Regionalization of streamflow characteristics for the Gulf-Atlantic Rolling Plains using leverage-guided region-of-influence regression.

Ken Eng[1], Jery R. Stedinger[2], and Andrea M. Gruber[3]

[1] Research Hydrologist, National Research Program, U.S. Geological Survey, 12201 Sunrise Valley Drive, Mail Stop 430, Reston, VA 20192; PH (703) 648-5843. FAX (703) 648-5484; Email: keng@usgs.gov
[2] Professor, School of Civil & Environmental Engineering, Cornell University, Hollister Hall, Ithaca, NY 14853-3501 USA; PH (607) 255-2351; Email: jrs5@cornell.edu.
[3] Graduate Research Assistant, School of Civil and Environmental Engineering, Cornell University, 220 Hollister Hall, Ithaca, NY 14853-3501; Email: amg66@cornell.edu.

Abstract. Multivariate regression models are often applied to hydrologic regions to estimate peak flows at ungauged basins in an area. Such regression models can be derived using a region of influence (RoI), which is a selected set of basins that are hydrologically similar to the ungauged basin for which peak flow estimates are required. These regions can be poorly defined resulting in unstable parameter estimates because observations at a few basins may disproportionately influence the parameters. Conventional treatment is to drop the basin, if the problem is even recognized. We propose a leverage-guided RoI regression approach that redefines the region of influence. This new procedure uses two newly defined RoI leverage and influence metrics. The proposed approach is applied to 996 streamflow gauging stations in the southeast United States to estimate the 50-year peak flow. The new leverage-guided RoI regression approach resulted in lower root-mean-square estimation errors, produced fewer observations with large leverage, and eliminated all influential observations.

1. Introduction

Hydrologists, engineers, state and local agencies, and the general public often require information on peak streamflow at locations where there are no streamflow-gauging stations (henceforth referred to as gauges). For example, these estimates are often used for flood insurance mapping to assess risk. Peak-streamflow characteristics at such ungauged basins are often inferred from flood flow records at similar, nearby gauges. A method to calculate peak-streamflow characteristics for ungauged basins 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 streamflow.

Regional regression models are applied to regions that are often defined by physiographic boundaries or from residuals from an overall regression (*Wandle*, 1977). Regions can also be defined by a collection of watersheds whose basin characteristics are similar, by some overall measure, to those at the ungauged site of interest. This "region-of-influence" (RoI) approach defines a unique region for each ungauged basin (*Burn*, 1990). The RoI can also be defined as a collection of the geographically closest basins.

For the different approaches to define regions, such as predictor-variable proximity and physiographic boundaries, there can be basins in these regions whose characteristics have an unusually large impact ("influence") on the estimation of the regression model parameters. Conventional treatment of these types of basins is to either to ignore the problems with the analysis or to remove the troublesome basins from the analysis, so that they do not unduly affect the parameter estimates of the regional regression models. With the RoI regression approach, influential observations have generally been ignored and left in the analysis (e.g., *Eng et al.*, 2005).

Unusually influential observations can occur due to construction of a poorly defined region, perhaps containing too few basins to support a statistical analysis or if the region is in some sense not representative. An alternative to the practice of removing or ignoring influential observations is to redefine the region.

The objectives for this study are to (1) develop new and specialized RoI leverage and influence metrics for GLS regression, (2) to use these metrics in a proposed leverage-guided RoI regression approach to redefine the region of application, and (3) to evaluate the proposed approach when used to estimate the 50-year peak flow characteristic, $\hat{Q}_{50}$. Our study employs 996 continuous gauging stations in the southeast United States.

2. Data and Study Area

For this analysis 996 streamflow gauges were selected because they are contained in a single physiographic region, the Gulf-Atlantic Rolling Plains (*Hammond*, 1964) (Figure 1). The record lengths at these gauges range from 10 to 103 years. Drainage areas range from 0.13 to 2,564 km^2, with a median of 402 km^2.

Eight basin characteristics are available for all gauged basins: drainage area, A_d, 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. Using the entire data set, the most statistically significant basin characteristics are A_d, S, and P determined by a best subsets selection made on the basis of the *Mallows* C_p statistic (*Mallows*, 1995; *Eng et al.*, 2005).

Because $\hat{Q}_{50}$ is to be estimated by regression against A_d, S, and P the analysis requires estimates of these variables for all the basins. Estimates of $\hat{Q}_{50}$ are derived from peak streamflow data obtained from the USGS National Water Information System (NWISWeb, http://nwis.waterdata.usgs.gov/usa/nwis/), which also provides A_d values. The $\hat{Q}_{50}$ values are 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 choose to examine a single return period, the 50-year return period, because it lies within the range commonly used in hydrologic analyses.

Isohyetal maps (*U.S. Department of Commerce*, 1976-1978) are used to obtain P. The values of S are calculated as the average channel slope (elevation difference divided by distance along the main channel) between points located 10 and 85 percent of the distance from the gaging station to the basin divide.

3. Regional Regression Models of Peak-Flow Characteristics

3.1. Generalized Least Squares (GLS) Parameter Fit

We consider log-linear regression models of the form,

$$\log(Q_{50}) = \beta_0 + \beta_{A_d}\log(A_d) + \beta_S\log(S) + \beta_p\log(P) + \delta \qquad (1)$$

where Q_{50} is the 50-year peak flow, β_0, β_{A_d}, β_S, and β_P are constants, and δ is model error, with mean zero and variance σ_δ^2.

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

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

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. This means that values of η for different sites are cross-correlated. Substituting (2) into (1) yields

$$\log(\hat{Q}_{50}) = \beta_0 + \beta_{A_d}\log(A_d) + \beta_S\log(S) + \beta_p\log(P) + \varepsilon, \qquad (3)$$

where $\varepsilon = \delta + \eta$ is the sum of the model and the time sampling errors.

A GLS parameter estimation technique is used to perform the regression in the presence of cross-correlation of η, following the assumption that model error ε is not spatially correlated (*Stedinger and Tasker*, 1985). Estimates of β_0, β_{A_d}, β_S, and β_P are $\hat{\beta}_0$, $\hat{\beta}_{A_d}$, $\hat{\beta}_S$, and $\hat{\beta}_P$, respectively. The GLS estimator $\hat{\boldsymbol{\beta}}$ of the parameter vector is

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\hat{\boldsymbol{\Lambda}}^{-1}\mathbf{X}^T\right)^{-1}\mathbf{X}^T\hat{\boldsymbol{\Lambda}}^{-1}\hat{\mathbf{Y}}, \qquad (4)$$

where $\mathbf{X}$ is a (J x 4) matrix of log(A_d), log(S), and log(P) values at J sites in the region of influence, augmented by a column of ones, J is the number of gauged basins in the region of influence, $\hat{\mathbf{Y}}$ is a (J x 1) vector of $\log(\hat{Q}_{50})$ values, and $\hat{\boldsymbol{\Lambda}}$ is a matrix containing the estimates of the covariance of ε across basins in the selected region of influence. The main diagonal elements of $\hat{\boldsymbol{\Lambda}}$ thus include a part associated with δ, and all elements include the effect of η. Following *Tasker and Stedinger* (1989), $\hat{\boldsymbol{\Lambda}}$ is given as

$$\hat{\Lambda}_{pq} = \begin{cases} \sigma_\delta^2 + \dfrac{\hat{s}_p^2\left[1 + K_p\hat{g}_p + 0.5K_p^2\left(1 + 0.75\hat{g}_p^2\right)\right]}{m_p} & (p = q) \\ \dfrac{\hat{r}_{pq}\hat{s}_p\hat{s}_q m_{pq}\left[1 + 0.5K_p\hat{g}_p - 0.5K_q\hat{g}_q + 0.5K_pK_q\left(\hat{r}_{pq} + 0.75\hat{g}_p\hat{g}_q\right)\right]}{m_p m_q} & (p \neq q) \end{cases},$$

(5)

where the subscripts p and q are indices of gauged basins in the region of influence, K_p and K_q are the Log-Pearson Type III distribution standard deviate for basins p and

q, $\hat{g}_p$ and $\hat{g}_q$ are the skewness coefficients for basins p and q determined by weighted least squares regression on basin attributes as described by *Tasker and Stedinger* (1986), m_p and m_q are basin specific record lengths, m_{pq} is the concurrent record length for basins p and q, $\hat{s}_p$ and $\hat{s}_q$ are estimates of the standard deviation of annual peaks estimated by methods by *Tasker and Stedinger* (1989), and r_{pq} is the sample cross-correlation of annual peaks at basins p and q estimated by methods in *Tasker and Stedinger* (1989) and coefficients from *Pope et al.* (2001). Equation (5) neglects the error in weighted skewness estimators, which is reasonable if the regional skewness estimator is relative precise (*Griffis and Stedinger*, 2006).

3.2. Region of Influence

The regression model parameter estimates are calculated by GLS regression using data for the gauged basins within a RoI for the ungauged ('estimation') basin. A RoI is formed in two different ways: predictor-variable (PRoI) and geographic space (GRoI). A typical RoI consists of n gauged basins closest in either space. The value of n is determined by methods described in section 3.3. The predictor-variable space proximity of an ungauged basin to a gauged basin, j, is defined as

$$R_j=\left[\left(\frac{\log(A_d)-\log(A_d)_j}{\sigma_{\log(A_d)}}\right)^2+\left(\frac{\log(S)-\log(S)_j}{\sigma_{\log(S)}}\right)^2+\left(\frac{\log(P)-\log(P)_j}{\sigma_{\log(P)}}\right)^2\right]^{1/2}, \quad (6)$$

where $\sigma_{\log(A_d)}$, $\sigma_{\log(S)}$, and $\sigma_{\log(P)}$ are the sample standard deviation of log(A_d), log(S), and log(P), respectively (computed from data from the entire study region). The geographic space proximity is simply measured by the geographic distance between the ungauged basin and each gauged basin.

3.3. Performance Metrics

The root-mean-square error (*RMSE*) of estimation of $\hat{Q}_{50}$ is used to evaluate the precision of the RoI regression procedures. In percentage terms (*Aitchison and Brown*, 1957; modified for use of common logarithms),

$$RMSE = 100\left\{e^{\left[(\ln 10)^2\varepsilon_{obs}^2\right]}-1\right\}^{1/2}, \quad (7)$$

where ε_{obs}^2 is the observed mean squared error,

$$\varepsilon_{obs}^2=\frac{1}{N}\sum_{i=1}^{N}\left[\log\left(\hat{Q}_{50}\right)_i-\log\left(\hat{Q}_{R50}\right)_i\right]^2, \quad (8)$$

where $\hat{Q}_{50}$ is the estimate of Q_{50} computed from the observed annual series of peak flows, $\hat{Q}_{R50}$ is the GLS-regressed estimate of $\hat{Q}_{50}$, and N is the total number of gauged basins for which predictions are made as if they were ungauged.

The location in predictor-variable space of the basin attributes of a gauged basin relative to those from the other basins is important. If the attributes of one basin are relatively far away from the centroid of the other basins in a region, then this unusual basin may significantly change or influence the parameter values obtained in a RoI regression. The leverage metric measures how far away observations are from the centroid of the characteristics of other basins. For conventional multiple linear regression using GLS, *Tasker and Stedinger* (1989) defined leverage for the i^{th} site as

$$h_{ii} = \left[\mathbf{X}\left(\mathbf{X}^T \boldsymbol{\Lambda}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T \boldsymbol{\Lambda}^{-1} \right]_{ii} . \tag{9}$$

We propose a more suitable form of (9) for RoI regression models using GLS, which is

$$\mathbf{h}_0^T = \left[\mathbf{x}_0\left(\mathbf{X}^T \boldsymbol{\Lambda}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T \boldsymbol{\Lambda}^{-1} \right], \tag{10}$$

where $\mathbf{x}_0$ is a vector of basin attributes at the ungauged basin, $\mathbf{h}^T{}_{0}$ is a vector containing the leverage values for each site for the RoI regression model for site 0. A basin potentially has large influence if its leverage exceeds the criteria given by

$$h_{\text{limit}} = \frac{C_h}{J}\sum_{j=1}^{N_{RoI}} h_{o,j} , \tag{11}$$

where C_h is a constant. For conventional multiple linear regression, C_h is equal to 2 in (11) and reflects the observation that values twice the average can be considered as unusually large. Multipliers of 2, 4 and 8 are tested in the numerator of (11) to determine the number of RoI regression models that have at least one large leverage point (Table 1).

For leverage-guided RoI regression, a C_h value equal to 2 was found to be too small resulting in the addition of too many basins to a region, and thus roughly 90% and 70% of all GRoI and PRoI regression models, respectively, are identified as having at least one large leverage point. Conversely, when C_h is equal to 8, roughly 15% and 2.5% of regression models for GRoI and PRoI, respectively, are identified as large leverage. A C_h value of 4 is chosen for the remainder of this study since it results in a moderate amount (20% and 45% for PRoI and GRoI, respectively) of the regressions as having at least one large leverage point.

The leverage metric only indicates if an observation is unusual from the others in predictor-variable space. Such unusual observations may or may not have any significant impact on the derived parameters in (3). The influence metric, such as Cook's D (*Cook*, 1977), indicates if an unusual observation had large influence over the parameter values. A Cook's D for multiple linear regression using GLS is given by (*Tasker and Stedinger*, 1989)

$$D_j = \frac{\hat{\varepsilon}_{r,j}^2 K_{jj}}{p\left(\Lambda_{jj} - K_{jj}\right)^2} , \tag{12}$$

where p is the dimension of β, Λ_{jj} is the j^{th} main diagonal of the Λ cross-correlation matrix, K_{jj} is the j^{th} main diagonal of $\mathbf{X}(\mathbf{X}^T\Lambda^{-1}\mathbf{X})^{-1}\mathbf{X}^T$, and $\hat{\varepsilon}_{r,j}$ is the j^{th} residual. For RoI, we propose the leverage statistic

$$D_{0j} = \left[\left(\left| \frac{h_{0j}}{H_{jj}} \right| \right) \frac{\hat{\varepsilon}_{r,j}^2 K_{jj}}{p\left(\Lambda_{jj} - K_{jj}\right)^2} \right], \tag{13}$$

where h_{0j} is the j^{th} component of (10), and H_{jj} is the j^{th} main diagonal of $\mathbf{H}=\mathbf{X}(\mathbf{X}^{T}\boldsymbol{\Lambda}^{-1}\mathbf{X})^{-1}\mathbf{X}^{T}\boldsymbol{\Lambda}^{-1}$. Identification of an observation that has caused large influence is if it exceeds the limit given by

$$D_{\text{limit}}=\left(\frac{C_D}{N_{RoI}}\right). \tag{14}$$

A C_D value of 4 is used for conventional multiple linear regression. Similar to the h_{limit}, the value of 4 may be inappropriate for RoI regression, so values of 4, 8, and 16 in the numerator of (14) are examined.

In a typical application of RoI regression, an optimal value of n would be determined by performing the RoI analyses for various fixed values of n (e.g., *Tasker et al.*, 1996). The optimal n value associated with the minimum *RMSE* value. In a previous study by *Eng et al.* (2005), the range of optimal n values for GRoI and PRoI were 10 to 20. In this study, we examined n =10, 15, and 20. As discussed in section 3.4, the leverage-guided region-of-influence approach in theory is not sensitive to the choice of n.

Table 1. Percent of RoI regression models that have at least one large leverage point.

Approach	n	Percentage of RoI regressions		
		C_h=2 & C_D=4	C_h=4 & C_D=8	C_h=8 & C_D=16
GRoI	10	90.7	47.5	13.6
	15	90.1	42.3	15.7
	20	88.0	43.2	15.3
PRoI	10	66.6	17.3	2.5
	15	72.1	18.3	2.5
	20	71.4	20.5	2.5

3.4. Leverage-Guided Region-of-Influence Regression Approach

As noted in the Introduction, potentially influential observations have often been ignored in RoI regression studies. Perhaps because the idea of RoI-regression is to identify regions of similar basins, one hopes that none will be unusual. However, the exercise of trying to select only basins that are similar creates an opportunity of a few basins that would not have been unusual if all basins were included in the analysis. As an alternative to ignoring or removing influential observations, here they are retained and new basins are strategically introduced to revise the region of influence so that it is better balanced.

For the leverage-guided region-of-influence approach, the first step is to conduct a conventional RoI analysis with a selected value of n, and if at least one large leverage basin using (10) a new basin is added. To add new observations to the RoI, the predictor-variable space is used. We focus on the two most significant basin attribute, A_d and S, and determine their maximum and minimum values. Only two predictor-variables are chosen to allow easy interpretation of the predictor-variable space by a two-dimensional graph. Basins whose A_d and S values fall within these ranges are added to the RoI one at a time, and the regression parameters and leverage values are recomputed. On one hand, if the added observation increases the magnitude of the leverage value(s) of any of the original potentially influential observation(s), it is removed from the RoI and the next candidate observation is added. On the other hand, if the added site decreases the magnitude of the leverage values then the new observation is retained, and another new observation is added. This process continues until any one of the three conditions is satisfied: (1) the number of new observations retained equals the original number of observations in the RoI, (2) all leverage values in the redefined RoI are less than the limit given by (11), or (3) the original variance of the leverage values in the RoI is reduced by nine tenths by the addition of the new ones.

RoI regressions were developed for all 996 basins in the database. The entire process for developing a RoI was repeated with all of the RoI regression models that have large leverage observations. We will focus on the comparison of leverage-guided RoI regression models that satisfy the second and third conditions mentioned previously to conventional RoI regression that use regions that are not redefined. Leverage-guided RoI regression models satisfying the first condition are incomplete, so they are not used for comparison. A value of this new procedure is that it identifies situations where RoI regression may be unstable, and there does not appear to be a simple solution.

Suitability of using a regression model to estimate streamflow characteristics at an ungaged location can be assessed by analysis of the two-dimensional graphs of the two most significant predictor variables. The collection of basins do not support a regression analysis if the basin attributes of the ungauged basin are far away from those of gauged basins in predictor-variable space.

4. Results

Figure 1 shows the locations of the 996 gauges used in the study. Figure 2 illustrates the basins selected by the leverage-guided GRoI and PRoI approaches. Both approaches start with the 10 closest gauged basins, as each defines close, to form their initial RoI. For the leverage-guided GRoI approach, the addition of two new observations reduces all the leverage values to be less than the h_{limit}. In addition, the variance of the leverage values reduces from 0.91 to 0.01, and the estimated *RMSE* computed using (7) decreases from 226% to 59%. For the leverage-guided PRoI approach, the addition of three observations did not reduce all leverage values to be less than h_{limit}. The variance of the leverage values reduces from 0.12 to 0.08, and the estimated *RMSE* from 48% to 22%.

The performances over the leverage-guided PRoI and GRoI regression approaches are summarized in Table 2. For GRoI, the proposed approach successfully rebalanced anywhere from 14 to 27% of the models that had at least one large leverage observation. Unlike GRoI, PRoI rebalanced 35 to 38% of the regression models. The conventional *RMSE* values for the GRoI regressions ranged from 63 to 120%. With leverage-guided GRoI, the *RMSE* values ranged from 50 to 53%. This reduction is consistent across different starting n values supporting the assumption that the leverage-guided GRoI regression is relatively insensitive to these values. For conventional PRoI, the *RMSE* values range from 75 to 187%, and the leverage-guided PRoI approach reduced this range from 72 to 140%. The leverage-guided PRoI is not as effective as the GRoI one for redefining regions that have at least one large leverage observation. This result is not a surprise since the range of predictor-variable values is significantly smaller from conventional PRoI approach than those from a GRoI analysis.

The numbers of large leverage and influence observations from conventional and leverage-guided RoI regression approaches are reported in Table 2. The leverage-guided GRoI regression approach roughly reduced the total number of large leverage observations on average by 67%. The average reduction by the leverage-guided PRoI approach is 5%. The leverage-guided RoI approach is much less effective for reducing large leverage observations for a RoI that is formed of basin closest in predictor-variable space. Both approaches, however, reduced the total number of large influence observations to zero.

Figure 3 illustrates the analysis of the two-dimensional graph of the two most significant predictor variables at station 02077210, Kilgore Tributary near Leasburg, North Carolina. The 20 geographically closest basins were initially chosen to form the RoI. The A_d and S values of the ungauged basin by themselves may fall within the range of values of the gauged basins, but their combination is unlike the others in the RoI. From further examination of the other 995 basins, there are no basins that are comparable in size and have a small slope. In this case and 6 others, applications of regression models to estimate peak characteristics were not pursued.

Table 2. Summary of performance of conventional GRoI and PRoI regression with leverage-guided RoI results with C_h equal to 4 and C_D equal to 8.

Approach	*n*	No. of Regressions	Conventional RoI			Leverage-Guided RoI		
			RMSE (%)	No. of Large Leverage Points	No. of Influential Points	*RMSE* (%)	No. of Large Leverage Points	No. of Influential Points
GRoI	10	126	120.3	263	18	50.5	96	0
	15	85	63.12	139	13	52.6	50	0
	20	62	86.7	78	8	52.2	23	0
PRoI	10	60	186.7	147	12	140.5	139	0
	15	66	82.8	177	8	77.2	169	0
	20	78	74.8	246	5	71.5	236	0

5. Discussion

The conventional strategy to address potentially influential observations in RoI regression models is to ignore them and leave them in the analysis. We propose an approach that retains potentially influential observations and rebalances the regression model by redefining the region of application, thus, maximizing the data that can be used for an analysis.

In this study, the parameters of each RoI regression model are determined by the local basin attributes in the RoI. These locally determined parameters could vary greatly in value and sign. To address these problems, a set of parameters for the basin attributes determined by the global data set could be used instead of the local ones.

The global parameter values could be used in every RoI regression, while the constant could be determined locally. This approach would solve the problem of nonphysical signs occurring and potentially reduce the root-mean-square estimation errors.

Although this study focused on RoI regression approaches, the procedures outlined can be readily applied to other types of regression approaches estimating other statistics in addition to flow characteristics.

References

Aitchison, J., and Brown, J. A. C. (1957). *The Lognormal Distribution*, Cambridge University Press, Cambridge, Massachusetts, 176 pp.

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

Cook, R. D. (1977). "Detection of influential observation in linear regression." *Technometrics*, 19, 15-18.

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.

Griffis, V. W., and Stedinger J. R. "The Use of GLS Regression in Regional Hydrologic Analyses." manuscript, Cornell University, July 2006.

Hammond, E. H. (1964). "Analysis of properties in land form geography: an application to broad-scale land form mapping." *Annals Assoc. American Geophys.*, 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, Virginia, 99 pp.

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

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." *U.S. Geological Survey Water-Resources Investigations Report 01-4207*, 44 pp.

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., and Stedinger, J. R. (1986). "Regional skew with weighted LS regression." *J. Water Resour. Plann. and Manage.*, 112(2), 225-237.

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*, no. 60, parts 1-52.

Wandle, S. W. (1977). "Estimating the magnitude frequency of floods on natural-flow streams in Massachusetts." *U.S. Geol. Surv. Water-Resour. Invest. Rep., 77-39*, 27 pp.

Figures

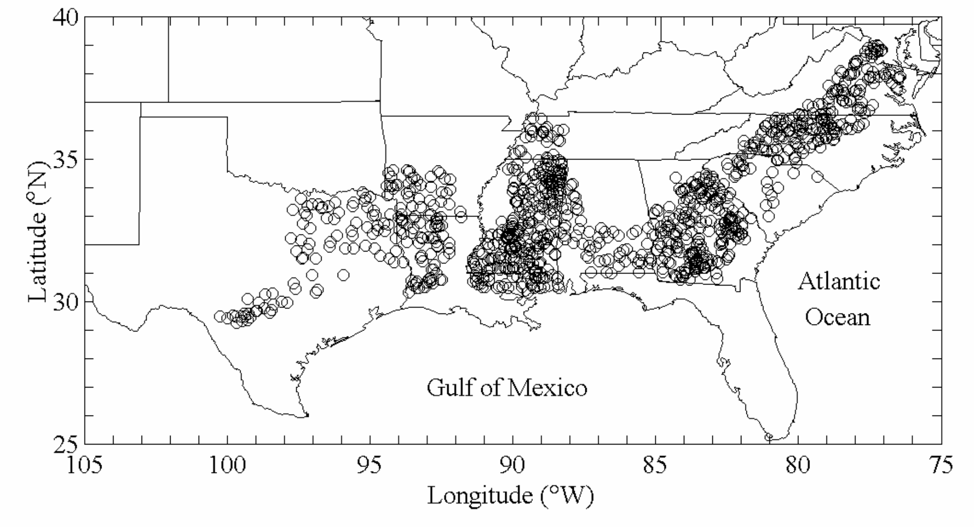

Figure 1. Southeastern United States. Circles represent 996 gauged basins.

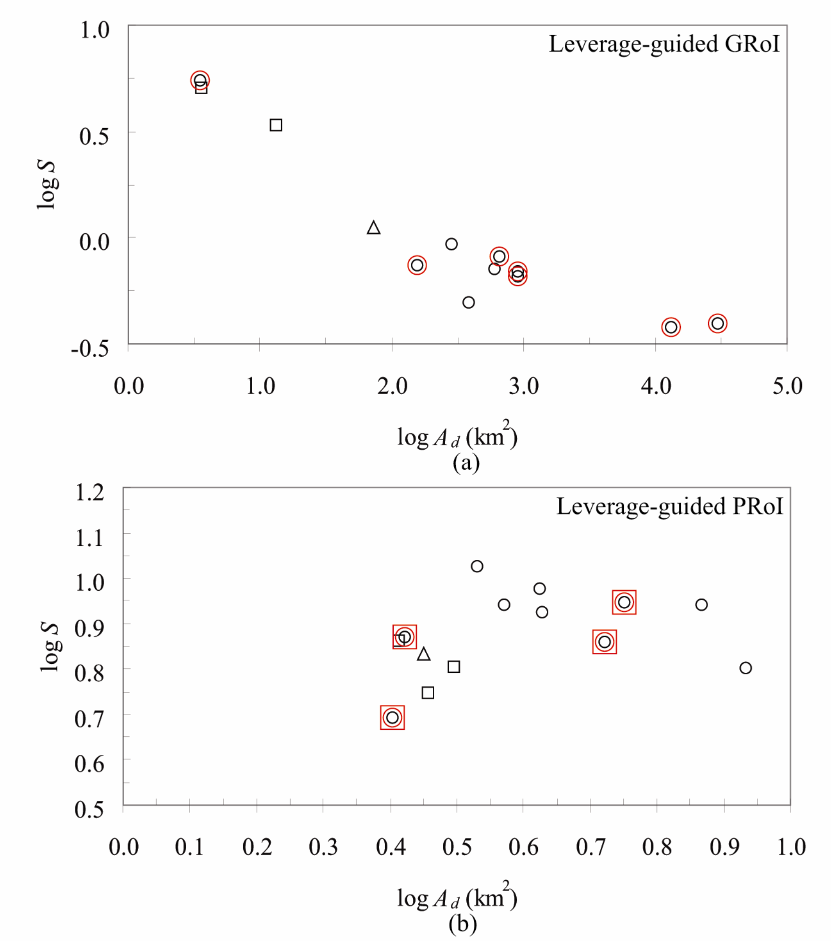

Figure 2. Predictor-variable space of gauged basins forming the RoI. (a) Leverage-guided GRoI approach (*n*=10, U.S. Geological Survey Station No. 02226700, Whitehead Creek near Denton, GA), and (b) leverage-guided PRoI approach (*n*=10,

U.S. Geological Survey Station No. 08031100, Bethlehem Branch near Van, TX). The black triangles represent the basin attributes at the ungauged basin, the black circles represent the basin attributes of the gauged basins in the unmodified RoI, the black squares represent the gauged basins added to the RoI, the red circles represent observations that have large leverage values in the unmodified RoI, and the red squares represent the observations that have large leverage values in the redefined RoI. Axis scales are very different in (a) and (b) reflecting points selected.

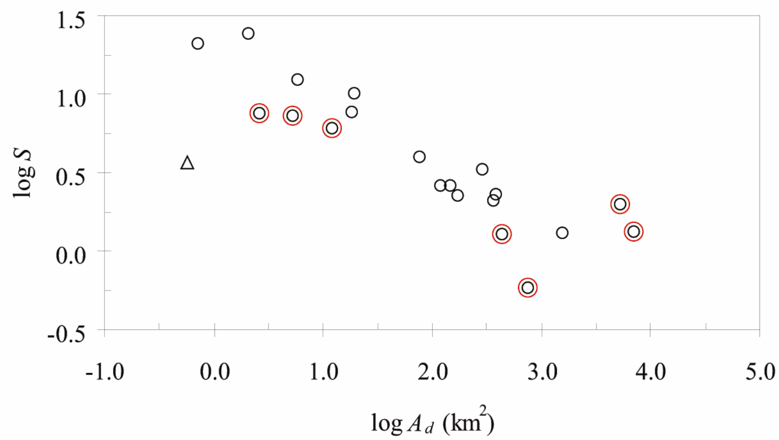

Figure 3. Predictor-variable space of gauged basins forming the GRoI (*n*=20, U.S. Geological Survey Station No. 02077210, Kilgore Tributary near Leasburg, NC). The black triangle represent the basin attributes at the ungauged basin, the black circles represent the basin attributes of the gauged basins in the unmodified RoI, and the red circles represent observations that have large leverage values in the unmodified RoI.