quantile estimation by the kernel method", Communications in Statistics - Simulation 20(1), 149-156.

Wahba, G. (1983) "Bayesian confidence intervals for the cross-validated smoothing spline" J. Roy. Statist. 10, 1040-1053.

Webb, R. H.and Betancourt, J. L. (1992) "Climatic variability and flood frequency of the Santa Cruz River, Pima county, Arizona", USGS Water-Supply Paper 2379, 1-40.

Zelterman, D. (1990) "Smooth nonparametric estimation of the quantile function" Journal of Statistical Planning and Inference 26, 339-352.

Table 1. Annual flood data(m^3/s), Santa Cruz River (1915-1986) at Tucson, Arizona from Webb and Betancourt (1992).

425	142	212	139	133	55	113	57	54	58	96	323
55	45	295	50	261	119	173	170	292	153	93	255
227	320	71	47	128	185	306	121	48	109	108	269
142	108	157	271	309	74	86	180	125	174	470	141
132	368	34	156	166	456	247	242	227	133	54	225
70	201	671	142	382	78	76	283	1493	283	n.a.	54

Table 2. Estimates for the 100-year flood on the Santa Cruz River at Tucson, Arizona reported Webb and Betancourt (1992)

Method or probability distribution	100-year flood(m^3/s)
Mixed-population analysis of floods cased by different storm types	1050
Curve comparison with floods in other watersheds	1280
Log-Pearson type III with method of moments fitting	575-1530
Log-Pearson type III with regression analysis	640-1810
Log-Pearson type III with envelope curve	572
Log-Boughton distribution with method of moments fitting	2180
Rain estimated from 100-year rainfall	1420
Log-Extreme Value distribution with method of moments fitting	2730-2780
Model estimated from rainfall-runoff model with 100-year	1330-1900

Table 3. Sensitivity of KQ Estimates to Plotting Formula for Santa Cruz's River (1915-1986).

Return Period	Expected Quantile Values (m3/s)			90% Confidence Interval
	Weibull For.	Adamowski For.	Hazen For.	for Adamowski. Form.
10 yrs (p=0.9)	364	360	356	300-451
20 yrs (p=0.95)	611	571	534	373-947
50 yrs (p=0.98)	1001	893	789	444-1527
100 yrs (p=0.99)	1244	1094	947	466-1658
200 yrs (p=0.995)	1403	1230	1053	470-1807

Table 4. Comparison of 90% confidence intervals and standard errors for KQ, QT, and PT1 at 100-year flood for Santa Cruz's River (1915-1986), cubic meter per second.

	Estimated 100-year flood	Confidence Interval	Standard Error
KQ	1094	(466, 1658)	414
QT	1102	(467, 1530)	348
PT1	873	(464, 1843)	451
EV1	818	(499,1132)	210
Lognormal	826	(614,1111)	154

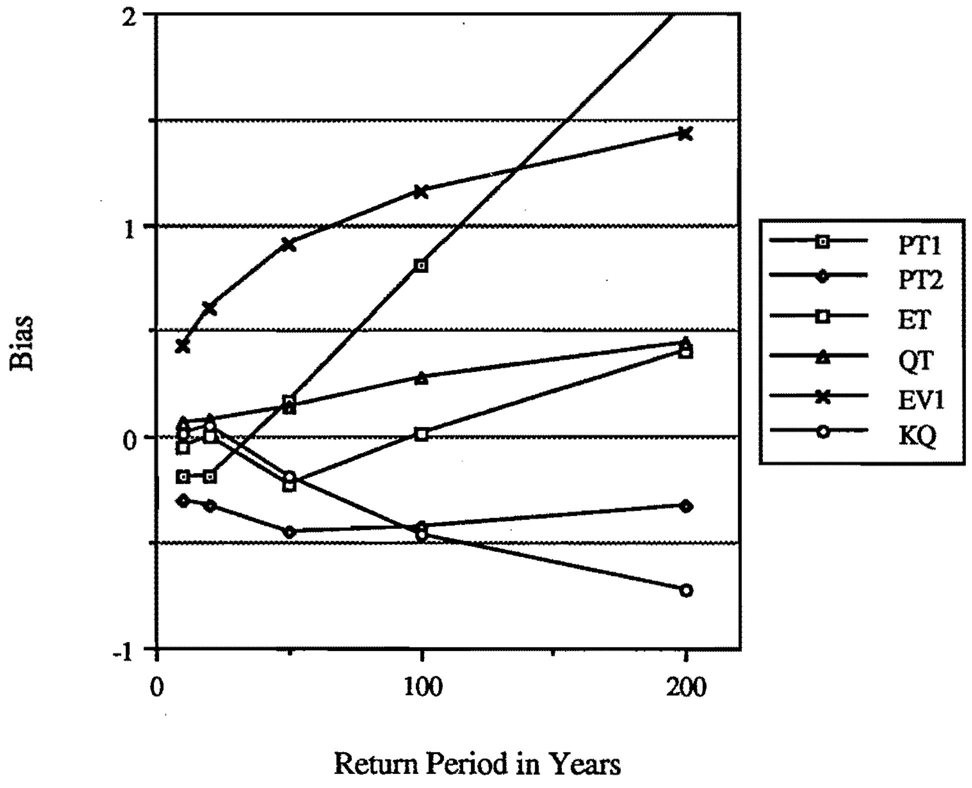


Figure 3

Bias for Normal data N(0,1), n=20

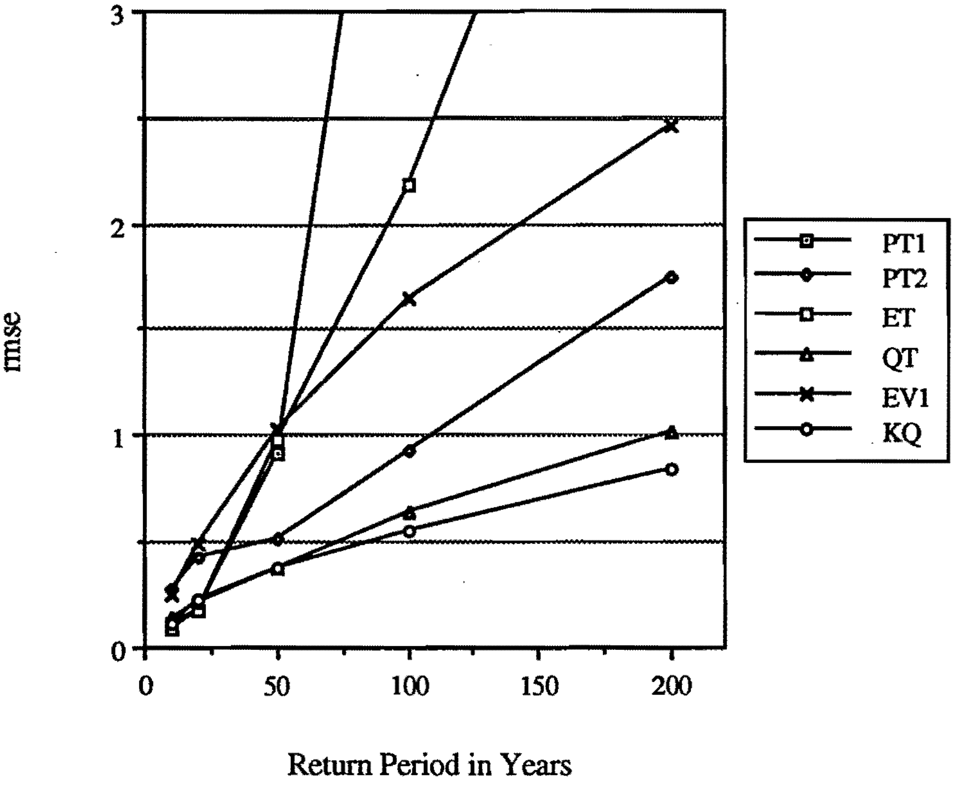


Figure 4

RMSE for Normal data N(0,1), n=20

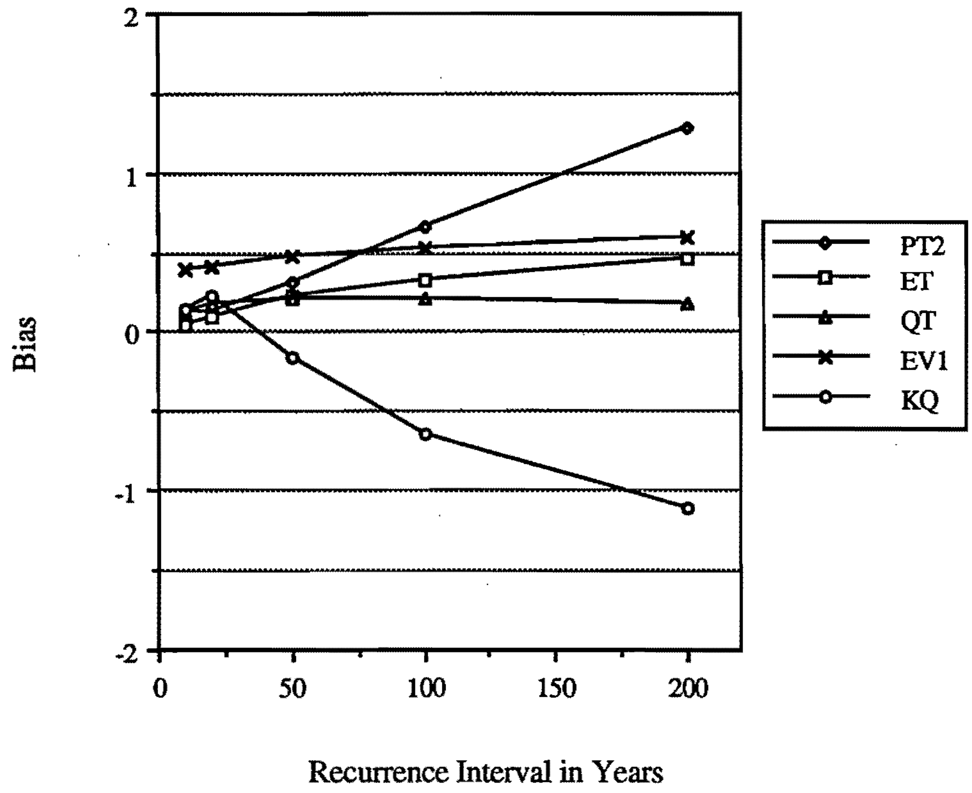


Figure 5

Bias for Pearson III (0,1,1) data, n = 20

(PT1 is not shown, because its bias values are off the scale used)

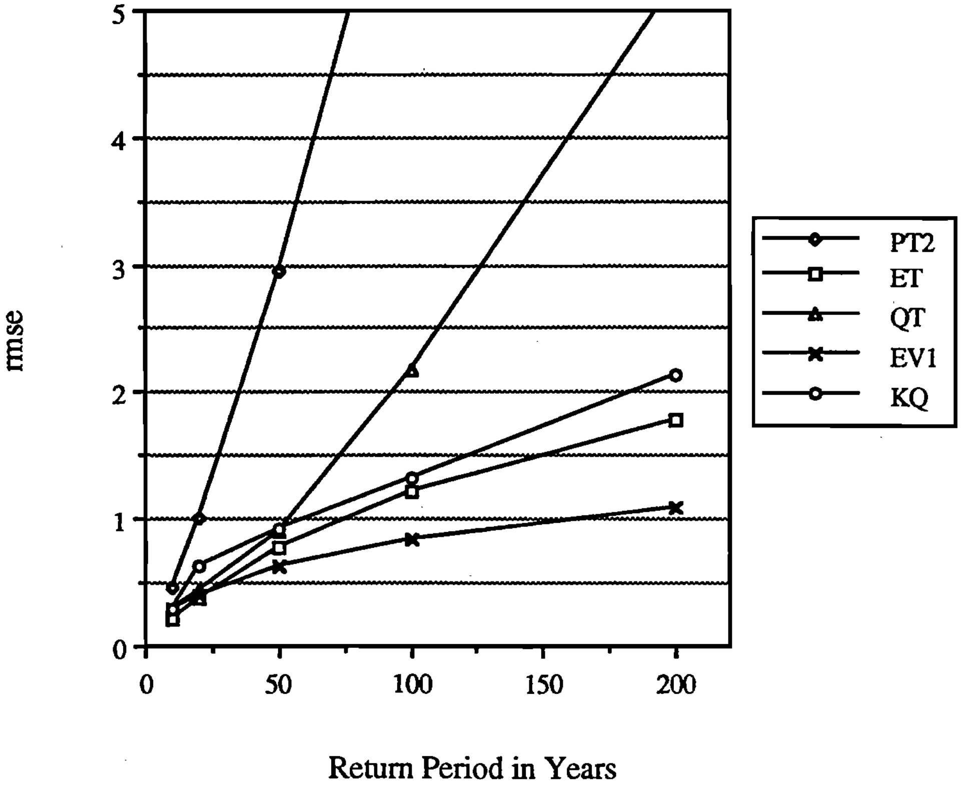


Figure 6

RMSE for Pearson III (0,1,1) data, n = 20

(PT1 is not shown, because its rmse values are off the scale used)

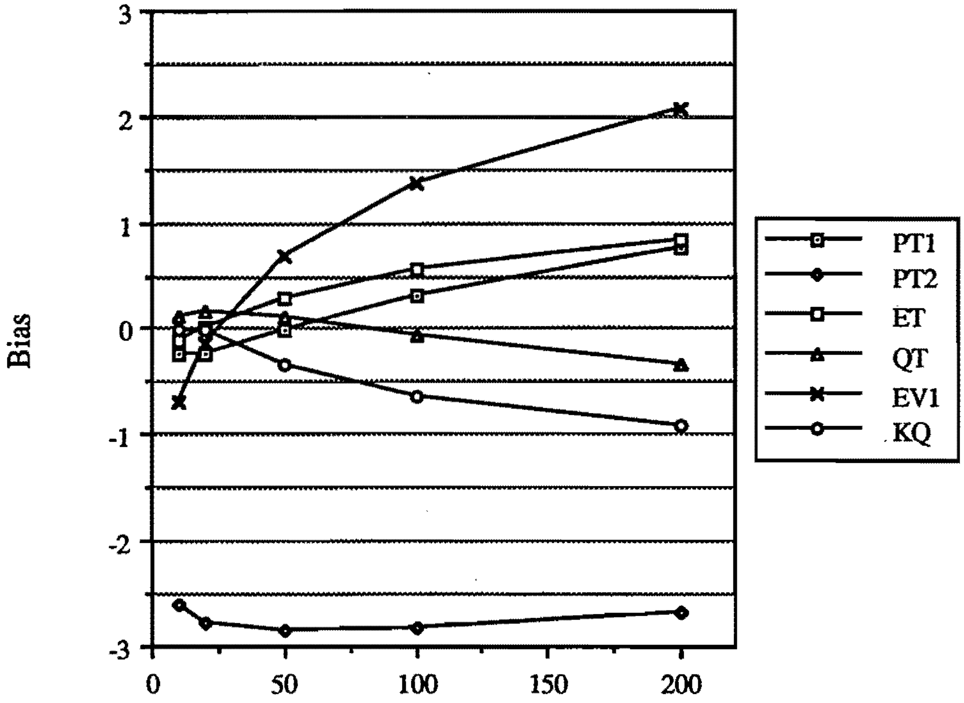


Figure 7

Bias for Mixture data {0.5N(0,1) + 0.5N(3,1)}, n = 20

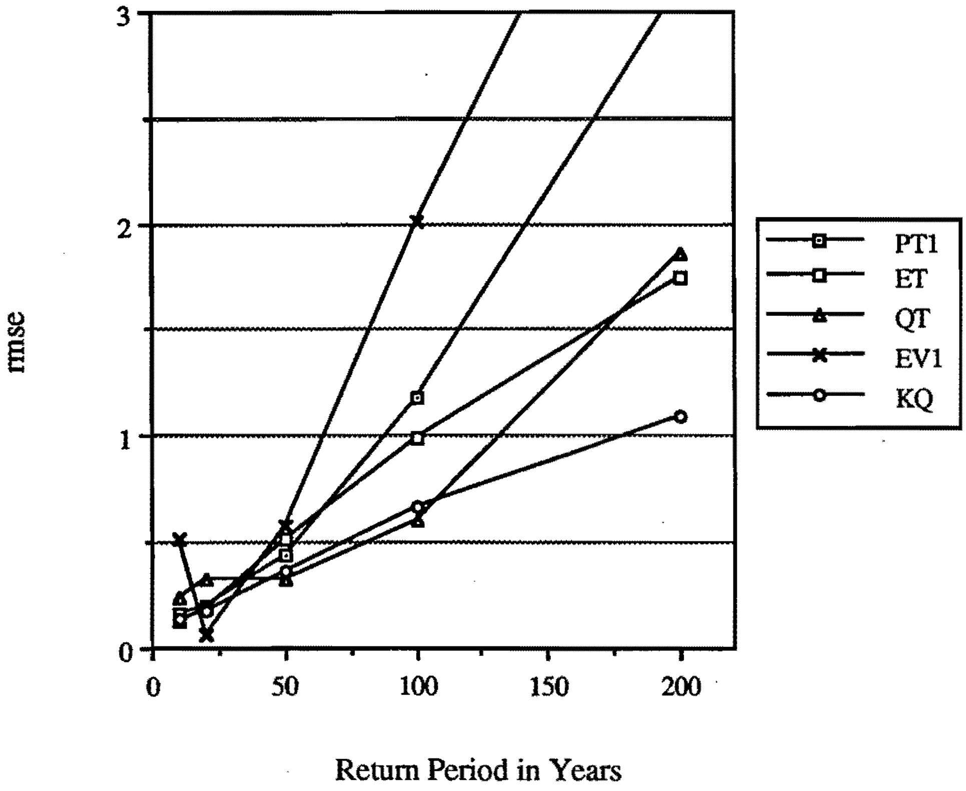


Figure 8

RMSE for Mixture data {0.5N(0,1) + 0.5N(3,1)}, n = 20

(PT2 is not shown, because its rmse values are off the scale used)

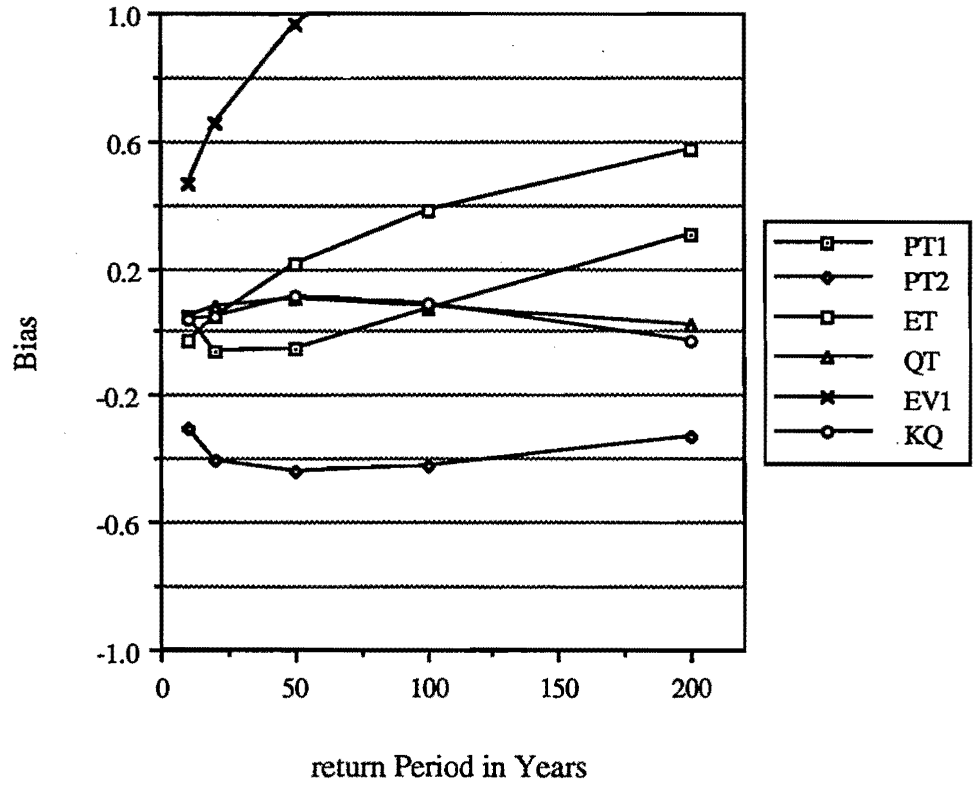


Figure 9

Bias for Normal data N(0,1), n=100

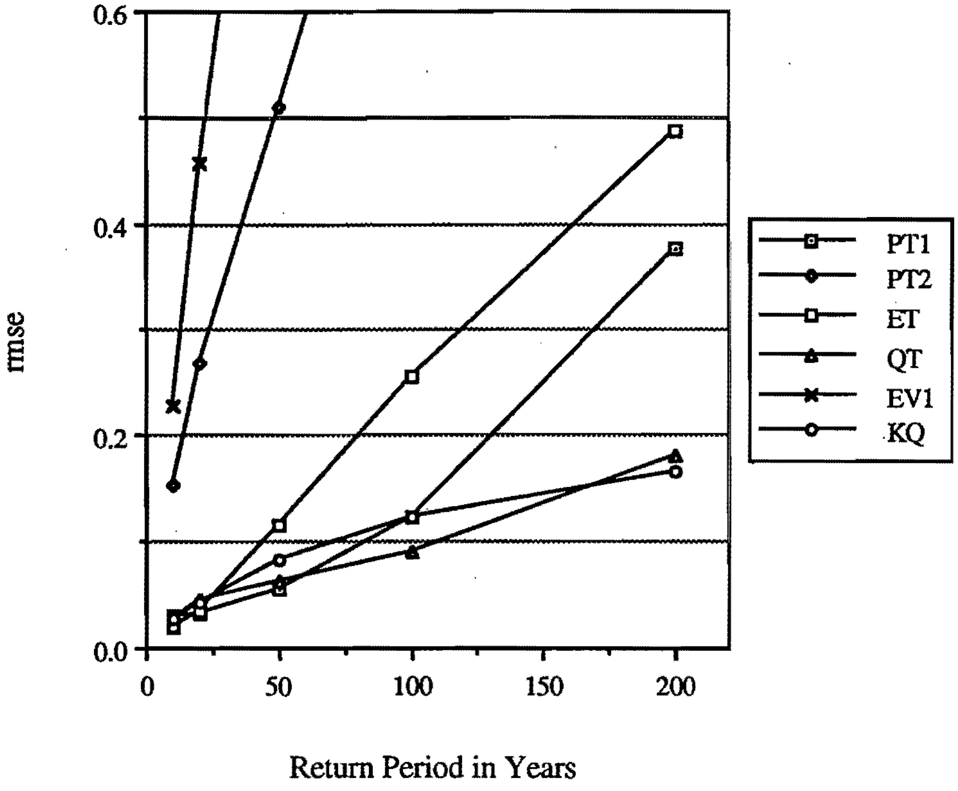


Figure 10

RMSE for Normal data N(0,1), n=100

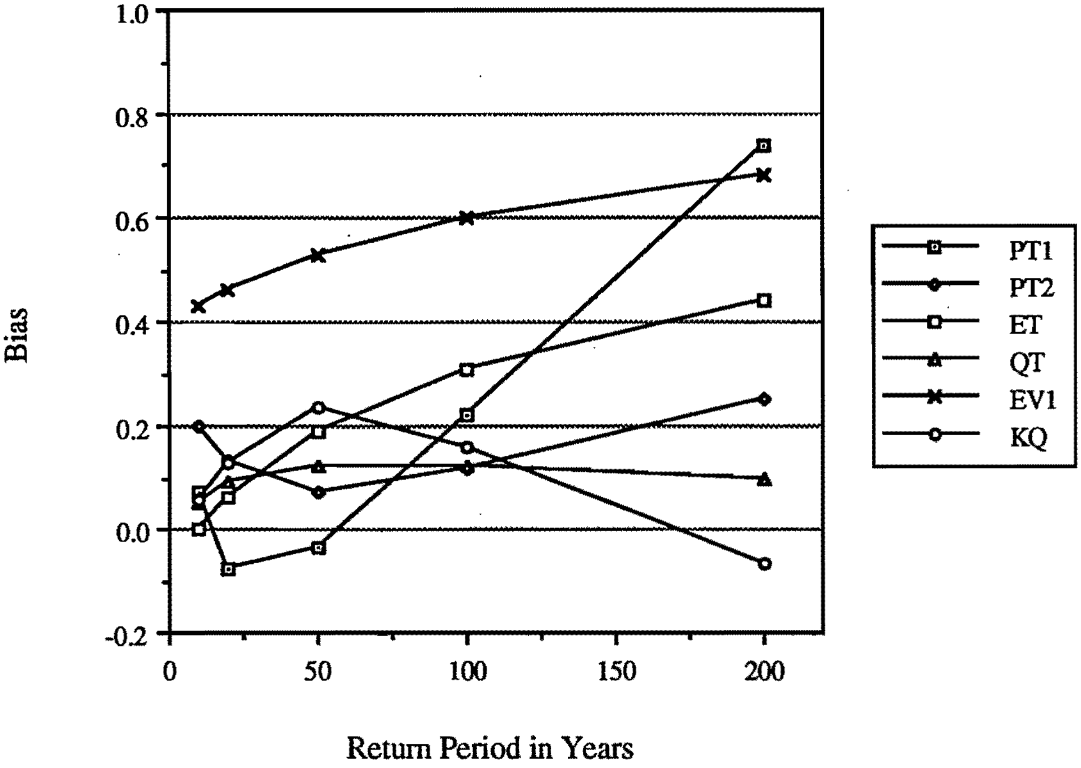


Figure 11

Bias for Pearson III (0,1,1) data, n = 100

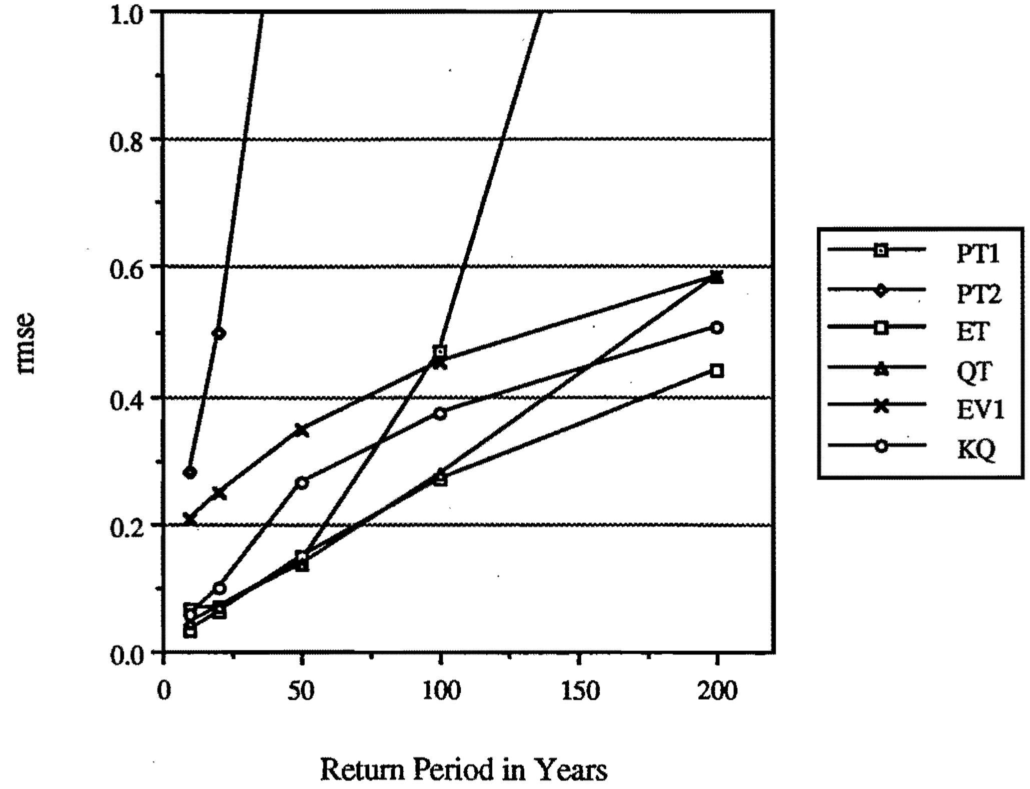


Figure 12
RMSE for Pearson III (0,1,1) data, n = 100

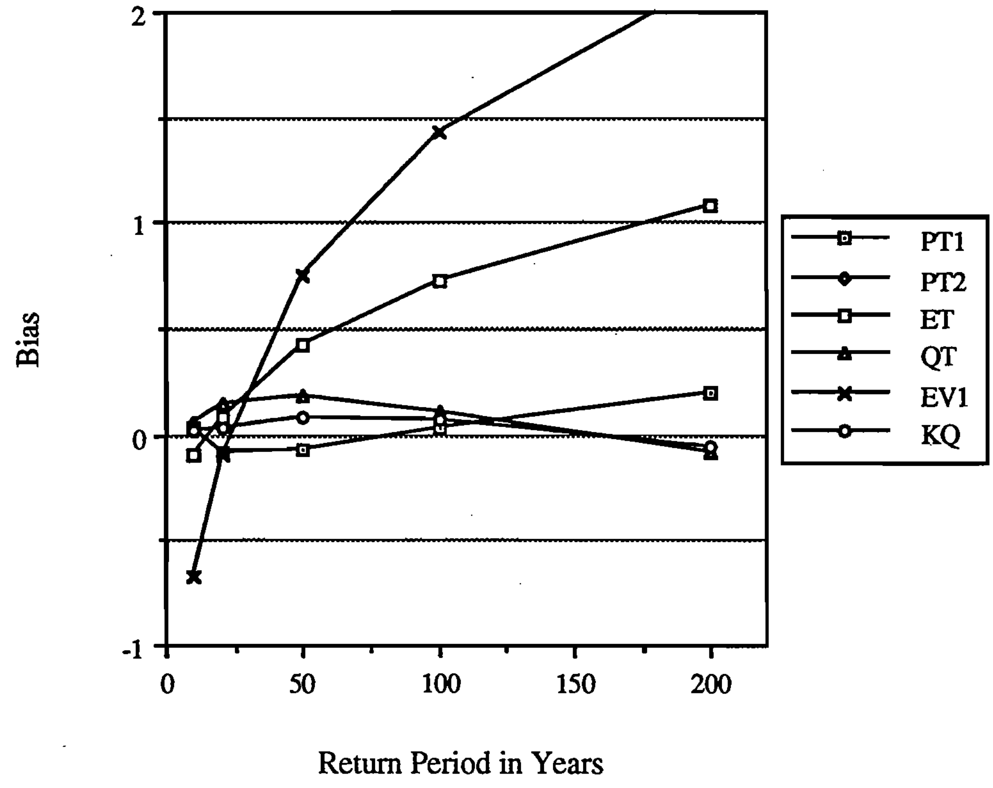
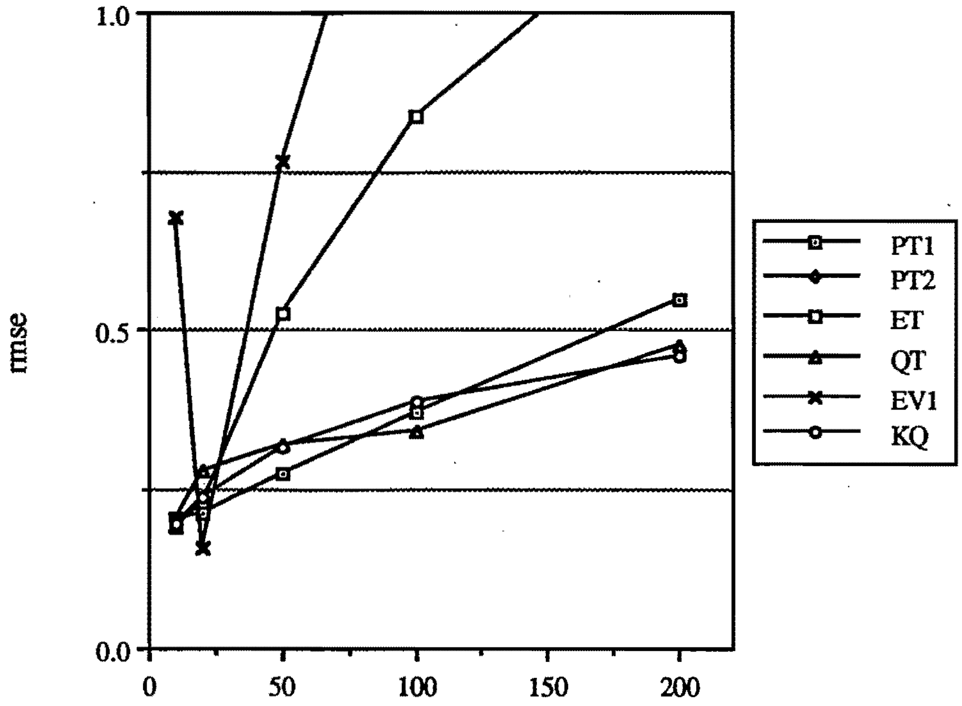


Figure 13

Bias for Mixture data {0.5N(0,1) + 0.5N(3,1)}, n = 100



Return Period in Years

Figure 14

RMSE for Mixture data {0.5N(0,1) + 0.5N(3,1)}, n = 100

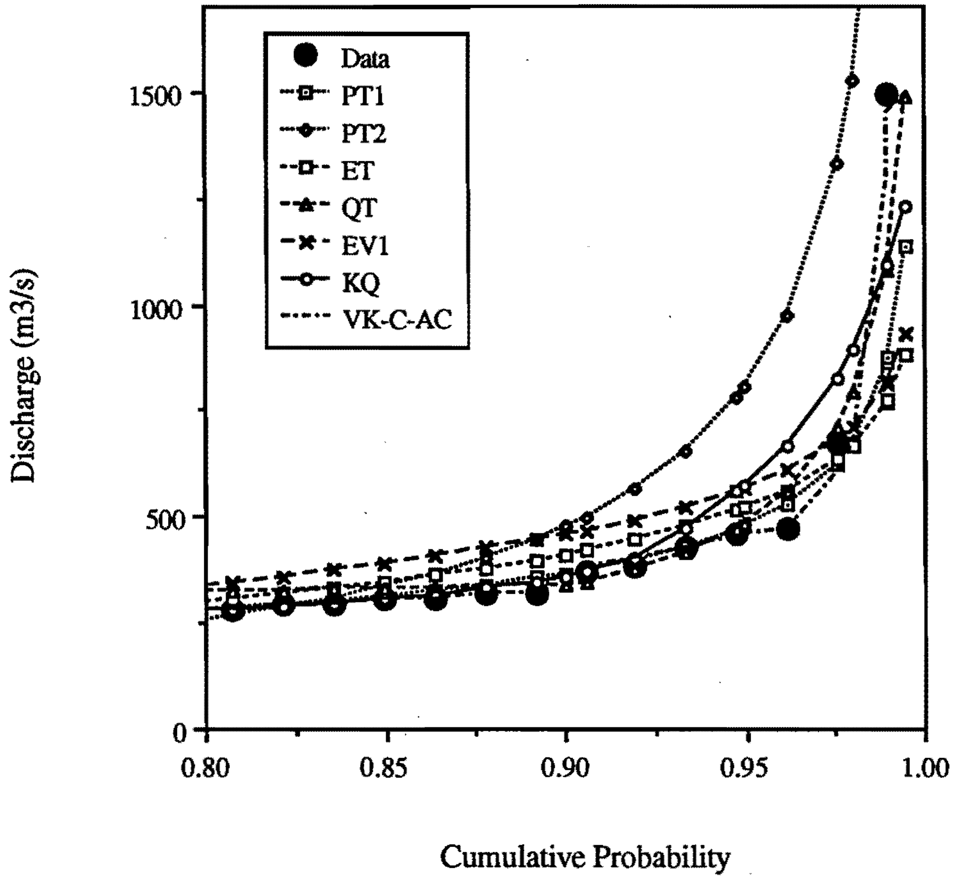


Figure 15

Quantile function estimates for Santa Cruz River annual maximum flood data

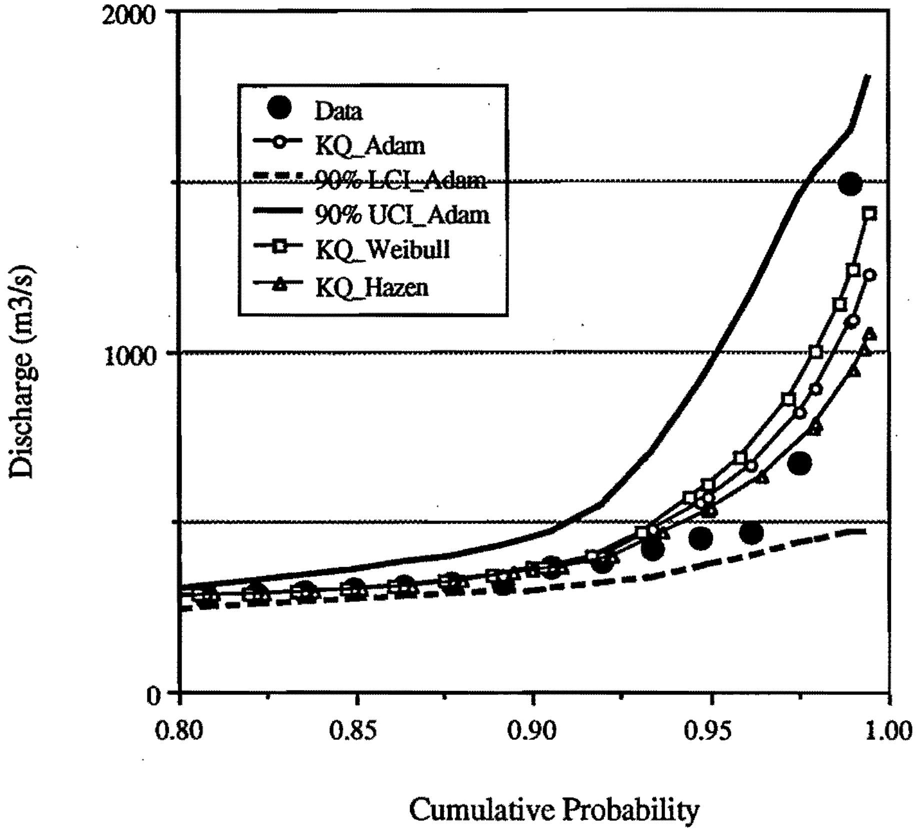


Figure 16

The 90% confidence band of KQ and the graphical comparison of KQ based on plotting formulas for Santa Cruz's River at Tucson, Arizona (1915-1986).The confidence band is constructed by bootstrap technique.

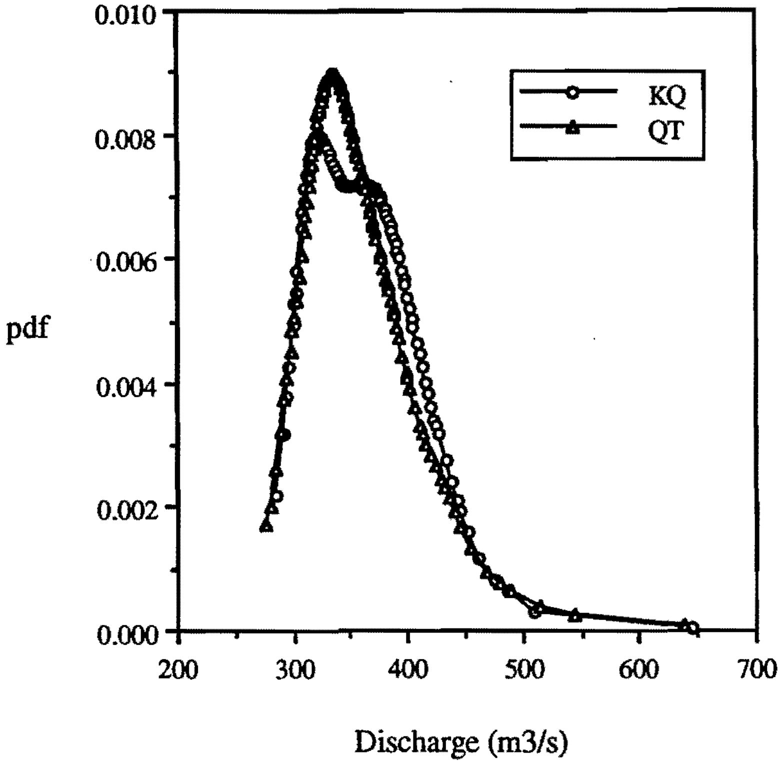


Figure 17

PDF of KQ and QT for estimated values of Bootstrap at p=0.9

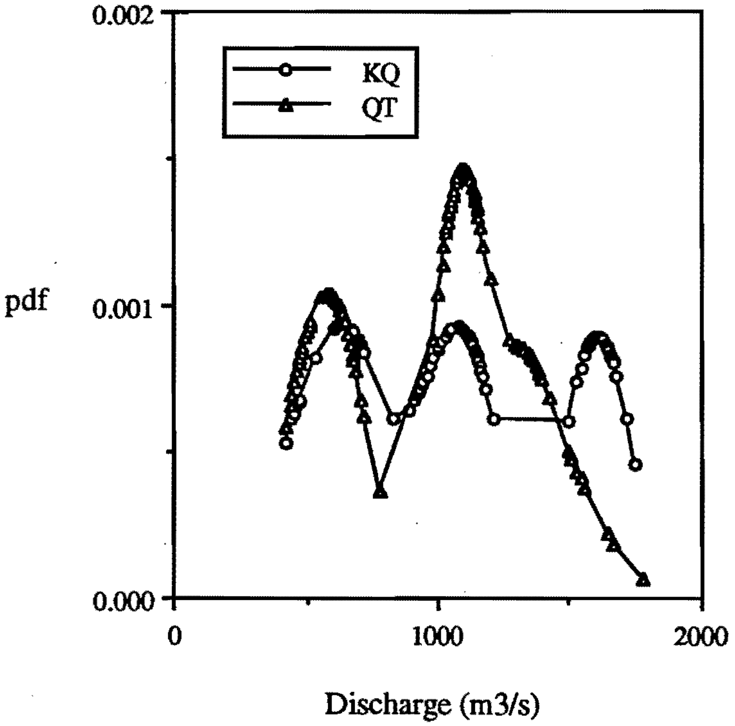


Figure 18

PDF of KQ and QT, for estimated values of Bootstrap at p=0.99