USGS
science for a changing world

Methods for Estimating Regional Skewness of Annual Peak Flows in Parts of the Great Lakes and Ohio River Basins, Based on Data Through Water Year 2013

Scientific Investigations Report 2019–5105

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

Methods for Estimating Regional Skewness of Annual Peak Flows in Parts of the Great Lakes and Ohio River Basins, Based on Data Through Water Year 2013

By Andrea G. Veilleux and Daniel M. Wagner

Scientific Investigations Report 2019–5105

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

U.S. Department of the Interior
DAVID BERNHARDT, Secretary

U.S. Geological Survey
James F. Reilly II, Director

U.S. Geological Survey, Reston, Virginia: 2019

Suggested citation:
Veilleux, A.G., and Wagner, D.M., 2019, Methods for estimating regional skewness of annual peak flows in parts of the Great Lakes and Ohio River Basins, based on data through water year 2013: U.S. Geological Survey Scientific Investigations Report 2019–5105, 26 p., https://doi.org/10.3133/sir20195105.

Associated data for this publication:
Wagner, D.M., and Veilleux, A.G., 2019, Annual peak-flow data, PeakFQ specification files, and PeakFQ output files for 368 selected streamflow gaging stations operated by the U.S. Geological Survey in the Great Lakes and Ohio River Basins that were used to estimate regional skewness of annual peak flows: U.S. Geological Survey data release, https://doi.org/10.5066/P9N7UAFJ.

ISSN 2328-0328 (online)

Contents

Figures

[Figures 1, 2, 3, and 5 are provided at https://doi.org/10.3133/sir20195105]

Tables

[Table 1 is provided at https://doi.org/10.3133/sir20195105]

Conversion Factors

International System of Units to U.S. customary units

Multiply	By	To obtain
	Length	
centimeter (cm)	0.3937	inch (in.)
kilometer (km)	0.6214	mile (mi)
	Volume	
cubic meter (m^3)	35.31	cubic foot (ft^3)
	Flow rate	
cubic meter per second (m^3/s)	35.31	cubic foot per second (ft^3/s)

Datum

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

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

Abbreviations

AEP	annual exceedance probability
ASEV	average sampling error variance
AVP_{new}	average variance of prediction
B17B	Bulletin 17B (see Interagency Advisory Committee on Water Data, 1982)
B17C	Bulletin 17C (see England and others, 2018)
B–GLS	Bayesian generalized least squares
B–WLS	Bayesian weighted least squares
DAR	drainage area ratio
EMA	Expected Moments Algorithm
EVR	error variance ratio
GAGES II	Geospatial Attributes of Gages for Evaluating Streamflow II
GLS	generalized least squares
LP–III	log-Pearson Type III distribution
*MBV**	Misrepresentation of the Beta Variance
MGBT	Multiple Grubbs-Beck test
MSE	mean square error
NWIS	National Water Information System of the USGS
OLS	ordinary least squares

PILF	potentially influential low flood
P_{RL}	pseudo record length
pseudo ANOVA	pseudo analysis of variance
SD	standardized distance
USGS	U.S. Geological Survey
WLS	weighted least squares

Methods for Estimating Regional Skewness of Annual Peak Flows in Parts of the Great Lakes and Ohio River Basins, Based on Data Through Water Year 2013

By Andrea G. Veilleux and Daniel M. Wagner

Abstract

Bulletin 17C (B17C) recommends fitting the log-Pearson Type III (LP−III) distribution to a series of annual peak flows at a streamgage by using the method of moments. The third moment, the skewness coefficient (or skew), is important because the magnitudes of annual exceedance probability (AEP) flows estimated by using the LP−III distribution are affected by the skew; interest is focused on the right-hand tail of the distribution, which represents the larger annual peak flows that correspond to small AEPs. For streamgages having modest record lengths, the skew is sensitive to extreme events like large floods, which cause a sample to be highly asymmetrical or "skewed." For this reason, B17C recommends using a weighted-average skew computed from the station skew for a given streamgage and a regional skew. This report generates an estimate of regional skew for a study area encompassing most of the Great Lakes Basin (hydrologic unit 04) and part of the Ohio River Basin (hydrologic unit 05). A total of 551 candidate streamgages that were unaffected by extensive regulation, diversion, urbanization, or channelization were considered for use in the skew analysis; after screening for redundancy and pseudo record length (P_{RL}) greater than 36 years, 368 streamgages were selected for use in the study. Flood frequencies for candidate streamgages were analyzed by employing the Expected Moments Algorithm (EMA), which extends the method of moments so that it can accommodate interval, censored, and historic/paleo flow data, as well as the Multiple Grubbs-Beck test to identify potentially influential low floods in the data series. Bayesian weighted least squares/Bayesian generalized least squares regression was used to develop a regional skew model for the study area that would incorporate possible variables (basin characteristics) to explain the variation in skew in the study area. Twelve basin characteristics were considered as possible explanatory variables; however, none produced a pseudo coefficient of determination (pseudo R^2_δ) greater than 5 percent; as a result, these characteristics did not help to explain the variation in skew in the study area. Therefore, a constant model having a regional skew coefficient of 0.086 and an average variance of prediction (AVP_{new}) (which corresponds to the mean square error [MSE]) of 0.13 at a new streamgage was selected. The AVP_{new} corresponds to an effective record length of 54 years, a marked improvement over the Bulletin 17B national skew map, whose reported MSE of 0.302 indicated a corresponding effective record length of only 17 years.

Introduction

Flood-frequency analysis of annual peak flows at streamflow-gaging stations (hereafter referred to as "streamgages") provides engineers, hydrologists, and many others estimates of the magnitudes and frequencies of floods for planning, design, and management of infrastructure along rivers and streams. The Subcommittee on Hydrology of the Federal Advisory Committee on Water Information recently published Bulletin 17C (herein referred to as "B17C," England and others, 2018), which comprises updated guidelines for flood-frequency analysis. The bulletin recommends use of the log-Pearson Type III (LP−III) distribution to fit a time series of annual peak flows measured by a streamgage to obtain estimates of flows corresponding to various annual exceedance probabilities (AEP). In the case of flood-frequency analysis, the LP−III distribution is described by three moments: the mean, the standard deviation, and the skewness coefficient of the logarithms of the flows. The third moment, the skewness coefficient (hereafter referred to as the "skew"), is a measure of the asymmetry of the distribution as shown by the thicknesses of the tails of the distribution. In flood-frequency analysis, the skew is important because the magnitudes of AEP flows estimated by using the LP−III distribution are affected by the skews of the annual peak flows at specific streamgages (hereafter referred to as "station skew"); interest is focused on the right-hand tail of the distribution, which represents annual peak flows corresponding to small AEPs of the larger flood flows.

For streamgages having modest record lengths, approximately in the range of 25 to 100 years, the skewness coefficient is sensitive to unusually large or small annual peak flows because they cause a sample of such flows to be asymmetrical

or skewed (Griffis and Stedinger, 2007). Thus, B17C guidelines recommend using a weighted-average skew that is computed from the skew of the station's annual peak flows and the regional skew. Using the weighted-average skew reduces the sensitivity of the station skew to extreme events, particularly for streamgages with short record lengths of less than approximately 25 years.

The B17C guidelines recommend using the Bayesian weighted least squares/Bayesian generalized least squares (B–WLS/B–GLS) method to estimate regional skew (England and others, 2018, p. 30). Using this procedure, the regional skew is estimated based on the station skew of the logarithms of annual peak-flow data. The B–WLS/B–GLS procedure first uses an ordinary least squares (OLS) regression analysis to generate an initial regional-skew model that is used to compute the variance of the station skew for each streamgage. Next, B–WLS is used to generate estimators of the regional skew model parameters. Finally, B–GLS is used to estimate the precision of the B–WLS parameter values, to estimate the model error variance and its precision, and to compute some diagnostic statistics. The B–WLS/B–GLS method can account for the complexities introduced by the Expected Moments Algorithm (hereinafter referred to as "EMA," Cohn and others, 1997), the B17C recommended generalization of the method of moments approach for flood-frequency analysis of the annual peak flows from streamgages, and the cross correlation between annual peak flows at pairs of streamgages (Veilleux, 2011; Veilleux and others, 2011).

To date, the B–WLS/B–GLS method has been used to generate estimates of regional skew for several regions around the Nation (Parrett and others, 2011; Eash and others, 2013; Olson, 2014; Paretti and others, 2014; Southard and Veilleux, 2014; Curran and others, 2016; Mastin and others, 2016; Wagner and others, 2016). In this study, the B–WLS/B–GLS procedure was used to estimate skew for a region encompassing parts of the Great Lakes and Ohio River Basins (hydrologic units 04 and 05, respectively; see fig. 1*A* at https://doi.org/10.3133/sir20195105) to improve estimates of regional skew and flows corresponding to various AEPs across the region.

Purpose and Scope

The purpose of this report is to present the results of a B–WLS/B–GLS analysis of regional skew for parts of the Great Lakes and Ohio River Basins (fig. 1*A*). The scope of the project includes 368 streamgages, 187 in the Great Lakes Basin (hydrologic unit 04) and 181 in the Ohio River Basin (hydrologic unit 05) located in the States of Illinois, Indiana, Kentucky, Michigan, Minnesota, New York, Ohio, Pennsylvania, Vermont, West Virginia, and Wisconsin (see fig. 1*B* at https://doi.org/10.3133/sir20195105). Flood-frequency analyses for the 368 streamgages were based on annual peak-flow data through water year 2013 (a water year is described as the period October 1–September 30, named for the year in which it ends) and were performed using the U.S. Geological Survey (USGS) peak-flow analysis software (PeakFQ version 7.2, Veilleux and others, 2014). The results were used to analyze the regional skew. Streamgages in 4-digit hydrologic units 0511 and 0513 in Kentucky and Tennessee and in Canada were not considered because USGS Water Science Center offices only in the States of Illinois, Michigan, Minnesota, New York, Ohio, Pennsylvania, and Wisconsin actively participated in the study.

A summary of output from the flood-frequency analyses for each streamgage used in the regional skew analysis and a description of the basin characteristics considered as potential explanatory variables in the study are provided in the tables in this report. Peak-flow input files (.txt), PeakFQ setup files (.psf), and PeakFQ output (.PRT) files for the 368 streamgages used in the analysis and corresponding metadata are provided in a data release associated with this report (Wagner and Veilleux, 2019).

Description of Study Area

The study area encompasses most of the Great Lakes Basin (hydrologic unit 04) and part of the Ohio River Basin (hydrologic unit 05) and includes the States of Indiana, Michigan, and Ohio, and parts of the States of Illinois, Minnesota, New York, Pennsylvania, and Wisconsin (fig. 1*A*). The study area spans approximately 1,600 kilometers from east to west from northeastern Minnesota to the New York-Vermont border and approximately 1,200 kilometers from north to south from northeastern Minnesota near Lake Superior to the Ohio River on the southern boundary of Illinois.

The study area contains parts of the Laurentian Upland, Appalachian Highlands, and Interior Plains physiographic divisions (Fenneman, 1938). The northwestern part of the study area is in the Laurentian Upland, characterized by gently rolling hills and small mountain remnants of the Canadian Shield, which is underlain by granitic rocks of Precambrian age (U.S. Environmental Protection Agency and Government of Canada, 1995). The southern part of the Great Lakes Basin and northern part of the Ohio River Basin in the study area are in a part of the Interior Plains that is characterized by relatively flat glacial-till plains and glacial deposits. The eastern part of the Ohio River Basin and far northeastern part of the Great Lakes Basin, which are respectively in Pennsylvania and New York, are characterized by the mountainous terrain of the Appalachian Highlands.

The study area exhibits two climate types—humid subtropical in southern Illinois, Indiana, Ohio, and Pennsylvania; and humid continental in the rest of the study area. Mean annual precipitation in the study area ranges from 40 to 50 inches (102 to 127 centimeters) in the south near the Ohio River to 25 to 30 inches (64 to 76 centimeters) in northeastern Minnesota and northern Michigan (Arguez and others, 2012).

Based on the 2011 National Land Cover Database, the study area is approximately 36 percent forested, 38 percent agricultural (crops and pasture), 11 percent developed, and 10 percent wetlands, with the remaining 5 percent

including open water, barren land, shrub/scrub, and grassland/herbaceous categories (Homer and others, 2015).

Methods

Streamgage Selection

A suite of 551 candidate streamgages were considered for use in the regional skew analysis (see table 1 at https://doi.org/10.3133/sir20195105). Annual peak flows for these streamgages were obtained from the USGS National Water Information System (NWIS; U.S. Geological Survey, 2015). Only streamgage records unaffected by extensive regulation, diversion, urbanization, or channelization (based on coding of annual peaks in the peak-flow files) and having 25 or more gaged peaks were considered for use in the regional skew analysis. Using these criteria, USGS employees who had local knowledge and experience in each State that participated in the study selected candidate streamgages. Finally, streamgages that were deemed redundant were then screened and removed from the larger dataset (see "Redundancy Screening" section for more information).

Redundancy Screening

Two streamgages may be redundant if their drainage basins are nested and similar in size; the drainage basins are considered nested if one entire drainage area is inside the other. If streamgages are redundant, a statistical analysis incorporating data from both streamgages incorrectly represents the information content in the regional dataset (Gruber and Stedinger, 2008). Instead of providing two spatially independent observations that depict how the characteristics of each basin are related to skew, the basins will be assumed to exhibit similar hydrologic responses to a given storm and thus represent only one spatial observation. To determine whether two streamgages are redundant and thus represent the same watershed for the purposes of developing a regional hydrologic model, two types of information are considered: (1) the standardized distance (*SD*) between the centroids of the basins and (2) the ratio of the drainage areas of the basins.

The *SD* between two basin centroids is used to determine the likelihood that the basins are redundant. *SD* is defined as

$$SD_{ij} = \frac{D_{ij}}{\sqrt{0.5\left(DRNAREA_i + DRNAREA_j\right)}}, \qquad (1)$$

where

D_{ij} is the distance between centroids of basin *i* and basin *j*, in miles;
$DRNAREA_i$ is the drainage area at streamgage *i*, in square miles; and
$DRNAREA_j$ is the drainage area at streamgage *j*, in square miles.

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

$$DAR = Max\left[\frac{DRNAREA_i}{DRNAREA_j}, \frac{DRNAREA_j}{DRNAREA_i}\right], \qquad (2)$$

where

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

Previous studies suggest that streamgage pairs having *SD* less than or equal to 0.50 and *DAR* less than or equal to 5.0 are likely to be redundant for purposes of determining regional skew (Veilleux, 2009). If *DAR* is large enough, even nested streamgages will reflect different hydrologic responses because storms of different sizes and duration typically affect sites differently.

All possible combinations of streamgage pairs from the 551 streamgages were considered in the redundancy analysis. All streamgage pairs with $SD \leq 0.5$ and $DAR \leq 5.0$ were identified as possibly redundant. The drainage area of each streamgage was then investigated to determine if one of the two drainage areas was nested inside the other; if this was true, the preference was generally for the streamgage having the smaller drainage area and the longer record length. The procedure identified 123 possibly redundant streamgage pairs; of these, 77 were found to be redundant and removed from the analysis, after which 474 were left for use in the regional skew study (table 1).

Basin Characteristics

Basin characteristics for the streamgages used in the skew analysis were either obtained from the USGS Geospatial Attributes of Gages for Evaluating Streamflow (GAGES II) database or generated. The GAGES II database consists of a subset of USGS streamgages having at least 20 years of discharge record since 1950 or that were active as of water year 2009 and whose watersheds lie within the United States (Falcone, 2011). For streamgages that were used in the skew analysis but not in the GAGES II database, the suite of basin characteristics was generated by using the ArcHydro package in Esri ArcGIS software version 10.3.1 (Esri, 2009; Eash and others, 2013; Wagner and others, 2016). This procedure ensured that a consistent suite of basin characteristics was available for all 368 streamgages used in the skew analysis.

Basin characteristics selected to potentially explain the variation in skew in the study area included morphometric (drainage area, latitude and longitude of basin centroid, mean basin slope, mean basin elevation, and basin compactness ratio), climatological (basin-average mean annual precipitation), and pedologic or geologic (areal percentages of open

water and forest, and average soil permeability) characteristics (table 2). In addition to these 10 basin characteristics, the basin-average 24-hour, 100-year precipitation intensity was determined for each streamgage (National Oceanic and Atmospheric Administration, 2014), as was the physiographic division within which the basin centroid was located (either the Laurentian Upland, Interior Plains, or Appalachian Highlands; Fenneman, 1938).

Annual Exceedance Probability Analyses

To estimate regional skew for parts of the Great Lakes and Ohio River Basins, a flood-frequency analysis must first be conducted for each streamgage to determine the station skew and its associated mean square error (MSE). The B17C guidelines recommend fitting the log-Pearson Type III (LP−III) distribution to a series of annual peak flows at a streamgage by using the method of moments (England and others, 2018). In doing so, it is recommended that the EMA is employed to extend the method of moments to accommodate interval, censored, and historical or paleo flood data, as well as the use of the Multiple Grubbs-Beck test (MGBT) to identify potentially influential low floods (PILFs) in the data series. In this study, the USGS software PeakFQ version 7.2 was used to analyze the flood frequencies (Veilleux and others, 2014; https://water.usgs.gov/software/PeakFQ/).

Hydrologists in the USGS Water Science Centers in Illinois, Michigan, Minnesota, New York, Ohio, Pennsylvania, and Wisconsin used EMA with PeakFQ version 7.2 for candidate streamgages in their respective States and used EMA with PeakFQ version 7.2 for candidate streamgages in Indiana, Kentucky, Vermont, and West Virginia. Flood frequencies were analyzed by using the station-skew option in PeakFQ software and, with few exceptions (such as a fixed threshold for PILFs that yielded a superior fit of the flood-frequency model to the dataset), the MGBT for PILFs. Historical peaks were included in the analysis; annual peak flows coded as urban or regulated were not. Hydrologists in the participating States assigned perception thresholds to the entire historical periods (from the start year to the end year of the record, including years with missing peaks and periods of crest-stage gage operation) and flow intervals to uncertain annual peak flows as appropriate.

Bayesian Weighted Least Squares/Bayesian Generalized Least Squares Analysis

Prior to analyzing regional skew by the B−WLS/B−GLS method, three preliminary steps were completed: (1) calculation of the pseudo record length for each streamgage, given the number of censored observations and concurrent record lengths; (2) correction for structural bias in the estimate of station skew and its MSE; and (3) development of a cross-correlation model of concurrent annual peak flows between streamgages.

Calculating Pseudo Record Length

The pseudo record length of the annual peak-flow series at each streamgage is used in the regional skew study in several steps, including unbiasing the station skew and its mean square error, determining the concurrent record length between two streamgages, and computing the cross correlation of the station skews. Because the dataset includes censored data and historical information, the effective record length used to compute the precision of the skewness estimators is no longer simply the number of annual peak flows at a streamgage. Instead, a more complex calculation based on the availability of historical information and censored values is used. Whereas historical information and records of censored peaks provide valuable information, they often provide less information than records of an equal number of years of gaged peaks (Stedinger and Cohn, 1986). The calculations described in the following paragraphs yield a pseudo record length (P_{RL}) associated with skew, which appropriately accounts for all types of peak-flow data available from a streamgage. If no interval, censored, historical data are present in the annual peak-flow record of a streamgage, P_{RL} is equal to the gaged record length.

The P_{RL} is defined as the number of years of gaged record that would be required to yield the same mean square error of the skew ($MSE\left(\tilde{G}\right)$) as would the combination of the historical and gaged records actually available at a streamgage; thus, the P_{RL} of the skew is a ratio of the MSE of the station skew when only the gaged record is analyzed ($MSE\left(\tilde{G}_S\right)$) to the MSE of the station skew when the entire record, including historical and censored data, is analyzed ($MSE\left(\tilde{G}_C\right)$):

$$P_{RL} = \frac{P_S \times MSE\left(\tilde{G}_S\right)}{MSE\left(\tilde{G}_C\right)}, \tag{3}$$

where

P_{RL} is the pseudo length of the entire record at the streamgage, in years;

P_S is the number of years with gaged peaks in the record;

$MSE\left(\tilde{G}_S\right)$ is the estimated MSE of the skew when only the gaged record is analyzed; and

$MSE\left(\tilde{G}_C\right)$ is the estimated MSE of the skew when the entire record, including historical and censored data, is analyzed.

Because the P_{RL} is an estimate, the following conditions must also be met to ensure a valid approximation. The P_{RL} must be nonnegative. If P_{RL} is greater than P_H (the length of the historical period), then P_{RL} should be set to equal P_H. Also, if P_{RL} is less than P_S, then P_{RL} is set to P_S. This ensures that the P_{RL} will not be larger than the complete P_H or less than the P_S.

As stated in B17C, the station skew is sensitive to extreme events; therefore, accurate estimates of skew require longer periods of record, typically 50 years or greater; however, 50 years of record are not available for

Table 2. Basin characteristics considered for use as explanatory variables in the regional skew analysis.

[GIS, geographic information system; DEM, digital elevation model; NAD83, North American Datum of 1983; NHD, National Hydrography Dataset; NLCD, National Land Cover Database; NOAA, National Oceanic and Atmospheric Administration; PRISM, Parameter Regression on Independent Slopes Model]

Basin characteristic	Units	Source
Drainage area of streamgage basin, delineated by using GIS	square kilometers	Derived from 30-meter NHDPlus data, http://www.horizon-systems.com/nhdplus/.
Latitude of basin centroid	decimal degrees NAD83	Determined from zonal statistics of grids derived from basin polygons in Esri ArcGIS, version 10.3.1.
Longitude of basin centroid	decimal degrees, NAD83	Determined from zonal statistics of grids derived from basin polygons in Esri ArcGIS, version 10.3.1.
Mean basin elevation	meters	Determined from 10-meter DEM, National Elevation Dataset, https://www.usgs.gov/core-science-systems/national-geospatial-program/national-map.
Mean basin slope	percent	Derived from 100-m resolution National Elevation Dataset, https://www.usgs.gov/core-science-systems/national-geospatial-program/national-map, or obtained from USGS GAGES II database (Falcone, 2011).
Basin compactness ratio (area/perimeter^^2×100); higher number indicates more compact shape	unitless	Calculated in Esri ArcGIS, version 10.3.1, by using drainage area and perimeter of GIS-delineated basin polygons.
Basin-averaged mean annual precipitation for the 30-year period 1971 to 2000	centimeters	800-meter PRISM data, Oregon State University, http://www.prism.oregonstate.edu/.
Basin-averaged soil permeability	inches per hour	Wolock, 1997 (http://water.usgs.gov/GIS/metadata/usgswrd/XML/muid.xml) and U.S. Department of Agriculture, 2008 (http://www.soils.usda.gov/survey/geography/statsgo/).
Percentage of streamgage basin in forested land use categories	percentage of streamgage basin surface area	2006 NLCD, sum of classes 41, 42, and 43, https://www.mrlc.gov/data?f%5B0%5D=year%3A2006.
Percentage of streamgage basin in open water	percentage of streamgage basin surface area	2006 NLCD, class 11, https://www.mrlc.gov/data?f%5B0%5D=year%3A2006.
Basin-averaged, 24-hour precipitation intensity (10-year recurrence interval)	inches	NOAA Atlas 14 precipitation frequency estimates, https://hdsc.nws.noaa.gov/hdsc/pfds/pfds_gis.html.
Basin-averaged, 24-hour precipitation intensity (100-year recurrence interval)	inches	NOAA Atlas 14 precipitation frequency estimates, https://hdsc.nws.noaa.gov/hdsc/pfds/pfds_gis.html.

most streamgages, and therefore a minimum of 35 years has been used in recent studies (Eash and others, 2013; Paretti and others, 2014; Southard and Veilleux, 2014; Wagner and others, 2016). Thus, after adequate geographic and hydrologic coverage was ensured, streamgages in the dataset having a P_{RL} less than 36 years were removed from the study. Of the 474 sites remaining after the 77 redundant sites were removed, 106 were removed for having a P_{RL} less than 36 years, leaving 368 streamgages from which a regional skew model was developed (table 1; see fig. 2 at https://doi.org/10.3133/sir20195105).

Removing the Bias of the At-Site Estimators

The station skew estimates were debiased by using the correction factor developed by Tasker and Stedinger (1986) and employed by Reis and others (2005). The unbiased station skew estimated by using the P_{RL} is

$$\hat{\gamma}_i = \left[1 + \frac{6}{P_{RL,i}}\right] G_i , \tag{4}$$

where

$\hat{\gamma}_i$ is the unbiased station skew estimate for site i,

$P_{RL,i}$ is the pseudo record length in years for site i as calculated in equations 1 and 2, and

G_i is the traditional biased station skew estimator based on the flood-frequency analysis for site i.

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

$$Var\left[\hat{\gamma}_i\right] = \left[1 + \frac{6}{P_{RL,i}}\right]^2 Var\left[G_i\right], \tag{5}$$

where

$Var[G_i]$ is calculated by using the formula (Griffis and Stedinger, 2009).

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

where

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

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

$$c\left(P_{RL}\right) = -\frac{7.31}{P_{RL}^{\ 0.59}} + \frac{45.90}{P_{RL}^{\ 1.18}} - \frac{86.50}{P_{RL}^{\ 1.77}} .$$

For the 368 streamgages in the study area used in the skew analysis, the unbiased station skew ranged from −1.37 to 4.13 in log units (table 1; see fig. 3 at https://doi.org/10.3133/sir20195105).

Estimating the Mean Square Error of the Skew

There are several possible ways to estimate $MSE\left(\tilde{G}\right)$. The approach used by EMA (taken from equation 55 in Cohn and others, 2001) generates a first-order estimate of the $MSE\left(\tilde{G}\right)$, which should perform well when interval data are available. Another option is to use the Griffis and Stedinger (2009) formula in equations 1–7 (the variance is equated to the MSE) by employing either the gaged-record length or the length of the entire historical period (from the beginning year to the ending year of the record); however, this method does not account for censored data and can lead to an inaccurate and underestimated $MSE\left(\tilde{G}\right)$. This issue was addressed by using the P_{RL} instead of the length of the historical period; the P_{RL} accounts for the effects of the censored data and the number of recorded gaged peaks. Thus, the unbiased $MSE\left(\tilde{G}\right)$ was used in the regional skewness model because it is more stable and relatively independent

of the station skew estimator (Griffis and Stedinger, 2009). This method also was used in previous regional skew studies (Parrett and others, 2011; Eash and others, 2013; Paretti and others, 2014; Southard and Veilleux, 2014; Wagner and others, 2016).

Cross-Correlation Model

A critical step for the GLS analysis is the estimation of the cross correlation of the station skew coefficient estimators. Martins and Stedinger (2002) used Monte Carlo experiments to derive a relation between the cross correlation of the skew estimators for two streamgages (i and j) as a function of the cross correlation of concurrent annual peak-flows (ρ_{ij}):

$$\hat{\rho}\left(\hat{\gamma}_i, \hat{\gamma}_j\right) = Sign\left(\hat{\rho}_{ij}\right) cf_{ij} \left|\hat{\rho}_{ij}\right|^k, \quad (7)$$

where

$\hat{\rho}_{ij}$ is the cross correlation of concurrent annual peak-flow for two streamgages,
k is a constant between 2.8 and 3.3, and
cf_{ij} is a factor that accounts for the sample size difference between the concurrent record lengths of the two streamgages and is defined as follows:

$$cf_{ij} = CY_{ij} / \sqrt{\left(P_{RL,i}\right)\left(P_{RL,j}\right)}, \quad (8)$$

where

CY_{ij} is the pseudo concurrent record length and
$P_{RL,i}$, $P_{RL,j}$ are the pseudo record lengths corresponding to streamgages i and j, respectively.

As shown in equation 8, the pseudo concurrent record length, CY_{ij}, is used to compute the cross correlation of station skews. The pseudo concurrent record length depends upon the years of common historical records between the two streamgages as well as the ratio of the pseudo record length to the historical record length (H_i) for each streamgage. Because censored and historical data are used, calculation of the effective concurrent record length is more complex than simply determining the years during which the two streamgages both recorded peaks.

To compute CY_{ij}, the years of historical record in common between the two streamgages are first determined. For the years in common, the following equation that includes the beginning year (YB_{ij}) and ending year (YE_{ij}) is then used to calculate the concurrent years of record between two streamgages (i and j):

$$CY_{ij} = \left(YE_{ij} - YB_{ij} + 1\right)\left(\frac{P_{RL,i}}{H_i}\right)\left(\frac{P_{RL,j}}{H_j}\right). \quad (9)$$

A cross-correlation model for the annual peak flows in the study area was developed by using the base-10 logarithms of annual peak flows from 54 streamgages that generated 1,036 streamgage pairs with at least 85 years of concurrent gaged peaks. As shown in figure 4*A*, a logit model, termed the Fisher Z Transformation ($Z = \log[(1+r)/(1-r)]$), provides a convenient transformation of the sample correlations r_{ij} from the (−1, +1) range to the (−∞, +∞) range (Fisher, 1915, 1921). Models relating the cross correlations of the concurrent annual peak flows at two streamgages (ρ_{ij}) to various basin characteristics were considered. The adopted model, which uses only one explanatory variable for estimating the cross correlations of concurrent annual peak flows between two streamgages, is based on the distance, in miles, between basin centroids (D_{ij}):

$$\rho_{ij} = \frac{exp\left(2Z_{ij}\right) - 1}{exp\left(2Z_{ij}\right) + 1}, \quad (10)$$

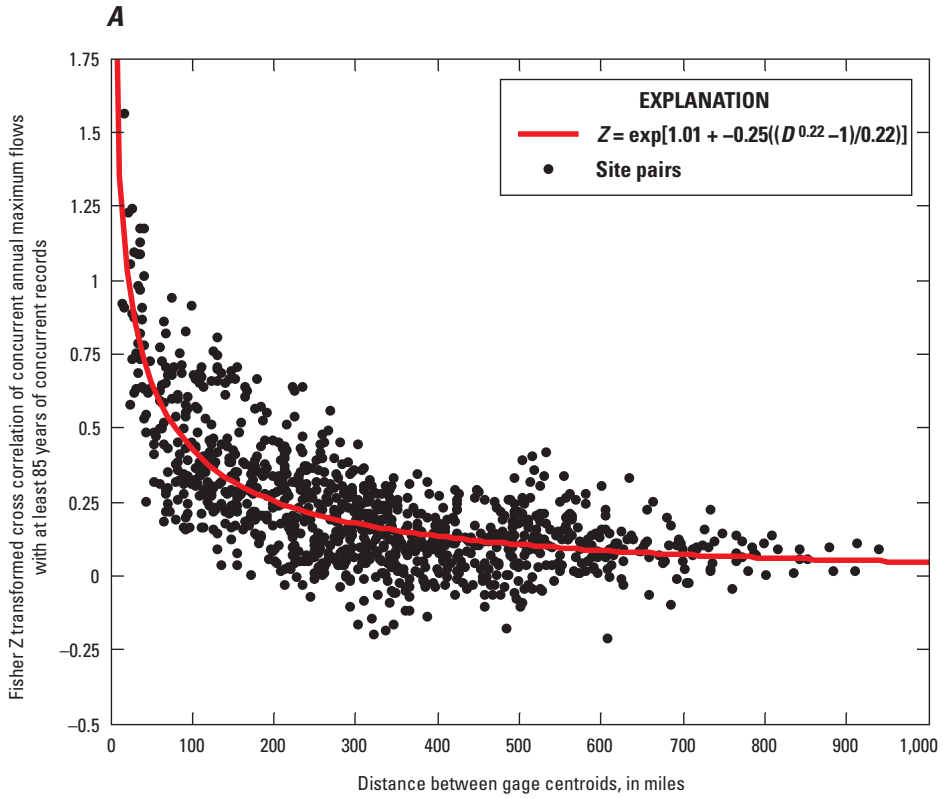
where

$$Z_{ij} = exp\left(0.89 - 0.18\left(\frac{D_{ij}^{0.29} - 1}{0.29}\right)\right) \quad (11)$$

An OLS regression analysis based on 1,036 streamgage pairs with at least 85 years of concurrent record indicated that this cross-correlation model is as accurate as having 119 years of concurrent annual peak flows from which to calculate cross correlation. As is the norm in an OLS analysis, each station pair in the model was given equal weight. By setting the concurrent-years threshold to 85, the model allowed the complete range of data in the study to be represented, while also minimizing the influence of station pairs with less accuracy and (or) less data. The fitted OLS regression relation between Z and the distance between basin centroids from the 1,036 streamgage pairs (fig. 4*A*) shows an exponential decline in the cross correlation for streamgages within 100 miles of each other. A similar decline is found in the cross correlation and distance between basin centroids for the untransformed streamgage pairs (fig. 4*B*). This model was used to estimate cross correlation for concurrent annual peak flows between all streamgage pairs used in the regional skew study.

Regression Analyses

The B−WLS/B−GLS method for computing a regional skew begins with an OLS analysis to develop a regional skew model that is used to generate an estimate of regional skew for each streamgage (Veilleux, 2011; Veilleux and others, 2011; Veilleux and others, 2012). The OLS-based regional skew estimate is the basis for computing the variance of the skew for each streamgage used in the WLS analysis. Next, B−WLS is used to generate estimators of the regional skew model parameters. Finally, B−GLS is used to estimate the precision of the WLS parameter values and the model error variance and its precision, and to compute various diagnostic statistics.



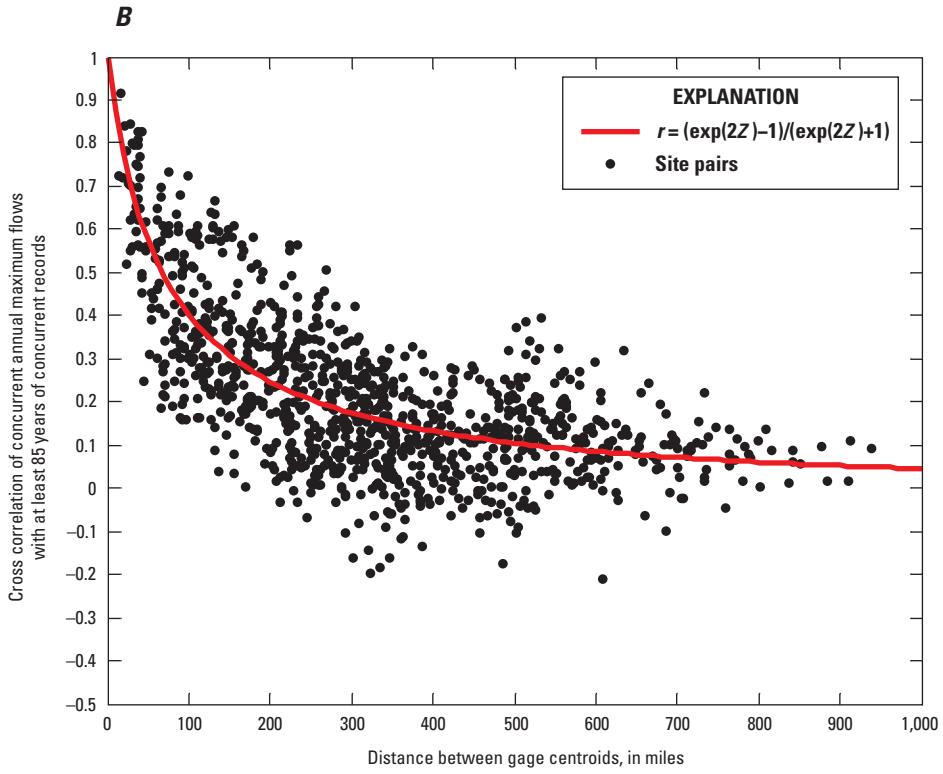


Figure 4. Graphs showing cross correlation of annual peak flows in the study area. *A*, Relation between Fisher Z transformed cross correlation of logarithms of annual peak flows and distances between basin centroids based on 1,036 streamgage pairs with concurrent record lengths greater than or equal to 85 years from 54 streamgages in the study area and *B*, Relation between untransformed cross correlation of logarithms of annual peak flows and distances between basin centroids, based on 1,036 streamgage pairs with concurrent record lengths greater than or equal to 85 years from 54 streamgages in the study area. Abbreviations: *r*, cross correlation of concurrent annual maximum flows; *D*, distance between gage centroids, in miles; and *Z*, Fisher Z transformed cross correlation of concurrent annual maximum flows

Ordinary Least Squares Analysis

The first step in the B−WLS/B−GLS regional skew analysis is to prepare an initial regional skew model by using OLS regression. The OLS regression analysis yields parameters (such as $\hat{\boldsymbol{\beta}}_{OLS}$) and a model that can be used to generate unbiased regional estimates of the skew for all streamgages:

$$\tilde{\boldsymbol{y}}_{OLS} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} \ , \tag{12}$$

where

$\tilde{\boldsymbol{y}}_{OLS}$ are the estimated regional skew values,
$\boldsymbol{X}$ is an ($n \times k$) matrix of basin characteristics,
$\hat{\boldsymbol{\beta}}_{OLS}$ is an ($k \times 1$) vector of estimated regression parameters,
n is the number of streamgages, and
k is the number of basin parameters, including a column of ones, to estimate the regression constant.

The estimated regional skew values $\tilde{\boldsymbol{y}}_{OLS}$ are then used to calculate unbiased streamgage regional skew variances by using equation 8 in Griffis and Stedinger (2009). These variances are based on the OLS estimator of the regional skew coefficient instead of the station skew estimator, making the weights in the subsequent steps relatively independent of the station skew estimates.

Weighted Least Squares Analysis

A WLS analysis is used to develop estimators of the regression coefficients for the regional skew model. The WLS analysis explicitly reflects variations in record length, but intentionally neglects cross correlations, thereby avoiding problems experienced with GLS parameter estimators (Veilleux, 2011; Veilleux and others, 2011).

The first step in the WLS analysis is to estimate the model error variance ($\sigma^2_{\delta,B-WLS}$) (Reis and others, 2005). Using a B−WLS approach to estimate the model error variance precludes the pitfall of estimating the model error variance as zero, which can occur when the method of moments WLS is used. Although the B−WLS analysis produces an estimate of the distribution of the model error variance, only the mean model error variance estimator is considered. Given the model error variance estimator, a B−WLS analysis is used to generate the weight matrix ($\boldsymbol{W}$) needed to compute estimates of the final regression parameters ($\hat{\boldsymbol{\beta}}_{WLS}$). To compute $\boldsymbol{W}$, a diagonal covariance matrix [$\boldsymbol{\Lambda}_{\boldsymbol{WLS}}\left(\sigma^2_{\delta,B-WLS}\right)$] is created (eq. 13). The diagonal elements of the covariance matrix are the sum of the estimated model error variance and the variance of the unbiased station skew ($Var\left[\hat{\gamma}_i\right]$), which depends upon on the record length and the estimate of the previously calculated OLS regional skew ($\tilde{\boldsymbol{y}}_{OLS}$). The off-diagonal elements of $\boldsymbol{\Lambda}_{\boldsymbol{WLS}}\left(\sigma^2_{\delta,B-WLS}\right)$ are zero because cross correlations among sets of streamgage data are not considered in the B−WLS analysis. Thus, the ($n \times n$) covariance matrix, $\boldsymbol{\Lambda}_{\boldsymbol{WLS}}\left(\sigma^2_{\delta,B-WLS}\right)$ is given by

$$\boldsymbol{\Lambda}_{\boldsymbol{WLS}}\left(\sigma^2_{\delta,B-WLS}\right) = \sigma^2_{\delta,B-WLS}\boldsymbol{I} + diag\left(Var\left[\hat{\gamma}\right]\right), \tag{13}$$

where

$\sigma^2_{\delta,B-WLS}$ is the model error variance,
$\boldsymbol{I}$ is an ($n \times n$) identity matrix,
n is the number of streamgages in the study, and
$diag\left(Var\left[\hat{\gamma}\right]\right)$ is the ($n \times n$) matrix containing the variance of the unbiased station skew, $Var\left[\hat{\gamma}_i\right]$, on the diagonal and zeros on the off-diagonal.

By using the covariance matrix, the WLS weights are calculated as

$$\boldsymbol{W} = \left[\boldsymbol{X}^T \boldsymbol{\Lambda}_{\boldsymbol{WLS}} \left(\sigma^2_{\delta,B-WLS} \right)^{-1} \boldsymbol{X} \right]^{-1} \boldsymbol{X}^T \boldsymbol{\Lambda}_{\boldsymbol{WLS}} \left(\sigma^2_{\delta,B-WLS} \right)^{-1}, \tag{14}$$

where

$\boldsymbol{W}$ is the ($k \times n$) matrix of weights,
$\boldsymbol{X}$ is the ($n \times k$) matrix of explanatory basin parameters,
$\boldsymbol{\Lambda}_{\boldsymbol{WLS}} \left(\sigma^2_{\delta,B-WLS} \right)$ is the ($n \times n$) covariance matrix, and
k is the number of basin parameters, including a column of ones, to estimate the regression constant.

These weights are used to compute the final estimates of the regression parameters ($\hat{\boldsymbol{\beta}}$) as

$$\hat{\boldsymbol{\beta}}_{WLS} = \boldsymbol{W} \hat{\gamma}, \tag{15}$$

where

$\hat{\boldsymbol{\beta}}_{WLS}$ is the ($k \times 1$) vector of estimated regression parameters.

Generalized Least Squares Analysis

After the regression model coefficients ($\hat{\boldsymbol{\beta}}_{WLS}$) and weights ($\boldsymbol{W}$) have been determined by using a B−WLS analysis, the degrees of precision of the fitted model and the regression coefficients also are estimated by using a B−GLS analysis. Using the B−GLS regression framework for regional skew, Reis and others (2005) developed the posterior probability-density function for model error variance described as

$$f\left(\sigma^2_{\delta,B-GLS} \mid \hat{\gamma}, \hat{\boldsymbol{\beta}}_{WLS}\right) \propto \xi\left(\sigma^2_{\delta,B-GLS}\right) \times \left|\boldsymbol{\Lambda}_{\mathbf{GLS}}\left(\sigma^2_{\delta,B-GLS}\right)\right|^{-0.5} \times \exp\left[-0.5\left(\hat{\gamma} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{WLS}\right)^T \left(\boldsymbol{\Lambda}_{\mathbf{GLS}}\left(\sigma^2_{\delta,B-GLS}\right)\right)^{-1} \left(\hat{\gamma} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_{WLS}\right)\right] \tag{16}$$

where

$\hat{\gamma}$ represents the skew data, and
$\xi\left(\sigma^2_{\delta,B-GLS}\right)$ is the exponential prior for the model error variance defined as

$$\xi\left(\sigma^2_{\delta,B-GLS}\right) = \lambda e^{-\lambda\left(\sigma^2_{\delta,B-GLS}\right)}, \sigma^2_{\delta,B-GLS} > 0 \cdot \tag{17}$$

The value 10 was adopted for lambda (λ) on the basis of a mean model error variance of 1/10. That prior assigns a 63-percent probability to the interval (0, 0.1), 86-percent probability to the interval (0, 0.2), and 95-percent probability to the interval (0, 0.3).

The mean B-GLS model error variance ($\sigma^2_{\delta,B-GLS}$) can then be used to compute the precision of the regression parameters ($\hat{\boldsymbol{\beta}}_{WLS}$) that were based on the B−WLS weights ($\boldsymbol{W}$). The B−GLS covariance matrix for the B−WLS estimator ($\hat{\boldsymbol{\beta}}_{WLS}$) is simply

$$\Sigma\left(\hat{\boldsymbol{\beta}}_{WLS}\right) = \boldsymbol{W}\boldsymbol{\Lambda}_{\boldsymbol{GLS}}\left(\sigma^2_{\delta,B-GLS}\right)\boldsymbol{W}^T, \tag{18}$$

where

$\boldsymbol{W}$ is the ($k \times n$) matrix of weights determined by B−WLS analysis, and
$\boldsymbol{\Lambda}_{\boldsymbol{GLS}}\left(\sigma^2_{\delta,B-WLS}\right)$ is an ($n \times n$) GLS covariance matrix calculated as

$$\boldsymbol{\Lambda}_{\boldsymbol{GLS}}\left(\sigma^2_{\delta,B-GLS}\right) = \sigma^2_{\delta,B-GLS}\boldsymbol{I} + \Sigma\left(\hat{\gamma}\right), \tag{19}$$

where

$\boldsymbol{I}$ is an ($n \times n$) identity matrix, and

$\Sigma(\hat{\gamma})$ is a full ($n \times n$) matrix containing the sampling variances of the streamflow record's unbiased skew, $Var[\hat{\gamma}_i]$, and the covariances of the skew $\hat{\gamma}_i$.

The off-diagonal values of $\Sigma(\hat{\gamma})$ are determined by the cross correlation of concurrent gaged annual peak flows and the *cf* factor, which accounts for the size differences between pairs of samples collected at different streamgages and their concurrent record length (see eq. 8; Martins and Stedinger, 2002). In the calculation of the *cf* factor by using the ratio of the number of concurrent peak flows at streamgage pairs to the total number of annual peak flows at both streamgages, only the gaged records and historical peaks are considered. Thus, any additional information provided by perception thresholds and censored peaks in the EMA analysis is neglected in the calculation of the cross correlation of annual peak flows and the *cf* factor. Precision metrics include (1) the standard error of the regression parameters [$SE(\hat{\beta}_{WLS})$]; (2) the model error variance ($\sigma^2_{\delta,B-GLS}$); (3) pseudo coefficient of determination (pseudo R^2_δ); and (4) the average variance of prediction at a streamgage not used in the regional model (AVP_{new}).

Results and Discussion

Final Bayesian Weighted Least Squares/ Bayesian Generalized Least Squares Regression Model

A constant B−WLS/B−GLS model having a skew of 0.086 and developed by using data from 368 streamgages with at least 36 years of P_{RL} each, produced the only statistically significant model of skew in the study area (table 3). A constant model does not explain any variability in skew; therefore, the pseudo R^2_δ, a diagnostic statistic that describes the percentage of the variability in the skew from streamgage to streamgage that is estimated by the model (Gruber and others, 2007; Parrett and others, 2011), is 0 percent. All available basin characteristics were evaluated as possible explanatory variables in the B−WLS/B−GLS regression analysis; however, the addition of any of the available basin characteristics or combinations thereof did not produce a pseudo R^2_δ greater than 5 percent, indicating that they did not explain the variation in station skews in the study area. Thus, the addition of basin characteristics as explanatory variables was not warranted because the increase in complexity did not result in a gain in precision.

The posterior mean of the constant model error variance (σ^2_δ) is 0.13. The average sampling error variance (*ASEV*) of the constant model is 0.0031, which represents the average error in the regional skew as calculated from the station skew values measured at streamgages used in the analysis. The average variance of prediction at a new streamgage (AVP_{new}) corresponds to the MSE used in B17B to describe the precision of the generalized skew map. The constant model has an AVP_{new} of 0.13, which corresponds to an effective record length of 54 years. An AVP_{new} of 0.13 is a marked improvement over the B17B national skew map, whose reported MSE of 0.302 has a corresponding effective record length of only 17 years (Interagency Advisory Committee on Water Data, 1982). Measured by effective record length, the new regional model includes more than three times the information of that of the B17B map. Appendix 1 provides a graphical assessment of the B−WLS/B−GLS model of regional skew.

Bayesian Weighted Least Squares/ Bayesian Generalized Least Squares Regression Diagnostics

To determine whether a regression model is a good representation of the data and which regression parameters, if any, should be included in the model, diagnostic statistics have been developed to evaluate how well a model fits a regional hydrologic dataset (Griffis, 2006; Gruber and Stedinger, 2008). In a regional skew study, potential explanatory variables are statistically evaluated to ensure an accurate prediction of skew while also keeping the model as simple as possible.

A pseudo analysis of variance (pseudo ANOVA) contains regression diagnostics and goodness-of-fit statistics that describe how much of the variation in the observations can be attributed to the regional model, and how much of the variation in the residuals can be attributed to modeling and sampling error (table 4; see fig. 5 at https://doi.org/10.3133/sir20195105). Determining these quantities is difficult; the

Table 3. Regional skew model and model fit for parts of the Great Lakes and Ohio River Basins.

[Standard deviations are in parentheses. σ^2_δ, model error variance; *ASEV*, average sampling error variance; AVP_{new}, average variance of prediction for a new site; Pseudo R^2_δ, fraction of the variability in the station skews explained by each model (Gruber and others, 2007)]

Model	Regression constant	σ^2_δ	*ASEV*	AVP_{new}	Pseudo R^2_δ (percent)
Constant	0.086	0.13	0.0031	0.13	0
	(0.055)	(0.015)			

modeling errors cannot be resolved because the values of the sampling errors (η_i) for each streamgage (i) are not known. However, the total sampling error sum of squares (SS) can be described by its mean value ($\Sigma_{i=1}^{n} Var\left[\hat{\gamma}_i\right]$), as there are n equations, and the total variation caused by 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 $\Sigma_{i=1}^{n} Var\left[\hat{\gamma}_i\right]$, and the residual variation attributed to the model error is $n\sigma_\delta^2(k)$.

For a model with no explanatory parameters other than the mean (the constant model), the estimated model error variance ($\sigma_\delta^2(0)$) describes all of the variation in $\gamma_i = \mu + \delta_i$, where μ is the mean of the estimated station skews. Thus, the total variation resulting from model error (δ_i) and sampling error ($\eta_i = \hat{\gamma}_i - \gamma_i$) in the expected sum of squares should equal $\sigma_\delta^2(0) + \Sigma_{i=1}^{n} Var\left(\hat{\gamma}_i\right)$. For a model type other than constant, the expected sum of squares attributed with k parameters equals $n\left[\sigma_\delta^2(0) - \sigma_\delta^2(k)\right]$ because the sum of the model error variance $n\sigma_\delta^2(k)$ and the variance explained by the model must equal $n\sigma_\delta^2(0)$. 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 determined from the computed residual errors and the observed model predictions, while not accounting for the effect of correlation on the sampling errors.

The error variance ratio (*EVR*) is a diagnostic statistic used to determine whether a simple OLS regression analysis would be sufficient, or a more sophisticated WLS or GLS analysis would be more appropriate. The *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 error variance is not negligible when compared to the model error variance, suggesting that a WLS or GLS regression analysis is appropriate. The *EVR* is calculated as

$$EVR = \frac{\mathrm{SS}(\text{sampling error})}{\mathrm{SS}(\text{model error})} = \frac{\Sigma_{i=1}^{n} Var\left(\hat{\gamma}_i\right)}{n\sigma_\delta^2(k)}. \qquad (20)$$

The constant model has a sampling error variance of 0.0031 (table 3) and an *EVR* of 1 (table 4), indicating that the sampling error variance is not negligible when compared to the model error variance, and that a WLS or GLS regression analysis was appropriate. Thus, an OLS model that neglects sampling error in the station skew would not provide a statistically reliable analysis of the data. Given the diagnostic statistics and the range of record lengths among streamgages,

Table 4. Pseudo analysis of variance (ANOVA) table for the constant model of regional skew in parts of the Great Lakes and Ohio River Basins.

[k, number of estimated regression parameters not including the constant; n, number of streamgages 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; $Var(\hat{\gamma}_i)$, variance of the estimated sample skew at site i; *EVR*, error variance ratio; *MBV**, misrepresentation of the beta variance; GLS, generalized least squares; WLS, weighted least squares; b_0^{WLS}, regression constant from WLS analysis; Λ, covariance matrix; 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	$n\left[\sigma_\delta^2(0) - \sigma_\delta^2(k)\right]$	0
Model error	$n-k-1$	367	$n\left[\sigma_\delta^2(k)\right]$	47
Sampling error	n	368	$\Sigma_{i=1}^{n} Var\left(\hat{\gamma}_i\right)$	46
Total	$2n-1$	735	$n\left[\sigma_\delta^2(k)\right] + \Sigma_{i=1}^{n} Var\left(\hat{\gamma}_i\right)$	93
$EVR = \frac{\Sigma_{i=1}^{n} Var\left(\hat{\gamma}_i\right)}{n\sigma_\delta^2(k)}$				1.0
$MBV^* = \frac{Var\left[b_0^{WLS} \mid GLS\ analysis\right]}{Var\left[b_0^{WLS} \mid WLS\ analysis\right]} = \frac{w^T \Lambda w}{w^T v}$			where $w_i = \frac{1}{\sqrt{\Lambda_{ii}}}$, $v = (n \times 1)$ vector of ones	4.7
Pseudo $R_\delta^2 = 1 - \frac{\sigma_\delta^2(k)}{\sigma_\delta^2(0)}$				0%

a WLS or GLS analysis was warranted to evaluate the final precision of the model.

The Misrepresentation of the Beta Variance (*MBV**) diagnostic statistic is used to determine whether a WLS regression is sufficient, or if a GLS regression is more appropriate to determine the precision of the estimated regression parameters (Griffis, 2006; Veilleux, 2011). 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 ($\hat{\gamma}_i$) generally has its greatest effect 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; conversely, if the *MBV** is not substantially greater than 1, a WLS analysis is sufficient. The *MBV** is calculated as

$$MBV^* = \frac{Var\left[b_0^{WLS} \mid GLS\ analysis\right]}{Var\left[b_0^{WLS} \mid WLS\ analysis\right]} = \frac{w^T \Lambda w}{\sum_{i=1}^{n} w_i}\ , \quad (21)$$

where $w_i = \frac{1}{\sqrt{\Lambda_{ii}}}$.

The *MBV** is equal to 4.7 for the constant model (table 4), indicating that the cross correlation among the skew estimators has an effect on the precision with which the regional skew can be estimated. If a WLS precision analysis were used for the estimated constant in the model, the variance would be underestimated by a factor of 4.7. Thus, a WLS analysis alone would misrepresent the variance of the constant in the skew model. Moreover, a WLS model would underestimate the variance of prediction, given that the sampling error in the constant term in both models was sufficiently large to make an appreciable contribution to the average variance of prediction.

Leverage and Influence

Diagnostic statistics for leverage and influence can be used to identify atypical observations and to address lack-of-fit when skew coefficients are estimated. The leverage statistics identify those streamgages in the analysis for which the observed streamflow values have a large impact on the fitted (or predicted) values (Hoaglin and Welsch, 1978). Generally, leverage statistics can determine 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 points that have the ability or potential to affect the fit of the regression, influence attempts to describe those points that have an unusual effect 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. If p is the number of estimated regression coefficients (p=1 for a constant model), and n is the sample size (or number of streamgages in the study), then leverage values have a mean of p/n, and values greater than $2p/n$ are generally considered large. Influence values greater than $4/n$ are typically considered large (Veilleux, 2011; Veilleux and others, 2011).

For the constant model of skew in the study area, influence greater than 0.011 (p/n = 4/368) and leverage greater than 0.005 [(2×1)/368] were considered high. No sites in the study area exhibited high leverage; therefore, the differences in the leverage values for the constant model reflect the variation in record lengths among sites. Eighteen streamgages in the study area exhibited high influence, and thus had an unusual effect on the fitted regression (table 5). These streamgages also had 18 of the 31 largest residuals (in magnitude) among the 368 streamgages used in the B−WLS/B−GLS analysis.

Summary

Bulletin 17C (B17C) guidelines recommend fitting the log-Pearson Type III (LP−III) distribution to a series of annual peak flows at a station by using the method of moments. The LP−III distribution is described by three moments: the mean, the standard deviation, and the skewness coefficient. The third moment, the skewness coefficient (hereinafter referred to as "skew"), is a measure of the asymmetry of the distribution or, in other words, the thickness of the tails of the distribution. In flood-frequency analysis, the skew is important because the magnitude of annual exceedance probability (AEP) flows for a streamgage estimated by using the LP−III distribution are affected by the skew of the annual peak flows (hereinafter referred to as "station skew"); interest is focused on the right-hand tail of the distribution, which represents annual peak flows corresponding to small AEPs and the larger flood flows. For streamgages having modest record lengths, the skew is sensitive to extreme events, such as large floods, as they cause a sample to be highly asymmetrical, or skewed. Thus, B17C recommends using a weighted-average skew computed from the station skew for a given streamgage and a regional skew. These choices reduce the sensitivity of the station skew to extreme events, particularly for streamgages with short record lengths. An estimate of regional skew is generated for a study area encompassing most of the Great Lakes Basin (hydrologic unit 04) and part of the Ohio River Basin (hydrologic unit 05), including the States of Indiana, Michigan, and Ohio and parts of the States of Illinois, Minnesota, New York, Pennsylvania, and Wisconsin. The study area spans approximately 1,600 kilometers from east to west, from northeastern Minnesota to the New York-Vermont border, and approximately 1,200 kilometers north to south, from northeastern Minnesota

Table 5. Gages with high influence on the constant model of regional skew for parts of the Great Lakes and Ohio River Basins.

[High influence is defined as Cook's D values greater than 4/*n* (or 4/368=0.011). Each of the 368 sites in the regional skew study was assigned a value from 1 to 368 signifying its relative rank, where a rank of 1 corresponds to the largest positive value in each category. The table is sorted from the largest to smallest value of influence. ERL, effective record length; MSE, mean square error; IL, Illinois; IN, Indiana; KY, Kentucky; MI, Michigan; NY, New York; OH, Ohio; PA, Pennsylvania; VT, Vermont; WI, Wisconsin; WV, West Virginia]

Index number	USGS streamgage number	State in which streamgage is located	Cook's D	Leverage	Pseudo ERL (years)		Unbiased at-site skew		Unbiased MSE (at-site skew)		Residual	
					Value	Rank	Value	Rank	Value	Rank	Value	Rank
681	03219500	OH	0.123	0.0035	115	12	2.3	2	0.36	5	2.2	2
664	03139000	OH	0.046	0.0031	83	74	1.7	4	0.38	3	1.6	4
246	03368000	IN	0.025	0.0026	58	201	1.6	5	0.37	4	1.5	5
974	03204000	WV	0.024	0.0027	61	178	−1.4	368	0.27	17	−1.5	368
23	03345500	IL	0.023	0.0037	128	8	−0.83	360	0.09	306	−0.91	360
376	04114498	MI	0.022	0.0028	67	154	−1.2	365	0.22	29	−1.3	365
964	03070500	WV	0.021	0.0034	100	29	1.1	12	0.14	151	1.0	12
432	04156000	MI	0.020	0.0037	129	7	−0.75	357	0.08	322	−0.84	357
203	03335700	IN	0.017	0.0027	60	186	−1.2	364	0.19	64	−1.3	364
888	04079000	WI	0.016	0.0032	89	52	−0.84	361	0.12	197	−0.93	361
866	04288000	VT	0.015	0.0038	150	3	0.79	29	0.07	346	0.7	29
284	03250000	KY	0.015	0.0024	49	263	−1.3	367	0.29	12	−1.4	367
677	03159540	OH	0.014	0.0024	48	269	1.4	6	0.35	6	1.3	6
682	03220000	OH	0.013	0.0029	71	131	1.0	13	0.18	77	0.95	13
843	03049800	PA	0.012	0.0025	51	249	1.3	7	0.29	10	1.2	7
750	04200500	OH	0.012	0.0029	69	142	1.0	15	0.18	72	0.91	15
612	04256000	NY	0.011	0.0029	71	132	1.0	17	0.16	108	0.89	17
166	03274650	IN	0.011	0.0022	43	311	−1.3	366	0.32	7	−1.3	366

near Lake Superior to the Ohio River on the southern boundary of Illinois.

Candidate streamgages in the study area were selected by the USGS in the respective States. Only streamgage records unaffected by extensive regulation, diversion, urbanization, or channelization, and having 25 or more years of gaged record were considered.

As recommended in B17C, the flood frequency for each candidate streamgage was determined by employing the Expected Moments Algorithm (EMA), which extends the method of moments so that it can accommodate interval, censored, and historical/paleo data, as well as use the Multiple Grubbs-Beck test (MGBT) to identify potentially influential low floods (PILFs) in the data series.

A total of 551 candidate streamgages were initially considered for use in the skew analysis; after screening for redundancy and sufficient pseudo record length (P_{RL}), 368 streamgages were selected. The Bayesian weighted least squares/Bayesian generalized least squares (B−WLS/B−GLS) regression method was used to develop a regional skew model for the study area that would incorporate possible explanatory variables (basin characteristics) to explain the variation in skew in the study area. Basin characteristics for candidate streamgages were obtained from the GAGES II database or generated by using the ArcHydro package in Esri ArcGIS version 10.3.1. Twelve basin characteristics were considered as possible explanatory variables in the B−WLS/B−GLS regression analysis; however, none produced a pseudo coefficient of determination greater than 5 percent, indicating that they did not explain the variation in station skews in the study area. Therefore, a constant skew model was selected. The constant model has a regional skew coefficient of 0.086 and an average variance of prediction (AVP_{new}) of 0.13, which corresponds to the mean square error (MSE). An AVP_{new} of 0.13 corresponds to an effective record length of 54 years, which is a marked improvement over the Bulletin 17B (B17B) national skew map, whose reported MSE of 0.302 has a corresponding effective record length of only 17 years. Measured by effective record length, the new regional model provides more than three times the amount of information provided by the B17B map.

Acknowledgments

The authors would like to acknowledge the following U.S. Geological Survey employees for selecting streamgages from their respective States and, in some cases, parts of adjacent States for use in the regional skew and flood-frequency analyses: Greg Koltun, Ohio-Kentucky-Indiana Water Science Center; Dave Holtschlag, Rob Waschbush, Chris Sanoki, Danny Morel, and Dave Lorenz, Upper Midwest Water Science Center; Doug Burns and Gary Wall, New York Water Science Center; Tom Over and David Soong, Central Midwest Water Science Center; and Mark Roland, Pennsylvania Water Science Center. The authors would also like to acknowledge Chris Sanoki and Danny Morel in the Upper Midwest Water Science Center for generating basin characteristics for streamgages that were not in the GAGES II database.

References Cited

Arguez, A., Durre, I., Applequist, S., Vose, R.S., Squires, M.F., Yin, X., Heim, R.R., Jr., and Owen, T.W., 2012, NOAA's 1981–2010 U.S. Climate Normals—An overview: Bulletin of the American Meteorological Society, v. 93, no. 11, p. 1687–1697, accessed October 1, 2019, at https://doi.org/10.1175/BAMS-D-11-00197.1.

Belsley, D.A., Kuh, E., and Welsch, R.E., 1980, Regression diagnostics—Identifying influential data and sources of collinearity: Hoboken, N.J., John Wiley & Sons, Inc., 300 p. [Also available at https://doi.org/10.1002/0471725153.

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 October 1, 2019, at https://doi.org/10.1029/97WR01640.

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. [Also available at https://doi.org/10.1029/2001WR900016.]

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

Curran, J.H., Barth, N.A., Veilleux, A.G., and Ourso, R.T., 2016, Estimating flood magnitude and frequency at gaged and ungaged sites on streams in Alaska and conterminous basins in Canada, based on data through water year 2012: U.S. Geological Survey Scientific Investigations Report 2016–5024, 47 p. [Also available at https://doi.org/10.3133/sir20165024.]

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.

England, J.F., Jr., Cohn, T.A., Faber, B.A., Stedinger, J.R., Thomas, W.O., Jr., Veilleux, A.G., Kiang, J.E., and Mason, R.R., Jr., 2018, Guidelines for determining flood flow frequency Bulletin 17C: U.S. Geological Survey Techniques and Methods, book 4, chap. B5, 148 p. [Also available at https://doi.org/10.3133/tm4B5.]

Environmental Systems Research Institute (Esri), 2009, ArcGIS desktop help, accessed November 23, 2015, at http://desktop.arcgis.com/en/desktop/.

Falcone, J.A., 2011, GAGES–II: Geospatial Attributes of Gages for Evaluating Streamflow: U.S. Geological Survey dataset. [Also available at https://pubs.er.usgs.gov/publication/70046617.]

Fenneman, N.M., 1938, Physiography of the Eastern United States (1st ed.): New York, McGraw-Hill Book Co., 714 p.

Fisher, R.A., 1915, Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population: Biometrika, v. 10, no. 4, p. 507–521. [Also available at https://www.jstor.org/stable/2331838.]

Fisher, R.A., 1921, On the "probable error of a coefficient of correlation deduced from a small sample": Metron, v. 1, p. 3–32.

Griffis, V.W., 2006, Flood-frequency analysis—Bulletin 17, regional information, and climate change: Cornell University, Ph.D. dissertation, 241 p.

Griffis, V.W., and Stedinger, J.R., 2007, Evolution of flood frequency analysis with Bulletin 17: Journal of Hydrologic Engineering, v. 12, no. 3, p. 283–297, accessed October 1, 2019, at https://doi.org/10.1061/(ASCE)1084-0699(2007)12:3(283).

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 (Amsterdam), v. 14, no. 2, p. 121–130.

Gruber, A.M., Reis, D.S., Jr., and Stedinger, J.R., 2007, Models of regional skew based on Bayesian GLS regression, *in* Kabbes, K.C., ed., Proceedings of the 2007 World Environmental and Water Resources Congress—Restoring our natural habitat, Tampa, Fla., May 15–18, 2007: Reston, Va., American Society of Civil Engineers, 10 p., accessed October 3, 2019, at https://doi.org/10.1061/40927(243)400.

Gruber, A.M., and Stedinger, J.R., 2008, Models of LP3 regional skew, data selection, and Bayesian GLS regression, *in* Babcock, R.W., and Walton, R., eds., World Environmental and Water Resources Congress 2008—Ahupua'A—Proceedings of the congress, May 12–16, 2008, Honolulu, Hawai'i: Reston, Va., American Society of Civil Engineers, p. 5575–5584. [Also available at https://doi.org/10.1061/40976(316)563.]

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.

Homer, C.G., Dewitz, J.A., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N.D., Wickham, J.D., and Megown, K., 2015, Completion of the 2011 National Land Cover Database for the conterminous United States—Representing a decade of land cover change information: Photogrammetric Engineering and Remote Sensing, v. 81, no. 5, p. 345–354, accessed October 12, 2018, at https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NERL&dirEntryId=309950.

Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood flow frequency—Bulletin 17B (revised and corrected): Washington, D.C., Hydrologic Subcommittee, 28 p.

Martins, E.S., and Stedinger, J.R., 2002, Cross correlations among estimators of shape: Water Resources Research, v. 38, no. 11, p. 34-1–34-7. [Also available at https://doi.org/10.1029/2002WR001589.]

Mastin, M.C., Konrad, C.P., Veilleux, A.G., and Tecca, A.E., 2016, Magnitude, frequency, and trends of floods at gaged and ungaged sites in Washington, based on data through water year 2014 (ver. 1.2, November 2017): U.S. Geological Survey Scientific Investigations Report 2016–5118, 70 p. [Also available at https://doi.org/10.3133/sir20165118.]

National Oceanic and Atmospheric Administration, 2014, Atlas 14 precipitation frequency estimates in GIS compatible format: National Weather Service Hydrometeorological Design Studies Center, Precipitation Frequency Data Server, accessed February 1, 2018, at https://hdsc.nws.noaa.gov/hdsc/pfds/pfds_gis.html.

Olson, S.A., 2014, Estimation of flood discharges at selected annual exceedance probabilities for unregulated, rural streams in Vermont, *with a section on* Vermont regional skew regression, by Veilleux, A.G.: U.S. Geological Survey Scientific Investigations Report 2014–5078, 27 p. plus appendixes. [Also available at http://dx.doi.org/10.3133/sir20145078.]

Parrett, C., 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. [Also available at https://doi.org/10.3133/sir20105260.]

Paretti, N.V., Kennedy, J.R., Turney, L.A., and Veilleux, A.G., 2014, Methods for estimating magnitude and frequency of floods in Arizona, developed with unregulated and rural peak-flow data through water year 2010: U.S. Geological Survey Scientific Investigations Report 2014–5211, 61 p. [Also available at https://doi.org/10.3133/sir20145211.]

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, no. 10, W10419. [Also available at https://doi.org/10.1029/2004WR003445.]

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. [Also available at https://doi.org/10.3133/sir20145165.]

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. [Also available at https://doi.org/10.1029/WR022i005p00785.]

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. [Also available at https://doi.org/10.1029/WR021i009p01421.]

Tasker, G.D., and Stedinger, J.R., 1986, Regional skew with weighted LS regression: Journal of Water Resources Planning and Management, v. 112, no. 2, p. 225–237. [Also available at https://doi.org/10.1061/(ASCE)0733-9496(1986)112:2(225).]

Tasker, G.D., and Stedinger, J.R., 1989, An operational GLS model for hydrologic regression: Journal of Hydrology, v. 111, nos. 1–4, p. 361–375. [Also available at https://doi.org/10.1016/0022-1694(89)90268-0.]

U.S. Department of Agriculture, 2008, U.S. general soil map (STATSGO): U.S. Department of Agriculture, Natural Resources Conservation Service database, accessed October 1, 2015, at http://www.soils.usda.gov/survey/geography/statsgo/.

U.S. Environmental Protection Agency and Government of Canada, 1995, The Great Lakes—An environmental atlas and resource book (3d ed.): U.S. Environmental Protection Agency 905–B–95–001, 46 p.

U.S. Geological Survey, 2015, USGS water data for the Nation: U.S. Geological Survey National Water Information System database, accessed October 1, 2015, at https://doi.org/10.5066/F7P55KJN. [Peak streamflow information directly accessible at https://nwis.waterdata.usgs.gov/nwis/peak.]

Veilleux, A.G., 2009, Bayesian GLS regression for regionalization of hydrologic statistics, floods and Bulletin 17 skew: Cornell University, M.S. thesis, 155 p.

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

Veilleux, A.G., Cohn, T.A., Flynn, K.M., Mason, 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. [Also available at https://doi.org/10.3133/fs20133108.]

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, *in* Loucks, E.D., ed., Proceedings of the World Environmental and Water Resources Congress 2012—Crossing boundaries, Albuquerque, N. Mex., May 20–24, 2012: Reston, Va., American Society of Civil Engineers, p. 2253–2263.

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, *in* Beighley, R.E., II, and Killgore, M.W., eds., Proceedings of the World Environmental and Water Resources Congress 2011—Bearing knowledge for sustainability, Palm Springs, Calif., May 22–26, 2011: Reston, Va., American Society of Civil Engineers, p. 3103–3123.

Wagner, D.M., Krieger, J.D., and Veilleux, A.G., 2016, Methods for estimating annual exceedance probability discharges for streams in Arkansas, based on data through water year 2013: U.S. Geological Survey Scientific Investigations Report 2016–5081, 136 p. [Also available at https://doi.org/10.3133/sir20165081.]

Wagner, D.M., and Veilleux, A.G., 2019, Annual peak-flow data, PeakFQ specification files, and PeakFQ output files for 368 selected streamflow gaging stations operated by the U.S. Geological Survey in the Great Lakes and Ohio River Basins that were used to estimate regional skewness of annual peak flows: U.S. Geological Survey data release, https://doi.org/10.5066/P9N7UAFJ.

Wolock, D.M., 1997, STATSGO soil characteristics for the conterminous United States: U.S. Geological Survey Open-File Report 97–656, accessed October 7, 2019, at https://doi.org/10.3133/ofr97656.

Appendix

Appendix 1. Assessment of a regional skew model for parts of the Great Lakes and Ohio River Basins by using Monte Carlo simulations

This appendix provides a graphical assessment of the Bayesian weighted least squares/Bayesian generalized least squares (B–WLS/B–GLS) model of regional skew that is described in this report for parts of the Great Lakes and Ohio River Basins. Observed, unbiased station skews are depicted in figure 1.1 along with contour lines and shading to provide a sense of geographic patterns in the skews. The contouring algorithm used to generate figure 1.1 shows a substantial amount of structure in the pattern of the unbiased station skews. The larger skews (positive skews) in eastern Ohio and western Pennsylvania might be a cause for concern.

Monte Carlo simulations were used to determine whether the apparent observed structure in the station skews is evidence of significant model misspecification or an artifact of random-sampling variability possibly confounded by the covariance structure of the errors. The Monte Carlo simulations were generated from a multivariate normal distribution with a mean equal to the constant from the regional skew model and a covariance matrix identical to the covariance matrix used in the regional skew model. The constant model of skew in the study area is:

$$\hat{\gamma}_{BWLS/BGLS} = 0.086 + \varepsilon \ , \tag{1.1}$$

where ε represents the total error and

$$\varepsilon \sim N\left(0, Var\left(\varepsilon\right)\right), \tag{1.2}$$

where N signifies a normal distribution of the total error in the constant regional skew model determined in the B–GLS analysis.

As described in equation 1.2, the $Var(\varepsilon)$ can be described as

$$\left[\varepsilon\varepsilon^{T}\right] = \Lambda_{GLS}\left(\sigma^{2}_{\delta,B-GLS}\right) = \sigma^{2}_{\delta,B-GLS}\boldsymbol{I} + \boldsymbol{\Sigma}\left(\hat{\gamma}\right), \tag{1.3}$$

where

$\Lambda_{GLS}\left(\sigma^{2}_{\delta,B-GLS}\right)$ is the $(n \times n)$ GLS covariance matrix,

$\sigma^{2}_{\delta,B-GLS}$ is the B–GLS variance of the underlying model error δ,

$\boldsymbol{I}$ is an $(n \times n)$ identity matrix, and

$\boldsymbol{\Sigma}\left(\hat{\gamma}\right)$ is the full $(n \times n)$ covariance matrix of the sampling errors for each streamgage (n).

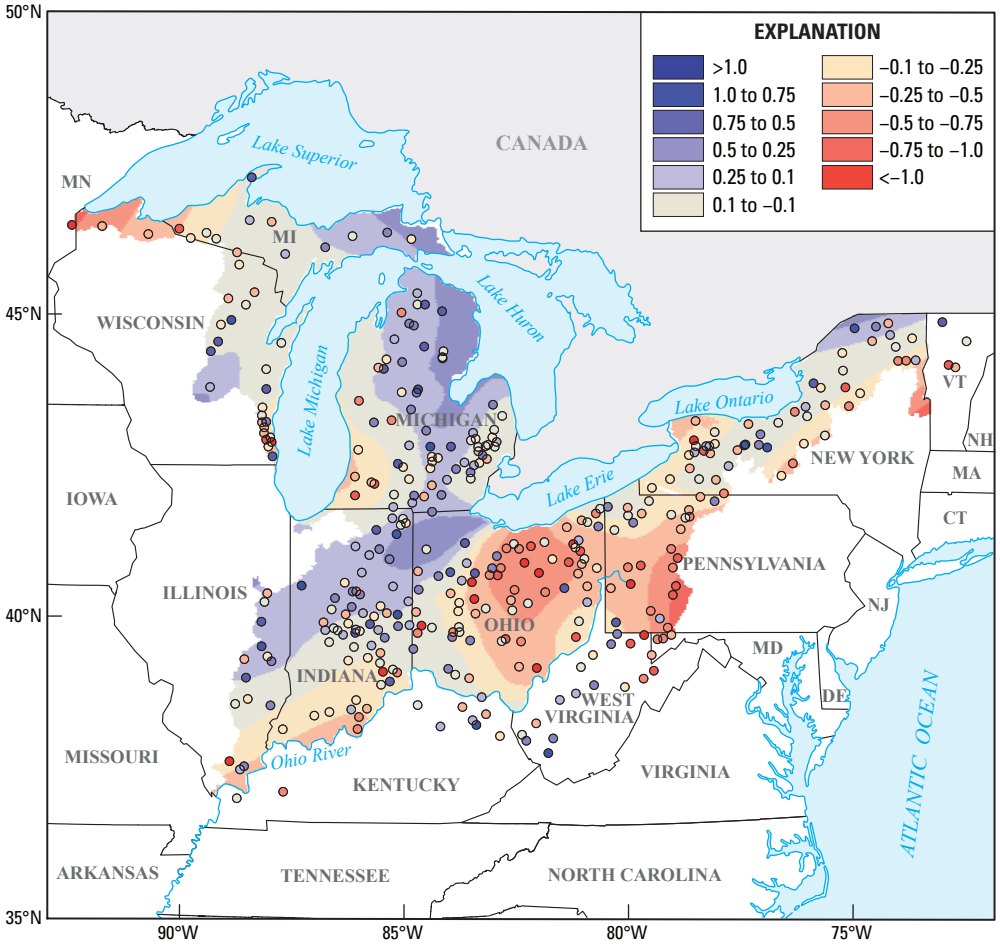


Figure 1.1. Contour map of unbiased station skews (unweighted) for the 368 streamgages used in the regional skew analysis for parts of the Great Lakes and Ohio River Basins.

The covariance matrix of the sampling errors is made up of the sampling variances of the unbiased station skew ($Var\left[\hat{\gamma}_i\right]$) and the covariances of the skewness estimators (γ_i). The off-diagonal values of $\boldsymbol{\Sigma}\left(\hat{\gamma}\right)$ are determined by the cross correlation of concurrent gaged annual peak flows and the *cf* factor (see eqs. 7 and 8 in report). The model error variance σ_δ^2 for the constant model is 0.13 (table 3) and was used in the Monte Carlo simulations. The covariance matrix $\boldsymbol{\Sigma}\left(\hat{\gamma}\right)$ used in the Monte Carlo simulations is the same as that used in the B−WLS/B−GLS regression analysis (see eqs. 13 and 18 in report).

The results of the Monte Carlo simulations are depicted graphically in 20 realizations of the expected patterns in the station skew if the station skews are normally distributed with a mean equal to 0.086 and the covariance matrix given by equation 1.3 (fig. 1.2). The Monte Carlo simulations reveal no consistent structure in the pattern of the station skews consistent with the observed pattern of the station skews in the constant model (fig. 1.1). Therefore, it seems reasonably safe to conclude that, based on the geographic patterns observed in the station skews, there is little evidence of a lack of fit.

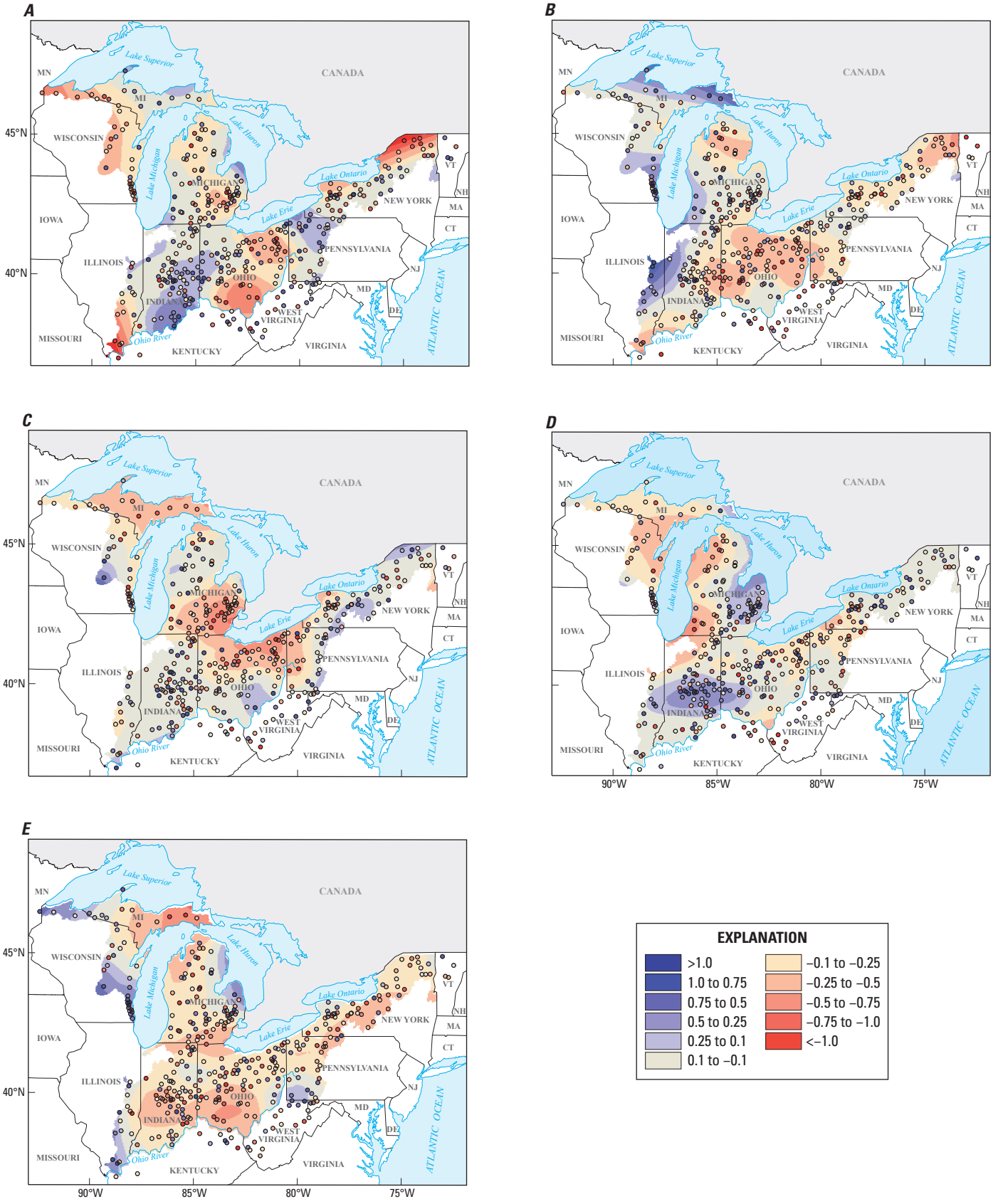


Figure 1.2. Contour maps showing results of 20 Monte Carlo simulations of skew at 368 streamgages in the Great Lakes and Ohio River Basins used in the regional skew analysis. Simulations are normally distributed to the constant skew model and covariance matrix.

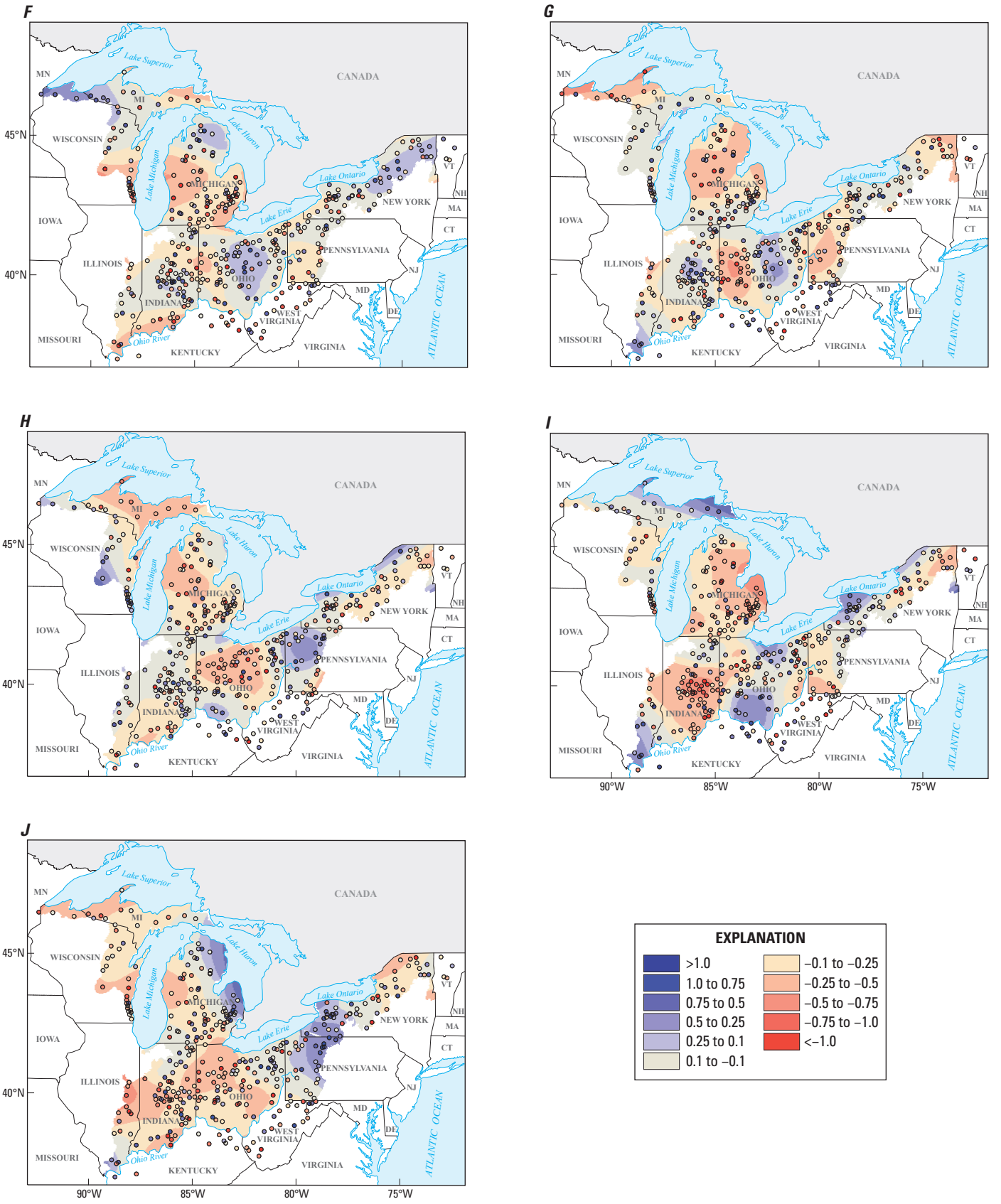


Figure 1.2. Continued.

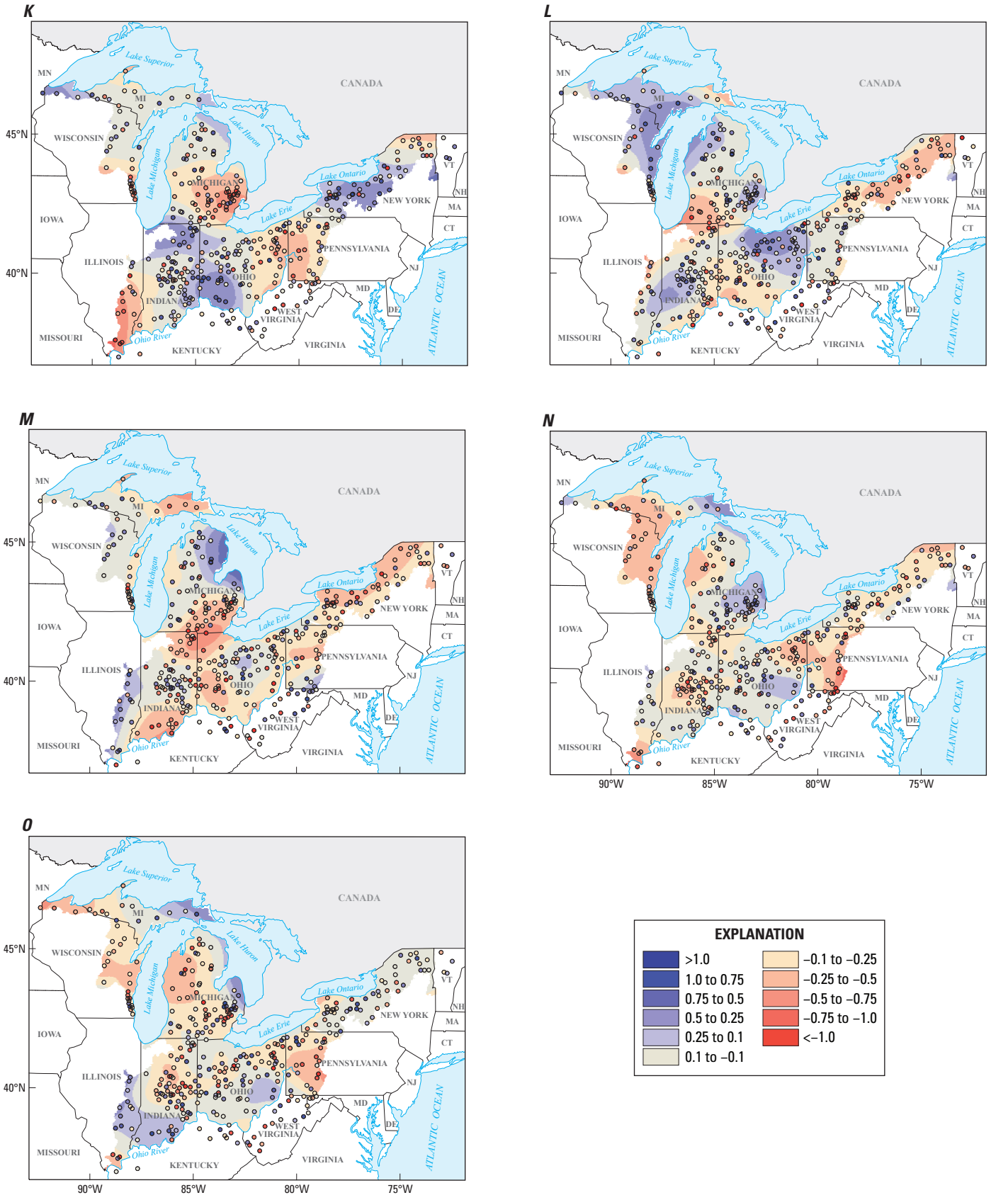


Figure 1.2. Continued.

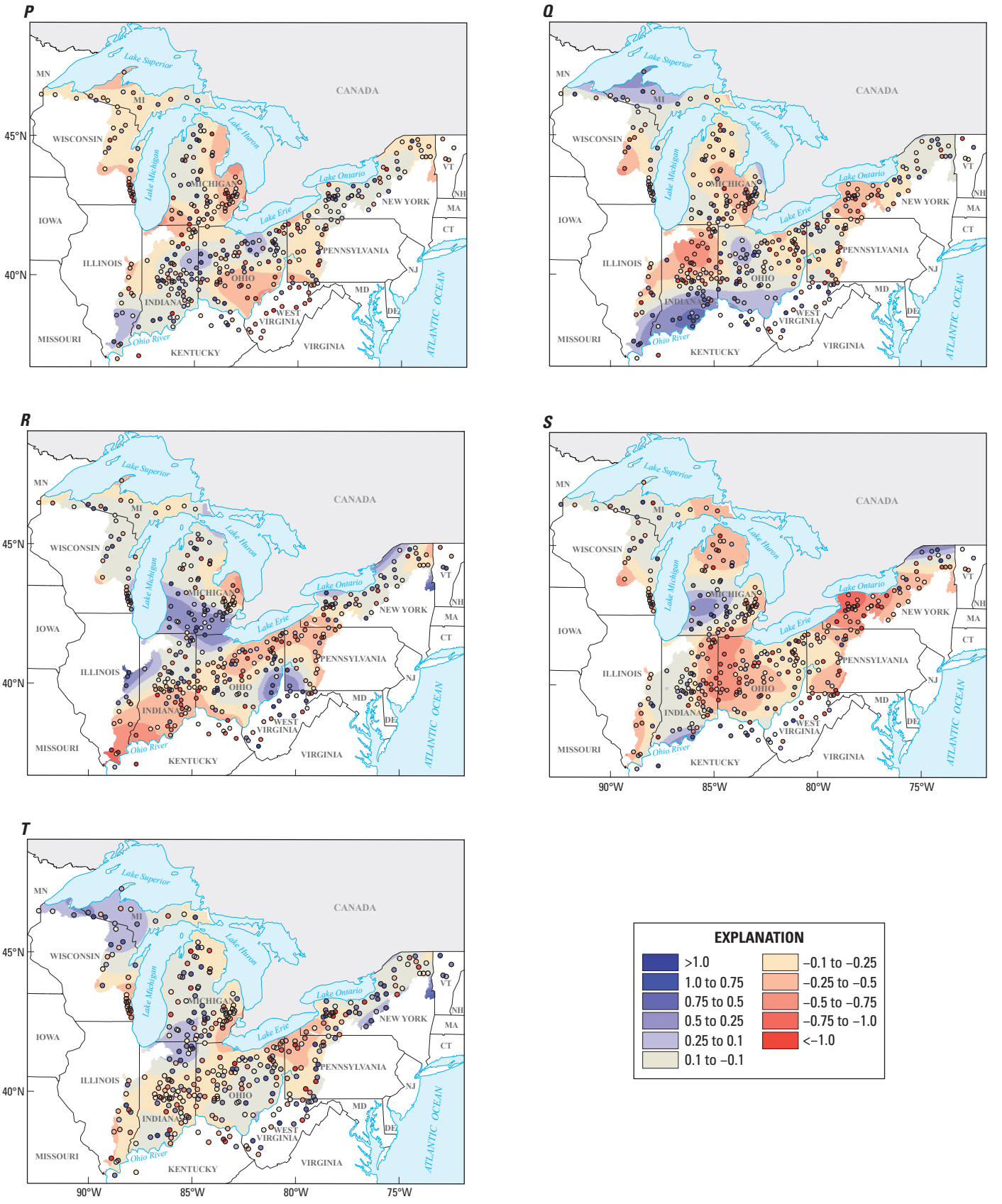


Figure 1.2. Continued.

ISSN 2328-0328 (online)
https://doi.org/10.3133/sir20195105