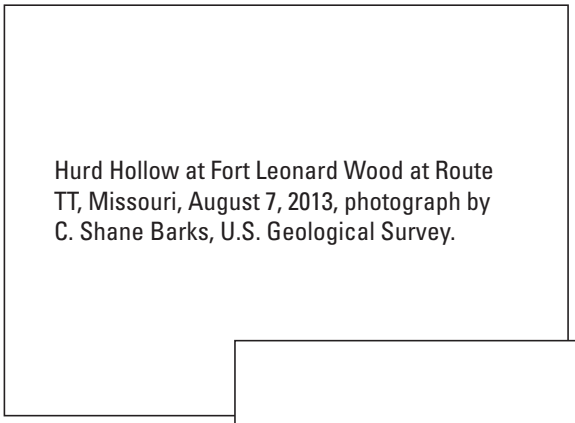
USGS
science for a changing world

Prepared in cooperation with the Missouri Department of Transportation and Federal Emergency Management Agency

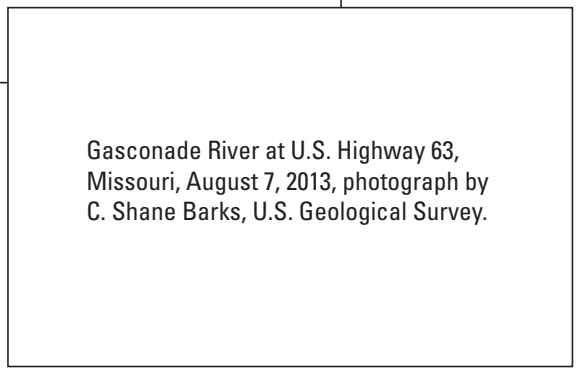
Methods for Estimating Annual Exceedance-Probability Discharges and Largest Recorded Floods for Unregulated Streams in Rural Missouri

Scientific Investigations Report 2014–5165

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



Hurd Hollow at Fort Leonard Wood at Route TT, Missouri, August 7, 2013, photograph by C. Shane Barks, U.S. Geological Survey.



Gasconade River at U.S. Highway 63, Missouri, August 7, 2013, photograph by C. Shane Barks, U.S. Geological Survey.

Back cover photograph: Gasconade River at Jerome, Missouri, August 8, 2013, photograph by C. Shane Barks, U.S. Geological Survey.

Methods for Estimating Annual Exceedance-Probability Discharges and Largest Recorded Floods for Unregulated Streams in Rural Missouri

By Rodney E. Southard and Andrea G. Veilleux

Prepared in cooperation with the Missouri Department of Transportation and Federal Emergency Management Agency

Scientific Investigations Report 2014–5165

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

U.S. Department of the Interior
SALLY JEWELL, Secretary

U.S. Geological Survey
Suzette M. Kimball, Acting Director

U.S. Geological Survey, Reston, Virginia: 2014

Suggested citation:
Southard, R.E., and Veilleux, A.G., 2014, Methods for estimating annual exceedance-probability discharges and largest recorded floods for unregulated streams in rural Missouri: U.S. Geological Survey Scientific Investigations Report 2014–5165, 39 p., *http://dx.doi.org/10.3133/sir20145165*.

2328-031X (print)
2328-0328 (online)
ISBN 978-1-4113-3845-6

Contents

Abstract 1
Introduction....1
Purpose and Scope 2
Description of Study Area 2
Previous Studies 5
Methods for Data Development for Streamgages 5
Site Selection....5
Annual Exceedance-Probability Analyses 5
Regional Skew Analysis 6
Expected Moments Algorithm Analysis 6
Multiple Grubbs-Beck Test for Detecting Potentially Influential Low Floods....7
Basin Characteristics....8
Regional Regression Analyses to Estimate Annual Exceedance-Probability Discharges....9
Definition of Flood Regions 9
Development of Regional Regression Equations 9
Ordinary-Least-Squares Regression....10
Generalized-Least-Squares Regression....10
Final Regional Regression Equations 11
Application and Limitations of Regression Equations 16
Streamgage Locations....16
Drainage-Area Ratio 20
Regional Regression Equations....20
Largest Recorded Floods in Missouri....20
Summary....24
References Cited....25
Appendix—Introduction to Statistical Analysis of Regional Skew....29

Figures

1. Map showing location of streamgages and hydrologic region boundaries in Missouri and in neighboring States of Missouri 3
2. Map showing regions used to determine annual exceedance-probability discharges and physiographic provinces in Missouri 4
3. Screen capture of the weighted-multiple-linear regression program (WREG) smoothing function for generalized-least-squares (GLS) correlation of annual peak flows as a function of the distance between 135 streamgages in Region 2 with 40 years of concurrent peak-flow record 11
4. Graphs showing relation between 1-percent annual exceedance-probability discharges computed from at-site streamflow to those predicted from generalized-least squares regression equations for flood regions in Missouri....14
5. Graph showing relation of Regions 1, 2, and 3 for the 1-percent ($Q_{1\%}$) annual exceedance-probability discharge using a factor of eight for basin shape and drainage areas from 3 to 2,100 square miles in the regional regression equations....15

6. Graph showing difference in at-site flood frequency estimates from Alexander and Wilson (1995; 17B method) and using the Expected Moments Algorithm with multiple Grubbs-Beck test method at USGS streamgage 07066000 (map number 241) 16
7. Graphs showing relation of basin characteristics to residuals from regression analyses for each region for the 1-percent ($Q_{1\%}$) annual exceedance-probability discharge 17
8. Graph showing percent difference by drainage area between 1-percent annual exceedance-probability estimates computed using regional regression equations developed in this study to those developed in Alexander and Wilson (1995) for 278 streamgages used in this study....19
9. Graphs showing relation between largest peak flow and drainage area (DRNAREA) for streams in Region 1, Region 2, and Region 3 22
10. Graph showing ratio of maximum annual peak flow to the largest direct or indirect discharge measurement for 66 streamgages used in this study that are currently active (2012) in Missouri....23
11. Graph showing largest floods in Missouri based on U.S. Geological Survey peak-flow data, water years 1915 through 2008 24
1–1. Graph showing relation between Fisher Z transformed (Z) cross-correlation of logs of annual peak discharge and distance (D) between basin centroids for 651 station-pairs with concurrent record lengths greater than or equal to 60 years from 42 streamgages in Missouri and neighboring States....35
1–2. Graph showing relation between untransformed cross-correlation of logs of annual peak discharge and distance, (D), between basin centroids based for 651 station-pairs with concurrent record lengths greater than or equal to 60 years from 42 streamgages in Missouri and neighboring States....37

Tables

1 Description of streamgages located in Missouri and selected streamgages in neighboring States of Missouri that were evaluated for use in the regional frequency regressions for Missouri 5
2. Basin characteristics of streamgages in Missouri and selected streamgages in neighboring States of Missouri 8
3. Flood-frequency statistics for annual peak flow data from 278 U.S. Geological Survey streamgages in Missouri and selected streamgages in neighboring States of Missouri 9
4. Annual exceedance-probability discharges for streamgages in Missouri and selected streamgages in neighboring States of Missouri....10
5. Summary of streamgages in Missouri and selected streamgages in neighboring States of Missouri that were considered for use in the regional regression analysis....10
6. Regional variables for the correlation smoothing function in the weighted-multiple-linear-regression model program for Missouri....11
7. Regression equations for estimating annual exceedance-probability discharges (AEPD) for unregulated streams in Region 1 in rural Missouri 12
8. Regression equations for estimating annual exceedance-probability discharges (AEPD) for unregulated streams in Region 2 in rural Missouri 13

9. Regression equations for estimating annual exceedance-probability discharges (AEPD) for unregulated streams in Region 3 in rural Missouri 13
10. Values needed to determine the 90-percent prediction intervals for estimates obtained from regional regression equations using covariance matrices in Missouri 13
11. Comparison of average standard error of prediction from Alexander and Wilson (1995) and those determined for this study 18
12. Variance of prediction values for streamgages in Missouri and selected streamgages in neighboring States of Missouri 18
13. Regional exponents and constants determined from regional regression of log-transformed drainage area for area-weighting method to estimate annual exceedance-probability discharges for ungaged sites on gaged streams 20
14. Range of basin-characteristic values used to develop regional annual exceedance-probability regression equations for unregulated streams in rural Missouri 21
1–1. Streamgages located in Missouri and in neighboring States that were evaluated for use in the regional skew analysis for Missouri 32
1–2. Regional skewness models for Missouri study area 34
1–3. Pseudo analysis of variance (ANOVA) for the Missouri regional skew study for the constant model 36

Conversion Factors and Datum

Inch/Pound to SI

Multiply	By	To obtain
	Length	
inch (in.)	2.54	centimeter (cm)
foot (ft)	0.3048	meter (m)
mile (mi)	1.609	kilometer (km)
	Area	
square mile (mi^2)	2.590	square kilometer (km^2)
	Flow rate	
foot per second (ft/s)	0.3048	meter per second (m/s)
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)
	Hydraulic gradient	
foot per mile (ft/mi)	0.1894	meter per kilometer (m/km)

Vertical coordinate information is referenced to North American Vertical Datum of 1988 (NAVD 88).

Horizontal coordinate information is referenced to North American Datum of 1983 (NAD 83).

Water year is defined as the 12-month period from October 1 through September 30 of the following year.

Abbreviations

17B	Bulletin 17B
ΔA	absolute value of the difference between the drainage areas of a streamgage and an ungaged site
Adj-R^2	adjusted coefficient of determination
AEPD	annual exceedance-probability discharge
$A_{(g)}$	drainage area for a streamgage
$A_{(u)}$	drainage area for an ungaged site
AVP	average variance of prediction
B-GLS	Bayesian- Generalized Least Squares regression
B-WLS	Bayesian-Weighted Least Squares regression
B-WLS/B-GLS	Bayesian-Weighted Least Squares/Bayesian-Generalized Least Squares regression
BSHAPE	basin shape
BSLDEM10M	mean basin slope
CSL1085LFP	main channel slope measured between the 10- and 85-percent points along the longest flow path
DAR	drainage-area ratio
DEM	digital elevation model
DRNAREA	geographic information system-determined drainage area
EMA	expected moments algorithm
EMA/MGB	expected moments algorithm with the multiple Grubbs-Beck test
FEMA	Federal Emergency Management Agency
FORESTNLCD01	forest class from National Land Cover Data 2001
GB	Grubbs-Beck test
GIS	geographic information system
GLS	Generalized Least Squares regression
HUC	hydrologic unit code
LFPLENGTH	longest flow length
LOW_PERM	minimal permeability
LOWESS	locally weighted scatter plot smoother
LP3	log-Pearson Type III
Mallow's *Cp*	measure of the total squared error for a subset model containing the number (*n*) of independent variables
MEV	model error variance
MGB	multiple Grubbs-Beck test
MSE	mean-square error
MSE_G	MSE of regional skew
MSE_S	MSE of station skew
NAWQA	National Water-Quality Assessment
NHD	National Hydrography Dataset

NRCS	Natural Resources Conservation Service
NWIS	National Water Information System
OLS	ordinary-least-squares regression
PASTURENLCD01	pasture class from National Land Cover Data 2001
PeakFQ	U.S. Geological Survey Peak flow FreQuency analysis program
POR	period of record
PRESS	Predicted Residual Sum of Squares
P_{RL}	pseudo record length
Pseudo ANOVA	pseudo Analysis of Variance
Pseudo-R^2	pseudo coefficient of determination
$Q_{90\%}$	annual exceedance-probability discharge of 90 percent (1.1-year recurrence-interval flood discharge)
$Q_{83\%}$	annual exceedance-probability discharge of 83 percent (1.2-year recurrence-interval flood discharge)
$Q_{50\%}$	annual exceedance-probability discharge of 50 percent (2-year recurrence-interval flood discharge)
$Q_{20\%}$	annual exceedance-probability discharge of 20 percent (5-year recurrence-interval flood discharge)
$Q_{10\%}$	annual exceedance-probability discharge of 10 percent (10-year recurrence-interval flood discharge)
$Q_{4\%}$	annual exceedance-probability discharge of 4 percent (25-year recurrence-interval flood discharge)
$Q_{2\%}$	annual exceedance-probability discharge of 2 percent (50-year recurrence-interval flood discharge)
$Q_{1\%}$	annual exceedance-probability discharge of 1 percent (100-year recurrence-interval flood discharge)
$Q_{0.5\%}$	annual exceedance-probability discharge of 0.5 percent (200-year recurrence-interval flood discharge)
$Q_{0.2\%}$	annual exceedance-probability discharge of 0.2 percent (500-year recurrence-interval flood discharge)
Q_{hist}	historical flood discharge
$Q_{P(g)r}$	regional regression equation estimate of flood discharge for AEPD for a streamgage
$Q_{P(g)w}$	weighted independent estimates of flood discharge for AEPD for a streamgage
$Q_{P(u)aw}$	area-weighted estimate of flood discharge for AEPD for an ungaged site
$Q_{P(u)r}$	regional regression equation estimate of flood discharge for AEPD for an ungaged site
$Q_{P(u)rw}$	regression-weighted estimate of flood discharge for AEPD for an ungaged site
R^2	coefficient of determination
RMSE	root mean square error, also referred to as SEE
RRE	regional regression equation
SD	standardized distance
SEE	average standard error of estimate, also referred to as RMSE

SEM	standard error of model
SEP	average standard error of prediction
SINKHOLES	number of sink holes
SOILASSURGO	soil type A class from National Land Cover Data 2001
SOILCSSURGO	soil type C class from National Land Cover Data 2001
SOILDSSURGO	soil type D class from National Land Cover Data 2001
SPRINGS	number of springs
SSURGO	Soil Survey Geographic
SWSTAT	U. S. Geological Survey Surface-Water Statistics computer program
U	covariance matrix
USACE	U.S. Army Corps of Engineers
USDA	U. S. Department of Agriculture
USGS	U.S. Geological Survey
VIF	variance inflation factor
WBD	Watershed Boundary Data set
WETLAND	open water wetlands
WIE	weighted independent estimates
WLS	weighted Least Squares regression
WREG	weighted-multiple-linear regression program

Methods for Estimating Annual Exceedance-Probability Discharges and Largest Recorded Floods for Unregulated Streams in Rural Missouri

By Rodney E. Southard and Andrea G. Veilleux

Abstract

Regression analysis techniques were used to develop a set of equations for rural ungaged stream sites for estimating discharges with 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probabilities, which are equivalent to annual flood-frequency recurrence intervals of 2, 5, 10, 25, 50, 100, 200, and 500 years, respectively. Basin and climatic characteristics were computed using geographic information software and digital geospatial data. A total of 35 characteristics were computed for use in preliminary statewide and regional regression analyses. Annual exceedance-probability discharge estimates were computed for 278 streamgages by using the expected moments algorithm to fit a log-Pearson Type III distribution to the logarithms of annual peak discharges for each streamgage using annual peak-discharge data from water year 1844 to 2012. Low-outlier and historic information were incorporated into the annual exceedance-probability analyses, and a generalized multiple Grubbs-Beck test was used to detect potentially influential low floods. Annual peak flows less than a minimum recordable discharge at a streamgage were incorporated into the at-site station analyses.

An updated regional skew coefficient was determined for the State of Missouri using Bayesian weighted least-squares/generalized least squares regression analyses. At-site skew estimates for 108 long-term streamgages with 30 or more years of record and the 35 basin characteristics defined for this study were used to estimate the regional variability in skew. However, a constant generalized-skew value of -0.30 and a mean square error of 0.14 were determined in this study.

Previous flood studies indicated that the distinct physical features of the three physiographic provinces have a pronounced effect on the magnitude of flood peaks. Trends in the magnitudes of the residuals from preliminary statewide regression analyses from previous studies confirmed that regional analyses in this study were similar and related to three primary physiographic provinces. The final regional regression analyses resulted in three sets of equations. For Regions 1 and 2, the basin characteristics of drainage area and basin shape factor were statistically significant. For Region 3, because of the small amount of data from streamgages, only drainage area was statistically significant. Average standard errors of prediction ranged from 28.7 to 38.4 percent for flood region 1, 24.1 to 43.5 percent for flood region 2, and 25.8 to 30.5 percent for region 3. The regional regression equations are only applicable to stream sites in Missouri with flows not significantly affected by regulation, channelization, backwater, diversion, or urbanization. Basins with about 5 percent or less impervious area were considered to be rural. Applicability of the equations are limited to the basin characteristic values that range from 0.11 to 8,212.38 square miles (mi^2) and basin shape from 2.25 to 26.59 for Region 1, 0.17 to 4,008.92 mi^2 and basin shape 2.04 to 26.89 for Region 2, and 2.12 to 2,177.58 mi^2 for Region 3.

Annual peak data from streamgages were used to qualitatively assess the largest floods recorded at streamgages in Missouri since the 1915 water year. Based on existing streamgage data, the 1983 flood event was the largest flood event on record since 1915. The next five largest flood events, in descending order, took place in 1993, 1973, 2008, 1994 and 1915. Since 1915, five of six of the largest floods on record occurred from 1973 to 2012.

Introduction

Floods are common in Missouri and can be caused by excessive rainfall in local areas resulting in flash flooding on small streams or by persistent precipitation patterns that result in excessive rainfall, which leads to long duration flood events in the Missouri or Mississippi River Basins. In Missouri, 2008 was the wettest calendar year on record since 1895 with an average rainfall of 57.34 inches (Southard, 2013). In 2011, record flooding in the Central United States caused 33 fatalities and approximately $4.2 billion in damages (Holmes and others, 2013).

Engineers and planners need the best hydrologic information possible to assess the adequacy of existing bridge structures and properly design new bridge structures. The Missouri Department of Transportation (MoDOT) is responsible

for maintaining about 10,400 bridges (Missouri Department of Transportation, 2013). Of the 10,400 bridges, less than 500 have sufficient data to compute at-site flood-frequency estimates. Also, flood-plain management requires the most accurate estimates of the 1- and 0.2-percent annual exceedance-probability discharges (AEPDs) available to define flood risk for land and home owners for the National Flood Insurance Program administered by the Federal Emergency Management Agency (FEMA; Federal Emergency Management Agency, 2002). Accurate estimates related to AEPDs for flood risk is essential for defining a water-surface profile and the area that can be inundated by the 1-percent AEPD (or base flood). Historical flood information is non-existent at most bridges and along stream reaches in Missouri.

The U.S. Geological Survey (USGS) has published a series of reports to provide users with updated and more accurate means of computing AEPDs at ungaged sites on rural streams. The last flood frequency report for Missouri was completed by Alexander and Wilson (1995). Since that publication, additional digital geospatial data is available for computing basin characteristics; Bayesian weighted least-squares/Bayesian generalized least-squares regression (B-WLS/B-GLS) is available for revision of the generalized skew map from Bulletin 17B (Bulletin 17B or 17B; U.S. Interagency Advisory Committee on Water Data, 1982), and an expected moments algorithm (EMA) statistical method is available to estimate AEPDs from annual-peak discharge record. Also, 19 years of additional peak flow data have been collected since the 1995 report was completed. The U.S. Geological Survey, in cooperation with the MoDOT and FEMA, initiated a statewide study in 2010 to update the Alexander and Wilson (1995) rural flood-frequency report using data about annual flood peaks from water year 1844 to 2012. A water year is the 12-month period October 1 through September 30 designated by the calendar year in which it ends.

Purpose and Scope

The purpose of this report is to present an updated set of regression equations for estimating AEPDs for use in Missouri. The regression equations relate AEPDs to size and shape of drainage basins. This report presents three sets of equations to estimate eight selected AEPD statistics that have probabilities of 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent, which are equivalent to annual flood-frequency recurrence intervals of 2, 5, 10, 25, 50, 100, 200, and 500 years, respectively. Hereafter, in this report, these statistics are denoted as $Q_{50\%}$ (in percent [%]), $Q_{20\%}$, $Q_{10\%}$, $Q_{4\%}$, $Q_{2\%}$, $Q_{1\%}$, $Q_{0.5\%}$, and $Q_{0.2\%}$, respectively. Data from streamgages used in the equations were unaffected by urbanization, regulation, backwater, and diversions. Annual peak-discharge data were compiled for water years 1844 through 2012 and streamgages with 10 or more years of record were evaluated for use in this study. The limitations and accuracy of the regression equations also are presented in this report.

Information in this report builds upon the work originally done in cooperation with the Missouri Department of Natural Resources by using the suite of basin characteristics that were considered potentially significant for the low-flow frequency study (Southard, 2013) and for flood-frequency analyses. A primary component of this study includes the application of a new B-WLS/B-GLS analysis to compute a new regional skew for Missouri; this methodology also was recently used to develop a new regional skew model for Iowa (Eash and others, 2013). To compute selected AEPDs, the EMA method was applied to data from 278 streamgages with at least 10 years of annual peak-discharge record. Data were included from water year 1844 to 2012. This report provides a set of regression equations that supersedes the equations in Alexander and Wilson (1995) for estimating selected AEPDs for an ungaged site on unregulated streams in rural Missouri. The independent variables may be computed using geographic information system (GIS) tools and digital geospatial data, which standardizes the computation and removes any potential bias from different manual techniques.

The scope of this report is to update rural flood frequency equations for Missouri and includes: (1) a review and compilation of annual peak-discharge data, (2) compilation of basin characteristics, (3) computation of AEPDs at each streamgage, (4) update of regional skew coefficients for Missouri, (5) definition of hydrologically similar regions, (6) development of regional regression equations, and (7) qualitative assessment of the largest recorded floods in Missouri. The regional skew map was updated by using B-WLS/B-GLS. The AEPDs for each streamgage were analyzed by implementing a new Expected Moments Algorithm (EMA) technique on annual flow series (Cohn and others, 1997). Application of EMA addresses several methodological concerns identified in 17B, but retains the essential structure and moments-based approach of the existing 17B procedures for determining flood frequency. EMA can accommodate interval data, which simplifies analysis of data sets containing censored observations, historic and (or) paleodata, low outliers, and uncertain data points, and also provides enhanced confidence intervals on the estimated discharges (Veilleux and others, 2014).

Description of Study Area

The study area consisted of the State of Missouri and selected gaged basins draining into or out of Missouri (fig. 1). The data from the 278 streamgages used in the study are located in Missouri, Iowa, Kansas, Oklahoma, and Arkansas. The streamgages within Missouri are located in three primary physiographic provinces: the Central Lowlands, the Ozark Plateaus, and the Mississippi Alluvial Plain (Fenneman, 1938; fig. 2).

The Central Lowlands (Region 1) are located in the northern part of the State (fig. 2), north of the Missouri River covering about 49 percent of Missouri. Local relief along the

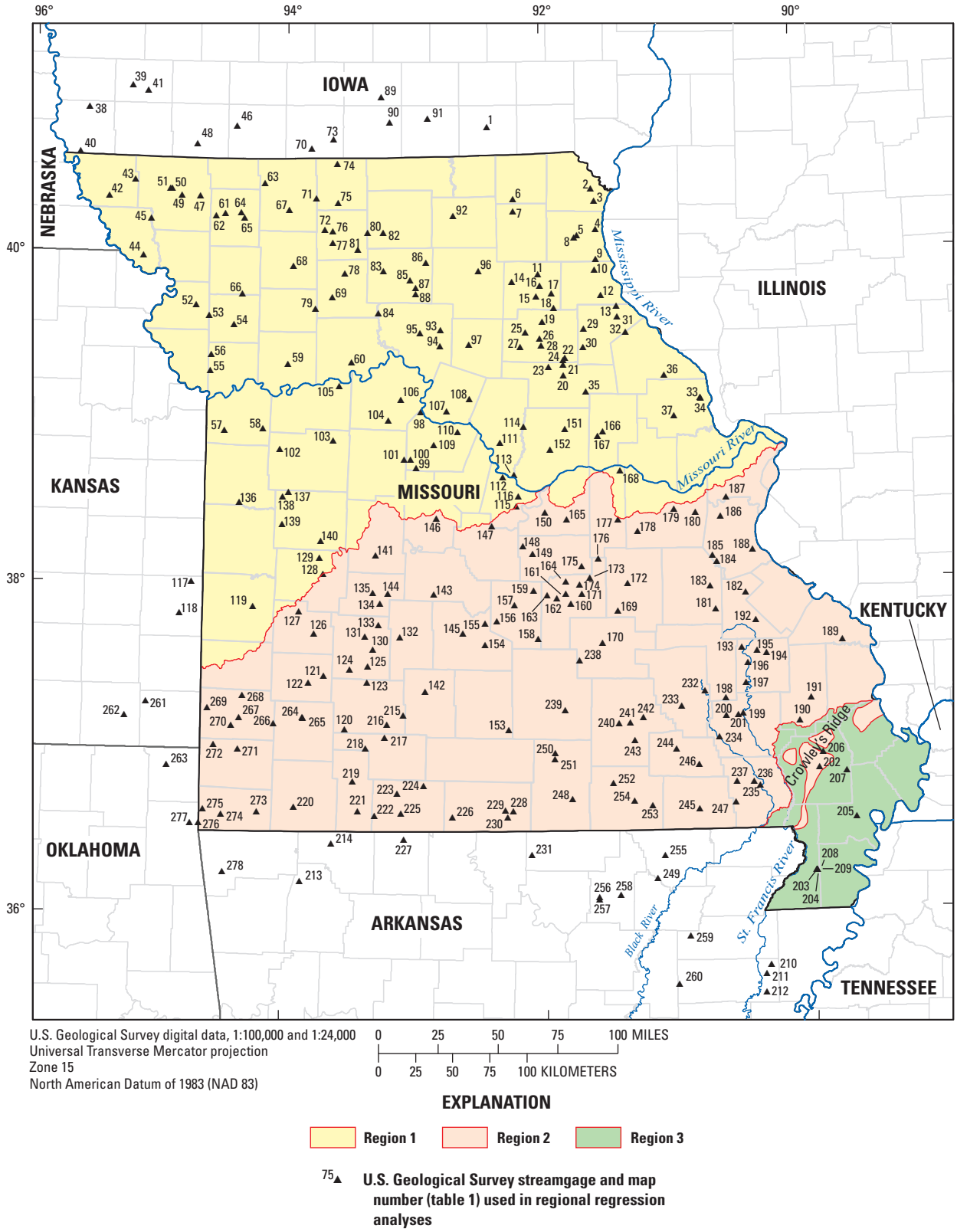


Figure 1. Location of streamgages and hydrologic region boundaries in Missouri and in neighboring States of Missouri.

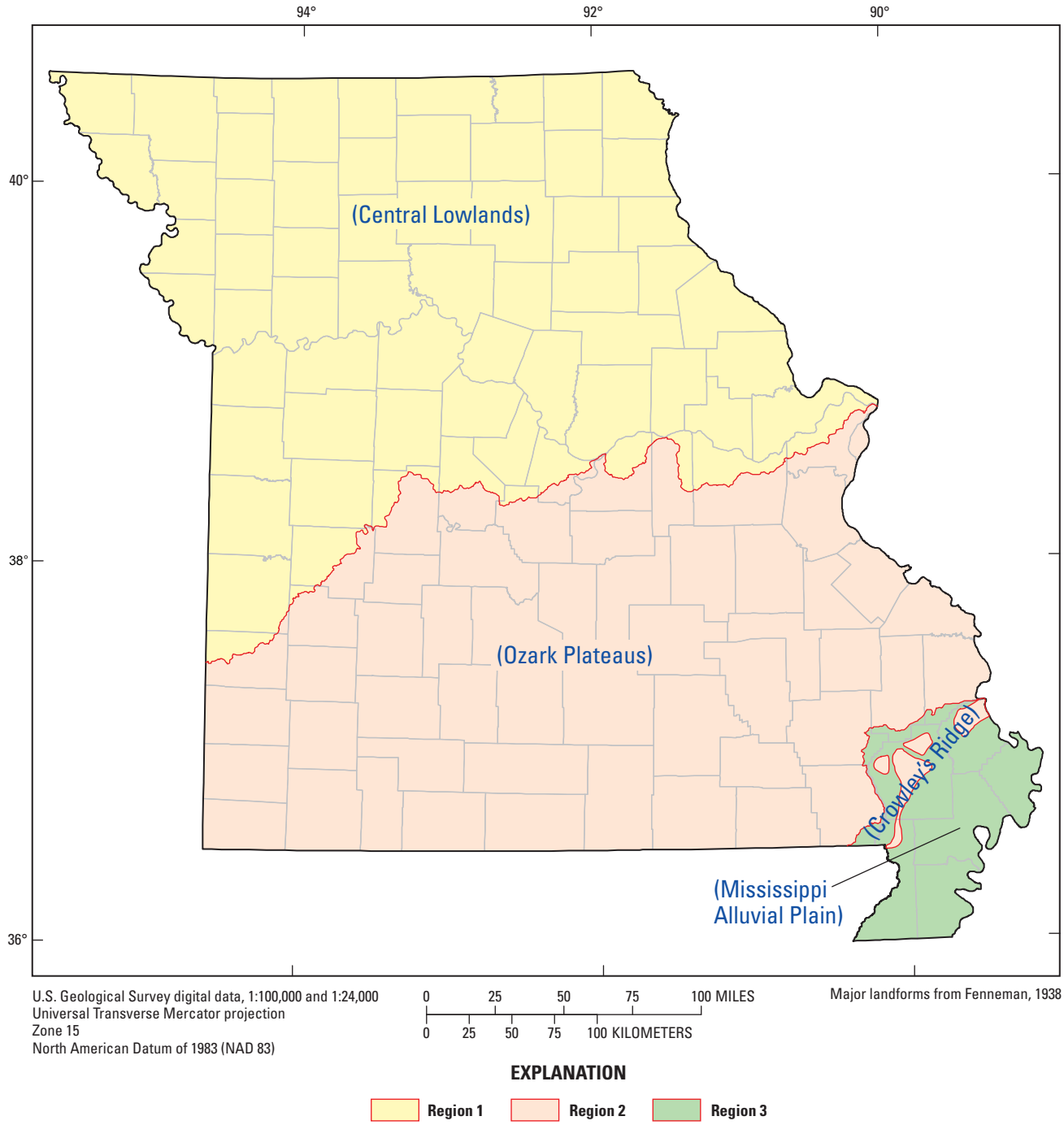


Figure 2. Regions used to determine annual exceedance-probability discharges and physiographic provinces in Missouri.

meandering streams in the wide and flat valleys generally ranges from 50 to 150 feet (Alexander and Wilson, 1995). Local relief within the physiographic region is about 600 feet above the North American Vertical Datum of 1988 (NAVD 88) near the Mississippi River to about 1,200 feet above NAVD 88 in the northwestern parts of the State.

The Ozark Plateaus (Region 2) are located in the southern part of the State (fig. 2). This area covers about 46 percent of Missouri. Narrow valleys 200 to 500 feet deep exist in this area with local relief generally ranging from 100 feet to 500 feet. Main channel gradients generally are steeper than elsewhere in Missouri. Karst features, such as springs, sinkholes, and losing streams, are prominent throughout this region. Elevations range from 800 to about 1,700 feet above NAVD 88 (Alexander and Wilson, 1995). Region 2 also includes the geologic formation known as Crowley's Ridge which rises 250 to 550 feet above the Mississippi Alluvial Plain. This unique and separate geologic formation reflects

hydrologic and hydraulic characteristics similar to Region 2 (fig. 2).

The Mississippi Alluvial Plain (Region 3), in the southeastern part of the State, covers about 5 percent of Missouri (fig. 2). This area is a relatively flat and is predominantly agricultural farmland. Local relief seldom exceeds 30 feet and slope of area streams averages about 1.5 feet per mile (Alexander and Wilson, 1995). Basin boundaries are very difficult to determine within the alluvial plain in the absence of accurate elevation data. Elevations range from 200 to 300 feet above NAVD 88.

Previous Studies

This report is the fifth in a series of reports that describe flood characteristics for Missouri streams. The first report (Searcy, 1955) used data through 1952. AEPDs were computed for selected intervals from $Q_{90\%}$ to $Q_{2\%}$ using data from 112 streamgages. The procedure used in Searcy (1955) consisted of determining the mean-annual flood at a site and the ratio of mean-annual flood to the selected recurrence interval using a set of predetermined curves. The final AEPD was determined by multiplying the mean annual flood by the ratio. The second report (Sandhaus and Skelton, 1968) used data through water year 1965. AEPDs were computed for selected intervals from $Q_{83\%}$ to $Q_{2\%}$ using data from 208 streamgages. Regression equations were developed that related the AEPDs to the basin characteristics of drainage area and main-channel slope. The third report (Hauth, 1974) used data through water year 1970. AEPDs were computed for selected intervals from $Q_{50\%}$ to $Q_{1\%}$ using data from 152 streamgages. The same independent variables of drainage area and main-channel slope also were used by Hauth (1974) to develop regression equations as in the second report. Hauth also used rainfall-runoff modeling to extend the small amount of data available for streams with a drainage area of less than 10 mi^2. The fourth report (Alexander and Wilson, 1995) used data through water year 1993. AEPDs were computed for selected intervals from $Q_{50\%}$ to $Q_{0.2\%}$ using data from 278 streamgages. The same independent variables of drainage area and main-channel slope were used to develop regression equations as those that were used in the 1968 and 1974 reports.

Methods for Data Development for Streamgages

Data preparation for the regional regression analyses included site selection, annual exceedance-probability analyses, and computation of basin characteristics. The annual exceedance-probability analyses included computation of a new regional skew for the State of Missouri, application of the EMA, and use of the multiple Grubbs-Beck test (MGB). The following is a description of each step of the data developed for the selected streamgages.

Site Selection

Data used in the analyses for this report were collected for 278 active and inactive streamgages located in Missouri (245) and in the neighboring States of Iowa (12), Kansas (4), Oklahoma (2), and Arkansas(15) (fig. 1, table 1; *http://pubs.usgs.gov/sir/2014/5165/Downloads/table_1.xlsx*). Streamgage data affected by regulation were not used in regression analyses resulting in unbiased computations of AEPDs. Streamgages were required to have a minimum of 10 years of record for frequency computations. Selection of streamgages in neighboring states was based on basins having similar basin and flow characteristics to basins in Missouri. Streamgages from neighboring States were included to improve the representativeness of the selected AEPDs and basin characteristics for areas near Missouri and to provide better estimates of the error of the regression equations for ungaged sites near the State border. Annual peak data collected through water year 2012 were retrieved for the 278 streamgages from the USGS National Water Information System (U.S. Geological Survey, 2012) database for use in computing selected AEPDs.

Peak-flow data at a streamgage were considered for use in the development of these regression equations if a significant trend did not exist, and if the data represented natural flow conditions, or if the data were affected minimally by anthropogenic activities. Anthropogenic activities that may affect flow statistics include, but are not limited to, regulation, diversions, storage, and urbanization. Basins with impervious areas greater than 5 percent of the total area of the basin were considered urbanized. Streamgages in urban areas generally were excluded from the analysis because of channel improvements, impervious area, and basin development. Data from streamgages also were removed for analysis if considerable storage from small impoundments and water-supply lakes were located in the basin upstream from the streamgage. Data from streamgages were removed if backwater conditions existed at the site. Decisions about inclusion or exclusion of streamgage data also were made using professional judgment.

Annual Exceedance-Probability Analyses

Annual exceedance-probability estimates were computed from an annual series of peak-flow data at continuous-record streamgages and crest-stage gages. Previous analyses were completed following 17B procedures according to methods recommended by the Hydrology Subcommittee of Interagency Advisory Committee on Water Data (IACWD; 1982). The previous rural flood frequency report by Alexander and Wilson (1995) expressed flood-frequency estimates in terms of *T*-year discharges, where *T* is the recurrence interval representing, on average, the number of years between a discharge equal to or greater than a given magnitude. For this report, the flood-frequency estimates are expressed in terms of exceedance probabilities, which are the reciprocals of the recurrence intervals. Exceedance probability can be further expressed as

a percentage, and a particular flood-frequency estimate is then termed the "*P*-percent chance discharge," where *P* is the probability, expressed as a percentage, that the discharge will be equaled or exceeded in any year. For example, a 100-year flood discharge is the same as a discharge having a 0.01 AEPD; this flood discharge also is described as a 1-percent chance flood discharge or $Q_{1\%}$. This report includes an update of the regional skew map for Missouri, and application of the EMA and multiple Grubbs-Beck test.

The IACWD recommends determining flood-frequency estimates by fitting a log-Pearson Type III (LP3) frequency distribution to the logarithms of the annual-peak flows (U.S. Interagency Advisory Committee on Water Data, 1982). Fitting the distribution requires calculating the mean, standard deviation, and skew coefficient of the logarithms of the annual peak-flow series. The mean, standard deviation, and skew coefficient describe the mid-point, slope, and curvature of the peak-flow frequency curve, respectively (Gotvald and others, 2012). Estimates of the *P*-percent AEPDs for each streamgage are computed by inserting the three statistics of the frequency distribution into the equation:

$$\log Q_p = \overline{X} + K_p S \quad (1)$$

where

Q_p is the *P*-percent annual exceedance-probability discharge, in cubic feet per second;

$\overline{X}$ is the mean of the logarithms (base 10) of the annual peak discharges;

K_p is a factor based on the skew coefficient and the given percent annual exceedance probability and is obtained from appendix 3 in Bulletin 17B (U.S. Interagency Advisory Committee on Water Data, 1982); and

S is the standard deviation of the logarithms of the annual peak discharges, which is a measure of the degree of variation of the annual values about the mean value.

The skew coefficient is a measure of the asymmetry of the frequency distribution and is strongly affected by the presence of high or low outliers (annual peaks that are substantially higher or lower than the trend of the data). Large positive skews typically are the result of high outliers, and large negative skews typically are the result of low outliers.

Regional Skew Analysis

The station skew coefficient is sensitive to outliers; therefore, the station skew coefficient for short records may not provide an accurate estimate of the data or true skew coefficient (Gotvald and others, 2009; Feaster and others, 2009; Weaver and others, 2009). Thus, guidelines in Bulletin 17B (U.S. Interagency Advisory Committee on Water Data, 1982) recommend that the skew coefficient calculated from streamgage data (station skew) be weighted with a generalized, or regional, skew determined from an analysis of selected long-term streamgages in the study area (Gotvald and others, 2012). The weighted skew is determined by weighting the station skew and the regional skew and is inversely proportional to their respective mean square errors, as shown in the following equation (U.S. Interagency Advisory Committee on Water Data, 1982):

$$G_W = \left[MSE_{G_R}(G_S) + MSE_{G_S}(G_R)\right] / (MSE_{G_R} + MSE_{G_S}) \quad (2)$$

where

G_W is the weighted skew,

G_S is the station skew,

G_R is the regional skew, and

MSE_{G_R} and MSE_{G_S} are the mean square errors of the regional and station skew, respectively.

The national generalized skew map (plate I, Bulletin 17B [U.S. Interagency Advisory Committee on Water Data, 1982]) is based on streamgage data through water year 1973. Nearly 40 additional years of streamgage data have been collected since the completion of Bulletin 17B and more rigorous statistical procedures are currently (2014) available to generate more accurate estimates of generalized-skew coefficients. Veilleux (2011), Veilleux and others, (2011), and Eash and others, (2013) have applied a B-WLS/B-GLS methodology that relates observed skewness coefficient estimators to basin characteristics in conjunction with diagnostic statistics. The same methodology used in Eash and others (2013) to update the regional skew coefficients in Iowa was applied to data in Missouri to update the regional skew map and to be consistent with procedures used in Iowa. Based on the B-WLS/B-GLS regression analysis using data from 108 long-term streamgages, a constant generalized-skew value of -0.30 was determined to be the best model to predict the generalized skew in the study area for this report. The mean square error (MSE) associated with the new constant generalized skew model for Missouri is 0.14. One difference between the regional skew studies in Missouri and Iowa was the minimum record length for a streamgage to be included in the analyses: the study in Iowa used 25 years and the study in Missouri used a minimum length of 30 years. A detailed description of the updated regional skew analyses for Missouri is presented in appendix 1 of this report.

Expected Moments Algorithm Analysis

In this study, the EMA with the multiple Grubbs-Beck test (EMA/MGB) method was used to compute LP3 exceedance-probability estimates for all 278 streamgages evaluated to develop regression equations for Missouri. EMA addresses several concerns about the methods in the procedures specified by 17B, while retaining the essential structure and moments-based approach of the existing 17B procedures to determine

flood frequency. EMA can accommodate interval data, which simplifies analysis of datasets containing censored observations, historic data, low outliers, and uncertain data points, while also providing enhanced confidence intervals on the estimated discharges. Unlike 17B, which recognizes two categories of data—systematic peaks (annual peaks observed during systematic streamgaging at the station) and historic peaks (peaks observed outside the range of systematic streamgaging)—EMA employs a more general description of the historical period (the length of time that includes both systematic and historic peaks). This is accomplished through the use of flow intervals to describe the knowledge of the peak flow Q_Y in each year Y and through the use of perception thresholds to describe the range of measurable potential discharges in each year Y. Flow intervals and perception thresholds must be defined for every year of the historical period regardless of whether a peak is recorded for a year. In the framework of EMA, the hydrologist's knowledge of the peak flow Q_Y is described by the flow interval ($Q_{Y,lower}$, $Q_{Y,upper}$), which are the upper and lower bounds of the peak flow. When running EMA, a flow interval must be specified for each year in the historical record, including any gaps for which no discharge is recorded, as well as for censored and interval peaks. Interval peaks, or censored peaks, are those peaks that only have a record that the flow was greater than some value and less than another value within a defined interval (flow values between $Q_{Y,lower}$ and $Q_{Y,upper}$). EMA distinguishes among sampling properties by using perception thresholds denoted as ($T_{Y,lower}$, $T_{Y,upper}$), which reflect the range of flows that could have been measured or recorded during an event and are independent of the actual peak discharges. The lower bound, $T_{Y,lower}$, represents the smallest peak flow that would result in a measured flow, whereas the upper bound, $T_{Y,upper}$, represents the largest peak flow that would result in a measured flow (Veilleux and others, 2014).

During a period outside of the systematic record, there may be evidence that floods never exceeded a discharge that would have overtopped certain bridges or roads. During these years, the annual peaks can be represented with a flow interval (0, $Q_{overtop\ the\ road}$). For streamgages that have continuous systematic annual peak discharge records with no low outliers, no censored data, and no historical flood information, the EMA/MGB method provides identical estimates of the three LP3 statistics (mean, standard, deviation, and skew coefficient) as the standard LP3 method described in 17B (Gotvald and others, 2012). A complete description and application of the algorithm is given in Cohn and others (1997) and computation of the confidence intervals for the EMA flood quantile estimates is given in Cohn and others, (2001).

Multiple Grubbs-Beck Test for Detecting Potentially Influential Low Floods

Bulletin 17B recommends the use of the Grubbs-Beck test (Grubbs and Beck, 1972) to statistically identify low outliers in a flood series. As described by Cohn and others (2013), the MGB is a generalization of the Grubbs-Beck method that allows for a standard procedure for identifying multiple Potentially Influential Low Floods (PILFs). In flood-frequency analysis, PILFs are annual peaks that meet three criteria; their magnitude is much smaller than the flood quantile of interest; they occur below a statistically significant break in the flood-frequency plot; and they have excessive influence on the estimated frequency of large floods. When an observation is identified as a PILF, the value of the smallest observation in the data set determined to not be a PILF (Qs) is used as the censoring threshold in the EMA analysis. All annual peaks smaller than this value will be treated as censored observations with flow intervals equal to (0, Qs) and perception thresholds equal to (Qs, *inf*). Identifying PILFs and recording them as censored peaks can greatly improve estimator robustness with little or no loss of efficiency. Thus, the use of the MGB test can improve the fit of the small annual exceedance probabilities, while minimizing lack-of-fit due to unimportant PILFs in an annual peak series (Cohn and others, 2013; Veilleux and others, 2013).

For clarity, it is important to distinguish between low outliers and PILFs. Low outlier typically refers to one or possibly two values in a data set that are assumed to be homogenous and that do not conform to the trend of the other observations. In contrast, PILFs may constitute one-half or more of the observations and are assumed to result from physical processes that are not relevant to the processes associated with large floods. Consequently, the actual magnitudes of PILFs, because they reflect physical processes that are not relevant to large floods, reveal little about the upper right-hand tail of the frequency distribution representing large flood events, and thus, should not have an effect when estimating the risk of large floods. The term "low outlier" has been replaced with the term "PILF" to more accurately describe the situation (U.S. Interagency Advisory Committee on Water Data, 2014).

The USGS computer program PeakFQ version 7.0 was used to compute the flood-frequency estimates for streamgages presented in this report (table 1). PeakFQ automates the EMA/MGB procedure described in this section of the report (PeakFQ version 7.1 was released March 14, 2014, and is now available for public use at *http://water.usgs.gov/software/PeakFQ/*. In this study, when computing flood-frequency estimates for streamgages, the process consisted of the following steps:

1. Retrieve the annual time-series data for peak flows for the streamgage from NWIS (on-line database *http://nwis.waterdata.usgs.gov/usa/nwis/peak*);

2. Consult staff of the USGS Water Science Center for the state where the streamgage is located, complete a literature search, or both, to obtain any at-site hydrologic information that can be used as a basis for inclusion of historic data;

3. Plot the annual-time series to find unusual observations that will require further investigation and to visually detect monotonic or step trends;

4. Evaluate the Kendall's tau test on the time-series data of each streamgage to determine if monotonic trends are statistically significant (Helsel and Hirsch, 2002) and then evaluate the trends to make adjustments to the time series data for trends or eliminate the streamgage data from further analyses;

5. Run EMA/MGB in the PeakFQ software program using the new B-WLS/B-GLS regional skew value of -0.3 and 0.37 for the generalized skew standard error (mean square error of 0.14) to obtain initial at-site flood-frequency estimates for the streamgage; and

6. Review the flood-frequency curve to determine if it adequately fits the annual peak data and evaluate the PILFs when identified by using the MGB test.

Basin Characteristics

Basin characteristics were selected for use as potential independent variables in the regression analyses on the basis of their theoretical relation to peak flows, results of previous studies in similar hydrologic regions, and the ability to measure the basin characteristics using digital data sets and GIS technology. The ability to measure the basin characteristics by using GIS methods helped automate the process of measuring basin characteristics and solving the regression equations for ungaged sites. In a previous study by Southard (2013), a list of 35 basin characteristics pertaining to low- and high-flows were compiled and computed for every streamgage in Missouri except for those streamgages on springs and for those on the Missouri and Mississippi Rivers.

Computation of basin characteristics using GIS software and an increasing number of digital geospatial data have substantially added to the number of possible independent variables for use in regression analyses. All basin and climatic characteristics evaluated for use in this study were computed from digital geospatial data to allow for the automated computation of the characteristic. A review of previous peak-flow studies in Missouri and in other States was completed to denote which characteristics were likely to be statistically significant in regression equations. A list of characteristics was compiled and additional characteristics that could be computed from the same data source were included in the list. The completed list included 35 basin and climatic characteristics (table 2, *http://pubs.usgs.gov/sir/2014/5165/Downloads/table_2.xlsx*). Digital geospatial data were assembled to cover most of the 278 streamgages listed in table 1. For some digital geospatial data types, the data were not available in States bordering Missouri. Every effort was made to make each digital geospatial data type as complete as possible by compiling additional data where available and appended to the existing digital geospatial data. The basin and climatic characteristics can be categorized into four categories: morphometric (physical or shape) characteristics, hydrologic characteristics, pedologic (soils)/geologic/land-use characteristics, and climatic characteristics.

Morphometric characteristics were derived from a USGS digital elevation model (DEM; U.S. Geological Survey, 2011) with a 10-meter resolution (1/3 arc-second National Elevation Data set) available in 2011. The DEM data are updated on a regular basis as more recent and accurate elevation data become available (L.A. Phillips, U.S. Geological Survey, oral commun., 2011). The latest DEM data set may be retrieved from the USGS National Elevation Data set (*http://ned.usgs.gov/*; Gesch, 2007). With higher resolution and more accurate DEM data sets, there are slight differences in the computation of drainage area at some streamgages compared to previously published data. For consistency, the GIS-derived drainage area values (table 2) were used in the regression analyses. The elevation data also were used to define the morphometric characteristics of the stream network of a basin. Characteristics such as stream length, density, and total miles of streams may be computed for a basin. The statistical method used to represent the basin shape (BSHAPE) in this study was to divide the distance of the longest flow path or stream length squared by the drainage area. BSHAPE effects the magnitude and arrival time of a peak discharge. Basins with elongated shapes will have longer duration hydrographs and lower peak discharges than a circular basin that will have shorter duration hydrographs and higher peak discharges.

The Mississippi Alluvial Plain in southeastern Missouri (Region 3, fig. 2) is a flat area that is drained by a series of man-made drainage ditches. Existing (2011) DEM data are not accurate enough to automatically define surface and channel features from the data. Thus, the basin boundaries may be less accurate in Region 3. To improve the interpretation of the basin boundaries in Region 3, the U. S. Department of Agriculture Natural Resources Conservation Service (NRCS) Watershed Boundary Data set (1:24,000 scale, using 12-digit hydrologic unit codes [HUCs]); U.S. Geological Survey and U.S. Department of Agriculture Natural Resources Conservation Service, 2009) and the USGS National Hydrography Data set (NHD, 1:24,000 scale; U.S. Geological Survey, 2012; Simley and Carswell, 2009) were implemented to define the basin boundaries for select streamgages in the Mississippi River Alluvial Plain.

Pedologic (soils)/geologic/land-use characteristics were computed from the NRCS Soil Survey Geographic (SSURGO) Database (Natural Resources Conservation Service, 2012; Multi-Resolution Land Characteristics Consortium, 2012) and geospatial data obtained from the Missouri Department of Natural Resources (2007). Additional information about springs (Branner, 1937) was added to the digital geospatial data to provide more complete information on spring locations in the study area. The basin characteristics from these sources

were processed using the National Water-Quality Assessment (NAWQA) Area-Characterization Toolbox (NACT.tbx) developed by Price and others (2010).

Mean annual precipitation was obtained from the Parameter-Elevation Regressions on Independent Slopes Model (PRISM) Climate Group (Parameter-Elevation Regressions on Independent Slopes Model Climate Group, 2008). Climatic characteristics were digitized and rectified from Hershfield (1961). The digitized contours were converted to a raster surface for processing by the NACT.tbx in ArcGIS version 9.3 (Esri, 2009).

Regional Regression Analyses to Estimate Annual Exceedance-Probability Discharges

The Interagency Advisory Committee on Water Data (1982) recommends that the LP3 distribution be used as the standard flood-frequency technique for Federal planning involving water and related land resources. This technique uses the method-of-moments to relate the annual maximum discharge data to selected frequencies. For this study, the frequencies reported for hydrologically similar regions are the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual-exceedance probabilities.

Definition of Flood Regions

The effect of the three physiographic provinces in Missouri (fig. 2) on the streamflow characteristics has been discussed and documented by Skelton (1970, 1976) and Alexander and Wilson (1995). Alexander and Wilson (1995) defined flood regional boundaries and developed a set of regression equations for each physiographic province. Southard (2013) defined low-flow regional boundaries and developed a set of regression equations for low-flow frequency statistics similar to the regional boundaries defined by Alexander and Wilson (1995). For this study, the definition of regions was based on the previous work of Alexander and Wilson (1995) and the primary basin boundaries that were defined by Southard (2013) to establish the low-flow regional boundaries. Streamgages in the vicinity of the boundaries were evaluated to determine if geographic bias existed and thus to verify the final flood regions.

Development of Regional Regression Equations

Regression equations were developed for use in estimating peak flows for selected annual exceedance probabilities of 50 to 0.2 percent at gaged and ungaged locations for rural basins within the State of Missouri. Ordinary-Least Squares (OLS) regression techniques were performed to select the basin characteristics for use as independent variables. Linear relations between the independent variables and the dependent variable are required for OLS regression. To satisfy this criterion, variables often are transformed, and in hydrologic analyses, typically the log-transformation is used. The dependent response variable is the *P*-percent AEPD and the independent explanatory variables are the basin characteristics that describe the variability determined in the AEPDs. All variables were transformed to base 10 logarithms except for variables representing a percentage such as impervious area.

Homoscedasticity (a constant variance in the dependent variable for the range of the independent variables) about the regression line and normality of residuals also are criteria for OLS regression. Transformation of the *P*-Percent AEPD and certain other variables to logarithms can enhance the homoscedasticity of the data about the regression line. Linearity, homoscedasticity, and normality of residuals were examined in residual plots.

The hydrologic model used in the regression analysis is of the form:

$$Q_P = aA^bB^c \quad (3)$$

where

Q_P is the dependent variable, *P*-percent annual exceedance-probability discharge (AEPD), in cubic feet per second;

A, B, are explanatory (independent) variables; and

a, b, c are regression coefficients.

If the dependent variable Q_P and the independent variables *A* and *B* are logarithmically transformed then the hydrologic model has the following linear form:

$$\mathrm{Log}Q_P = \log(a) + b(\log A) + c(\log B) \quad (4)$$

where the variables are as previously defined.

The OLS results were evaluated on the basis of (Mallow's *Cp*), (TIBCO Software Inc., 2008), statistical significance of the explanatory variables, coefficient of determination (R^2), multicollinearity (correlation among the candidate explanatory variables), and the variance inflation factor (VIF) (Gotvald and others, 2012). The basin characteristics (table 2) selected were determined to not be affected by multicollinearity.

The statistics used to fit the LP3 distribution to the logarithms of observed annual peak flows for each streamgage are the mean of the logarithms, standard deviation of the logarithms, and skew of the logarithms (table 3, *http://pubs.usgs.gov/sir/2014/5165/Downloads/table_3.xlsx*). These statistics describe the midpoint, slope, and curvature of the peak-flow frequency curve, respectively. Individual annual-peak flows that are substantially higher or lower than the trend of the other peaks in the annual flood series are considered outliers, and these peak flows strongly affect the skew parameter. The final estimated flood discharges for selected annual

exceedance probabilities for each streamgage are shown in table 4 (*http://pubs.usgs.gov/sir/2014/5165/Downloads/table_4.xlsx*).

Ordinary-Least-Squares Regression

To evaluate regions for this study (fig. 2), a preliminary statewide regression analysis was implemented using OLS regression with selected streamgages with 10 or more years of record and the 35 basin characteristics listed in table 2. The $Q_{1\%}$ AEPD was chosen for the regression analyses because it is a commonly used statistic by FEMA and the MoDOT for flood frequency analyses in Missouri. The residual value (differences between flood frequency statistics computed from observed peak flow and those predicted from the regression equations) from the preliminary statewide regression analyses were mapped at each streamgage location to identify spatial trends in the predictive accuracy of the preliminary regression equation. Residuals from the statewide analyses confirmed that characteristics of the physiographic provinces affected the results of the preliminary statewide $Q_{1\%}$ AEPD regression equation. OLS regression analyses implemented using subsets of the statewide data set were computed separately for each primary physiographic province to compare regional and statewide predictive accuracies. An improvement in accuracy was made by partitioning the streamgage data by primary physiographic province and then by basin divides. The difference in regional boundaries from Alexander and Wilson (1995) and Southard (2013) were minimal with Southard (2013) using GIS-derived basin boundaries and available streamgage data to define the regional boundaries. An exception to this minimal difference was Crowley's Ridge, which lies in the Mississippi Alluvial Plain region (fig. 2). The topographic relief in Crowley's Ridge is more reflective of relief present in the Ozark Plateaus region. With limited hydrologic and digital data available for improved definition of Crowley's Ridge, the boundary defined in Alexander and Wilson (1995) was included in this study without alteration. The regional boundaries from the low-flow study (Southard, 2013) were evaluated to define the regional boundaries for this study, with the addition of Crowley's Ridge, to allow for consistent application of low-flow and flood-frequency equations for the State of Missouri. The three previously defined low-flow regions, with the inclusion of Crowley's Ridge with the analyses of the Ozark Plateaus (Region 2), were then evaluated for regional flood-frequency regression analyses (fig. 2, table 5; *http://pubs.usgs.gov/sir/2014/5165/Downloads/table_5.xlsx*).

The number and spatial distribution of the streamgages in this study limited the definition of separate hydrologic regions for the State. Additional streamgages outside of Missouri were used to supplement the Missouri streamgages to increase the range of applicability of the regression equations for basins throughout the State. The residuals from the OLS analyses were plotted on a state map. The residuals are the differences between the streamgage flood-frequency estimates and the corresponding OLS regression equation results. The magnitude and numerical sign of the residuals were checked for possible regional biases.

Generalized-Least-Squares Regression

The generalized-least squares (GLS) multiple-linear regression was used to compute the final regression coefficients and the measures of accuracy for the regression equations using the computer program weighted-multiple-linear-regression model program (WREG; Eng and others, 2009). Stedinger and Tasker (1985) compared ordinary, weighted, and GLS regression techniques. The results of their study determined that weighted and generalized regression techniques provided better estimates of the accuracy of the equations than OLS. Improvements in the weighted equations occurred when streamflow records at streamgages are of different lengths. Also, improvements were noted when concurrent flows at different streamgages were correlated. GLS regression, as described by Stedinger and Tasker (1985), Tasker and Stedinger (1989), and Griffis and Stedinger (2007), is a method that weights data from streamgages in the regression analysis according to differences in streamflow reliability (record lengths) and variability (record variance) and according to spatial cross correlations of concurrent streamflow among streamgages. Compared to OLS regression, GLS regression provides improved estimates of streamflow statistics and improved estimates of the predictive accuracy of the regression equations (Stedinger and Tasker, 1985).

The correlation smoothing function used by WREG to compute a weighting matrix for the data from 135 streamgages included in the development of the GLS regression equation for estimating AEPDs for flood region 2 with 40 years of concurrent flow is shown in figure 3. The smoothing function relates the correlation between annual-peak discharges at two streamgages to the geographic distance between the streamgages for every paired combination of the 135 streamgages with 40 years of concurrent flow. Strong evidence of cross correlation is shown in figure 3 because of the abundance of paired points for 40 years of concurrent flow that form the tail of the curve that extends towards the bottom right side of the graph. Final GLS regression models were selected primarily on the basis of minimizing values of the standard error of model (SEM) and the standard error of estimate (SEP), and maximizing values of the pseudo coefficient of determination (pseudo-R^2). The computed annual exceedance probabilities of 50, 20, 10, 4, 2, 1, 0.5, and 0.2 percent for the at-site estimates (EMA/MGB), regional-regression estimates (RRE), and weighted-independent estimates (WIE) values are included in table 4. The regional variables for the correlation smoothing function used in WREG for each region are presented in table 6.

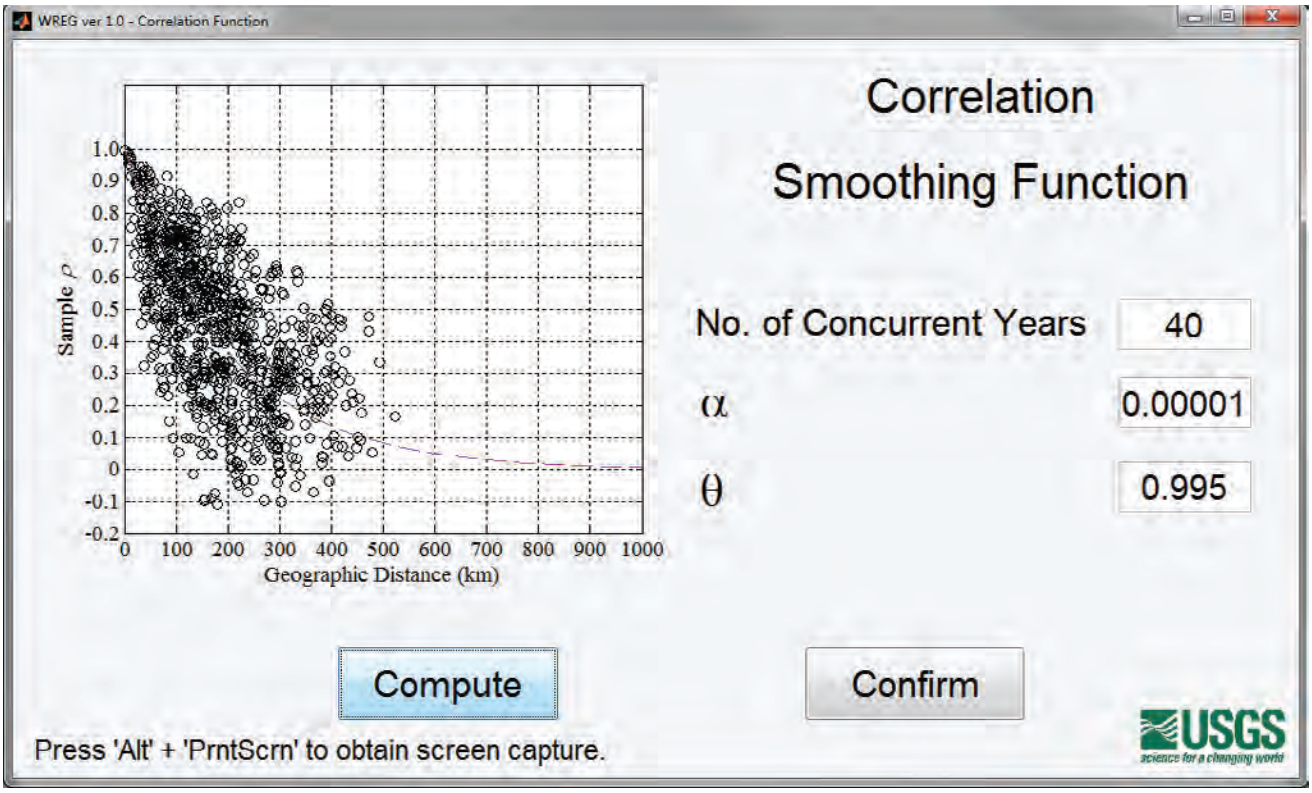


Figure 3. Screen capture of the weighted-multiple-linear regression program (WREG) smoothing function for generalized-least-squares (GLS) correlation of annual peak flows as a function of the distance between 135 streamgages in Region 2 with 40 years of concurrent peak-flow record.

Final Regional Regression Equations

A set of regional regression equations was developed for each region. The selection of basin and climatic characteristics and the evaluation of the accuracy of the regional equations were based on the $Q_{1\%}$ AEPD statistic. Data from the streamgages were subdivided into separate data sets by regions with consideration given to primary drainage basin boundaries at the 4-digit HUC level and the low-flow boundaries defined by Southard (2013). The physiographic provinces were referenced using the terminology of Alexander and Wilson (1995) with the Central Lowlands as Region 1, Ozark Plateaus as Region 2, and Mississippi Alluvial Plain as Region 3 for the development of the regression equations (fig. 2). Boundaries of Region 1, 2, and 3 approximate the location of the physiographic province boundaries (Fenneman, 1938, fig. 2) but the two are not coincident. Streamgages were identified by region. The number of streamgages selected for Region 1, 2, and 3 were 131, 135, and 12, respectively. The area encompassed by each region is approximately 34,100 mi^2 for Region 1; 32,300 mi^2 for Region 2; and 3,300 mi^2 for Region 3.

To identify basin characteristics that are statistically significant for inclusion in the regression analyses, the Efroymson stepwise-selection method (Efroymson, 1960) was used to define potential explanatory variables from the list of 35 characteristics. The procedure is similar to forward selection, which tests basin characteristics one by one and identifies those that are statistically significant; however, as each different basin characteristic is identified as being significant, partial correlations are checked to see if any previously identified variables can be deleted (Ahearn, 2010). Highly correlated characteristics were included in the Efroymson selection method one at a time to avoid problems with multicollinearity. Important characteristics were defined for each region in Missouri. The statistical analyses were implemented using Spotfire S+ statistical software (TIBCO Software Inc., 2008).

Table 6. Regional variables for the correlation smoothing function in the weighted-multiple-linear-regression model program for Missouri.

Regional variable	Region 1	Region 2	Region 3
		(fig. 2)	
Streamflow record, in years	40	40	10
Alpha[1]	0.0001	0.00001	0.0001
Beta[1]	0.994	0.995	0.990

[1]Dimensionless parameters estimated from the peak flow data.

For Region 1, the characteristics of longest flow length (LFPLENGTH), drainage area (DRNAREA), main channel slope measured between the 10- and 85-percent points along the longest flow path (CSL1085LFP), basin shape (BSHAPE), soil type A (SOILASSURGO), soil type D (SOILDSSURGO), open water wetlands (WETLAND), forest (FORESTNLCD01), pasture (PASTURENLCD01), minimal permeability (LOW_PERM), number of sinkholes (SINKHOLES), and number of springs (SPRINGS) were determined to be statistically significant. To evaluate the combination of characteristics to use for the ordinary-least squares regression, a linear model subset selection was used to identify the "best" three linear-regression model combinations for each of the one-variable to six-variable regression equations. The variables used in the analyses were log-transformed except for SOILASSURGO, SOILDSSURGO, WETLAND, FORESTNLCD01, PASTURENLCD01, and LOW-PERM. These six variables were not log-transformed because they represent a percentage with numerical limits of 0 to 100. The final regression model was based on the following performance metrics:

1. Adjusted-*R* Squared (Adj-R^2) is the adjusted coefficient of determination and an alternative to *R*-Squared (R^2) in which the percent of variation in the dependent variable ($Q_{1\%}$) can be explained by the variation of the independent variables in the model. In contrast to R^2, Adj-R^2 is adjusted for the number of parameters in the model (number of streamgages and number of independent variables [basin characteristics]; Freund and Littell, 2000).

2. Mallow's *Cp* statistic is a measure of the total squared error for a subset model containing the number (*n*) of independent variables (Freund and Littell, 2000). Mallow's *Cp* is an indicator of model bias (Cavalieri and others, 2000). Models with a large *Cp* are biased because they contain independent variables that are not important in the population.

3. Predicted Residual Sum of Squares (PRESS) statistic is the sum of squares of residuals using models obtained by estimating the equation with all observations except for the *i*th observation (Freund and Littell, 2000) and is an estimate of PRESS. The PRESS statistic measures how well the regression model predicts the ith observation as though it were a new observation (Cavalieri and others, 2000).

The Adj-*R2* statistic is maximized and the Mallow's *Cp* and PRESS statistics are minimized with better combinations of independent variables in a regression model that explain more of the variance in the dependent variable. Incremental improvements in the performance metrics also were evaluated with the addition of another independent variable to the model. A subset of the linear modeling indicated that two independent variables provided the best model to use in the GLS analyses based on the above criteria.

For Region 1, the combination of DRNAREA and BSHAPE resulted in the lowest standard error of prediction of 29.8 percent for the $Q_{1\%}$ statistic for Region 1. The final regional regression equations for Region 1 for the $Q_{50\%}$, $Q_{20\%}$, $Q_{10\%}$, $Q_{4\%}$, $Q_{2\%}$, $Q_{1\%}$, $Q_{0.5\%}$, and $Q_{0.2\%}$ AEPD statistics are presented in table 7. The SEP ranged from 28.7 percent for the $Q_{2\%}$ AEPD statistic to 38.4 percent for the $Q_{50\%}$ AEPD statistic for Region 1.

For Region 2, the most statistically significant independent variables from the Efroymson selection method were LFPLENGTH, DRNAREA, CSL1085LFP, mean basin slope (BSLDEM10M), BSHAPE, SOILASSURGO, soil type C (SOILCSSURGO), WETLAND, SINKHOLES and SPRINGS. The linear-model subset results indicated that a two-variable equation was the most efficient model to use in the GLS analyses. The two-variable equation with the lowest SEP included the independent variables of DRNAREA and BSHAPE. These two variables resulted in a standard error of prediction of 24.4 percent for the $Q_{1\%}$ AEPD statistic and

Table 7. Regression equations for estimating annual exceedance-probability discharges (AEPD) for unregulated streams in Region 1 in rural Missouri.

[SEP, average standard error of prediction; pseudo-R^2, pseudo coefficient of determination; SEM, average standard error of model; AVP, average variance of prediction; $Q_{x\%}$, annual exceedance-probability discharge of x percent; DRNAREA, geographic information system drainage area; BSHAPE, longest flow path squared divided by drainage area]

Annual exceedance-probability equation	SEP (percent)	Pseudo-R^2 (percent)	SEM (percent)	AVP (log $ft^3/s)^2$
Data from 131 streamgages used to develop equations				
$Q_{50\%}=(10^{2.594})\ (DRNAREA^{0.618})\ (BSHAPE^{-0.282})$	38.4	95.7	37.3	0.026
$Q_{20\%}=(10^{2.861})\ (DRNAREA^{0.593})\ (BSHAPE^{-0.266})$	30.8	97.0	29.7	0.017
$Q_{10\%}=(10^{2.990})\ (DRNAREA^{0.580})\ (BSHAPE^{-0.258})$	29.1	97.2	27.8	0.015
$Q_{4\%}=(10^{3.120})\ (DRNAREA^{0.568})\ (BSHAPE^{-0.248})$	28.8	97.1	27.3	0.015
$Q_{2\%}=(10^{3.199})\ (DRNAREA^{0.560})\ (BSHAPE^{-0.242})$	28.7	97.0	27.1	0.015
$Q_{1\%}=(10^{3.266})\ (DRNAREA^{0.554})\ (BSHAPE^{-0.236})$	29.8	96.7	28.1	0.016
$Q_{0.5\%}=(10^{3.324})\ (DRNAREA^{0.550})\ (BSHAPE^{-0.231})$	31.0	96.4	29.1	0.017
$Q_{0.2\%}=(10^{3.391})\ (DRNAREA^{0.544})\ (BSHAPE^{-0.226})$	33.2	95.7	31.2	0.020

a range from 24.1 percent for the $Q_{4\%}$ AEPD statistic to 43.5 percent for the $Q_{50\%}$ AEPD statistic (table 8).

For Region 3, a limited number of streamgages (12) with sufficient record length were available for regional regression analyses. Linear-model subset results were unable to define statistically significant variables; therefore, the variables that were determined to be statistically significant in the Region 1 and 2 analyses were evaluated in the GLS regression for Region 3. Only DRNAREA was determined to be statistically significant. The standard error of prediction was 26.9 percent for the $Q_{1\%}$ AEPD statistic for the one-variable model. For region 3, the standard error of prediction ranges from 25.8 percent for the $Q_{10\%}$ AEPD statistic to 30.5 percent for the $Q_{50\%}$ AEPD statistic (table 9). The parameters needed to determine the 90-percent prediction intervals for estimates obtained from the three sets of eight regional regression equations in Missouri are presented in table 10 (*http://pubs.usgs.gov/sir/2014/5165/Downloads/table_10.xlsx*). A summary of the input data used in the development of the regional regression equations is presented in table 5.

The at-site $Q_{1\%}$ AEPD values were plotted against the estimated values from the regional GLS equations presented in tables 7–9 in figure 4. Data for all three regions document a fairly uniform distribution around the line of equality.

A comparison of the $Q_{1\%}$ frequency statistic for the three regions is shown in figure 5. Curves from the final regional regression equations (fig. 5) use representative basin characteristics, such as a range of drainage areas from 3 to 2,100 mi^2 and a basin shape factor of 8 for Regions 1 and 2. The highest $Q_{1\%}$ estimates were determined for Region 2, for a given size of drainage area, and Region 1 is slightly higher than Region 3. Use of different combinations of characteristics may

Table 8. Regression equations for estimating annual exceedance-probability discharges (AEPD) for unregulated streams in Region 2 in rural Missouri.

[SEP, average standard error of prediction; Pseudo-R^2, pseudo coefficient of determination; SEM, average standard error of model; AVP, average variance of prediction; $Q_{x\%}$, annual exceedance-probability discharge of x percent; DRNAREA, geographic information system drainage area; BSHAPE, longest flow path squared divided by drainage area]

Annual exceedance-probability equation	SEP (percent)	Pseudo-R^2 (percent)	SEM (percent)	AVP (log $ft^3/s)^2$
Data from 135 streamgages used to develop equations				
$Q_{50\%}=(10^{2.493})\ (DRNAREA^{0.686})\ (BSHAPE^{-0.222})$	43.5	96.0	42.2	0.033
$Q_{20\%}=(10^{2.801})\ (DRNAREA^{0.679})\ (BSHAPE^{-0.251})$	31.8	97.7	30.4	0.018
$Q_{10\%}=(10^{2.955})\ (DRNAREA^{0.676})\ (BSHAPE^{-0.268})$	28.0	98.2	26.5	0.014
$Q_{4\%}=(10^{3.113})\ (DRNAREA^{0.673})\ (BSHAPE^{-0.287})$	24.1	98.7	22.2	0.011
$Q_{2\%}=(10^{3.205})\ (DRNAREA^{0.671})\ (BSHAPE^{-0.296})$	24.2	98.7	22.1	0.011
$Q_{1\%}=(10^{3.282})\ (DRNAREA^{0.669})\ (BSHAPE^{-0.302})$	24.4	98.6	22.1	0.011
$Q_{0.5\%}=(10^{3.349})\ (DRNAREA^{0.668})\ (BSHAPE^{-0.307})$	24.6	98.6	22.1	0.011
$Q_{0.2\%}=(10^{3.422})\ (DRNAREA^{0.667})\ (BSHAPE^{-0.311})$	27.0	98.3	24.3	0.013

Table 9. Regression equations for estimating annual exceedance-probability discharges (AEPD) for unregulated streams in Region 3 in rural Missouri.

[SEP, average standard error of prediction; pseudo-R^2, pseudo coefficient of determination; SEM, average standard error of model; AVP, average variance of prediction; $Q_{x\%}$, annual exceedance-probability discharge of x percent; DRNAREA, geographic information system drainage area]

Annual exceedance-probability equation	SEP (percent)	Pseudo-R^2 (percent)	SEM (percent)	AVP (log $ft^3/s)^2$
Data from 12 streamgages used to develop equations				
$Q_{50\%}=(10^{1.933})\ (DRNAREA^{0.665})$	30.5	95.7	27.8	0.017
$Q_{20\%}=(10^{2.026})\ (DRNAREA^{0.681})$	26.4	96.9	24.0	0.013
$Q_{10\%}=(10^{2.070})\ (DRNAREA^{0.689})$	25.8	97.1	23.3	0.012
$Q_{4\%}=(10^{2.113})\ (DRNAREA^{0.698})$	26.2	97.1	23.6	0.013
$Q_{2\%}=(10^{2.139})\ (DRNAREA^{0.703})$	26.6	97.1	23.8	0.013
$Q_{1\%}=(10^{2.162})\ (DRNAREA^{0.708})$	26.9	97.1	24.1	0.013
$Q_{0.5\%}=(10^{2.182})\ (DRNAREA^{0.713})$	28.3	96.9	25.3	0.015
$Q_{0.2\%}=(10^{2.204})\ (DRNAREA^{0.718})$	28.7	96.9	25.5	0.015

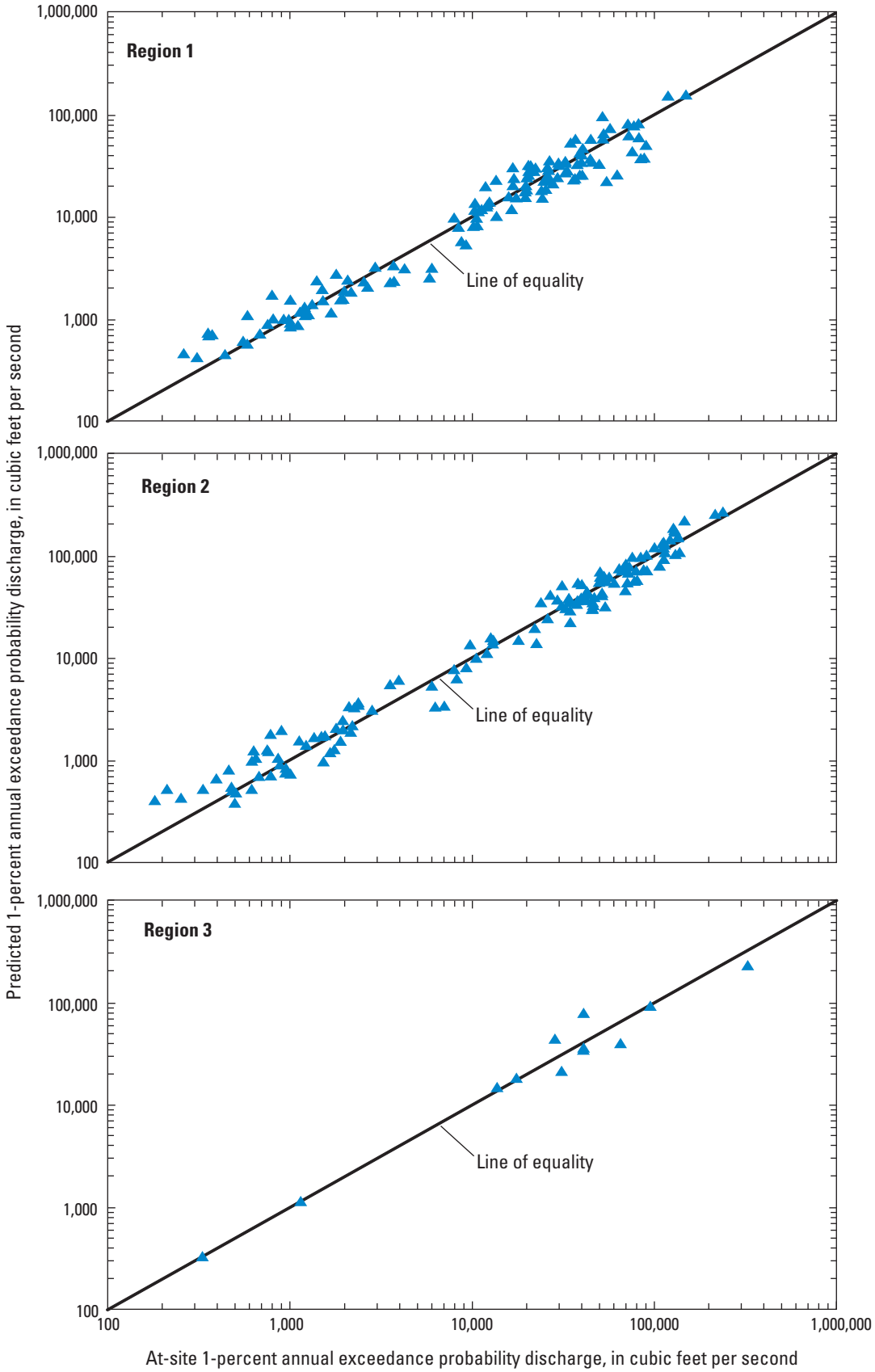


Figure 4. Relation between 1-percent annual exceedance-probability discharges computed from at-site streamflow to those predicted from generalized-least squares regression equations for flood regions in Missouri.

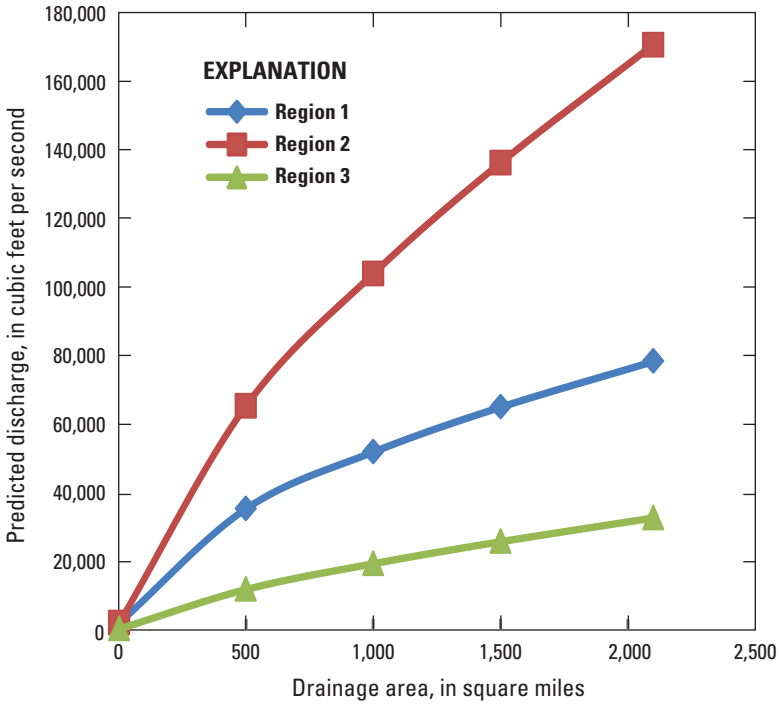


Figure 5. Relation of Regions 1, 2, and 3 for the 1-percent ($Q_{1\%}$) annual exceedance-probability discharge using a factor of eight for basin shape and drainage areas from 3 to 2,100 square miles in the regional regression equations.

produce slightly different results. The magnitude of the frequency statistics are the highest in Region 2 where relief is the greatest and main-channel gradients are steeper than elsewhere in Missouri (Alexander and Wilson, 1995).

Differences in the results of the methodology used in Alexander and Wilson (1995) and the application of the EMA algorithm using the MGB test for data from streamgage 07066000 (map number 241, fig. 1) to detect PILFs is shown in figure 6. At this streamgage, 20 flows were recorded as censored values in the MGB test and a slightly different regional skew value of -0.30 was used in this study, compared to -0.31 in the 1995 study. By removing the 20 potentially influential low flows from the at-site analyses, the estimated regression estimates were raised for the upper AEPDs. Alexander and Wilson (1995) estimated the $Q_{0.2\%}$ AEPD to be 102,000 cubic feet per second and the EMA/MGB analysis estimate for this study was 117,000 cubic feet per second. Only the 1994 peak flow (58,500 cubic feet per second) was greater than a $Q_{10\%}$ AEPD since the 1995 report was published. The plot of the station frequency curves for both studies is shown in figure 6 and the EMA/MGB analyses seems to align better with the historic flood of 1904.

A plot of the basin characteristics and residuals of the $Q_{1\%}$ AEPD for each region is shown in figure 7. The magnitude and numerical sign of the residuals were checked for possible regional biases and none were determined. The random scatter of the points above and below the zero reference line indicates that the models were satisfactorily meeting the assumption of multiple regression techniques.

An analysis of the potential change in regression equations from Alexander and Wilson (1995) to this study was done using the $Q_{1\%}$ AEPD estimates. The $Q_{1\%}$ AEPD was computed for all 278 streamgages using the regional regression equations in Alexander and Wilson (1995) and those equations contained in tables 7–9. A percent difference was computed by subtracting the 1995 regression equation estimates from the estimates of this study, dividing by the estimates of this study, then multiplying by 100. Graphs showing the percentage differences by region are shown in figure 8. A logarithmic trend line also is shown on the log-linear plots. The trend line indicates the equations in table 7 will provide slightly lower estimates than the previous equations for Region 1. In Region 2, the trend line appears to indicate the equations in table 8 will provide slightly higher estimates than the previous equations. With limited data in Region 3, the trend line appears to indicate the equations in table 9 will provide lower estimates than previous equations for drainage areas less than about 500 mi^2, and slightly higher for larger basins compared to the 1995 equations by Alexander and Wilson.

Derivation of a new skew coefficient with a much lower standard error, application of the EMA with the MGB test technique, and additional streamgage data all contributed to improved estimates of the average standard error of prediction for the regional regression equations in each region. In table 11, the average standard errors of prediction are presented for Alexander and Wilson (1995) and those determined for this study. Except for the $Q_{50\%}$ AEPD equation for Region 1 and 2, the errors presented in this study are lower and some substantially lower than in Alexander and Wilson (1995).

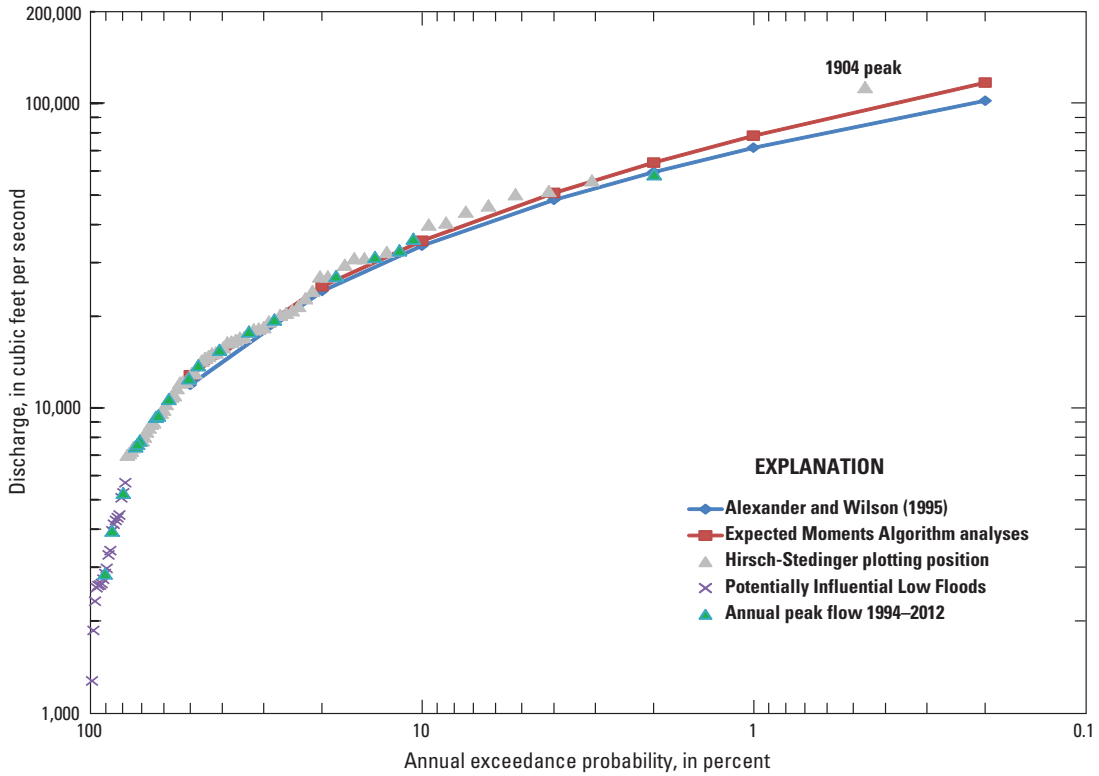


Figure 6. Difference in at-site flood frequency estimates from Alexander and Wilson (1995; 17B method) and using the Expected Moments Algorithm with multiple Grubbs-Beck test method at USGS streamgage 07066000 (map number 241).

Application and Limitations of Regression Equations

Three methods are presented below to estimate AEPDs at an ungaged site. The best method may depend on several factors: (1) sufficient length of record at a streamgage needed to compute reliable AEPDs, if the ungaged site is on the same stream as a streamgage, (2) differences in size of the drainage areas of the ungaged site and the streamgage, and (3) if the streamgage data are representative of the flow characteristics at the ungaged site.

Streamgage Locations

Improved estimates of AEPDs at streamgages can be obtained by weighting the AEPD EMA/MGB estimate with the RRE estimate. The variance of prediction is considered to be a measure of the uncertainty of each estimate and can be used to lower the uncertainty of the weighted estimate by weighting the variance of prediction of each estimate that is inversely proportional to their associated estimates. The EMA/MGB and RRE estimates are assumed to be independent. The variance of the weighted estimate will be less than the variance of either of the independent estimates. Optimal weighted estimates of AEPDs were computed for this study using the Weighted Independent Estimates (WIE) computer program available at *http://water.usgs.gov/usgs/osw/swstats/freq.html*. Information about this computer program is presented by Cohn and others (2012).

Once the variances have been computed, the two independent annual exceedance-probability estimates can be weighted using the following equation (Verdi and Dixon, 2011; Cohn and others, 2012; Gotvald and others, 2012).

$$logQ_{P(g)w} = \frac{VP_{P(g)r} \log Q_{P(g)s} + VP_{P(g)s} \log Q_{P(g)r}}{VP_{P(g)s} + VP_{P(g)r}}, \quad (5)$$

where

$Q_{P(g)w}$ is the weighted independent estimate of annual peak flow for the selected *P*-percent annual exceedance probability for a streamgage, *g*, in cubic feet per second;

$VP_{P(g)r}$ is the variance of prediction at the streamgage derived from the applicable regional-regression equations for the selected *P*-percent annual exceedance probability (from table 12, *http://pubs.usgs.gov/sir/2014/5165/Downloads/table_12.xlsx*), in log units;

$Q_{P(g)s}$ is the at-site estimate from the expected moments algorithm or multiple Grubbs-Beck log-Pearson Type III analysis for the selected *P*-percent annual exceedance probability (from table 4) for a streamgage, *g*, in cubic feet per second;

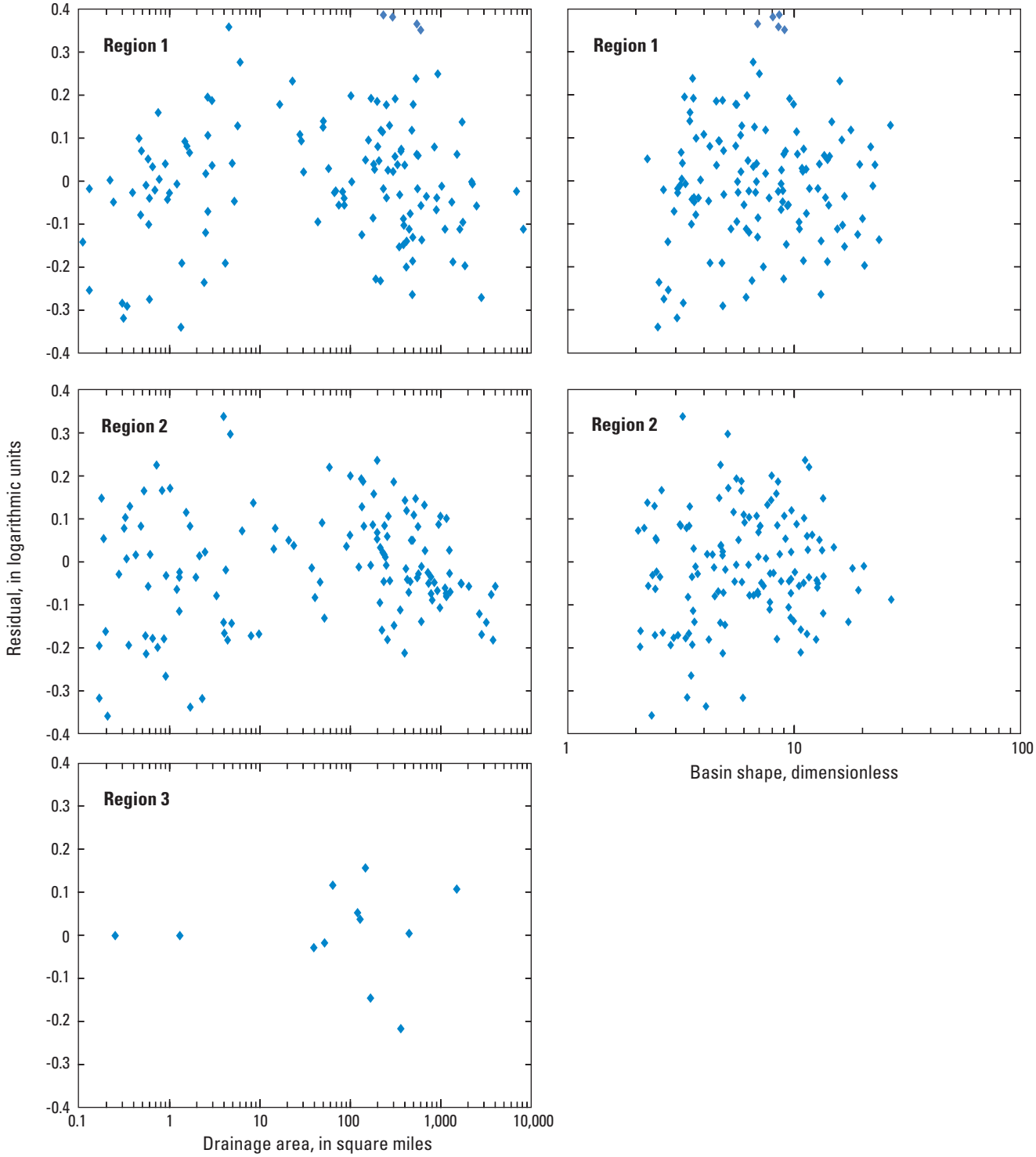


Figure 7. Relation of basin characteristics to residuals from regression analyses for each region for the 1-percent ($Q_{1\%}$) annual exceedance-probability discharge.

Table 11. Comparison of average standard error of prediction from Alexander and Wilson (1995) and those determined for this study.

[$Q_{x\%}$, annual exceedance-probability discharge of x percent; NA, not available]

Annual exceedance probability	Alexander and Wilson (1995)		Current study	
	Number of streamgages used in regression analysis	Average standard error of prediction (percent)	Number of streamgages used in regression analysis (tables 7–9)	Average standard error of prediction (percent) (tables 7–9)
		Region 1		
$Q_{50\%}$	118	34	131	38
$Q_{20\%}$	118	32	131	31
$Q_{10\%}$	118	34	131	29
$Q_{4\%}$	118	36	131	29
$Q_{2\%}$	118	38	131	29
$Q_{1\%}$	118	40	131	30
$Q_{0.5\%}$	NA	NA	131	31
$Q_{0.2\%}$	118	45	131	33
		Region 2		
$Q_{50\%}$	143	43	135	44
$Q_{20\%}$	143	36	135	32
$Q_{10\%}$	143	34	135	28
$Q_{4\%}$	143	32	135	24
$Q_{2\%}$	143	31	135	24
$Q_{1\%}$	143	32	135	24
$Q_{0.5\%}$	NA	NA	135	25
$Q_{0.2\%}$	143	34	135	27
		Region 3		
$Q_{50\%}$	17	34	12	30
$Q_{20\%}$	17	36	12	26
$Q_{10\%}$	17	38	12	26
$Q_{4\%}$	17	41	12	26
$Q_{2\%}$	17	44	12	27
$Q_{1\%}$	17	46	12	27
$Q_{0.5\%}$	NA	NA	12	28
$Q_{0.2\%}$	17	54	12	29

$VP_{P(g)s}$ is the variance of prediction at the streamgage from the expected moments algorithm or multiple Grubbs-Beck log-Pearson Type III analysis for the selected *P*-percent annual exceedance probability (from table 12), in log units; and

$Q_{P(g)r}$ is the peak flow estimate for the selected *P*-percent annual exceedance probability at the streamgage derived from the applicable regional-regression equations (from table 4), in cubic feet per second.

Weighting the variances inversely proportional minimizes the effect of an estimate with high uncertainty. Likewise, if the uncertainty is low then the weight of the estimate is large. The computed variance of prediction associated with the weighted estimate, $VP_{P(g)w}$, is shown in the following equation (Verdi and Dixon, 2011; Gotvald and others, 2012):

$$VP_{P(g)w} = \frac{VP_{P(g)s} VP_{P(g)r}}{VP_{P(g)s} + VP_{P(g)r}}, \quad (6)$$

where the variables are previously defined. The weighted AEPDs estimates that were computed from equation 5 are listed in table 4. The variance of prediction values for the 278 streamgages included in this study are listed in table 12.

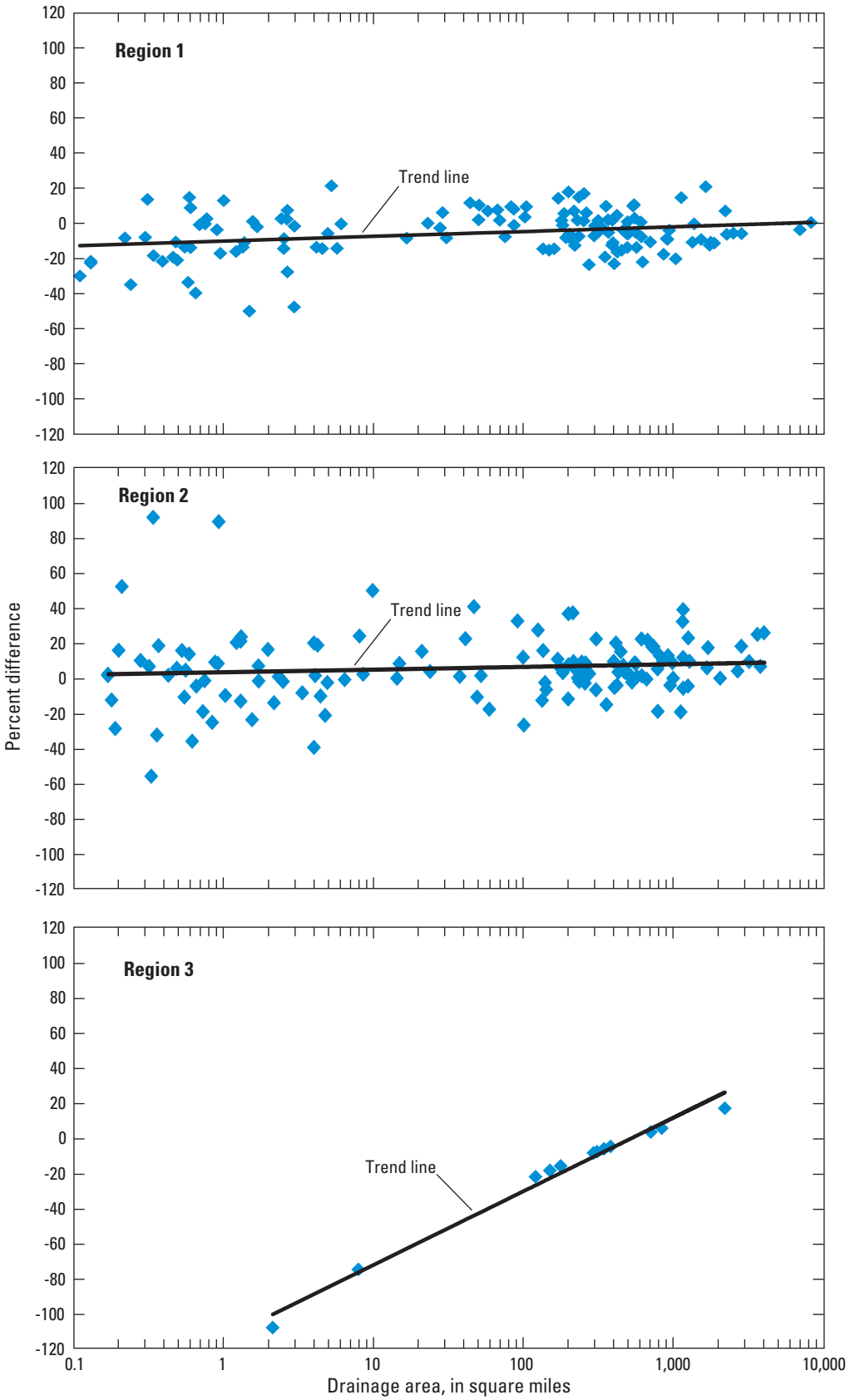


Figure 8. Percent difference by drainage area between 1-percent annual exceedance-probability estimates computed using regional regression equations developed in this study to those developed in Alexander and Wilson (1995) for 278 streamgages used in this study.

Drainage-Area Ratio

Locations on streams with streamgages may have estimates determined by area weighting the AEPDs on the basis of the drainage-area ratio between an ungaged site and a streamgage on the same stream. The weighting procedure is not applicable when the drainage-area ratio is less than 0.5 or greater than 1.5, or when the flood characteristics substantially change between sites (Eash and others, 2013). To compute the area-weighted estimate at the ungaged site, the WIE estimate for the streamgage must be computed then the area-weighted AEPD for the ungaged site, $Q_{P(u)aw}$, is then computed using the following equations:

$$Q_{P(u)aw} = \left(\frac{A_{(u)}}{A_{(g)}}\right)^{b} Q_{P(g)w}, \quad (7)$$

where

$Q_{P(u)aw}$ is the area-weighted estimate of flood discharge for the selected P-percent annual exceedance probability for the ungaged site, u, in cubic feet per second;

$A_{(u)}$ is the drainage area of the ungaged site, in square miles;

$A_{(g)}$ is the drainage area of the gaged site, in square miles;

$Q_{P(g)w}$ is described for equation 5; and

b is the exponent of drainage area from the appropriate P-percent annual exceedance probability regional exponent for the region (table 13).

A GLS analyses using only drainage area (DRNAREA) as an independent variable was performed to define the regional exponent for area-weighted estimates. Regional exponents ranged from 0.515 to 0.580 for Region 1, from 0.623 to 0.654 for Region 2, and from 0.665 to 0.718 for Region 3 (table 13).

Regional Regression Equations

The regional regression equations can be used if the ungaged site meets the criteria for use of this method and if the site is not at a streamgage or within a drainage-area ratio of 0.5 to 1.5 on the same stream. The equations presented in tables 7–9 are applicable for streams that are minimally affected by anthropogenic activities. The applicable range of basin characteristics for the equations for each region is listed in table 14. These equations are to be used with caution for the determination of statistics at ungaged locations for which the basin characteristics are outside the range of those used to develop the regression equations. Region 1 has two basin-characteristic ranges for applying the regional equations. For Region 1, the applicable range for drainage area is from 0.11 to 8,212.38 mi^2 and the applicable range for basin shape is from 2.25 to 26.59. Region 2 also has two basin-characteristics ranges for applying the regional equations. For Region 2, the applicable range for drainage area is from 0.17 to 4,008.92 mi^2 and the applicable range for basin shape is from 2.04 to 26.89. For Region 3, where only one basin characteristic was significant, the applicable range for drainage area is from 2.12 to 2,177.58 mi^2.

Largest Recorded Floods in Missouri

The largest recorded peak flow at a streamgage may be qualitatively assessed by comparison to the 0.2-percent AEPD regional regression with drainage area and the regional envelope curve. Relation between the largest recorded peak flow and drainage area for each of the three flood regions in Missouri is shown in figure 9. Peak discharges may be determined in one of three ways: (1) a direct measurement is made at or near the peak discharge (Rantz and others, 1982; Turnipseed and Sauer, 2010); (2) an indirect measurement is made after the flood event (Benson and Dalrymple, 1967); or (3) the stage-discharge rating is extended above the highest

Table 13. Regional exponents and constants determined from regional regression of log-transformed drainage area for area-weighting method to estimate annual exceedance-probability discharges for ungaged sites on gaged streams.

Annual exceedance probability (percent)	Region 1		Region 2		Region 3	
	Exponent *b*	Constant	Exponent *b*	Constant	Exponent *b*	Constant
50	0.580	2.421	0.654	2.372	0.665	1.933
20	0.557	2.697	0.643	2.663	0.681	2.026
10	0.546	2.832	0.637	2.806	0.689	2.070
4	0.536	2.968	0.632	2.949	0.698	2.113
2	0.529	3.050	0.629	3.036	0.703	2.139
1	0.524	3.121	0.626	3.110	0.708	2.162
0.5	0.520	3.182	0.625	3.173	0.713	2.182
0.2	0.515	3.253	0.623	3.246	0.718	2.204

Table 14. Range of basin-characteristic values used to develop regional annual exceedance-probability regression equations for unregulated streams in rural Missouri.

[GIS-derived, drainage area derived from a geographic information system; DRNAREA, GIS-derived drainage area; BSHAPE, basin shape; NA, not applicable—basin characteristic not used to develop regional regression equations]

	GIS-derived drainage area, DRNAREA (square mile)	Basin shape, BSHAPE (dimensionless)
Region 1		
Minimum	0.11	2.25
Maximum	8,212.38	26.59
Mean	465.42	8.41
Median	178.47	6.81
Number of sites	131	131
Region 2		
Minimum	0.17	2.04
Maximum	4,008.92	26.89
Mean	440.31	7.06
Median	141.84	5.98
Number of sites	135	135
Region 3		
Minimum	2.12	NA
Maximum	2,177.58	NA
Mean	456.16	NA
Median	298.29	NA
Number of sites	12	NA

measurement previously made but does not exceed two times this measurement.

An envelope curve based on annual peak flow data used in this study is presented for each region (fig. 9) along with the largest peak flow at each streamgage. The envelope curves indicate the largest peak flow potential based on recorded streamgage data (Crippen and Bue, 1977).

In Region 1, the largest peak flows from four streamgages (map numbers 33, 73, 90, and 118; table 1, fig. 9*A*) are shown greater than the envelope curves developed for this study. All four maximum flood peaks are based on indirect measurements and the estimation for these peaks was somewhat uncertain. Based on other peak flows derived from direct measurements, the curves were drawn slightly below the four peaks shown, but within the error range of the indirect measurements. Also, streamgages 73, 90, and 118 are located outside of the State of Missouri (fig. 1).

The regional regression equation curves using drainage area as the only independent variable for the $Q_{0.2\%}$ AEPD are shown for each flood region in figure 9. In Region 1, 27 out of 131 streamgages have maximum flood peaks plotting above the $Q_{0.2\%}$ AEPD curve and in Region 2, 23 out of 135 streamgages have maximum flood peaks plotting above the $Q_{0.2\%}$ AEPD curve. Region 3 with a limited data set, 2 out of 12 streamgages have maximum flood peaks plotting above the $Q_{0.2\%}$ AEPD curve. In all three regions with drainage areas greater than 1 square mile, extreme storm events have resulted in large peak flows in excess of the $Q_{0.2\%}$ AEPD when drainage area (DRNAREA) is the only variable used to define the relation.

A second set of envelope curves from Crippen and Bue (1977) also is shown based on all available data for all active or discontinued, unregulated streamgages and ungaged sites through water year 1974. The envelope curves developed by Crippen and Bue (1977) for Missouri include Regions 9, 8, and 3 in a nationwide study using 883 sites with drainage areas less than 10,000 mi^2. The Nation was grouped into 17 regions based initially on physiographic type, variations in rainfall intensity, and hydrologic judgment. Crippen and Bue (1977) divided Missouri into three flood-region boundaries similar geographically to the three regions defined in this study. Data from sites from Oklahoma, Kansas, and Texas resulted in higher envelope curves than the set of curves developed solely on data from Missouri (fig. 9).

Each method used to determine the maximum flood discharge has uncertainty in the accuracy of the computed peak flow. Direct measurements have less uncertainty than indirect measurements. A direct measurement, rated as fair for accuracy, is considered to be plus or minus 8 percent of the actual discharge (Rantz and others, 1982). An indirect measurement with a similar rating is considered to be plus or minus 15 percent of the actual flow (Benson and Dalrymple, 1967). The accuracy of a peak flow determined from a stage-discharge rating extension will be dependent on the type of measurements used to develop the stage-discharge rating and the consistency of the flow characteristics at the stage of the highest measurement used to define the stage-discharge rating and the stage of the peak flow. A histogram showing the ratio of maximum peak flow above the largest discharge measurement to direct or indirect discharge measurements for the 66 active (2012) streamgages in Missouri with no historical peaks is shown in figure 10. The median extension of the rating curves is 1.25 times the measurement. The minimum extension above the measurement is 0.9 and the maximum extension is 2.1 times the measurement. Of the 66 streamgages evaluated, 18 streamgages have rating extensions 1.5 times or greater than the highest measurement (fig. 10).

The largest peak flows maybe viewed in terms of their magnitude chronologically and spatially. Large widespread flood events will affect many streamgages within an area. For this study, the largest recorded peak flow was determined for all streamgages and the number of years of record for each streamgage was summed for each water year the largest peak flow took place. As an example, for streamgage 07013000 (map number 172, table 1) the largest recorded peak flow was in water year 1915 and the station has 98 years of record (table 5). Another streamgage 07016000 (map number 178) also had the largest recorded peak flow in 1915 and this streamgage had 68 years of record (table 5). Thus, the

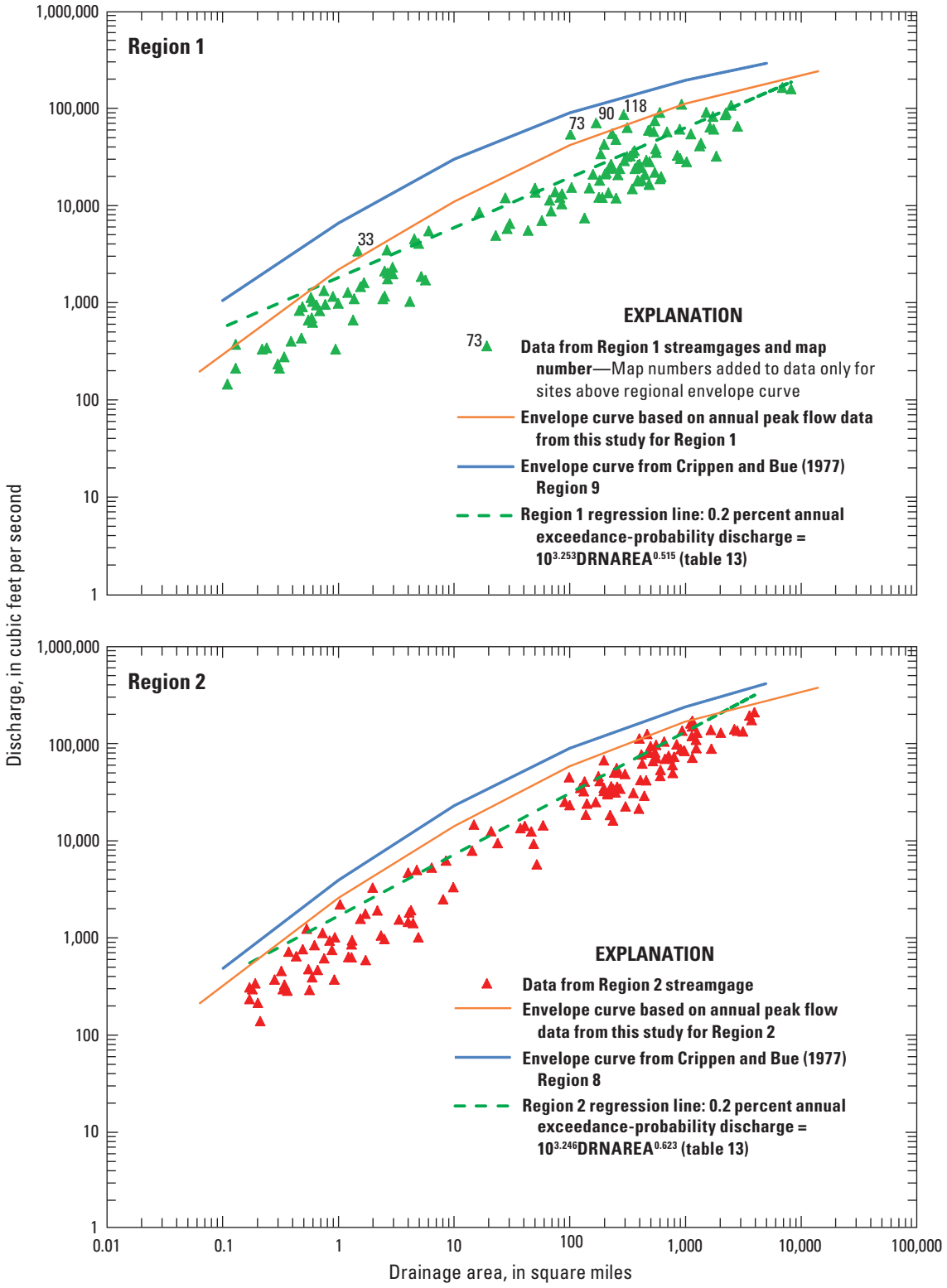


Figure 9. Relation between largest peak flow and drainage area (DRNAREA) for streams in Region 1, Region 2, and Region 3.

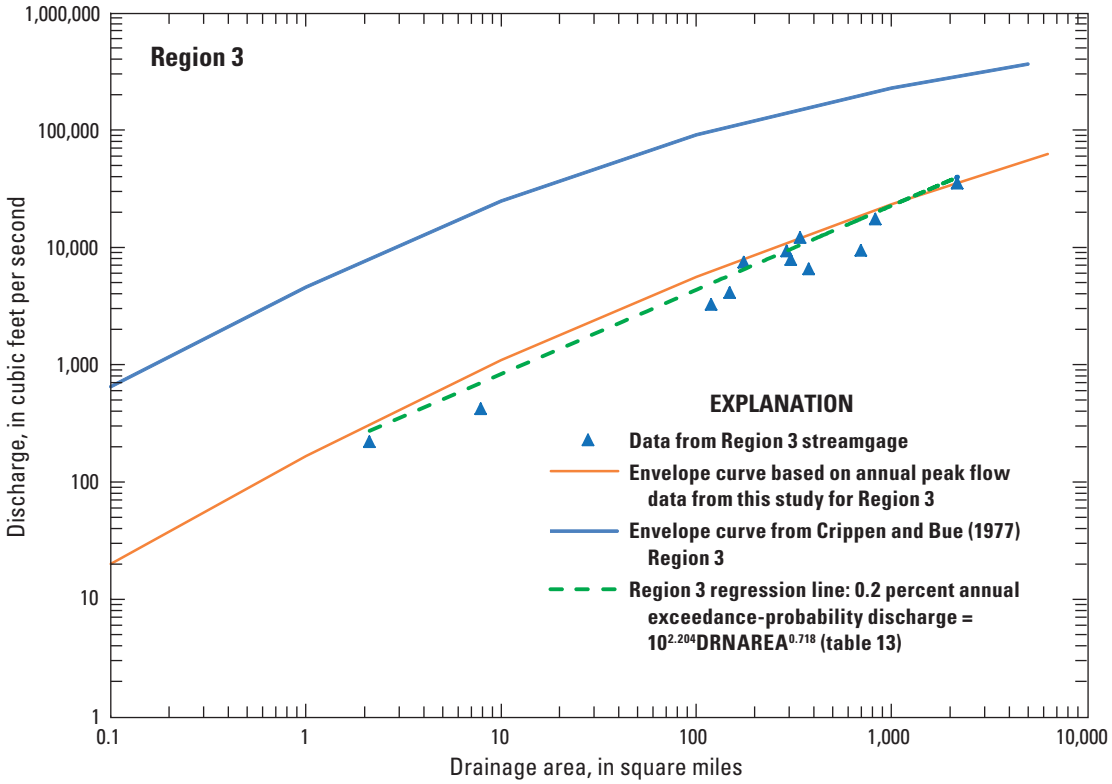


Figure 9. Relation between largest peak flow and drainage area (DRNAREA) for streams in Region 1, Region 2, and Region 3.—Continued

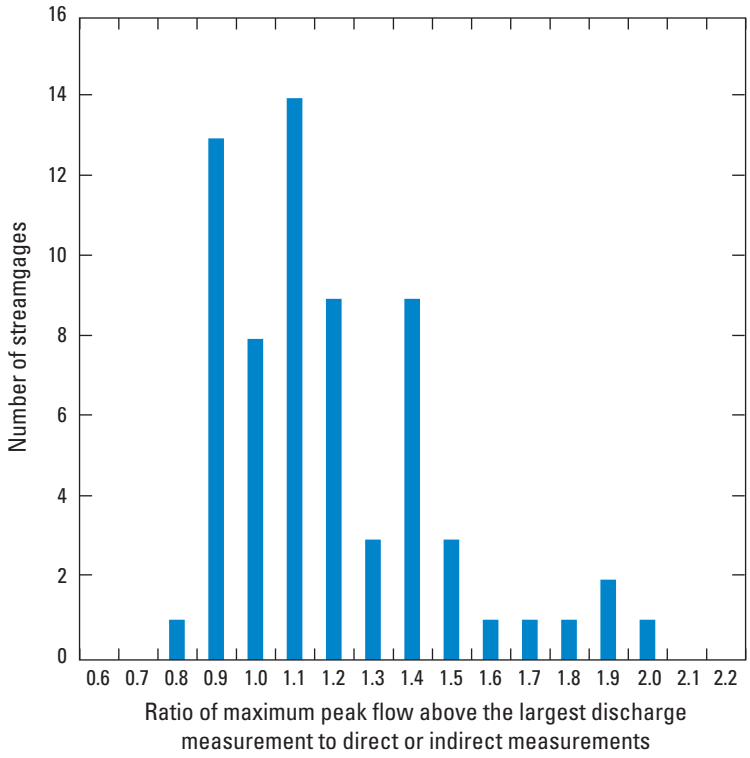


Figure 10. Ratio of maximum annual peak flow to the largest direct or indirect discharge measurement for 66 streamgages used in this study that are currently active (2012) in Missouri.

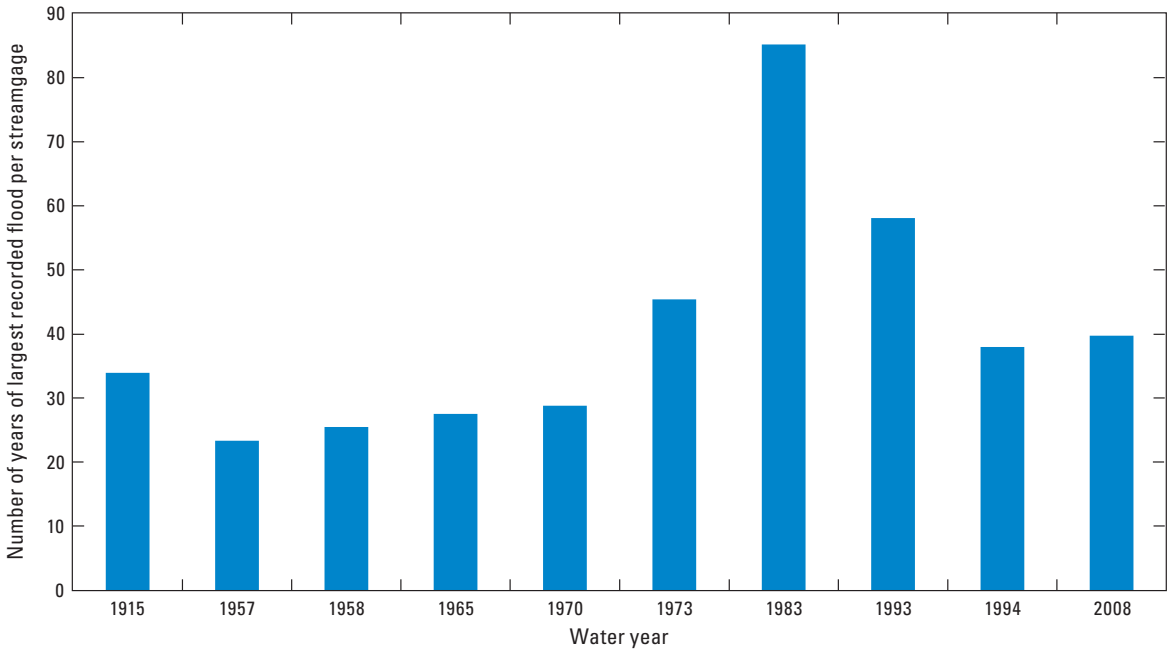


Figure 11. Largest floods in Missouri based on U.S. Geological Survey peak-flow data, water years 1915 through 2008.

total-station years of record for the 1915 peak flow was 166 years. Summing the station years for 21 stations used in this study and dividing by the number of stations where the 1915 event was the largest known flood results in an average of 34 years per station (fig. 11) and is the sixth largest flood event since 1915. The 2008 flood event was the fourth largest and most recent flood event recorded in Missouri. The largest flood event, in years per station, was the 1983 flood at 85 years per station. In decreasing magnitude, the following list summarizes the six largest floods in the last 100 years: 1983, 1993, 1973, 2008, 1994, and 1915. Five of the six largest floods happened in the last 40 years (1973–2012).

Summary

In 2008, Missouri experienced the wettest calendar year on record since 1895. The U.S. Geological Survey, in cooperation with the Missouri Department of Transportation and the Federal Emergency Management Agency, initiated a statewide study in 2010 to update the rural flood frequency equations using annual flood peaks through water year 2012. To improve the final results of the regression equations, a Bayesian weighted least-squares/generalized least-squares regression was implemented to update the generalized skew map for Missouri. Also, two new statistical techniques, the expected moments algorithm and the multiple Grubbs-Beck test, were used to determine the at-site magnitude and frequency estimates for peak-flow data. The multiple Grubbs-Beck test allowed for the detection of multiple potentially influential low floods (PILFs) compared to the Grubbs-Beck test as is currently (2014) recommended in Bulletin 17B. The expected moments algorithm is an updated method for fitting the frequency curve that has been shown to be a more effective means of incorporating historical flood information into a flood-frequency analysis and EMA is consistent with Bulletin 17B methods.

Preliminary statewide regression analyses indicated that the three primary physiographic provinces (Central Lowlands, Ozark Plateaus, and Mississippi Alluvial Plain) had a pronounced effect on peak flow values. Regional regression analyses were initially performed using the low-flow regional boundaries defined by Southard (2013) and the regional boundary defined by Alexander and Wilson (1995) for Crowley's Ridge located in the Mississippi Alluvial Plain. Residuals computed from observed minus predicted peak flow values were examined for possible bias and no geographic bias was determined to exist. The basin characteristics used in the analyses were from Southard (2013) and for Region 1 (Central Lowlands) and Region 2 (Ozark Plateaus) the statistically significant independent variables were drainage area and basin shape. For Region 3 (Mississippi Alluvial Plain) the only statistically significant variable was drainage area. A total of 278 streamgages were used in the regional analyses with 131 streamgages in Region 1, 135 streamgages in Region 2, and 12 streamgages in Region 3.

The generalized-least squares multiple-linear regression was used to compute the final regression coefficients and the measures of accuracy for each set of regional equations. The program weighted-multiple-linear-regression model program was used to perform the generalized-least squares regression technique in each region. Regression analyses were performed on for the selected annual exceedance probabilities of 50, 20, 10, 4, 2, 1, 0.5, and 0.2 percent. The standard error of predictions ranged from 28.7 to 38.4 percent in Region 1, 24.1 to

43.5 percent for Region 2, and 25.8 to 30.5 percent for Region 3. Comparing $Q_{1\%}$ estimates for a range of 3 to 2,100 mi^2 and a basin shape factor of 8 for Regions 1 and 2, the magnitude of the frequency statistic is the highest in Region 2 where relief is the greatest and main-channel gradients are the highest. Comparing the results of the regression equations for the $Q_{1\%}$ statistic from the 1995 flood frequesnvy study indicates that slightly lower estimates for this study exists in Region 1, slightly higher estimates for this study in Region 2, and lower estimates for this study in Region 3 for basins smaller than about 500 mi^2.

Three methods are proposed for computing estimates of annual exceedance probabilities at a site. If the site is at a streamgage with 10 or more years of record, improved estimates for the site can be obtained by weighting the annual exceedance probability log-Pearson Type III estimate with the regional regression-equation estimate by weighting the variance of prediction of each estimate. If the ungaged site is located on the same stream as a streamgage with 10 or more years of record, the estimate at the streamgage can be transferred to the site using a drainage area ratio (DAR). The DAR must range from 0.5 to 1.5. If the site does not meet the two above conditions, the regional regression equations presented in this report may be used assuming the basin characteristics of the site are within the ranges used to develop the regression equations. The equations developed for this study are applicable for streams that are not appreciably affected by storage, regulation, urbanization, or diversion.

The largest peak flows were compared to the 0.2-percent annual exceedance-probability discharge and envelope curves developed using data in this study and to curves from a previous study. In Region 1, 27 out of 131 streamgages have maximum flood peaks greater than the $Q_{0.2\%}$ estimate. Similarly, in Region 2, 23 out of 135 streamgages have greater than the $Q_{0.2\%}$ estimate, and in Region 3, 2 out of 12 streamgages have maximum flood peaks greater than the $Q_{0.2\%}$ estimate. The maximum flood event since 1915, based on a summation of years a given flood was the largest flood, was the 1983 flood in Missouri. The second and third largest floods in Missouri happened 1993 and 1973.

References Cited

Ahearn, E.A., 2010, Regional regression equations to estimate flow-duration statistics at ungaged stream sites in Connecticut: U.S. Geological Survey Scientific Investigations Report 2010–5052, 45 p. at *http://pubs.usgs.gov/sir/2010/5052/*.

Alexander, T.W., and Wilson, G.L., 1995, Techniques for estimating the 2- to 500-year flood discharges on unregulated streams in rural Missouri: U.S. Geological Survey Water-Resources Investigations Report 95–4231, 33 p.

Benson, M.A., and Dalrymple, Tate, 1967, General field and office procedures for indirect discharge measurements: U.S. Geological Survey Techniques of Water-Resources Investigations, book 3, chap. A1, 30 p. at *http:// pubs.usgs.gov/twri/twri3-a1/*.

Branner, G.C., 1937, Data on springs in Arkansas: Little Rock, Ark., Arkansas Geological Survey, 163 p.

Cavalieri, P., Jayawickrama, J., Luca, R., Patetta, M., Scott, K., and Walsh, S., 2000, Statistics I—Introduction to ANOVA, regression, and logistic regression: Cary, N. C., SAS Institute, Inc., 504 p.

Cohn, T.A., Lane, W.L., and Baier, W.G., 1997, An algorithm for computing moments-based flood quantile estimates when historical flood information is available: Water Resources Research, v. 33, no. 9, p. 2089–2096, accessed March 15, 2013, at *http://onlinelibrary.wiley.com/doi/10.1029/97WR01640/pdf*.

Cohn, T.A., Lane, W.L., and Stedinger, J.R., 2001, Confidence intervals for expected moments algorithm flood quantile estimates: Water Resources Research, v. 37, no. 6, p. 1695–1706, accessed March 15, 2013, at *http://timcohn.com/Publications/CohnLaneSted2001WR900016.pdf*.

Cohn, T.A., Berenbrock, Charles, Kiang, J.E., and Mason, R.R., Jr., 2012, Calculating weighted estimates of peak streamflow statistics: U.S. Geological Survey Fact Sheet 2012–2038, 4 p. at *http://pubs.usgs.gov/ fs/2012/3038/*.

Cook, R.D., 1977, Detection of influential observation in linear regression: Technometrics, v. 19, p. 15–18, accessed March 15, 2013, at *http://www.ime.usp.br/~abe/lista/pdfWiH1zqnMHo.pdf*.

Crippen, J.R. and Bue, C. D., 1977, Maximum floodflows in the conterminous United States: U.S. Geological Survey Water-Supply Paper 1887, 52 p.

Eash, D.A., Barnes, K.K., and Veilleux, A.G., 2013, Methods for estimating annual exceedance-probability discharges for streams in Iowa, based on data through water year 2010: U.S. Geological Survey Scientific Investigations Report 2013–5086, 63 p. with appendix.

Efroymson, M.A., 1960, Multiple regression analysis, *in* Ralston, A., and Wilf, H.S., eds., Mathematical methods for digital computers: New York, John Wiley and Sons, Inc., p. 191–203.

Eng, Ken, Chen, Yin-Yu, and Kiang, J.E., 2009, User's guide to the weighted-multiple-linear-regression program (WREG version 1.0): U.S. Geological Survey Techniques and Methods, book 4, chap. A8, 21 p. (Also available at *http://pubs.usgs.gov/tm/tm4a8/.)*

Esri, Inc., 2009, ArcGIS desktop help, accessed March 15, 2013, at *http://webhelp. esri.com/arcgisdesktop/9.3*.

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 3, South Carolina: U.S. Geological Survey Scientific Investigations Report 2009–5043, 238 p.

Federal Emergency Management Agency, 2002, National flood insurance program description: FEMA Federal Insurance and Mitigation Administration, 41 p., accessed March 15, 2013, at *http://www.fema.gov/library/ viewRecord.do?id=1480.*

Fenneman, N.M., 1938, Physiography of eastern United States: New York, McGraw-Hill, 714 p.

Freund, R.J., and Littell, R.C., 2000, SAS system for regression (3rd ed.): Cary, N. C., 235 p.

Gesch, D.B., 2007, The national elevation data set, *in* Maune, D. B., ed., Digital elevation model technologies and applications—The DEM user's manual (2d ed.): Bethesda, Md., American Society for Photogrammetry and Remote Sensing, p. 99–118, accessed March 15, 2013, at *http://topotools.cr.usgs.gov/pdfs/Gesch_Chp_4_Nat_Elev_Data_2007.pdf* .

Gotvald, A.J., Feaster, T.D., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey Scientific Investigations Report 2009–5043, 120 p. (Also available at *http://pubs.usgs.gov/sir/2009/5043/.*)

Gotvald, A.J., Barth, N.A., Veilleux, A.G., and Parrett, Charles, 2012, Methods for determining magnitude and frequency of floods in California, based on data through water year 2006: U.S. Geological Survey Scientific Investigations Report 2012–5113, 38 p., 1 pl. (Also available at *http://pubs.usgs.gov/sir/2012/5113/.*)

Griffis, V.W., and Stedinger, J.R., 2007, The use of GLS regression in regional hydrologic analyses: Journal of Hydrology, v. 344, p. 82–95, accessed March 15, 2013, at *http://www.sciencedirect.com/science/article/pii/S0022169407003848.*

Grubbs, F.E., and Beck, Glenn, 1972, Extension of sample sizes and percentage points for significance tests of outlying observations: Technometrics, v. 14, no. 4, p. 847–854. (Also available at *http://www.jstor.org/stable/1267134*.)

Hauth, L.D., 1974, Technique for estimating the magnitude and frequency of Missouri floods: U.S. Geological Survey Open-File Report 91–89, 20 p.

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. (Also available at *http://pubs.usgs.gov/twri/twri4a3/html/pdf_new.html.*)

Hershfield, D.M., 1961, Rainfall frequencies atlas of the United States for durations from 30 minutes to 24 hours and return periods from 1 to 100 years: Washington, D.C., U.S. Weather Bureau, Technical Paper No. 40, 115 p.

Holmes, R.R., Jr., Wiche, G.J., Koenig, T.A., and Sando, S.K., 2013, Peak streamflows and runoff volumes for the Central United States, February through September, 2011: U.S. Geological Survey Professional Paper 1798–C, 60 p., *http://pubs.usgs.gov/pp/1798c/*.

Missouri Department of Natural Resources, 2007, Missouri environmental geology atlas: Rolla, Mo., Division of Geology and Land Survey, CD-ROM.

Missouri Department of Transportation, 2013, The road ahead, 2013 Report to the joint committee on transportation oversight, City, State, Missouri Department of Transportation, 8 p. accessed December 31, 2013, at *http://www.modot.org/newsandinfo/reports/2013/index.htm*

Multi-Resolution Land Characteristics Consortium (MRLC), 2012, National Land Cover Database (NLCD): U.S. Geological Survey, accessed March 15, 2013, at *http://www.mrlc.gov/index.php.*

Natural Resources Conservation Service, 2012, Geospatial data gateway: U.S. Department of Agriculture, accessed March 30, 2012, at *http://datagateway.nrcs.usda.gov/.*

Parameter-Elevation Regressions on Independent Slopes Model Climate Group (PRISM), 2008, Normal annual precipitation grid for the conterminous United States, accessed March 15, 2013, at *http://www.prism.oregonstate.edu/state_products/maps.phtml?id=US, http://www.prism.oregonstate.edu/pub/prism/docs/prisguid.pdf.*

Price, C.V., Nakagaki, Naomi, and Hitt, K.J., 2010, National Water-Quality Assessment (NAWQA) area-characterization toolbox, release 1.0: U.S. Geological Survey Open-File Report 2010–1268, accessed November 20, 2012, at *http://pubs.usgs.gov/of/2010/1268*.

Rantz, S.E., and others, 1982, Measurement and computation of streamflow—Volume 1, Measurement of stage and discharge, and volume 2, Computation of discharge: U.S. Geological Survey Water-Supply Paper 2175, 631 p. (Also available at *http://pubs.usgs.gov/wsp/wsp2175/.*)

Sandhaus, E.H. and Skelton, John, 1968, Magnitude and frequency of Missouri floods: Division of Geology and Land Survey Water Resources Report 23, 276 p., 1 pl.

Searcy, J.K., 1955, Floods in Missouri, magnitude and frequency: U.S. Geological Survey Circular 370, 126 p.

Searcy, J.K., 1959, Flow-duration curves: U.S. Geological Survey Water Supply Paper 1542-A, 33 p.

Simley, J.D., and Carswell, W.J., Jr., 2009, The National Map—Hydrography: U.S. Geological Survey Fact Sheet 2009–3054, 4 p. (Also available at *http://pubs.usgs.gov/fs/2009/3054/, also see http://nhd.usgs.gov/*.)

Skelton, J.K., 1970, Base flow recession characteristics and seasonal low-flow frequency characteristics for Missouri streams: Missouri Geological Survey and Water Resources, Water Resources Report 25, 43 p.

Skelton, J.K., 1976, Missouri stream and springflow characteristics—Low-flow frequency and flow duration: Missouri Department of Natural Resources, Geological Survey, Water Resources 32, 76 p.

Southard, R.E., 2013, Computed statistics at streamgages, and methods for estimating low-flow frequency statistics and development of regional regression equations for estimating low-flow frequency statistics at ungaged locations in Missouri: U.S. Geological Survey Scientific Investigations Report 2013–5090, 28 p., *http://pubs.usgs.gov/sir/2013/5090/*.

Stedinger, J.R., and Tasker, G.D., 1985, Regional hydrologic analysis 1—Ordinary, weighted, and generalized least square compared: Water Resources Research, v. 21, no. 9, p. 1421–1432, accessed March 15, 2013, at *http://www.agu.org/journals/wr/v021/i009/WR021i009p01421/WR021i009p01421.pdf*.

TIBCO Software Inc., 2008, TIBCO Spotfire S+ 8.1 for Windows, user's guide: Palo Alto, California, 582 p., accessed March 15, 2013, at *http://stn.spotfire.com/pdfud/TIB_sf_splus_8.2.0_win_user_guide.pdf*

Turnipseed, D.P., and Sauer, V.B., 2010, Discharge measurements at gaging stations: U.S. Geological Survey Techniques and Methods book 3, chap. A8, 87 p. (Also available at *http://pubs.usgs.gov/tm/tm3-a8/*.)

U.S. Geological Survey, 2012, Peak streamflow for the Nation: U.S. Geological Survey, National Water Information System—Web Interface, data available on the World Wide Web, accessed March 15, 2012, at *http://nwis.waterdata.usgs.gov/usa/nwis/peak*.

U.S. Geological Survey, 2012, National hydrography data set: U.S. Geological Survey, accessed March 15, 2012, at *http://nhd. usgs.gov/*.

U.S. Geological Survey, 2011, National elevation data set: U.S. Geological Survey, accessed March 15, 2011, at *http://ned. usgs.gov/*.

U.S. Geological Survey and U.S. Department of Agriculture, Natural Resources Conservation Service, 2009, Federal guidelines, requirements, and procedures for the National Watershed Boundary Data set: U.S. Geological Survey Techniques and Methods, book 11, chap. A3, 55 p. (Also available at *http://pubs.usgs.gov/tm/tm11a3/*.) [also see *http://datagateway.nrcs.usda.gov/*.]

U.S. Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood flow frequency— Bulletin 17-B of the Hydrology Subcommittee: Reston, Va., U.S. Geological Survey, Office of Water Data Coordination 183 p. and appendixes, accessed March 15, 2013, at *http://water.usgs. gov/osw/bulletin17b/dl_flow.pdf*.

U.S. Interagency Advisory Committee on Water Data, 2014, Subcommittee on Hydrology, Hydrologic Frequency Analysis Work Group, Determining Flood Frequency using EMA, Frequency Asked Questions accessed July 30, 2014 *http://acwi.gov/hydrology/Frequency/b17_swfaq/EMAFAQ.html*

Veilleux, A.G., Stedinger, J.R., and Eash, D.A., 2012, Bayesian WLS/GLS regression for regional skewness analysis for regions with large crest stage gage networks, paper 2253, World Environmental and Water Resources Congress 2012: Crossing Boundaries, American Society of Civil Engineers, Albuquerque, N. M., May 20–24, 2012: Albuquerque, New Mexico, American Society of Civil Engineering, Paper 227, p. 2253–2263. (Also available at *http//ia.water.usgs.gov/media/pdf/report/Veilleux-Stedinger-Eash-EWRI-2012-227R.pdf*).

Veilleux, A.G., 2011, Bayesian GLS regression, leverage and influence for regionalization of hydrologic statistics: Cornell, Cornell University, Ph. D. dissertation, 184 p.

Veilleux, A.G., Stedinger, J.R., and Lamontagne, J.R., 2011, Bayesian WLS/GLS regression for regional skewness analysis for regions with large cross-correlations among flood flows, paper 1303, *in* World Environmental and Water Resources Congress 2011—Bearing Knowledge for Sustainability, Palm Springs, Calif., May 22–26, 2011, American Society of Civil Engineers, p. 3103–3112.

Veilleux, A.G., Cohn, T.A., Flynn, K.M., Mason Jr., R.R., and Hummel, P.R., 2014, Estimating Magnitude and Frequency of Floods using the PeakFQ 7.0 Program: U. S. Geological Survey Fact Sheet 2013–3108, 2 p., *http://dx.doi.org/10.3133/fs20133108*.

Verdi, R. J., and Dixon, J. F., 2011, Magnitude and frequency of floods for rural streams in Florida, 2006: U.S. Geological Survey Scientific Investigations Report 2011–5034, 69 p., 1 pl. at *http://pubs.usgs.gov/sir/2011/5034/*.

Weaver, J.C., Feaster, T.D., and Gotvald, A.J., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 2, North Carolina: U.S. Geological Survey Scientific Investigations Report, 2009–5158, 111p. (Also available at: *http://pubs.usgs.gov/sir/2009/5158/*.)

Appendix

Introduction to Statistical Analysis of Regional Skew

For the log-transformation of annual peak discharges, Bulletin 17B (U.S Interagency Advisory Committee on Water Data, 1982) recommends using a weighted average of the station-skew coefficient and a regional skew coefficient (equation 2 in this report) to help improve estimates of annual exceedance probability discharges or AEPDs. Bulletin 17B presents a national map, but also promotes development of maps that are more specific to local areas and for local relations. Since the first map was published in 1976, additional information has been collected and compiled, and better spatial estimation procedures have been developed (Stedinger and Griffis, 2008).

Tasker and Stedinger (1986) developed a weighted least-squares (WLS) procedure for estimating regional skew coefficients based on sample skew coefficients for the logarithms of annual peak-discharge data. Their method of regional analysis of skewness estimators accounts for the precision of the estimate of the skew coefficient for each streamgage or station, which depends on the length of record for each streamgage and the accuracy of an ordinary least-squares (OLS) regional mean skewness. More recently, Reis and others (2005), Gruber and others (2007), and Gruber and Stedinger (2008) developed a Bayesian generalized least-squares (GLS) regression model for regional skewness analyses. The Bayesian methodology allows for the computation of a posterior distribution of both the regression parameters and the model error variance. As shown in Reis and others (2005), for cases in which the model error variance is small compared to the sampling error of the station estimates, the Bayesian posterior distribution provides a more reasonable description of the model error variance than both the GLS method-of-moments and maximum likelihood point estimates (Veilleux, 2011). Whereas WLS regression accounts for the precision of the regional model and the effect of the record length on the variance of skew-coefficient estimators, GLS regression also considers the cross-correlations among the skew-coefficient estimators. In some studies the cross-correlations have had a large effect on the precision attributed to different parameter estimates (Eash and others, 2013; Parrett and others, 2011; Feaster and others, 2009; Gotvald and others, 2009; Weaver and others, 2009).

Because of complications introduced by the use of the expected moments algorithm (EMA) with the censoring of potentially influential low floods (PILFs) identified by the multiple Grubbs-Beck (MGB) test (Cohn and others, 1997) and large cross-correlations between annual peak discharges at pairs of streamgages, an alternate regression procedure was developed to provide both stable and defensible results for regional skewness (Veilleux and others, 2012; Veilleux, 2011; Lamontange and others, 2012). This alternate procedure is referred to as the Bayesian WLS/Bayesian GLS (B-WLS/B-GLS) regression framework (Veilleux and others, 2012; Veilleux, 2011; Veilleux and others, 2011). It uses an OLS analysis to fit an initial regional skewness model; that OLS analysis (model) is then used to generate a stable regional skew-coefficient estimate for each site. That stable regional estimate is the basis for computing the variance of each station skew-coefficient estimator employed in the WLS analysis. Then, B-WLS is used to generate estimators of the regional skew-coefficient model parameters. Finally, B-GLS is used to estimate the precision of those WLS parameter estimators, to estimate the model error variance and the precision of that variance estimator, and to compute various diagnostic statistics.

The U.S. Geological Survey (USGS) operates a large network of crest-stage gages (CSGs) that only record discharge above a minimum recording threshold and thus produce a censored data record. The CSGs are different from continuous-record streamgages, which measure almost all discharges and have been used in previous B-GLS and B-WLS/B-GLS regional skew studies. The Missouri regional skew study described herein did not exhibit large cross-correlations among annual-peak discharges; the study did make extensive use of EMA to estimate the station skew and its mean square error. Because EMA allows for the censoring of PILFs—as well as the use of estimated interval discharges for missing, censored, and historic data—it complicates the calculations of effective record length (and effective concurrent record length) used to describe the precision of sample estimators because the peak discharges are no longer solely represented by single values. To properly account for these complications, the new B-WLS/B-GLS procedure was used. The steps of this alternative procedure are described the following section, "Methodology for Regional Skewness Model".

Methodology for Regional Skewness Model

This section provides a brief description of the B-WLS/B-GLS methodology (Veilleux and others, 2012). Veilleux and others (2011) and Veilleux (2011) provide a more detailed description of the methodology.

Ordinary Least Squares Analysis

The first step in the B-WLS/B-GLS regional skewness analysis is the estimation of a regional skewness model using OLS. The OLS regional regression yields parameters $\hat{\boldsymbol{\beta}}_{OLS}$ and a model that can be used to generate unbiased and relatively stable regional estimates of the skewness data from all streamgages:

$$\tilde{\mathbf{y}}_{OLS}=\mathbf{X}\hat{\boldsymbol{\beta}}_{OLS} \quad \text{(A1)}$$

Here are the estimated regional skewness values, $\mathbf{X}$ is a (n x k) matrix of basin characteristics, n is the number of streamgages, and k is the number of basin parameters

including a column of ones to estimate the constant. These estimated regional skewness values $\tilde{\mathbf{y}}_{OLS}$ are then used to calculate unbiased station-regional skewness variances using the equations reported in Griffis and Stedinger (2009). These station-regional skewness variances are based on the regional OLS estimator of the skewness coefficient instead of the station skewness estimator, which makes the weights in the subsequent steps relatively independent of the station skewness estimates.

Weighted Least Squares Analysis

The B-WLS analysis is used to develop estimators of the regression coefficients for each regional skewness model (Veilleux, 2011; Veilleux and others, 2011). The WLS analysis explicitly reflects variations in record length, but does not take into account cross correlations thereby avoiding the problems experienced with GLS parameter estimators (Veilleux, 2011; Veilleux and others, 2011).

Generalized Least Squares Analysis

After the regression model coefficients, $\hat{\boldsymbol{\beta}}_{WLS}$, are determined with a WLS analysis, the precision of the fitted model and the precision of the regression coefficients are estimated using a B-GLS analysis (Veilleux, 2011; Veilleux and others, 2011). Precision metrics include the standard error of the regression parameters, $SE(\hat{\boldsymbol{\beta}}_{WLS})$, the model error variance, $\sigma^2_{\delta,B\text{-}GLS}$, pseudo-$R^2_\delta$ as well as the average variance of prediction at a streamgage that is not used in the regional model, AVP_{new}.

Data Analysis

The statistical analysis of the data requires several steps. This section describes the redundant site analysis, the calculations for both pseudo record length for each site given the number of censored observations and concurrent record lengths, as well as the development of the model of cross-correlations of concurrent annual-peak discharges.

Data for Missouri Regional Skew Study

This study is based on annual peak-discharge data from 302 streamgages in Missouri and the surrounding states of Iowa, Arkansas and Kansas. The annual peak-discharge data through September 30, 2010 were downloaded from the USGS National Water Information System (NWIS) database (U.S. Geological Survey, 2012). In addition to the peak-discharge data, 34 basin characteristics for each of the 302 sites were available as explanatory variables in the regional skew study. The basin characteristics available include three physiographic regions (fig. 2), as well as the more standard morphometric characteristics such as location of the basin centroid, drainage area, main basin slope, and mean channel elevation among others.

Station Skewness Estimators

To estimate the station logarithm base 10 (log10) skew coefficient, G_S, and its mean square error, (MSE_{GS}) the analysis used the EMA (Cohn and others, 1997; Griffis and others, 2004). EMA provides a straightforward and efficient method for incorporating historic information and censored data, such as those from a CSG, contained in the record of annual peak discharges for a streamgage. PeakfqSA, an EMA software program developed by Cohn (2011), is used to generate the station log10 estimates of G_S and its MSE_{GS}, assuming an LP3 distribution and using a MGB test for PILF screening. EMA estimates, based on annual peak-discharge data through September 30, 2010, of G_S and its MSE_{GS} are listed in table 1–1 for the 302 streamgages evaluated for the Missouri regional skew study (see sections "Expected Moments Algorithm (EMA) Analyses" and "Multiple Grubbs-Beck Test for Detecting PILFs" in the main part of this report for more detail regarding EMA and the multiple Grubbs-Beck test.)

Pseudo Record Length

Because the data set includes censored data and historic information, the effective record length used to compute the precision of the skewness estimators is no longer simply the number of annual peak- discharge data collected at a streamgage. Instead, a more complex calculation should be used to take into account the availability of historic information and censored values. Historic information and censored peaks provide valuable information, but they can provide less information than an equal number of years with systematically recorded annual peaks (Stedinger and Cohn, 1986). The following calculations provide a pseudo record length, P_{RL}, which appropriately accounts for all peak-discharge data types available for a site. P_{RL} equals the systematic record length if such a complete record is all that is available for a site.

The first step is to run EMA with all available information, including historic information and censored peaks (denoted EMA_C, for EMA complete). From the EMA run, the station skewness without regional information, $\hat{G}_C$, and the MSE of that skewness estimator, $MSE(\hat{G}_C)$ are extracted, as well as the year the historical period begins, YB_C, the year the historical period ends YE_C, and the length of the historical period H_C. (YB_C, YE_C, and H_C are used in equation A11.)

The second step is to run EMA with only the systematic peaks (denoted EMA_S, for EMA systematic). From the EMA_S analysis, the station skewness without regional information, $\hat{G}_S$, and the MSE of that skewness estimator, $\hat{G}_{sys}$, are extracted, as well as the number of peaks, P_{sys} (P_{sys} is used in equation A4.)

The third step is to represent, from both EMA_C and EMA_{sys}, the precision of the skewness estimators as two record lengths, RL_C and RL_{sys}, based upon the estimated skew and MSE. The corresponding record lengths (RL) are calculated using equations from Griffis and others (2004) and Griffis and Stedinger (2009):

$$MSE(\widehat{G}) = \left[\frac{6}{\mathrm{RL}} + a(RL)\right] * \left[1 + \left(\frac{9}{6} + b(RL)\right)\widehat{G}^2 + \left(\frac{15}{48} + c(RL)\right)\widehat{G}^4\right] \quad \text{(A2)}$$

$$a(RL) = -\frac{17.75}{RL^2} + \frac{50.06}{RL^3}$$

$$b(RL) = \frac{3.93}{RL^{0.3}} - \frac{30.97}{RL^{0.6}} + \frac{37.1}{RL^{0.9}}$$

$$c(RL) = \frac{6.16}{RL^{0.56}} - \frac{36.83}{RL^{1.12}} + \frac{66.9}{RL^{1.68}}$$

where RL and $\widehat{G}$ use appropriate subscripts for RL_C and RL_{sys}; for example, RL_C uses $\widehat{G}_C$ and $MSE(\widehat{G}_C)$, and RL_{sys} uses $\widehat{G}_{sys}$ and $MSE(\widehat{G}_{sys})$, and a(RL), $b(RL)$, and $c(RL)$ are variables described in the equation A2.

Next, the difference between RL_C and RL_{sys} is used as a measure of the extra information provided by the historic and/or censored information that was included in the EMA_c analysis, but not in the EMA_{sys} analysis:

$$RL_{diff} = RL_C - RL_{sys} \quad \text{(A3)}$$

The pseudo record length for the entire record for the streamgage, P_{RL}, is calculated using RL_{diff} from equation A3 and the number of systematic peaks P_{sys}:

$$P_{RL} = RL_{diff} + P_{sys} \quad \text{(A4)}$$

P_{RL} must be nonnegative. If P_{RL} is greater than H_C, then P_{RL} should be set to equal H_C. Also, if P_{RL} is less than P_{sys}, then P_{RL} is set to P_{sys}. This ensures that the pseudo record length will not be larger than the complete historical period or less than the number of systematic peaks.

As stated in Bulletin 17B, the skew coefficient of the station skew is sensitive to extreme events and more accurate estimates can be obtained from longer records. Thus, after ensuring adequate special and hydrologic coverage those gage sites that do not have a minimum of 30 years of pseudo record length were removed from the regional skew study. Of the 302 sites, 69 were removed because their P_{RL} was less than 20 years, 46 were removed because their P_{RL} was between 20 and 24 years, and 45 were removed because their P_{RL} was between 25 and 29 years. Thus, data from 142 streamgages remained from which to build a regional skewness model for the Missouri study area.

Redundant Sites

Redundancy results when the drainage basins of two streamgages are nested, meaning that one basin is contained inside the other and the two basins are of similar size. Instead of providing two independent spatial observations that depict how drainage basin characteristics are related to skew (or AEPs), these two basins will have the same hydrologic response to a given storm, and thus represent only one spatial observation. When streamgages in basins (site pairs) are redundant, a statistical analysis using both streamgages incorrectly represents the information in the regional data set (Gruber and Stedinger, 2008). To determine if two sites are redundant and thus represent the same hydrologic conditions, two types of information are considered: (1) whether their basins are nested, and (2) the ratio of the basin drainage areas.

The standardized distance (SD), is used to determine the likelihood that the basins are nested. The standardized distance between two basin centroids, SD is defined as:

$$SD_{ij} = \frac{D_{ij}}{\sqrt{0.5(DRNAREA_i + DRNAREA_j)}} \quad \text{(A5)}$$

where

D_{ij} is the distance between centroids of basin i and basin j; and
$DRNAREA_i$ is the drainage area at site i; and
$DRNAREA_j$ is the drainage area at site j.

The drainage area ratio (DAR) is used to determine if two nested basins are sufficiently similar in size to conclude that they are, or are at least in large part, the same watershed for the purposes of developing a regional hydrologic model. The DAR is defined as (Veilleux, 2009):

$$DAR = Max\left[\frac{DRNAREA_i}{DRNAREA_j}, \frac{DRNAREA_j}{DRNAREA_i}\right] \quad \text{(A6)}$$

where

DAR is the Max (maximum) of the two values in brackets;
$DRNAREA_i$ is the drainage area at site i; and
$DRNAREA_j$ is the drainage area at site j.

Two basins might be expected to have possible redundancy if the basin sizes are similar and the basins are nested. Previous studies suggest that site pairs having SD less than or equal to 0.50 and DAR less than or equal to 5 were likely to have possible redundancy problems for purposes of determining regional skew. If DAR is large enough, even if the site pairs are nested, they will reflect different hydrologic responses because storms of different sizes and durations will affect each site differently.

Table 1–1 (*http://pubs.usgs.gov/sir/2014/5165/Downloads/table_1-1.xlsx*) shows the results of the redundant site screening on the Missouri regional skew data. All possible combinations of site pairs from the 142 streamgages were considered in the redundancy analysis. In order to be conservative, all site pairs with $SD < 0.75$ and $DAR < 8$ were identified as possible redundant site-pairs. All site pairs identified as redundant were then investigated to determine if, in fact, one site of the pair is nested inside the other. For site pairs that are nested, one site from the pair was removed from the regional skew analysis. Sites removed from the Missouri regional skew study because of redundancy are identified in table 1–1.

From the 104 identified possible redundant site-pairs, 78 were determined to be redundant but only 34 sites were actually removed from the analyses as the same sites appeared in multiple site pairs. Thus, of the 142 sites, 34 were removed because of redundancy, which left 108 sites to use in the Missouri regional skew study.

Unbiasing the Station Estimators

By using the correction factor developed by Tasker and Stedinger (1986) and used by Reis and others (2005), unbiased estimates of the station skewness can be calculated. The unbiased station skewness estimator using the pseudo record length can be determined:

$$\hat{Y}_i = \left[1 + \frac{6}{P_{RL,i}}\right] G_i \tag{A7}$$

Here $\hat{Y}_i$ is the unbiased station sample skewness estimate for site i, $P_{RL,i}$ is the pseudo record length for site i as calculated in equation A6, and G_i is the traditional biased station skewness estimator for site i from EMA.

The variance of the unbiased station skewness also includes the correction factor developed by Tasker and Stedinger (1986):

$$VAR[\hat{Y}_i] = \left[1 + \frac{6}{P_{RL,i}}\right]^2 \mathrm{Var}[G_i] \tag{A8}$$

where

$VAR[G_i]$ is calculated using (Griffis and Stedinger, 2009)

$$Var(\hat{G}) = \left[\frac{6}{P_{RL}} + a(P_{RL})\right] * \left[1 + \left(\frac{9}{6} + b(P_{RL})\right)\hat{G}^2 + \left(\frac{15}{48} + c(P_{RL})\right)\hat{G}^4\right] \tag{A9}$$

where

$$a(P_{RL}) = -\frac{17.75}{P_{RL}^{\ 2}} + \frac{50.06}{P_{RL}^{\ 3}}$$

$$b(P_{RL}) = \frac{3.92}{P_{RL}^{\ 0.3}} - \frac{31.10}{P_{RL}^{\ 0.6}} + \frac{34.86}{P_{RL}^{\ 0.9}}$$

$$c(P_{RL}) = \frac{7.13}{P_{RL}^{\ 0.59}} - \frac{45.90}{P_{RL}^{\ 1.18}} + \frac{86.50}{P_{RL}^{\ 1.77}}$$

Estimating the Mean Square Error of the Skewness Estimator

There are several possible ways to estimate MSE_{G_S}. The approach used by EMA (Cohn and others (2001; equation 55) generates a first order estimate of the MSE_{G_S}, which should perform well when interval data are used. Another option is to use the Griffis and Stedinger (2009) formula in equation A8 (the variance is equal to the MSE), using either the systematic record length or the length of the whole historical period. However, this method does not account for censored data, and thus, can lead to inaccurate and underestimated. This issue has been addressed by using the pseudo record length instead of the length of the historical period; the pseudo record length reflects the effect of the censored data and the number of recorded systematic peaks. Thus, the unbiased Griffis and Stedinger (2009) MSE_G is used in the regional skewness model because it is more stable and relatively independent of the station skewness estimator. This methodology was used in the Iowa regional skew study (Each and others, 2013).

Cross-Correlation Models

A critical step for a GLS analysis is estimation of the cross-correlation of the skewness coefficient estimators. Martins and Stedinger (2002) used Monte Carlo statistical methods to derive a relation between the cross-correlation of the skewness estimators at two stations, i and j,as a function of the cross-correlation of concurrent annual maximum flows, ρ_{ij}:

$$\hat{\rho}(\hat{y}_i, \hat{y}_i) = Sign(\hat{\rho}_{ij}) cf_{ij} \left|\hat{\rho}_{ij}\right|^K \tag{A10}$$

where

$\hat{\rho}_{ij}$ is the cross-correlation of concurrent annual peak discharge for two streamgages;

κ is the constant between 2.8 and 3.3, and

cf_{ij} is the factor that accounts for the sample size difference between stations and their concurrent record length, and is defined as follows:

$$cf_{ij} = CY_{ij} / \sqrt{(P_{RL,i})(P_{RL,j})} \tag{A11}$$

where

CY_{ij} is the pseudo record length of the period of concurrent record; and

$P_{RL,i}$, $P_{RL,j}$ is the pseudo record length corresponding to sites i and j, respectively (see equation A6).

Pseudo Concurrent Record Length

After calculating the P_{RL} for data from each streamgage in the study, the pseudo concurrent record length between site pairs can be calculated. Because of the use of censored data and historic data, the effective concurrent record length calculation is more complex than determining in which years the two streamgages (site pairs) both have recorded systematic peaks.

The years of historical record in common between the two streamgages is first determined. For the years in common, with beginning water year YB_{ij} and ending water year YE_{ij}, the following equation is used to calculate the concurrent years of record between site i and site j:

$$CY_{ij} = (YE_{ij} - YB_{ij} + 1)\left(\frac{P_{RL,i}}{H_{C,i}}\right)\left(\frac{P_{RL,j}}{H_{C,j}}\right) \tag{A12}$$

The computed pseudo concurrent record length depends on the years of historical record in common between the two streamgages, as well as the ratios of the pseudo record length to the historical record length for each of the two streamgages.

Missouri Study Area Cross-Correlation Model of Concurrent Annual Peak Discharge

A cross-correlation model for the logarithm of the annual peak discharges in the Missouri study area were developed using 42 sites with at least 60 years of concurrent systematic peak discharge, zero discharge not included, (which resulted in 651 station pairs). Various models relating the cross-correlation of the concurrent annual peak discharge at two sites, $\rho_{ij,}$ to various basin characteristics were considered. A logit model, termed the Fisher Z Transformation ($Z = \log[(1+r)/(1-r)]$), provided a convenient transformation of the sample correlations r_{ij} from the (-1, +1) range to the (negative infinity, positive infinity) range (Devore, 2004). The adopted model for estimating the cross-correlations of concurrent annual peak discharge at two stations, which used the distance between basin centroids, D_{ij}, as the only explanatory variable, is

$$\rho_{ij} = \frac{\exp(2Z_{ij}) - 1}{\exp(2Z_{ij}) + 1} \quad \text{(A13)}$$

where

$$Z_{ij} = exp\left(0.59 - 0.066\left(\frac{D_{ij}^{0.53}-1}{0.53}\right)\right) \quad \text{(A14)}$$

An OLS regression analysis based on the 651 station pairs indicated that this model is as accurate as having 230 years of concurrent annual peak discharges from which to calculate cross correlation. Figure 1–1 shows the relation between the Fisher Z transformed cross-correlation of logs of annual peak discharge and distance between basin centroids for the 651 station pairs (Figure 1–2 shows the functional relation between the untransformed cross correlation and distance between basin centroids together with the plotted sample data from the 651 station pairs of data. The cross correlation model was used to estimate site-to-site cross correlations for concurrent annual peak discharges at all pairs of sites in the regional skew study.

Missouri Regional Skew Study Results

The results of the Missouri regional skew study using the B-WLS/B-GLS regression methodology are provided below. All of the available basin characteristics were initially considered as explanatory variables in the regression analysis for regional skew. There are a wide array of basin characteristics available for use in the regional skew study including: morphometric (drainage area, basin slope, shape factor, longest flow path, stream slope) pedologic or geologic (soil, hydraulic conductivity, wetlands, permeability, overburden, sinkholes, and springs), precipitation (mean annual, maximum 24 hours over several years), urban (impervious area). None of the basin characteristics were statistically significant in explaining the site-to-site variability in skewness. Thus, the best model, as classified by having the smallest model error variance, , and pseudo (the fraction of the variability in the true skews explained by the model), is the Constant model. Table 1–2 provides the final results for the constant skewness.

Table 1–2. Regional skewness models for Missouri study area.

[Standard deviation in parentheses. $\hat{y}$, estimated regional skew; b_1, estimated regression parameter; σ_δ^2, model error variance; ASEV, average sampling error variance; AVP_{new}, average variance of prediction for a new site; Pseudo-R_δ^2, fraction of the variability in the true skews explained by each model (Gruber and others, 2007); %, percent]

Model	Regression parameter b_1	σ_δ^2	ASEV	AVP_{new}	Pseudo-R_δ^2
Constant: $\hat{y} = b_1$	-0.3	0.13	0.01	0.14	0%
	(0.1)	(0.03)			

Table 1–2 includes data about the pseudo-R_δ^2, which describes the estimated fraction of the variability in the true skewness from site-to-site explained by each model (Gruber and others, 2007; Parrett and others, 2011). A constant model does not explain any variability, so the pseudo-R_δ^2 equals 0. The posterior mean of the model error variance, σ_δ^2, for the Constant model is $\sigma_\delta^2 = 0.13$.

The addition of any of the available basin characteristics (none of which are statistically significant) did not produce a pseudo-R_δ^2 greater than 5 percent or decrease the model error variance. This indicates that the inclusion of a basin characteristic as an explanatory variable in the regression did not help explain the variability in the true skewness. The addition of a basin characteristic is not warranted as the increased model complexity does not result in a gain in model precision. Thus, the Constant model is chosen as the best regional skewness model for the Missouri study area. The average sampling error variance (ASEV) in table 1–2 is the average error in the regional skewness estimator at the sites in the dataset. The average variance of prediction at a new site (AVP_{new}) corresponds to the mean square error (MSE) used in Bulletin 17B to describe the precision of the generalized skewness. The Constant model has an AVP_{new}, equal to 0.14, which corresponds to an effective record length of 54 years. An AVP_{new} of 0.14 is a marked improvement over the Bulletin 17B national skew map, whose reported MSE is 0.302 (Interagency Committee on Water Data, 1982) for a corresponding effective record length of only 17 years. Thus, the new regional model has three times the information content (as measured by effective record length) of that calculated for the Bulletin 17B map.

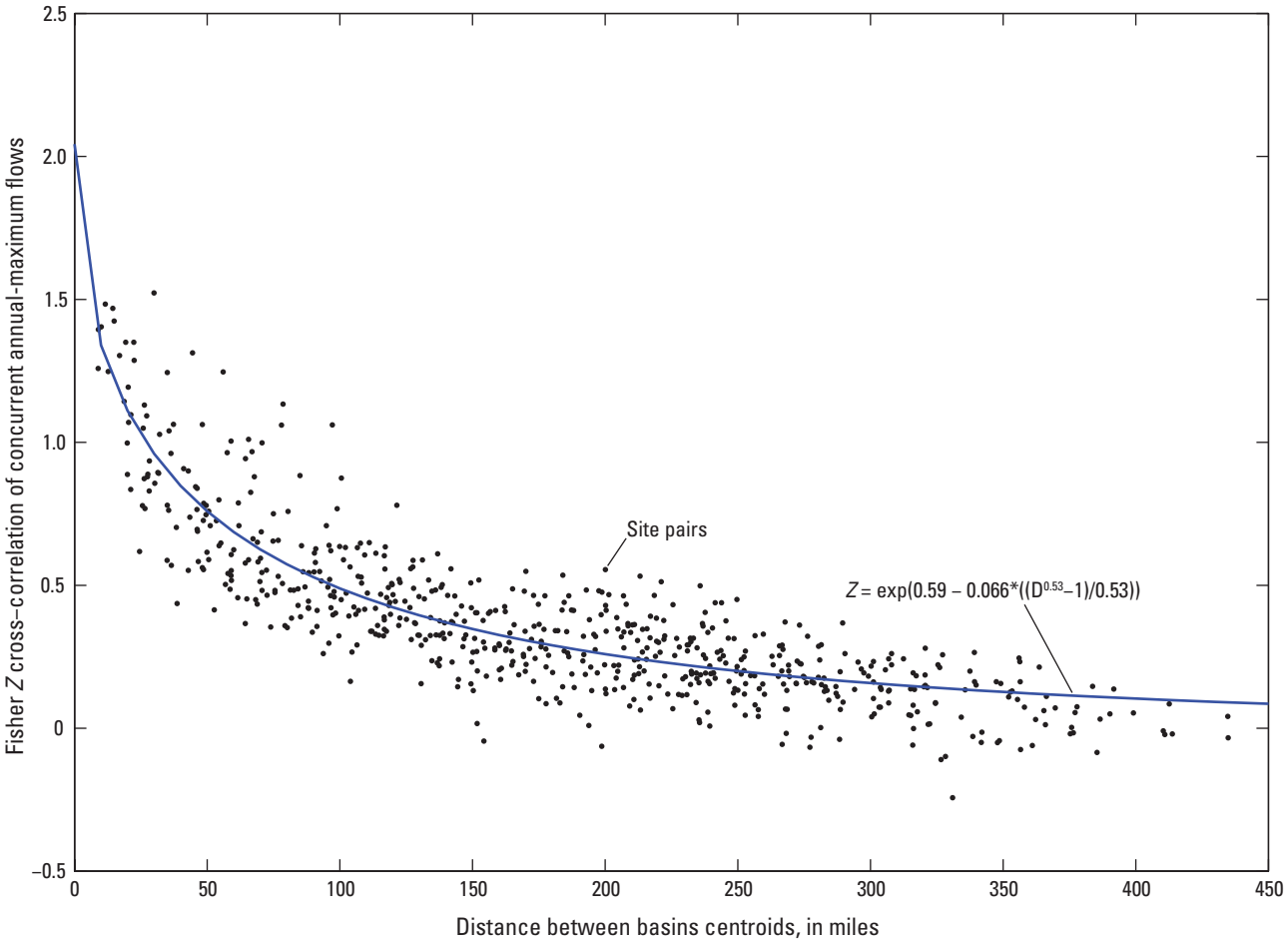


Figure 1–1. Relation between Fisher Z transformed (Z) cross-correlation of logs of annual peak discharge and distance (D) between basin centroids for 651 station-pairs with concurrent record lengths greater than or equal to 60 years from 42 streamgages in Missouri and neighboring States.

Bayesian-Weighted Least Squares/Bayesian-Generalized Least Squares Regression Diagnostics

To determine if a model is a good representation of the data and which regression parameters, if any, should be included in a regression model, diagnostic statistics have been developed to evaluate how well a model fits a regional hydrologic data set (Griffis, 2006; Gruber and others, 2008). In this study, the goal was to determine the set of possible explanatory variables that best fit annual peak discharges for the Missouri study area affording the most accurate skew prediction and keeping the model as simple as possible. This section presents the diagnostic statistics for a B-WLS/B-GLS analysis, and discusses the specific values obtained for the Missouri regional skew study.

Table 1–3 presents a pseudo Analysis of Variance (pseudo ANOVA) table for the Missouri regional skew analysis. The table contains regression diagnostics and goodness of fit statistics, which are explained below.

In particular, the table describes how much of the variation in the observations can be attributed to the regional model, and how much of the residual variation can be attributed to model error and sampling error, respectively Determining these quantities is difficult. The model errors cannot be resolved because the values of the sampling errors η_i for each site *i,* are not known. However, the total sampling error sum of squares can be described by its mean value, $\sum_{i=1}^{n} Var[\hat{\gamma}_i]$. Because there are *n* equations, the total variation due to the model error δ for a model with *k* parameters has a mean equal to $n\sigma_\delta^2(k)$. Thus, the residual variation attributed to the sampling error is, and the residual variation attributed to the model error is $n\sigma_\delta^2(k)$.

For a model with no parameters other than the mean (that is the Constant model), the estimated model error variance $\sigma_\delta^2(0)$ describes all of the anticipated variation in $\gamma_i = \mu + \delta_i$, where μ is the mean of the estimated station sample skews. The total expected sum of squares variation (table 1–3) due to model error δ_i and due to sampling error in expectation should equal $n\sigma_\delta^2(0) + \sum_{i=1}^{n} Var(\hat{\gamma}_i)$. Therefore, the expected sum of squares attributed to a regional skew model with *k* parameters equals $n[\sigma_\delta^2(0) - \sigma_\delta^2(k)]$, because the sum of

Table 1–3. Pseudo analysis of variance (ANOVA) for the Missouri regional skew study for the constant model.

[k, number of estimated regression parameters not including the constant; n, number of observations (gage sites) used in regression; $\sigma_\delta^2(0)$, model error variance of a constant model; $\sigma_\delta^2(k)$, model error variance of a model with k regression parameters and a constant; $Var(\hat{y}_i)$, variance of the estimated sample skew at site i; *EVR*, error variance ratio; *MBV**, misrepresentation of the beta variance; pseudo-R_δ^2, fraction of variability in the true skews explained by each model (Gruber and others, 2007); %, percent]

Source	Degrees-of-freedom		Equations	Sum of squares
Model	k	0	$n[\sigma_\delta^2(0) - \sigma_\delta^2(k)]$	0
Model error	n-k-1	107	$n[\sigma_\delta^2(0)]$	14
Sampling error	n	108	$\sum_{i=1}^{n} Var(\hat{y}_i)$	16
Total	2n-1	215	$n[\sigma_\delta^2(0)] + \sum_{i=1}^{n} Var(\hat{y}_i)$	31
EVR				1.1
*MBV**				3.8
Pseudo-R_δ^2	$R_\delta^2 = 1 - \frac{\sigma_\delta^2(k)}{\sigma_\delta^2(0)}$			0%

the model error variance $n\sigma_\delta^2(k)$ and the variance explained by the model must sum to $n\sigma_\delta^2(0)$. Table 1–3 considers a model with $k = 0$ (a constant model). This division of the variation in the observations is referred to as a pseudo ANOVA because the contributions of the three sources of error are estimated or constructed, rather than being determined from the computed residual errors and the observed model predictions, while also ignoring the impact of correlation among the sampling errors.

Table 1–3 contains the pseudo ANOVA results for the Constant model. The Constant model does not have any explanatory variables; thus the variation attributed to the model is 0. As shown in table 1–3, the Constant model has a sampling error variance larger than its model error variance.

The Error Variance Ratio (*EVR*) is a modeling diagnostic used to evaluate if a simple OLS regression is sufficient, or a more sophisticated WLS or GLS analysis is appropriate. *EVR* is the ratio of the average sampling error variance to the model error variance. Generally, an *EVR* greater than 0.20 indicates that the sampling variance is not negligible when compared to the model error variance, suggesting the need for a WLS or GLS regression analysis. The *EVR* is calculated as

$$EVR = \frac{SS\,(\text{sampling error})}{SS\,(\text{model error})} = \frac{\sum_{i=1}^{n} Var\,(\hat{y}_i)}{n\sigma_\delta^2(k)} \qquad \text{(A15)}$$

where

SS is the sum of squares.

For the Missouri study-area data, *EVR* had a value of 1.1 for the Constant model. The sampling variability in the sample skewness estimators was larger than the error in the regional model. Thus, an OLS model that neglects sampling error in the station skewness estimators may not provide a statistically reliable analysis of the data. Given the variation of record lengths from site-to-site, it is important to use a WLS or GLS analysis to evaluate the final precision of the model, rather than a simpler OLS analysis.

The misrepresentation of the Beta Variance (*MBV**) statistic is used to determine whether a WLS regression is sufficient, or if a GLS regression is appropriate to determine the precision of the estimated regression parameters (Veilleux, 2011; Griffis, 2006). The *MBV** describes the error produced by a WLS regression analysis in its evaluation of the precision of b_0^{WLS}, which is the estimator of the constant β_0^{WLS}, because the covariance among the estimated station skews generally has its greatest impact on the precision of the constant term (Stedinger and Tasker, 1985). If the *MBV** is substantially greater than 1, then a GLS error analysis should be employed. The *MBV** is calculated as,

$$MBV^* = \frac{Var[b_0^{WLS} \mid GLS\ analysis]}{Var[b_0^{WLS} \mid WLS\ analysis]} = \frac{w^T \Lambda w}{\sum_{i=1}^{n} w_i} \qquad \text{(A16)}$$

where

w_i is $\frac{1}{\sqrt{\Lambda_{ii}}}$.

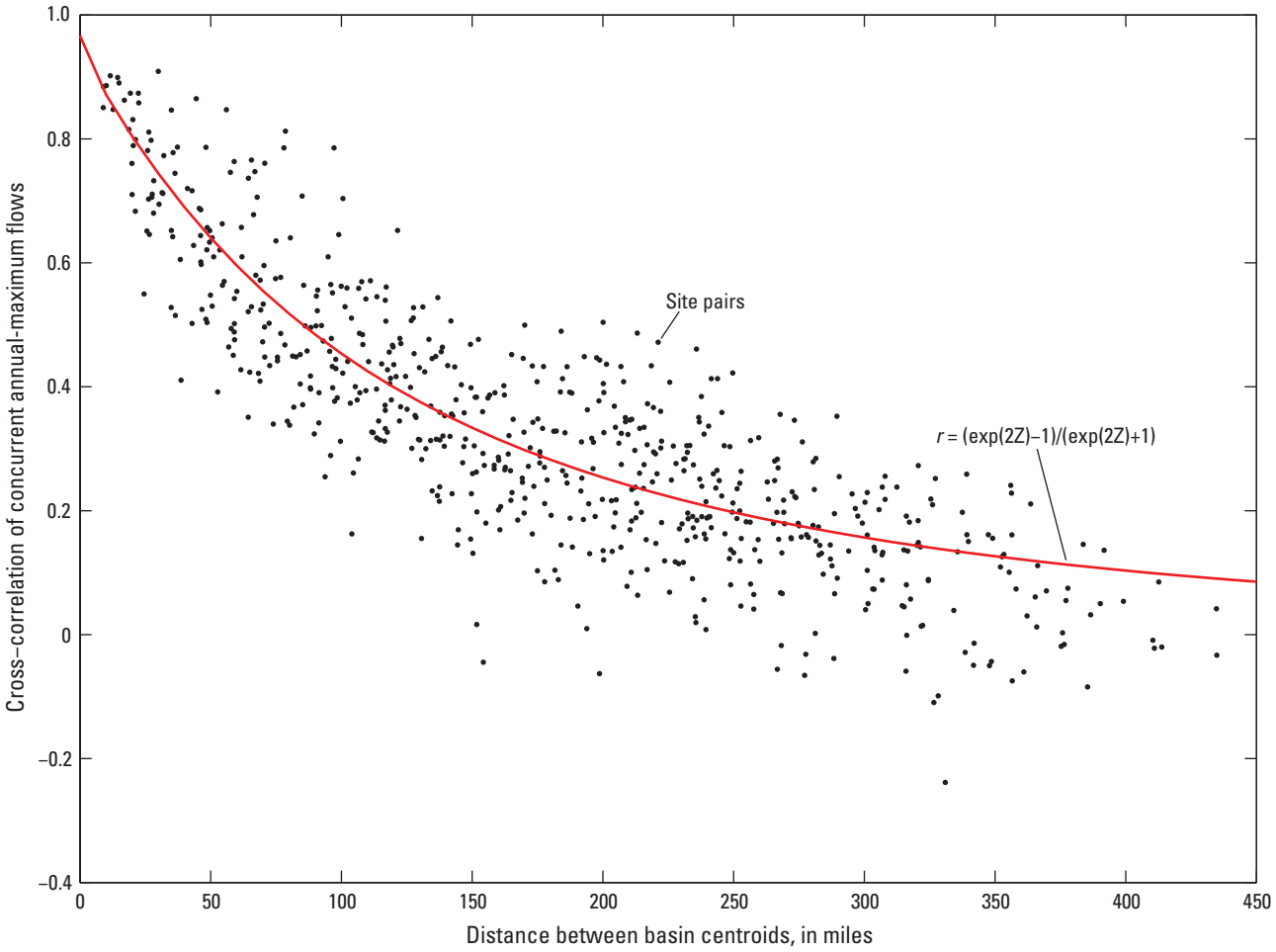


Figure 1–2. Relation between untransformed cross-correlation of logs of annual peak discharge and distance, (D), between basin centroids based for 651 station-pairs with concurrent record lengths greater than or equal to 60 years from 42 streamgages in Missouri and neighboring States.

For the Missouri regional skew study, the *MBV** is equal to 3.8 for the Constant model. This is a large value indicating the cross-correlation among the skewness estimators has had an impact on the precision with which the regional average skew coefficient can be estimated. If a WLS precision analysis were used for the estimated constant parameter in the Constant model, the variance would be underestimated by a factor of 3.8. Thus, a WLS analysis would misrepresent the variance of the constant in the Constant model. Moreover, a WLS model would have resulted in underestimation of the variance of prediction, given that the sampling error in the constant term in both models was sufficiently large enough to make an appreciable contribution to the average variance of prediction.

Leverage and Influence

Leverage and influence diagnostics statistics can be used to identify rogue observations and to effectively address lack-of-fit when estimating skew coefficients. Leverage identifies those streamgages in the analysis where the observed values have a large impact on the fitted (or predicted) values (Hoaglin and Welsch, 1978). Generally, leverage considers whether an observation, or explanatory variable, is unusual, and thus likely to have a large effect on the estimated regression coefficients and predictions. Unlike leverage, which highlights observations that have the ability or potential to affect the fit of the regression, influence attempts to describe observations that do have an unusual impact on the regression analysis (Belsley and others, 1980; Cook and Weisberg, 1982; Tasker and Stedinger, 1989). An influential observation is one with an unusually large residual that has a disproportionate effect on the fitted regression relations. Influential observations often have high leverage. For a detailed description of the equations used to determine leverage and influence for a B-WLS/B-GLS analysis see Veilleux and others (2011) and Veilleux (2011).

The leverage and influence values for the B-WLS/B-GLS constant regional skew model for the Missouri study area are described here. Only two sites in the B-WLS/B-GLS constant regional skew model for the Missouri study area have high influence, and thus have an unusual impact on the fitted regression relation. No sites in the regression

had high leverage; the differences in leverage values for the constant model reflect the variation in record lengths among sites. Streamgage 06813000 (regional skew index number 43, table 1–1) has the highest influence value due to its large residual, the largest magnitude residual in the study (that is the fourth smallest unbiased station skew = -1.45). Streamgage 07043500 (regional skew index number 227) has the highest influence value due to its large residual, the second largest magnitude residual in the study (that is, the third smallest unbiased station skew = -1.54).

References Cited

Belsley, D.A., Kuh, E., and Welsch, R.E.,1980, Regression diagnostics—identifying influential data and sources of collinearity, chap. 2: New York, John Wiley and Sons, Inc., p. 6–84.

Cohn, T.A., 2011, PeakfqSA [Flood-frequency analysis with expected moments algorithm], version 0.960 [software]: PeakfqSA/EMA-Peak, accessed July 30, 2014, at *http://www.timcohn.com/TAC_Software/PeakfqSA/*.

Cohn, T.A., Lane, W.L., and Stedinger, J.R., 2001, Confidence intervals for Expected Moments Algorithm flood quantile estimates: Water Resources Research, v. 37, no. 6, p. 1695–1706.

Cohn, T.A., Lane, W.L., and Baier, W.G., 1997, An algorithm, for computing moments-based flood quantile estimates when historical flood information is available: Water Resources, v. 33, no. 9, p. 2089–2096.

Cook, R.D., and Weisberg, S., 1982, Residuals and influence in regression: New York, Chapman and Hall, 230 p.

Devore, J.L., 2004, Probability and Statistics for Engineering and the Sciences: Belmont, Calif., Brooks/Cole, , 795 p.

Eash, D.A., Barnes, K.K., and Veilleux, A.G., 2013, Methods for estimating annual exceedance-probability discharges for streams in Iowa, based on data through water year 2010: U.S. Geological Survey Scientific Investigations Report 2013–5086, 63 p. with appendix.

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 3, South Carolina: U.S. Geological Survey Scientific Investigations Report 2009–5043, 238 p.

Gotvald, A.J., Feaster, T.D., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey Scientific Investigations Report 2009–5156, 120 p.

Griffis, V.W., and Stedinger, J.R., 2009, Log-Pearson type 3 distribution and its application in flood frequency analysis, III: sample skew and weighted skew estimators: Journal of Hydrology, v. 14, no. 2, p. 121–130.

Griffis, V.W., 2006. Flood Frequency Analysis: Bulletin 17, Regional Information, and Climate Change: Cornell, Cornell University, Ph.D. Dissertation, 246 p.

Griffis, V.W., Stedinger, J.R., and Cohn, T.A., 2004, Log Pearson type 3 Quantile estimators with regional skew information and low outlier adjustments: Water Resources Research, v. 40, W07503, doi:1029/2003WR002697, 17 p.

Gruber, Andrea M., and Stedinger, J.R., 2008, Models of LP3 regional skew, data selection and Bayesian GLS regression, Paper 596, *in* Babcock, R.W. and Watson, R., eds., World Environmental and Water Resources Congress – Ahupua'a, Honolulu, Hawai'i, May 12–16, 2008, 10 p.

Gruber, A.M., Reis, D.S., Jr., and Stedinger, J.R., 2007, Models of regional skew based on Bayesian GLS regression, Paper 40927-3285, *in* Kabbes, K.C., ed., Restoring our natural habitat: Proceedings of the World Environmental and Water Resources Congress, May 15–18, 2007,Tampa, Fla.: American Society of Civil Engineers, Reston, VA., 10 p.

Hoaglin, D.C., and Welsch, R.E., 1978, The Hat Matrix in regression and ANOVA: The American Statistician, v. 32, no. 1, p. 17–22.

U.S. Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood flow frequency—Bulletin 17B of The Hydrology Subcommittee: U.S Geological Survey, Office of Water Data Coordination, 183 p.

Lamontagne, J.R., Stedinger, J.R., Berenbrock, Charles, Veilleux, A.G., Ferris, J.C., and Knifong, D.L., 2012, Development of regional skews for selected flood durations for the Central Valley Region, California, based on data through water year 2008: U.S. Geological Survey Scientific Investigations Report 2012–5130, 60 p.

Martins, E.S., and Stedinger, J.R., 2002, Cross-correlation among estimators of shape: Water Resources Research, v. 38, no. 11, doi: 10.1029/2002WR001589, p. 34–1 to 34–7.

Parrett, Charles, Veilleux, A.G., Stedinger, J.R., Barth, N.A., Knifong, D.L., and Ferris, J.C., 2011, Regional skew for California, and flood frequency for selected sites in the Sacramento–San Joaquin River Basin, based on data through water year 2006: U.S. Geological Survey Scientific Investigations Report 2010–5260, 94 p.

Reis, D.S., Jr., Stedinger, J.R., and Martins, E.S., 2005, Bayesian generalized least squares regression with application to the log Pearson type III regional skew estimation: Water Resources Research, v. 41, W10419, doi:10.1029/2004WR003445, 14 p.

Stedinger, J.R., and Cohn, T.A., 1986, Flood frequency analysis with historical and paleoflood information: Water Resources Research, v. 22, no. 5, p. 785–793.

Stedinger, J.R., and Tasker, G.D., 1985, Regional hydrologic analysis, 1. Ordinary, weighted and generalized least squares compared: Water Resources Research, v. 21, no. 9, p. 1421–1432.

Tasker, G.D., and Stedinger, J.R., 1989, An operational GLS model for hydrologic regression: Journal of Hydrology, v. 111, no. (1–4), p. 361–375.

Tasker, G.D., and Stedinger, J.R., 1986, Regional skew with weighted LS regression: Journal of Water Resources Planning and Management, ASCE, v.112, no. 2, p. 225–237.

Veilleux, A.G., Stedinger, J.R., and Eash, D.A., 2012, Bayesian WLS/GLS regression for regional skewness analysis for regions with large crest stage gage networks, paper 2253, World Environmental and Water Resources Congress 2012: Crossing Boundaries, American Society of Civil Engineers, Albuquerque, N. M., May 20–24, p. 2253–2263.

Veilleux, A.G., 2011, Bayesian GLS regression, leverage and influence for regionalization of hydrologic statistics: Cornell, Cornell University, Ph.D. dissertation, 184 p.

Veilleux, A.G., Stedinger, J.R., and Lamontagne, J.R., 2011, Bayesian WLS/GLS regression for regional skewness analysis for regions with large cross-correlations among flood flows, paper 1303, *in* World Environmental and Water Resources Congress 2011—Bearing Knowledge for Sustainability, Palm Springs, Calif., May 22–26, 2011, American Society of Civil Engineers, p. 3103–3112.

Veilleux, A.G. 2009, Bayesian GLS regression for regionalization of hydrologic statistics,Floods and Bulletin 17 Skew: Cornell, Cornell Univeristy, M.S. Thesis, 155 p.

Weaver, J.C., Feaster, T.D., and Gotvald, A.J., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 2, North Carolina: U.S. Geological Survey Scientific Investigations Report, 2009–5158, 111 p. (Also available at: *http://pubs.usgs.gov/sir/2009/5158/*).

Publishing support provided by:

Rolla Publishing Service Center

For more information concerning this publication, contact:
Director, Missouri Water Science Center
U.S. Geological Survey
1400 Independence Road, MS-100
Rolla, MO 65401
(573) 308-3667

Or visit the Missouri Water Science Center Web site at:
http://mo.water.usgs.gov/

ISBN 978-1-4113-3845-6

2328-031X (print)
2328-0328 (online)
http://dx.doi.org/10.3133/sir20145165