WATER RESOURCES RESEARCH, VOL. 39, NO. 5, 1134, doi:10.1029/2002WR001593, 2003

Flood quantiles in a changing climate: Seasonal forecasts and causal relations

A. Sankarasubramanian and Upmanu Lall[1]

International Research Institute for Climate Prediction, Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York, USA

Received 17 July 2002; revised 16 December 2002; accepted 4 March 2003; published 21 May 2003.

[1] Recognizing that the frequency distribution of annual maximum floods at a given location may change over time in response to interannual and longer climate fluctuations, we compare two approaches for the estimation of flood quantiles conditional on selected "climate indices" that carry the signal of structured low-frequency climate variation, and influence the atmospheric mechanisms that modify local precipitation and flood potential. A parametric quantile regression approach and a semiparametric local likelihood approach are compared using synthetic data sets and for data from a streamflow gauging station in the western United States. Their relative utility in different settings for seasonal flood risk forecasting as well as for the assessment of long-term variation in flood potential is discussed. *INDEX TERMS:* 1821 Hydrology: Floods; 1833 Hydrology: Hydroclimatology; 4522 Oceanography: Physical: El Niño; *KEYWORDS:* teleconnection, seasonal flood forecasting, local likelihood, quantile regression

Citation: Sankarasubramanian, A., and U. Lall, Flood quantiles in a changing climate: Seasonal forecasts and causal relations, *Water Resour. Res.*, *39*(5), 1134, doi:10.1029/2002WR001593, 2003.

1. Introduction

[2] A traditional assumption underlying flood frequency analysis is that the underlying stochastic process is stationary in time, and that the annual maximum flood corresponds to an independent identically distributed (iid) process. However, it is now widely acknowledged that both climate and land use changes modify flood frequency. *Hirschboeck* [1987a, 1987b, 1988] recognized that annual maximum floods at a given site could be due to markedly different rainfall or climate mechanisms that occur in different seasons, and explored the use of mixture models for estimating flood frequency. Changes in flood frequency over paleotimescales [*Porparto and Ridolfi*, 1998; *Knox*, 1993] have also been reported. The recognition that there is interannual to decadal organization in climate, as well as systematic, anthropogenic climate changes that affect flood mechanisms and hence lead to structured temporal variations in flood frequency is more recent [*Katz and Brown*, 1992; *Jain and Lall*, 2000; *Pizaro and Lall*, 2002; *Milly et al.*, 2002; *Franks and Kuczera*, 2002]. The management dilemma posed by the apparent increase in the flood threat to Sacramento, California, from the American River [*National Research Council* (*NRC*), 1995, 1999] is an example of such issues.

[3] This paper focuses on the treatment of changing flood frequency at a given site, where the nonstationarity is derived primarily from structured low-frequency climate variations, and surrogate records of climate indices that represent the essential modes of the underlying climate variations are available. Given these indices, one can (1) consider a diagnostic analysis (as in the work of *Jain and Lall* [2000, 2001]) that relates the flood series to appropriate climate indices; (2) carryout a prognostic analysis that uses climate indices to forecast season ahead (or longer) flood risk; (3) reconstruct flood quantiles over periods prior to the period covered by the historical flood record; and (4) improve regional flood frequency curves by recognizing that the nonoverlapping periods of record across sites may reflect different, yet identifiable climate conditions. Here, the second and third of these analyses are considered in the framework of a regression approach for estimating conditional flood quantiles. The focus is on documenting the relative merits of two methods for conditional flood quantile estimation as a building block toward these two objectives. The flood variable in such a setting may be the peak flow rate over the period of interest, or the n-day (e.g., 3 day) flood volume.

[4] A brief overview of the interconnections between low-frequency oscillations in the climate signals and flood variability is provided in the next section. The conditional flood quantile estimation problem is then introduced in this context and selected methodologies for estimation are reviewed in the third section. A Monte Carlo investigation with synthetic data used to compare two of these methods follows. An application to data from a basin in Montana is then presented.

2. Low-Frequency Climate Variability and Its Relation to Flood Process

[5] Large-scale moisture delivery pathways and their variability often determine the flood potential of a region [*Wendland and Bryson*, 1981; *Hirschboeck*, 1991]. Recent progress in understanding ocean-atmosphere interactions

[1]Also at Department of Earth and Environmental Engineering, Columbia University, New York, New York, USA.

0043-1397/03/2002WR001593$09.00

shows that there are well organized modes of interannual and interdecadal variability in climate that modulate these dominant moisture delivery pathways and has significant projections on continental scale rainfall and flood patterns [*Trenberth and Guillemot*, 1996; *Cayan et al.*, 1999]. Interannual modes such as the El Nino-Southern Oscillation (ENSO) resulting from sea surface conditions in the tropical Pacific Ocean primarily determine the interannual variability in precipitation over North and South America [*Rasmusson and Carpenter*, 1982; *Ropeleweski and Halpert*, 1987; *Halpert and Ropelewski*, 1992]. There are also other dominant decadal and interdecadal climatic modes such as Pacific Decadal Oscillation (PDO) and North Atlantic Oscillation (NAO) that putatively govern the interannual variability in climate over the North America.

[6] ENSO is a quasi-oscillatory mode of coupled ocean-atmosphere interactions in the tropical Pacific with a characteristic narrow band periodicity in the 3–7 year band. During the two phases of ENSO, El Nino and La Nina, anomalous sea surface conditions in the tropical pacific are communicated to the extra-tropics through ocean-atmospheric circulation in the form of upper tropospheric divergence anomalies. These translate into a modulation of the storm tracks over the extra-tropics and exhibit teleconnections influencing the distribution of temperature and precipitation across the globe. Several researchers have found that the interannual variability in hydrologic extremes over the western U.S. is associated with the state of ENSO [*Trenberth and Guillemot*, 1996; *Piechota and Dracup*, 1996; *Jain and Lall*, 2000; *Barlow et al.*, 2001; *Pizaro and Lall*, 2002]. *Cayan et al.* [1999] show that the frequency distribution of daily winter precipitation and winter and spring daily streamflow in the western U.S. exhibits strong and systematic responses to the two phases of ENSO (El Nino and La Nina). *Haston and Michaelsen* [1994] found that extremes in precipitation over the coastal regions of California occur during El Nino conditions based on 600 yearlong reconstructed annual rainfall from tree ring chronology. *Pizaro and Lall* [2002] show that the annual maximum peak over the western U.S. is significantly correlated to the modes of ENSO and PDO. *Jain and Lall* [2000] illustrate that ENSO may actually represent many timescales of long-term variability and hence floods over any period of record may not adequately represent the frequency of floods in a subsequent period of similar length. This will invariably lead to a ''surprise'' for users of the frequency curves estimated from the existing record. However, if multicentury ENSO dynamics were well understood and the fluctuation of flood potential was associated with these dynamics, then one may be able to characterize the nature of this ''surprise''. Developing such an association is a goal of the current paper.

[7] *Mantua et al.* [1997] identified a pattern of variability in the ocean-atmosphere interactions over the Pacific Ocean having a characteristic timescale of 18–22 years, which they called the Pacific Decadal Oscillation (PDO). This North Pacific climate mode putatively influences the snowpack variability and winter surface climate over the western U.S, thereby influencing the timing and magnitude of flood peaks [*Mantua et al.*, 1997; *Cayan*, 1996; *Jain and Lall*, 2000, 2001; *Pizaro and Lall*, 2002]. Several investigators have also tried to understand the combined effect of ENSO and PDO on the interannual climatic variability over the U.S. [*Gershunov et al.*, 1998; *Gershunov and Barnett*, 1998; *McCabe and Dettinger*, 1999]. The PDO phase may modulate the effects of ENSO that can change the sign and strength of the ENSO effects on the streamflow over the western U.S. In other words, extra-tropical interdecadal North Pacific oscillations can substantially modulate the mean position of the jet stream that brings moisture into the continents, thereby reducing or enhancing the influence of tropical oscillations like El Nino. *Jain and Lall* [2001] identified space-time-frequency patterns that connect floods at multiple locations in the western United States with concurrent hemispheric Sea Surface Temperature and Sea Level Pressure patterns. Quasi-biennial, interannual and interdecadal joint modes of variation with a distinct spatial correlation structure in each frequency band are identified. Thus low-frequency climate and flood variations have been connected to each other.

[8] A number of papers have been published on the potential and observed impacts of anthropogenic climate change at secular timescales on flood potential. We do not consider these factors and mechanisms here other than their potential manifestation through changes in the modes of low-frequency climate variability considered here. The methods considered will allow the consideration of specific measures of land use change or surrogate measures of climate change as predictors in addition to the indices of low-frequency climate variability.

3. Methods

[9] Define Q as a flood variable of interest, e.g., the annual maximum flow, the annual maximum n-day flow, or the number of days in a season or a year for which the flow Q exceeds a threshold q*. The inference of the pth quantile, Q_{pt} (quantile) of Q, for year t, conditional on some set of m climatic indices (or other predictors), $\mathbf{X}_t = [x_{1t}\ x_{2t} \cdots x_{mt}]$, is of interest. To do this, an estimate of the conditional probability density function $f(Q_t|\mathbf{X}_t)$, or the conditional distribution function $F(Q_t|\mathbf{X}_t)$ from the historical data $\{Q_t, \mathbf{X}_t, t = 1\ldots n\}$:

$$F_p(Q_t|X_t) = \int_{-\infty}^{Q_{pt}} f(Q_t|X_t)dQ = p \tag{1a}$$

$$Q_{pt} = F_p^{-1}(Q_t|X_t) \tag{1b}$$

[10] The conventional approach to estimate the conditional distribution function $F(Q_t|\mathbf{X}_t)$ is to assume that the joint probability density function $f(Q_t, \mathbf{X_t})$ follows a particular distribution and then to estimate its parameters. Quantile estimates obtained from this approach vary widely as it depends primarily on the nature of the tails of the conditional probability density function $f(Q_t|\mathbf{X}_t)$. *Jain and Lall* [2000] tried to overcome this by assuming $f(Q_t|\mathbf{X}_t)$ to be lognormal, with its mean and variance varying in time conditional on the state of ENSO and PDO over a 30 year moving window. Another approach would be to estimate the conditional distribution function $F(Q_t|\mathbf{X}_t)$ nonparametrically using kernel and k-nearest neighbor (k-NN) methods [*Yu and*

Jones, 1998; *Bhattacharya and Gangopadhyay*, 1990]. Both these methods have limitations. The kernel based approach of conditional quantile estimation suggested by *Yu and Jones* [1998] may be difficult to implement in practice, whereas the nearest neighbor approach of *Bhattacharya and Gangopadhyay* [1990] performs poorly near the boundaries of predictors [*Yu*, 1999]. *Yu* [1999] suggested a two-step approach to overcome these limitations by first estimating the quantile using a k−NN approach and then smoothing the estimated quantiles using a kernel function. However, the additive model structure used by the combination approach required for higher dimension data (for increased number of predictors) tends to be computationally intensive (Yu and Lu, personal communication, 2002). A second, but a different approach that focuses on developing a regression relationship between Q_{pt} in (1b) and the predictors $\mathbf{X_t}$, is the quantile regression developed by *Koenkar and Bassett* [1978]. Quantile regression is a parametric method to estimate conditional quantiles by minimizing the sum of asymmetrically weighted absolute deviation by giving different weights for positive and negative residuals using simple optimization techniques. The advantage of this method is that it is easy to implement and can be extended even under nonlinear situations [*Koenkar and Park*, 1996]. Recently, *Davison and Ramesh* [2000] suggested a semiparametric approach to estimate the trend in the quantiles using a local−likelihood smoothing. This approach is similar to the ad hoc approach of *Jain and Lall* [2000], but the emphasis was on the time trend in the quantiles. In this study, we consider two approaches, the parametric quantile regression approach of *Koenkar and Bassett* [1978] and the semiparametric approach of *Davison and Ramesh* [2000] for estimating flood quantiles conditioned on the climatic indices that carry the signal of low−frequency climate variation. The performance of these two methods is first compared on a synthetic data set and then evaluated for potential application in issuing seasonal flood forecasts in a hydrologic basin.

3.1. Quantile Regression

[11] The first method considered is quantile regression as implemented by *Koenkar and Bassett* [1978]. Define the pth conditional quantile through the regression:

$$Q_{pt} = \Psi_p(\mathbf{X_t}) + \nu_{pt} \tag{2}$$

where $\psi_p(.)$ is a linear or nonlinear function relating the pth conditional quantile to the climatic indices and ν_{pt} is a noise process with the pth quantile zero and variance σ_p^2. The noise process may in general be homoskedastic (σ_p^2 = a constant) or heteroskedastic (i.e., $\mathrm{Var}(\nu_{pt}) = \sigma_p^2(\mathbf{X}_t)$). *Koenkar and Bassett* [1978] consider the homoskedastic case. The function $\Psi_p(\mathbf{X}_t)$ is estimated by solving the following minimization problem

$$\min_{\Psi_p(\mathbf{X_t})} \sum_{t=1}^{n} R_p\left(Q_{pt} - \Psi_p(\mathbf{X_t})\right) \tag{3}$$

where, $R_p(u) = u(p - I\{u\}) = \frac{|u|+(2p-1)u}{2}$ and $I\{u\}$ denotes the indicator function with

$$I\{u\} = \begin{cases} 1 \text{ if } u < 0 \\ 0 \text{ otherwise} \end{cases}.$$

[12] As an example, if the regression function ψ_p in (2) is linear, and we consider the median (p = 0.5), then the regression is defined through

$$\Psi_p(\mathbf{X_t}) = \beta_p \mathbf{X}_t \tag{4}$$

where β_p is an m*1 vector of regression coefficients for the pth quantile, and the minimization in (3) corresponds to least absolute deviation regression. *Koenkar and D'Orey* [1987] provide an algorithm to estimate β using linear programming techniques for any given 'p'. Fortran subroutines for implementing the quantile regression in (3) are available in Statlib (http://lib.stat.cmu.edu/). Bayesian extensions that incorporate parameter uncertainty into the estimation of β were pursued by *Yu and Moyeed* [2001]. *Koenkar and Park* [1996] present optimization algorithms for estimating the parameters of (3) if $\psi(.)$ follows a specific nonlinear function. Semiparametric approaches that minimize the check function with a penalized likelihood function have also been pursued to estimate conditional quantiles [*Koenker et al.*, 1992]. Here, we used the parametric approach of *Koenkar and D'Orey* [1987], with $\psi(.)$ taken to be linear.

3.2. Local Likelihood Model

[13] *Davison and Ramesh* [2000] present an alternate semiparametric method that estimates the parameters of the assumed flood frequency distribution conditional on predictors using local likelihood estimation, based on local neighborhood in the predictor state-space. They were concerned with a time trend in the parameters and used a time index as a predictor. Here, we extend this approach to consider multiple climate indices as predictors.

[14] Consider the conditional pdf $f(Q_t|\mathbf{X}_t)$ with parameters $\theta(\mathbf{X_t})$. The parameters $\theta(\mathbf{X_t})$ carrying the conditional information regarding the probability model $f(Q_t|\mathbf{X}_t)$ are approximated through a linear function in the neighborhood of $\mathbf{X_t}$. For instance, in the case of a two parameter distribution, if $\theta(\mathbf{X}_t) = [\mu(\mathbf{X}_t)\ \sigma(\mathbf{X}_t)]$ represent the location ($\mu(\mathbf{X_t})$) and the scale ($\sigma(\mathbf{X_t})$) parameters of the distribution, then $\mu(\mathbf{X}_t) = \lambda_0 + \sum_{k=1}^{m}\lambda_k\left(x_{kj} - x_{kt}\right)$ and $\sigma(\mathbf{X}_t) = \gamma_0 + \sum_{k=1}^{m}\gamma_k\left(x_{kj} - x_{kt}\right)$ can be represented as a linear function of m predictors where j denotes all the data points ($\mathbf{X}$) receiving weights w_j. The local likelihood method estimates $\theta(\mathbf{X_t})$ by maximizing the likelihood of the sample in such a way that data points ($\mathbf{X_j}$) lying closer to $\mathbf{X_t}$ receive more weightage. To assign appropriate weightage w_j for each $\mathbf{X_j}$, which lies closer to $\mathbf{X_t}$, a kernel function that receives finite support about the point of estimate $\mathbf{X_t}$ is used. A product form of the Epanechnikov kernel in (5) was used [*Pagan and Ullah*, 1999].

$$w_j(\mathbf{X}_t) = \left\{ \prod_{k=1}^{m}\left(1 - u_{kj}^2\right) \quad \text{if} \left|u_{kj}\right| \le 1 \quad \forall k \right\} \\ 0 \quad \text{otherwise} \tag{5}$$

where $u_{kj} = \left\{\frac{x_{kt} - x_{kj}}{h_k}\right\}$ and h_k is a bandwidth associated with the kth predictor.

[15] The parameters of the method are thus m bandwidths and then 2m+2 coefficients (λ_k, γ_k, $k = 0, 1, \ldots, m$ in estimating $\theta(\mathbf{X}_t) = [\mu(\mathbf{X}_t)\ \sigma(\mathbf{X}_t)]$) for the neighborhood of the point of estimate. The bandwidths h_k can be obtained by specifying that each point of estimate have at least (usually substantially higher) 2m+2 observations. Cross-validated

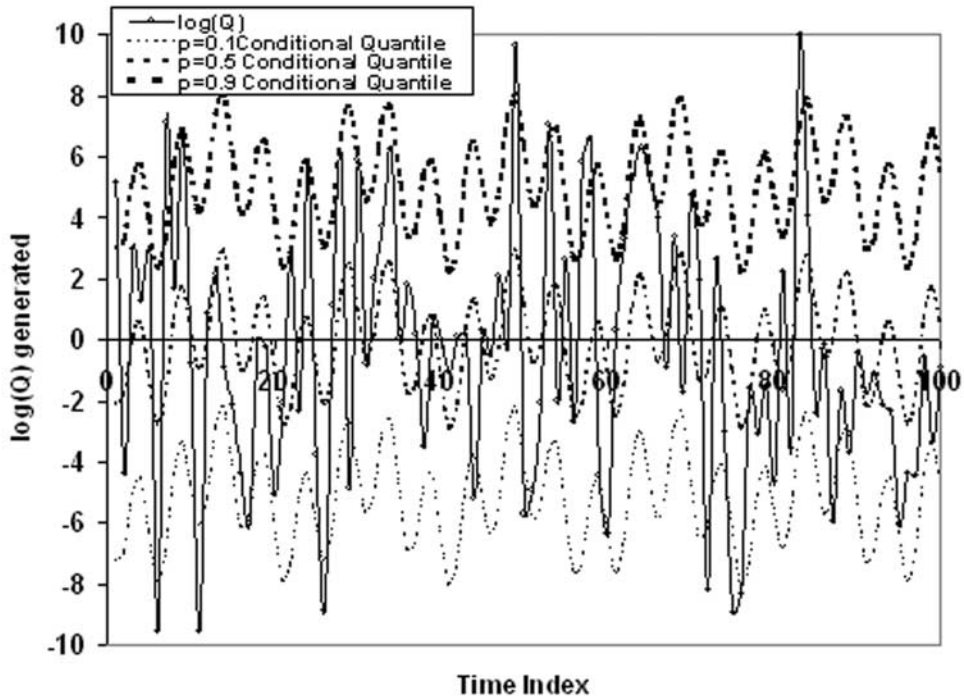

Figure 1. Illustration of conditional flood quantile estimation. Figure 1 shows a realization of log(Q) generated under homoscedastic variance using equation (8). The quantiles shown are the population quantiles for p = 0.1, 0.5, and 0.9 for each time step. Note that the actual flood peak in a given year does not necessarily match with the direction of departure of the conditional quantile from the corresponding unconditional quantile.

maximum likelihood in (6) is also commonly used to choose the bandwidths.

[16] Leave one-out cross validation is carried out by leaving out the response (Q_t) and predictors ($\mathbf{X}_t$) from the observed data set (Q_t, $\mathbf{X}_t$, t = 1, 2, . . ., n) and the parameters ($\boldsymbol{\theta}_{-t}(\mathbf{X}_t)$,−t denoting leave one-out cross validation estimate) are estimated using the rest of the (n − 1) observations where n is the total length of observed records in a given site. The entire set of parameters and bandwidths can be obtained by maximizing the cross-validated local log likelihood

$$\ell_{CV}\left(\hat{\boldsymbol{\theta}}_{-t}(\mathbf{X}_t), \hat{\mathbf{h}}\right) = \sum_{\substack{j=1 \\ j\neq t}}^{n} w_j(\mathbf{X}_t) \log\left(f_{Q_t|\mathbf{X}_t;\hat{\boldsymbol{\theta}}_{-t}(\mathbf{X}_t)}\right) \qquad (6)$$

with respect to $\hat{\boldsymbol{\theta}}_{-t}(\mathbf{X}_t)$ and $\hat{\mathbf{h}}$. The cross-validated local log likelihood in (6) estimates the distribution of Q_t conditioned on the predictors $\mathbf{X}_t$ by estimating the parameters $\boldsymbol{\theta}_{-t}(\mathbf{X}_t)$. The shuffled complex evolution algorithm [*Duan et al.*, 1992] was used to perform the maximization of (6) at each candidate point of estimate $\mathbf{X}_t$. Thus the bandwidths and the parameters of the local distribution are estimated locally at each point of estimate $\mathbf{X}_t$. The cross-validated conditional flood quantile $[\hat{Q}_{pt}]_{-t}$ is estimated by assuming the local density function to be lognormal with the locally estimated parameters $\hat{\boldsymbol{\theta}}_{-t}(\mathbf{X}_t)$ and $\hat{\mathbf{h}}$.

4. Conditional Flood Quantile Estimation: A Simulation Study

[17] A Monte Carlo simulation experiment with synthetic data is used to compare the two methods described in the previous section in an idealized setting designed to replicate the cyclostationary behavior (periodic modes with incommensurate frequencies) expected to be present under ENSO and PDO. Two cases are considered: (1) nonstationarity in the mean of the annual flood variable with a constant variance of the noise process (homoscedastic), and (2) nonstationarity in the mean and variance of the annual maximum peak (heteroskedastic).

4.1. Experiment Design

[18] Consider that the annual maximum flood Q_t in year t arises from a lognormal distribution. This corresponds to a model:

$$y_t \approx N\left(\mu_y(t), \sigma_y(t)\right) \qquad (7)$$

where $y_t = \log(Q_t)$, $\mu_Q(t) = \exp(\mu_y(t) + \sigma_y^2(t))$ and $\sigma_Q^2(t) = \exp^2(\mu_y(t))[\exp(\sigma_y^2(t) * \exp\sigma_y^2(t) - 1)]$. Then for the first case (homoskedastic), the parameters of the distribution are assumed to vary as:

$$\begin{aligned} \mu_{yt} &= C_1 x_{1t} + C_2 x_{2t} \\ \sigma_{yt} &= C \end{aligned} \qquad (8)$$

where C is a constant variance, C_1 and C_2 are coefficients, and x_1 and x_2 are two climate predictors. For the second case (heteroskedastic), the corresponding population parameters are:

$$\begin{aligned} \mu_{yt} &= C_1 x_{1t} + C_2 x_{2t} \\ \sigma_{yt} &= C_v \mu_{yt} \end{aligned} \qquad (9)$$

where C_v is a constant coefficient of variation.

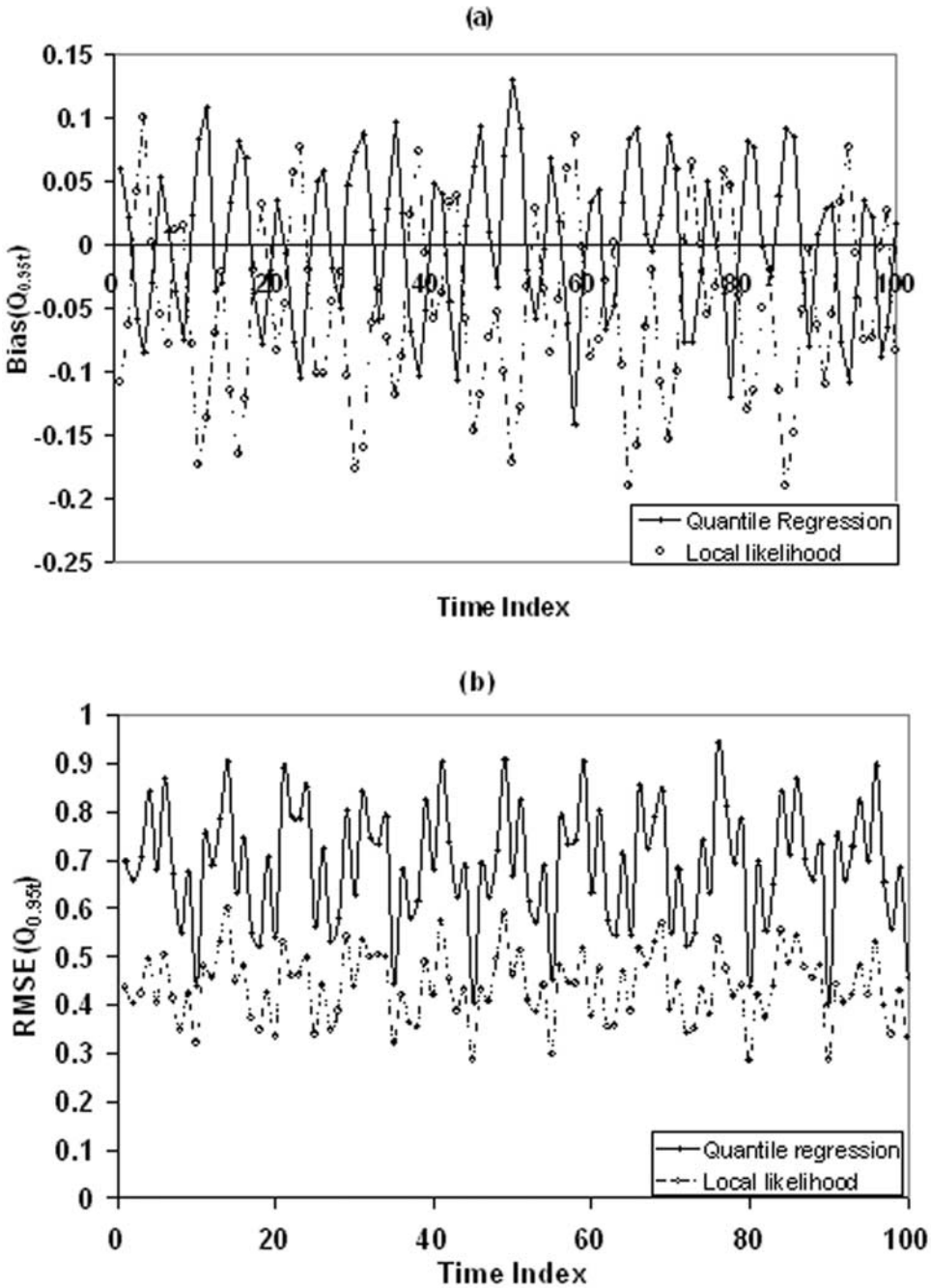

Figure 2. Monte Carlo performance comparison of two methods under leave one out cross validation for the homoscedastic synthetic model in the study: (a) bias($Q_{0.95t}$) and (b) root mean square error ($Q_{0.95t}$). The average bias and RMSE relative to the population conditional quantiles across the 100 years are 0.01 and 0.35 for quantile regression, and are −0.03 and 0.21 for local likelihood. The bias and RMSE are averaged over 1000 realizations.

[19] The predictors are modeled as periodic modes with incommensurate frequencies ω_1 and ω_2:

$$x_{1t} = a\ \sin(\omega_1 t + \phi_1)$$
$$x_{2t} = b\ \sin(\omega_2 t + \phi_2) \quad (10)$$

where ϕ_1 and ϕ_2 are the phase angles and 'a' and 'b' are the amplitudes of two climate signals. For the example here, a 5 year (center of the ENSO band) and an 18 year period (approximately the PDO band) was used for these two predictors, with $\phi_1 = 180$ and $\phi_2 = 0$, $a = 1.352$, $b = 1.743$, $C_1 = 1.352$, and $C_2 = -0.678$. The amplitudes were estimated from a Fourier analysis of the NINO3 and the PDO series, and the coefficients C_1 and C_2 correspond to those estimated for the Blacksmith Fork, Hyrum streamflow (analyzed by *Jain and Lall* [2000]) conditional on NINO3 and PDO. For the heteroskedastic case, C_v was taken to be 0.12. Selected

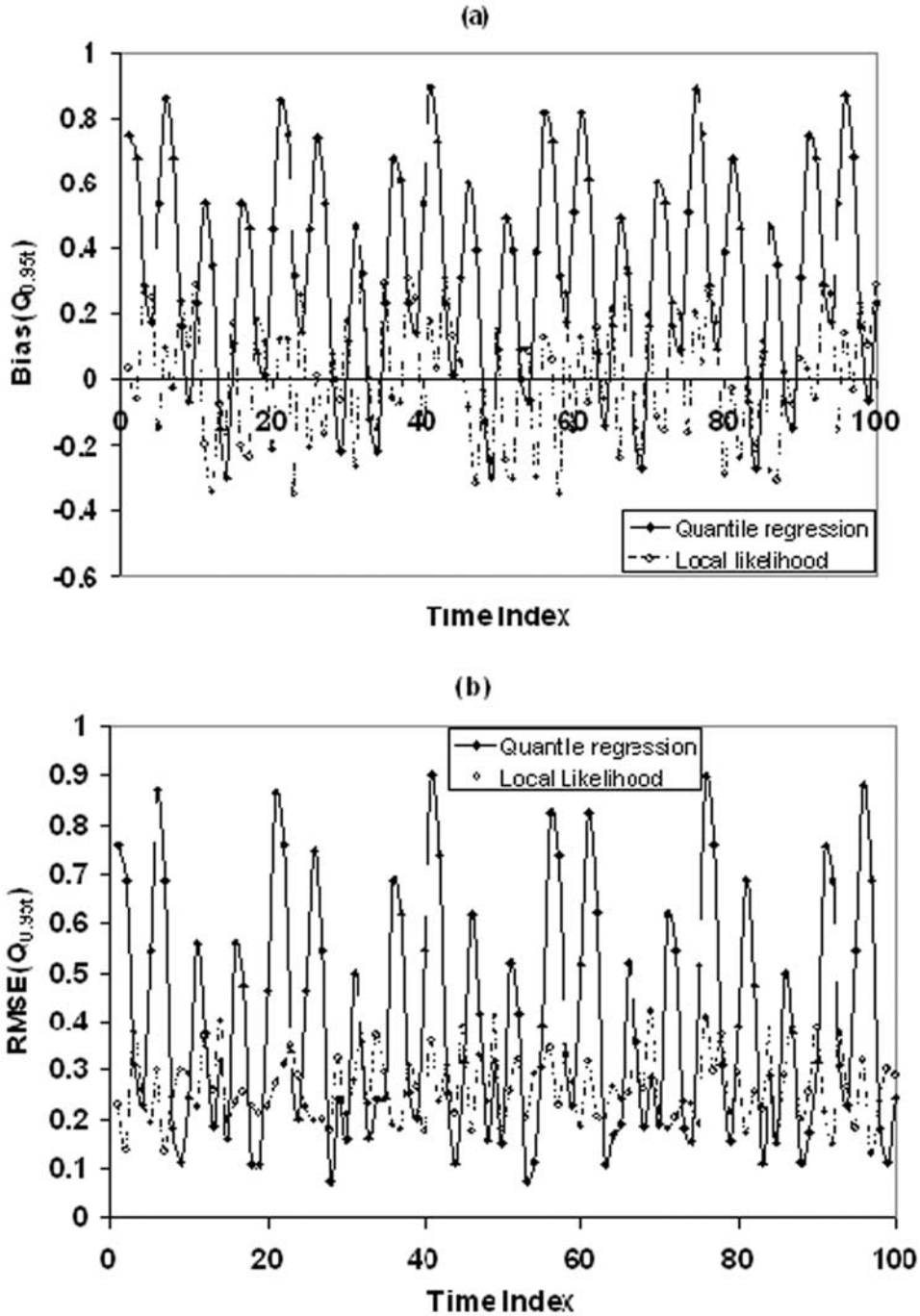

Figure 3. Monte Carlo performance comparison of two methods under leave one out cross validation for the heteroskedastic synthetic model in the study: (a) bias($Q_{0.95t}$) and (b) root mean square error ($Q_{0.95t}$). The average bias and RMSE relative to the population conditional quantiles across the 100 years are −0.12 and 0.68 for quantile regression and are 0.03 and 0.64 for local likelihood. The bias and RMSE are averaged over 1000 realizations.

population quantiles and one realization from this model for the homoskedastic case (with C = 4.0) are illustrated in Figure 1.

[20] Using these parameters, 1000 realizations of Q_t, x_{1t} and x_{2t} with record length n = 100 are generated and cross-validated estimates of the pth quantile $[\hat{Q}_{pt}]_{-t}$ are obtained using both quantile regression and the local likelihood method. This implies that under each realization, we obtain 100 cross-validated estimates of $[\hat{Q}_{pt}]_{-t}$ that correspond to each year. The data was log transformed in both cases before application. The estimation techniques are compared in terms of their cross-validated bias and root mean square error (rmse) in estimating Q_{pt}.

[21] The cross-validated bias and root mean square error averaged over the 1000 realizations are computed at each time t:

$$Bias\left(\hat{Q}_{pt}\right) = \frac{1}{1000}\sum_{i=1}^{1000}\left[\hat{Q}_{pt}\right]_{-it} - Q_{pt} \qquad (11)$$

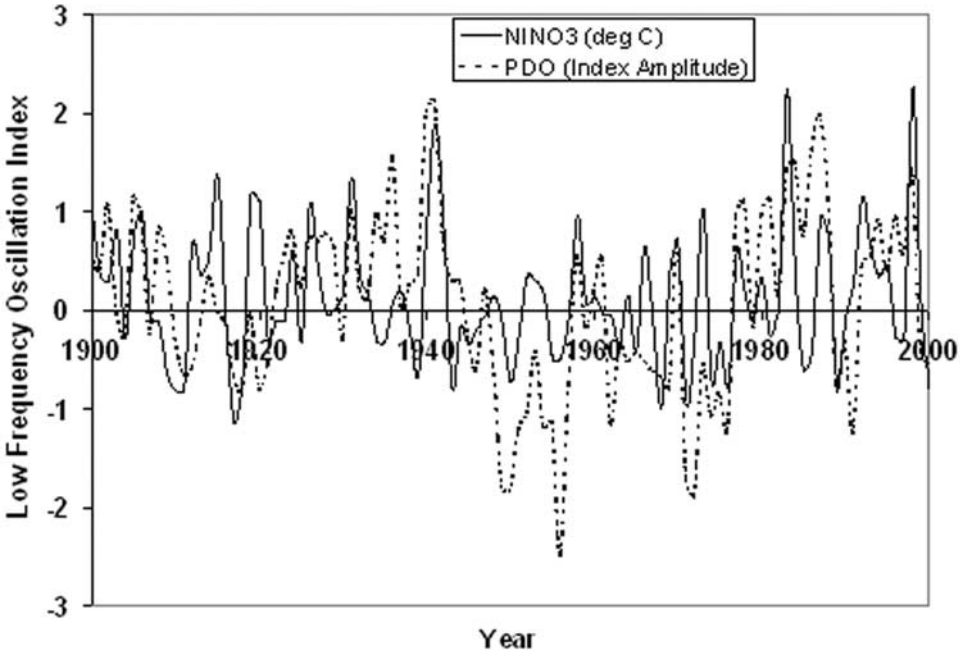

Figure 4. Wintertime average (January–February–March–April) of the NINO3 and PDO indices.

$$Rmse\left(\hat{Q}_{pt}\right) = \sqrt{\frac{1}{1000}\sum_{i=1}^{1000}\left(\left[\hat{Q}_{pt}\right]_{-it} - Q_{pt}\right)^2} \qquad (12)$$

4.2. Results of the Monte Carlo Experiment

4.2.1. Homoskedastic Case

[22] The cross-validated performance of the two methods in terms of the two performance measures is illustrated in Figure 2, for p = 0.95, the 20 year flood. The average bias and average RMSE relative to the population conditional quantiles across the entire 100 years are 0.010 and 0.348 respectively, for quantile regression and −0.030 and 0.214, respectively for local likelihood. The higher absolute bias of local likelihood is manifest at points of high curvature in the target function, as expected. The bias in quantile regression is purely due to sampling [*Buchinsky*, 1995]. However, somewhat surprisingly, the RMSE performance of local likelihood is better. Results for p = 0.1, 0.5, 0.75 and 0.9 are similar.

4.2.2. Heteroskedastic Case

[23] From Figure 3 we observe that the local likelihood estimator now outperforms quantile regression in terms of both bias and rmse. The average bias and RMSE relative to the population conditional quantiles across the 100 years are −0.120 and 0.679 for quantile regression, and are 0.034 and 0.635 for local likelihood.

[24] While the bias and variance of quantile regression increase as expected; the bias of the local likelihood is similar, while the variance is higher reflecting the greater complexity of this setting. Thus in terms of cross validated RMSE, for the case of linear model structure, it appears that local likelihood is a more effective method since it is competitive in both situations considered. If the relationship between the flood quantiles and the climate predictors was nonlinear (as illustrated by *Jain and Lall* [2000]), then the local likelihood method would still be directly applicable as a somewhat biased (the local bias^2 is proportional to the local curvature of the target function) estimator, while the parametric quantile regression approach would require exploration of different nonlinear functions in a multivariate setting, as well as special treatment for heteroskedasticity of the noise process. Another difference is that since each quantiles is estimated independently by the quantile regression process for each value of p, it is conceivable that the estimated quantile regression curves will cross for different values of p. While this is understandable in the context of sampling variability, it is an undesirable outcome. Local likelihood does not suffer from this malady, since the quantiles at a particular predictor state are estimated from a common local density function. However, as one moves in the neighborhood of a point, again due to separate optimizations, there may be marked differences in the estimated quantiles due to sampling variability and its effects on parameter selection. A Bayesian approach following *Holmes and Adams* [2002] would be useful to formally address such uncertainties, but was not pursued in this work.

[25] For the local likelihood method, selection of larger bandwidths increases the potential estimation bias and smaller bandwidths increase the variance. Methods other than cross validation are also available to choose the bandwidth. A plug-in method that minimizes the asymptotic mean square error of the estimated quantile is presented by *Loader* [1999]. The most significant issue is that of choosing the bandwidth locally or globally. The procedure described in (6) leads to a large number of parameters being estimated. The Monte Carlo experiment described earlier was modified to consider global (i.e., m bandwidth parameters common to all points of estimate) rather than local bandwidths. The resulting bias and RMSE performance was comparable and local overfitting was consider-

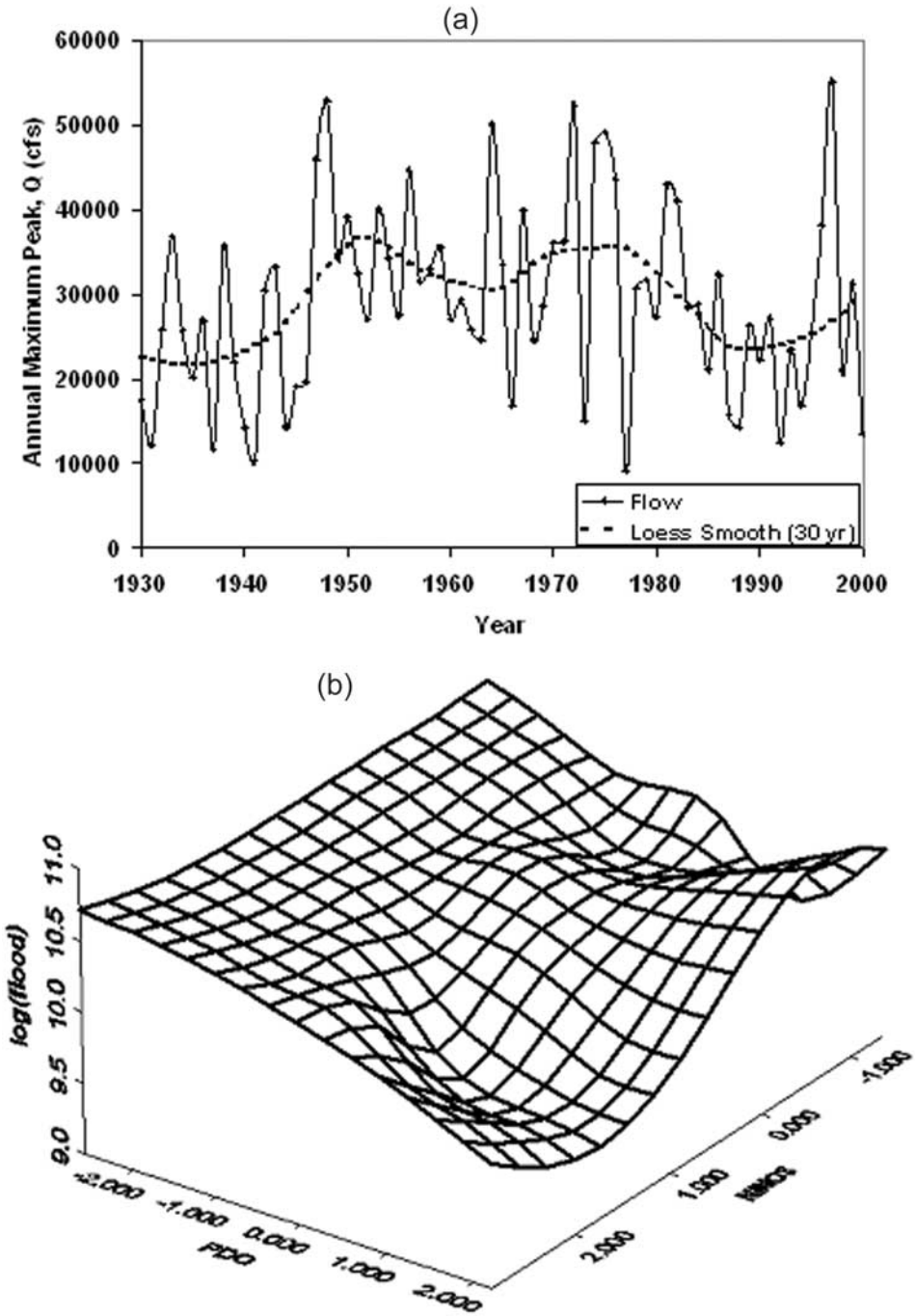

Figure 5. Observed annual maximum peak at Clark Fork River below Missoula, Montana, for the period 1930–2000. (a) Time series of observed flows with a 30-year Loess smooth to illustrate the temporal variation in the mean flood. (b) Loess smooth of log(flood) as a function of NINO3 and PDO illustrates the nonlinearity of the relationship.

ably reduced. Consequently we used global bandwidths in subsequent analyses. Another option would be to index the local bandwidths to the distance to k-nearest neighbors (as in the smoother Loess, or in the semiparametric approach illustrated by *De Souza and Lall* [2003], and then solve for a global bandwidth parameter (e.g., $\mathbf{H}_i = \mathbf{H}d_{ik}$, where $\mathbf{H}_i$ is a local bandwidth matrix, $\mathbf{H}$ is a global bandwidth matrix, and d_{ik} is the distance from the ith point of estimate in predictor space to its kth nearest neighbor). This extension was not pursued here.

5. Application

[26] An example application of the two methods was performed with data from the gage at Clark Fork River (CFR) below Missoula, MT (USGS Station No: 12353000),

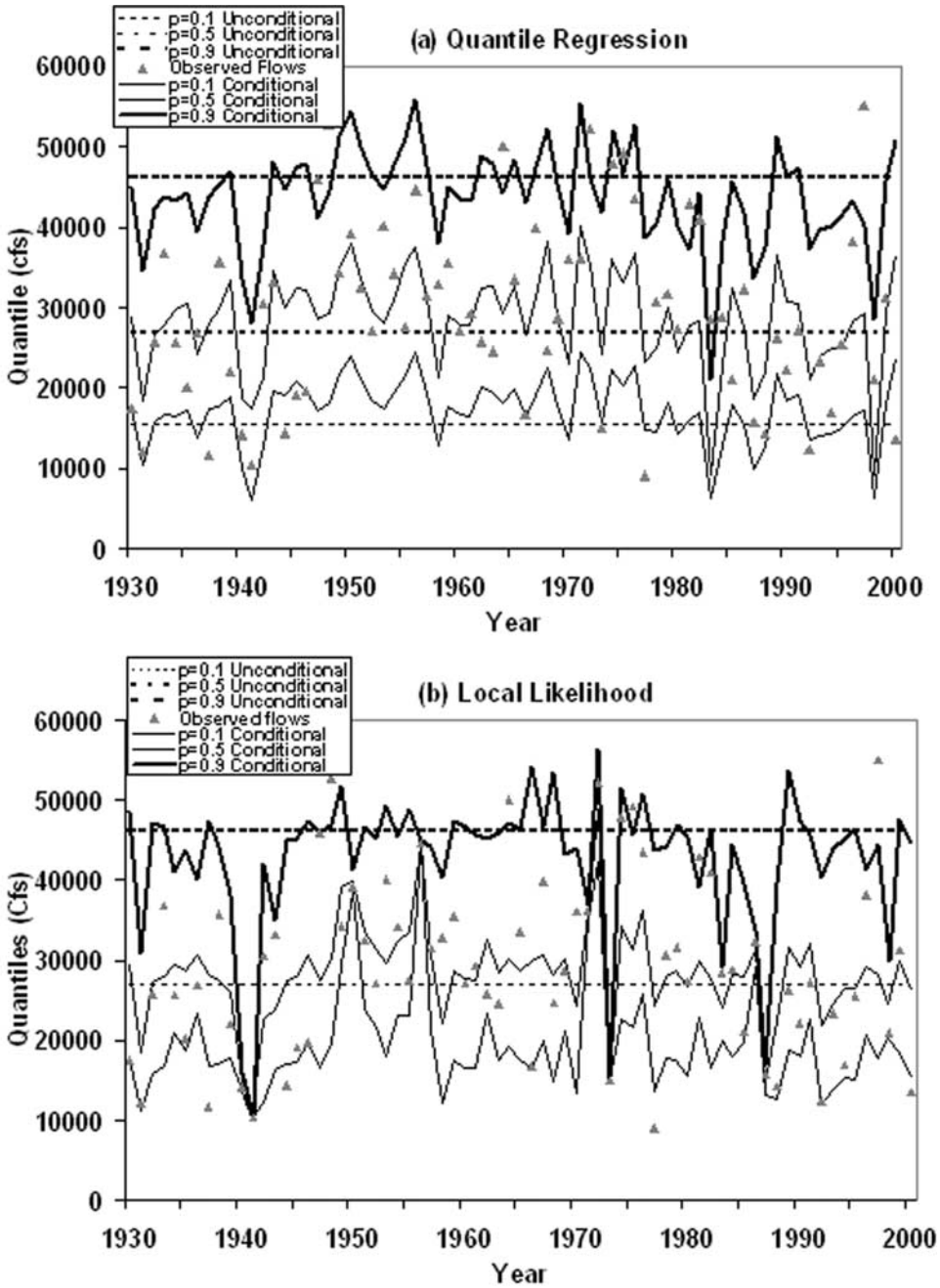

Figure 6. Cross-validated conditional flood quantile estimates for the Clark Fork River below Missoula. (a) Quantile regression. (b) Local likelihood smoothing.

located at 46°52′09″N, 114°07′33″W and an elevation of 3083 feet above mean sea level. The drainage area of the largely undisturbed mountain watershed with national forest, rangeland and recreation use is 23,336 Km^2. The quality of data in this basin is ''at least good'' according to USGS standards and the recorded flow at the gauging stations are minimally affected by upstream activities, diversions and human influence [*Slack et al.*, 1993]. Daily streamflow records are available from 1930 to 2000. The annual maximum flood was taken to be the target variable.

[27] As predictors, we consider ENSO and PDO. For ENSO, the sea surface temperature anomaly in the ''NINO3'' region in the eastern equatorial Pacific (5°N–5°S and 150°W–90°W) was used as the index. The NINO3 data set was obtained from the IRI data library (http://ingrid.ldeo.columbia.edu/SOURCES/.Indices/.nino/.EXTENDED/.NINO3/). The PDO index developed by *Mantua et al.* [1997] is the leading principal component of the gridded, monthly SST anomalies in the North Pacific Ocean, poleward of 20°N. The PDO data sets were acquired from the University of Washington (http://tao.atmos.washington.edu/pdo/). The winter (January–February–March–April) averages of the NINO3 and PDO indices were used as the predictors of the flood flows. The time series of these indices are provided in Figure 4, and their relationship with the flood series is explored in Figure 5b. The flood season at this location is predominantly April–May–June. Pearson's correlation coefficients between the flow Q and the winter

Table 1. Performance of Conditional Flood Quantile Estimation in Terms of Correlation Between the Cross-Validated Conditional Flood Quantiles and the Observed Annual Maximum Peak ($\hat{\rho}_{Q,Q_p}$) for the Clark Fork River below Missoula

	$\hat{\rho}_{Q,Q_p}$		
	p = 0.1	p = 0.5	p = 0.9
Quantile regression	0.40	0.42	0.33
Local likelihood	0.51	0.58	0.39

averages of NINO3 and PDO are −0.37 and −0.39 respectively for the 71 year record. The partial correlations cor(Q, NINO3|PDO) and cor(Q, PDO|NINO3) are −0.23 and −0.26 respectively. The null hypothesis of zero correlation is rejected at the 5% significance level for each of these estimates. Figure 5b shows that PDO mainly influences the anomalous conditions in the annual maximum peak, though anomalous conditions in both NINO3 and PDO result in extreme values of annual maximum peak at CFR basin. For instance, positive anomalous conditions in both Nino3 (El Nino) and PDO result in low flows, whereas negative anomalous conditions in Nino3 (La Nino) and PDO correspond to high values of annual maximum peak. On the other hand, normal conditions in PDO usually produce annual maximum peaks closer to climatology (median) irrespective of conditions in the tropical Pacific Ocean (Nino3), whereas flows vary quite substantially under normal conditions of Nino3. Thus anomalous conditions in PDO influence the anomalous conditions in the flows than the anomalous conditions in Nino3.

5.1. Conditional Flood Quantile Estimation for the Clark Fork River Below Missoula, MT

[28] Cross-validated conditional flood quantiles for p = 0.1, 0.5 and 0.9 estimated by quantile regression and by local likelihood applied to log transformed flows, and the corresponding unconditional quantiles assuming a log normal distribution for the flow data are presented in Figure 6. The correlation between the observed peaks and the conditional quantiles ($\hat{\rho}_{Q,Q_p}$) for four percentiles (p = 0.1, 0.5 and 0.9) is provided in Table 1. As expected, the correlation is highest (0.58 for local likelihood) with the estimated median flood, and given the apparent nonlinearity in the relationship illustrated in Figure 5b, local likelihood performs somewhat better than quantile regression.

[29] There are a number of years in which the cross-validated quantiles estimated by either method exhibit dramatic shifts from the unconditional values, and in several of these years, the "forecasts" correspond to anomalous floods of the right sign. For instance, in Figure 6b, years 1941 (NINO3 = 2.03 and PDO = 2.21) and 1987 (NINO3 = 1.25 and PDO = 1.91) correspond to positive anomalous conditions in both the tropical and extra tropical Pacific Ocean that result in reduced flows at the CFR basin and the predicted conditional flood quantiles also respond correspondingly with low values. Similarly, year 1972 (NINO3 = −0.19 and PDO = −1.77) relate to negative anomalous conditions in NINO3 and PDO that result in increased annual maximum peak, thereby higher values of predicted conditional flood quantiles.

5.2. Reconstruction of Flood Records

[30] To further illustrate the potential for forecasting flood risk, we considered a reconstruction of the conditional flood quantiles using NINO3 and PDO and the local likelihood method for 1900–1929, a period prior to the earliest year of record at the CFR site used. Annual maximum peak data from two nearby sites on the Clark Fork River (USGS Stations: 12354500; and 1238900) is available for part of the prior period for a pseudovalidation. Inflows recorded at station 1238900 are reported to be significantly affected by diversions from Clark Fork River below Missoula, MT from 1938 onwards. However, the correlation between the annual

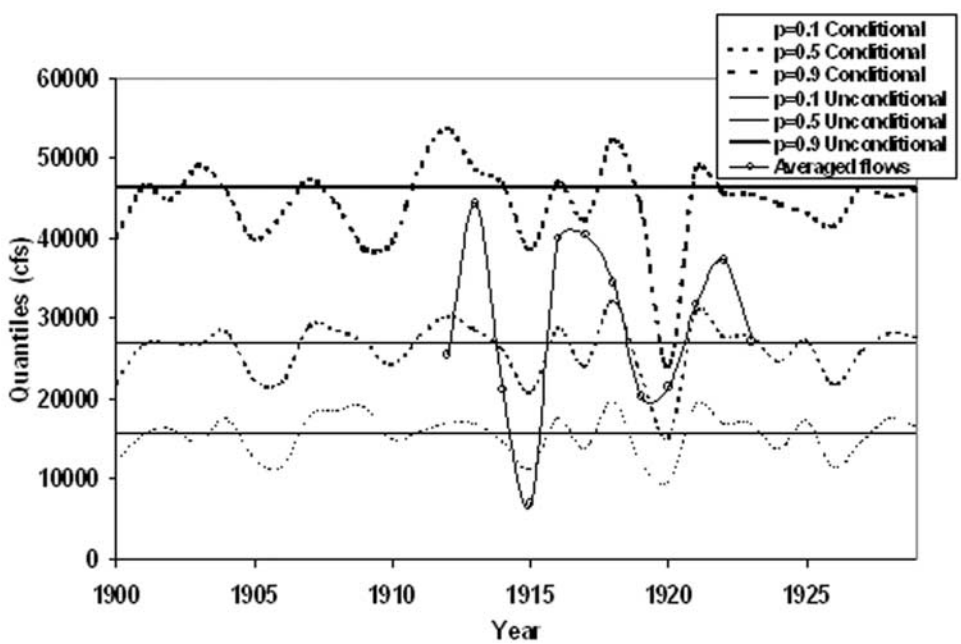

Figure 7. Reconstructed conditional flood quantiles for p = 0.1,0.5 and 0.9 using the local likelihood method with NINO3 and PDO for the period 1900–1929.

Table 2. Correlation of Reconstructed Flood Quantiles With the Observed Annual Maximum Peak at the Nearby Sites on the Clark Fork River

Station	Drainage Area, km^2	Longitude	Latitude	Period of Record Considered for Validation	p = 0.1	p = 0.5	p = 0.9
1238900	51,731	114°51′18″	47°25′47″	1912–1929	0.64	0.54	0.33
12354500	27,757	115°05′11″	47°18′07″	1911–1923	0.53	0.48	0.37

maximum peak at Clark Fork River below Missoula, MT (site considered for the study) and the annual maximum peak at Clark Fork River near Plains, MT (12389000) is 0.986 over the 1930–2000 period. Annual maximum peaks observed between 1900–1938 were not affected by the diversion from the Clark Fork River below Missoula, MT. Similarly, the correlation between the annual maximum peaks at Clark Fork River below Missoula, MT and the annual maximum peak at Clark Fork River at St. Regis, MT (12389000) observed during the period 1930–2000 is 0.911. The conditional flood quantiles reconstructed at the study site for the 1900–1929 period are shown in Figure 7, and their correlation with the two sites that have data for part of the period is provided in Table 2. These correlations are similar in strength to those noted during leave one out cross validation.

[31] Thus there is promise for forecasting flood risk, based on the season-ahead forecasts of climatic predictors (e.g., the leave one out cross validation experiment), as well as for reconstructing past variations in flood risk, contingent on the identification of appropriate climate prognostic variables. Here winter averages of two climate indices were used for both purposes. In practice, using the knowledge of the underlying moisture transport mechanisms that lead to floods at a site, one would explore appropriate predictors that may be observed values of variables such as Sea Surface Temperature, or forecasted precipitation from a numerical climate model. Likewise, for record extension the predictor may be a variable chosen in the flood season (i.e., concurrent to floods), while for the near term forecast, it could be a variable that is recorded in the season or two prior to the flood season.

6. Summary and Conclusions

[32] There is now growing evidence that particularly for frontally and snowmelt driven flood processes, such as in the western United States, an identification of indices of low-frequency climate variability is useful for understanding changes in local/regional flood frequency and for forecasting the flood risk in its season of occurrence. Two methods that allow such an estimation framework to be developed were compared here. The quantile regression approach has the advantage that it directly allows the computation of conditional quantiles without an explicit assumption as to the underlying density function of the conditional distribution. However, the need to assume and test a parametric form (potentially for each quantile to be estimated) for the regression poses logistical problems that translate into issues of the statistical identifiably and consistency of the resulting estimator. The second method tested, relies on local likelihood estimation. The conditional density function of the flood process is estimated locally at the point of estimate, and adapts to nonlinearity and heteroskedasticity of the relationship between floods and predictor variables. This is a semiparametric approach that is expected to be deficient as the number of predictors increases since the effective degrees of freedom will decrease rapidly. Bandwidth selection for this method is typically plagued by high variability, yet the resulting estimates are not terribly sensitive in terms of RMSE to bandwidths chosen over a reasonably wide range. Consequently, the application of this method with log transformed data with a modest number of predictors appears to lead to superior results over conditional quantile estimation, even in conditions where the quantile regression approach may be expected to have an advantage. For higher dimensional predictor space, a semiparametric treatment as by *De Souza and Lall* [2003] may be effective in this context. Bayesian approaches that can effectively characterize model and parameter uncertainty in this context need to be pursued.

[33] The application with the Montana flood series demonstrates that such methods with properly chosen additional predictors may offer prospects for reconstructing past flood series, as well as for short term forecasting. Such reconstructions may in turn allow for better identification of regional flood frequency curves and of regime like variations in local and regional flood risk. Work in these directions is currently under progress.

[34] As these methods become accessible and tested, we can expect that flood hazard insurance and mitigation programs can actively use the forecasts of flood risk, for reservoir operation, relief effort planning, premium setting and the like. We can also expect that the prior period reconstructions can be used to better understand the nature of temporal variations in flood risk and thus used to guide investments in long term flood risk reduction. These are evolving areas in an exciting area of study.

References

Barlow, M., S. Nigam, and E. H. Berbery, ENSO, Pacific decadal variability, and U.S. summertime precipitation, drought, and stream flow, *J. Clim.*, *14*, 2105–2128, 2001.

Bhattacharya, P. K., and A. K. Gangopadhyay, Kernel and nearest-neighbor estimation of a conditional quantile, *Ann. Stat.*, *18*(3), 1400–1415, 1990.

Buchinsky, M., Estimating the asymptotic covariance-matrix for quantile regression-models—A Monte-Carlo study, *J. Econometrics*, *68*(2), 303–338, 1995.

Cayan, D. R., Interannual climate variability and snowpack in the western United States, *J. Clim.*, *9*, 928–948, 1996.

Cayan, D. R., K. T. Redmond, and L. G. Riddle, ENSO and hydrologic extremes in the western United States, *J. Clim.*, *12*, 2881–2893, 1999.

Davison, A. C., and N. I. Ramesh, Local likelihood smoothing of sample extremes, *J. R. Stat. Soc., Ser. B.*, *62*, 191–208, 2000.

De Souza, F., and U. Lall, Seasonal to interannual ensemble streamflow forecasts for Ceara, Brazil: Applications of a multivariate, semiparametric algorithm, *Water Resour. Res.*, doi:10.1029/2002WR001373, in press, 2003.

Duan, Q. Y., S. Sorooshian, and V. Gupta, Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*(4), 1015–1031, 1992.

Franks, S. W., and G. Kuczera, Flood frequency analysis: Evidence and implications of secular climate variability, New South Wales, *Water Resour. Res.*, *38*(5), 1062, 2002.

Gershunov, A., and T. P. Barnett, Interdecadal modulation of ENSO teleconnections, *Bull. Am. Meteorol. Soc.*, *79*(12), 2715–2725, 1998.

Gershunov, A., J. Michaelsen, and C. Gautier, Large-scale coupling between the tropical greenhouse effect and latent heat flux via atmospheric dynamics, *J. Geophys. Res.*, *103*(D6), 6017–6031, 1998.

Halpert, M. S., and C. F. Ropelewski, Surface-temperature patterns associated with the Southern Oscillation, *J. Clim.*, *5*, 577–593, 1992.

Haston, L., and J. Michaelsen, Long-term central coastal California precipitation variability and relationships to El Nino Southern Oscillation, *J. Clim.*, *7*, 1373–1387, 1994.

Hirschboeck, K. K., Catastrophic flooding and atmospheric circulation anomalies, in *Catastrophic Flooding*, edited by L. Mayer and D. Nash, pp. 23–56, Allen and Unwin, Concord, Mass., 1987a.

Hirschboeck, K. K., Hydroclimatically defined mixed distributions in partial duration flood series, in *Hydrologic Frequency Modeling*, edited by V. P. Singh, pp. 192–205, D. Reidel, Norwell, Mass., 1987b.

Hirschboeck, K. K., Flood hydroclimatology, in *Flood Geomorphology*, John Wiley, New York, 1988.

Hirschboeck, K. K., Climate and floods, in *National Water Summary 1988–89–Hydrologic Events and Floods and Droughts*, compiled by R. W. Paulson et al., *U.S. Geol. Surv. Water Supply Pap.*, *2375*, 99–104, 1991.

Holmes, C. C., and N. M. Adams, A probabilistic nearest neighbor method for statistical pattern recognition, *J. R. Stat. Soc., Ser. B*, *64*, 295–306, 2002.

Jain, S., and U. Lall, Magnitude and timing of annual maximum floods: Trends and large-scale climatic associations for the Blacksmith Fork River, Utah, *Water Resour. Res.*, *36*(12), 3641–3651, 2000.

Jain, S., and U. Lall, Floods in a changing climate: Does the past represent the future?, *Water Resour. Res.*, *37*(12), 3193–3205, 2001.

Katz, R. W., and B. G. Brown, Extreme events in a changing climate—Variability is more important than averages, *Clim. Change*, *21*(3), 289–302, 1992.

Knox, J., Large increase in flood magnitude in response to modest changes in climate, *Nature*, *361*, 430–432, 1993.

Koenkar, R., and G. S. Bassett, Regression quantiles, *Econometrica*, *46*, 33–50, 1978.

Koenkar, R., and V. D'Orey, Computing regression quantiles, *Appl. Stat.*, *36*(3), 383–393, 1987.

Koenkar, R., and B. J. Park, An interior point algorithm for nonlinear quantile regression, *J. Am. Stat. Assoc.*, *94*, 1296–1309, 1996.

Koenker, R., S. Portnoy, and P. Ng, Nonparametric estimation of conditional quantile functions, in *Statistical Analysis and Related Methods*, edited by Y. Dodge, pp. 217–229, Elsevier Sci., New York, 1992.

Loader, C., *Local Regression and Likelihood*, Springer-Verlag, New York, 1999.

Mantua, N. J., S. R. Hare, Y. Zhang, J. M. Wallace, and R. C. Francis, A Pacific Interdecadal Climate Oscillation With Impacts on Salmon Production, *Bull. Am. Meteorol. Soc.*, *78*, 1069–1079, 1997.

McCabe, G. J., and M. D. Dettinger, Decadal variations in the strength of ENSO teleconnections with precipitation in the western United States, *Int. J. Climatol.*, *19*(13), 1399–1410, 1999.

Milly, P. C. D., R. T. Wetherald, K. A. Dunne, and T. L. Delworth, Increasing risk of great floods in a changing climate, *Nature*, *415*(6871), 514–517, 2002.

National Research Council (NRC), *Flood Risk Management and the American River Basin: An Evaluation*, 235 pp., Natl. Acad. Press, Washington, D. C., 1995.

National Research Council (NRC), *Improving American River Flood Frequency Analyses*, 120 pp., Natl. Acad. Press, Washington, D. C., 1999.

Pagan, A., and A. Ullah, *Nonparametric Econometrics*, Cambridge Univ. Press, New York, 1999.

Piechota, T. C., and J. A. Dracup, Drought and regional hydrologic variation in the United States: Associations with the El Nino Southern Oscillation, *Water Resour. Res.*, *32*(5), 1359–1373, 1996.

Pizaro, G., and U. Lall, El Niño and Floods in the US West: What can we expect?, *Eos Trans. AGU*, *83*(32), 349–352, 2002.

Porparto, A., and L. Ridolfi, Influence of weak trends on exceedance probability, *Stochastic Hydrol. Hydraul.*, *12*(1), 1–15, 1998.

Rasmusson, E. M., and T. H. Carpenter, Variations in tropical sea-surface temperature and surface wind fields associated with the Southern Oscillation El-Nino, *Mon. Weather Rev.*, *110*(5), 354–384, 1982.

Ropeleweski, C. F., and M. S. Halpert, Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation, *Mon. Weather Rev.*, *115*, 1606–1626, 1987.

Slack, J. R., A. M. Lumb, and J. M. Landwehr, Hydroclimatic Data Network (HCDN): A U.S. Geological Survey streamflow data set for the United States for the study of climate variation, 1874–1988, *U.S. Geol. Surv. Water Resour. Invest. Rep., 93-4076*, 1993.

Trenberth, K. E., and C. J. Guillemot, Physical processes involved in the 1988 drought and 1993 floods in North America, *J. Clim.*, *9*, 1288–1298, 1996.

Wendland, W. M., and R. A. Bryson, Northern Hemisphere airstream regions, *Mon. Weather Rev.*, *109*(2), 255–270, 1981.

Yu, K., Smoothing regression quantiles by combining k-NN estimation with local linear kernel fitting, *Stat. Sinica*, *9*, 759–774, 1999.

Yu, K. M., and M. C. Jones, Local linear quantile regression, *J. Am. Stat. Assoc.*, *93*(441), 228–237, 1998.

Yu, K., and R. A. Moyeed, Bayesian quantile regression, *Stat. Prob. Lett.*, *54*, 437–447, 2001.

U. Lall and A. Sankarasubramanian, International Research Institute for Climate Prediction, Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA. (ula2@columbia.edu; sankar@iri.columbia.edu)