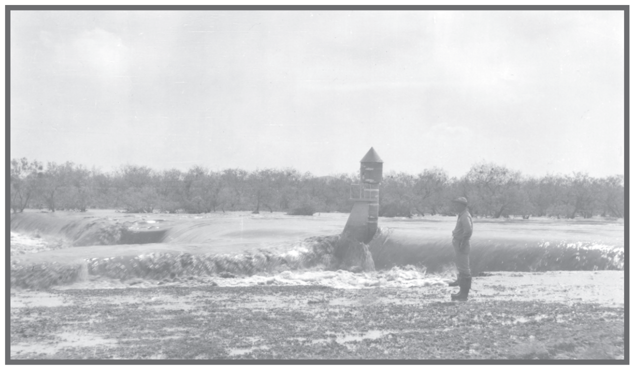
science for a changing world

In cooperation with the Texas Department of Transportation

Alternative Regression Equations for Estimation of Annual Peak-Streamflow Frequency for Undeveloped Watersheds in Texas using PRESS Minimization



Scientific Investigations Report 2008–5084
(Texas Department of Transportation Research Report 0–5521–2)

U.S. Department of the Interior
U.S. Geological Survey

—Technical Report Documentation Page—

1. Report No. FHWA/TX–08/0–5521–2	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle Alternative Regression Equations for Estimation of Annual Peak-Streamflow Frequency for Undeveloped Watersheds in Texas using PRESS Minimization		5. Report Date June 2008
7. Authors(s) William H. Asquith and David B. Thompson		8. Performing Organization Report No. USGS SIR 2008–5084
9. Performing Organization Name and Address U.S. Geological Survey 8027 Exchange Drive Austin, Texas 78754		10. Work Unit No. (TRAIS) 11. Contract or Grant No. Project 0–5521
12. Sponsoring Agency Name and Address Texas Department of Transportation Research and Technology Implementation Office P.O. Box 5080 Austin, Texas 78731		13. Type of Report and Period Covered Technical report on research from 2007 to 2008 14. Sponsoring Agency Code
15. Supplementary Notes Project conducted in cooperation with the Texas Department of Transportation and the Federal Highway Administration		
16. Abstract The U.S. Geological Survey, in cooperation with the Texas Department of Transportation and in partnership with Texas Tech University, investigated a refinement of the regional regression method and developed alternative equations for estimation of peak-streamflow frequency for undeveloped watersheds in Texas. A common model for estimation of peak-streamflow frequency is based on the regional regression method. The current (2008) regional regression equations for 11 regions of Texas are based on $\log_{10}$ transformations of all regression variables (drainage area, main-channel slope, and watershed shape). Exclusive use of $\log_{10}$-transformation does not fully linearize the relations between the variables. As a result, some systematic bias remains in the current equations. The bias results in overestimation of peak streamflow for both the smallest and largest watersheds. The bias increases with increasing recurrence interval. The primary source of the bias is the discernible curvilinear relation in $\log_{10}$ space between peak streamflow and drainage area. Bias is demonstrated by selected residual plots with superimposed LOWESS trend lines. To address the bias, a statistical framework based on minimization of the PRESS statistic through power transformation of drainage area is described and implemented, and the resulting regression equations are reported. Compared to $\log_{10}$-exclusive equations, the equations derived from PRESS minimization have PRESS statistics and residual standard errors less than the $\log_{10}$-exclusive equations. Selected residual plots for the PRESS-minimized equations are presented to demonstrate that systematic bias in regional regression equations for peak-streamflow frequency estimation in Texas can be reduced. Because the overall error is similar to the error associated with previous equations and because the bias is reduced, the PRESS-minimized equations reported here provide alternative equations for peak-streamflow frequency estimation.		
17. Key Words Flood Frequency, Peak Streamflow, Texas	18. Distribution Statement No restrictions. This document is available to the public through the National Technical Information Service, Springfield Virginia, 22161, www.ntis.gov	
19. Security Classif. (of report) Unclassified	20. Security Classif. (of this page) Unclassified	21. No. of Pages 40 — 22. Price

Form DOT F 1700.7 (8-72) Reproduction of complete page authorized

Front cover:

U.S. Geological Survey station 08134000 North Concho River near Carlsbad, Tex., May 12, 1953. Photographs by J.D. Yost.

Top: Panorama from left bank looking directly across channel showing gage shelter and control (dam); flow about 7 feet over dam at gage height of about 9.7 feet.

Bottom: From left bank looking across channel showing extent of washout at the station.

Back cover:

U.S. Geological Survey station 08171000 Blanco River near Wimberley, Tex., June 28, 1928. Photographer unknown.

Left: Gage shelter and stilling well.

Right: Gage shelter at far left (to left of automobiles).

Alternative Regression Equations for Estimation of Annual Peak-Streamflow Frequency for Undeveloped Watersheds in Texas using PRESS Minimization

By William H. Asquith and David B. Thompson

In cooperation with the Texas Department of Transportation

Scientific Investigations Report 2008–5084
(Texas Department of Transportation Research Report 0–5521–2)

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
DIRK KEMPTHORNE, Secretary

U.S. Geological Survey
Mark D. Myers, Director

U.S. Geological Survey, Reston, Virginia: 2008

For product and ordering information:
World Wide Web: http://www.usgs.gov/pubprod
Telephone: 1-888-ASK-USGS

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment:
World Wide Web: http://www.usgs.gov
Telephone: 1-888-ASK-USGS

Suggested citation:
Asquith, W.H., and Thompson, D.B., 2008, Alternative regression equations for estimation of annual peak-streamflow frequency for undeveloped watersheds in Texas using PRESS minimization: U.S. Geological Survey Scientific Investigations Report 2008–5084, 40 p. [http://pubs.usgs.gov/sir/2008/5084]

Typeset on July 9, 2008

Contents

Figures

Tables

Conversion Factors

Inch/Pound to SI

Multiply	By	To obtain
	Length	
inch (in.)	25.4	millimeter (mm)
foot (ft)	.3048	meter (m)
mile (mi)	1.609	kilometer (km)
	Area	
square mile (mi^2)	2.590	square kilometer (km^2)
	Flow	
cubic feet per second (ft^3/s)	.02832	cubic meters per second (m^3/s)

SI to Inch/Pound

Multiply	By	To obtain
	Length	
millimeter (mm)	0.03937	inch (in.)
meter (m)	3.281	foot (ft)
	Area	
square kilometer (km^2)	.3861	square mile (mi^2)
	Flow	
cubic meters per second (m^3/s)	35.31	cubic feet per second (ft^3/s)

Alternative Regression Equations for Estimation of Annual Peak-Streamflow Frequency for Undeveloped Watersheds in Texas using PRESS Minimization

By William H. Asquith and David B. Thompson[1]

Abstract

The U.S. Geological Survey, in cooperation with the Texas Department of Transportation and in partnership with Texas Tech University, investigated a refinement of the regional regression method and developed alternative equations for estimation of peak-streamflow frequency for undeveloped watersheds in Texas. A common model for estimation of peak-streamflow frequency is based on the regional regression method. The current (2008) regional regression equations for 11 regions of Texas are based on $\log_{10}$ transformations of all regression variables (drainage area, main-channel slope, and watershed shape). Exclusive use of $\log_{10}$-transformation does not fully linearize the relations between the variables. As a result, some systematic bias remains in the current equations. The bias results in overestimation of peak streamflow for both the smallest and largest watersheds. The bias increases with increasing recurrence interval. The primary source of the bias is the discernible curvilinear relation in $\log_{10}$ space between peak streamflow and drainage area. Bias is demonstrated by selected residual plots with superimposed LOWESS trend lines. To address the bias, a statistical framework based on minimization of the PRESS statistic through power transformation of drainage area is described and implemented, and the resulting regression equations are reported. Compared to $\log_{10}$-exclusive equations, the equations derived from PRESS minimization have PRESS statistics and residual standard errors less than the $\log_{10}$-exclusive equations. Selected residual plots for the PRESS-minimized equations are presented to demonstrate that systematic bias in regional regression equations for peak-streamflow frequency estimation in Texas can be reduced. Because the overall error is similar to the error associated with previous equations and because the bias is reduced, the PRESS-minimized equations reported here provide alternative equations for peak-streamflow frequency estimation.

[1]Texas Tech University, Lubbock, Tex.

Introduction

Peak-streamflow frequency estimates are needed for flood-plain management; for objective assessment of flood risk; for cost-effective design of dams, levees, and other flood-control structures; and for design of roads, bridges, and culverts. Peak-streamflow frequency represents the peak streamflow for recurrence intervals of 2, 5, 10, 25, 50, and 100 years.

Beginning in 2003, the U.S. Geological Survey (USGS), in cooperation with the Texas Department of Transportation (Research Project 0–4405) and in partnership with Texas Tech University, began a 3-year investigation of the influence of hydrologic scale (represented by drainage area for this report) on hydrologic model performance (Asquith and Thompson, 2005; Thompson, 2006). Hydrologic models for estimation of design floods are in widespread use by TxDOT engineers and the broader hydrologic engineering community. A common model for estimation of peak-streamflow frequency is based on the regional regression method. This method is the subject of this report.

Bias exists in the regional regression equations for estimation of peak-streamflow frequency in Texas (Asquith and Slade, 1997), hereinafter referred to as AS1997. The source of the bias is the discernible curvilinear relation between peak streamflow and drainage area—the bias is graphically illustrated in this report. The current regional regression equations might overestimate peak-streamflow for both the smallest and largest watersheds represented in the AS1997 investigation. The bias is scale-dependent (depends on the size of the drainage area) and can be reduced.

Purpose and Scope

The primary purpose of this report, which parallels the discussion of Asquith and Thompson (2005), is to use an alternative statistical framework to develop regression equations with potentially less bias and therefore enhanced prediction capability—in particular, enhanced prediction capability for small watersheds. For this report, a small watershed is defined as having a contributing drainage area less than about 10 square miles. Peak-streamflow frequency estimation for small undeveloped (rural) and ungaged watersheds in Texas is a major concern for TxDOT engineers. The alternative framework uses a technique involving the minimization of the PRESS (PRediction Error Sum of Squares) statistic (Helsel and Hirsch, 2002, p. 247). The secondary purpose of this report is to present regression equations based on PRESS minimization for the estimation of peak-streamflow frequency at ungaged sites in undeveloped watersheds in Texas. Finally, the tertiary purpose of this report is to present "statewide" regression equations for Texas lacking a context of specific geographic regions.

The scope of the report is limited to the at-site peak-streamflow frequency values for 664 USGS streamflow-gaging stations used in AS1997 and digitally tabulated in Asquith and Slade (1999, file `tx664.dat`). The alternative regression equations presented here are based on the entire study area (Texas and slight overlap with surrounding states) of AS1997. The scope of the report does not include consideration of the spatial dependence of peak-streamflow frequency beyond its association with mean annual precipitation.

Acknowledgments

The authors recognize the following colleagues of TxDOT for their contributions and support of this work: George R. Herrmann, 0–5521 Project Director and Amy J. Ronnfeldt, 0–5521 Project Advisor.

Current (2008) Regional Regression Equations for Peak-Streamflow Frequency Estimation in Texas

The current (2008) regional regression equations are provided by AS1997, who provide 96 equations to estimate the 2-, 5-, 10-, 25-, 50-, and 100-year annual peak discharge (the peak-streamflow frequency curve) for undeveloped watersheds in Texas. The equations use the watershed characteristics of drainage area, main-channel slope, and watershed shape as predictor variables. AS1997 divides Texas into 11 regions. The mean number of stations used for each equation is 36. For each region, 6 or 12 weighted-least-squares regression equations were developed using a forward stepwise procedure. The distinction between 6- and 12-equation regions is elaborated upon later in this section.

The AS1997 statistical analysis is sound, with innovative methods of equation development and presentation, and widely used (in a second printing); however, three observations regarding the AS1997 procedural framework are important for this report. The observations are important because they relate to application or implementation of the AS1997 equations by end users involved in public and private infrastructure design. The observations gradually developed over the years since publication of AS1997 and were refined for this investigation. The three observations are described in the following sections.

Inconsistent Peak-Streamflow Frequency Curves by Regional Regression

For a given region, watershed characteristics used to develop the AS1997 regression equations for the 2- through 100-year equations are inconsistent; a fact that can be attributed to statistically inconsistent peak-streamflow frequency curve for some watersheds. By definition, a peak-streamflow frequency curve must monotonically increase with increasing recurrence interval. The term inconsistent in this context means that the computed discharge for a recurrence interval exceeds the discharge for a larger recurrence interval. For example, the 50-year peak streamflow is computed to be greater than the 100-year peak streamflow. The source of the peak-streamflow inconsistency is the inconsistent use of watershed characteristics within an equation ensemble (a set of equations for a given region).

The inconsistency in watershed characteristics exists because AS1997 used a forward stepwise regression procedure and did not specifically force predictor variables into the equations. For example, the equations for region 11 of AS1997 (southeastern Texas) are listed in table 1. Main-channel slope is not used for the 2-year recurrence interval, but it is for larger recurrence intervals. Watershed shape is used for the 2- through 10-year recurrence intervals, but it is not used for larger recurrence intervals. Although difficult to visualize, combinations of watershed characteristics can be substituted into the equations listed in table 1 to produce an inconsistent frequency curve.

AS1997 explicitly discusses the potential for inconsistent peak-streamflow frequency curves from the equations (Asquith and Slade, 1997, p. 11). When the equations and guidance on equation application originally were developed, the authors (Asquith and Slade) anticipated that end users would apply "hydrologic engineering judgement" to manually mitigate peak-streamflow inconsistencies. However, numerous end users have communicated to the senior author a degree of confusion or frustration in regard to application of the AS1997 equations, which indicates a need for alternative equations that will not produce, or have a greatly reduced potential for producing, inconsistent peak-streamflow frequency curves.

Regional Regression Equation Applicability and Implementation

AS1997 provides numerous figures (Asquith and Slade, 1997, figs. 4–14) in which the relations between drainage area, main-channel slope, and watershed shape are graphically depicted for each of the 11 regions. Superimposed on these plots are generalized "convex hulls"[2] representing the "approximate [parameter space] defined by [watershed] characteristics" for each region. For watersheds having coordinates of drainage area, main-channel slope, and watershed shape outside the convex hull, the applicability of the equations for the region is uncertain, and the potential for an inconsistent peak-streamflow frequency curve increases.

Since publication of AS1997, the senior author has learned from interaction with end users that the convex hulls presented in AS1997 commonly are underutilized.

[2]Quoting from http://en.wikipedia.org/wiki/Convex_hull accessed on February 19, 2008: "For planar objects, [those] lying in the plane, the convex hull may be easily visualized by imagining an elastic band stretched open to encompass the given object; when released, it will assume the shape of the required convex hull."

Table 1. Asquith and Slade (1997) regression equations for region 11 (southeast Texas).

[Q_T, peak streamflow for T-year recurrence interval in cubic feet per second; A, drainage area in square miles; S, main-channel slope in feet per mile; and H, dimensionless watershed shape. Sixty-six stations were used in the regression development.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$
$Q_2 = 159A^{0.669}H^{-0.262}$	0.91	0.18
$Q_5 = 191A^{0.696}S^{0.130}H^{-0.186}$	.91	.18
$Q_{10} = 199A^{0.718}S^{0.221}H^{-0.151}$	.90	.20
$Q_{25} = 201A^{0.713}S^{0.313}$	.88	.22
$Q_{50} = 207A^{0.735}S^{0.380}$	.86	.24
$Q_{100} = 213A^{0.755}S^{0.442}$	.85	.26

Furthermore, some end users abstracted (reproduced) for application only the equations from AS1997. As a result, important context that contributes to optimum use of the equations is lost.

The apparent lack of full adherence to the entire procedural framework and caveats of the AS1997 regional regression equations is understandable given that AS1997 provides 96 separate equations and considerable detail. Therefore, a simpler regional regression method for estimation of peak-streamflow frequency in Texas would be useful.

Biased Peak-Streamflow Frequency Values

The multiple linear regional regression equations of AS1997 are exclusively based on $\log_{10}$ transformations of observed peak-streamflow frequency values, drainage area, main-channel slope, and watershed shape. Multiple linear regression is based on a linear relation between the regressor variable (peak-streamflow frequency) and the predictor variables (drainage area, main-channel slope, watershed shape, and others). AS1997 (Asquith and Slade, 1997, p. 8) notes that, for some regions, peak-streamflow values (for example, the 100-year peak streamflow) have a discernible curvilinear relation with drainage area in $\log_{10}$ space. AS1997 addresses the nonlinearity (and thus mitigates the bias) by classifying watersheds into two ranges of drainage area. Separate regional regression analyses were done for watersheds with drainage areas less than 32 square miles and for watersheds with areas greater than 32 square miles. The 32-square-mile break point was determined by data interpretation. The drainage-area distinction and bias mitigation is explicitly discussed in AS1997 (Asquith and Slade, 1997, p. 13).

The drainage-area classification was not made for six of the 11 regions because either the number of watersheds was small (degrees of freedom for regression) within a region or an absence of a discernible curvilinear relation between $\log_{10}$-transformed peak streamflow and drainage area was perceived. For a region in which the drainage-area classification was made, 12 equations for the region were developed—six equations for watersheds with drainage areas less than 32 square miles and six equations for watersheds with drainage areas greater than 32 square miles. Conversely, six equations were developed for regions in which nonlinearity was not apparent and no drainage-area classification was made.

The drainage-area classification complicates application of the equations for watersheds near the 32-square mile break point. AS1997 (Asquith and Slade, 1997, p. 12) provides an ad hoc procedure to prorate estimates for watersheds of 10 to 100 square miles between the equation ensemble for drainage areas less than or equal to 32 square miles and the ensemble for drainage areas greater than or equal to 32 square miles. If the proration procedure is not followed, "jumps" in peak streamflow at 32 square miles will result.

The nonlinearity is apparent in the graphical depiction of the 32-square-mile classification technique to mitigate for nonlinearity (Asquith and Slade, 1997, figs. 3 and 15). Despite the measures to address the nonlinearity and thus mitigate bias, the AS1997 equations still have the potential to overestimate peak-streamflow frequency values for both the smallest and largest watersheds. As noted, eliminating or reducing the potential for inconsistent peak-streamflow frequency curves and making the regional regression equation method easier for end users to apply are ancillary reasons to develop the alternative equations shown in this report; but another reason for development of alternative equations is to remove the bias inherent in the AS1997 equations.

Typical regression practice to reduce underestimation or overestimation (peak-streamflow frequency values for this report) is to seek an alternative transformation on the regressor variable (Maindonald and Braun, 2003, p. 126–127). Some readers might question why an alternative transformation on drainage area (a predictor variable) is sought rather than an alternative transformation on the 2- through 100-year peak-streamflow values (regressor variables). The authors chose to assess an alternative transformation on drainage area so that the residual standard errors ($\log_{10}$ units of streamflow) reported are directly comparable to those from AS1997.

Alternative Regression Equations for Estimation of Peak-Streamflow Frequency for Watersheds in Texas

Regression Equations Based on Logarithmic Transformation of Drainage Area

The traditional practice for development of regression equations to estimate peak-streamflow frequency is to transform regressor variables (the at-site peak-streamflow frequency values, such as the 2- through 100-year peak streamflows) and all the predictor variables (Stedinger and others, 1993, p. 18.35) by the $\log_{10}$ function. Drainage area, a measure of watershed slope, and other characteristics are common predictor variables. AS1997 considered six characteristics: 2-year 24-hour precipitation, mean annual precipitation (1951–80)[3], drainage area, stream length, basin shape factor, and main-channel slope. The precipitation statistics reported in AS1997 are for the approximate watershed centroid. However, for the equations reported in AS1997, only drainage area, main-channel slope, and watershed shape are used. For this report, only drainage area, mean annual precipitation, and main-channel slope are used.

Because of the ubiquitous nature of $\log_{10}$ transformation in hydrologic analyses, important comparative analysis for this report is facilitated by developing regression equations using $\log_{10}$ transformation on the same data used in AS1997. However, no designation of geographic region is used for this report. AS1997 considered data for 664 USGS streamflow-gaging stations. From preliminary data analysis (results not presented here), eight stations were identified as outliers on the basis of the relative change of exploratory regressions and associated diagnostics when each of the outlying stations were individually dropped. These stations were eliminated from further analysis. The summary statistics (table 2) of drainage area, mean annual precipitation, and main-channel slope were computed after the removal of the eight stations listed in table 3.

Weighted-least squares regression on the 2-, 5-, 10-, 25-, 50-, and 100-year peak-streamflow values for the remaining 656 streamflow-gaging stations is accomplished using drainage area, mean annual precipitation, and main-channel slope as predictor variables. For comparison, the mean number of stations per equation in AS1997 is 36. Therefore, the degrees of freedom for the regression equations reported here are about 18 times larger than those of AS1997.

Analysis of collinearity through variance inflation factors and statistical significance (results not reported here) strongly indicated that inclusion of watershed shape in the regression equations in addition to drainage area, mean

[3]This period coincides with the most undeveloped peak-streamflow data and streamflow-gaging stations in Texas.

Table 2. Summary statistics of basin characteristics used in regression analysis described in this report.

[*A*, drainage area, in square miles; *P*, mean annual precipitation, in inches (1951–1981); *S*, main-channel slope, in feet per mile]

	Mininum	1st Quartile	Median	3rd Quartile	Maximum
A	0.10	7.2	128	6,960	174,000
P	8	22	31	41	57
S	.38	7.35	13.4	33.6	371

Table 3. U.S. Geological Survey streamflow-gaging stations identified as outliers and removed from analysis.

Station no.	Station name	Drainage area (square miles)
08039900	Little Sandy Creek tributary near Jasper, Texas	0.46
08080700	Running Water Draw at Plainview, Texas	382
08089100	Elm Creek tributary near Graford, Texas	1.10
08210400	Lagarto Creek near George West, Texas	155
08383200	Pintada Arroyo tributary near Clines Corners, New Mexico	29.20
08393600	North Spring River at Roswell, New Mexico	19.50
08405050	Last Chance Canyon tributary near Carlsbad Caverns, New Mexico	.20
08434000	Toyah Creek below Toyah Lake near Pecos, Texas	3,709

annual precipitation, and main-channel slope is not appropriate. The six regression equations are listed in table 4. For all six equations, the p-values for the coefficients on the watershed characteristics are less than .0001, which means that the variables and intercept are all highly statistically significant.

A simple comparison between 100-year residual standard error listed in table 4 and the weighted-mean 100-year residual standard error from AS1997 provides perspective. The weighted-mean 100-year residual standard error from AS1997 is computed by weighting the errors in AS1997 (Asquith and Slade, 1997, table 2) by the number of stations for each region. The weighted-mean, 100-year, residual standard error from AS1997 is about 0.27; the 100-year residual standard error listed in table 4 is 0.34. These two residual standard errors are of similar magnitude. Additional comparisons of residual standard errors listed in table 4 to those in AS1997 indicate that all have about the same magnitude, although overall the errors are greater for the equations reported here. The conclusion from this comparison is that the six equations in table 4 have approximately the same residual standard error as the 96 equations reported in AS1997.

For the equations in table 4, inclusion of mean annual precipitation for the watershed is useful. Mean annual precipitation becomes a surrogate for spatial location that replaces the concept of geographic region designation associated with the equations in AS1997. Mean annual precipitation was not used in AS1997 for the final equations shown in that report.

Bias in multiple linear regression is well depicted in a residual (observed minus predicted) graph in which the residual for a particular data point is plotted on the vertical axis and the corresponding fitted value is plotted on the horizontal axis. If there is a discernible trend or shape in the graph—that is, a tendency for residuals to plot above or below the zero-residual line, then bias in the equation exists.

Residuals for the 100-year peak-streamflow equation listed in table 4 are graphed in figure 1. A LOWESS (LOcally WEighted Scatterplot Smoothing) trend line (Cleveland, 1979) through the data is superimposed. The `lowess()` function of the R software package (R Development Core Team, 2006) with default settings was used. The concave-down shape of the LOWESS trend line indicates systematic bias in the regression. The negative magnitudes of the left and right segments of the LOWESS trend line indicate that overestimation of the 100-year peak-streamflow occurs for watersheds with small and large fitted values (the smallest and largest watersheds, respectively).

The LOWESS trend line is only an indicator of bias and does not represent a true bias correction; however, interpretation of the line as a bias measure is useful. For example, referring to figure 1, for a fitted value of about 2.5 (316 cubic feet per second) and a LOWESS-indicated bias of about −0.25, a more appropriate value might be $2.5 - 0.25 = 2.25$ (178 cubic feet per second). Therefore, the bias-corrected value, albeit ad hoc, is about 44 percent less than the fitted value. In general, the $\log_{10}$-exclusive regressions in table 4 have concave-down trend lines through the residuals. The concavity of the LOWESS trend line (interpreted as bias in the equations) increases with increasing recurrence interval (results not presented here).

Hydrologic scale typically is measured by drainage area. Therefore, it is informative to develop second and third sets of $\log_{10}$-transformed regression equations on the same 656 stations using drainage area and mean annual precipitation (table 5) and only drainage area (table 6) as predictor variables.

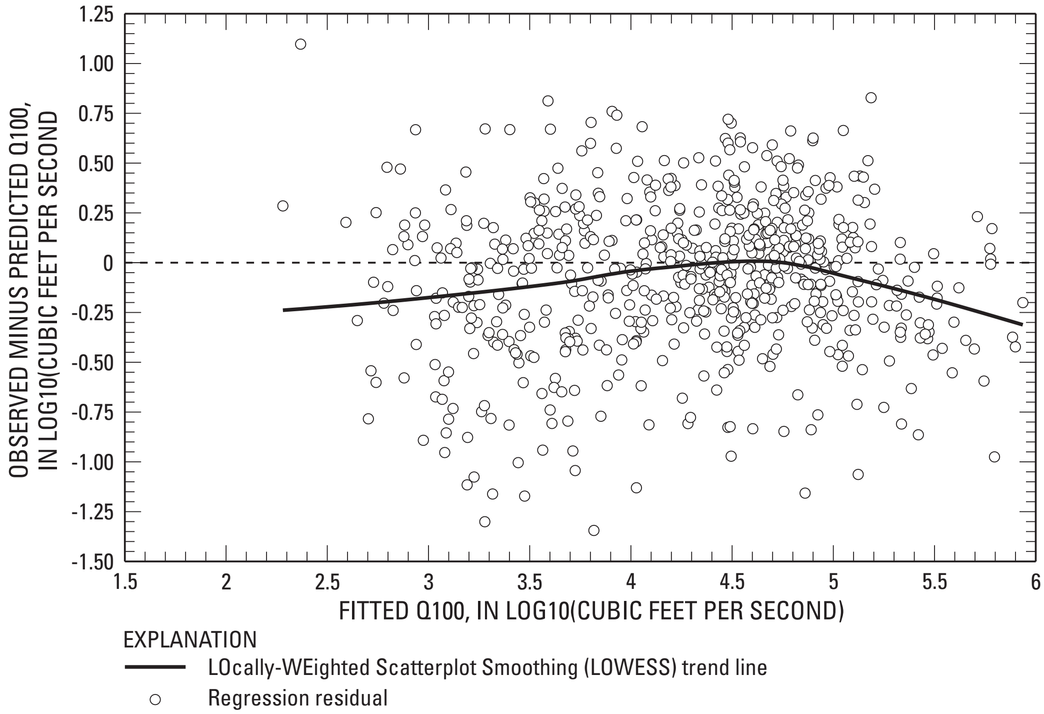


Figure 1. Residual plot of regression of 100-year peak streamflow using logarithmic transformation of drainage area using three predictor variables.

For these 12 equations, the p-values for the coefficients on the intercept, drainage area, and mean annual precipitaiton are less than .0001. The residual standard errors associated with the equations in tables 5 and 6 are all greater than those listed in table 4 and because one and two fewer predictor variables are in the equations in tables 5 and 6, respectively.

Residuals for the 100-year peak-streamflow equation listed in table 6 are shown in figure 2. The LOWESS trend line superimposed through the data has considerable downward concavity similar to the trend line in figure 1. The interpretations of the regressions in table 6 using the LOWESS trend line on the residual plot are the same as those for the regression equations in table 4. Specifically, peak streamflow is overestimated for watersheds with small fitted values (the smallest watersheds) and for watersheds with large fitted values (the largest watersheds). The bias is considerable. The concavity of the LOWESS trend line increases with increasing recurrence interval (results not presented here).

In conclusion, systematic bias is present in the regression equations reported in tables 4–6, and by general method association, bias is present in the AS1997 equations. The bias exists because of the curvilinear relation between $\log_{10}$-transformed peak streamflow and drainage area. The bias is mitigated in the AS1997 analysis by separating regressions into two groups on the basis of watershed drainage area, less than or greater than 32 square miles. The relation between $\log_{10}$-transformed peak streamflow and drainage area becomes increasingly curvilinear with increasing recurrence interval.

Regression Equations Based on PRESS Minimization and Power Transformation of Drainage Area

The PRESS statistic generally is regarded as a measure of regression performance when the model is used to predict new data (Montgomery and others, 2001, p. 153). Prediction of new data is what analysts and engineers do when they estimate peak streamflow from a regression equation. Regression equations with small PRESS values are desirable. Thus, PRESS minimization is an appropriate goal. Helsel and Hirsch (2002, p. 247) state that, “Minimizing PRESS means that the equation produces the least error when making new predictions.” Conceptually, PRESS minimization identifies the appropriate transformation to “press” the bias out of the equations (fig. 3).

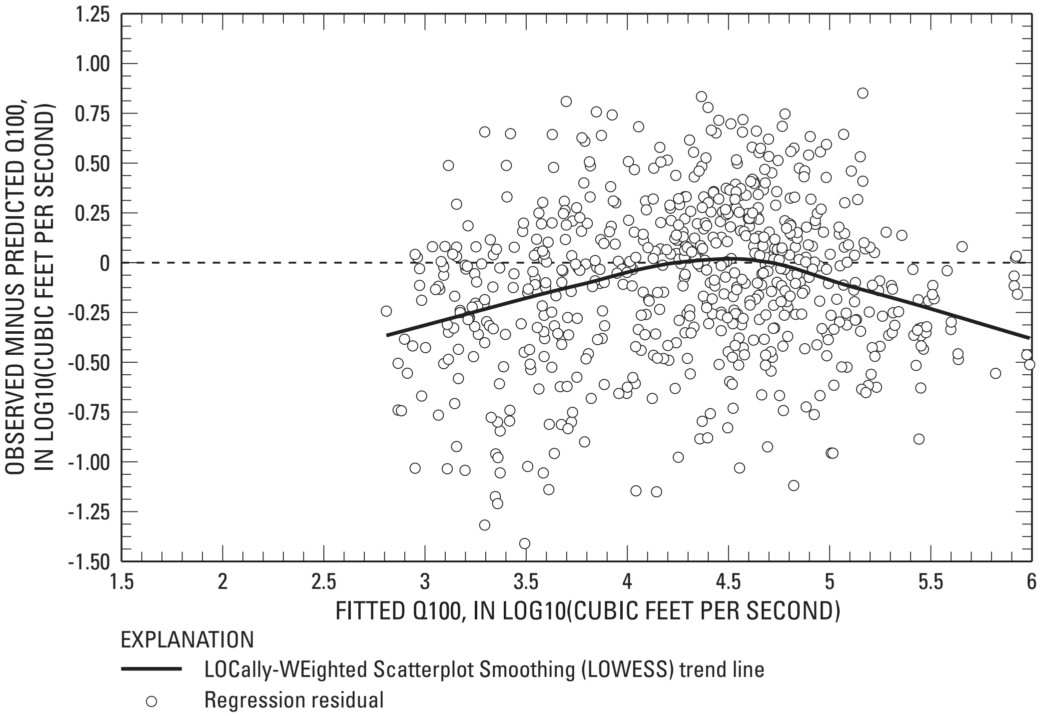


Figure 2. Residual plot of regression of 100-year peak streamflow using logarithmic transformation of drainage area using drainage area as the only predictor variable.

The PRESS statistic is computed from the PRESS residuals, which are defined as

$$e_{(i)} = y_i - y_i', \tag{1}$$

where $e_{(i)}$ is the PRESS residual, y_i is the observed ith peak-streamflow value, and y_i' is the predicted value based on the remaining $n-1$ sample points. In other words, the ith station (data point) is not used to generate the ith regression equation. Thus, PRESS residuals are regarded as validation statistics. The PRESS statistic, with inclusion of the regression weight factor (w_i), is

$$\text{PRESS} = \sum_{i=1}^{n} w_i e_{(i)}^2. \tag{2}$$

Equation 2 is computationally intensive (n regressions are required). A more efficient computation of PRESS is made using regression residuals (e_i) and leverage (h_{ii}). (The double subscript ii refers to the diagonal of the hat matrix.[4]) These values are readily available from modern regression software packages. The PRESS computation is made by

$$\text{PRESS} = \sum_{i=1}^{n} w_i \left(\frac{e_i}{1-h_{ii}}\right)^2. \tag{3}$$

Because the PRESS statistic is an overall measure of regression fit (like residual standard error) and is a validation statistic (unlike residual standard error), minimization of PRESS is desirable. The most "valid" regression is produced when the PRESS statistic is minimized. The following transformation on drainage area was selected after exploratory analysis:

$$A' = A^{\lambda}, \tag{4}$$

where A' is the transformed value for the regression, A is drainage area, and λ is a real number. The transformation is referred to in this report as the power transformation.

Three computer programs were written in R (R Development Core Team, 2006) to loop through successive non-integer values of λ and record the value that yields a minimum PRESS for each of the six recurrence intervals. Tens of thousands of regressions were done in the process of exploratory data analysis and for the final minimization reported here. The first program implemented the

[4]Quoting from Montgomery and others (2001, p. 75): The $n \times n$ matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$ is usually called the hat matrix. It maps the vector of observed values into a vector of fitted values. The hat matrix and its properties play a central role in regression analysis.

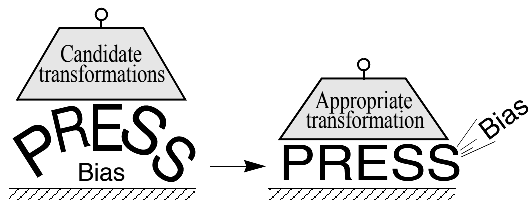


Figure 3. Conceptual display of the PRESS statistic minimization.

watershed characteristics drainage area, mean annual precipitation, and main-channel slope as predictor variables; the second program implemented drainage area and mean annual precipitation; and the third program implemented only drainage area as a predictor variable.

The programs and incremental output are provided in appendixes 1–3. The programs and output are included in the report to provide an archive of the PRESS minimization algorithm and the regression analysis results summarized in tables 4–6 as well as tables 7–9.

The results of the power transformation of drainage area using the three predictor variables are listed in table 7. The value of λ is the exponent on A in the equations. The value of λ increases in absolute magnitude with increasing recurrence interval; the larger the absolute value of λ, the larger the amount of concavity in the trend line of residuals (systematic bias) that is reduced.

In all six equations, the p-values for the coefficients on the watershed characteristics are less than .0001. The diagnostic statistics of adjusted R-squared and residual standard error in the table are greater and less than, respectively, those in table 4. Therefore, the equations using the power transformation have less uncertainty. However, the PRESS statistic is the more important statistic to compare.

The PRESS statistic for a given recurrence interval is less when the power transformation on drainage area is used instead of the $\log_{10}$ transformation. The percentage changes in the PRESS statistics associated with the power transformation (table 7) compared to those associated with the $\log_{10}$ transformation (table 4) show that, as recurrence interval increases, the power transformation produces an increasingly more valid regression.

Residual standard errors of the PRESS-minimized equations in table 7 are similar to those of the equations reported in AS1997. For example, the 100-year residual standard error is about 0.33 and the AS1997 weighted value is 0.27 for the 11 regions collectively.

Residuals for the 100-year peak-streamflow equations using the three predictor variables (table 7) are shown in figure 4. Downward concavity of the superimposed LOWESS trend line is not present, unlike the LOWESS trend line in figure 1. In fact, the LOWESS trend line is essentially flat, which indicates that systematic bias in the equation is reduced through use of the specified power transformation. The power transformation on drainage area effectively linearizes the relation between 100-year peak streamflow and drainage area. Minimization of the PRESS statistic effectively removes systematic bias. Similar results (not reported here) were obtained for the other five recurrence intervals.

The results of the power transformation of drainage area using drainage area and mean annual precipitation and only drainage area as predictor variables are listed in tables 8 and 9. Again, the value of λ is the exponent on A in the equations. In all 12 equations, the p-values for the coefficients on the watershed characteristics are less than .0001. Adjusted R-squared and residual standard error for regression based on power transformation are greater and less than, respectively, for those regressions based exclusively on $\log_{10}$ transformation (tables 5 and 6). Therefore, the equations using the power transformation have less uncertainty. The PRESS statistic for a given recurrence interval is less when the power transformation on drainage area is used instead of the $\log_{10}$ transformation. The percentage changes in the PRESS statistic associated with the power transformation (tables 8 and 9) compared to those associated with the $\log_{10}$ transformation (tables 5 and 6) show that, as recurrence interval increases, the power transformation produces an increasingly more valid regression.

Residuals for the 100-year peak-streamflow equations in table 9 are graphed in figure 5. The concave-down shape of the superimposed LOWESS trend line in the residuals graph from the $\log_{10}$ transformation (fig. 2) is not present in the graph derived from the power transformation. In fact, the LOWESS trend line is essentially flat (fig. 5), which indicates that systematic bias in the equation has been reduced. The authors conclude that the power transformation on drainage area effectively linearizes the relation between 100-year peak streamflow and drainage area. Minimization of PRESS effectively removes systematic bias. Similar results (not reported here) were obtained for the other five recurrence intervals.

The PRESS statistics for the equations in tables 7 and 9 are shown graphically by recurrence interval in figure 6 (the PRESS statistics for the equations in table 8 are not shown). From the figure it is clear that the power transformation with PRESS minimization produces PRESS statistics less than those from the $\log_{10}$-exclusive equations. PRESS minimization becomes increasingly important as recurrence interval increases because the $\log_{10}$ transformation does not produce a linear relation between peak streamflow and drainage area for the larger recurrence-interval events.

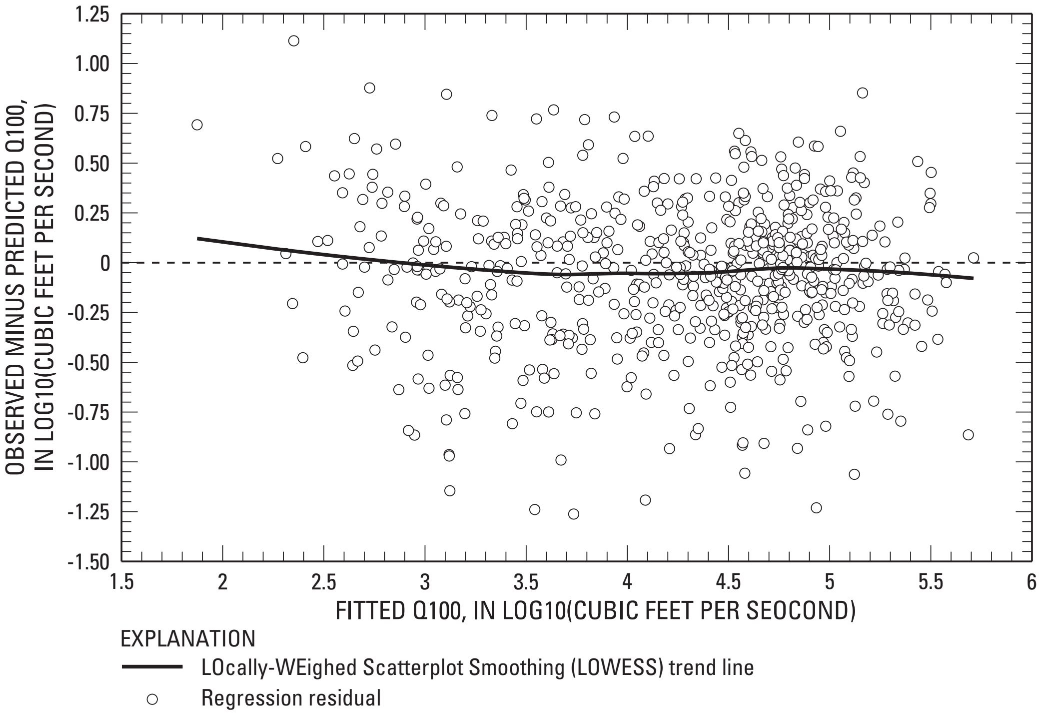


Figure 4. Residual plot of regression of 100-year peak streamflow using power transformation of drainage area using three predictor variables.

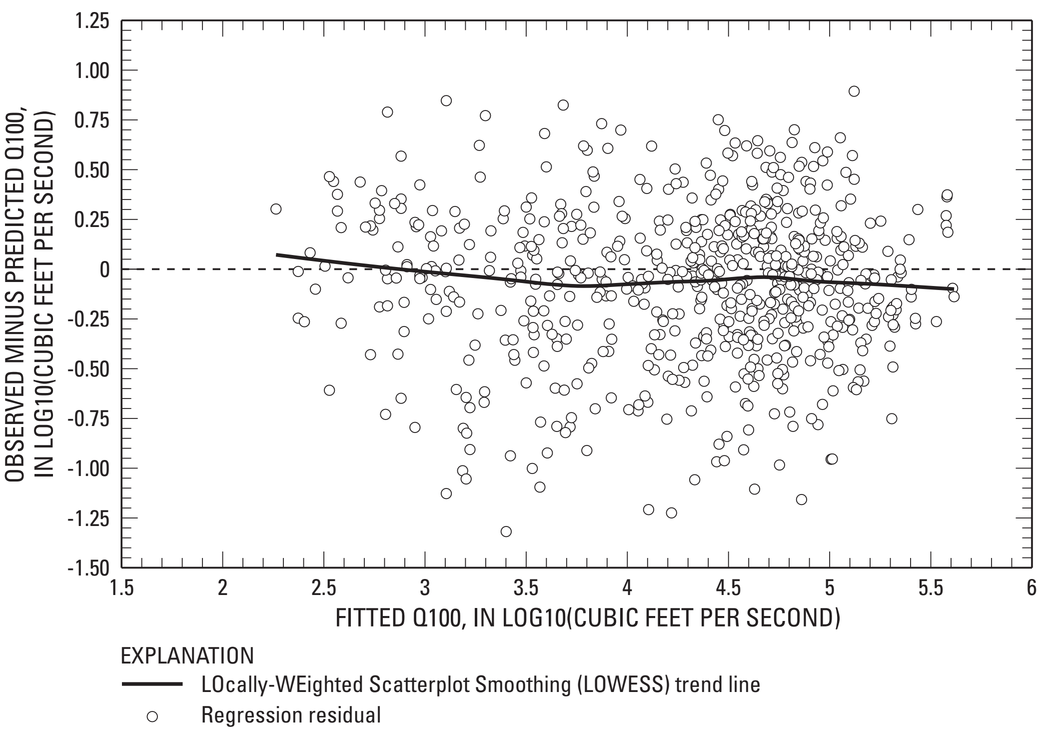


Figure 5. Residual plot of regression of 100-year peak streamflow using power transformation of drainage area using drainage area as the only predictor variable.

Although the smallest PRESS statistics occur for the 5-year recurrence interval, the PRESS statistics for the 2-year recurrence interval are not exceeded until the 25-year and larger recurrence intervals are reached. An interpretation of the PRESS statistic is that estimation of the 2-year peak streamflow using watershed characteristics is more difficult than estimation for the 5-year and 10-year peak streamflow. This observation is consistent with residual standard errors reported in AS1997.

Finally, the magnitude and extent of the bias between the $\log_{10}$-exclusive regression and the PRESS-minimized regression is informative. The magnitude of the bias can be expressed as the ratio (the bias ratio) of the $\log_{10}$ equations (tables 4 or 6) to the PRESS-minimized equations (tables 7 or 9). For example, the bias ratio for the 100-year peak streamflow for the drainage-area-only equations is

$$\frac{Q_{100}^{\log_{10}}}{Q_{100}^{\mathrm{PRESS}}} = \frac{10^{3.318}A^{0.5094}}{10^{6.800-3.659A^{-0.0934}}} \equiv \text{y-axis in figure 7.} \qquad (5)$$

When the ratio is greater than 1, the $\log_{10}$-exclusive regression overestimates peak streamflow relative to the PRESS-minimized regression. Similar equations of the bias ratio for other recurrence intervals are easily defined. Together, the six equations defining the bias ratio document the inherent differences between the $\log_{10}$-exclusive peak-streamflow equations and the PRESS-minimized equations.

The extent of the bias ratio is shown by the ratio as a function of drainage area. An example, by recurrence interval, for the regressions using drainage area as the only predictor variable is shown in figure 7 (see eq. 5). An interpretation of the figure is that the $\log_{10}$-exclusive regressions overestimate peak-streamflow frequency for drainage areas less than about 8 square miles and drainage areas greater than about 2,000 square miles. The overestimation for drainage areas less than about 2 square miles is substantial. The overestimation for drainage areas less than about 0.5 square mile exceeds 100 percent for all but the 2-year peak streamflow. Alternatively, the $\log_{10}$-exclusive regressions slightly underestimate peak-streamflow frequency for drainage areas between about 8 and 2,000 square miles.

Summary

Peak-streamflow frequency estimates are needed for flood-plain management; for objective assessment of flood risk; for cost-effective design of dams, levees, other flood-control structures; and for design of roads, bridges, and culverts. Peak-streamflow frequency represents the collective peak streamflow for recurrence intervals of 2, 5, 10, 25, 50, and 100 years.

The U.S. Geological Survey (USGS), in cooperation with the Texas Department of Transportation and in partnership with Texas Tech University, investigated a refinement of the regional regression method and developed alternative equations for estimation of peak-streamflow frequency for undeveloped watersheds in Texas. A common model for estimation of peak-streamflow frequency is based on the regional regression method, which relates peak-streamflow frequency to watershed characteristics.

The current (2008) regional regression equations (96 separate equations) for 11 geographic regions of Texas are based on $\log_{10}$ transformations on all regression variables (the peak-streamflow values and the watershed characteristics of drainage area, main-channel slope, and watershed shape). The $\log_{10}$ transformation does not fully linearize the relations between the variables, which is a major assumption in linear regression analysis. As a result, some systematic bias remains in the current equations. The primary source of the bias is the discernible curvilinear relation between peak streamflow and drainage area in $\log_{10}$ space. The bias results in overestimation of peak streamflow for both the smallest and largest watersheds, and the bias increases with increasing recurrence interval.

To demonstrate the extent of the bias, equations using $\log_{10}$(drainage area) for the study area (Texas and slight overlap with surrounding states) are reported. Separate regional distinction is not made for this report. Mean annual precipitation provides a surrogate for spatial location that replaces the concept of geographic region designation associated with the current equations. The use of mean annual precipitation reduces the number of equations for a given number of predictor variables (three, two, or one) from 96 to 6—one equation for each of the six recurrence intervals. To address the bias, a statistical framework based on minimization of the PRESS statistic through power transformation on drainage area is described.

The PRESS statistic is an important measure of regression performance. It is a validation-type statistic, and small values are desirable. Minimization of PRESS is appropriate for peak-streamflow frequency analysis because the equations are used in hydrologic engineering practice to predict new data.

Compared to $\log_{10}$(drainage area) equations, the equations derived from PRESS minimization have PRESS statistics and residual standard errors less than the $\log_{10}$(drainage area) equations. Selected residual plots for the PRESS-minimized equations demonstrate that the systematic bias in regional regression equations for peak-streamflow frequency estimation in Texas can be reduced. Because the overall error is similar to the overall error associated with the equations currently in use and bias is reduced, the PRESS-minimized equations reported here provide alternative equations for peak-streamflow frequency estimation.

References

Asquith, W.H., and Slade, R.M., 1997, Regional equations for estimation of peak-streamflow frequency for natural basins in Texas: U.S. Geological Survey Water-Resources Investigations Report 96–4307, 68 p. [`http://pubs.usgs.gov/wri/wri964307/`]

Asquith, W.H., and Slade, R.M., 1999, Site-specific estimation of peak-streamflow frequency using generalized least-squares regression for natural basins in Texas: U.S. Geological Survey Water-Resources Investigations Report 99–4172, 19 p. [`http://pubs.usgs.gov/wri/wri994172/`]

Asquith, W.H., and Thompson, D.B., 2005, Alternative regression equations for estimation of annual peak-streamflow frequency for undeveloped watersheds in Texas using PRESS minimization: Lubbock, Center for Multidisciplinary Research in Transportation, Texas Tech University, Texas Department of Transportation Research Report 0–4405–2, 27 p. [`http://library.ctr.utexas.edu/pdf/4405-2.pdf`]

Cleveland, W.W., 1979, Robust locally weighted regression and smoothing scatterplots: Journal of the American Statistical Association, v. 80, p. 829–836.

Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chap. A3, 510 p. [`http://pubs.usgs.gov/twri/twri4a3/`]

Maindonald, J., and Braun, J., 2003, Data analysis and graphics using R—An example-based approach: Cambridge, United Kingdom, Cambridge University Press, 362 p.

Montgomery, D.C., Peck, E.A., and Vining, G.G., 2001, Introduction to linear regression analysis (3rd ed.): New York, Wiley, 641 p.

R Development Core Team, 2006, R—A language and environment for statistical computing: Vienna, Austria, R Foundation for Statistical Computing, ISBN 3–900051–07–0, `http://www.R-project.org` and direct link to report is [`http://www.cran.r-project.org/doc/manuals/refman.pdf`]

Stedinger, J.R., Vogel, R.M., and Foufoula-Georgiou, E., 1993, Frequency analysis of extreme events, chapter 18, *in* Maidment, D.A., ed., Handbook of hydrology: New York, McGraw-Hill, p. 18.1–18.66.

Thompson, D.B., 2006, The rational method, regional regression equations, and site-specific flood frequency relations: Lubbock, Center for Multidisciplinary Research in Transportation, Texas Tech University, Texas Department of Transportation Research Report 0–4405–1, 92 p., `http://www.techmrt.ttu.edu/reports.php` and direct link to report is [`http://www.techmrt.ttu.edu/Reports/Complete%20Reports/4405-1.pdf`]

Table 4. Regression equations based on logarithmic transformation of drainage area using three predictor variables.

[Q_T, peak streamflow for T-year recurrence interval in cubic feet per second; A, drainage area in square miles; P, mean annual precipitation in inches; and S, main-channel slope in feet per mile.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$	PRESS statistic
$Q_2 = 10^{-0.5240} A^{0.6565} P^{1.474} S^{0.3525}$	0.8282	0.2866	54.36
$Q_5 = 10^{-0.2204} A^{0.6790} P^{1.376} S^{0.4828}$	.8414	.2686	47.76
$Q_{10} = 10^{-0.04207} A^{0.6896} P^{1.317} S^{0.5421}$	.8310	.2778	51.12
$Q_{25} = 10^{0.1501} A^{0.7005} P^{1.256} S^{0.6005}$	.8086	.2993	59.34
$Q_{50} = 10^{0.2748} A^{0.7073} P^{1.218} S^{0.6359}$	.7887	.3186	67.25
$Q_{100} = 10^{0.3879} A^{0.7133} P^{1.183} S^{0.6660}$	.7675	.3393	76.25

Table 5. Regression equations based on logarithmic transformation of drainage area using two predictor variables.

[Q_T, peak streamflow for T-year recurrence interval in cubic feet per second; A, drainage area in square miles; and P, mean annual precipitation in inches.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$	PRESS statistic
$Q_2 = 10^{0.8330} A^{0.5534} P^{0.9732}$	0.8153	0.2971	58.29
$Q_5 = 10^{1.639} A^{0.5378} P^{0.6896}$	.8157	.2895	55.32
$Q_{10} = 10^{2.045} A^{0.5311} P^{0.5469}$	.7987	.3032	60.69
$Q_{25} = 10^{2.462} A^{0.5249} P^{0.4029}$	.7698	.3282	71.10
$Q_{50} = 10^{2.723} A^{0.5214} P^{0.3140}$	.7463	.3491	80.44
$Q_{100} = 10^{2.952} A^{0.5186} P^{0.2366}$	.7224	.3707	90.72

Table 6. Regression equations based on logarithmic transformation of drainage area using one predictor variable.

[Q_T, peak streamflow for T-year recurrence interval in cubic feet per second; and A, drainage area in square miles.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$	PRESS statistic
$Q_2 = 10^{2.339} A^{0.5158}$	0.7642	0.3357	74.22
$Q_5 = 10^{2.706} A^{0.5111}$	.7889	.3099	63.28
$Q_{10} = 10^{2.892} A^{0.5100}$	.7820	.3156	65.63
$Q_{25} = 10^{3.086} A^{0.5093}$	.7612	.3343	73.67
$Q_{50} = 10^{3.209} A^{0.5092}$	.7414	.3525	81.88
$Q_{100} = 10^{3.318} A^{0.5094}$	.7199	.3724	91.39

Table 7. Regression equations based on power transformation of drainage area using three predictor variables.

[Q_T, peak streamflow for T-year recurrence interval in cubic feet per second; A, drainage area in square miles; P, mean annual precipitation in inches; and S, main-channel slope in feet per mile. The exponent of A is the power λ.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$	PRESS statistic	Percent change from PRESS in table 4
$Q_2 = 10^{35.60-36.09A^{-0.0082}} P^{1.448} S^{0.3472}$	0.8286	0.2863	54.27	−0.17
$Q_5 = 10^{11.16-11.28A^{-0.0299}} P^{1.279} S^{0.4640}$	.8461	.2646	46.37	−2.9
$Q_{10} = 10^{9.047-8.950A^{-0.0400}} P^{1.188} S^{0.5172}$	.8396	.2707	48.57	−5.0
$Q_{25} = 10^{7.949-7.628A^{-0.0497}} P^{1.096} S^{0.5699}$	.8217	.2889	55.32	−6.8
$Q_{50} = 10^{7.554-7.090A^{-0.0553}} P^{1.039} S^{0.6021}$	.8048	.3062	62.18	−7.5
$Q_{100} = 10^{7.307-6.714A^{-0.0601}} P^{0.9883} S^{0.6295}$	.7862	.3253	70.19	−7.9

Table 8. Regression equations based on power transformation of drainage area using two predictor variables.

[Q_T, peak streamflow for T-year recurrence interval in cubic feet per second; A, drainage area in square miles; and P, mean annual precipitation in inches. The exponent of A is the power λ.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$	PRESS statistic	Percent change from PRESS in table 5
$Q_2 = 10^{17.36-16.51A^{-0.0157}} P^{0.9429}$	0.8162	0.2965	58.02	−0.46
$Q_5 = 10^{8.080-6.403A^{-0.0451}} P^{0.6065}$	.8226	.2841	53.27	−3.7
$Q_{10} = 10^{7.200-5.107A^{-0.0596}} P^{0.4397}$	.8103	.2943	57.18	−5.8
$Q_{25} = 10^{6.849-4.329A^{-0.0740}} P^{0.2729}$	.7870	.3157	65.80	−7.5
$Q_{50} = 10^{6.777-3.991A^{-0.0828}} P^{0.1706}$	.7670	.3345	73.88	−8.2
$Q_{100} = 10^{6.776-3.758A^{-0.0903}} P^{0.08212}$	.7458	.3548	82.98	−8.5

Table 9. Regression equations based on power transformation of drainage area using one predictor variable.

[Q_T, peak streamflow for T-year recurrence interval in cubic feet per second; and A, drainage area in square miles. The exponent of A is the power λ.]

Regression equation	Adjusted R-squared	Residual standard error $\log_{10}(Q_T)$	PRESS statistic	Percent change from PRESS in table 6
$Q_2 = 10^{8.280-6.031A^{-0.0465}}$	0.7710	0.3309	72.06	−2.9
$Q_5 = 10^{7.194-4.614A^{-0.0658}}$	.8030	.2994	59.00	−6.8
$Q_{10} = 10^{6.961-4.212A^{-0.0749}}$	.8002	.3021	60.10	−8.4
$Q_{25} = 10^{6.840-3.914A^{-0.0837}}$	.7834	.3184	66.77	−9.4
$Q_{50} = 10^{6.806-3.766A^{-0.0890}}$	.7659	.3354	74.08	−9.5
$Q_{100} = 10^{6.800-3.659A^{-0.0934}}$	.7462	.3545	82.78	−9.4

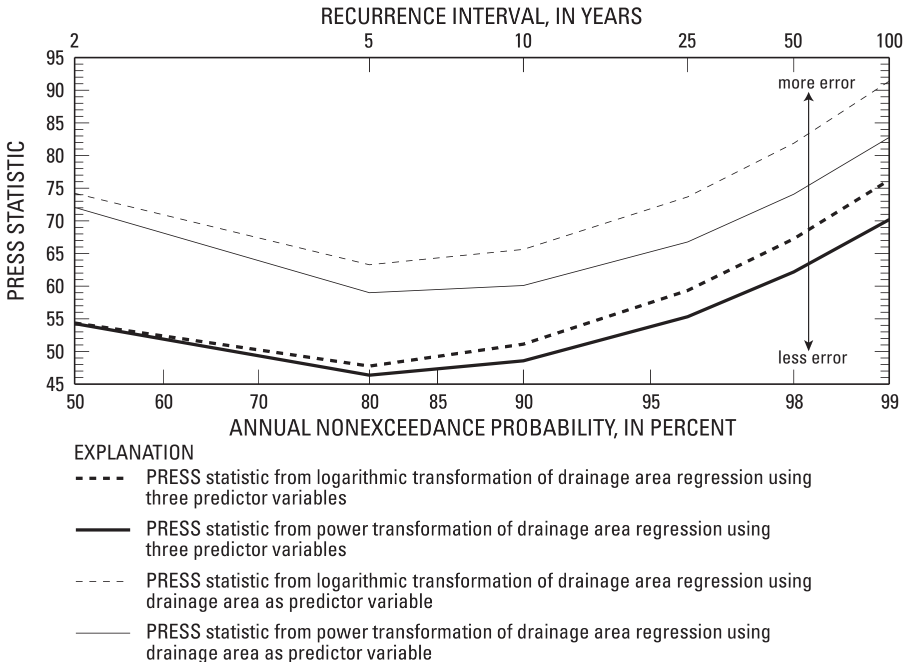


Figure 6. Comparison of PRESS statistics from regression analysis.

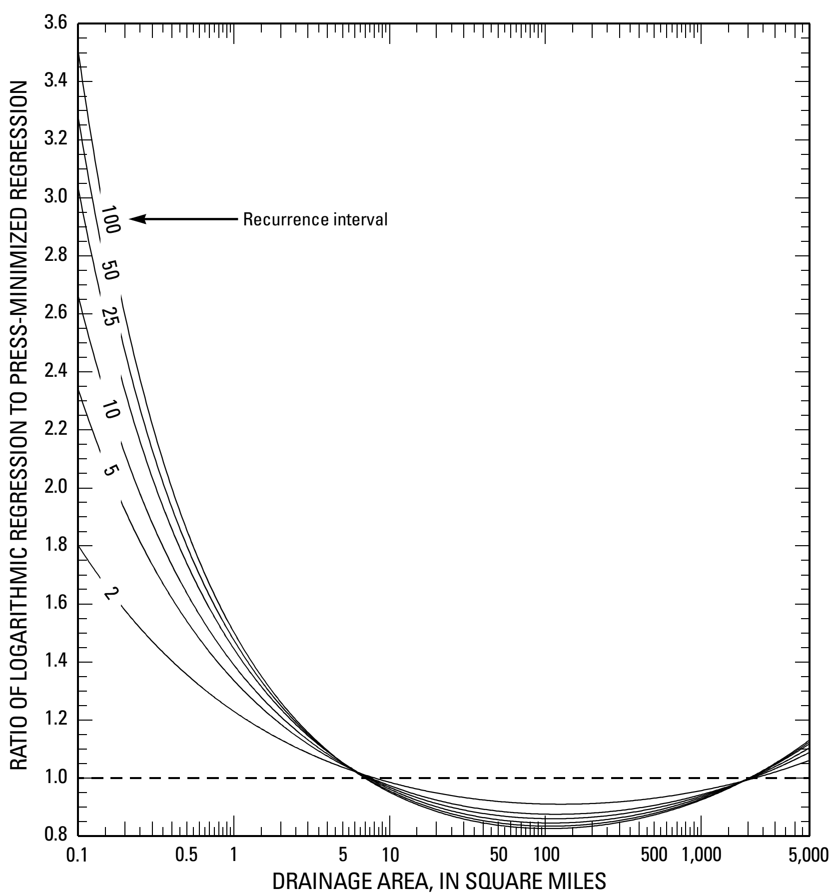


Figure 7. Relation between bias ratio and drainage area by recurrence interval for the regressions using drainage area as the only predictor variable.

Appendix 1—Computational Script using PRESS Minimization and Drainage Area, Mean Annual Precipitation, and Main-Channel Slope

Blank Page

```
R : Copyright 2006, The R Foundation for Statistical Computing
Version 2.3.0 (2006-04-24)
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> # In R, it appears necessary to ensure that the sum of the
> # weight factors equals the length of the weight factor
> # vector.  Otherwise, one artificially inflates the
> # residual standard error and hence prediction limits.  Other
> # diagnostics are changed, but remain in relative proportion
> # with one another--so influence conclusions remain the same?
> # Also, the coefficients of the regression are correct regardless
> # of the summation constraint needed on the weights.
> MLRweights <- function(vector) {
+   tmp = length(vector)/sum(vector)
+   return (tmp*vector)
+ }
>
> # PRESS statistics
> PRESS <- function(model) {
+   if(is.null(model$terms)) stop("invalid␣'lm'␣object:␣no␣terms")
+   sum( (weighted.residuals(model)/(1-hatvalues(model)))^2 )
+ }
>
>
> DATA <- read.csv("tx664.csv",header=T)
> attach(DATA)
> names(DATA)
 [1] "Station" "LatD"    "LatM"    "LatS"    "LonD"    "LonM"    "LonS"
 [8] "EqYrs"   "CDA"     "MAP"     "P224"    "Slope"   "Shape"   "Q2"
[15] "Q5"      "Q10"     "Q25"     "Q50"     "Q100"    "C2"      "C25"
[22] "C100"
> outliers <- c(212,323,358,602,614,620,628,637)
> CDA  <- CDA[-outliers]
> Q2   <- Q2[-outliers]
> Q5   <- Q5[-outliers]
> Q10  <- Q10[-outliers]
> Q25  <- Q25[-outliers]
> Q50  <- Q50[-outliers]
> Q100 <- Q100[-outliers]
> MAP  <- MAP[-outliers]
> Slope <- Slope[-outliers]
> WEIGHTS <- MLRweights(EqYrs[-outliers])
>
>
> WLS2_2.OUT   <- lm(Q2~CDA+MAP+Slope,  weights=WEIGHTS)
> WLS2_5.OUT   <- lm(Q5~CDA+MAP+Slope,  weights=WEIGHTS)
> WLS2_10.OUT  <- lm(Q10~CDA+MAP+Slope, weights=WEIGHTS)
> WLS2_25.OUT  <- lm(Q25~CDA+MAP+Slope, weights=WEIGHTS)
> WLS2_50.OUT  <- lm(Q50~CDA+MAP+Slope, weights=WEIGHTS)
> WLS2_100.OUT <- lm(Q100~CDA+MAP+Slope, weights=WEIGHTS)
>
> PRESS(WLS2_2.OUT)
[1] 54.36495
> summary(WLS2_2.OUT)

Call:
lm(formula = Q2 ~ CDA + MAP + Slope, weights = WEIGHTS)

Residuals:
      Min        1Q    Median        3Q       Max
-1.119685 -0.171447 -0.009884  0.188213  1.104813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) -0.52401    0.22148  -2.366   0.0183 *
CDA          0.65645    0.01762  37.250  < 2e-16 ***
MAP          1.47417    0.09926  14.851  < 2e-16 ***
Slope        0.35245    0.04981   7.076 3.84e-12 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2866 on 652 degrees of freedom
Multiple R-Squared: 0.829,      Adjusted R-squared: 0.8282
F-statistic:  1054 on 3 and 652 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_5.OUT)
[1] 47.75871
> summary(WLS2_5.OUT)

Call:
lm(formula = Q5 ~ CDA + MAP + Slope, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97622 -0.17392 -0.01896  0.14456  0.86452

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.22042    0.20756  -1.062    0.289
CDA          0.67897    0.01652  41.111   <2e-16 ***
MAP          1.37583    0.09303  14.790   <2e-16 ***
Slope        0.48282    0.04668  10.344   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2686 on 652 degrees of freedom
Multiple R-Squared: 0.8422,     Adjusted R-squared: 0.8414
F-statistic:  1160 on 3 and 652 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_10.OUT)
[1] 51.11918
> summary(WLS2_10.OUT)

Call:
lm(formula = Q10 ~ CDA + MAP + Slope, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.01418 -0.19171 -0.01982  0.13608  0.84651

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.04207    0.21472  -0.196    0.845
CDA          0.68963    0.01708  40.365   <2e-16 ***
MAP          1.31742    0.09623  13.690   <2e-16 ***
Slope        0.54210    0.04829  11.227   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2778 on 652 degrees of freedom
Multiple R-Squared: 0.8318,     Adjusted R-squared: 0.831
F-statistic:  1075 on 3 and 652 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_25.OUT)
[1] 59.34262
> summary(WLS2_25.OUT)

Call:
lm(formula = Q25 ~ CDA + MAP + Slope, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.03979 -0.20721 -0.02706  0.13507  1.02047

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.15008    0.23132   0.649    0.517
CDA          0.70048    0.01841  38.057   <2e-16 ***
MAP          1.25642    0.10368  12.119   <2e-16 ***
Slope        0.60053    0.05202  11.544   <2e-16 ***
---
```

```
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2993 on 652 degrees of freedom
Multiple R-Squared: 0.8095,     Adjusted R-squared: 0.8086
F-statistic: 923.4 on 3 and 652 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_50.OUT)
[1] 67.2532
> summary(WLS2_50.OUT)

Call:
lm(formula = Q50 ~ CDA + MAP + Slope, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09952 -0.22645 -0.03173  0.13983  1.09056

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.27484    0.24625   1.116    0.265
CDA          0.70729    0.01959  36.097   <2e-16 ***
MAP          1.21778    0.11037  11.034   <2e-16 ***
Slope        0.63587    0.05538  11.482   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3186 on 652 degrees of freedom
Multiple R-Squared: 0.7896,     Adjusted R-squared: 0.7887
F-statistic: 815.8 on 3 and 652 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_100.OUT)
[1] 76.25372
> summary(WLS2_100.OUT)

Call:
lm(formula = Q100 ~ CDA + MAP + Slope, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.16119 -0.23060 -0.03985  0.14572  1.20525

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.38788    0.26221   1.479    0.140
CDA          0.71333    0.02086  34.190   <2e-16 ***
MAP          1.18327    0.11752  10.069   <2e-16 ***
Slope        0.66604    0.05897  11.295   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3393 on 652 degrees of freedom
Multiple R-Squared: 0.7686,     Adjusted R-squared: 0.7675
F-statistic: 721.7 on 3 and 652 DF,  p-value: < 2.2e-16

>
>
> doQt <- function(Q,type) {
+   smallpress <- 10000
+   smallpower <- 10000
+   #Q <- 10^Q
+   for(power in seq(-.1,.1,by=0.0001)) { # 0.007, 0.08, by=0.0001
+     if(power == 0) next
+     power <- -1 * power
+     CDA1 <- 10^CDA
+     CDA1 <- CDA1^power
+     WLS.OUT <- lm(Q~CDA1+MAP+Slope, weights=WEIGHTS)
+     press <- PRESS(WLS.OUT)
+     if(press < smallpress) {
+       smallpress <- press
+       smallpower <- power
+     }
+     #plot(fitted(WLS.OUT),residuals(WLS.OUT),ylim=c(-2,2),col=2)
+     #sm <- lowess(fitted(WLS.OUT),y=residuals(WLS.OUT))
+     #lines(sm,lwd=2)
+     #lines(c(-10,10),c(0,0))
+   }
+   print(c(type,smallpower,smallpress))
+   return(c(type,smallpower,smallpress))
```

```
+ }
>
>
> vals2   <- doQt(Q2,2)
[1]  2.00000 -0.00820 54.26502
> vals5   <- doQt(Q5,5)
[1]  5.00000 -0.02990 46.36806
> vals10  <- doQt(Q10,10)
[1] 10.00000 -0.04000 48.56881
> vals25  <- doQt(Q25,25)
[1] 25.00000 -0.04970 55.31596
> vals50  <- doQt(Q50,50)
[1] 50.0000 -0.0553 62.1783
> vals100 <- doQt(Q100,100)
[1] 100.00000  -0.06010  70.18746
>
> vals2
[1]  2.00000 -0.00820 54.26502
> vals5
[1]  5.00000 -0.02990 46.36806
> vals10
[1] 10.00000 -0.04000 48.56881
> vals25
[1] 25.00000 -0.04970 55.31596
> vals50
[1] 50.0000 -0.0553 62.1783
> vals100
[1] 100.00000  -0.06010  70.18746
>
>
> finalQt <- function(Q,power,type) {
+   CDA1 <- 10^CDA
+   CDA1 <- CDA1^power
+   WLS.OUT <- lm(Q~CDA1+MAP+Slope, weights=WEIGHTS)
+   #plot(CDA,residuals(WLS.OUT),ylim=c(-3,3))
+   #plot(fitted(WLS.OUT),residuals(WLS.OUT),pch=16,col=2,ylim=c(-1.5,1))
+   #sm <- lowess(fitted(WLS.OUT),y=residuals(WLS.OUT))
+   #lines(sm,lwd=2)
+   #lines(c(-10,10),c(0,0))
+   cat(c("POWER:␣",power,"\n"))
+   print(summary(WLS.OUT))
+
+   W <- diag(WEIGHTS)
+   X = model.matrix(WLS.OUT)
+   Xt = t(X)
+
+   # Perform manual WLS regression and hat matrix
+   tmp <- chol2inv( chol( Xt %*% W %*% X ) )
+   wlshat1 <- X %*% tmp %*% Xt
+   print(tmp) # inverted covariance matrix
+   print(max(diag(wlshat1))) # the maximum leverage
+   m.wls.out <- tmp %*% Xt %*% W %*% Q
+   print(m.wls.out) # the regression coefficients
+
+   PRESS(WLS.OUT)
+   return(WLS.OUT)
+ }
>
>
> F2.OUT   <- finalQt(Q2,vals2[2],2)
POWER:  -0.0082

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
      Min        1Q    Median        3Q       Max
-1.134364 -0.169019 -0.008506  0.190237  1.097330

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.59867    0.78972  45.078  < 2e-16 ***
CDA1        -36.09370    0.96760 -37.302  < 2e-16 ***
MAP           1.44788    0.09866  14.676  < 2e-16 ***
Slope         0.34723    0.04963   6.996 6.53e-12 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2863 on 652 degrees of freedom
Multiple R-Squared: 0.8293,     Adjusted R-squared: 0.8286
```

```
F-statistic:  1056 on 3 and 652 DF,  p-value: < 2.2e-16

           [,1]       [,2]        [,3]        [,4]
[1,]  7.6093394 -9.2195944  0.55243205  0.36807256
[2,] -9.2195944 11.4232854 -0.80660576 -0.48359566
[3,]  0.5524321 -0.8066058  0.11875864  0.04243762
[4,]  0.3680726 -0.4835957  0.04243762  0.03005193
[1] 0.1053207
           [,1]
[1,]  35.5986711
[2,] -36.0936961
[3,]   1.4478828
[4,]   0.3472258
> F5.OUT   <- finalQt(Q5,vals5[2],5)
POWER:  -0.0299

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
      Min        1Q    Median        3Q       Max
-0.933747 -0.170028 -0.005652  0.142126  0.836051

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.16386    0.14688   76.01   <2e-16 ***
CDA1        -11.27975    0.26878  -41.97   <2e-16 ***
MAP           1.27941    0.08982   14.24   <2e-16 ***
Slope         0.46404    0.04544   10.21   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2646 on 652 degrees of freedom
Multiple R-Squared: 0.8468,     Adjusted R-squared: 0.8461
F-statistic:  1201 on 3 and 652 DF,  p-value: < 2.2e-16

            [,1]       [,2]        [,3]        [,4]
[1,]  0.30819904 -0.3799637 -0.01210666  0.03055440
[2,] -0.37996370  1.0321015 -0.23486075 -0.14338346
[3,] -0.01210666 -0.2348608  0.11524761  0.04091835
[4,]  0.03055440 -0.1433835  0.04091835  0.02949867
[1] 0.1047414
            [,1]
[1,]  11.1638574
[2,] -11.2797452
[3,]   1.2794083
[4,]   0.4640428
> F10.OUT  <- finalQt(Q10,vals10[2],10)
POWER:  -0.04

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
      Min        1Q    Median        3Q       Max
-0.949383 -0.169673 -0.007528  0.145502  0.782993

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.04660    0.11863   76.26   <2e-16 ***
CDA1        -8.94992    0.21392  -41.84   <2e-16 ***
MAP          1.18822    0.09121   13.03   <2e-16 ***
Slope        0.51724    0.04625   11.18   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2707 on 652 degrees of freedom
Multiple R-Squared: 0.8403,     Adjusted R-squared: 0.8396
F-statistic:  1143 on 3 and 652 DF,  p-value: < 2.2e-16

             [,1]       [,2]        [,3]         [,4]
[1,]  0.192009986 -0.1216301 -0.06356396 -0.000679897
[2,] -0.121630080  0.6244004 -0.17970515 -0.110637868
[3,] -0.063563963 -0.1797051  0.11352360  0.040132667
[4,] -0.000679897 -0.1106379  0.04013267  0.029183281
[1] 0.1041505
           [,1]
[1,]  9.0465969
[2,] -8.9499203
[3,]  1.1882238
```

```
[4,]  0.5172355
> F25.OUT  <- finalQt(Q25,vals25[2],25)
POWER:  -0.0497

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.01532 -0.18941 -0.01757  0.14541  0.94158

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.94932    0.11881   66.91   <2e-16 ***
CDA1        -7.62804    0.19053  -40.03   <2e-16 ***
MAP          1.09555    0.09660   11.34   <2e-16 ***
Slope        0.56992    0.04906   11.62   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2889 on 652 degrees of freedom
Multiple R-Squared: 0.8225,     Adjusted R-squared: 0.8217
F-statistic:  1007 on 3 and 652 DF,  p-value: < 2.2e-16

            [,1]        [,2]        [,3]        [,4]
[1,]  0.16917355 -0.01930396 -0.09202358 -0.01816880
[2,] -0.01930396  0.43507942 -0.14753723 -0.09156878
[3,] -0.09202358 -0.14753723  0.11183416  0.03934196
[4,] -0.01816880 -0.09156878  0.03934196  0.02885127
[1] 0.1034083
           [,1]
[1,]  7.9493152
[2,] -7.6280350
[3,]  1.0955540
[4,]  0.5699208
> F50.OUT  <- finalQt(Q50,vals50[2],50)
POWER:  -0.0553

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.10808 -0.20353 -0.02085  0.14901  1.00273

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.55371    0.12581   60.04   <2e-16 ***
CDA1        -7.08958    0.18527  -38.27   <2e-16 ***
MAP          1.03865    0.10195   10.19   <2e-16 ***
Slope        0.60210    0.05183   11.62   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3062 on 652 degrees of freedom
Multiple R-Squared: 0.8057,     Adjusted R-squared: 0.8048
F-statistic: 901.1 on 3 and 652 DF,  p-value: < 2.2e-16

            [,1]        [,2]        [,3]        [,4]
[1,]  0.16878887  0.01314134 -0.10338009 -0.02523111
[2,]  0.01314134  0.36601926 -0.13398330 -0.08354382
[3,] -0.10338009 -0.13398330  0.11084897  0.03887227
[4,] -0.02523111 -0.08354382  0.03887227  0.02864815
[1] 0.1029075
           [,1]
[1,]  7.5537081
[2,] -7.0895811
[3,]  1.0386544
[4,]  0.6021032
> F100.OUT <- finalQt(Q100,vals100[2],100)
POWER:  -0.0601

Call:
lm(formula = Q ~ CDA1 + MAP + Slope, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21129 -0.21871 -0.03113  0.15723  1.08687

Coefficients:
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.30661    0.13476  54.218   <2e-16 ***
CDA1        -6.71364    0.18421 -36.445   <2e-16 ***
MAP          0.98827    0.10790   9.159   <2e-16 ***
Slope        0.62951    0.05489  11.468   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3253 on 652 degrees of freedom
Multiple R-Squared: 0.7872,     Adjusted R-squared: 0.7862
F-statistic: 803.9 on 3 and 652 DF,  p-value: < 2.2e-16

             [,1]        [,2]        [,3]        [,4]
[1,]  0.17159632  0.03242555 -0.11114151 -0.03010189
[2,]  0.03242555  0.32062551 -0.12431136 -0.07782179
[3,] -0.11114151 -0.12431136  0.11000107  0.03846329
[4,] -0.03010189 -0.07782179  0.03846329  0.02846809
[1] 0.1024386
           [,1]
[1,]  7.3066130
[2,] -6.7136406
[3,]  0.9882687
[4,]  0.6295080
```

Blank Page

Appendix 2—Computational Script Using PRESS Minimization and Drainage Area and Mean Annual Precipitation

```
R : Copyright 2006, The R Foundation for Statistical Computing
Version 2.3.0 (2006-04-24)
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> # In R, it appears necessary to ensure that the sum of the
> # weight factors equals the length of the weight factor
> # vector.  Otherwise, one artificially inflates the
> # residual standard error and hence prediction limits.  Other
> # diagnostics are changed, but remain in relative proportion
> # with one another--so influence conclusions remain the same?
> # Also, the coefficients of the regression are correct regardless
> # of the summation constraint needed on the weights.
> MLRweights <- function(vector) {
+   tmp = length(vector)/sum(vector)
+   return (tmp*vector)
+ }
>
> # PRESS statistics
> PRESS <- function(model) {
+   if(is.null(model$terms)) stop("invalid␣'lm'␣object:␣no␣terms")
+   sum( (weighted.residuals(model)/(1-hatvalues(model)))^2 )
+ }
>
>
> DATA <- read.csv("tx664.csv",header=T)
> attach(DATA)
> names(DATA)
 [1] "Station" "LatD"    "LatM"    "LatS"    "LonD"    "LonM"    "LonS"
 [8] "EqYrs"   "CDA"     "MAP"     "P224"    "Slope"   "Shape"   "Q2"
[15] "Q5"      "Q10"     "Q25"     "Q50"     "Q100"    "C2"      "C25"
[22] "C100"
> outliers <- c(212,323,358,602,614,620,628,637)
> CDA  <- CDA[-outliers]
> Q2   <- Q2[-outliers]
> Q5   <- Q5[-outliers]
> Q10  <- Q10[-outliers]
> Q25  <- Q25[-outliers]
> Q50  <- Q50[-outliers]
> Q100 <- Q100[-outliers]
> MAP  <- MAP[-outliers]
> Slope <- Slope[-outliers]
> WEIGHTS <- MLRweights(EqYrs[-outliers])
>
>
> WLS2_2.OUT   <- lm(Q2~CDA+MAP,  weights=WEIGHTS)
> WLS2_5.OUT   <- lm(Q5~CDA+MAP,  weights=WEIGHTS)
> WLS2_10.OUT  <- lm(Q10~CDA+MAP, weights=WEIGHTS)
> WLS2_25.OUT  <- lm(Q25~CDA+MAP, weights=WEIGHTS)
> WLS2_50.OUT  <- lm(Q50~CDA+MAP, weights=WEIGHTS)
> WLS2_100.OUT <- lm(Q100~CDA+MAP, weights=WEIGHTS)
>
> PRESS(WLS2_2.OUT)
[1] 58.28556
> summary(WLS2_2.OUT)

Call:
lm(formula = Q2 ~ CDA + MAP, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09718 -0.18756 -0.01553  0.18068  1.10729

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  0.83297    0.11490    7.25 1.19e-12 ***
CDA          0.55340    0.01029   53.78  < 2e-16 ***
MAP          0.97323    0.07215   13.49  < 2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2971 on 653 degrees of freedom
Multiple R-Squared: 0.8159,     Adjusted R-squared: 0.8153
F-statistic:  1447 on 2 and 653 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_5.OUT)
[1] 55.31802
> summary(WLS2_5.OUT)

Call:
lm(formula = Q5 ~ CDA + MAP, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86588 -0.20147 -0.02265  0.14310  0.86740

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.63852    0.11196  14.635   <2e-16 ***
CDA          0.53779    0.01003  53.638   <2e-16 ***
MAP          0.68959    0.07030   9.809   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2895 on 653 degrees of freedom
Multiple R-Squared: 0.8163,     Adjusted R-squared: 0.8157
F-statistic:  1451 on 2 and 653 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_10.OUT)
[1] 60.68633
> summary(WLS2_10.OUT)

Call:
lm(formula = Q10 ~ CDA + MAP, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91736 -0.22632 -0.02745  0.14438  1.00583

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.04511    0.11726  17.441  < 2e-16 ***
CDA          0.53112    0.01050  50.577  < 2e-16 ***
MAP          0.54692    0.07363   7.428 3.47e-13 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3032 on 653 degrees of freedom
Multiple R-Squared: 0.7993,     Adjusted R-squared: 0.7987
F-statistic:  1300 on 2 and 653 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_25.OUT)
[1] 71.10338
> summary(WLS2_25.OUT)

Call:
lm(formula = Q25 ~ CDA + MAP, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-0.97147 -0.24335 -0.03816  0.13714  1.17710

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.46222    0.12692  19.400  < 2e-16 ***
CDA          0.52489    0.01137  46.181  < 2e-16 ***
MAP          0.40288    0.07970   5.055 5.59e-07 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3282 on 653 degrees of freedom
Multiple R-Squared: 0.7705,     Adjusted R-squared: 0.7698
```

```
F-statistic:  1096 on 2 and 653 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_50.OUT)
[1] 80.44262
> summary(WLS2_50.OUT)

Call:
lm(formula = Q50 ~ CDA + MAP, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00636 -0.26998 -0.04396  0.13322  1.28007

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.72306    0.13499  20.173  < 2e-16 ***
CDA          0.52137    0.01209  43.130  < 2e-16 ***
MAP          0.31401    0.08476   3.705 0.000230 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3491 on 653 degrees of freedom
Multiple R-Squared: 0.7471,     Adjusted R-squared: 0.7463
F-statistic: 964.5 on 2 and 653 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_100.OUT)
[1] 90.72161
> summary(WLS2_100.OUT)

Call:
lm(formula = Q100 ~ CDA + MAP, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.03776 -0.28890 -0.05912  0.15130  1.41052

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.95226    0.14334  20.596  < 2e-16 ***
CDA          0.51858    0.01284  40.398  < 2e-16 ***
MAP          0.23661    0.09001   2.629  0.00877 **
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3707 on 653 degrees of freedom
Multiple R-Squared: 0.7233,     Adjusted R-squared: 0.7224
F-statistic: 853.4 on 2 and 653 DF,  p-value: < 2.2e-16

>
>
> doQt <- function(Q,type) {
+   smallpress <- 10000
+   smallpower <- 10000
+   #Q <- 10^Q
+   for(power in seq(-.1,.1,by=0.0001)) { # 0.007, 0.08, by=0.0001
+     if(power == 0) next
+     power <- -1 * power
+     CDA1 <- 10^CDA
+     CDA1 <- CDA1^power
+     WLS.OUT <- lm(Q~CDA1+MAP, weights=WEIGHTS)
+     press <- PRESS(WLS.OUT)
+     if(press < smallpress) {
+       smallpress <- press
+       smallpower <- power
+     }
+     #plot(fitted(WLS.OUT),residuals(WLS.OUT),ylim=c(-2,2),col=2)
+     #sm <- lowess(fitted(WLS.OUT),y=residuals(WLS.OUT))
+     #lines(sm,lwd=2)
+     #lines(c(-10,10),c(0,0))
+   }
+   print(c(type,smallpower,smallpress))
+   return(c(type,smallpower,smallpress))
+ }
>
>
> vals2   <- doQt(Q2,2)
[1]  2.00000 -0.01570 58.02393
> vals5   <- doQt(Q5,5)
```

```
[1]   5.00000  -0.04510 53.26502
> vals10  <- doQt(Q10,10)
[1] 10.00000  -0.05960 57.17684
> vals25  <- doQt(Q25,25)
[1] 25.00000  -0.07400 65.79596
> vals50  <- doQt(Q50,50)
[1] 50.00000  -0.08280 73.87932
> vals100 <- doQt(Q100,100)
[1] 100.00000   -0.09030  82.98081
>
> vals2
[1]   2.00000  -0.01570 58.02393
> vals5
[1]   5.00000  -0.04510 53.26502
> vals10
[1] 10.00000  -0.05960 57.17684
> vals25
[1] 25.00000  -0.07400 65.79596
> vals50
[1] 50.00000  -0.08280 73.87932
> vals100
[1] 100.00000   -0.09030  82.98081
>
>
> finalQt <- function(Q,power,type) {
+    CDA1 <- 10^CDA
+    CDA1 <- CDA1^power
+    WLS.OUT <- lm(Q~CDA1+MAP, weights=WEIGHTS)
+    #plot(CDA,residuals(WLS.OUT),ylim=c(-3,3))
+    #plot(fitted(WLS.OUT),residuals(WLS.OUT),pch=16,col=2,ylim=c(-1.5,1))
+    #sm <- lowess(fitted(WLS.OUT),y=residuals(WLS.OUT))
+    #lines(sm,lwd=2)
+    #lines(c(-10,10),c(0,0))
+    cat(c("POWER:␣",power,"\n"))
+    print(summary(WLS.OUT))
+
+    W <- diag(WEIGHTS)
+    X = model.matrix(WLS.OUT)
+    Xt = t(X)
+
+    # Perform manual WLS regression and hat matrix
+    tmp <- chol2inv( chol( Xt %*% W %*% X ) )
+    wlshat1 <- X %*% tmp %*% Xt
+    print(tmp) # inverted covariance matrix
+    print(max(diag(wlshat1))) # maximum leverage
+    m.wls.out <- tmp %*% Xt %*% W %*% Q
+    print(m.wls.out) # the regression coefficients
+
+    PRESS(WLS.OUT)
+    return(WLS.OUT)
+ }
>
>
> F2.OUT    <- finalQt(Q2,vals2[2],2)
POWER:  -0.0157

Call:
lm(formula = Q ~ CDA1 + MAP, weights = WEIGHTS)

Residuals:
     Min        1Q    Median        3Q       Max
-1.11654  -0.18172  -0.01350  0.18093  1.09514

Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)  17.35866    0.27228    63.75    <2e-16 ***
CDA1        -16.51148    0.30614   -53.93    <2e-16 ***
MAP           0.94286    0.07183    13.13    <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2965 on 653 degrees of freedom
Multiple R-Squared: 0.8167,      Adjusted R-squared: 0.8162
F-statistic:  1455 on 2 and 653 DF,  p-value: < 2.2e-16

             [,1]         [,2]         [,3]
[1,]  0.84355557  -0.88041997  -0.02485273
[2,] -0.88041997   1.06646053  -0.06598735
[3,] -0.02485273  -0.06598735   0.05871136
[1] 0.02790216
```

```
            [,1]
[1,]  17.3586632
[2,] -16.5114820
[3,]   0.9428574
> F5.OUT   <- finalQt(Q5,vals5[2],5)
POWER:  -0.0451

Call:
lm(formula = Q ~ CDA1 + MAP, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91606 -0.20427 -0.02218  0.14384  0.83370

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.07960    0.11743  68.806   <2e-16 ***
CDA1        -6.40329    0.11665 -54.894   <2e-16 ***
MAP          0.60654    0.06857   8.846   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2841 on 653 degrees of freedom
Multiple R-Squared: 0.8231,     Adjusted R-squared: 0.8226
F-statistic:  1519 on 2 and 653 DF,  p-value: < 2.2e-16

            [,1]        [,2]        [,3]
[1,]  0.17084538 -0.09552241 -0.06531660
[2,] -0.09552241  0.16858758 -0.02473015
[3,] -0.06531660 -0.02473015  0.05825606
[1] 0.0277843
           [,1]
[1,]  8.0796013
[2,] -6.4032928
[3,]  0.6065365
> F10.OUT  <- finalQt(Q10,vals10[2],10)
POWER:  -0.0596

Call:
lm(formula = Q ~ CDA1 + MAP, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98237 -0.22307 -0.01623  0.14879  0.93836

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.19977    0.11028  65.284  < 2e-16 ***
CDA1        -5.10664    0.09729 -52.490  < 2e-16 ***
MAP          0.43974    0.07091   6.201 9.94e-10 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2943 on 653 degrees of freedom
Multiple R-Squared: 0.8109,     Adjusted R-squared: 0.8103
F-statistic:  1400 on 2 and 653 DF,  p-value: < 2.2e-16

            [,1]        [,2]        [,3]
[1,]  0.14038398 -0.05084317 -0.07034334
[2,] -0.05084317  0.10924813 -0.01930732
[3,] -0.07034334 -0.01930732  0.05804055
[1] 0.02768903
          [,1]
[1,]  7.199766
[2,] -5.106641
[3,]  0.439738
> F25.OUT  <- finalQt(Q25,vals25[2],25)
POWER:  -0.074

Call:
lm(formula = Q ~ CDA1 + MAP, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05059 -0.23898 -0.02263  0.15452  1.09630

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.84851    0.11304  60.586  < 2e-16 ***
CDA1        -4.32907    0.08916 -48.551  < 2e-16 ***
```

```
MAP           0.27291    0.07593   3.594  0.00035 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3157 on 653 degrees of freedom
Multiple R-Squared: 0.7877,     Adjusted R-squared: 0.787
F-statistic:  1211 on 2 and 653 DF,  p-value: < 2.2e-16

            [,1]        [,2]        [,3]
[1,]  0.12817981 -0.03022921 -0.07326873
[2,] -0.03022921  0.07975385 -0.01598832
[3,] -0.07326873 -0.01598832  0.05783358
[1] 0.02757124
          [,1]
[1,]  6.8485138
[2,] -4.3290741
[3,]  0.2729137
> F50.OUT  <- finalQt(Q50,vals50[2],50)
POWER:  -0.0828

Call:
lm(formula = Q ~ CDA1 + MAP, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11361 -0.24415 -0.02832  0.14935  1.19162

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.77654    0.11793  57.461   <2e-16 ***
CDA1        -3.99103    0.08744 -45.643   <2e-16 ***
MAP          0.17064    0.08037   2.123   0.0341 *
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3345 on 653 degrees of freedom
Multiple R-Squared: 0.7677,     Adjusted R-squared: 0.767
F-statistic:  1079 on 2 and 653 DF,  p-value: < 2.2e-16

            [,1]        [,2]        [,3]
[1,]  0.12426816 -0.02270511 -0.07450586
[2,] -0.02270511  0.06831574 -0.01451140
[3,] -0.07450586 -0.01451140  0.05771085
[1] 0.02748834
          [,1]
[1,]  6.7765427
[2,] -3.9910263
[3,]  0.1706351
> F100.OUT <- finalQt(Q100,vals100[2],100)
POWER:  -0.0903

Call:
lm(formula = Q ~ CDA1 + MAP, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18053 -0.25296 -0.04201  0.15569  1.27209

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.77645    0.12386  54.711   <2e-16 ***
CDA1        -3.75755    0.08748 -42.953   <2e-16 ***
MAP          0.08212    0.08509   0.965    0.335
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3545 on 653 degrees of freedom
Multiple R-Squared: 0.7469,     Adjusted R-squared: 0.7461
F-statistic: 963.4 on 2 and 653 DF,  p-value: < 2.2e-16

            [,1]        [,2]        [,3]
[1,]  0.12205140 -0.01801320 -0.07534347
[2,] -0.01801320  0.06088392 -0.01347027
[3,] -0.07534347 -0.01347027  0.05760862
[1] 0.02741139
            [,1]
[1,]  6.77645185
[2,] -3.75754823
[3,]  0.08211505
```

Appendix 3—Computational Script Using PRESS Minimization and Drainage Area

Blank Page

```
R : Copyright 2006, The R Foundation for Statistical Computing
Version 2.3.0 (2006-04-24)
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> # In R, it appears necessary to ensure that the sum of the
> # weight factors equals the length of the weight factor
> # vector.  Otherwise, one artificially inflates the
> # residual standard error and hence prediction limits.  Other
> # diagnostics are changed, but remain in relative proportion
> # with one another--so influence conclusions remain the same?
> # Also, the coefficients of the regression are correct regardless
> # of the summation constraint needed on the weights.
> MLRweights <- function(vector) {
+   tmp = length(vector)/sum(vector)
+   return (tmp*vector)
+ }
>
> # PRESS statistics
> PRESS <- function(model) {
+   if(is.null(model$terms)) stop("invalid␣'lm'␣object:␣no␣terms")
+   sum( (weighted.residuals(model)/(1-hatvalues(model)))^2 )
+ }
>
>
> DATA <- read.csv("tx664.csv",header=T)
> attach(DATA)
> names(DATA)
 [1] "Station" "LatD"    "LatM"    "LatS"    "LonD"    "LonM"    "LonS"
 [8] "EqYrs"   "CDA"     "MAP"     "P224"    "Slope"   "Shape"   "Q2"
[15] "Q5"      "Q10"     "Q25"     "Q50"     "Q100"    "C2"      "C25"
[22] "C100"
> outliers <- c(212,323,358,602,614,620,628,637)
> CDA  <- CDA[-outliers]
> Q2   <- Q2[-outliers]
> Q5   <- Q5[-outliers]
> Q10  <- Q10[-outliers]
> Q25  <- Q25[-outliers]
> Q50  <- Q50[-outliers]
> Q100 <- Q100[-outliers]
> MAP  <- MAP[-outliers]
> Slope <- Slope[-outliers]
> WEIGHTS <- MLRweights(EqYrs[-outliers])
>
>
> WLS2_2.OUT   <- lm(Q2~CDA,  weights=WEIGHTS)
> WLS2_5.OUT   <- lm(Q5~CDA,  weights=WEIGHTS)
> WLS2_10.OUT  <- lm(Q10~CDA, weights=WEIGHTS)
> WLS2_25.OUT  <- lm(Q25~CDA, weights=WEIGHTS)
> WLS2_50.OUT  <- lm(Q50~CDA, weights=WEIGHTS)
> WLS2_100.OUT <- lm(Q100~CDA, weights=WEIGHTS)
>
> PRESS(WLS2_2.OUT)
[1] 74.22438
> summary(WLS2_2.OUT)

Call:
lm(formula = Q2 ~ CDA, weights = WEIGHTS)

Residuals:
      Min        1Q    Median        3Q       Max
-1.391650 -0.189314  0.009216  0.194396  1.339501

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept)  2.33920    0.03059   76.47   <2e-16 ***
CDA          0.51577    0.01119   46.09   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3357 on 654 degrees of freedom
Multiple R-Squared: 0.7646,     Adjusted R-squared: 0.7642
F-statistic:  2124 on 1 and 654 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_5.OUT)
[1] 63.27671
> summary(WLS2_5.OUT)

Call:
lm(formula = Q5 ~ CDA, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.05357 -0.20297 -0.03195  0.16029  1.03193

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.70576    0.02824   95.82   <2e-16 ***
CDA          0.51113    0.01033   49.48   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3099 on 654 degrees of freedom
Multiple R-Squared: 0.7892,     Adjusted R-squared: 0.7889
F-statistic:  2449 on 1 and 654 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_10.OUT)
[1] 65.62669
> summary(WLS2_10.OUT)

Call:
lm(formula = Q10 ~ CDA, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99571 -0.22175 -0.03330  0.15717  0.95890

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.89156    0.02875  100.57   <2e-16 ***
CDA          0.50998    0.01052   48.48   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3156 on 654 degrees of freedom
Multiple R-Squared: 0.7823,     Adjusted R-squared: 0.782
F-statistic:  2351 on 1 and 654 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_25.OUT)
[1] 73.67088
> summary(WLS2_25.OUT)

Call:
lm(formula = Q25 ~ CDA, weights = WEIGHTS)

Residuals:
    Min      1Q  Median      3Q     Max
-1.0580 -0.2404 -0.0421  0.1519  1.1425

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.08574    0.03046   101.3   <2e-16 ***
CDA          0.50931    0.01114    45.7   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3343 on 654 degrees of freedom
Multiple R-Squared: 0.7616,     Adjusted R-squared: 0.7612
F-statistic:  2089 on 1 and 654 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_50.OUT)
```

```
[1] 81.88232
> summary(WLS2_50.OUT)

Call:
lm(formula = Q50 ~ CDA, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09259 -0.27013 -0.04421  0.14452  1.25313

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.20903    0.03212   99.92   <2e-16 ***
CDA          0.50923    0.01175   43.34   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3525 on 654 degrees of freedom
Multiple R-Squared: 0.7418,     Adjusted R-squared: 0.7414
F-statistic:  1879 on 1 and 654 DF,  p-value: < 2.2e-16

>
> PRESS(WLS2_100.OUT)
[1] 91.3865
> summary(WLS2_100.OUT)

Call:
lm(formula = Q100 ~ CDA, weights = WEIGHTS)

Residuals:
    Min      1Q  Median      3Q     Max
-1.1200 -0.2832 -0.0552  0.1460  1.3960

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.31846    0.03393   97.81   <2e-16 ***
CDA          0.50944    0.01241   41.05   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3724 on 654 degrees of freedom
Multiple R-Squared: 0.7204,     Adjusted R-squared: 0.7199
F-statistic:  1685 on 1 and 654 DF,  p-value: < 2.2e-16

>
>
> doQt <- function(Q,type) {
+   smallpress <- 10000
+   smallpower <- 10000
+   #Q <- 10^Q
+   for(power in seq(-.1,.1,by=0.0001)) { # 0.007, 0.08, by=0.0001
+     if(power == 0) next
+     power <- -1 * power
+     CDA1 <- 10^CDA
+     CDA1 <- CDA1^power
+     WLS.OUT <- lm(Q~CDA1, weights=WEIGHTS)
+     press <- PRESS(WLS.OUT)
+     if(press < smallpress) {
+       smallpress <- press
+       smallpower <- power
+     }
+     #plot(fitted(WLS.OUT),residuals(WLS.OUT),ylim=c(-2,2),col=2)
+     #sm <- lowess(fitted(WLS.OUT),y=residuals(WLS.OUT))
+     #lines(sm,lwd=2)
+     #lines(c(-10,10),c(0,0))
+   }
+   print(c(type,smallpower,smallpress))
+   return(c(type,smallpower,smallpress))
+ }
>
>
> vals2   <- doQt(Q2,2)
[1]  2.00000 -0.04650 72.06471
> vals5   <- doQt(Q5,5)
[1]  5.00000 -0.06580 59.00068
> vals10  <- doQt(Q10,10)
[1] 10.00000 -0.07490 60.09731
> vals25  <- doQt(Q25,25)
[1] 25.0000 -0.0837 66.7667
> vals50  <- doQt(Q50,50)
```

```
[1] 50.0000 -0.0890 74.0839
> vals100 <- doQt(Q100,100)
[1] 100.00000  -0.09340  82.78375
>
> vals2
[1]  2.00000 -0.04650 72.06471
> vals5
[1]  5.00000 -0.06580 59.00068
> vals10
[1] 10.00000 -0.07490 60.09731
> vals25
[1] 25.0000 -0.0837 66.7667
> vals50
[1] 50.0000 -0.0890 74.0839
> vals100
[1] 100.00000  -0.09340  82.78375
>
>
> finalQt <- function(Q,power,type) {
+   CDA1 <- 10^CDA
+   CDA1 <- CDA1^power
+   WLS.OUT <- lm(Q~CDA1, weights=WEIGHTS)
+   #plot(CDA,residuals(WLS.OUT),ylim=c(-3,3))
+   #plot(fitted(WLS.OUT),residuals(WLS.OUT),pch=16,col=2,ylim=c(-1.5,1))
+   #sm <- lowess(fitted(WLS.OUT),y=residuals(WLS.OUT))
+   #lines(sm,lwd=2)
+   #lines(c(-10,10),c(0,0))
+   cat(c("POWER:␣",power,"\n"))
+   print(summary(WLS.OUT))
+
+   W <- diag(WEIGHTS)
+   X = model.matrix(WLS.OUT)
+   Xt = t(X)
+
+   # Perform manual WLS regression and hat matrix
+   tmp <- chol2inv( chol( Xt %*% W %*% X ) )
+   wlshat1 <- X %*% tmp %*% Xt
+   print(tmp) # inverted covariance matrix
+   print(max(diag(wlshat1))) # maximum leverage
+   m.wls.out <- tmp %*% Xt %*% W %*% Q
+   print(m.wls.out) # the regression coefficients
+
+   PRESS(WLS.OUT)
+   return(WLS.OUT)
+ }
>
>
> F2.OUT   <- finalQt(Q2,vals2[2],2)
POWER:  -0.0465

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.43150 -0.18074  0.01207  0.20405  1.28629

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.2799     0.1002   82.63   <2e-16 ***
CDA1         -6.0308     0.1284  -46.97   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3309 on 654 degrees of freedom
Multiple R-Squared: 0.7713,     Adjusted R-squared: 0.771
F-statistic:  2206 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]       [,2]
[1,]  0.09169293 -0.1165254
[2,] -0.11652540  0.1505865
[1] 0.01885060
          [,1]
[1,]  8.279869
[2,] -6.030849
> F5.OUT   <- finalQt(Q5,vals5[2],5)
POWER:  -0.0658

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)
```

```
Residuals:
      Min         1Q    Median        3Q       Max
-1.064167 -0.195349  0.003766  0.160790  0.957834

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.19379    0.06350  113.29   <2e-16 ***
CDA1        -4.61402    0.08928  -51.68   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.2994 on 654 degrees of freedom
Multiple R-Squared: 0.8033,     Adjusted R-squared: 0.803
F-statistic:  2671 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]        [,2]
[1,]  0.04499556 -0.06218392
[2,] -0.06218392  0.08895185
[1] 0.02071799
          [,1]
[1,]  7.193790
[2,] -4.614025
> F10.OUT  <- finalQt(Q10,vals10[2],10)
POWER:  -0.0749

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)

Residuals:
      Min         1Q    Median        3Q       Max
-1.069176 -0.212532 -0.002709  0.158296  0.889868

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.96115    0.05610  124.08   <2e-16 ***
CDA1        -4.21241    0.08222  -51.23   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3021 on 654 degrees of freedom
Multiple R-Squared: 0.8005,     Adjusted R-squared: 0.8002
F-statistic:  2625 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]        [,2]
[1,]  0.03449016 -0.04941901
[2,] -0.04941901  0.07408408
[1] 0.02164636
          [,1]
[1,]  6.961146
[2,] -4.212410
> F25.OUT  <- finalQt(Q25,vals25[2],25)
POWER:  -0.0837

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.14066 -0.23002 -0.01563  0.16378  1.06665

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.84023    0.05276  129.64   <2e-16 ***
CDA1        -3.91357    0.08038  -48.69   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3184 on 654 degrees of freedom
Multiple R-Squared: 0.7838,     Adjusted R-squared: 0.7834
F-statistic:  2370 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]        [,2]
[1,]  0.02746451 -0.04066401
[2,] -0.04066401  0.06374535
[1] 0.02257363
          [,1]
[1,]  6.840234
[2,] -3.913570
> F50.OUT  <- finalQt(Q50,vals50[2],50)
```

```
POWER:  -0.089

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.18085 -0.24244 -0.02337  0.16170  1.17321

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.80600    0.05219   130.4   <2e-16 ***
CDA1        -3.76582    0.08133   -46.3   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3354 on 654 degrees of freedom
Multiple R-Squared: 0.7662,     Adjusted R-squared: 0.7659
F-statistic:  2144 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]        [,2]
[1,]  0.02422098 -0.03653889
[2,] -0.03653889  0.05882340
[1] 0.02314619
          [,1]
[1,]  6.806003
[2,] -3.765824
> F100.OUT <- finalQt(Q100,vals100[2],100)
POWER:  -0.0934

Call:
lm(formula = Q ~ CDA1, weights = WEIGHTS)

Residuals:
     Min       1Q   Median       3Q      Max
-1.21299 -0.25175 -0.03704  0.15891  1.26325

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.80019    0.05252  129.49   <2e-16 ***
CDA1        -3.65879    0.08336  -43.89   <2e-16 ***
---
Signif. codes:  0 "***", 0.001 "**", 0.01 "*", 0.05 ".", 0.1 "␣", 1

Residual standard error: 0.3545 on 654 degrees of freedom
Multiple R-Squared: 0.7465,     Adjusted R-squared: 0.7462
F-statistic:  1926 on 1 and 654 DF,  p-value: < 2.2e-16

            [,1]        [,2]
[1,]  0.02194667 -0.03360600
[2,] -0.03360600  0.05530056
[1] 0.0236296
          [,1]
[1,]  6.800189
[2,] -3.658793
```

Publishing support provided by USGS Lafayette Publishing Service Center

Information regarding water resources in Texas is available at
http: //tx.usgs.gov/

Blank Page