Regional estimation of flood quantiles: Parametric versus nonparametric regression models

Marco Latraverse
IREQ, Hydro-Quebec's Research Institute, Varennes, Quebec, Canada

Peter F. Rasmussen
Department of Civil Engineering, University of Manitoba, Winnipeg, Manitoba, Canada

Bernard Bobée
INRS-Eau, University of Quebec, Sainte-Foy, Quebec, Canada

Received 21 May 2001; revised 12 November 2001; accepted 12 November 2001; published 2 July 2002.

[1] A recent trend in regional frequency analysis is to consider floating regions where only basins that are sufficiently similar to the design site are considered for information transfer. Similarity is measured in some suitable metric of catchment characteristics. This paper discusses the analogy between this idea and nonparametric regression. Some of the techniques developed recently in the area of nonparametric regression are employed to develop improved regional flood estimators. The additive model used here to a large extent overcomes the curse of dimensionality often associated with nonparametric regression on multivariate predictor space. The application of the proposed methodology to selected areas of the United States suggests that there can be substantial gains over the traditional log linear models currently employed. *INDEX TERMS:* 1821 Hydrology: Floods; 1860 Hydrology: Runoff and streamflow; 1869 Hydrology: Stochastic processes; *KEYWORDS:* regionalization, floods, nonlinear, nonparametric, regression, frequency analysis

1. Introduction

[2] Regional flood frequency analysis is commonly used to obtain estimates of design flows at ungauged sites. Statistical methods for regional frequency analysis have been employed for several decades and have evolved from rudimentary methods involving simple area ratio scaling to advanced methods that attempt to capture the complex relationship between catchment characteristics and the distribution of peak runoff. Two methods have been particularly favored by practitioners, the index flood method and the quantile regression method. *Dalrymple*'s [1960] index flood method is based on the assumption that within hydrological regions the distribution of annual floods is the same at all sites except for a scaling factor related primarily to the size of the basin. This hypothesis is equivalent to assuming that moment ratios such as the coefficients of variation and skewness are constant throughout the region. This has often been termed regional homogeneity. Hence, when delineating hydrologic regions for use in the index flood method, a primary concern is to ensure that sites within each region have approximately the same coefficients of variation and skewness of annual floods, with differences no larger than what can be attributed to sampling variability. This may be difficult to achieve within geographical regions because moment ratios generally depend on basin area. For that reason, several countries have adopted quantile regression procedures as standard for national agencies. This is the case for example in the United States where regression relationships relating flood quantiles to catchment characteristics have been developed for each state and each hydrologic region [*Jennings et al.*, 1994]. Regional quantile regression involves less restrictive assumptions regarding the homogeneity of regions than the index flood method. Regions have typically be determined by fitting a regression function to a large data set and then identifying subregions where regression residuals have similar sign and magnitude. It is generally required that regions must be homogeneous in terms of catchment characteristics that are not included as explanatory variables in the regression equations. For example, climatic factors such as mean annual precipitation and/or mean snow accumulation are expected to have an influence on peak runoff. Hence one should either include these variables in the regression model or use regions where they can be considered constant.

[3] A recent trend has been to consider regions that are not geographically contiguous. *Tasker* [1982] identified homogeneous regions based on cluster analysis of watershed characteristics and used discriminant analysis to determine the probability of an ungauged site belonging to a particular cluster of stations. This concept of region appears useful and overcomes some of the problems associated with geographical regions. However, fixed regions, whether geographical or noncontiguous, suffer from the lack of continuity over region boundaries.

[4] *Acreman and Wiltshire* [1987, 1989] proposed what has become known as the "region-of-influence" approach, a term owed to *Burn* [1990a, 1990b]. This approach dispenses with the traditional concept of regions by associating an individual set of stations with each ungauged site. The selection of stations for inclusion in the region of influence of an ungauged site can be based on the Euclidean

0043-1397/02/2001WR000677$09.00

distance in some suitable space of watershed characteristics. *Burn* [1990a, 1990b] and *Zrinji and Burn* [1994] employed the region of influence in conjunction with the index flood method. In a recent study, *Tasker et al.* [1996] compared five regression methods for regional estimation of 50-year floods in Arkansas. The methods differed in the way regions were determined. Three methods involved traditional geographical regions with different subdivisions (state divided into one, two, and four regions), one method was based on the cluster/discriminant analysis technique developed by *Tasker* [1982], and one method involved a new approach based on the region-of-influence concept. To implement the last method, a distance metric based on catchment characteristics was defined. *Tasker et al.* [1996] specifically defined the distance d_{ij} between sites i and j as

$$d_{ij} = \left[\sum_{k=1}^{K} \left(\frac{C_{k,i} - C_{k,j}}{\mathrm{std}(\mathbf{C}_k)} \right)^2 \right]^{1/2}, \quad (1)$$

where $C_{k,i}$ denotes the value of the kth watershed characteristic for site i, K is the number of attributes considered, $\mathbf{C}_k$ is the vector of values of the kth characteristic for all sites, and std($\mathbf{C}_k$) is the standard deviation of the elements of $\mathbf{C}_k$. To estimate a regression model for a particular site, *Tasker et al.* [1996] considered the 34 stations closest to the site according to the above distance metric. A conventional log linear model was then estimated. In the split sample experiment conducted by *Tasker et al.* [1996], the region-of-influence (ROI) approach performed better than the other models. Tasker et al. argued that the ROI is intuitively appealing because the estimation involves only stations whose characteristics are close to that of the ungauged site and that extrapolation errors therefore are largely avoided. Moreover, problems of nonlinearity are likely to be reduced with the ROI approach.

[5] It is argued here that there is a clear analogy between the region-of-influence approach and nonparametric regression. It would seem reasonable then to take advantage of some of the powerful modeling tools developed in recent years in the area of nonparametric regression to improve regional flood estimation procedures. In particular, current applications of the ROI approach employ some often arbitrary choices of parameters. The nonparametric framework eliminates the need for arbitrary choices by optimizing model parameters in a formal way.

[6] The paper is organized as follows. In section 2, we review different aspects of nonparametric regression for a single explanatory variable and discuss the link between the region-of-influence approach and nonparametric regression. In section 3, we consider the case of multidimensional explanatory variables and describe a procedure to deal with the curse of dimensionality often attributed to multivariate nonparametric regression. In section 4, the proposed approach is illustrated with an application to two regions in the US. Finally, section 5 provides a discussion of the potential of the proposed method.

2. Nonparametric Regression With One Independent Variable

[7] Nonparametric regression is a way to circumvent some of the problems that occasionally arise in conventional linear regression. In particular, nonparametric regression does not impose a particular functional form on the relationship between the dependent and independent variables. It can handle fairly easily possible heteroscedasticity of residuals and can be fine tuned in an objective way to particular situations. In cases where the basic hypotheses of linear regression are verified, nonparametric regression often performs almost as well as its parametric counterpart.

[8] There exists a number of nonparametric regression procedures, some of which have been used in hydrology. Several textbooks provide overview of nonparametric regression procedures, for example, *Härdle* [1990], *Wand and Jones* [1995], *Green and Silverman* [1994], and *Loader* [1999]. Here we adopt the so-called local polynomial regression approach which is a generalization of *Tasker et al.*'s [1996] region-of-influence regression procedure. The method is based on the well-known kernel method. This section presents local polynomial regression for the case where there is only one independent variable, although it can be generalized to the multivariate case. As described in section 3, our approach to handling the case of multiple explanatory variables is based on the additive model which involves fitting a number of univariate regression relationships.

2.1. Local Polynomial Regression

[9] Consider the observation of a set of data (x_i, y_i), $i = 1, \ldots, n$. It is assumed that the data have been organized so that the x variable occurs in ascending order, i.e., $x_1 \leq x_2 \leq \ldots \leq x_n$ and that the observations are independent of each other. As in conventional linear regression, the objective of nonparametric regression is to determine the mean value function

$$f(x) = E[Y \mid X = x]. \quad (2)$$

The mean value function provides the best (in the sense of mean square error) estimate of y associated with the point x. Linear regression is based on the hypothesis that f is a linear function of x. In contrast, nonparametric regression requires no prior assumption regarding the form of $f(x)$.

[10] All nonparametric regression methods share the common feature that estimation of f at a point x_0 is based on observations whose x values are in the vicinity of x_0. In the kernel method, a weight function (the kernel) is used to assign weights to the observations so that the closer they are to x_0, the more weight they receive. The kernels K considered here are continuous, positive, symmetric functions that integrate to 1; that is

$$\int K(u)du = 1. \quad (3)$$

Examples of some frequently used kernels are given in Table 1. The kernel defines the sequence of weights $\omega_i(x_0)$, $i = 1,\ldots,n$ that will be associated with each of the n observations for estimation of the mean value function at x_0. More specifically, the weights are defined as

$$\omega_i(x_0) = \frac{K_h(x_i - x_0)}{\sum_{j=1}^{n} K_h(x_j - x_0)}, \quad (4)$$

where $K_h(u) = h^{-1}K(u/h)$. The denominator in the above expression ensures that the sequence of weights sum to 1.

Table 1. Common Kernel Functions

Name	$K(u)$	Support
Rectangular	1	$\|u\| < 1$
Triangular	$1 - \|u\|$	$\|u\| < 1$
Epanechnikov	$\frac{3}{4}(1-u^2)$	$\|u\| < 1$
Biweight	$\frac{15}{16}(1-u^2)^2$	$\|u\| < 1$
Tricube	$\frac{7}{20}(1-\|u\|^3)^3$	$\|u\| < 1$
Triweight	$\frac{35}{32}(1-u^2)^3$	$\|u\| < 1$
Normal	$\frac{1}{\sqrt{2\pi}}\exp(-u^2/2)$	$u \in R$

Note that for most of the kernels given in Table 1, the weight assigned to observation i will be zero if $|x_i - x_0| \geq h$. The estimation of $f(x_0)$ is given by

$$\hat{f}(x_0) = \sum_{i=1}^{n} \omega_i(x_0) y_i. \tag{5}$$

Two factors determine how smooth the mean value function will be when evaluated over a range of x values: the type of kernel K and the smoothing parameter h. It is generally recognized that the choice of kernel is relatively unimportant compared to the choice of smoothing parameter.

[11] As described above, the kernel method involves estimation of a zero-degree polynomial, i.e., a constant, based on the data in the vicinity of x_0. A natural extension is to consider the fitting of a p-degree polynomial

$$\beta_0 + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + \cdots + \beta_p(x - x_0)^p \tag{6}$$

to the data close to x_0. The parameters $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ may be determined by minimizing the weighted sum of squared deviations

$$\sum_{i=1}^{n} \omega_i(x_0)\left[y_i - \beta_0 - \beta_1(x - x_0) - \beta_2(x - x_0)^2 - \cdots - \beta_p(x - x_0)^p\right]^2, \tag{7}$$

where the weights are obtained from the kernel. Depending on the type of kernel, some weights may be zero, implying that some data are not considered in the fitting of the polynomial. Defining $\mathbf{Y} = (y_1, y_2, \ldots, y_n)^T$, $\mathbf{W}_{x_0} = \mathrm{diag}[\omega_1(x_0), \omega_2(x_0), \ldots, \omega_n(x_0)]$, and

$$\mathbf{X}_{x_0} = \begin{pmatrix} 1 & x_1 - x_0 & \cdots & (x_1 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_0 & \cdots & (x_n - x_0)^p \end{pmatrix},$$

the classical weighted least squares estimator of β can be expressed as

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}_{x_0}^T \mathbf{W}_{x_0} \mathbf{X}_{x_0}\right)^{-1} \mathbf{X}_{x_0}^T \mathbf{W}_{x_0} \mathbf{Y}. \tag{8}$$

Because the polynomial is centered on x_0, the intercept $\hat{\beta}_0$ constitutes the estimate of f at x_0:

$$\hat{f}(x_0) = \hat{\beta}_0 = \mathbf{e}_1^T \hat{\boldsymbol{\beta}}, \tag{9}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)^T$ is a $(p + 1)$ unit vector. It should be emphasized that the assumption of a p-degree polynomial in the neighborhood of x_0 in no way implies that the regression function $f(x)$ will have a polynomial form. The adoption of a p-degree polynomial in the vicinity of x_0 serves only to obtain a robust estimate of $f(x_0)$. As one proceeds to another point x_0', the regression parameters will change because the kernel weights used in the estimation change and the function $f(x)$ will take a form consistent with the data. The fitting of a p-degree polynomial rather than a constant has several advantages, notably increased robustness near the end points of the data. If the data exhibit a nonzero slope near an end point, the simple kernel regression will invariably be biased in this region because data will be clustered on one side of the estimation point. By fitting a p-degree polynomial this bias is largely avoided.

[12] The degree of the polynomial should be chosen reasonably low in order to avoid overfitting and ensure a robust estimate of β. Experience has shown that polynomials of order one to three typically will do a satisfactory job.

[13] The local polynomial regression is robust to heteroscedasticity. Because only data in the vicinity of x_0 are used to fit the polynomial, any systematic dependence between the residual variance and the independent variable can for most practical purposes be ignored. In many cases the y observations represent estimated quantities, for example, flood quantiles. If the estimation variance is known, this information can be readily introduced in the nonparametric regression. If $\mathbf{V} = \mathrm{diag}(1/\sigma_1^2, 1/\sigma_2^2, \ldots, 1/\sigma_n^2)$ represent the diagonal matrix with the inverse estimation variance of the ny values on the diagonal, then the only modification to the above described procedure is to replace the weight matrix $\mathbf{W}_{x_0}$ with the product $\mathbf{W}_{x_0}\mathbf{V}$. *Cleveland* [1979] proposed a method, known as the LOWESS method, that is robust to outliers.

2.2. Relationship Between Local Polynomial Regression and the Region-of-Influence Method

[14] From the above discussion the analogy between nonparametric regression and *Tasker et al.*'s [1996] region-of-influence approach should be clear. Their approach with one explanatory variable is equivalent to fitting a first-degree polynomial (a straight line) to the data surrounding x_0 using a rectangular kernel, i.e., equal weight is given to the data included in the neighborhood. Tasker et al.'s procedure involves a variable-size neighborhood because in their specific application, 34 stations were systematically used at all estimation points. This corresponds to the so-called k nearest neighbors method which is an alternative to the fixed kernel approach.

[15] Treating regional quantile regression more formally as a problem of nonparametric regression offers potential advantages over the region-of-influence method. *Tasker et al.*'s [1996] choice of a variable-sized neighborhood of 34 stations was made somewhat arbitrarily. Although good results were obtained, one would expect that through careful application of nonparametric estimation procedures, even better results could be achieved. The tapering of weights and the use of higher-order polynomials would be expected to yield more robust regression estimates and to reduce or eliminate the abrupt changes in the regression function associated with a uniform kernel.

2.3. Linear Nonparametric Smoothers

[16] An issue of practical concern in nonparametric regression is the degree of freedom of the model. Models

with a large degree of freedom provide a good description of the data but may be unsuitable for prediction. In order to assess the degree of freedom of a nonparametric regression the concept of a linear smoother can be employed. A nonparametric regression function is called a linear smoother if it is a linear function of the observations y_i, $i = 1, \ldots, n$, that is, if the regression function can be written as

$$\hat{f}(x_0) = \sum_{i=1}^{n} l_i(x_0) y_i = \mathbf{L}_{x_0}^T \mathbf{Y}, \tag{10}$$

where $\mathbf{L}_{x_0} = [l_1(x_0), l_2(x_0), \ldots, l_n(x_0)]$ is a weight sequence that does not depend on the y observations. From equations (8) and (9) it can be seen that the local polynomial regression is a linear smoother with

$$\mathbf{L}_{x_0}^T = \mathbf{e}_1^T \left(\mathbf{X}_{x_0}^T \mathbf{W}_{x_0} \mathbf{X}_{x_0} \right)^{-1} \mathbf{X}_{x_0}^T \mathbf{W}_{x_0} \tag{11}$$

The vector $\hat{\mathbf{f}}$ of fitted values at the observations $x_1, x_2, \ldots, x_n$ is given by

$$\hat{\mathbf{f}} = \begin{bmatrix} \hat{f}(x_1) \\ \hat{f}(x_2) \\ \vdots \\ \hat{f}(x_n) \end{bmatrix} = \begin{bmatrix} \mathbf{L}_{x_1}^T \\ \mathbf{L}_{x_2}^T \\ \vdots \\ \mathbf{L}_{x_n}^T \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \mathbf{S}_\Omega \mathbf{Y}, \tag{12}$$

where $\mathbf{S}_\Omega = [\mathbf{L}_{x_1}, \mathbf{L}_{x_2}, \ldots, \mathbf{L}_{x_n}]^T$ is called the smoother matrix. In general, this matrix depends on a set of smoothing parameters Ω.

[17] The smoother matrix is useful for determining the effective degree of freedom of a nonparametric regression function and the associated noise. To put things in perspective, consider the case of a classical linear regression of a variable y on $p - 1$ explanatory variables, with coefficients estimated by ordinary least squares. The smoother matrix is given by $\mathbf{S}_\Omega = \mathbf{X} (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$, where $\mathbf{X}$ is the design matrix with dimension $n \times p$ (including intercept). Since there are p parameters, $\mathbf{S}_\Omega$ projects the y observations onto a p-dimensional space. Hence the rank of $\mathbf{S}_\Omega$ or, equivalently, the trace of $\mathbf{S}_\Omega$ equals p, the model degree of freedom. Correspondingly, the degree of freedom for estimating the noise is $\mathrm{tr}(\mathbf{I} - \mathbf{S}_\Omega) = n - p$. For the nonparametric regression function, we may define the model degree of freedom in a similar way. *Buja et al.* [1989] suggested the following definition of the model degree of freedom for a nonparametric regression function:

$$\nu = \mathrm{Tr}(\mathbf{S}_\Omega). \tag{13}$$

The degree of freedom is useful for comparing different models and will be discussed further in section 4.

2.4. Selection of Smoothing Parameters

[18] Three factors influence the degree of smoothing of a nonparametric function estimated by local polynomial regression: the kernel type (K), the smoothing parameter h that determines the spread of the kernel, and the degree p of the local polynomial. The degree of smoothing is a tradeoff between bias and variance. A highly fluctuating function represents a case of low bias and large variance, whereas a smooth function represents low variance and possibly high bias. The mean square error of residuals is a frequently used measure of precision that combines bias and variance. The vector of smoothing parameters $\Omega = (h, p, K)$ may be determined by minimizing the following quantity, called the generalized cross validation (GVC) mean square error:

$$\mathrm{GCV}(\Omega) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i - \hat{f}_\Omega(x_i)}{1 - \mathrm{tr}(\mathbf{S})/\mathrm{n}} \right]^2. \tag{14}$$

The rationale for this objective function is outlined in Appendix A.

[19] As mentioned previously, the choice of kernel type has relatively little influence on the degree of smoothing. For that reason, it seems acceptable to make a prior choice of an appropriate kernel. Furthermore, the degree of the local polynomial can for most purposes also be preselected, and estimation effort can be concentrated on the smoothing parameter. The rectangular kernel should be avoided because it tends to yield a rugged regression function, but any other kernel given in Table 1 would be acceptable.

[20] The parameter h has a strong impact on the degree of smoothing. For kernels with finite support, only observations with $|x_i - x_0| < h$ are included in the estimation of $f(x_0)$. As h increases, more and more observations will be included in the estimation of $f(x_0)$ and observations will tend to get increasingly similar weights. As h approaches infinity, all observations will be included and with equal weight. The regression function is not defined at x_0 if the denominator in equation (4) is zero. Hence, for the regression function to be defined over the entire range of observations $[x_1; x_n]$, we must have $h > \max(x_{i+1} - x_i)$ when a kernel with finite support is used.

3. Nonparametric Regression With Several Independent Variables

[21] In regional flood frequency analyses one will typically be interested in including several explanatory variables in the regression model. It is not uncommon to find three to five catchment characteristics that are statistically significant in linear regression models of quantiles. In the context of nonparametric regression, we are interested in finding a function f such that

$$f(\mathbf{x}) = E[Y \mid x_1, x_2, \ldots, x_d] = f(x_1, x_2, \ldots, x_d), \tag{15}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ is a vector of explanatory variables and d is the dimension of the predictor space. (In the following, whenever we refer to x_i, we mean the ith predictor, not the ith observation.)

[22] Local polynomial regression with kernel weights has a natural extension to the multivariate case [*Härdle*, 1990]. The kernels are then multidimensional and so are the local regression surface. However, kernel regression in high-dimensional predictor space suffers from the curse of dimensionality, resulting in slow rate of convergence. To handle the multivariate case, it is, in practice, necessary to impose some constraints on the regression surface. A logical simplification is to assume that the contribution from each

of the independent variables to the regression function is additive [*Hastie and Tibshirani*, 1990]:

$$f(\mathbf{x}) = E[Y \mid x_1, x_2, \ldots, x_d] = f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d), \quad (16)$$

where f_i represent a univariate nonparametric smoother associated with the dependent variable x_i. This model does not suffer from the problem of dimensionality and has the additional advantage of allowing visualization of the relationship between the regression function and each independent variable, something that is not possible with the multivariate kernel method. This can be quite useful, for example, to assess possible linearity between the response variable and certain independent variables. A detailed study of the applicability of nonparametric additive regression models to regional flood estimation was conducted by *Latraverse* [2000].

3.1. Estimation of Additive Models

[23] The estimation of the additive model in equation (16) involves the determination of d univariate functions f_i, $i = 1, \ldots, d$. The problem is not just to find optimal smoothing parameters for each f_i but also to determine how much each f_i should contribute to f. The so-called back fitting algorithm is an efficient technique to determine the contribution of each f_i for a fixed set of smoother matrices.

[24] Assume that a set of observations $(x_{1,j}, x_{2,j}, \ldots, x_{d,j}, y_j), j = 1, \ldots, n$ is available. The relationship between the dependent and independent variables may be expressed as

$$y_j = \mu_y + f_1(x_{1,j}) + f_2(x_{2,j}) + \cdots + f_d(x_{d,j}) + \epsilon_j. \quad (17)$$

The unconditional expectation of Y, μ_y, is included explicitly in the model, so that we can assume $E[f_i] = 0$. In this way, floating constants are eliminated from the f_i values. Residuals are assumed to have zero mean and variance σ_f^2. In matrix form the fitted values at the observation points are given by

$$\hat{\mathbf{f}} = \bar{y} + \hat{\mathbf{f}}_1 + \hat{\mathbf{f}}_2 + \cdots + \hat{\mathbf{f}}_d. \quad (18)$$

We will assume that each f_i is a linear smoother with parameter Ω_i, i.e., $\hat{\mathbf{f}}_i$ is of the form given in equation (12).

[25] The back fitting algorithm involves the following steps.

1. Initialize $\mathbf{f}_i$, $i = 1, \ldots, d$. For example, set them equal to linear regression estimates.
2. For $i = 1, \ldots, d$, determine the vector

$$\mathbf{f}_i = \mathbf{S}_i\left(\mathbf{y} - \bar{y} - \sum_{\substack{k=1 \\ k \neq i}}^{d} \mathbf{f}_k\right),$$

where $\mathbf{S}_i(\mathbf{u})$ is the smooth of $\mathbf{u}$ on variable x_i. In the summation, use the most recent value of $\mathbf{f}_k$.

3. Repeat step 2 until there is no more change in the $\mathbf{f}_i$ values over subsequent cycles.

[26] The motivation for the back fitting procedure is easily appreciated. Here $(\mathbf{y} - \bar{y} - \Sigma_{k \neq i}\, \mathbf{f}_k)$ represents the residuals when x_i is not included in the model. In step 2 we try to explain as much as possible of the variation of these residuals using x_i as explanatory variable. This is done in turn for each independent variable. The whole sequence is repeated until the $\mathbf{f}_i$ values become stable. *Buja et al.* [1989] examined conditions for convergence of the back fitting algorithm and concluded that in most cases of practical interest the algorithm will converge rapidly.

[27] Because the additive regression function f is a sum of linear smoothers, it is itself a linear smoother. *Hastie and Tibshirani* [1990] suggest using the following expression to calculate the approximate degree of freedom of the model:

$$\nu \simeq 1 + \sum_{i=1}^{d} [\mathrm{tr}(\mathbf{S}_i) - 1], \quad (19)$$

which is an extension of equation (13).

[28] Fitting an additive model involves finding the set of parameters $\mathbf{\Omega} = (\Omega_1, \Omega_2, \ldots, \Omega_d)$. This can be accomplished using a performance measure such as the generalized cross validation mean square error. A reasonable approximation to the GCV for the additive model is

$$\mathrm{GCV}(\Omega) = \frac{1}{n}\sum_{j=1}^{n} \frac{\left[y_j - \bar{y} - \sum_{i=1}^{d} \hat{f}_{i,\Omega_i}(x_{i,j})\right]^2}{\left[1 - \frac{1}{n}\left(1 + \sum_{i=1}^{d} \{\mathrm{tr}(\mathbf{S}_{\Omega_i}) - 1\}\right)\right]^2}. \quad (20)$$

Because of the dimension of the problem it may be difficult to enumerate and compare all parameter combinations. A rational search algorithm is therefore needed.

[29] Before addressing the problem of finding an optimal parameter set, we note that quite often the traditional log linear regression model performs satisfactorily in regional frequency analyses. It is therefore of interest to be able to include linear terms in the additive model if the relationship between the dependent variable and some of the independent variables exhibit distinct linearity. For practical implementation the selection must be based on objective considerations. This problem is closely related to that of estimating the smoothing parameters for the individual smoothers in the additive model. The linear regression term $\beta_j x_j$ may be seen as one candidate among all possible smoothers $f_j(x_j)$. In the following, we outline a procedure for estimating the components of the additive model including possible linear terms. The method is similar in nature to the method of stepwise selection of variables in linear regression models and significantly reduces estimation time compared to the case of complete enumeration.

[30] The vector of fitted values at the observation points may be written

$$\hat{\mathbf{f}}(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n) = \bar{y} + \mathbf{S}_{\Omega_1}\mathbf{Y} + \mathbf{S}_{\Omega_2}\mathbf{Y} + \cdots + \mathbf{S}_{\Omega_d}\mathbf{Y}, \quad (21)$$

where $\mathbf{S}_{\Omega_i}$ is the smoother matrix associated with variable i and $\Omega_i = (h_i, p_i, K_i)$ is a vector containing the parameters of the ith smooth, that is, the size of smoothing window, the degree of the local polynomial, and the kernel type, respectively. We shall assume that the kernel type has been fixed in advance and also that the degree of the local polynomial has been fixed to one. These assumptions are not very restrictive and should not have any significant implications for the performance of the method for reasons already discussed. The only parameters to be determined are

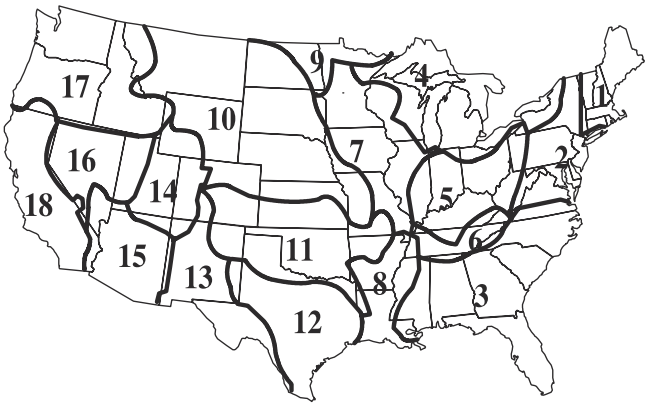

Figure 1. Hydrologic regions of the United States [after *Slack et al.*, 1993].

the size of the smoothing window for each individual variable. Rather than using a fixed kernel, we use a variable-sized neighborhood based on the k nearest neighbors, where k now is the smoothing parameter to be determined. The parameter vector is thus given by $\Omega_i = (k_i, 1, K)$. Since k_i must be between 1 and n, there is a total of n possible models. In addition, we consider the option of including a strictly linear term in the model and not including the ith variable at all for a total of $n + 2$ models. The model candidates for the smoothing of the ith variable can be enumerated as follows:

$$\left[\Omega_i^{(1)}, \Omega_i^{(2)}, \ldots, \Omega_i^{(n)}, \Omega_i^{(n+1)}, \Omega_i^{(n+2)}\right] = [(1,1,K),(2,1,K),\ldots,(n,1,K),(n,1,\text{rect}),(0,1,K)]. \quad (22)$$

The first n models are nonparametric fits based on different numbers of nearest neighbors (1 to n). Model $n + 1$, i.e., $\Omega_i^{(n+1)} = (n, 1, \text{rect})$, uses all n observations and a rectangular kernel to fit a first order polynomial. This corresponds to including a linear term in the model. Model $n + 2$ uses no observations at all and corresponds to omitting the ith variable from the regression. The models appear in order of decreasing degrees of freedom. The first model has n degrees of freedom (reproduces the observations exactly), the degrees of freedom for the following models decrease with increasing k, the linear term has one degree of freedom, and the omission of the ith variable corresponds to the case of zero degrees of freedom.

[31] To estimate the parameters of the additive model, $\boldsymbol{\Omega} = \left[\Omega_1^{(k_1)}, \Omega_2^{(k_2)}, \ldots, \Omega_d^{(k_d)}\right]$, where k_i represent the number of the model associated with the ith variable, we suggest the following procedure. Start with a reasonable initial guess, for example, based on multiple linear regression. Then consider the $2d$ models arising by decreasing and increasing each k_i by 1, i.e., consider models $\left[\Omega_1^{(k_1-1)}, \Omega_2^{(k_2)}, \ldots, \Omega_d^{(k_d)}\right]$, $\left[\Omega_1^{(k_1+1)}, \Omega_2^{(k_2)}, \ldots, \Omega_d^{(k_d)}\right]$, $\left[\Omega_1^{(k_1)}, \Omega_2^{(k_2-1)}, \ldots, \Omega_d^{(k_d)}\right]$, …, $\left[\Omega_1^{(k_1)}, \Omega_2^{(k_2)}, \ldots, \Omega_d^{(k_d+1)}\right]$, each one estimated using the back fitting algorithm. These models may be compared with the initial model using the GCV criterion. The best model becomes the new initial model. The search continues until there is no further improvement in the model fit.

[32] It is possible to evaluate the significance of model improvement using test techniques similar to those employed in conventional regression analyses. For example, we may test the null hypothesis that the true model is Ω_0 versus the alternative Ω_1 by computing the F statistic:

$$F = \frac{[\text{SSE}(\boldsymbol{\Omega}_0) - \text{SSE}(\boldsymbol{\Omega}_1)]/(\nu_0 - \nu_1)}{\text{SSE}(\boldsymbol{\Omega}_1)/(n - \nu_o)}, \quad (23)$$

where SSE denotes the sum of squared errors and ν_0 and ν_1 denote the degree of freedom of the two models.

4. Application

4.1. Data

[33] The performance of the proposed method in comparison with traditional log linear regression methods and the region-of-influence approach by *Tasker et al.* [1996] is investigated through an application to selected regions of the United States. Our comparison serves only as illustration as we do not believe it is possible to prove the general superiority of one method of regression over another. The performance of different methods depends on the underlying data and the measures used for comparison. Therefore comparisons will invariably be local in nature.

[34] A critical factor affecting the relative performance of different methods is the linearity between the dependent and independent variables (or their log transforms). Figure 1 shows the 21 hydrological regions defined in the Hydroclimatic Data Network of U.S. Geological Survey (USGS) described by *Slack et al.* [1993]. There is a total of 1659 stations included in the network. Flow data are largely unregulated and presumably of good quality. To get a preliminary idea of the linearity of flow quantiles and explanatory variables, we considered the regression of log Q_{50} on log A. Q_{50} was estimated from annual peak flows using the GEV distribution and the method of probability weighted moments [*Hosking et al.*, 1985]. The GEV distribution was chosen somewhat arbitrarily; being a three-parameter distribution, it is relatively flexible and should not introduce any significant bias in the estimation of quantiles with return periods <50 years. The catchment area A is by far the most important explanatory variable and occurs in all published regression relationships. The hypothesis of linearity between log Q_{50} on log A may be confronted with the alternative of a nonlinear relationship, estimated using nonparametric regression as described in section 2. The p value of the F test of linearity versus nonlinearity is given in Table 2 for each of the 21 regions. Values significant at the 5% level are indicated in bold. As it can be seen from Table 2, several regions did not pass the F test, suggesting that significant nonlinearities may be present.

[35] In the following, we further investigate regions 11 and 12. The Texas-Gulf of Mexico region represents a case of possible nonlinearities, whereas the Arkansas region represents a case with apparent linearity, at least between log Q_{50} and log A. The Arkansas region was also used in *Tasker et al.*'s [1996] study, which further motivates its inclusion here. In fact, we employed exactly the same data as Tasker et al. in order to allow a fair comparison with their study.

[36] It should be noted that the application of the proposed methodology within two geographical regions involves a

Table 2. Preliminary Analysis of the Log Linearity of Q_{50} Versus Catchment Area[a]

Region Number	Region	n	p value
1	New England	71	0.29
2	mid-Atlantic	167	**0.04**
3	South Atlantic-Gulf of Mexico	193	**0.01**
4	Great Lakes	57	0.75
5	Ohio	108	**0.00**
6	Tennessee	44	0.23
7	upper Mississippi	127	0.06
8	lower Mississippi	23	0.76
9	Souris-red-rainy	39	0.25
10	Missouri	144	**0.02**
11	Arkansas-white-red	87	0.24
12	Texas-Gulf of Mexico	90	**0.02**
13	Rio Grande	22	0.08
14	upper Colorado	44	0.33
15	lower Colorado	17	0.06
16	Great Basin	32	0.29
17	Pacific Northwest	191	**0.00**
18	California	115	0.72
19	Alaska	31	**0.02**
20	Hawaii	42	0.95
21	Caribbean	15	0.12

[a] The p value of the F test is for the null hypothesis of log linearity versus the alternative of a nonlinear relationship. Values significant at the 5% level are indicated in bold.

first level of (geographical) regionalization. This is not strictly needed in the proposed method; however, for comparison with current practice, the analysis was limited to data from established hydrologic regions.

4.2. Description of Models and Their Implementation

[37] All models considered here requires the selection of explanatory variables to be included in the regression. For the three models described in the following, we use the jackknife RMSE as criterion to select the optimal subspace of explanatory variables:

$$\mathrm{RMSE} = \left[\frac{1}{N}\sum_i \left(\log Q_{T,i} - \log \hat{Q}_{T,i}^{-i}\right)^2\right]^{1/2}. \qquad (24)$$

Here $Q_{T,i}$ represents the value of Q_T for site i as estimated from observed annual maximum flood data, and $\hat{Q}_{T,i}^{-i}$ is the regression estimate of Q_T ignoring flow information at site i. For the Texas region we used the GEV/probability weighted moments (PWM) method to quantile estimation, whereas for the Arkansas region we used the quantiles already calculated by *Tasker et al.* [1996]. The above summation is taken over all stations in the appropriate region. The logarithms of quantiles are considered in order to give approximately equal weight to small and large drainage basins. From an initial set of explanatory variables, we select the subset that minimizes the RMSE.

4.2.1. Log linear model

[38] The log linear models considered in the comparison are estimated by the method of least squares. Because the dependent variables Q_T are estimated from records of different length, a weighted least squares scheme is used in which each site is weighted with its record length. The combination of explanatory variables that minimizes the RMSE in equation (24) is retained for comparison with the other models.

4.2.2. ROI model

[39] The ROI approach involves the application of the traditional log linear model described above to a subset of stations in the region. There is an interaction between the number of explanatory variables and the size of the neighborhood. To determine the optimal combination, we proceed in two steps. In the first step the results for the ordinary log linear model are used to determine the best subset of $d = 1, 2, \ldots, 5$ explanatory variables from an initial array of five candidate variables. That is, we retain the best model considering one variable, the best model considering two variables, and so forth. Again the predictive RMSE is employed as comparison criterion. Next, for the models with $d = 1, 2, \ldots, 5$ explanatory variables, we vary the number k of stations in the neighborhood from 1 to N, the total number of stations in the region. The distance between two stations i and j is defined as

$$d_{ij} = \left(\mathbf{x}^i - \mathbf{x}^j\right)^T \Sigma^{-1} \left(\mathbf{x}^i - \mathbf{x}^j\right), \qquad (25)$$

where $\mathbf{x}^i$ is the vector of explanatory variables at site i and Σ is the covariance matrix of the explanatory variables. The combination of d and k that yields the smallest RMSE is retained as the best model.

4.2.3. Additive model

[40] Finally, we employ the proposed additive model. Assuming a first-order local polynomial and a tricube kernel function (see Table 1), the estimation problem reduces to that of finding the optimal size of a neighborhood in five dimensions, $\mathbf{K} = (k_1, k_2, \ldots, k_5)$. For the practical implementation, we consider the span parameter $\mathbf{H}$ defined as $\mathbf{H} = (h_1, h_2, \ldots, h_5)$, where $h_i = k_i/N$, rather than $\mathbf{K}$. Here h_i is the fraction of the total number of stations included in the neighborhood when regressing on the ith variable, and it takes values between zero and one.

[41] Initially, $\mathbf{H}$ is set to [0.5, 0.5, ..., 0.5]; that is, 50% of stations are included in each of the five univariate smooths. The search algorithm is then employed by, in turn, modifying each element of $\mathbf{H}$ by ±0.1 and chancing the parameter that leads to the smallest RMSE. When $h_i = 1$, a rectangular kernel is used, so that $h_i = 1$ corresponds to including the linear term $\beta_i x_i$ in the model. When $h_i = 0$, the variable x_i is not considered in the model. Hence, at the outset, all explanatory variables are included in the model, but some may be eliminated during optimization.

[42] Our experimentation showed that the step size affected the convergence of the algorithm. The value 0.1 was found to be a reasonable choice. The search ends when there is no further model improvement according to the GCV criterion. The estimation of additive models was accomplished using the gam and locfit functions in S-Plus.

4.3. Texas Region

[43] In the case study, regression models were developed and compared for quantiles of return period $T = 2, 5, 10, 25$, and 50. By restricting the study to low-return periods the uncertainty involved in estimating quantiles from local data is reduced, and attention can be focused on regression

Table 3. Physiographic and Climatological Data for Texas

Variable	Symbol	Description
Area	A	area of drainage basin (km^2)
Slope	S	slope of main channel (m/km)
Annual precipitation	P	basin mean annual precipitation (cm)
Elevation	E	mean elevation of drainage basin over MSL (m)
Channel length	L	length of main channel from divide to gauge (km)

model errors. Quantiles were estimated from observed records of annual floods using the GEV/PWM model as described by *Hosking et al.* [1985].

[44] Table 3 shows the explanatory variables considered by USGS for Texas. There are 90 sites in the region; however, because of missing data, only 69 sites are considered in this study. The RMSE criterion described above was used to choose the best combination of variables for the log linear model. The results are given in Table 4. For example, in the case of 50-year flood, minimum RMSE was obtained when considering the four variables A, S, P, and E, and the estimated model is

$$\log Q_{50} = 2.01 + 0.59 \log A + 0.35 \log S + 0.70 \log P - 0.21 \log E. \quad (26)$$

[45] The ROI model for the Texas data was estimated as described in section 4.2.2. For each of the considered quantiles the combination of variables and number of stations that minimizes the predictive RMSE is given in Table 5.

[46] The estimation of the additive model proceeded according to the procedure described in section 4.2.3, resulting in the span parameters given in Table 6. Figure 2 shows the smooth of log Q_{50} on each of the explanatory variables.

[47] The predictive RMSE values obtained with the three models are summarized in Table 7. The additive model systematically outperforms the other models for all quantiles considered. The relative gain in RMSE of the additive models over the other models appears to increase with increasing return period. The gain over the log linear model is of the order of 20–25% for the largest quantiles, whereas the gain over the ROI model is relatively modest, on the order of 10–15% for the largest quantiles.

4.4. Arkansas Region

[48] The Arkansas region represents a case where the log linear model is expected to perform relatively well since linearity of log A and log Q_{50} was not rejected (Table 2). Results for this region are included here to illustrate the performance of the additive model when the conditions for the classical log linear model, at least at a first glance, appear to be good.

[49] In order to reproduce the results of *Tasker et al.*'s [1996] study, we employed their basin shape factor defined as basin area divided by the square of the length of the main channel. This explanatory variable replaced main channel length.

[50] The results are summarized in Table 8. Again, the additive model outperforms the other two models in terms of predictive RMSE, with the ROI method a close second. The log linear model does significantly worse than the other two models. The additive model improves the RMSE of ~20–25% over the log linear model. A closer inspection of the relationship between dependent and independent variables reveals a distinct nonlinearity between log Q_{50} and channel slope which is most likely the cause of the poor performance of the log linear model.

5. Conclusions

[51] The log linear model is widely used in regional frequency analyses. Its simplicity has made it the preferred technique in many countries, including the United States, for estimating flood quantiles at ungauged sites. It is well suited for institutional implementation because regression equations for each state or hydrologic region can be developed and published. Application of such published relationships to a particular site is straightforward and only requires knowledge of the explanatory variables involved in the equation.

[52] It can be argued that the simplicity of the traditional log linear model is also its primary shortcoming. Assumed linearities may not always be present, and abrupt discontinuities across region boundaries may have little hydrologic justification. The ROI approach developed by *Tasker et al.* [1996] overcomes some of the limitations of conventional regression models. While preserving the

Table 4. Parameters of Log Linear Models for Texas

T	β_0	β_A	β_S	β_P	β_E	β_L
2	0.053	0.548	0.294	1.339	−0.263	0.262
5	0.862	0.531	0.321	1.051	−0.236	0.211
10	1.265	0.535	0.340	0.919	−0.230	0.172
25	1.675	0.594	0.317	0.817	−0.194	–
50	2.013	0.587	0.352	0.697	−0.205	–

Table 5. Optimal ROI Models for Texas

T	Variables	Number of Stations k
2	A, S, P, E	62
5	A, P, S, E, L	64
10	A, P, S, E, L	62
25	A, P, S, E, L	60
50	A, P, S, E, L	60

Table 6. Parameters of Additive Models for Texas

	Span Parameter				
Quantile	A	S	P	E	L
Q_2	1.0	1.0	0.5	0.5	0.9
Q_5	1.0	0.6	0.7	0.3	0.7
Q_{10}	1.0	0.4	0.8	0.3	0.6
Q_{25}	0.4	1.0	0.9	0.3	0.6
Q_{50}	0.4	1.0	1.0	0.4	0.5

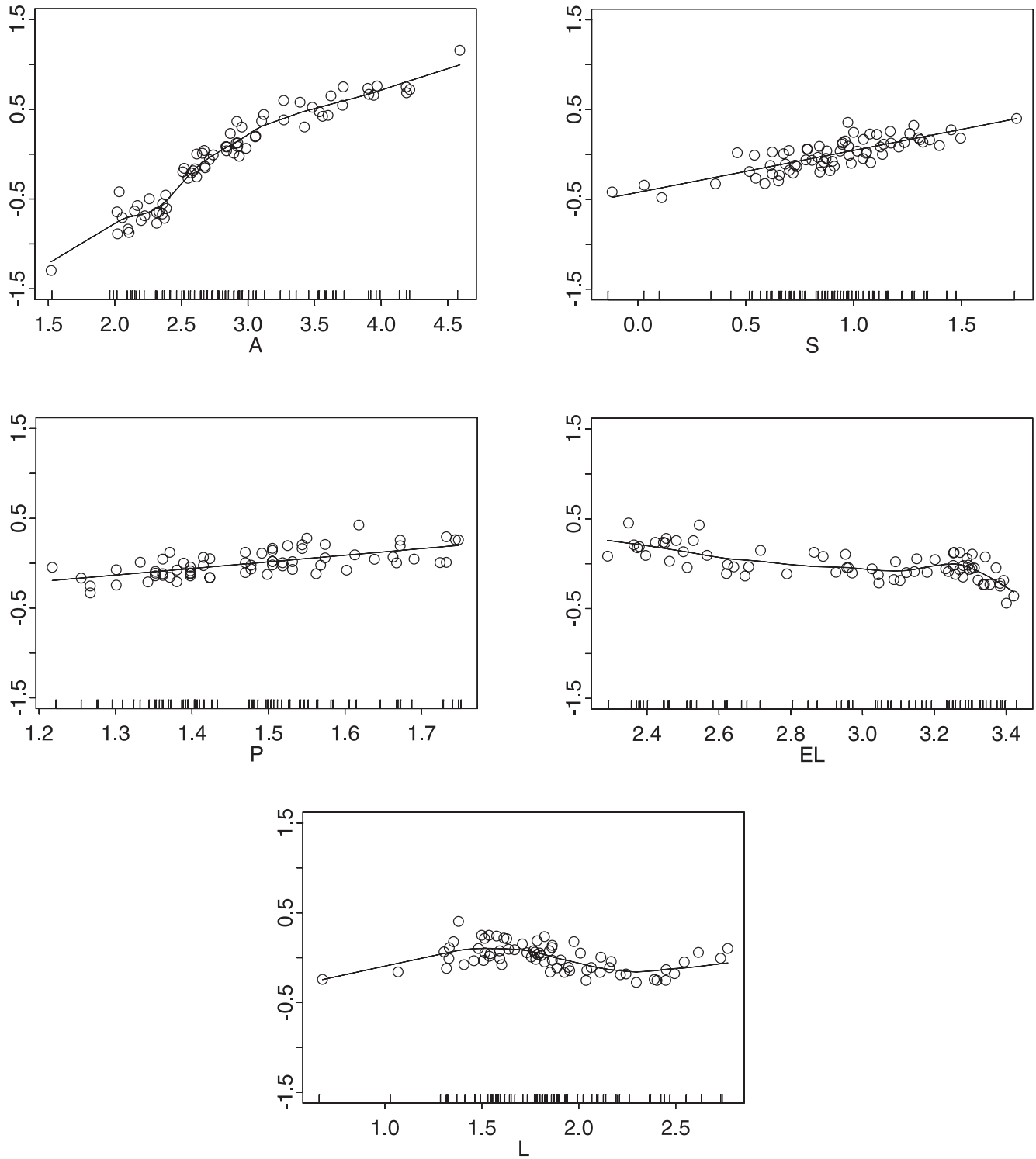

Figure 2. Additive nonparametric regression of Q_{50} on five basin characteristics (Texas region).

simplicity of the log linear model, it avoids the problem of discontinuities across region boundaries and, to some degree, also circumvents the problems of nonlinearities over the range of explanatory variables. As described in this paper, the parameters of the model can be estimated in a rational and objective way. It should be noted that our application of the approach does not give full credit to the method since we employed it within confined regions (Texas and Arkansas). This may still yield some discontinuities over boundaries of the greater regions. Ideally, there should be no region boundaries, and the method should be allowed to select automatically (according to the proximity criterion) the stations to be included in the region of influence.

[53] A potential further improvement in regional flood estimation is the additive nonparametric regression model

presented in detail in this paper. It builds on the ROI principle in the sense that only basins that are similar to the ungauged basin are considered in the estimation. However, it adds increased flexibility in several ways. First, it does not impose linearity or log linearity between dependent and independent variables. Secondly, the proximity criterion inherent in the nonparametric regression model has a more rational foundation than the one used in the ROI approach (equation (1)). For example, equation (1) does not discriminate between important and unimportant variables. A potentially useful station may be excluded from the ROI based on differences in an unimportant explanatory variable (of course, a sensible use of equation (1) would involve a careful selection of variables). In contrast to this, the additive model weights the importance of each variable. The extent of the region is defined individually for each explanatory variable. For example, to determine the contribution of drainage area to the quantile, it may turn out to be advantageous to consider all stations in the greater region, while another variable, say the channel slope, would involve only half of the stations. In addition to this, the kernel method used to estimate the univariate contributions to the regression function provides robustness by tapering the weights so that sites are weighted according to their similarity to the ungauged site.

[54] Our application and comparison of the log linear model and the additive model suggest that the latter may have significant advantages in terms of predictive ability. For the particular cases studied here, improvements in RMSE on the order of 20–25% over the traditional log linear model were observed. The additive model also performed better than the ROI approach, although the differences were less pronounced. For several reasons, we avoid making any general conclusions about the relative performance of the methods. The performance of the models depends on the particularities of the data and the degree to which various assumptions are verified. Furthermore, the choice of comparison figures is always subject to questioning. However, we do believe the potential gain by the additive nonparametric regression model fully justifies its increased complexity compared with the traditional log linear model and its incorporation into hydrological practice.

Appendix A: Generalized Cross Validation Mean Square Error

[55] To select the best smoothing parameter, one could envision minimizing the mean square error (MSE) of the residuals. This quantity can be estimated from the observations as

$$\mathrm{MSE}(\Omega) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_\Omega(x_i)\right]^2, \tag{A1}$$

where $\Omega = (h, p, K)$ are the smoothing parameters. The above estimator is, however, an overly optimistic goodness of fit measure because y_i is used to estimate $f(x_i)$. In fact, MSE(Ω) can be made equal to zero by selecting a vector of smoothing parameters that interpolates the observations. A more practical measure is the cross validation mean square error (CVMSE) which is given by

$$\mathrm{CVMSE}(\Omega) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}_\Omega^{-i}(x_i)\right]^2, \tag{A2}$$

where $\hat{f}_\Omega^{-i}(x_i)$ is the estimate of $f_\Omega(x_i)$ obtained by ignoring the observation (x_i,y_i). The parameters Ω may be determined in an objective way by minimizing the CVMSE. An inconvenience associated with equation (A2) is the need to perform n nonparametric regressions. Fortunately, this can be avoided. From equation (4) it can be seen that the weight sequence $\omega_j(x_i) = S_{ij}$, $j = 1,\ldots, n$ for estimating $f_\Omega(x_i)$ sum to 1. The jackknifed estimate $\hat{f}_\Omega^{-i}(x_i)$ corresponds to setting S_{ii} equal to zero and adjusting the remaining weights so that they sum to one. Therefore we have

$$\hat{f}_\Omega^{-i}(x_i) = \sum_{\substack{j=1\\ j\neq i}}^{n}\frac{S_{ij}}{1-S_{ii}}y_j. \tag{A3}$$

A few manipulations allow us to rewrite this expression as

$$y_i - \hat{f}_\Omega^{-i}(x_i) = \frac{y_i - \hat{f}_\Omega(x_i)}{1-S_{ii}}, \tag{A4}$$

which upon insertion in equation (A2) yields

$$\mathrm{CVMSE}(\Omega) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}_\Omega(x_i)}{1-S_{ii}}\right]^2. \tag{A5}$$

The generalized cross validation mean square error is obtained by replacing S_{ii} by its average value tr(**S**)/n:

$$\mathrm{GCV}(\Omega) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}_\Omega(x_i)}{1-\mathrm{tr}(\mathbf{S})/\mathrm{n}}\right]^2. \tag{A6}$$

Table 7. Comparison of Predictive RMSE for Texas

	Predictive RMSE		
Quantile	Log Linear	ROI	Additive
Q_2	0.178	0.166	0.165
Q_5	0.173	0.150	0.146
Q_{10}	0.179	0.149	0.144
Q_{25}	0.197	0.162	0.143
Q_{50}	0.216	0.182	0.160

Table 8. Comparison of Predictive RMSE for Arkansas

	Predictive RMSE		
Quantile	Log Linear	ROI	Additive
Q_2	0.234	0.190	0.189
Q_5	0.209	0.176	0.156
Q_{10}	0.207	0.160	0.158
Q_{25}	0.213	0.178	0.164
Q_{50}	0.219	0.185	0.172

References

Acreman, M., and S. E. Wiltshire, Identification of regions for regional flood frequency analysis (abstract), *Eos Trans. AGU*, *68*(44), 1262, 1987.

Acreman, M., and S. Wiltshire, The regions are dead; long live the regions, in *Methods of Identifying and Dispensing With Regions for Flood Frequency Analysis*, IAHS Publ., *187*, 175–188, 1989.

Buja, A., T. J. Hastie, and R. J. Tibshirani, Linear smoothers and additive models (with discussion), *Ann. Stat.*, *17*, 453–555, 1989.

Burn, D. H., Evaluation of regional flood frequency analysis with a region of influence approach, *Water Resour. Res.*, *26*, 2257–2265, 1990a.

Burn, D. H., An appraisal of the "region of influence" approach to flood frequency analysis, *Hydrol. Sci. J.*, *35*, 149–165, 1990b.

Cleveland, W. S., Robust locally weighted regression and smoothing scatterplots, *J. Am. Stat. Assoc.*, *74*, 829–836, 1979.

Dalrymple, T., Flood-frequency analyses, *Manual of hydrology*, part 3, *Flood-Flow Techniques*, U.S. Gov. Print. Off., Washington, D.C., 1960.

Green, P. J., and B. W. Silverman, *Nonparametric regression and generalized linear models*, Chapman and Hall, New York, 1994.

Härdle, W., *Applied Nonparametric Regression*, Cambridge Univ. Press, New York, 1990.

Hastie, T. J., and R. J. Tibshirani, *Generalized Additive Models*, Chapman and Hall, New York, 1990.

Hosking, J. R. M., J. R. Wallis, and E. F. Wood, Estimation of the generalized extreme value distribution by the method of probability weighted moments, *Technometrics*, *27*, 251–261, 1985.

Jennings, M. E., W. O. Thomas, and H. C. Riggs, Nationwide summary of U.S. Geological Survey regional regression equations for estimating magnitude and frequency of floods for ungauged sites, 1993, *U. S. Geol. Surv. Water Resour. Invest. Rep.*, *94-4002*, 196, 1994.

Latraverse, M., Modelisation additive par polynomes locaux pour la regionalisation des quantiles de crues: Approche optimale de regression par region d'influence, Ph.d. thesis, INRS-Eau, Univ. of Quebec, Montreal, Quebec, Canada, 2000.

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

Slack, J. R., A. M. Lumb, and J. M. Landwehr, Hydro-climatic data network (hcdn): Streamflow data set, 1874–1988, *U.S. Geol. Surv. Water Resour. Invest. Rep.*, *93-4076*, 1993.

Tasker, G. D., Comparing methods of hydrologic regionalization, *Water Resour. Res.*, *18*, 965–970, 1982.

Tasker, G. D., S. A. Hodge, and C. S. Barks, Region of influence regression for estimating the 50-year flood at ungauged sites, *Water Resour. Bull.*, *32*, 163–170, 1996.

Wand, M. P., and M. C. Jones, *Kernel Smoothing*, Chapman and Hall, New York, 1995.

Zrinji, Z., and D. H. Burn, Flood frequency analysis for ungauged sites using a region of influence approach, *J. Hydrol.*, *153*, 1–21, 1994.

M. Latraverse, IREQ, Hydro-Quebec's Research Institute, 1800 Boulevard Lionel-Boulet, Varennes, Quebec, Canada J3X 1S1. (latraverse.marco@ireq.ca)

P. F. Rasmussen, Department of Civil Engineering, University of Manitoba, Winnipeg, Manitoba, Canada R3T 5V6. (rasmusse@cc.umanitoba.ca)

Bernard Bobée, INRS-Eau, University of Quebec, 2800 rue Einstein, CP 7500 Sainte-Foy, Quebec, Canada G1V 4C7. (bernard_bobee@inrs-eau.uquebec.ca)