USGS
science for a changing world

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

Development of Regression Equations for the Estimation of Flood Flows at Ungaged Streams in Pennsylvania

Scientific Investigations Report 2019–5094
Supersedes USGS Scientific Investigations Report 2008–5102
Version 1.1, December 2020

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

Front cover: Photograph of inundated bridge downstream of U.S. Geological Survey streamgage 01573695 Conewago Creek near Bellaire, PA, after flooding associated with rainfall in July 2018. Photograph by Jason Ferree, U.S. Geological Survey.

Back cover: Photograph of vertical staff and crest-stage gage at U.S. Geological Survey streamgage 01573695 Conewago Creek near Bellaire, PA, after flooding associated with rainfall in July 2018. Photograph by Jason Ferree, U.S. Geological Survey.

Development of Regression Equations for the Estimation of Flood Flows at Ungaged Streams in Pennsylvania

By Mark A. Roland and Marla H. Stuckey

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

Scientific Investigations Report 2019–5094
Supersedes USGS Scientific Investigations Report 2008–5102
Version 1.1, December 2020

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
Supersedes USGS Scientific Investigations Report 2008–5102
First release: 2019
Revised: December 2020

Suggested citation:
Roland, M.A., and Stuckey, M.H., 2020, Development of regression equations for the estimation of flood flows at ungaged streams in Pennsylvania (ver. 1.1, December 2020): U.S. Geological Survey Scientific Investigations Report 2019–5094, 36 p., https://doi.org/10.3133/sir20195094. [Supersedes USGS Scientific Investigations Report 2008–5102]

Associated data for this publication:
Roland, M.A., and Stuckey, M.H., 2019, Data in support of development of regression equations for the estimation of flood flows at ungaged streams in Pennsylvania: U.S. Geological Survey data release, https://doi.org/10.5066/P9YHIU6G.

ISSN 2328-0328 (online)

Contents

Figures

Tables

Conversion Factors

U.S. customary units to International System of Units

Multiply	By	To obtain
	Length	
inch (in.)	2.54	centimeter (cm)
foot (ft)	0.3048	meter (m)
mile (mi)	1.609	kilometer (km)
	Area	
acre	0.004047	square kilometer (km^2)
square foot (ft^2)	0.09290	square meter (m^2)
square mile (mi^2)	2.590	square kilometer (km^2)
	Volume	
gallon (gal)	0.003785	cubic meter (m^3)
cubic foot (ft^3)	0.02832	cubic meter (m^3)
acre-foot (acre-ft)	1,233	cubic meter (m^3)
	Flow rate	
foot per year (ft/yr)	0.3048	meter per year (m/yr)
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)
gallon per day (gal/d)	0.003785	cubic meter per day (m^3/d)

Temperature in degrees Celsius (°C) may be converted to degrees Fahrenheit (°F) as follows:

$$°F = (1.8 \times °C) + 32.$$

Temperature in degrees Fahrenheit (°F) may be converted to degrees Celsius (°C) as follows:

$$°C = (°F – 32) / 1.8.32) / 1.8.$$

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).

Altitude, as used in this report, refers to distance above the vertical datum.

Abbreviations

AEP	annual exceedance probability
AVP	average variance of prediction
EMA	expected moments algorithm
EPA	U.S. Environmental Protection Agency
FEMA	Federal Emergency Management Agency
GIS	geographic information system
GLS	generalized least squares
HUC	hydrologic unit code
IACWD	Interagency Advisory Committee on Water Data
LP3	log-Pearson Type III
MGBT	Multiple Grubbs-Beck test
MSE	mean squared error
NWIS	National Water Information System
OLS	ordinary least squares
PADOT	Pennsylvania Department of Transportation
PeakFQ	U.S. Geological Survey peak flow analysis program
PILF	potentially influential low flow
USGS	U.S. Geological Survey
WLS	weighted least squares
WREG	weighted multiple linear regression

Development of Regression Equations for the Estimation of Flood Flows at Ungaged Streams in Pennsylvania

By Mark A. Roland and Marla H. Stuckey

Abstract

Regression equations, which may be used to estimate flood flows at select annual exceedance probabilities, were developed for ungaged streams in Pennsylvania. The equations were developed using annual peak flow data through water year 2015 and basin characteristics for 285 streamflow gaging stations across Pennsylvania and surrounding states. The streamgages included active and discontinued continuous-record stations, as well as crest-stage partial-record stations, and required a minimum of 10 years of annual peak streamflow data for inclusion in the study. Explanatory variables significant at the 95-percent confidence level for one or more regression equations included the following basin characteristics: drainage area, maximum basin elevation, mean basin slope, percent storage, and the percentage of carbonate bedrock within a basin. The State was divided into five regions, and regional regression equations were developed to estimate flood flows associated with the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probabilities (which correspond to the 2-, 5-, 10-, 25-, 50-, 100-, 200-, and 500-year recurrence intervals, respectively). Although the regression equations can be used to estimate the magnitude of flood flows for most streams in the State, they are not valid for streams with drainage areas generally greater than 1,500 square miles or with substantial regulation, diversion, or mining activity within the basin. The regional regression equations will be incorporated into the U.S. Geological Survey StreamStats application (https://water.usgs.gov/osw/streamstats/).

Additionally, annual peak flow data for 356 streamgages initially considered for inclusion in the analysis for development of updated flood-flow regression equations were analyzed for the existence of trends; estimates of flood-flow magnitude and frequency were also computed for these streamgages. Estimates of flood-flow magnitude and frequency for streamgages substantially affected by upstream regulation are also presented.

Introduction

Information regarding the magnitude and frequency of floods is critical to engineers and planners for the design of bridges, culverts, and other structures near streams and rivers, as well as for flood insurance studies and flood-plain management. Commonly used estimates of flood magnitude are associated with the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probabilities (AEPs). Respectively, these AEPs correspond to flood flows occurring, on average, once every 2, 5, 10, 20, 50, 100, 200, and 500 years. These flood flows are estimates based on statistical probabilities or frequencies, and the recurrence intervals associated with each flood flow refer to the average number of years between the floods (Dinicola, 1996). For example, the 100-year flood has a 1 in 100 chance (or 1 percent probability) that a flood of this magnitude will occur in any given year (table 1).

Regression equations for estimating the magnitude and frequency of flood flows were last developed for Pennsylvania by the U.S. Geological Survey (USGS) using annual peak-flow data generally through water year[1] 2005 (Roland and Stuckey, 2008). Flood events occurring in Pennsylvania since 2005, most notably the result of Hurricane Irene and Tropical Storm Lee in 2011, as well as advances in geospatially derived basin characteristics warranted updating the flood-flow regression equations for Pennsylvania. The USGS, in cooperation with the Federal Emergency Management Agency (FEMA) and the Pennsylvania Department of Transportation (PADOT) developed updated regression equations to estimate flood flows associated with the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent AEPs for ungaged streams in Pennsylvania not subject to substantial regulation, diversion, or mining activity. This report discusses the methodology used and presents the

[1]Water year is defined as a 12-month period beginning October 1 and ending September 30. The water year is designated by the calendar year in which it ends.

Table 1. Annual exceedance probabilities with corresponding recurrence intervals.

[Q_{AEP}, flood-flow magnitude for the indicated annual exceedance probability (AEP)]

Annual exceedance probability (percent)	Q_{AEP}	Probability of occurrence in any given year	Recurrence interval (years)
50	Q_{50}	1 in 2	2
20	Q_{20}	1 in 5	5
10	Q_{10}	1 in 10	10
4	Q_{4}	1 in 25	25
2	Q_{2}	1 in 50	50
1	Q_{1}	1 in 100	100
0.5	$Q_{0.5}$	1 in 200	200
0.2	$Q_{0.2}$	1 in 500	500

results of the regression analysis. In addition to the regression analysis, annual peak flow data at streamgages were examined for the presence of trends.

Purpose and Scope

This report presents the estimated magnitude of flood flows associated with the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent AEPs for 356 streamflow gaging stations (referred to hereafter as streamgages) across Pennsylvania and surrounding states not subject to substantial flow regulation, diversion, or mining activity. The magnitude and frequency of flood flows at the streamgages were computed using the expected moments algorithm (EMA) methodology (England and others, 2018), which fits a log-Pearson Type III (LP3) probability distribution curve to annual peak streamflow data. The streamgages included in the study had a minimum of 10 years of data through water year 2015. The report also documents a regional skew analysis, trends in annual peak streamflow data, and development of regional flood-flow regression equations from the relation of flood-flow and basin characteristic data. The resultant regression equations allow for the computation of flood flows at ungaged basins across Pennsylvania by means of select basin characteristics that are significantly correlated with streamflow. Estimates of flood-flow magnitude and frequency for streamgages substantially affected by upstream regulation are also presented. The limitations of the study and uncertainties of the flood-flow estimates are also discussed.

Previous Studies

Regression equations used to predict flood frequency-magnitude relations for ungaged streams in Pennsylvania were first published by Flippo in 1977 and were updated by Flippo in 1982. The equations published in 1982 were evaluated by Ehlke and Reed (1999) based on a comparison between flood flows calculated from the regression equations and peak-flow data collected through the 1996 water year. Regression equations for estimating magnitude of flood flows in Pennsylvania for selected recurrence intervals were subsequently published in 2000 using data through the 1997 water year (Stuckey and Reed, 2000) and most recently in 2008 using annual peak streamflow data generally through water year 2005 (Roland and Stuckey, 2008). Selected flood-flow statistics for streamgage locations in and near Pennsylvania using data collected through 2008 were reported by Stuckey and Roland (2011).

Study Area

The study area for developing updated flood-flow regression equations for Pennsylvania includes the Commonwealth of Pennsylvania (fig. 1) and parts of surrounding states where hydrologic unit code (HUC8) subwatersheds overlap state borders. According to the U.S. Census Bureau (2018), Pennsylvania has a total area of approximately 46,054 square miles (mi^2) with approximately 1,312 mi^2 of that area covered by perennial water. Pennsylvania has three major river basins and three smaller river basins containing more than 98,100 linear

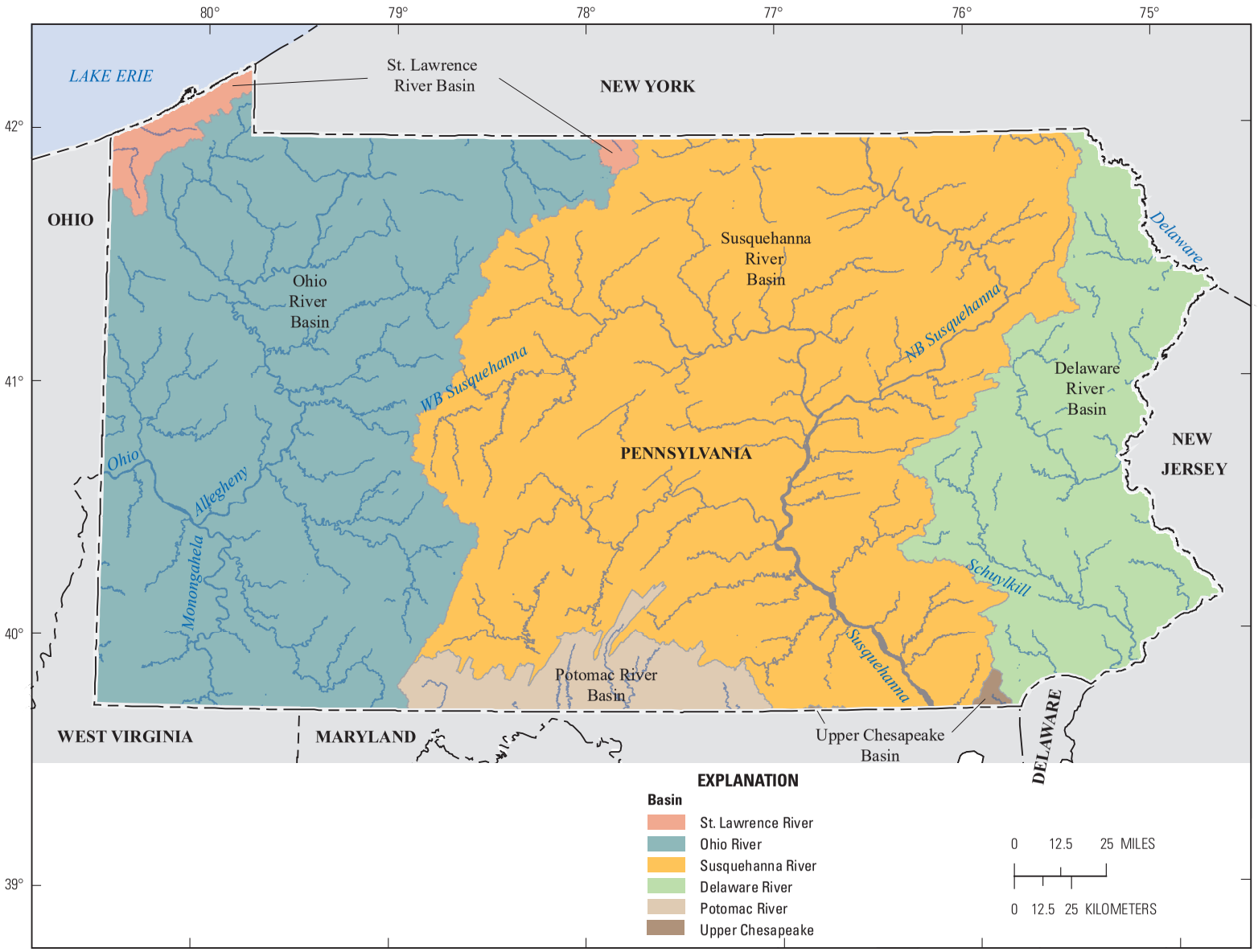


Figure 1. Location of study area and drainage basins in Pennsylvania.

miles of streams in the State (Sloto and others, 2017). The Delaware River forms the boundary between Pennsylvania and New Jersey on the east, and the river flows south into the Delaware Bay. The Susquehanna River Basin is in the central part of the State, and the river flows south from New York into the Chesapeake Bay in Maryland. The Ohio River Basin drains the western part of the State, including the Allegheny and Monongahela Rivers, and the river flows into the Mississippi River and ultimately the Gulf of Mexico (not shown). The three basins with small areas in Pennsylvania include the Potomac River Basin in the southcentral part of the State, the Saint Lawrence River Basin in the northwestern part of the State, and the Chesapeake Bay Basin in the southeastern part of the State that drains directly into the Chesapeake Bay.

There are 11 U.S. Environmental Protection Agency (EPA) Level III ecoregions across Pennsylvania (fig. 2). According to the EPA ecosystems research website (https://www.epa.gov/eco-research/ecoregion-download-files-state-region-3#pane-36), ecoregions denote areas of general similarity in ecosystems and in the type, quality, and quantity of environmental resources; they are designed to serve as a spatial framework for the research, assessment, management, and monitoring of ecosystems and ecosystem components. Ecological regions can be identified through the analysis of the spatial patterns and the composition of biotic and abiotic phenomena that affect or reflect differences in ecosystem quality and integrity (Wiken, 1986; Omernik, 1987, 1995). These phenomena include geology, physiography, vegetation, climate, soils, land use, wildlife, and hydrology. Explanations of the methods used to define the EPA's ecoregions are given in Gallant and others (1989), Griffith and others (1994), Omernik (1995), and Woods and Omernik (1996).

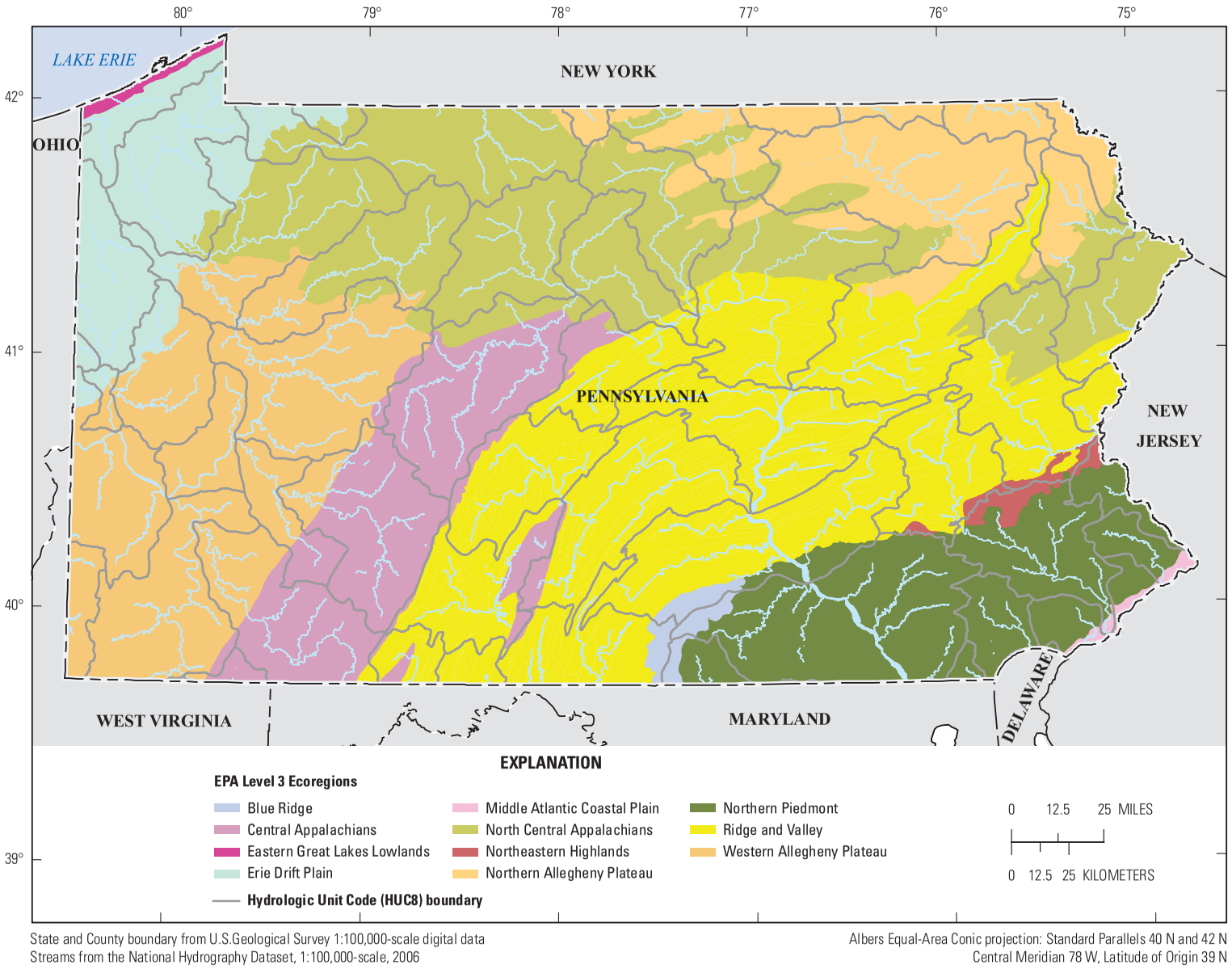
Streamgage Selection and Data Analysis

An initial list of 356 USGS streamgages was compiled to identify those that would be considered for analysis in the development of updated flood-flow regression equations for Pennsylvania. These streamgages were further evaluated based on considerations such as flow regulation, proximity to other streamgages, and period of record (the time when a streamgage is in operation). The streamflow data associated with a streamgage during its period of record is known as systematic data and the maximum instantaneous streamflow value recorded at a streamgage for an entire water year is known as the annual peak flow. The systematic annual peak flow data for all streamgages considered in the analysis were retrieved and carefully reviewed to assure the quality of the records.

Flow Data and Selection of Streamgages

The terms "peak flows" and "flood flows" may often be used interchangeably in common dialogue; however, within this report an attempt is made to maintain consistency regarding usage. The term "peak flow" is associated with an instantaneous peak measured at a streamgage that is associated with the water year in which it occurred. The term "flood flow" refers to a flood frequency magnitude computed for a site that is associated with an AEP or recurrence interval. In the context of this report, flood flows computed at streamgages using annual peak flow data will be referred to as "observed", whereas the flood flows computed from regression equations will be referred to as "predicted".

Initially, 356 active and discontinued streamgages were identified as potential unregulated candidate streamgages to have their respective streamflow and basin characteristic data incorporated into the analysis for the development of updated flood-flow regression equations (appendix 1). This initial selection was based on Pennsylvania streamgages (1) operating for a minimum of 10 years, with (2) drainage areas of less than approximately 1,500 mi^2, and (3) having flow conditions not substantially affected by regulation, diversions, or mining. In the event of substantial regulation, diversion, or mining occurring within a basin during the period of record, only the period of record prior to the event was used in the analysis. Streamgages in neighboring states that have drainage basins partly in or near Pennsylvania and that meet the above criteria were also considered. Annual peak flow data were retrieved for these streamgages from the USGS National Water Information System (NWIS) website (https://waterdata.usgs.gov/nwis) and carefully reviewed for completeness of record, evaluation of historic peaks, and qualification codes (https://nwis.waterdata.usgs.gov/usa/nwis/pmcodes/help?output_formats_help). In addition to the review of annual peak flow data, the streamgages were further evaluated based on proximity to other streamgages and period of record. Streamgages located within the same watershed can produce similar hydrologic information and introduce bias if included in the regression analysis by overemphasizing, or placing greater weight, on a watershed with multiple streamgages. Streamgages on the same stream and within 0.33 to 3 times the size of another streamgage drainage area (Sloto and others, 2017) were considered to potentially have the same or similar hydrologic characteristics and were further evaluated. Generally, when these situations were identified, the streamgage with a longer period of record and (or) more current period of record was retained for inclusion in the regression analysis and the other streamgage was removed from the analysis. Based on these analyses and during the exploratory development of the regression equations, the initial list of 356 streamgages was



State and County boundary from U.S.Geological Survey 1:100,000-scale digital data
Streams from the National Hydrography Dataset, 1:100,000-scale, 2006

Albers Equal-Area Conic projection: Standard Parallels 40 N and 42 N
Central Meridian 78 W, Latitude of Origin 39 N

Figure 2. U.S. Environmental Protection Agency level 3 ecoregions and hydrologic unit code boundaries (HUC8) in Pennsylvania.

reduced to 285 continuous[2]- and partial-record[3] streamgages on streams in Pennsylvania and surrounding states (New York, Ohio, Maryland, and West Virginia) that were used in the development of flood-flow regression equations for Pennsylvania (fig. 3).

Trends in Annual Peak Streamflow Data

The annual maximum peak flows used in the LP3 flood-frequency analysis for the 356 unregulated streamgages considered for inclusion in the development of updated flood-flow regression equations were analyzed to determine whether trends existed in their respective periods of record using a Mann-Kendall statistical test. The LP3 analysis assumes the peak-flow data are a reliable and representative time sample of random homogeneous events (England and others, 2018). The results of the Mann-Kendall test (appendix 1) showed that 46 streamgages, or 13 percent, exhibited a significant trend at the 95-percent confidence interval (fig. 4). Of those with a significant trend, 32 were positive, indicating annual peak flows were increasing over time, and 14 were negative, indicating annual peak flows were decreasing over time. The Kendall's tau mean absolute value for the streamgages with a significant trend was 0.31 (a value of 1 or -1 would indicate a perfect monotonically upward or downward trend, respectively) and the mean number of years of operation of the streamgages was 44 years. Generally, significant positive trends were found in the Delaware River Basin (88 percent of those streamgages with a significant trend were positive) and Susquehanna River Basin (70 percent of those streamgages with a significant trend were positive). There are a variety of factors that may be attributed to these trends, such as the length of time or period of record a streamgage was in operation, variations in precipitation over the streamgage period of record, and increased impervious surface or urbanization within the basin leading to additional stormwater entering receiving waterways instead of recharging groundwater. Owing to the lack of certainty regarding the cause of the significant trends, these streamgages were included in the exploratory regression analysis, and 38 were used in the development of the final regression equations. The topic of nonstationarity, particularly with regards to climate variability and change, is an active and ongoing area of study.

[2]A continuous-record station is a site where stage or streamflow is recorded at some interval on a continuous basis. The recording interval is usually 15 minutes, but may be less or more frequent.

[3]A partial-record station is a site where discrete measurements of one or more hydrologic parameters are obtained over a period of time without continuous data being recorded or computed. A common example is a crest-stage gage partial-record station at which only peak stages and flows are recorded.

Flood Magnitude and Frequency at Streamgages

The magnitude of floods for selected AEPs at streamgaging stations can be estimated on the basis of statistical properties (or moments) associated with its annual peak flow record. The Interagency Advisory Committee on Water Data (1982) recommends fitting a LP3 distribution to the logarithms (base 10) of annual peak flows at a streamgage. This method-of-moments technique is commonly referred to as the Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) method. By determining the mean, standard deviation, and skew of the log-transformed annual peak flow data, the following equation may be used to compute the magnitude of observed flood flow for a desired AEP:

$$\text{Log } Q_p = X + K_p \bullet S \quad (1)$$

where

Q_p is the flood magnitude at a selected exceedance probability p,
X is the mean of the logarithms of the annual peak flows,
K_p is a factor based on the skew coefficient and the selected percent AEP, and
S is the standard deviation of the logarithms of the annual peak flows.

At the time of this study, updates to Bulletin 17B were being documented in Bulletin 17C (England and others, 2018). Although it maintains the moments-based approach of the Bulletin 17B procedures, the method outlined in Bulletin 17C employs the EMA procedure (Cohn and others, 1997), which incorporates a generalized version of the Grubbs-Beck test for the detection of low outliers (Cohn and others, 2013). EMA can accommodate interval peak flow data, which simplifies analysis of datasets containing censored observations, historical data, low outliers, and uncertain data points, while also providing enhanced confidence intervals for the estimated streamflows (Veilleux and others, 2014). The EMA methodology outlined in Bulletin 17C has been incorporated into the USGS peak flow frequency analysis program, PeakFQ version 7.1 (Flynn and others, 2006; U.S. Geological Survey, 2014; Veilleux and others, 2014) and was used to compute observed flood flows for streamgages included in this study. Within the framework of EMA, flow intervals and perception thresholds are defined for each year in a streamgage's annual peak flow record. Flow intervals (defined with a lower and upper bound based on observations, written records, or physical evidence) are used to describe the knowledge of a peak flow value. For most peak flows during the systematic period of record, the default lower and upper bounds of the

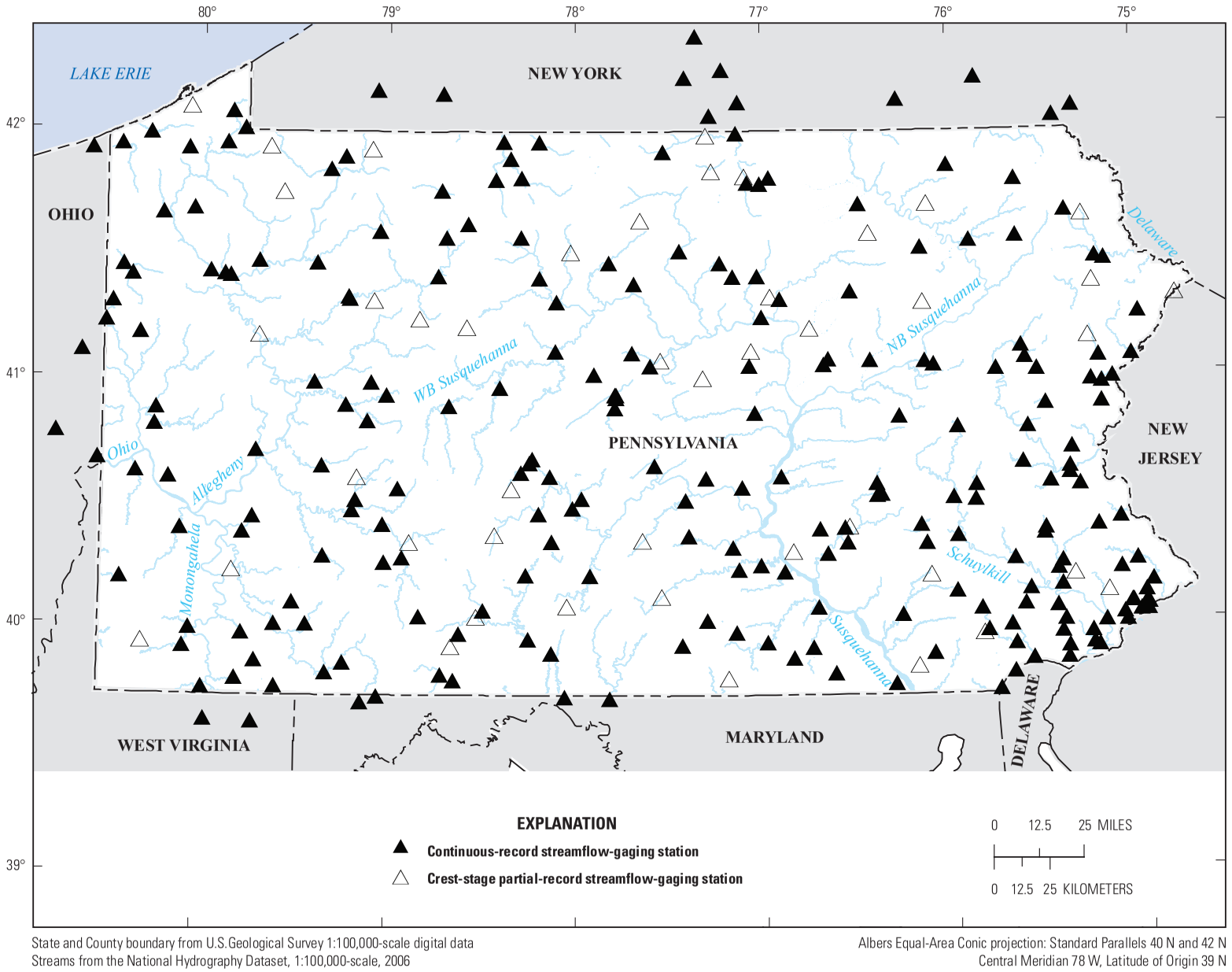


Figure 3. U.S. Geological Survey streamgaging stations used in the development of flood-flow regression equations for Pennsylvania streams.

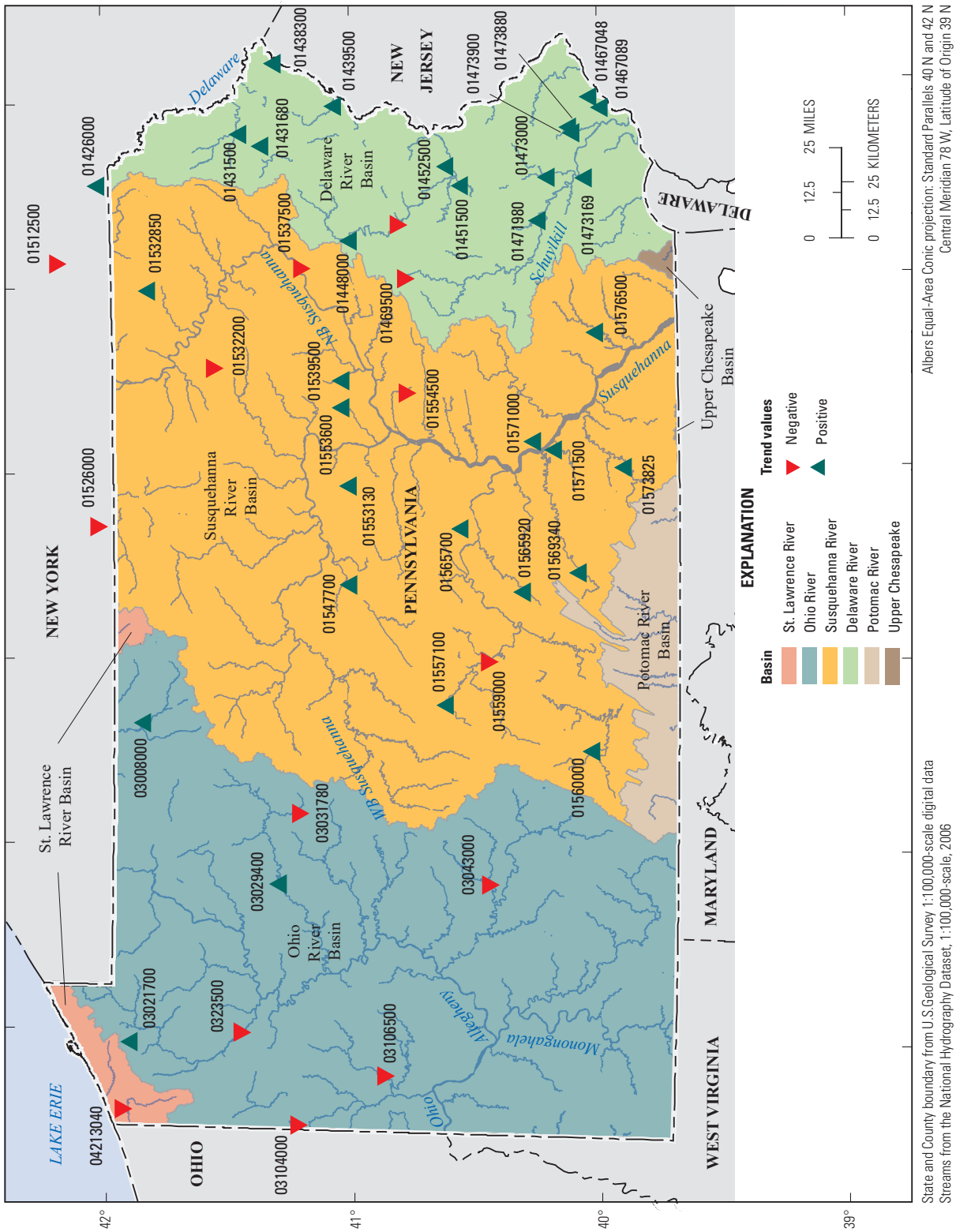


Figure 4. U.S. Geological Survey streamgaging stations with significant trends in annual maximum peak flows over their period of record.

flow interval both equal the observed peak flow, and for most years when no information has been recorded, the default lower and upper bounds are zero and infinity, respectively. If there is uncertainty in a peak flow value, the lower and upper bounds of the flow interval may be set to a range of probable flows. Perception thresholds (lower and upper) identify the range of potentially measurable flood flows whose magnitude would have been measured had they occurred. Generally, for annual peak flows recorded during a streamgage's systematic period of record, the default range of perception thresholds is from zero to infinity. At some streamgages, flows can be determined only when water in the stream reaches a certain minimum measurable level. An example is a crest stage gage, which consists of a mounted vertical steel pipe that houses a measuring stick and granulated cork. Intake holes at the bottom of the pipe, along with a vent hole at the top, allow water to enter the pipe when it reaches a certain level; permitting peak events to be documented when the cork floats and sticks to the measuring stick. It is possible that in some years the water won't reach that minimum level, consequently the lower perception threshold is the flow associated with the minimum measurable water level. For this study, years with measurable discharge were assigned the default perception thresholds of zero to infinity and years when water did not reach the minimum level of measurement were generally assigned a lower perception threshold of either the flow associated with the minimum measurable water level or the lowest flow associated with the systematic annual peak flow record; upper thresholds were set to infinity. In some instances, historic peaks are a documented part of an annual peak flow record, that is, a peak flow associated with a major flood event occurring outside of the streamgage's systematic period of record. For the ungaged years between the historic and systematic peaks, the lower perception threshold is typically set to the minimum flow the analyst thinks would be measured if it had occurred and the upper threshold is set to infinity. The flow interval and perception thresholds that were incorporated into the EMA analysis for each streamgage included in this study are provided in a separate document (https://doi.org/10.5066/P9YHIU6G).

Occasionally, a site would have a documented historic peak gage height (record of flood height occurring outside of the systematic period of record) with no associated peak flow. In these instances, a peak flow was estimated based on a comparison to other peak gage height and associated flow values occurring at the site of interest. If no similar peak gage height values existed for comparison, a nearby gage having similar hydrologic properties and period of record was identified and a simple regression equation was developed using coincident annual peak flow values from the two sites. This equation was then used to estimate a peak flow for the site of interest. The estimated peak flow and recorded gage height were then compared to systematic peak flow and gage height data. If the estimated peak flow was generally in agreement with the systematic data, it was incorporated into the EMA analysis by estimating a flow interval for that year in the historical record and setting perception thresholds accordingly for the ungaged years between the historic and systematic peaks.

In some instances, documented peak flows occurred outside of the systematic period of record, and were categorized as an opportunistic peak flow. Opportunistic peaks were collected based on factors other than the exceedance of a perception threshold and thus were not treated as historic peaks. Furthermore, these flows are not truly random as their sampling properties are unknown. Consequently, opportunistic peaks were not included in the flood-frequency analyses because of the potential to bias the sample.

As indicated in the Bulletin 17C guidelines (England and others, 2018) and Mastin and others (2016), flood frequency analyses commonly focus on larger floods. These floods typically have lower AEP probabilities and are near the upper end of the peak-flow distribution. When evaluating annual peak flows associated with a streamgage, attention is given to flows that are relatively lower, as they may be influential on the upper end of the peak-flow distribution. These flows are referred to as potentially influential low flows (PILFs). The Multiple Grubbs-Beck test (MGBT; Cohn and others, 2013) was an option within the PeakFQ software used to identify PILFs. Censoring of these data typically results in improved estimation of flood flows in the upper end of the frequency distribution. As recommended by Bulletin 17C guidelines, for instances when PILFs were identified by means of the MGBT, a careful analysis of the streamgage peak-flow data was conducted by applying local knowledge of the watershed and hydrologic considerations. This analysis was used to determine whether the use of the MGBT was appropriate for the identification of PILFs. In the rare instances when the MGBT was not used, the single Grubbs-Beck test (Grubbs and Beck, 1972) was used for low-outlier identification (Interagency Advisory Committee on Water Data, 1982).

Regional Skew Analysis

Skew is one of the three moments used to fit a LP3 frequency distribution to the data when estimating the magnitude of a flood flow for a given probability. Skew is sensitive to extreme peak flows in a station's period of record, particularly for streamgages with relatively short records. The skew value associated with a station's systematic peak flow record is referred to as the station skew. The Interagency Advisory Committee on Water Data (1982) states that a better estimate of skew can be obtained by combining the station skew with a regionalized value of skew (which incorporates information from surrounding streamgages) to determine a weighted skew. The purpose of the regionalized skew is to adjust the at-site station skew to better reflect regional and long-term conditions. In the past, the regionalized skew value could be obtained from a national map associated with the Bulletin 17B publication (International Advisory Committee on Water Data, 1982) that had an associated mean squared error (MSE) of 0.302. It is important to note that a national study to update

the regionalized skew for Pennsylvania and surrounding states was underway at the same time as this study; however; the timing of the national study did not correspond with the need for an updated regionalized skew for Pennsylvania. Because of this, a separate regional skew analysis for Pennsylvania was conducted as part of this peak flow regression update, which resulted in an MSE of 0.181.

A total of 356 unregulated streamgages were initially considered for inclusion in the regional skew analysis. Most of these were Pennsylvania gaging stations, however out-of-state streamgages from Maryland, New York, Ohio, and West Virginia that were included in previous regression reports were also considered. The initial list of streamgages was subsequently reduced to 123 when 233 streamgages did not meet the following criteria: (1) a minimum period of record of 35 years, (2) drainage basins having less than 10 percent urban land cover, and (3) drainage basins having less than 10 percent flow regulation. These criteria were based partly on regional skew studies associated with recent flood magnitude and frequency reports (Mastin and others, 2016; Curran and others, 2016) as well as the need to have representative streamgage coverage across the study area. The period of record is longer than the 10 years required for the computation of a flood-frequency estimate because the skew coefficient of the station skew is sensitive to extreme events and more accurate estimates can be obtained from streamgages with longer periods of record (Mastin and others, 2016). From the subset of 123 streamgages, 8 additional streamgages were removed from the regional skew analysis because they were located on the same stream as another streamgage and within 0.33 to 3 times the drainage area; this was done to avoid redundancy of hydrologic information. As a result, 115 streamgages were used in the regional skew analysis for Pennsylvania.

A comparison of drainage areas was done to determine whether the subset of 115 streamgages selected for the skew analysis was representative of the larger dataset of 356 streamgages that could potentially be included in the regression analysis. Overall, the range in drainage areas found in the larger dataset of streamgages was generally represented in the subset of streamgages, although there were more streamgages with small drainage areas in the larger dataset. The median of the larger set of streamgages is 61.1 mi^2, whereas the median for the subset used in the skew analysis is 153 mi^2. Additionally, the lower bound in the range of drainage areas for the subset of streamgages was slightly higher than that of the larger dataset of streamgages. The 115 streamgages selected for the skew analysis were also evaluated for adequate spatial distribution. With the possible exception of the south-central and southwestern parts of Pennsylvania where streamgage coverage is sparse, the subset of streamgages provided adequate spatial representation.

Magnitude and frequency of flood flows were computed for the 115 streamgages selected for skew analysis using the EMA methodology with annual peak flow data through water year 2015 and utilizing the station skew option. The relation between station skew and select basin characteristics (drainage area, mean elevation, precipitation, percent carbonate, percent urban, percent forest, latitude and longitude) found to be important in previous studies when estimating flood frequencies for Pennsylvania or surrounding states was evaluated using an ordinary least squares regression analysis. This was done to determine whether the basin characteristics could be used as explanatory variables to describe the variation of at-site station skews across Pennsylvania; however, none of the basin characteristics were significant at the 95-percent confidence interval. Therefore, a constant statistical model utilizing the mean value of all at-site station skews, was used to determine a regional skew value for Pennsylvania and the associated MSE. Computed residuals (the difference between the observed and computed values) for each of the 115 sites were plotted spatially using a geographic information system (GIS) and visually inspected for bias. From this plot, a grouping of eight gages in New York, in the upper headwaters of the Susquehanna River basin, exhibited high residuals compared to the rest of the study area. Considering no other patterns in residuals were observed throughout the rest of the study area, and that Pennsylvania was the focus for the regional skew analysis, these eight New York sites were removed from the subset of streamgages for skew analysis. The constant statistical model was then recomputed with the remaining 107 gages, resulting in a regional skew of 0.350 and MSE of 0.181. The station skews were plotted spatially, and no consistent pattern or bias was observed (fig. 5). The systematic period of record ranged from 36 to 113 years with a mean of 70 years; the historical period of record (which includes historic peaks outside of the systematic period of record) for the 107 gages, ranged from 36 to 128 years with a mean of 76 years. As a comparison, the MSE reported in Bulletin 17B for plate 1, which is a map of generalized skews across the United States, is 0.302 with a corresponding record length of 17 years (Griffis and Stedinger, 2007a). The values from the constant statistical model for skew were subsequently used in the EMA flood frequency computations for the unregulated streamgages associated with this publication.

It should be noted that upon completion of the regression equation analysis, it was discovered that the annual peak flows for 01603500 Evitts Creek near Centerville, PA were incorrect in the NWIS-retrieved peak flow file. Thus, incorrect flood frequencies were computed for 01603500 and incorporated into the regional skew and regression analyses. To determine the impact the incorrect peak flow file for streamgage 01603500 had on the regional skew analysis, the corrected flood frequency results for 01603500 were incorporated into a revised skew analysis. The results of this analysis lowered the regional skew slightly from 0.350 to 0.343 and the MSE from 0.181 to 0.178. Based on this information, a check was performed to determine the impact the results of the revised skew analysis (incorporating the corrected flood frequency results for 01603500) had on a subset of streamgages that were included in the development of the flood-flow regression equations. Flood frequency estimates were computed for 119 streamgages (using the weighted skew option) with

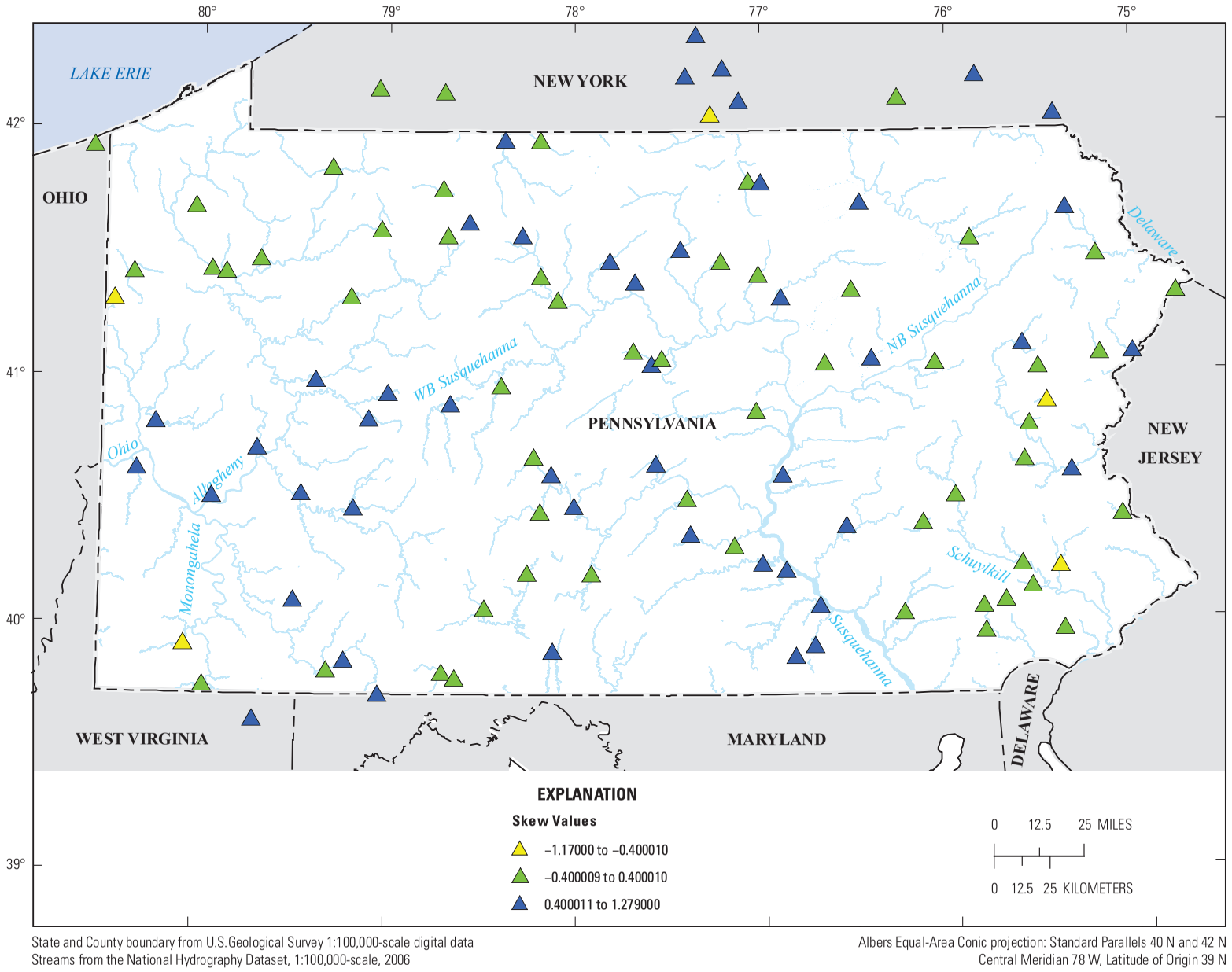


Figure 5. U.S. Geological Survey streamgaging stations used in development of regional skew for Pennsylvania.

a regional skew value of 0.350 (MSE of 0.181) and with a regional skew value of 0.343 (MSE of 0.178). A subsequent comparison of the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent AEPs indicated there was less than 1 percent difference in flood-flow estimates for all 119 streamgages. Because the difference in skew value and the impact to flood-frequency estimates was minimal, and the regression analysis had already been completed at the time the error was discovered, this change was not incorporated into the updated regression analysis. However, the information reported in appendixes 1 and 2 associated with streamgage 01603500 reflects the correct peak flow file.

Observed Flood-flow Computations

The observed flood-flow computations utilized annual peak flow data through the 2015 water year and incorporated the skew from the Pennsylvania regional skew analysis. The resultant observed flood flows computed by means of the EMA methodology were subsequently used in the development of the regression equations for selected AEPs. The observed flood flows, with their 95-percent confidence limits, are provided in appendix 2. Flood-frequency estimates for streamgages located outside of Pennsylvania are reported for informational purposes and are not intended to supersede state-specific sources of flood-frequency information.

When comparing updated flood flows to those previously published, in addition to the updated regional skew analysis, it is important to keep in mind the period of record of annual peak flows being used for the flood-frequency computation. That is, flood flows computed and used in the previous development of regression equations in Pennsylvania (Roland and Stuckey, 2008) used annual peak flow data generally through water year 2005; whereas, the flood flows computed for this study used annual peak flow data through water year 2015, if available. Incorporating an additional 10 years of annual peak flow data into the flood-frequency computations can have a substantial impact on the magnitude of flood flows, particularly for streamgages with shorter periods of record. Additionally, the methodology used in the computation of previously published flood flows was based on Bulletin 17B; whereas, the EMA methodology presented in Bulletin 17C was used to compute flood flows associated with this study. As previously discussed, the EMA incorporates the use of perception thresholds and flow intervals to help define flow values for each year in the streamgage's period of record. If there was uncertainty or limited information regarding a peak flow value, flow intervals and perception thresholds were set accordingly based on available data. Additionally, EMA incorporates a generalized MGBT for identifying potentially low outliers, which, with few exceptions, was selected for flood-frequency computations associated with this study.

Flood flow frequencies for regulated streamgages were computed using the EMA as described above, with the exception of using station skews rather than weighted skews. The skewness of annual peaks at regulated streamgages can differ substantially from the skewness at unregulated streamgages, and the regional skew used to determine the weighted skew would not be representative of those observed at regulated streamgages. The LP3 curve was evaluated closely for each regulated streamgage for fit to the observed peaks. Because many flood-control reservoirs are designed to control the 1-percent AEP and the runoff associated with the event, flood frequencies at regulated streamgages were not computed for AEPs greater in magnitude than the 1-percent AEP.

Basin and Climate Characteristics

Multiple basin and climate characteristics thought to affect streamflow were compiled for streamgages being considered for the development of updated regional regression equations. These characteristics were used as potential explanatory variables as they related to the observed flood-flow magnitudes (based on annual peak streamflow data at streamgages) to develop regression equations for estimating predicted flood flows. These variables, broadly characterized by land use, land cover, geology, terrain, and climate, were compiled from various GIS sources. The use of GIS-derived basin and climate characteristics in the development of regression equations results in improved consistency and reproducibility by utilizing a defined georeferenced drainage basin boundary for each streamgage included in the analysis. These GIS-derived drainage basins are then used with various basin and climate variables to extract basin-specific datasets, which are readable by the regression software programs. The basin and climate characteristics evaluated in the exploratory regression analysis are presented in table 2.

Table 2. Basin and climate characteristics selected for use as potential explanatory variables in the development of regression equations for flood-flow estimates in Pennsylvania.

[DEM, digital elevation model; NHD, National Hydrography Dataset; NLCD, National Land Cover Dataset; PRISM, Parameter-elevation Regressions on Independent Slopes Model]

Variable	Source	Reference	Unit	Description
10–85 Slope	Digital elevation model	U.S. Geological Survey (2000a)	Feet per mile	The 10–85 slope measure is calculated by first determining the length of the main channel between points at 10 and 85 percent of the length from the downstream end. The slope is then determined by subtracting the elevation at the 10-percent point from the elevation at the 85-percent point and dividing by the distance between the points.
24-hour maximum precipitation	National Oceanic and Atmospheric Administration - National Weather Service	Bonnin and others (2006)	Inches	Area average, 2-, 5-, 10-, 25-, 50-, 100-, 200-, and 500-year maximum precipitation (area average)
Annual temperature	Parameter-elevation Regressions on Independent Slopes Model 1981–2010	PRISM Climate Group (2017)	Degrees Fahrenheit	Mean, maximum, minimum
Barren	National Land Cover Datase 2011	Homer and others (2015)	Percent	Percent of basin area
Basin elevation	Digital elevation model	U.S. Geological Survey (2000a)	Feet	Mean, minimum, and maximum
Basin relief	Digital elevation model	U.S. Geological Survey (2000a)	Feet	Maximum basin elevation minus minimum basin elevation
Carbonate bedrock	U.S. Geological Survey Pennsylvania StreamStats application version 1 carbonate grid layer (post-2013)	Kolb and others (2020)	Percent	Percent of basin area
Coniferous	National Land Cover Dataset 2011	Homer and others (2015)	Percent	Percent of basin area
Cultivated	National Land Cover Dataset 2011	Homer and others (2015)	Percent	Percent of basin area
Developed	National Land Cover Dataset 2011	Homer and others (2015)	Percent	Percent of basin area
Drainage area	U.S. Geological Survey StreamStats application using the watershed delineation tool	U.S. Geological Survey (2017)	Square miles	Total upstream area of the streamgage that drains to a point on a stream
Elevation at centroid	Digital elevation model	U.S. Geological Survey (2000a)	Feet	Elevation of DEM at National Water Information System latitude/longitude
Elevation at site	Digital elevation model	U.S. Geological Survey (2000a)	Feet	Elevation of DEM at National Water Information System latitude/longitude

Table 2. Basin and climate characteristics selected for use as potential explanatory variables in the development of regression equations for flood-flow estimates in Pennsylvania.—Continued

[DEM, digital elevation model; NHD, National Hydrography Dataset; NLCD, National Land Cover Dataset; PRISM, Parameter-elevation Regressions on Independent Slopes Model]

Variable	Source	Reference	Unit	Description
Forested	National Land Cover Dataset 2011	Homer and others (2015)	Percent	Percent of basin area
Glaciated	U.S. Geological Survey - from modified geology maps	Soller and Packard (1998)	Percent	Percent of basin area
Ground-water head	Digital elevation model	U.S. Geological Survey (2000a)	Feet	Mean basin elevation minus minimum basin elevation
Impervious surface area	National Land Cover Dataset 2011	Xian and others (2011)	Percent	Impervious percent dataset
Lakes and open water	National Land Cover Dataset 2011	Homer and others (2015)	Percent	Percent of basin area; lakes greater than or equal to 0.2 square miles
Mean annual precipitation (1981–2010)	Parameter-elevation Regressions on Independent Slopes Model 1981–2010	PRISM Climate Group (2017)	Inches	Weighted-area average
Mean basin slope	Digital elevation model	U.S. Geological Survey (2000a)	Degrees	Calculated using an Environmental Systems Research Institute accelerated atan() function
Mean drain (dry) percent poorly drained	Soil Survey Geographic dataset	Soil Survey Staff, Natural Resources Conservation Service	Percent	Weighted-area average[1]
Mean drain (dry) score	Soil Survey Geographic dataset	Soil Survey Staff, Natural Resources Conservation Service	Score	Weighted-area average[2]
Mean drain (wet) percent poorly drained	Soil Survey Geographic dataset	Soil Survey Staff, Natural Resources Conservation Service	Percent	Weighted-area average[1]
Mean drain (wet) score	Soil Survey Geographic dataset	Soil Survey Staff, Natural Resources Conservation Service	Score	Weighted-area average[2]
Mean hydrologic group	Soil Survey Geographic dataset	Soil Survey Staff, Natural Resources Conservation Service	Score	Hydgrp is a code A,B,C,D,A/D,B/D,C/D which was converted into a 1–4 score[3] to be averaged by watershed. This was handled using the method outlined by Schwarz and Alexander (1995).
Mean population density	U.S. Census Bureau (2010)	Falcone (2016)	Population/square mile	Population/area calculation based on weighted-area average

Table 2. Basin and climate characteristics selected for use as potential explanatory variables in the development of regression equations for flood-flow estimates in Pennsylvania.—Continued

[DEM, digital elevation model; NHD, National Hydrography Dataset; NLCD, National Land Cover Dataset; PRISM, Parameter-elevation Regressions on Independent Slopes Model]

Variable	Source	Reference	Unit	Description
Percentage of drainage basin greater than or equal to 1,200 feet	Digital elevation model	U.S. Geological Survey (2000a)	Percentage	Percentages greater than or equal to 1,200 feet
Seasonal mean precipitation (1981–2010)	Parameter-elevation Regressions on Independent Slopes Model 1981–2010	PRISM Climate Group (2017)	Inches	Fall (September, October, November), Winter (December, January, February), Spring (March, April, May), Summer (June, July, August)
Seasonal temperature	Parameter-elevation Regressions on Independent Slopes Model 1981–2010	PRISM Climate Group (2017)	Degrees Fahrenheit	Fall (September, October, November), Winter (December, January, February), Spring (March, April, May), Summer (June, July, August); Mean, Maximum, Minimum
Shape factor	Digital elevation model	U.S. Geological Survey (2000a)	Unitless	Measure of the shape of a basin; computed as the ratio of the square of the main channel stream length or longest drainage path (in square miles) to its computed drainage area (in square miles)
Storage	National Land Cover Dataset 2011; National Hydrography Dataset, 1:24,000 scale	Homer and others (2015); U.S. Geological Survey (2000b)	Area in square miles	NHD (LakePond) + NLCD (Wetlands: 90 + 95)
Stream density	National Hydrography Dataset, 1:24,000 scale	U.S. Geological Survey (2000b)	Miles/square miles	Computed by dividing the total length of all streams in a basin by the drainage area
Urbanized area	National Land Cover Dataset 2011	Homer and others (2015)	Percent	Percent of basin area
Wetlands	National Land Cover Dataset 2011	Homer and others (2015)	Percent	Percent of basin area

[1]Drainage classes were determined under different moisture conditions (dry and wet) and assigned numeric values (0 or 1) based on their classification: excessively drained=0, somewhat excessively drained=0, well drained=0, moderately well drained=0, somewhat poorly drained=1, poorly drained=1, very poorly drained=1. Based on these values, a weighted-area average percent was computed over a basin providing an estimate of percentage of poorly drained soils.

[2]Drainage classes were determined under different moisture conditions (dry and wet) and assigned numeric values (1-7) based on their classification: excessively drained=1, somewhat excessively drained=2, well drained=3, moderately well drained=4, somewhat poorly drained=5, poorly drained=6, very poorly drained=7. Based on these values, a weighted-area average (score) was computed over a basin.

[3]The character codes defined in the STATSGO component file were converted into numeric codes according to Bill Battaglin's methods (C. Price, written commun., April 17, 2017). The coding transformations were A=1 (high infiltration, deep soils, well drained to excessively drained sands and gravels), B=2 (moderate infiltration rates, deep and moderately deep, moderately well and well drained soils with moderately coarse textures), C=3 (slow infiltration rates, soils with layers impeding downward movement of water, or soils with moderately fine or fine textures), D=4 (very slow infiltration rates, soils are clayey, have a high water table, or are shallow to an impervious layer). Mixture codes A/D, B/D, and C/D were assigned the value 4. Miscellaneous areas labeled as Dumps and Gullied Land were assigned the value 2.5, if the hydgrp value was missing. Areas denoted as Pits, Rock Outcrops, Terrace Escarpments, and Urban Land with missing hydgrp codes were assigned a value of 4.

Development of Regression Equations

Regression equations for the estimation of flood flows at ungaged streams in Pennsylvania were developed using observed flood magnitude and frequency estimates computed for select streamgages across the State as well as in neighboring states. Basin and climate characteristics were compiled for the drainage basins upstream from the streamgages, and regression techniques were utilized to relate the updated flood magnitude and frequency estimates (the dependent or response variable) to the basin and climate characteristics (the independent or explanatory variables) to produce regression equations. These equations can be used to compute flood flows by means of applicable basin and climate characteristics for selected AEPs for streams where no gaging-station data are available.

Regression Analysis and Regionalization

The observed flood frequencies at 294 streamgages were related to basin and climate characteristics using exploratory weighted least squares (WLS) and generalized least squares (GLS) regression techniques. To form a near-linear relation between the flood flows and basin characteristics, all independent and dependent variables were log-transformed prior to regression analysis. Because some basin characteristics are represented by percentages (for example, percent carbonate), they can have a value of zero, which would result in an error in the regression model output. To prevent this potential error from occurring, a value of 0.1 was added to all percentages of the basin characteristic before the log transformation. The flood frequencies were weighted using the following expression for the exploratory WLS regression techniques:

$$Q_{WLS_i} = Q_i \bullet \frac{\left(Y_i \bullet S\right)}{Y_T} \quad (2)$$

where

Q_{WLS_i} is the flood magnitude at a selected exceedance probability for streamgage *i* used in the exploratory WLS regression analysis weighted by its respective number of years of record of annual peak flow;

Q_i is the flood magnitude at a selected exceedance probability for streamgage *i* used in the exploratory regression analysis not weighted by its respective number of years of record of annual peak flow;

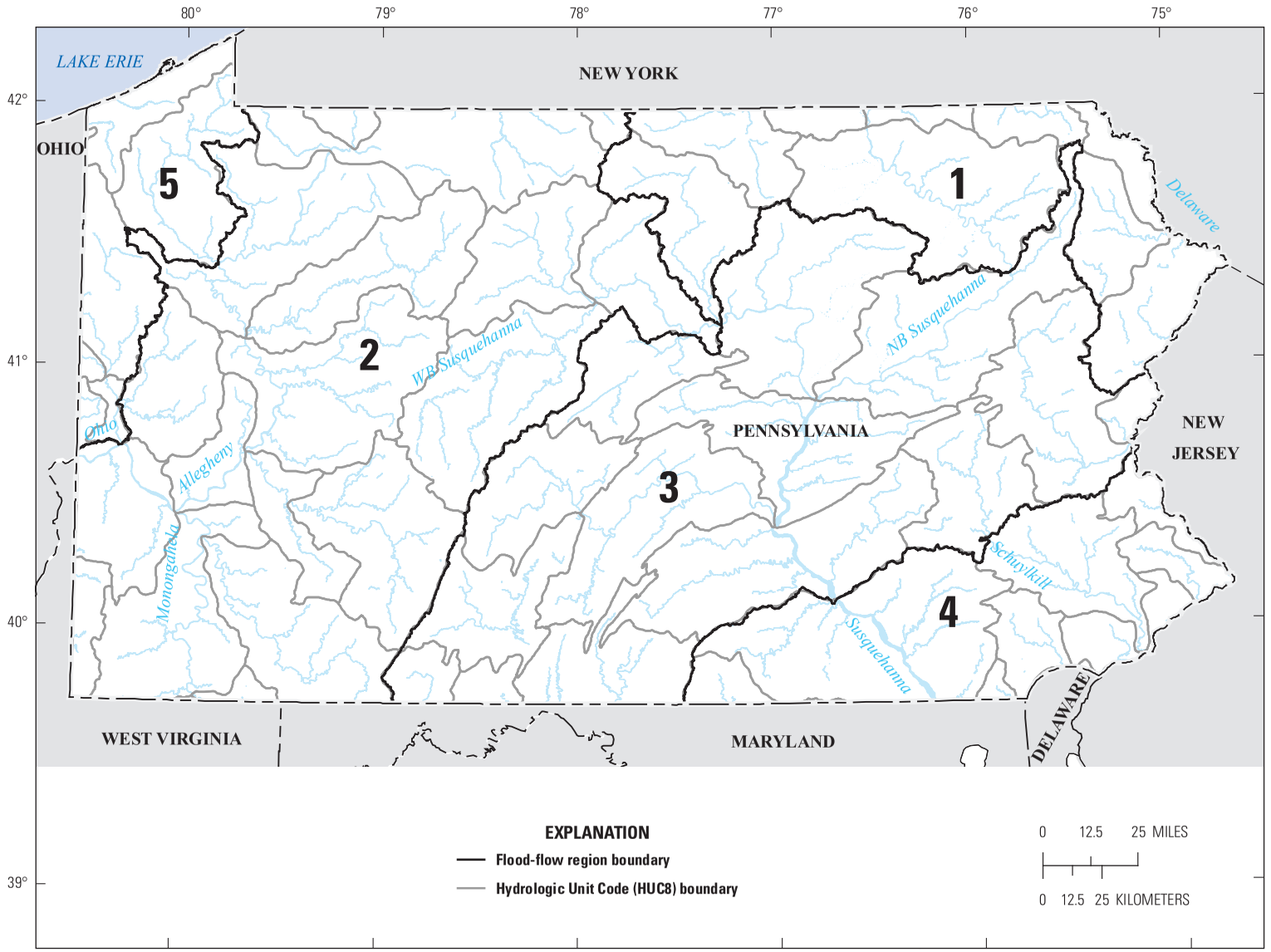
Y_i is the number of years of record used in the computation of flood magnitude at a selected exceedance probability at streamgage *i*;

S is the total number of streamgages considered in the exploratory analysis; and

Y_T is the total number of years of record used in the computation of flood magnitudes at a selected exceedance probability for all streamgages considered in the exploratory analysis.

Exploratory WLS regression iterations were performed using the statistical software package Spotfire S+ (TIBCO Software Inc., 2008) and final GLS regressions were performed using the USGS software package Weighted-Multiple-Linear Regression Program (WREG; Eng and others, 2009) version 2.02 (Farmer, 2017). Default values for the correlation-smoothing function as defined within WREG version 2.02 for each region (alpha, α, equal to 0.002 and theta, θ, equal to 0.98) were used with the exception of regions 4 and 5. Region 4 used a θ value equal to 0.97, and region 5 used an α equal to 0.004 and θ equal to 0.97. Based on visual interpretation, these adjustments to α and θ resulted in a better fit of the curve to the data points. Regression diagnostics used to evaluate potential regression equations during exploratory analysis included graphical relations, multicollinearity (correlation of coefficients and variance inflation factor), predicted residual error sum of squares (PRESS) statistic, standard error of prediction, coefficient of determination (R^2), residuals, and Cook's distance (Cook's D) (Helsel and Hirsch, 2002).

Exploratory analysis was done using WLS regression techniques and a preliminary statewide regression equation was developed. Residuals (the difference between observed and predicted flood frequencies) were plotted against HUCs, ecoregions, 1-percent AEP yields, and previously published flood-flow regions defined for Pennsylvania. From this analysis it was apparent the State needed to be regionalized, and five new flood-flow regions were developed based on ecoregions and HUC8 boundaries. HUC8 boundaries were followed wherever possible to avoid dividing flood-flow regions. However, it was necessary to divide several HUC8 polygons into different flood-flow regions where hydrologic and physiographic properties differed. In those instances, care was taken to ensure that basins with drainage areas within the range that were used to develop the regression equations were not split, so predicted flood flows would remain consistent. Exploratory regressions developed using these new regions were compared to exploratory regressions developed using previously published regions as well as statewide regressions using the PRESS statistic, R^2, standard error of prediction, and residual plots. The newly developed exploratory regressions based on the EPA Level III ecoregions performed better in all comparisons; as a result, the five new flood-flow regions (fig. 6) were used to develop new regression equations in WREG using GLS regression techniques. All explanatory variables were significant at the 95-percent confidence level for at least one of the AEPs. Residuals from the newly defined regions were plotted and evaluated for any bias within each region; no pattern or bias was observed within the regions. Each set of regression equations for a region was constrained to contain the same explanatory variables to ensure the predicted flood flows uniformly increased as the AEP decreased. Nine streamgages were removed from the initial list of 294 streamgages during exploratory analysis because of questionable data,



State and County boundary from U.S.Geological Survey 1:100,000-scale digital data
Streams from the National Hydrography Dataset, 1:100,000-scale, 2006

Albers Equal-Area Conic projection: Standard Parallels 40 N and 42 N
Central Meridian 78 W, Latitude of Origin 39 N

Figure 6. Flood-flow regions and hydrologic unit code boundaries (HUC8) in Pennsylvania.

abnormal basin characteristic values not representative of the region, or poorly defined regulation, resulting in a total of 285 streamgages used in the final regression analysis.

Flood-flow region 1 is in the northeastern part of the State in the glaciated area of the Allegheny Plateau ecoregion (fig. 7) and contains the upper reaches of the Delaware and Susquehanna River Basins. The rest of the Allegheny Plateau ecoregion, as well as the Appalachian ecoregions, is generally in flood-flow region 2, which contains much of the Ohio and West Branch Susquehanna River Basins (fig. 8). Flood-flow region 3 is in the central part of the State, making up the Ridge and Valley ecoregion (fig. 9). Flood-flow region 4 is in the southeastern part of the State in the Northern Piedmont ecoregion and contains the lower reaches of the Delaware and Susquehanna River Basins (fig. 10). The northwestern part of the State in the Erie Drift Plain ecoregion makes up flood-flow region 5 (fig. 11).

The resultant basin-characteristic variables and coefficients, along with model diagnostics, are shown in table 3. The regression model took the following form in log units:

$$\log Q_{reg_i} = A + [b \bullet \log(DA)] + [c \bullet \log(ME)] + [d \bullet \log (0.1+ C)] + [e \bullet \log(0.1 + St)] + [f \bullet \log(Sl)] \quad (3)$$

Or, in arithmetic space

$$Q_{reg_i} = 10^A (DA)^b (ME)^c (0.1 + C)^d (0.1 + St)^e (Sl)^f \quad (4)$$

where

log	is	log to base 10;
Q_{reg_i}	is	the regional regression predicted flood-flow estimate for a given AEP at site *i*, in cubic feet per second;
A	is	the intercept (estimated by GLS);
DA	is	drainage area of the basin, in square miles;
ME	is	maximum basin elevation, in feet;
C	is	basin underlain by carbonate bedrock, in percent;
St	is	storage in the basin, in percent;
Sl	is	mean basin slope, in degrees; and
b, *c*, *d*, *e*, and *f*	are	explanatory variable coefficients of regression estimated by GLS

The flood-flow regional regression equations were evaluated for fit, sensitivity, overall bias, and monotonic relation between frequency and magnitude. Residuals for the 1-percent AEP were plotted spatially and the Nash-Sutcliffe efficiency (NSE) value was computed between the two datasets. Particular attention was paid to region 5 because of the low number of streamgages used to develop the regression equations. No major issues were found with any of the equations. The mean NSE value computed for the regional 1-percent AEP regression equations was 0.90; NSE values ranged from 0.85 for region 4 to 0.93 for regions 1 and 2.

Example Using a Flood-flow Regression Equation.

Example 1. Calculate the 1-percent AEP flood flow for USGS streamgage 01516350, Tioga River near Mansfield, Pennsylvania, which is located in the northeastern part of Pennsylvania at latitude 41°47'49" and longitude 77°04'50". The drainage area is 153 mi^2 and the maximum basin elevation is 2,446 feet. The basin is unaffected by substantial regulation, diversion, or mining.

1. From figure 7 and the latitude and longitude, the site is in flood-flow region 1.

2. Using coefficients from table 3, the 1-percent AEP ($Q_{AEP,\ 1\text{-percent}}$) regional regression equation is

   $\log Q_{AEP,\ 1\text{-percent}} = -5.721 + [(0.6543) \bullet \log (DA)] + [(2.5552) \bullet \log (ME)]$

3. Substituting the basin characteristics for the site into the equation produces

   $\log Q_{AEP,\ 1\text{-percent}} = -5.721 + [(0.6543) \bullet \log (153)] + [(2.5552) \bullet \log (2{,}446)]$

   $\log Q_{AEP,\ 1\text{-percent}} = -5.721 + 1.4294 + 8.6582$

   $\log Q_{AEP,\ 1\text{-percent}} = 4.3666$

   $Q_{AEP,\ 1\text{-percent}} = 23{,}300$ ft^3/s (rounded to three significant figures)

Comparison to Previous Flood-flow Regression Equations

A comparison was done between results from the current (2019) flood-flow regression equations and the previous regressions equations (Roland and Stuckey, 2008) with respect to percent differences between predicted flood-flow magnitudes associated with the 50-, 20-, 10-, 2-, 1-, and 0.2-percent AEPs for streamgages included in both analyses. The comparison excluded the 4- and 0.5-percent AEPs, as the previous regression equations did not include these flood-flow probabilities. Statewide, the mean and median differences between the current and past regression equations ranged from -2 percent (for both the mean and median) for the 50-percent AEP to 2 percent and 5 percent, respectively, for the 0.2-percent AEP (table 4). The median percent differences for the 1-percent AEP are shown in figure 12. Overall, the largest percent differences are found in flood-flow region 5. Within flood-flow region 2, a grouping of negative percent differences is found in the southwestern part of the State that correspond to the previous flood-flow region 4, which indicates the current regression equations provide flood-flow estimates of the 1-percent AEP that are uniformly lower than the previous regression

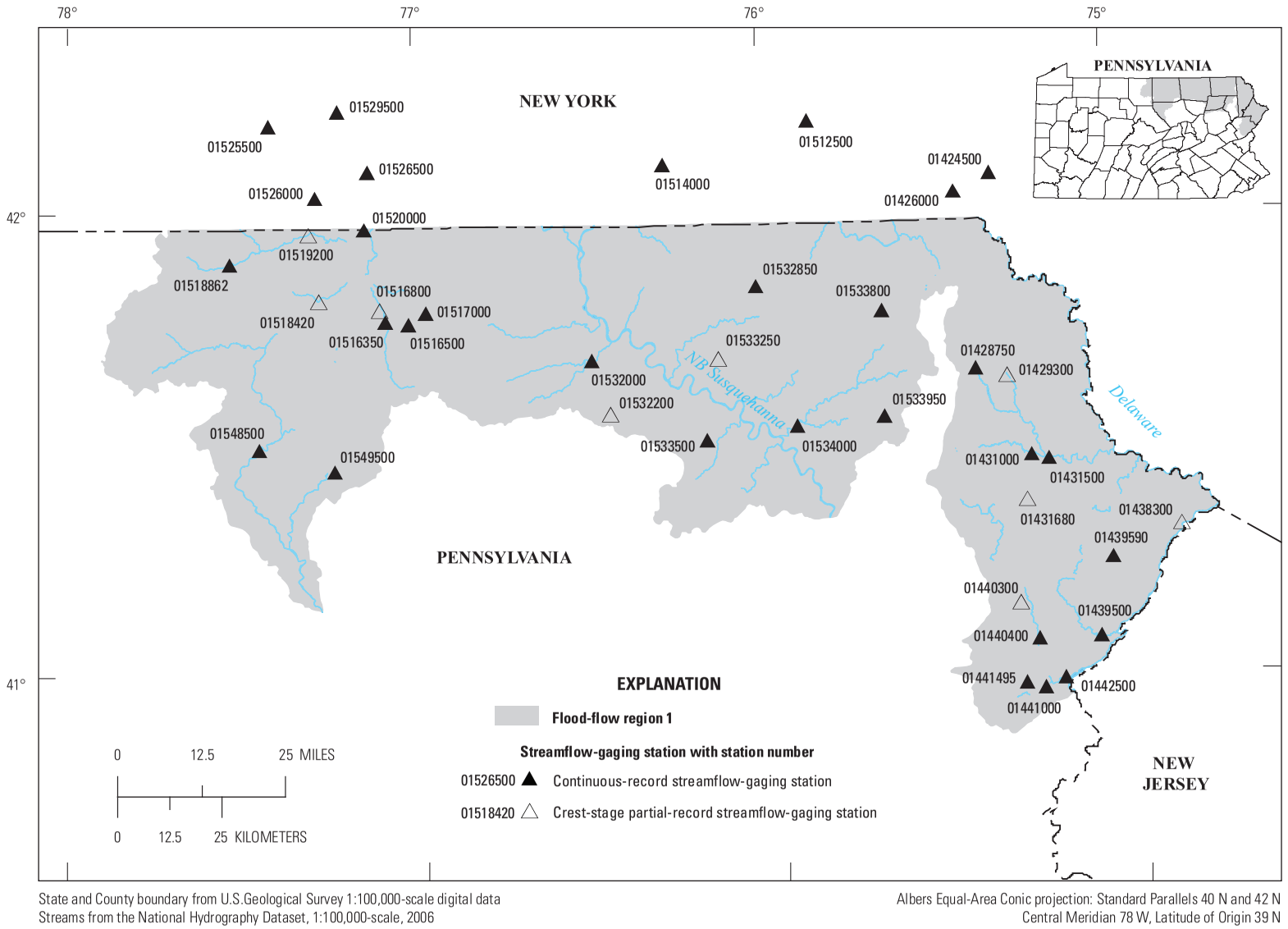


Figure 7. Flood-flow region 1 in Pennsylvania.

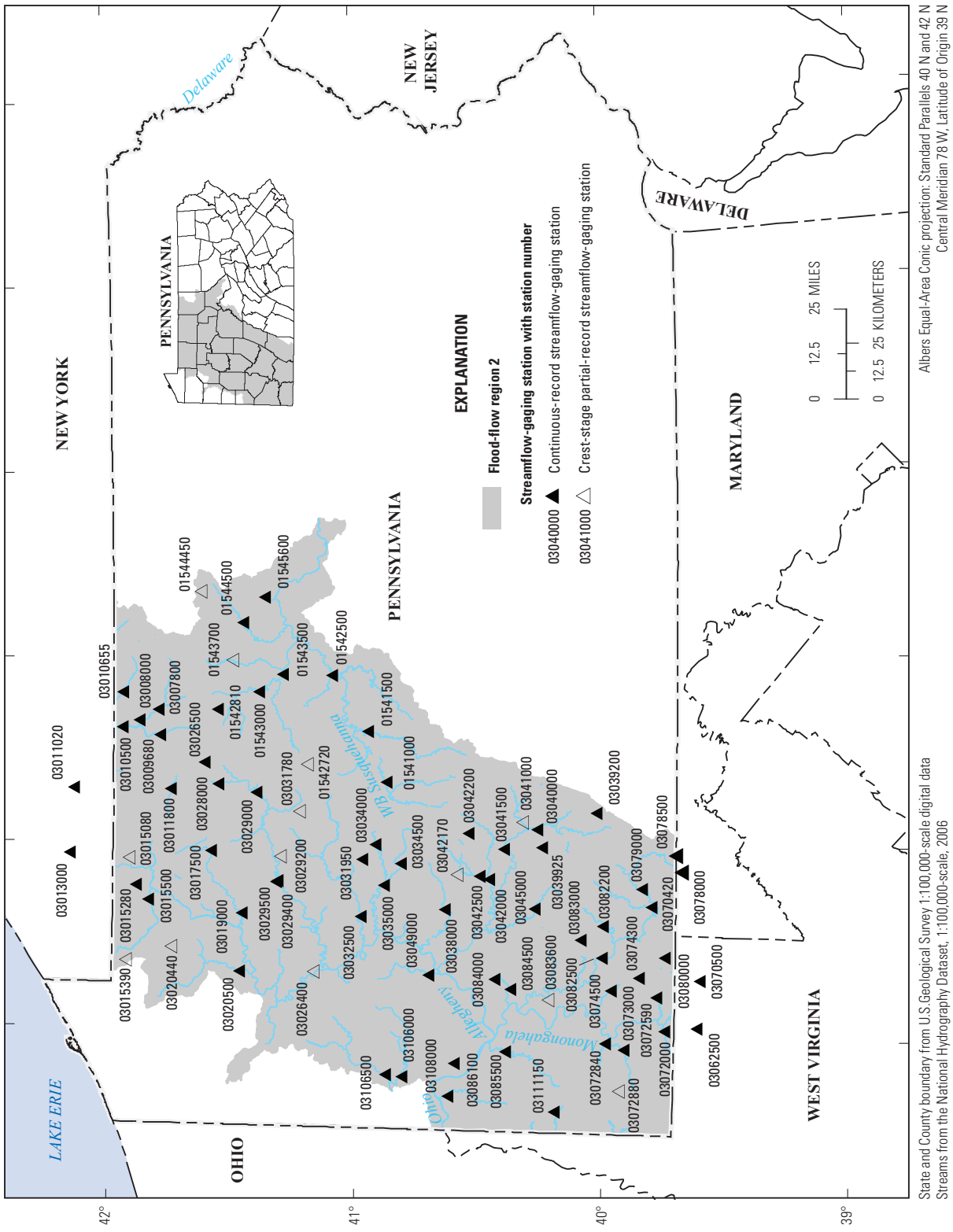
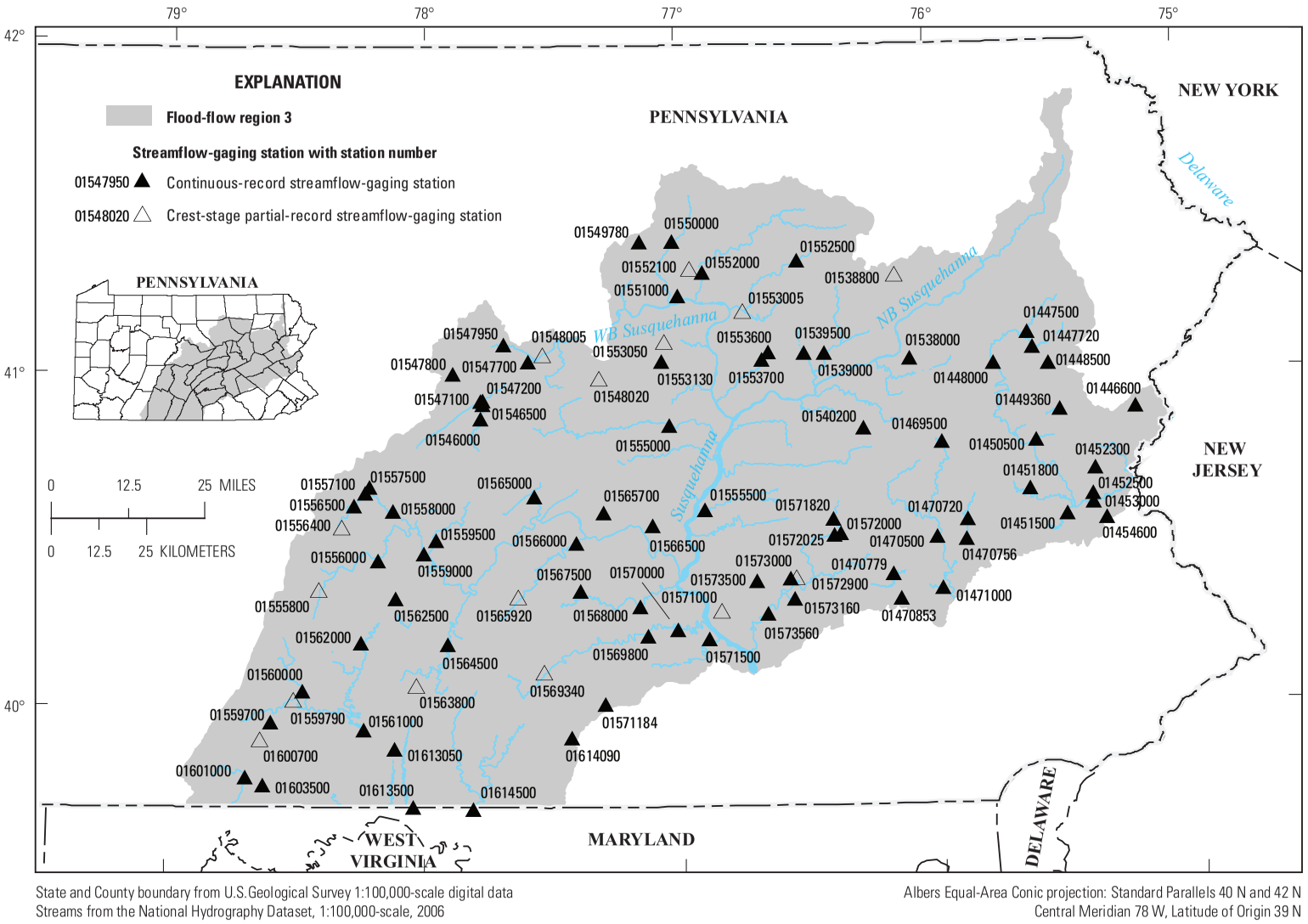


Figure 8. Flood-flow region 2 in Pennsylvania.



State and County boundary from U.S.Geological Survey 1:100,000-scale digital data
Streams from the National Hydrography Dataset, 1:100,000-scale, 2006

Albers Equal-Area Conic projection: Standard Parallels 40 N and 42 N
Central Meridian 78 W, Latitude of Origin 39 N

Figure 9. Flood-flow region 3 in Pennsylvania.

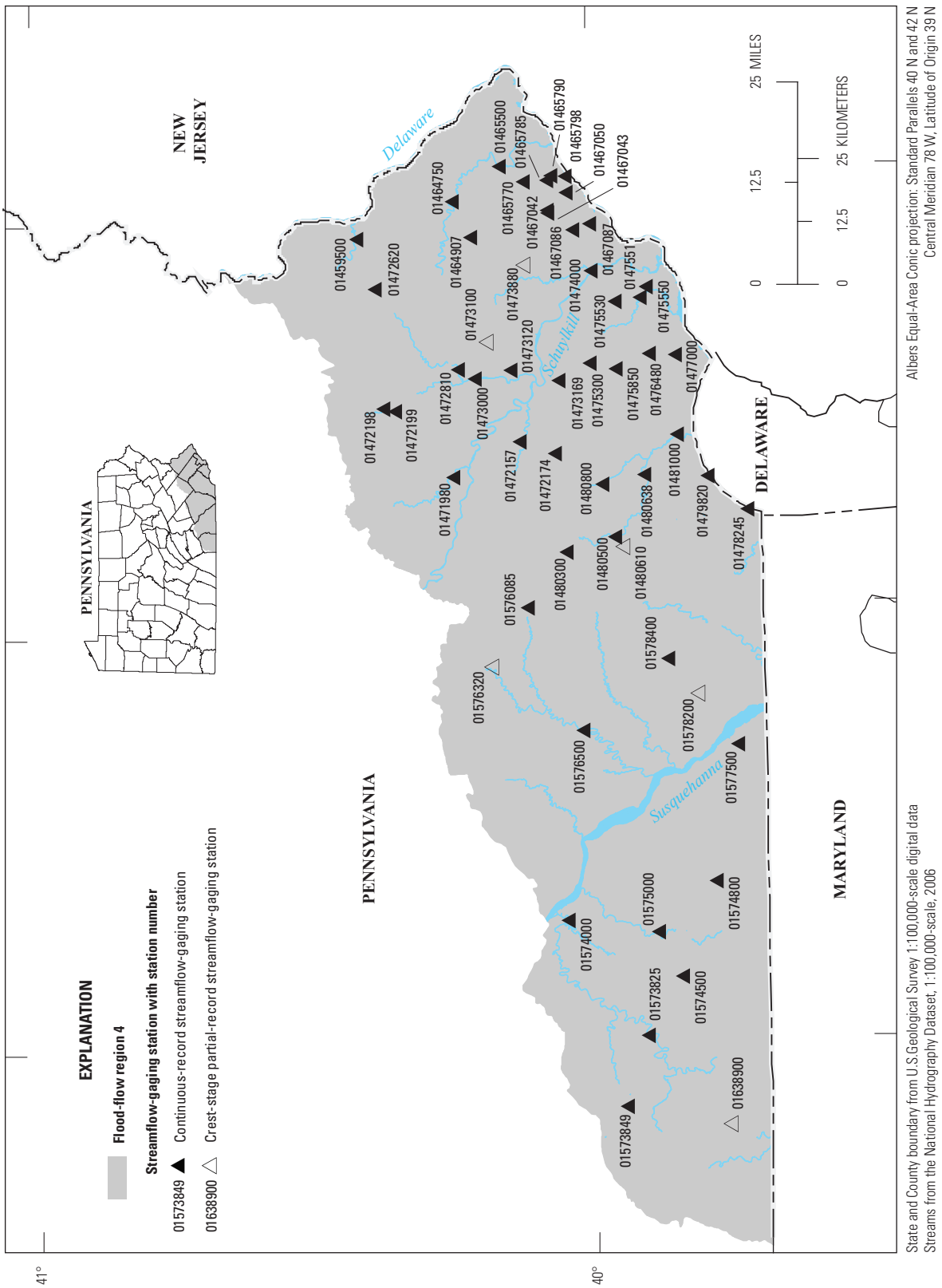


Figure 10. Flood-flow region 4 in Pennsylvania.

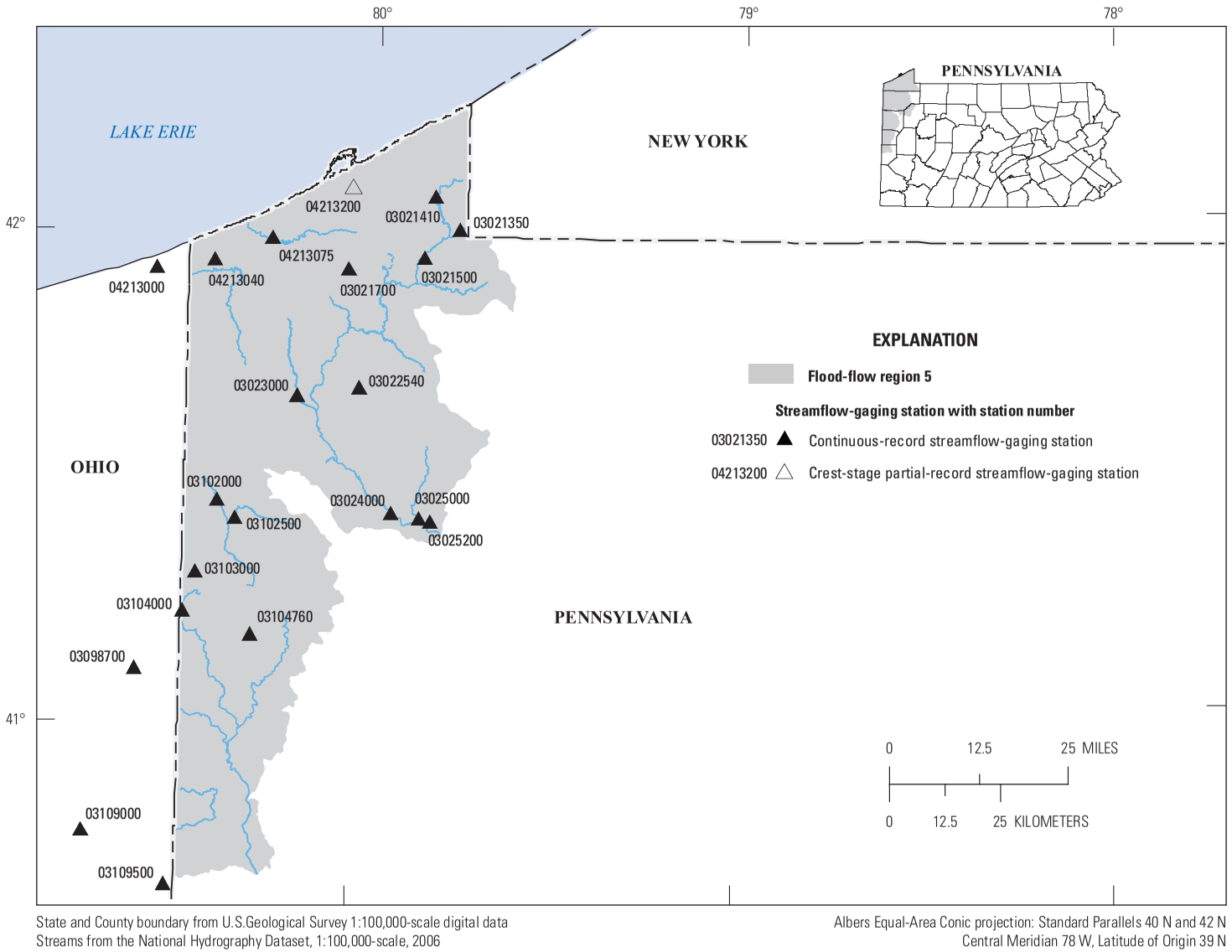


Figure 11. Flood-flow region 5 in Pennsylvania.

Table 3. Regression coefficients for use with flood-flow regression equations for Pennsylvania streams and model diagnostics.

[AEP, annual exceedance probability; Intercept, y-axis intercept of regression equation; Drainage area, drainage area of basin, in square miles; Maximum basin elevation, maximum elevation in the basin in feet; Percent storage, percentage of lakes, ponds, and wetlands in the basin; Percent carbonate bedrock, percentage of basin underlain by carbonate bedrock; Mean basin slope, average slope of the drainage basin, in degrees; $t_{(\alpha/2,\, n-p)}$, critical value from Student's t distribution for the 95-percent probability with $n-p$ degrees of freedom where n is the number of sites used in the regression equation and p is the number of variables plus 1; --, not a variable in regression equation for this flood-flow region]

Percent AEP	Number of stations used in analysis	Basin characteristic variables						Average variance of prediction (AVP)	Average standard error of prediction (SEP)	Coefficient of determination (pseudo R^2)	$t_{(\alpha/2,\, n-p)}$
		Intercept	Drainage area	Maximum basin elevation	Percent storage	Percent carbonate bedrock	Mean basin slope	Log units	Percent	Percent	
Region 1											
50	40	–4.0395	0.7199	1.8288	--	--	--	0.0116	25.2	97.12	2.024
20	40	–4.6287	0.6931	2.0831	--	--	--	0.0145	28.27	96.38	2.024
10	40	–4.9505	0.6803	2.2216	--	--	--	0.0166	30.31	95.86	2.024
4	40	–5.2953	0.6679	2.3706	--	--	--	0.0188	32.41	95.33	2.024
2	40	–5.5083	0.6606	2.4642	--	--	--	0.0201	33.56	95.05	2.024
1	40	–5.7210	0.6543	2.5552	--	--	--	0.0224	35.55	94.51	2.024
0.5	40	–5.9128	0.6489	2.6377	--	--	--	0.0248	37.51	93.97	2.024
0.2	40	–6.1649	0.6430	2.7438	--	--	--	0.0285	40.38	93.15	2.024
Region 2											
50	78	1.7840	0.8389	--	–0.0316	--	--	0.0124	26.07	97.6	1.992
20	78	1.9921	0.8253	--	–0.0530	--	--	0.0133	27.02	97.34	1.992
10	78	2.1038	0.8192	--	–0.0663	--	--	0.0144	28.19	97.08	1.992
4	78	2.2249	0.8137	--	–0.0813	--	--	0.0179	31.59	96.31	1.992
2	78	2.3042	0.8108	--	–0.0916	--	--	0.0215	34.76	95.56	1.992
1	78	2.3756	0.8088	--	–0.1011	--	--	0.0252	37.77	94.78	1.992
0.5	78	2.4422	0.8072	--	–0.1103	--	--	0.0300	41.55	93.73	1.992
0.2	78	2.5234	0.8058	--	–0.1219	--	--	0.0363	46.1	92.42	1.992
Region 3											
50	92	1.7719	0.8366	--	--	–0.0469	--	0.0303	41.74	92.72	1.987
20	92	2.0604	0.8058	--	--	–0.0435	--	0.0274	39.58	92.86	1.987
10	92	2.2165	0.7916	--	--	–0.0415	--	0.0258	38.28	93.02	1.987
4	92	2.3839	0.7782	--	--	–0.0398	--	0.0260	38.48	92.64	1.987
2	92	2.4916	0.7707	--	--	–0.0389	--	0.0266	38.92	92.27	1.987
1	92	2.5888	0.7644	--	--	–0.0380	--	0.0281	40.1	91.62	1.987
0.5	92	2.6772	0.7592	--	--	–0.0375	--	0.0297	41.29	90.98	1.987
0.2	92	2.7854	0.7531	--	--	–0.0369	--	0.0330	43.73	89.73	1.987
Region 4											
50	54	2.3029	0.6998	--	--	–0.0571	--	0.0285	40.38	87.38	2.007
20	54	2.5537	0.6727	--	--	–0.0558	--	0.0196	33.1	90.39	2.007
10	54	2.6911	0.6584	--	--	–0.0546	--	0.0172	30.86	91.23	2.007
4	54	2.8377	0.6450	--	--	–0.0538	--	0.0161	29.82	91.49	2.007
2	54	2.9320	0.6373	--	--	–0.0530	--	0.0166	30.36	91.03	2.007
1	54	3.0167	0.6309	--	--	–0.0521	--	0.0178	31.46	90.3	2.007

Table 3. Regression coefficients for use with flood-flow regression equations for Pennsylvania streams and model diagnostics.—Continued

[AEP, annual exceedance probability; Intercept, y-axis intercept of regression equation; Drainage area, drainage area of basin, in square miles; Maximum basin elevation, maximum elevation in the basin in feet; Percent storage, percentage of lakes, ponds, and wetlands in the basin; Percent carbonate bedrock, percentage of basin underlain by carbonate bedrock; Mean basin slope, average slope of the drainage basin, in degrees; $t_{(\alpha/2,\ n-p)}$, critical value from Student's *t* distribution for the 95-percent probability with $n-p$ degrees of freedom where n is the number of sites used in the regression equation and p is the number of variables plus 1; --, not a variable in regression equation for this flood-flow region]

Percent AEP	Number of stations used in analysis	Basin characteristic variables						Average variance of prediction (AVP)	Average standard error of prediction (SEP)	Coefficient of determination (pseudo R^2)	$t_{(\alpha/2,\ n-p)}$
		Intercept	Drainage area	Maximum basin elevation	Percent storage	Percent carbonate bedrock	Mean basin slope	Log units	Percent	Percent	
0.5	54	3.0934	0.6258	--	--	–0.0513	--	0.0191	32.68	89.5	2.007
0.2	54	3.1877	0.6199	--	--	–0.0502	--	0.0223	35.42	87.78	2.007
Region 5											
50	21	1.8137	0.7271	--	--	--	0.3445	0.0205	33.89	95.45	2.093
20	21	2.0159	0.6937	--	--	--	0.4053	0.0189	32.51	95.43	2.093
10	21	2.1283	0.6743	--	--	--	0.4412	0.0182	31.8	95.41	2.093
4	21	2.2533	0.6521	--	--	--	0.4803	0.0183	31.88	95.13	2.093
2	21	2.3384	0.6368	--	--	--	0.5030	0.0186	32.18	94.85	2.093
1	21	2.4185	0.6219	--	--	--	0.5229	0.0197	33.2	94.32	2.093
0.5	21	2.4949	0.6068	--	--	--	0.5413	0.0215	34.76	93.53	2.093
0.2	21	2.5866	0.5881	--	--	--	0.5674	0.0238	36.68	92.44	2.093

Table 4. Percent differences of predicted flood-flow magnitudes for select annual exceedancee probabilities between current and previous flood-flow regression equations.

	Annual exceedance probability (percent)						
	All	50	20	10	2	1	0.2
Statewide							
Mean	–0	–2	–2	–2	0	1	2
Median	1	–2	–1	0	2	2	5
Flood-flow region 1							
Mean	2	5	2	1	1	1	2
Median	5	7	4	2	5	6	8
Flood-flow region 2							
Mean	–7	–2	–6	–8	–9	–10	–10
Median	–9	–3	–10	–12	–12	–11	–9
Flood-flow region 3							
Mean	1	–9	–5	–2	4	7	11
Median	1	–11	–6	–2	6	7	9
Flood-flow region 4							
Mean	3	2	3	3	3	3	3
Median	4	11	7	3	4	5	1
Flood-flow region 5							
Mean	10	12	9	8	9	10	11
Median	13	15	11	11	13	15	16

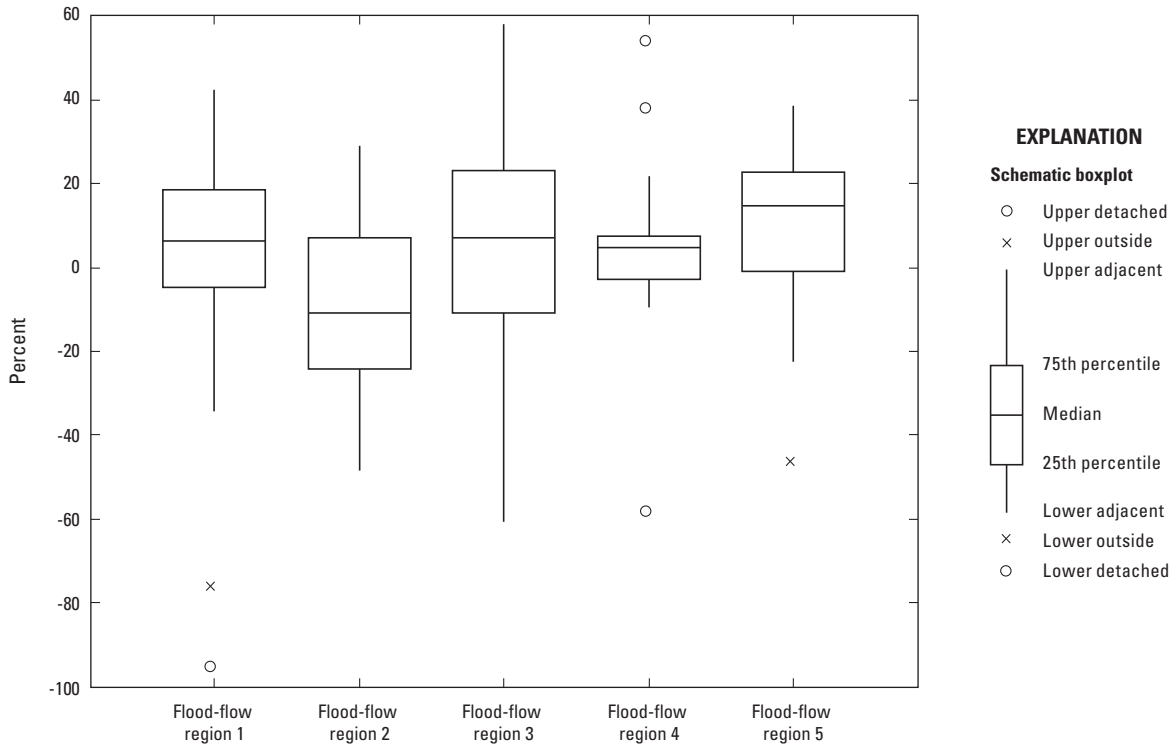


Figure 12. Boxplots showing percent differences between current and previous regression equation 1-percent annual exceedance probability flood-flow estimates.

estimates. The regression equations developed for the previous flood-flow region 4 used the fewest number of streamgages and had the highest standard error of prediction.

Uncertainty in the Regression Equations

When using regression equations to provide an estimate of flood flow at an ungaged site, it is important to understand the uncertainty associated with the estimate. This uncertainty is represented by the prediction interval, or the range of values within a specific confidence interval, within which the true value exists (Helsel and Hirsch, 2002). For example, for the 1-percent AEP flood, there is 95-percent confidence that the true value for that flood is within the minimum and maximum values of the 95-percent prediction interval. The following equations show how to compute the 95-percent prediction intervals for a flood-flow value estimated from regression equations.

$$PI_{U_i} = 10^{[Q_{reg_i} + t_{(\alpha/2,\, n-p)} (AVP_{reg})^{0.5}]} \quad (5)$$

$$PI_{L_i} = 10^{[Q_{reg_i} - t_{(\alpha/2,\, n-p)} (AVP_{reg})^{0.5}]} \quad (6)$$

where

Q_{reg_i} is the regional regression flood-flow estimate (in log base 10 units) for a given AEP at site *i*;

PI_{U_i}, PI_{L_i} are the upper and lower prediction interval confidence limits, respectively, for site *i*;

$t_{(\alpha/2,\, n-p)}$ is Student's *t* with a specified alpha (α) level and *n*-*p* degrees of freedom, where *n* is the number of sites used in the regression equation and *p* is the number of explanatory variables plus 1.0 (values are listed in table 3); and

AVP_{reg} is the average variance of prediction (in log base 10 units) for a given AEP (values are listed in table 3).

Specific to the variance of prediction, it is important to note that when applying equations 5 and 6 to ungaged sites and sites not used in the development of the regression equations, the average variance of prediction (AVP) value is used. However, when applying the equations to streamgages used in the development of the regression equations, the site-specific variance of prediction, as reported in appendix 2, is used in place of the AVP.

Accuracy and Limitations of Regression Equations

Regression equations are statistical models that utilize basin characteristics to estimate or predict flood flows for select AEPs. Certain coditions can limit the application of the regression equations presented in this report. Predicted

streamflow for basins with basin characteristics outside the ranges of those used to develop the regression equations may yield results inconsistent with the reported uncertainties shown in table 3. A summary of the range of all the variables is presented in table 5. It is worth noting that flood-flow region 4 in the southeastern part of the State has a maximum drainage area of 512 mi^2; basins with drainage areas greater than 512 mi^2 in flood-flow region 4 are not valid for use with the regression equations. Additionally, when applying the regression equations to estimate flood magnitude and frequency, it is important to maintain the same source of basin-characteristic data that was used in the development of the equations; otherwise, the predicted flood flows may not be valid. A description of the basin and climate characteristics incorporated into this study, as well as the source of the data, is provided in table 2. Impervious land cover associated with urbanization can increase the peak flows over similar areas lacking impervious land cover. Although streamgages in region 4 contained basins with increased percentages of urbanization, the median percent urban land cover for each region was less than 5 percent. None of the regional regression equations presented in this study contain a variable for percentage of urban development or impervious land cover, and as a result, effects from these variables may not be captured.

The regression equations should not be used to predict flood flows if streamflow at the site of interest is substantially affected by an upstream flood-control reservoir. The streamflow-gaging stations that were excluded from the regression analysis because of substantial upstream regulation are listed in appendix 3 with observed flood flows for specified recurrence intervals. The 500-year recurrence flow is not listed in the appendix because the storage capacity for some flood-control reservoirs may not be sufficient to store all the runoff associated with the 500-year flood event.

The accuracy of a regression model can be measured by how well the predicted flood-flow values represent the observed flood-flow values. This can be visualized in a plot of the observed values against the predicted values from the regression models for the 1-percent AEP (fig. 13). Other metrics of model fit generated for the GLS analysis by WREG include the pseudo-coefficient of determination (pseudo-R^2; Griffis and Stedinger, 2007b) and the average standard error of prediction (table 3). As discussed by Zarriello (2017) and Painter and others (2017), the pseudo-R^2 value is based on the variability of the dependent variable (flood flow) as determined by the regression after removing the effect of time sampling error. The pseudo-R^2 is similar to the standard regression coefficient of determination (R^2) in that the closer the value is to 1.0, the better the model fit and the greater the amount of variance that is explained by the regression. The pseudo-R^2 values for the updated regression equations ranged from 87.4 percent for the 50-percent AEP in region 4 to 97.6 percent for the 50-percent AEP in region 2, with values specific to the 1-percent AEP ranging from 90.3 to 94.8 percent across all flood-flow regions. The standard error of prediction is derived from the sum of the model error and the sampling error, and provides an estimate of reliability of the predicted flood flows (Helsel and Hirsch, 2002). Lower standard error of prediction values equate to less spread (or dispersion) of predicted values around the true value. The standard error of prediction for the flood-flow regression equations ranged from 25.2 percent for the 50-percent AEP in region 1 to 46.1 percent for the 0.2-percent AEP in region 2. The standard errors of prediction for the 1-percent AEP ranged from 31.5 to 40.1 percent across all flood-flow regions. Approximately two-thirds of the estimates obtained from the equations for ungaged sites will have errors less than the noted standard error of prediction (Ries and Dillow, 2006).

Table 5. Summary of the variables used to develop the flood-flow regional regression equations in Pennsylvania.

[mi^2, square miles; ft, feet]

Flood-flow region	Number of stations used	Range of basin characteristic variables				
		Drainage area (mi^2)	Maximum basin elevation (ft)	Mean basin slope (degrees)	Percent carbonate bedrock	Percent storage
1	40	3.04–1,490	1,470–2,690	3.63–11.7	0–3.14	0.26–26.6
2	78	0.92–1,610	1,300–3,370	4.27–16.1	0	0–8.9
3	92	1.42–1,280	669–3,150	2.58–13.0	0–100	0–23.0
4	54	1.20–512	176–2,090	1.77–7.48	0–68.5	0–6.92
5	21	2.27–1,030	897–1,890	1.55–6.82	0	0.13–19.2

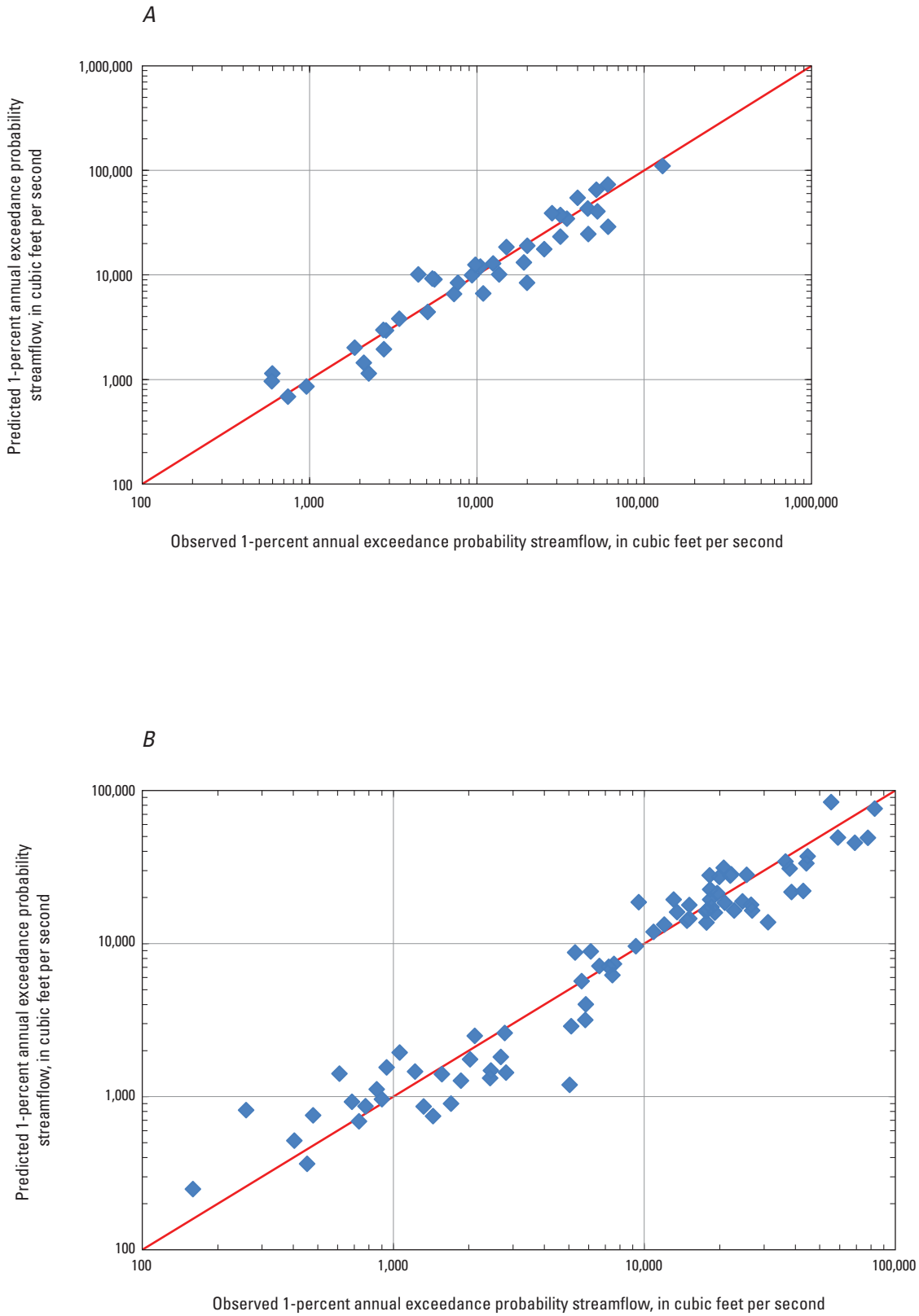


Figure 13. Comparison of the computed streamflows for the 1-percent annual exceedance probability using observed peak-flow data at streamgages and predicted data from the regional regression equations for the five flood-flow regions in Pennsylvania. *A*, Region 1; *B*, Region 2; *C*, Region 3; *D*, Region 4; and *E*, Region 5.

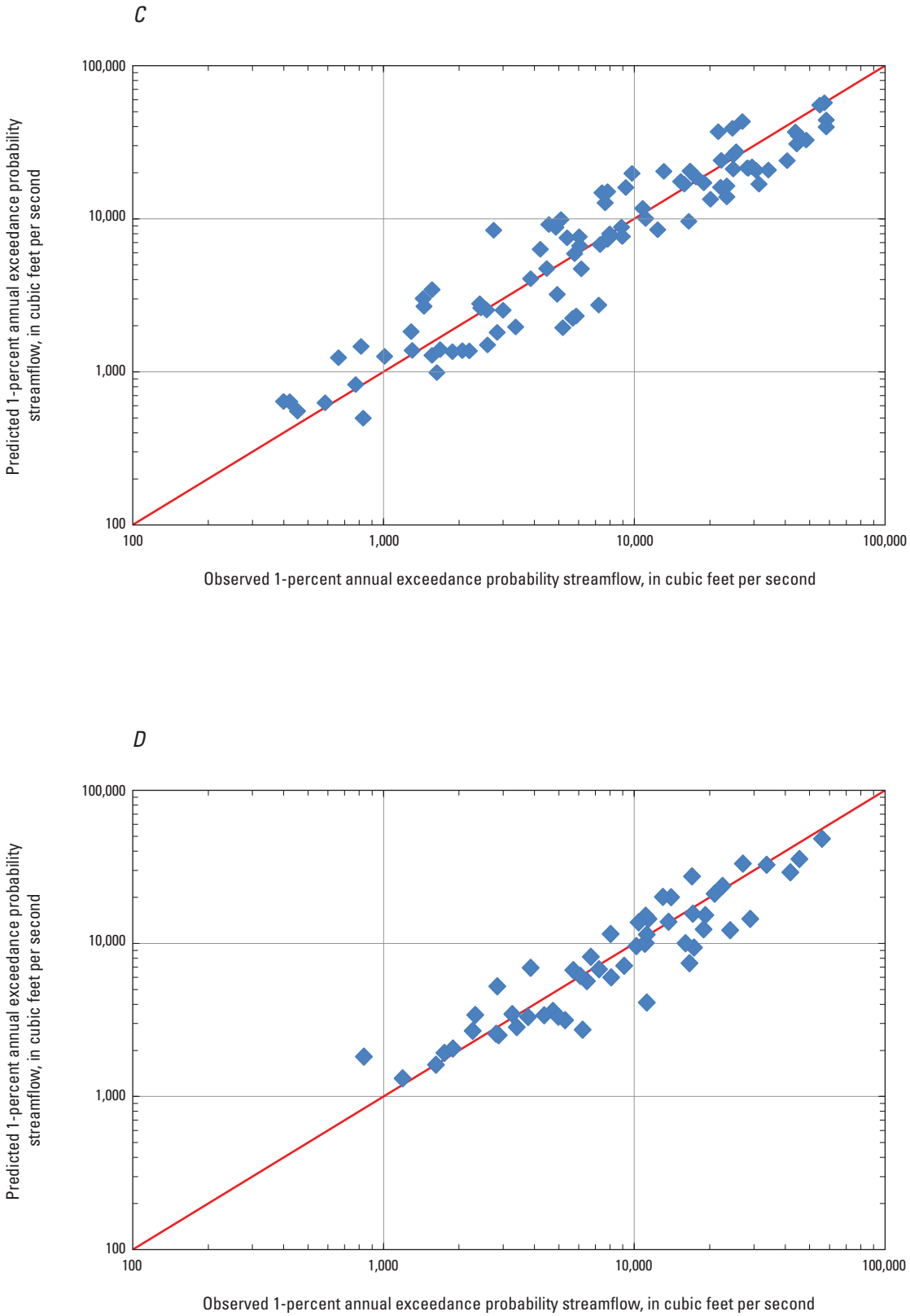


Figure 13. Comparison of the computed streamflows for the 1-percent annual exceedance probability using observed peak-flow data at streamgages and predicted data from the regional regression equations for the five flood-flow regions in Pennsylvania. *A*, Region 1; *B*, Region 2; *C*, Region 3; *D*, Region 4; and *E*, Region 5.—Continued

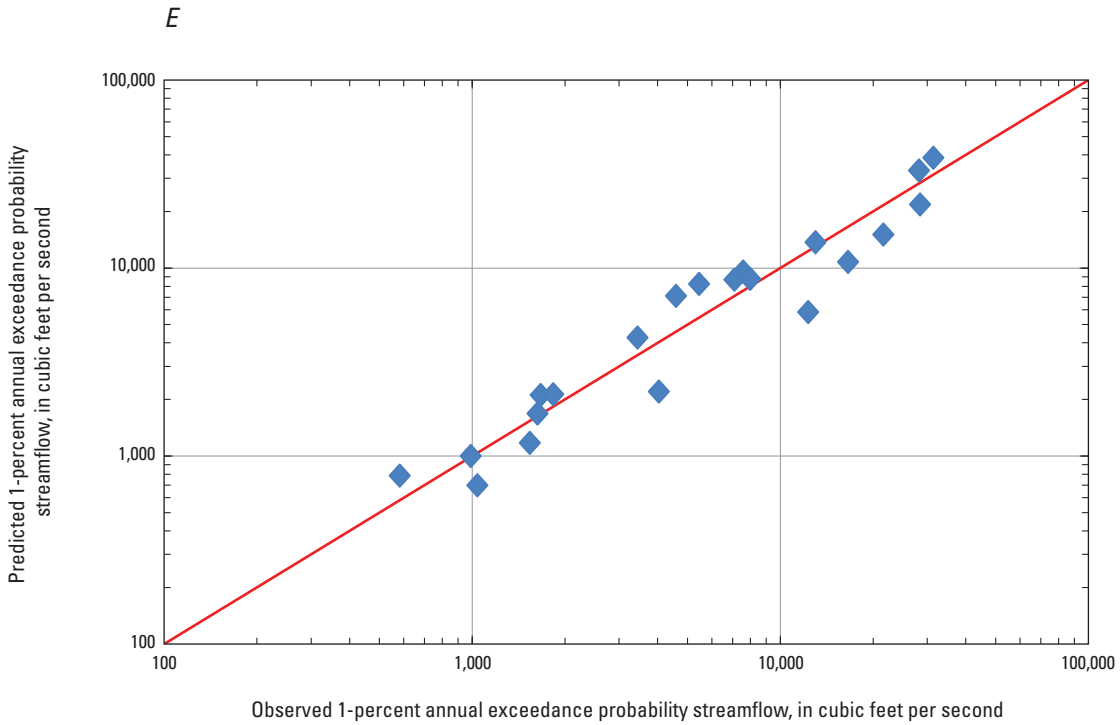


Figure 13. Comparison of the computed streamflows for the 1-percent annual exceedance probability using observed peak-flow data at streamgages and predicted data from the regional regression equations for the five flood-flow regions in Pennsylvania. *A*, Region 1; *B*, Region 2; *C*, Region 3; *D*, Region 4; and *E*, Region 5.—Continued

Computation of Weighted Flood-flow Estimates, Variances, and Confidence Limits at Gaged Sites

A weighted flood-flow estimate can be computed using the variance of independent estimates (observed and predicted), which results in reduced uncertainty. Bulletin 17C (England and others, 2018) describes a weighting method in which the observed at-site EMA and predicted regression estimates are weighted inversely proportional to their respective independent variances. The weighting equation is shown below (all variables in $\log_{10}$ units).

$$Q_{wtd_i} = \frac{\left(Q_{site_i} \cdot V_{reg_i}\right) + \left(Q_{reg_i} \cdot V_{site_i}\right)}{V_{site_i} + V_{reg_i}} \quad (7)$$

where

Q_{wtd_i} is the weighted flood-flow estimate for a given AEP at site *i*,

Q_{site_i} is the observed at-site EMA flood-flow estimate for a given AEP at site *i*,

V_{reg_i} is the variance of the regional regression estimate for a given AEP at site *i*,

Q_{reg_i} is the predicted regional regression flood-flow estimate for a given AEP at site *i*, and

V_{site_i} is the variance of the at-site EMA estimate for a given AEP at site *i*.

As shown below, a weighted variance (V_{wtd_i}) can be computed from the variances of each flood-flow estimate (all variables are in $\log_{10}$ units):

$$V_{wtd_i} = \frac{V_{site_i} \cdot V_{reg_i}}{V_{site_i} + V_{reg_i}} \quad (8)$$

Once Q_{wtd_i} and V_{wtd_i} have been determined (eqs. 7 and 8), the 95-percent confidence interval (95-percent_*CI*) can be computed using the following equations:

$$\text{95-percent\_}CI_{upper_i} = 10^{\left[Q_{wtd_i} + t_{(\alpha/2,\, n-p)} \left(V_{wtd_i}\right)^{0.5}\right]} \quad (9)$$

$$\text{95-percent\_}CI_{lower_i} = 10^{\left[Q_{wtd_i} - t_{(\alpha/2,\, n-p)} \left(V_{wtd_i}\right)^{0.5}\right]} \quad (10)$$

where

$t_{(\alpha/2,\, n-p)}$ is Student's *t* with a specified alpha (α) level and *n*-*p* degrees of freedom, where *n* is the number of sites used in the regional regression equation and *p* is the number of explanatory variables plus 1.0 (values are listed in table 3).

As noted by Zarriello (2017), the variables needed to calculate equations 7 and 8 are obtained directly from output from PeakFQ (ver. 7.1) and WREG (ver. 2.02) for sites used in the regional regression model. WREG computes a variance of prediction for each streamgage used in the regional regression analysis that is used for V_{reg_i}. For unregulated streamgages not used in the development of regional regression equations but having sufficient data for computation of at-site AEP flows, the AVP for the regression (table 3) can be substituted for V_{reg_i} to compute a weighted estimate of AEP flows for the site. Weighted estimates of AEP flows, variances, and confidence limits are reported in appendix 2.

Example of Weighting a Flood-flow Estimate with Observed and Predicted Values

Example 2. Calculate the 1-percent AEP weighted flood flow (Qwtd) for USGS streamgage 01516350, Tioga River near Mansfield, Pennsylvania, which is located in the northeastern part of Pennsylvania at latitude 41o47'49" and longitude 77o04'50". The drainage area is 153 mi2 and the maximum basin elevation is 2,446 feet. The basin is unaffected by substantial regulation, diversion, or mining.

1. From example 1, a predicted 1-percent AEP flood-flow estimate (Q_{reg}) of 23,300 cubic feet per second (ft^3/s) (converted to 4.367 ft^3/s log units) was computed using the appropriate regression equation for flood-flow region 1.

2. From appendix 2, the predicted variance (V_{reg}) for the regression estimate is reported as 0.0213 (log units).

3. From appendix 2, the observed at-site 1-percent AEP flood-flow estimate (Q_{site}) and corresponding at-site variance (V_{site}) are reported as 31,500 ft^3/s (4.498 ft^3/s log units) and 0.0065 (log units), respectively.

4. Using equation 7, a weighted flood-flow estimate can be computed as follows,

   Q_{wtd} = [(4.498 ft^3/s) • (0.0213)] + [(4.367 ft^3/s) • (0.0065)] / (0.0065 + 0.0213)

   Q_{wtd} = 0.1242 / 0.0278

   Q_{wtd} = 4.467

   Q_{wtd} = $10^{(4.467)}$

   Q_{wtd} = 29,300 ft^3/s (rounded to 3 significant figures).

Examples for Computing a Weighted Variance (V_{wtd}) and 95-percent Confidence Intervals (CI_{upper}, CI_{lower})

Example 3. Calculate the weighted variance for USGS streamgage 01516350, Tioga River near Mansfield, Pennsylvania.

1. From appendix 2, the at-site variance (V_{site}) for the observed estimate is reported as 0.0065 (log units) and the predicted variance (V_{reg}) for the regression estimate is reported as 0.0213 (log units).

2. Using equation 8, a weighted variance (V_{wtd}) can be computed as follows,

   V_{wtd} = [(0.0065) • (0.0213)] / (0.0065 + 0.0213)

   V_{wtd} = 0.00014 / 0.0278

   V_{wtd} = 0.0050.

Example 4. Calculate the 95-percent confidence intervals for USGS streamgage 01516350, Tioga River near Mansfield, Pennsylvania.

1. From the previous examples, the weighted flood-flow estimate (Q_{wtd}) and weighted variance (V_{wtd}) were computed as 29,300 ft^3/s (converted to 4.467 ft^3/s log units) and 0.0050 (log units), respectively.

2. From table 3, the Student's *t* distribution value for flood-flow region 1 is 2.024.

3. Using equation 9, an upper confidence level for the 1-percent AEP flood-flow (CI_{upper}), can be computed as follows

   CI_{upper} = $10^{[4.467 + (2.024) \cdot (0.0050)\text{^}0.5]}$

   CI_{upper} = $10^{(4.610)}$

   CI_{upper} = 40,700 ft^3/s.

4. Using equation 10, a lower confidence level for the 1-percent AEP flood-flow (CI_{lower}), can be computed as follows,

   CI_{lower} = $10^{[4.467 - (2.024) \cdot (0.0050)\text{^}0.5]}$

   CI_{lower} = $10^{(4.324)}$

   CI_{lower} = 21,100 ft^3/s.

It should be noted that as a result of rounding issues, the resultant computational values provided in these examples do not necessarily exactly match the values reported in appendix 2.

Estimating Flood Flows at Ungaged Sites Near a Streamgage

For unregulated streams, if the site of interest is at an ungaged location, but near a streamgage on the same stream having at least 10 years of record, a more accurate estimate of flood flow can be obtained by including the at-site flood frequency information from the streamgage, as opposed to using just the regional regression equation (Ries, 2007). This section outlines the weighting procedure assuming the ungaged site is within 0.5 to 1.5 times the drainage area of the streamgage. The first step is to obtain the streamgage-based estimate for the ungaged site, $Q_{(u)_g}$, based on flow per unit area using the equation

$$Q_{(u)_g} = \left(\frac{A_u}{A_g}\right)^b \bullet Q_{(g)_{wtd}} \quad (11)$$

where

A_u is the drainage area of the ungaged site,
A_g is the drainage area of the streamgage on the same stream as the ungaged site,
b is the exponent of the drainage area variable in the regional regression equation (table 3), and
$Q_{(g)_{wtd}}$ is the weighted estimate of flood flow for the desired AEP at the streamgage.

The next step involves computing the weighted flood-flow estimate for the ungaged site, $Q_{(u)_{wtd}}$, using the result from equation 11 and the following equation. As noted in Ries (2007) and Mastin (2016), this weighting algorithm gives full weight to the regression estimates when applied to ungaged locations 0.5 or 1.5 times the drainage area of the gaging station and increasing weight to the gaging station-based estimates as the drainage area ratio approaches 1. The weighting procedure should not be applied when the drainage area of the ungaged site is less than 0.5 or greater than 1.5 times the drainage area of the gaging station.

$$Q_{(u)_{wtd}} = \left(\frac{2\,|DA|}{A_g}\right) Q_{(u)_{reg}} + \left(1 - \frac{2\,|DA|}{A_g}\right) Q_{(u)_g} \quad (12)$$

where

$|DA|$ is the absolute value of the difference between the drainage areas of the streamgage and ungaged site, and
$Q_{(u)_{reg}}$ is the predicted flood-flow estimate for the ungaged site computed with the updated regional regression equations for the desired AEP.

General Guidelines for the Estimation of Magnitude and Frequency of Flood Flows

Methods for estimating the magnitude and frequency of flood flows for unregulated Pennsylvania streams depend on the data available for the site of interest. The methods are presented in this section include

- If a streamgage exists at the site of interest and has 10 or more years of record, techniques described in Bulletin 17C (England and others, 2018) should be used to compute at-site flood-flow estimates for the desired AEP flood quantile. It is generally recommended (particularly for streamgages having shorter periods of record) that the at-site estimates be combined with flood-flow estimates predicted from regional regression equations, to compute a weighted average.
- If the site of interest is ungaged and within 0.5 to 1.5 times the drainage area of a nearby streamgage located on the same unregulated stream, flood-flow estimates should be computed for the site using the regional regression equations. Flood-flow estimates should also be computed for the nearby streamgage. These estimates should then be used (along with the respective drainage areas of the site of interest and the nearby streamgage) to compute a drainage-area ratio weighted flood-flow estimate (utilizing eqs. 11 and 12) for the site of interest.
- If the site of interest is ungaged and there are no nearby streamgages on the same unregulated stream, regional regression equations based on basin characteristics may be used to estimate flood flow for the desired AEP flood quantile.

The USGS StreamStats web application (https://water.usgs.gov/osw/streamstats) contains flood-flow estimates for many streamgages across Pennsylvania and the Nation. The regional regression equations developed for Pennsylvania (2019) will be incorporated into StreamStats (along with supporting basin characteristic datasets) to facilitate computations at ungaged locations. For more information on StreamStats see Ries and others (2017).

Summary

A study was conducted by the U.S. Geological Survey, in cooperation with the Pennsylvania Department of Transportation and the Federal Emergency Management Agency to compute updated flood-frequency estimates for streamgages in and near the borders of Pennsylvania and to develop flood-flow regression equations for unregulated Pennsylvania streams.

Flood-flow statistics are crucial for the design of bridges and flood-control structures, floodplain management, and hazard preparedness to help minimize damages associated with flooding. Accessible methods that produce estimates of the frequency and magnitude of floods are important to engineers and planners working on such projects.

Sites initially selected for this study consisted of 356 active and discontinued continuous- and partial-record streamgages minimally impacted by flow regulation, diversion, and mining activity. Streamgages included in this study were also required to have a minimum of 10 years of annual peak flow record through water year 2015. After careful review of annual peak flow data and completion of a redundancy analysis, the number of streamgages selected for inclusion in the regression analysis was reduced to 285, although flood-frequency statistics were computed and reported for all 356 streamgages. Trends in annual peak flow data for all 356 streamgages were also included in the analysis.

A regional skew analysis was also included as part of this study. Skew is an important statistical measure used to fit a log-Pearson Type III frequency distribution to the data when estimating the magnitude of a flood flow for a given probability. In the past, the regionalized skew value could be obtained from a national map associated with the Bulletin 17B publication (Interagency Advisory Committee on Water Data, 1982) that had an associated mean squared error (MSE) of 0.302. For this study, a regional skew value was developed for Pennsylvania that had an MSE of 0.181. The updated regional skew value was incorporated into the expected moments algorithm methodology to estimate flood-frequency statistics at unregulated streamgages included in this study.

Regression equations were developed to estimate flood-flows for the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probabilities (which correspond to the 2-, 5-, 10-, 25-, 50-, 100-, 200-, and 500-year recurrence-intervals, respectively) for five flood-flow regions in Pennsylvania. The following basin characteristics were significant at the 95-percent confidence level for one or more regression equations: drainage area, maximum basin elevation, mean basin slope, percent storage, and the percentage of carbonate bedrock within a basin. The standard errors of prediction for the flood-flow regression equations ranged from 25.2 percent for the 50-percent annual exceedance probability (AEP) in region 1 to 46.1 percent for the 0.2-percent AEP in region 2. The standard errors of prediction for the 1-percent AEP ranged from 31.5 to 40.1 percent across all flood-flow regions. To minimize temporal bias that may be associated with a station, a weighting method is presented that incorporates the observed, as well as the predicted, flood flows into a weighted-average flood-frequency streamflow estimate.

Certain conditions can limit the application of the regression equations presented in this report. The equations should not be used if the site of interest has a contributing drainage area or basin characteristics outside of the ranges used to develop the regression equations, as they may not yield valid predicted streamflow estimates. The regression equations should not be used to predict flood-flow frequency statistics if streamflow at the site of interest is substantially affected by upstream flood-control regulation, diversion, or mining. Estimates of flood-flow magnitude for streamgages substantially affected by upstream regulation are also presented in this report.

Acknowledgments

The authors wish to thank the U.S. Geological Survey Pennsylvania Water Science Center (as well as the Maryland, New York, Ohio, and West Virginia) Hydrologic Surveillance Program staffs for their compilation and meticulous review of gaging station streamflow data that were used in the computation of flood frequency estimates, which were subsequently used in the development of flood-flow regression equations. Special thanks are given to Curtis V. Price and Scott A. Hoffman for their geographic information system (GIS) expertise and assistance in the compilation and validation of basin characteristics considered for this study. The authors also thank Linda F. Zarr for her diligent support in organizing and formatting the data files, tables, and appendixes incorporated into this study.

References Cited

Bonnin, G.M., Martin, D., Lin, B., Parzybok, T., Yekta, M., and Riley, D., 2006, Precipitation-frequency atlas of the United States: National Oceanic and Atmospheric Administration Atlas 14, v. 2, 295 p.

Cohn, T.A., England, J.F., Berenbrock, C.E., Mason, R.R., Stedinger, J.R., and Lamontagne, J.R., 2013, A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series: Water Resources Research, v. 49, no. 8, p. 5047–5058.

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.

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., accessed December 2017 at http://dx.doi.org/10.3133/sir20165024.

Dinicola, K., 1996, The “100-year flood”: U.S. Geological Survey Fact Sheet 229–96, 2 p.

Ehlke, M.H., and Reed, L.A., 1999, Comparison of methods for computing streamflow statistics for Pennsylvania streams: U.S. Geological Survey Water-Resources Investigations Report 99–4068, 80 p.

Eng, K., Chen, Y.-Y., and Kiang, J.E., 2009, User's guide to the weighted-multiple-linear-regression program (WREG ver. 1.0): U.S. Geological Survey Techniques and Methods, book 4, chap. A8, 21 p., accessed July 2016 at https://pubs.usgs.gov/tm/tm4a8.

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., 2018, Guidelines for determining flood flow frequency—Bulletin 17C: U.S. Geological Survey Techniques and Methods, book 4, chap. B5, 148 p., accessed October 2018 at https://dx.doi.org/10.3133/tm4B5.

Falcone, J.A., 2016, U.S. block-level population density rasters for 1990, 2000, and 2010: U.S. Geological Survey data release, accessed April 2017 at http://dx.doi.org/10.5066/F74J0C6M.

Farmer, W.H., 2017, Weighted multiple-linear-REGression (WREG) program: Github web page, accessed December 20, 2017, at https://github.com/wfarmer-usgs/WREG.

Flippo, H.N., Jr., 1977, Floods in Pennsylvania: Pennsylvania Department of Environmental Resources, Water Resources Bulletin No. 13, 59 p.

Flippo, H.N., Jr., 1982, Evaluation of the streamflow data program in Pennsylvania: U.S. Geological Survey Water-Resources Investigations 82-21, 56 p.

Flynn, K.M., Kirby, W.H., and Hummel, P.R., 2006, User's manual for program PeakFQ, annual flood frequency analysis using Bulletin 17B guidelines: U.S. Geological Survey Techniques and Methods, book 4, chap. B4, 42 p.

Gallant, A.L., Whittier, T.R., Larsen, D.P., Omernik, J.M., and Hughes, R.M., 1989, Regionalization as a tool for managing environmental resources: U.S. Environmental Protection Agency EPA/600/3-89/060, 152 p.

Griffis, V.W., and Stedinger, J.R., 2007a, Evolution of flood frequency analysis with Bulletin 17: Journal of Hydrologic Engineering, v. 12, no. 3, p. 283–297.

Griffis, V.W., and Stedinger, J.R., 2007b, The use of GLS regression in regional hydrologic analyses: Journal of Hydrology, v. 344, no. 1–2, p. 82–95.

Griffith, G.E., Omernik, J.M., Wilton, T.F., and Pierson, S.M., 1994, Ecoregions and subregions of Iowa—A framework for water quality assessment and management: The Journal of the Iowa Academy of Science, v. 101, no. 1, p. 5–13.

Grubbs, F.E., and Beck, G., 1972, Extension of sample sizes and percentage points for significance tests of outlying observations: Technometrics, v. 14, no. 4, p. 847–854.

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, 522 p.

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.

Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood-flow frequency: U.S. Geological Survey Bulletin 17B, 183 p.

Kolb, K.R., Steeves, P.A., and Hoffman, S.A., 2020, Basin characteristics rasters for Pennsylvania StreamStats 2020: U.S. Geological Survey data release, https://doi.org/10.5066/P9M47KLH.

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., accessed August 2018 at http://dx.doi.org/10.3133/sir20165118.

Omernik, J.M., 1987, Ecoregions of the conterminous United States (map supplement): Annals of the Association of American Geographers, v. 77, no. 1, p. 118–125, scale 1:7,500,000.

Omernik, J.M., 1995, Ecoregions—A framework for environmental management, *in* Davis, W.S., and Simon, T.P., eds., Biological assessment and criteria—Tools for water resource planning and decision making: Boca Raton, Fla., Lewis Publishers, p. 49–62.

Painter, C.C., Heimann, D.C., and Lanning-Rush, J.L., 2017, Methods for estimating annual exceedance-probability streamflows for streams in Kansas based on data through water year 2015 (ver. 1.1, September 2017): U.S. Geological Survey Scientific Investigations Report 2017–5063, 20 p., accessed February 2018 at https://doi.org/10.3133/sir20175063.

PRISM Climate Group, 2017, PRISM climate data: Oregon State University database, accessed April 2017 at http://prism.oregonstate.edu.

Ries, K.G., III, 2007, The national streamflow statistics program—A computer program for estimating streamflow statistics for ungaged sites: U.S. Geological Survey Techniques and Methods, book 4, chap. A6, 37 p.

Ries, K.G., III, and Dillow, J.J.A., 2006, Magnitude and frequency of floods on nontidal streams in Delaware: U.S. Geological Survey Scientific Investigations Report 2006–5146, 59 p.

Ries, K.G., III, Newson, J.K., Smith, M.J., Guthrie, J.D., Steeves, P.A., Haluska, T.L., Kolb, K.R., Thompson, R.F., Santoro, R.D., and Vraga, H.W., 2017, StreamStats—Version 4: U.S. Geological Survey Fact Sheet 2017–3046, 4 p. [Also available at https://pubs.usgs.gov/fs/2017/3046/fs20173046.pdf.]

Roland, M.A., and Stuckey, M.H., 2008, Regression equations for estimating flood flows at selected recurrence intervals for ungaged streams in Pennsylvania: U.S. Geological Survey Scientific Investigations Report 2008–5102, 57 p.

Schwarz, G.E., and Alexander, R.B., 1995, State Soil Geographic (STATSGO) data base for the conterminous United States: U.S. Geological Survey Open-File Report 95–449, accessed April 2017 at https://water.usgs.gov/GIS/metadata/usgswrd/XML/ussoils.xml.

Sloto, R.A., Stuckey, M.H., and Hoffman, S.A., 2017, Evaluation of the streamgage network for estimating streamflow statistics at ungaged sites in Pennsylvania and the Susquehanna River Basin in Pennsylvania and New York: U.S. Geological Survey Scientific Investigations Report 2016–5149, 102 p., accessed January 2018 at https://doi.org/10.3133/sir20165149.

Soller, D.R., and Packard, P.H., 1998, Digital representation of a map showing the thickness and character of Quaternary sediment in the glaciated United States east of the Rocky Mountains: U.S. Geological Survey Digital Data Series DDS-38, scale 1:1,000,000, accessed April 2017 at http://pubs.usgs.gov/dds/dds38/shape.html.

Stuckey, M.H., and Reed, L.A., 2000, Techniques for estimating magnitude and frequency of peak flows for Pennsylvania streams: U.S. Geological Survey Water-Resources Investigations Report 00–4189, 43 p.

Stuckey, M.H., and Roland, M.A., 2011, Selected streamflow statistics for streamgage locations in and near Pennsylvania: U.S. Geological Survey Scientific Investigations Report 2011–1070, 88 p.

TIBCO Software Inc., 2008, TIBCO Spotfire S+ 8.1 guide to packages: Palo Alto, Calif., 77 p.

U.S. Census Bureau, 2018, State area measurements and internal point coordinates: U.S. Census Bureau web page, accessed October 3, 2018, at https://www.census.gov/geo/reference/state-area.html.

U.S. Geological Survey, 2000a, US GeoData digital elevation models: U.S. Geological Survey Fact Sheet 040-00, 2 p., accessed April 2017 at https://doi.org/10.3133/fs04000.

U.S. Geological Survey, 2000b, The National Hydrography Dataset: U.S. Geological Survey database, accessed April 2017 at https://www.usgs.gov/core-science-systems/ngp/national-hydrography/national-hydrography-dataset?qt-science_support_page_related_con=0#qt-science_support_page_related_con.

U.S. Geological Survey, 2014, PeakFQ (ver. 7.1): U.S. Geological Survey software release, accessed March 2017 at http://water.usgs.gov/software/PeakFQ/.

U.S. Geological Survey, 2017, The StreamStats program: U.S. Geological Survey StreamStats Program web page, accessed December 2017 at http://streamstats.usgs.gov.

Veilleux, A.G., Cohn, T.A., Flynn, K.M., Mason, R.R., Jr., 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.

Wiken, E., 1986, Terrestrial ecozones of Canada: Environment Canada, Lands Directorate, Ecological Land Classification Series No. 19, 26 p. plus map.

Woods, A.J., and Omernik, J.M., 1996, Ecoregions of Pennsylvania: Pennsylvania Geographer, v. XXXIV, no. 2, p. 3–37.

Xian, G., Homer, C., Dewitz, J., Fry, J., Hossain, N., and Wickham, J., 2011, The change of impervious surface area between 2001 and 2006 in the conterminous United States: Photogrammetric Engineering and Remote Sensing, v. 77, no. 8, p. 758–762.

Zarriello, P.J., 2017, Magnitude of flood flows at selected annual exceedance probabilities for streams in Massachusetts: U.S. Geological Survey Scientific Investigations Report 2016–5156, 54 p., accessed December 2017 at https://doi.org/10.3133/sir20165156.

Appendixes 1, 2, and 3, available online as Excel files at https://doi.org/10.3133/sir20195094

Appendix 1

Unregulated streamgages considered for the development of updated flood-flow regression equations for Pennsylvania streams.

Appendix 2

Magnitude, variance, and confidence intervals of annual exceedance probability floods for select unregulated streamgages in Pennsylvania and surrounding states.

Appendix 3

Magnitude, variance, and confidence intervals of annual exceedance probability floods for select streamgages in Pennsylvania substantially affected by upstream regulation.

For additional information, contact:

Director, Pennsylvania Water Science Center

U.S. Geological Survey
215 Limekiln Road
New Cumberland, Pa. 17070

Or visit our website at:
https://www.usgs.gov/centers/pa-water

Publishing support provided by the
West Trenton Publishing Service Center

USGS

Roland and Stuckey—Development of Regression Equations for the Estimation of Flood Flows at Ungaged Streams in Pennsylvania—Scientific Investigations Report 2019–5094, ver 1.1

ISSN 2328-0328 (online)
https://doi.org/10.3133/sir20195094