Flood Regionalization: A Hybrid Geographic and Predictor-Variable Region-of-Influence Regression Method

Ken Eng[1]; P. C. D. Milly[2]; and Gary D. Tasker[3]

Abstract: To facilitate estimation of streamflow characteristics at an ungauged site, hydrologists often define a region of influence containing gauged sites hydrologically similar to the estimation site. This region can be defined either in geographic space or in the space of the variables that are used to predict streamflow (predictor variables). These approaches are complementary, and a combination of the two may be superior to either. Here we propose a hybrid region-of-influence (HRoI) regression method that combines the two approaches. The new method was applied with streamflow records from 1,091 gauges in the southeastern United States to estimate the 50-year peak flow (Q_{50}). The HRoI approach yielded lower root-mean-square estimation errors and produced fewer extreme errors than either the predictor-variable or geographic region-of-influence approaches. It is concluded, for Q_{50} in the study region, that similarity with respect to the basin characteristics considered (area, slope, and annual precipitation) is important, but incomplete, and that the consideration of geographic proximity of stations provides a useful surrogate for characteristics that are not included in the analysis.

DOI: 10.1061/(ASCE)1084-0699(2007)12:6(585)

CE Database subject headings: Floods; Regression analysis; Regional analysis; Statistics; Surface waters; Streamflow; Hybrid methods.

Introduction

Streamflow cannot possibly be monitored at every location on a river. Consequently, hydrologists and engineers, state and local agencies, and the general public often require information on streamflow characteristics at ungauged sites. As a solution, streamflow characteristics at ungauged sites are inferred from records at similar, nearby gauged sites. A method to calculate streamflow characteristics at ungauged sites is to use regional regression models that relate observable basin characteristics, such as drainage area, to streamflow characteristics, such as the 50-year-return peak discharge. (The 50-year-return peak is the annual peak flow that is expected to be exceeded on average in 1 out of 50 years; it is equivalent to the 98th percentile of the distribution of annual peak streamflows.)

The accuracy of these regression models is limited by the realism of their structure (i.e., the choice of predictor variables and the assumed functional relation between predictor variables and predictand) and by the accuracy of the estimates of the model parameters. Model structure is constrained both by scientific understanding of physical controls of the streamflow process and by the availability of data on the predictor variables that are suggested by such understanding. Because hydrologic understanding and data availability are incomplete, regression-model parameters vary in space, and estimates of these parameters from data at gauged sites are subject to temporal sampling errors, with better estimates corresponding to longer periods of measurement. The spatial variability of regression-model parameters implies a need to estimate them by analysis of streamflow data from hydrologically similar gauged basins. Temporal sampling errors and incomplete characterization of similarity imply a need to use as many similar sites as possible. In the selection of sites for a regression, the statistical advantage of increasing the number of sites must be traded off against the physical disadvantage of including increasingly dissimilar sites. Thus, two critical problems are how to define hydrologic similarity and how to choose an optimal number of similar sites for a regression.

A common solution to these two problems is to assume the hydrologic similarity (homogeneity of geological, topographic, and climatic characteristics) among sites at some sufficiently small spatial scale. For practical reasons, the analysis usually begins with a politically defined area (e.g., a state in the United States), and that region may then be divided further on the basis of hydrologic judgment. Once the regions are defined, the parameters of the model are determined for each region. Within a given region, a single regression equation is used for all ungauged sites.

An alternative is to define regions for regression in predictor-variable space; the regression is performed on a subset of stations for which the basin characteristics are, by some overall measure, closest to those at the ungauged site of interest. In the typical application of this "region-of-influence" (RoI) approach, further, a unique "region" is defined for each ungauged site (Burn 1990).

[1] Research Hydrologist, National Research Program, U.S. Geological Survey, 12201 Sunrise Valley Dr., Mail Stop 430, Reston, VA 20192 (corresponding author). E-mail: keng@usgs.gov.

[2] Research Hydrologist, National Research Program, U.S. Geological Survey, Geophysical Fluid Dynamics Laboratory/NOAA, Route 1, Forrestal Campus, Princeton, NJ 08542. E-mail: cmilly@usgs.gov.

[3] Scientist Emeritus, National Research Program, U.S. Geological Survey, 12201 Sunrise Valley Dr., Mail Stop 430, Reston, VA 20192. E-mail: gdtasker@usgs.gov.

Note. Discussion open until April 1, 2008. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on December 6, 2005; approved on February 5, 2007. This paper is part of the *Journal of Hydrologic Engineering*, Vol. 12, No. 6, November 1, 2007. ©ASCE, ISSN 1084-0699/2007/6-585–591/$25.00.

Yet another alternative is to define a region of influence analogously as a disk in geographic space. We thus qualify RoI approaches as predictor-variable (PRoI) or geographic (GRoI), depending on which space (predictor-variable or geographic) is used to establish proximity. One strength of any RoI approach is that it centers the estimation space over the ungauged site, maximizing hydrologic similarity between gauged and ungauged basins. Of course, any RoI method can also be applied piecewise, i.e., within subjectively predefined regions.

An advantage of using geographic proximity to define the RoI stems from the facts that the model is not likely to contain all controls on streamflow and that some of the missing controls are spatially coherent. Even if all the important predictor variables are included in the model, the functional form of the regression model may be incorrectly specified. For example, if streamflow characteristic y depends nonlinearly on basin characteristic x, but a linear relation is instead assumed because the nonlinear form is unknown, then it would be helpful to apply different linear relations over different ranges of x. If x is spatially correlated, this could be accomplished by a GRoI approach.

In fact, the basin characteristic x of gauged sites may not always be spatially correlated, or x may be correlated only on very short length scales (relative to inter-gauge distances). In such situations, PRoI has the advantage of pooling data from nonadjacent but hydrologically similar basins. On the other hand, a PRoI can unnecessarily restrict the range of some predictor, causing increased variance of the regression coefficient associated with that predictor. The restricted range is useful if the relation is nonlinear, but can be counterproductive if the relation is linear.

The PRoI approach has had varying degrees of success (Tasker and Slade 1994; Tasker et al. 1996; Pope et al. 2001; Berenbrock 2002; Feaster and Tasker 2002; Law and Tasker 2003; Eng et al. 2005). Eng et al. (2005) suggest that the PRoI performance is a function of the spatial scale of the study area; performance decreases with increasing scale, so that PRoI works best on spatial scales smaller than those of a medium-sized state (e.g., Georgia in the United States). Merz and Blöschl (2005) obtained their best estimates of flood statistics when they considered both predictor-variable and geographic proximity.

To combine the strengths of PRoI and GRoI, we propose here the use of both predictor-variable and geographic space to define a RoI. To minimize the problem of cross correlation in geographically proximal records, parameter estimation is based on generalized-least-squares (GLS) regression (Stedinger and Tasker 1985). This RoI approach is referred to as the hybrid RoI (HRoI) approach. A comparison of the performance of PRoI, GRoI, and HRoI is presented for the case of the 50-year peak discharge in the Gulf-Atlantic Rolling Plains.

Study Area and Data

Estimates of the 50-year peak discharge, $\hat{Q}_{50}$ (m^3/s), and basin characteristics for 1,091 streamflow-gauging stations in the southeastern United States (Fig. 1) were used in this study. The record lengths at these sites ranged from 10 to 103 years. The study area, gauged sites, and basin characteristics used in this paper were identical to those considered in Eng et al. (2005). These stations were selected because they were contained within the boundaries of a single physiographic region, the Gulf-Atlantic Rolling Plains (Hammond 1964). The $\hat{Q}_{50}$ values were estimated by the standard methods described in *Bulletin 17B* of the Hydrology Subcommittee of the Interagency Advisory Committee on Water Data (1982). In this preliminary study, we chose to examine a single return period, the 50-year return period, because it lies within the range commonly used in hydrologic analyses.

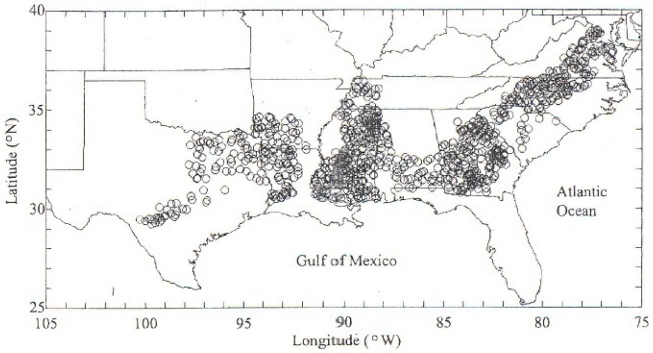

Fig. 1. Southeastern United States; circles represent gauged sites

Our analysis began with a set of eight basin characteristics available from a previous study (Eng et al. 2005). These were drainage area, A, main channel slope, S, mean basin elevation, E, forested area fraction, F, main-channel stream length, L, fractional area of basin occupied by reservoirs and lakes, SWB, mean annual precipitation, P, and mean minimum January temperature, JT. All predictors were logarithmically transformed; the fractional areas, some of which are zero, had an arbitrary constant of one percent added to allow this transformation. The A, S, E, F, L, and SWB values were estimated from U.S. Geological Survey 1:24,000 scale topographic maps. S was calculated as the average channel slope (elevation difference divided by distance along the main channel) between points located 10 and 85% of the distance from the gauging station to the basin divide. E was calculated as the average ground elevation above mean sea level from 20 to 80 points sampled in the basin. L was defined as the channel length from the gauged site to the basin divide. F and SWB were calculated by dividing the number of grid cells that contain either forest or surface water bodies by the total number of grid cells placed over the watershed. Isothermal maps (U.S. Department of Commerce 1976–1978) were used to get JT. Isohyetal maps (U.S. Department of Commerce 1976–1978) were used to obtain P.

From experience we knew that some basin characteristics were correlated with others, and we knew that some basin characteristics generally were not good predictors of flood characteristics. Thus, it was necessary to select a model (i.e., to choose a subset of the eight predictors to use in the regression) at some stage of the analysis. This selection could be made once, globally, at the start of the analysis, or once for every RoI (i.e., for more than a thousand cases) for each RoI method examined. Because we were not comfortable with any automated procedure for model selection, and because of our interest in computationally practical methods, we chose to make a single global selection of the predictors at the outset of the analysis. This selection was made on the basis of the Mallows (1973, 1995) C_p statistic. The methods described in detail by Mallows were applied to the 2^8 possible combinations of the eight available predictors. The analysis was performed on the entire set of data from 1,091 gauges in conjunction with ordinary-least-squares regression. The analysis indicated that the optimal model should use three predictors (A, S, and P). From ancillary computations, multicollinearity was determined to be insignificant among these three predictor variables (Eng et al. 2005). The correlation coefficient among the log (base 10) transformed $\hat{Q}_{50}$ and transformed A, S, E, F, L, SWB, P, and JT values were 0.9, −0.7, 0.1, 0.2, 0.8, 0.1, 0.3, and 0.1, respectively.

Methodology

The relation between the logarithmic (base 10) transforms of the peak discharges and the basin characteristics in the best-subsets regression model is

$$\log(Q_{50}) = \xi_0 + \xi_1 \log(A) + \xi_2 \log(S) + \xi_3 \log(P) + \varepsilon \qquad (1)$$

where Q_{50}=50-year peak flow; ξ_0, ξ_1, ξ_2, and ξ_3=regression parameters, A, S, and P are the drainage area, slope, and precipitation, respectively, and ε=model error, with mean equal to zero and variance equal to γ^2. For each ungauged site, Eq. (1) was applied at all sites within the RoI of the ungauged site to estimate the ξ parameters and then at the ungauged site to estimate Q_{50}.

The historical estimate of $\log(Q_{50})$ at gauged sites, $\log(\hat{Q}_{50})$, is derived from a sample of observed flows at each gauged site. The associated temporal sampling error, η, is defined by

$$\eta = \log(\hat{Q}_{50}) - \log(Q_{50}). \qquad (2)$$

Substituting Eq. (2) into Eq. (1) gives

$$\log(\hat{Q}_{50}) = \xi_0 + \xi_1 \log(A) + \xi_2 \log(S) + \xi_3 \log(P) + \upsilon \qquad (3)$$

where $\upsilon = \varepsilon + \eta$. Time-sampling errors from basins close together will generally be correlated, because the finite sample of observed flows at one site temporally overlaps the sample from another and temporal variations of flows are spatially correlated.

A GLS parameter estimation technique was used to perform the regression in the presence of cross correlation of υ, following the assumption that model error ε is not spatially correlated (Stedinger and Tasker 1985). Estimates of ξ_0, ξ_1, ξ_2, and ξ_3 are $\hat{\xi}_0$, $\hat{\xi}_1$, $\hat{\xi}_2$, and $\hat{\xi}_3$, respectively. The vector $\hat{\boldsymbol{\xi}}$ of these parameter estimates is given by

$$\hat{\boldsymbol{\xi}} = (\mathbf{X}_{\mathrm{RoI}}^T \hat{\boldsymbol{\Lambda}}^{-1} \mathbf{X}_{\mathrm{RoI}})^{-1} \mathbf{X}_{\mathrm{RoI}}^T \hat{\boldsymbol{\Lambda}}^{-1} \hat{\mathbf{Y}}_{\mathrm{RoI}} \qquad (4)$$

where $\mathbf{X}_{\mathrm{RoI}}$=$(J \times 4)$ matrix of $\log(A)$, $\log(S)$, and $\log(P)$ values at J sites in the RoI of the ungauged site, augmented by a column of ones; J=number of gauged basins in the RoI; $\hat{\mathbf{Y}}_{\mathrm{RoI}}$=$(J \times 1)$ vector of $\log(\hat{Q}_{50})$ values; and $\hat{\boldsymbol{\Lambda}}$=matrix containing the estimates of the correlation of υ across sites in the RoI. The main diagonal elements of $\hat{\boldsymbol{\Lambda}}$ thus include a part associated with ε, and all elements include the effect of η. Following Tasker and Stedinger (1989), $\hat{\boldsymbol{\Lambda}}$ is given as

$$\hat{\Lambda}_{pq} = \begin{cases} \gamma^2 + \dfrac{\hat{s}_p^2[1 + K_p g_p + 0.5K_p^2(1 + 0.75g_p^2)]}{m_p}, & (p = q) \\ \dfrac{r_{pq}\hat{s}_p\hat{s}_q m_{pq}[1 + 0.5K_p g_p - 0.5K_q g_q + 0.5K_p K_q(r_{pq} + 0.75g_p g_q)]}{m_p m_q}, & (p \neq q) \end{cases} \qquad (5)$$

where the subscripts p and q=indices of gauged sites in the RoI; K_p and K_q=log-Pearson Type III distribution standard deviate for stations p and q; g_p and g_q=skewness coefficients for stations p and q determined by procedures outlined in Bulletin 17B of the Hydrology Subcommittee of the Interagency Advisory Committee on Water Data (1982); m_p and m_q=site specific record lengths; m_{pq}=concurrent record length for stations p and q; $\hat{s}_p$ and $\hat{s}_q$=estimates of the standard deviation of annual peaks; and r_{pq}=sample cross correlation of annual peaks at stations p and q. The $\hat{s}_p$ and $\hat{s}_q$ values that had been computed from annual peak streamflow records were not used in Eq. (5), because these values would produce biased regression-parameter estimates, as explained by Tasker and Stedinger (1989). Instead, the $\hat{s}_p$ and $\hat{s}_q$ values were determined by ordinary least squares regressions against the basin characteristics in $\mathbf{X}_{\mathrm{RoI}}$ over the J sites in the RoI of the ungauged site

$$\log(\hat{s}_{p \text{ or } q}) = \kappa_0 + \kappa_1 \log(A) + \kappa_2 \log(S) + \kappa_3 \log(P) + \delta \qquad (6)$$

where κ_0, κ_1, κ_2, and κ_3=constants and δ=model error with mean equal to zero and variance equal to ψ^2. Values of the sample cross correlation were estimated approximately by (Tasker and Stedinger 1989)

$$r_{pq} = \Theta^{[d_{pq}/\alpha d_{pq} + d_o]} \qquad (7)$$

where d_{pq}=distance between gauges p and q (km); d_o=constant equal to 1 km; and Θ and α=dimensionless parameters. (Parameterization as a function of distance between basin centroids would be better than parameterization as a function of distance between gauges, but data were not readily available for the former approach.) The values of Θ=0.980 and α=0.00431 from a previous study in North Carolina (Pope et al. 2001) were assumed to be representative for our entire study area; we did not have the time series of annual-peak flow that would be needed to estimate r_{pq} for our data set. The cross-correlation values estimated by Eq. (7) ranged from 0.942 to 0.0312 at geographic distances of 3 and 660 km, respectively.

For every site at which the flow-estimation procedure was applied, parameter estimates were calculated by GLS regression on the gauged sites within the RoI of the estimation site. The RoI was formed in three different ways. GRoI used a RoI containing the n gauged sites geographically closest to the estimation site. PRoI used a RoI containing the n closest gauged sites in predictor-variable space. HRoI used a RoI containing the n closest gauged sites in predictor-variable space chosen from the subset of all gauged sites having a geographic distance less than D from the ungauged site; however, if fewer than n gauges were available within a distance D of a given estimation site, then the limit D was ignored and the n geographically closest gauges were used; i.e., HRoI reverts to GRoI in that situation. (Thus, in the limit as D approaches zero, HRoI reduces to GRoI, and in the limit as D becomes arbitrarily large, HRoI reduces to PRoI.) Distance in predictor-variable space from the ungauged site to the gauged site j, R_j, is defined as

$$R_j=\left[\left(\frac{\log(A)-\log(A)_j}{\sigma_{\log(A)}}\right)^2+\left(\frac{\log(S)-\log(S)_j}{\sigma_{\log(S)}}\right)^2+\left(\frac{\log(P)-\log(P)_j}{\sigma_{\log(P)}}\right)^2\right]^{1/2} \qquad (8)$$

where $\sigma_{\log(A)}$, $\sigma_{\log(S)}$, and $\sigma_{\log(P)}$=sample standard deviations of $\log(A)$, $\log(S)$, and $\log(P)$, respectively (computed from data from the entire study region). The decision to use the n closest gauged sites in either geographic or predictor-variable space to form a RoI in this study was guided by the results of Eng et al. (2005). They showed for OLS that a RoI formed of either the n closest geographic or predictor-variable space sites was superior to one containing all geographic or predictor-variable similar sites below a predetermined threshold value of distance.

Let $\hat{Q}_{R50}$ be the GLS-regressed estimate at a gauged site not used in the regression and treated as an ungauged site. The log-space metrics for model optimization and evaluation were the root-mean-square difference (RMSE) between $\hat{Q}_{R50}$ and $\hat{Q}_{50}$ (Aitchison and Brown 1957), expressed, in percent of the actual flow statistic, as

$$\text{RMSE}=100[(200.74)^{\sigma_e^2}-1]^{1/2} \qquad (9)$$

where

$$\sigma_e^2=\left\{\frac{\sum_{i=1}^{N}[\log(\hat{Q}_{50})_i-\log(\hat{Q}_{R50})_i]^2}{N}\right\} \qquad (10)$$

and the average error of $\hat{Q}_{R50}$ (BIAS) as

$$\text{BIAS}=\left\{\frac{\sum_{i=1}^{N}[\log(\hat{Q}_{50})_i-\log(\hat{Q}_{R50})_i]}{N}\right\} \qquad (11)$$

where $(\hat{Q}_{50})_i$=estimate of Q_{50} at site i based on streamflow records and N=total number of sites in the data set of interest.

The split-sampling and the optimization-evaluation procedures used in this study were identical to the procedures implemented by Eng et al. (2005). The set of 1,091 stations was split into three equally sized subsets in such a way that the three subsets had very

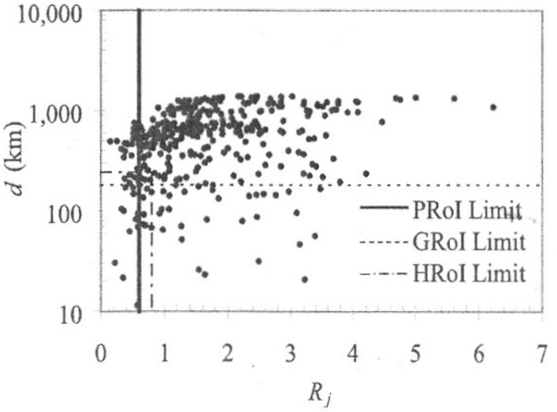

Fig. 2. Distances in predictor-variable (R_j) and geographic (d) space between an evaluation site in Mississippi (U.S. Geological Survey Station No. 02436000) and the other sites in its evaluation set; the horizontal and vertical lines define the extent of the three RoIs

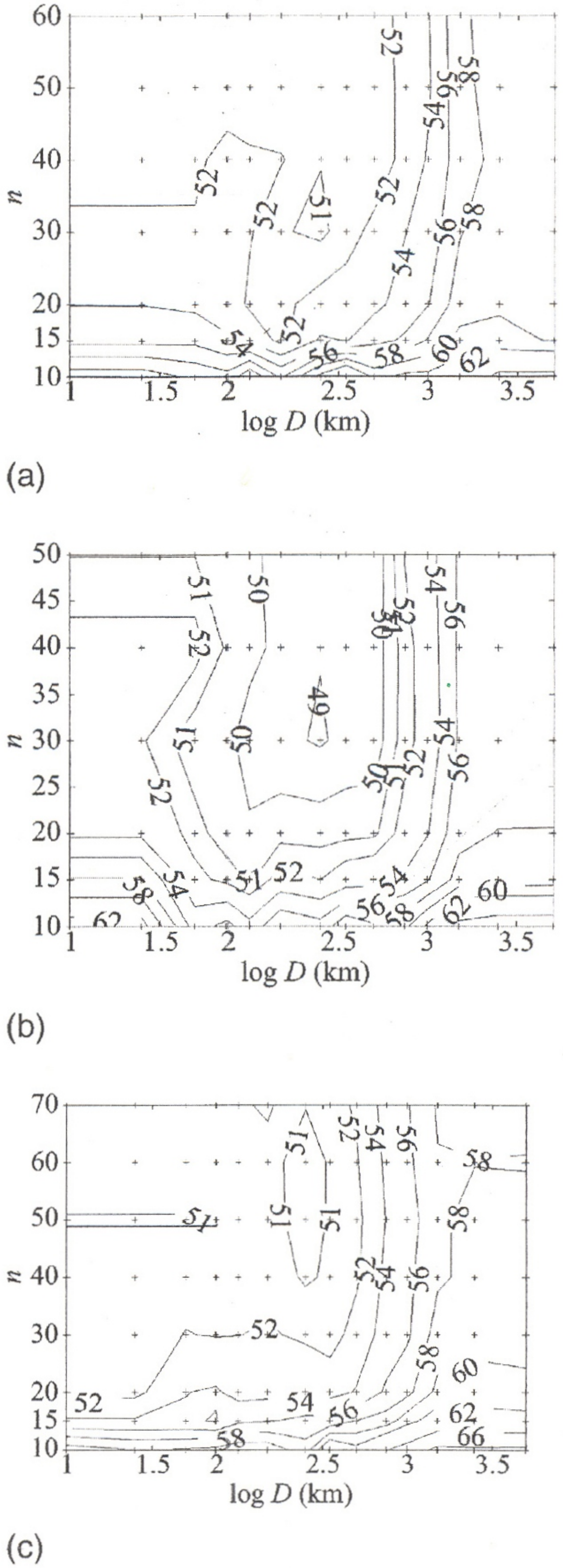

Fig. 3. Optimization RMSE (%) as a function of log of the target maximum region-of-influence (RoI) distance, D, and the number of sites contained in the RoI, n, for the HRoI approach. Computed values are indicated by the plus sign. Split-sampled sets (a) 2 and 3; (b) 1 and 3; and (c) 1 and 2.

Table 1. Summary of Split-Sample Optimization and Evaluation Results by Subset

		Optimization			Evaluation	
Approach	Evaluation subset	n	D (km)	RMSE (%)	RMSE (%)	Outlier count
PRoI	1	40	—	58.8	59.3	19
	2	30	—	56.3	60.9	17
	3	40	—	58.5	58.4	22
GRoI	1	50	—	51.0	55.8	17
	2	50	—	51.0	56.2	16
	3	50	—	50.9	50.3	17
HRoI	1	30	250	50.8	50.1	11
	2	30	250	48.9	55.6	15
	3	60	250	50.2	47.7	13

Note: The outlier count is the number of estimation sites having a residual error whose absolute value was more than twice the RMSE.

similar statistical distributions of the three predictor variables. Two of the three subsets were then combined and used in an optimization step to calculate RMSE values for various values of n and D for HRoI and for various values of n for both PRoI and GRoI. The lowest resulting values of RMSE and the corresponding values of n and (for HRoI) D were noted. The third subset was then used to evaluate model performance, by calculating the RMSE value associated with the optimal n and (for HRoI) D determined in the previous step. All three possible combinations of subsets for this optimization-evaluation procedure were employed, and an overall RMSE value was then computed as the root mean square value of the three individual values.

Results

Fig. 2 illustrates the formation of the three types of RoI for an example estimation site for a given set of RoI parameters. The PRoI contains the 40 closest gauges in predictor-variable space; the GRoI contains the 50 geographically closest gauges; and the HRoI contains the 30 closest gauges in predictor-variable space that are within 250 km of the example site. The PRoI contains some sites that are located far (500–800 km) from the estimation site, whereas the HRoI ranges no farther than 250 km from the estimation site. For the GRoI, the greatest distance in predictor-variable space between the estimation site and the gauged sites is more than four times the analogous distance for the HRoI. By using both distance measures in its definition, the HRoI avoids extreme values of either.

Table 2. Combined Split-Sample Evaluation Results

Approach	$\overline{\text{RMSE}}$	Estimated model-error RMSE (%)	Total outlier count
PRoI	59.5	47.6	58
GRoI	54.1	33.5	50
HRoI	51.1	35.4	39

Note: The $\overline{\text{RMSE}}$ (%) is the root-mean-square value across the three subsets of the evaluation RMSE (%) values from Table 1. The estimated model-error RMSE (%) values are the intercepts from the fitted lines in Fig. 5. The total outlier count is the sum of the number of outliers from the three split-sampled data sets.

The dependence of RMSE on n and D for the HRoI approach is shown in Fig. 3. For a given value of D, the RMSE generally decreases as n increases for n less than 30. For n greater than 30, the sensitivity to n is very small. For a fixed and sufficiently large value of n, however, the RMSE is minimized at an intermediate value of D. The split-sampled optimized n and D values and the optimization and evaluation RMSE values are presented in Table 1 for all three approaches. Evaluation results from Table 1, but aggregated across the three evaluation sets, can be found in Table 2, along with additional results to be described below. For PRoI and GRoI (Fig. 4, equivalent to large-D and small-D limits, respectively, of Fig. 3), the relation between RMSE and n is similar to that for HRoI. The HRoI approach resulted in the smallest overall evaluation RMSE (Table 2). All approaches had BIAS values smaller than 0.01. The optimal n and D values are not very sensitive to choice of data subset used for optimization, as shown in Table 1. For comparison, the PRoI overall evaluation RMSE values for much more geographically limited physiographic subregions inside individual states ranged from 46 to 53% for North Carolina (Pope and Tasker 1999), from 36 to 40% for South Carolina (Feaster and Tasker 2002), and from 37 to 48% for Tennessee (Law and Tasker 2003).

The RMSE values in Figs. 3 and 4 and Tables 1 and 2 include both model errors (ε) and temporal sampling errors (η). The former depends on the choice of RoI, whereas the latter should have a variance proportional to the inverse of the record length. We estimated the part of the evaluation RMSE associated with model error as follows: The values of $1/\sqrt{n_p}$ (where n_p=record length at the evaluation-set estimation site in years) were binned into several groups, so that all sites within any bin would have similar values of record length. Then, instead of computing RMSE over the full evaluation data set, we computed it over each of the subsets corresponding to these bins. Next, we scatterplotted the bin RMSE value against the central value of $1/\sqrt{n_p}$ associated with each bin (Fig. 5). Fitted lines were then extrapolated to infinite n_p (i.e., to the RMSE axis) to obtain estimates of the model-error RMSE (Table 2). Evidently, time-sampling errors contribute substantially to the overall RMSE values. Approximate removal of time-sampling error suggests that GRoI and HRoI generate better regression models than PRoI. Because the technique introduced here for removing time-sampling errors is crude, we doubt

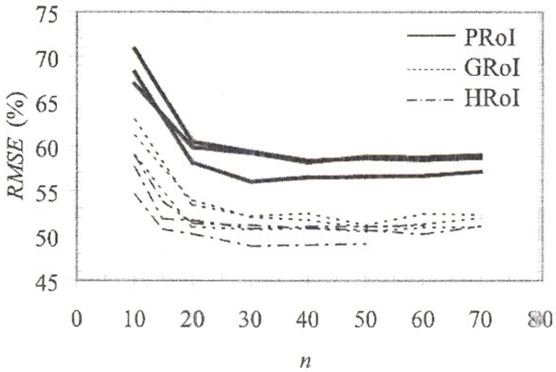

Fig. 4. Optimization RMSE (%), as a function of the number of sites contained in the region of influence (RoI), n. For the HRoI approach, plotted values are the minimum over all the target maximum RoI distances, D. The three curves for each approach correspond to the three optimization subsets.

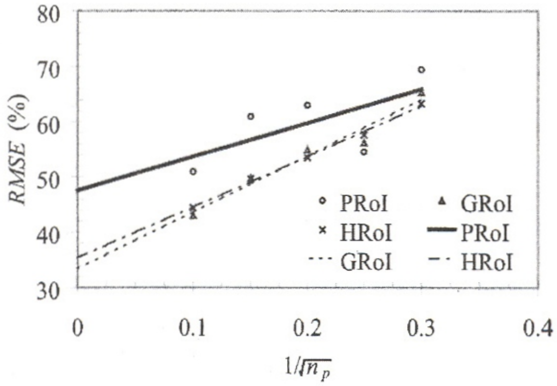

Fig. 5. Evaluation RMSE (%) as a function of the inverse of the square root of the record length at the evaluation site, $1/\sqrt{n_p}$. Data were binned by $1/\sqrt{n_p}$ values, with bins centered at 0.1, 0.15, 0.2, 0.25, and 0.3; the numbers of sites in each bin are 244, 362, 273, 107, and 105, respectively.

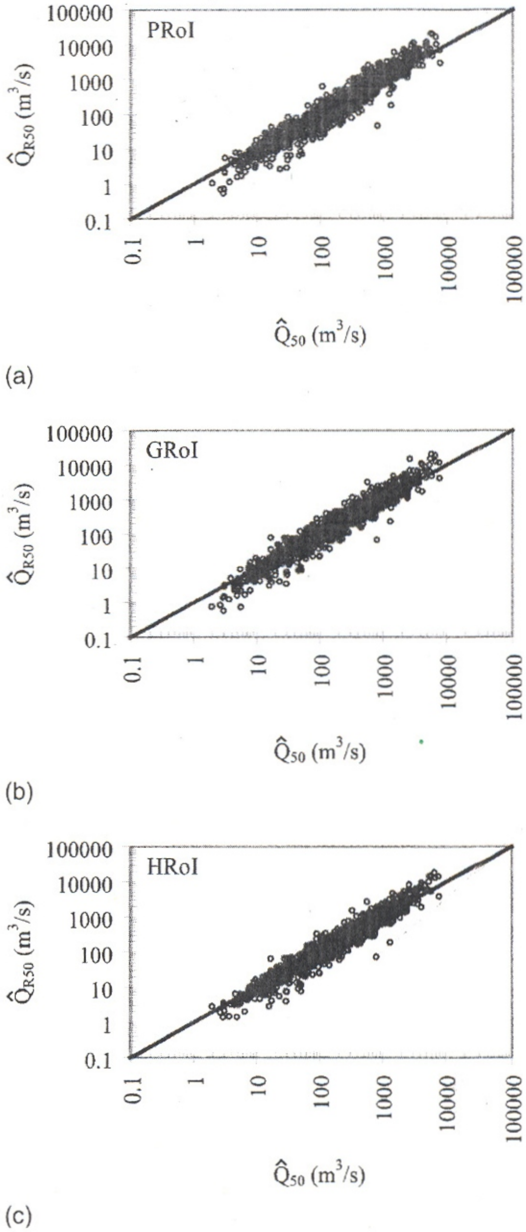

Fig. 6. Comparison of at-site $\hat{Q}_{50}$ values to the RoI-modeled estimates of $\hat{Q}_{50}$, $\hat{Q}_{R50}$. The solid lines are the 1:1 line. Approaches: (a) PRoI; (b) GRoI; and (c) HRoI. For each approach, all three evaluation subsets are plotted.

that the small difference in estimated model-error RMSE between GRoI and HRoI is statistically significant.

Because values of RMSE might not fully characterize the distribution of errors (i.e., total errors, υ), we report in Tables 1 and 2 the number of residuals with extreme values (absolute value more than twice the RMSE) for each evaluation split-sampled set. By this measure of worst-case performance, the HRoI approach is superior to the other two approaches.

Scatter plots of model predictions against at-site estimates from flow time series are presented in Fig. 6. It is difficult to distinguish differences in typical behavior of errors across methods from these plots, but the plots are useful for visually comparing the larger model errors. For all three methods, the model predictions tend to be positively biased for large flows and negatively biased for small flows. This tendency is reduced, however, with the HRoI method in comparison to the PRoI and GRoI methods. This observation is consistent with the outlier counts in Table 2.

Discussion

The optimal value of D in HRoI can be used as an indicator of the need to account for geographic proximity in the definition of a region of influence. Our optimized value of 250 km is consistent with the findings of Eng et al. (2005) that the PRoI approach worked best when applied to a region about the size of a medium-size state, such as Georgia. The need to limit geographic extent of the region indicates the importance of variables that are not included among the predictors. Presumably such neglected variables exhibit spatial correlation on a length scale represented by D.

Our findings confirm those of Merz and Blöschl (2005), who demonstrated the value of considering both predictor-variable and geographic proximity when developing regional statistical models of flood characteristics. An interesting question for future consideration is how the HRoI parameter D might be related to the length scale of the variogram employed in geostatistical studies.

The relatively poor performance of the PRoI approach in this study presents a serious challenge to this approach. In fairness, it should be noted that certain features of our analysis may have put PRoI at a disadvantage. For example, only the crudest measures of climate (mean annual precipitation and mean minimum January air temperature) were included as candidate predictors; additional predictive power could possibly be obtained by using such variables as mean precipitation in individual months, storm interarrival rate (Milly 1994), and potential evaporation. Two sites with same mean annual precipitation but very different temporal distributions of precipitation would naturally differ in their flood characteristics.

The PRoI results also may have been negatively affected by the arbitrary scaling (by standard deviations) of the various predictors in the definition of a scalar distance in predictor-variable space. Although this is the standard practice for the PRoI method, it has not been shown to provide the optimal scaling. It is not unreasonable to suggest that better performance could be obtained by use of scale factors either obtained by optimization or derived from physical considerations. Consideration of nonlinearities in predictor-predictand relations may lead to improved definitions of distance in predictor-variable space. Suppose, for example, that the predictand depends linearly on one predictor over its entire range of values, but depends nonlinearly on the other predictors. In that case, differences in the value of the first predictor would not indicate hydrologic dissimilarity and so should not be included in the definition of R_j.

The potential of the HRoI approach will be limited to some degree by the same factors that limit the PRoI approach. Any improvements in the selection and scaling of predictor variables will benefit the performance not only of PRoI, but also of HRoI.

We have shown that the contribution of temporal sampling error to our performance metric (the RMSE), which also serves as the objective function for parameter optimization, is substantial. The RMSE metric is commonly used in regionalization studies (e.g., Tasker et al. 1996) and will tend to overstate the magnitude of model errors as shown in the results of this study. The RMSE metric will also distort any optimization based on it, because sites with short records are weighted just as heavily as sites with long records. Consideration of this effect in the regression process (e.g., by weighting for differing temporal sampling errors) would probably yield more accurate models.

Conclusions

We have described and evaluated a hybrid approach to defining a region of influence for hydrologic regression. This approach considers both geographic proximity and proximity in predictor-variable space. We evaluated the proposed approach by application to estimation of the 50-year flood discharge. Our evaluation, based on the root-mean-square errors of estimation and number of large (outlier) errors, indicates that the performance of the hybrid approach is superior to that of less general approaches.

Acknowledgments

The writers wish to thank Wladmir Guimaraes, U.S. Geological Survey, Columbia S.C., for providing the estimates of 50-year flood peaks and basin characteristics.

References

Aitchison, J., and Brown, J. A. C. (1957). *The lognormal distribution*, Cambridge University Press, Cambridge, Mass.

Berenbrock, C. (2002). "Estimating the magnitude of peak flows at selected recurrence intervals for streams in Idaho." *Water-Resources Investigations Rep. No. 02-4170*, U.S. Geological Survey, ⟨http//pubs.er.usgs.gov/usgspubs/wri/wri20024170⟩.

Burn, D. H. (1990). "Evaluation of regional flood frequency analysis with a region of influence approach." *Water Resour. Res.*, 26(10), 2257–2265.

Eng, K., Tasker, G. D., and Milly, P. C. D. (2005). "An analysis of region-of-influence methods for flood regionalization in the Gulf-Atlantic Rolling Plains." *J. Am. Water Resour. Assoc.*, 41(1), 135–143.

Feaster, T. D., and Tasker, G. D. (2002). "Techniques for estimating the magnitude and frequency of floods in rural basins of South Carolina, 1999." *Water-Resources Investigations Rep. No. 02-4140*, U.S. Geological Survey, ⟨http//pubs.er.usgs.gov/usgspubs/wri/wri024140⟩.

Hammond, E. H. (1964). "Analysis of properties in land form geography: An application to broad-scale land form mapping." *Ann. Assoc. Am. Geogr.*, 54, 11–23.

Hydrology Subcommittee of the Interagency Advisory Committee on Water Data. (1982). *Guidelines for determining flood flow frequency bulletin 17B of the hydrology subcommittee*, Office of Water Data Coordination, U.S. Geological Survey, Reston, Va.

Law, G. S., and Tasker, G. D. (2003). "Flood-frequency prediction methods for unregulated streams of Tennessee, 2000." *Water-Resources Investigations Rep. No. 03-4176*, U.S. Geological Survey, ⟨http//pubs.er.usgs.gov/usgspubs/wri/wri034176⟩.

Mallows, C. L. (1973). "Some comments on C_p." *Technometrics*, 15(4), 661–675.

Mallows, C. L. (1995). "More comments on C_p." *Technometrics*, 37(4), 362–372.

Merz, R., and Blöschl, G. (2005). "Flood frequency regionalisation—Spatial proximity vs. catchment attributes." *J. Hydrol.*, 302(4), 283–306.

Milly, P. C. D. (1994). "Climate, soil-water storage, and the average annual water balance." *Water Resour. Res.*, 30(7), 2143–2156.

Pope, B. F., Tasker, G. D., and Robbins, J. C. (2001). "Estimating the magnitude and frequency of floods in rural basins of North Carolina—Revised." *Water-Resources Investigations Rep. No. 01-4207*, U.S. Geological Survey, ⟨http//pubs.er.usgs.gov/usgspubs/wri/wri014207⟩.

Stedinger, J. R., and Tasker, G. D. (1985). "Regional hydrologic analysis. 1: Ordinary, weighted, and generalized least squares compared." *Water Resour. Res.*, 21(9), 1421–1432.

Tasker, G. D., Hodge, S. A., and Barks, C. S. (1996). "Region of influence regression for estimating the 50-year flood at ungauged sites." *Water Resour. Bull.*, 32(1), 163–170.

Tasker, G. D., and Slade, R. M., Jr. (1994). "An interactive regional regression approach to estimating flood quantiles." *ASCE Proc. of the 21st Annual Conf. of the Water Resources Planning and Management Division*, D. G. Fontane and H. N. Tuvel, eds., 782–785.

Tasker, G. D., and Stedinger, J. R. (1989). "An operational GLS model for hydrologic regression." *J. Hydrol.*, 111, 361–375.

U.S. Department of Commerce, National Oceanic and Atmospheric Administration. (1976–1978). "Climates of the United States." *Climatology of the United States*, Washington, D.C., No. 60, Parts 1–52.