USGS
science for a changing world

Development of Regression Equations for the Estimation of the Magnitude and Frequency of Floods at Rural, Unregulated Gaged and Ungaged Streams in Puerto Rico Through Water Year 2017

Scientific Investigations Report 2021–5062

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

Cover. High water and debris from Hurricane Maria destroyed the Highway 404 bridge at U.S. Geological Survey (USGS) streamgage 50147800 Rio Culebrinas at Highway 404 near Moca. View is oriented northeast, from the right bank toward the left bank. Photograph by Mark E. Smith, USGS, October 23, 2017.

Development of Regression Equations for the Estimation of the Magnitude and Frequency of Floods at Rural, Unregulated Gaged and Ungaged Streams in Puerto Rico Through Water Year 2017

By Patrick J. Ryan, Anthony J. Gotvald, Cody L. Hazelbaker, Andrea G. Veilleux, and Daniel M. Wagner

Scientific Investigations Report 2021–5062

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

U.S. Geological Survey, Reston, Virginia: 2021

For more information on the USGS—the Federal source for science about the Earth, its natural and living resources, natural hazards, and the environment—visit https://www.usgs.gov or call 1–888–ASK–USGS.

For an overview of USGS information products, including maps, imagery, and publications, visit https://store.usgs.gov.

Suggested citation:
Ryan, P.J., Gotvald, A.J., Hazelbaker, C.L., Veilleux, A.G., and Wagner, D.M., 2021, Development of regression equations for the estimation of the magnitude and frequency of floods at rural, unregulated gaged and ungaged streams in Puerto Rico through water year 2017: U.S. Geological Survey Scientific Investigations Report 2021–5062, 37 p., https://doi.org/10.3133/sir20215062.

Associated data for this publication:
Ryan, P.J., and Hazelbaker, C.L., 2021, Data files for the development of regression equations for estimation of the magnitude and frequency of floods at rural, unregulated gaged and ungaged streams in Puerto Rico through water year 2017: U.S. Geological Survey data release, https://doi.org/10.5066/P91XT14B.

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	
square mile (mi^2)	2.590	square kilometer (km^2)
	Flow rate	
cubic foot per second (ft^3/s)	0.02832	cubic meter per second (m^3/s)
	Precipitation	
inch per year (ft/d)	2.54	centimeter per year (cm/yr)

Datum

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

Abbreviations

AEP	annual exceedance probability
ANOVA	analysis of variance
AVP	average variance of prediction
B–GLS	Bayesian generalized least-squares
B–WLS	Bayesian weighted least-squares
CSG	crest-stage gage
DAR	drainage area ratio
EMA	Expected Moments Algorithm
GIS	geographic information system
GLS	generalized least squares
HUC8	8-digit hydrologic unit code
MEV	model error variance
MGBT	Multiple Grubbs-Beck test
MSE	mean squared error
NOAA	National Oceanic and Atmospheric Administration
NRCS	Natural Resources Conservation Service
NWIS	National Water Information System
NWS	National Weather Service
OLS	ordinary least squares
PeakFQ	U.S. Geological Survey peak-flow analysis program
PILF	potentially influential low flow
PRISM	Parameter-elevation Regressions on Independent Slopes Model
QA/QC	quality assurance and quality control
R^2	coefficient of determination
$Radj^2$	adjusted coefficient of determination
SD	standardized distance
SEP	standard error of prediction
S_{MEV}	standard model error variance
USGS	U.S. Geological Survey
VIF	variance inflation factor
WLS	weighted least squares
WREG	Weighted Multiple Linear Regression (package)

Development of Regression Equations for the Estimation of the Magnitude and Frequency of Floods at Rural, Unregulated Gaged and Ungaged Streams in Puerto Rico Through Water Year 2017

By Patrick J. Ryan, Anthony J. Gotvald, Cody L. Hazelbaker, Andrea G. Veilleux, and Daniel M. Wagner

Abstract

The methods of computation and estimates of the magnitude of flood flows were updated for the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent chance exceedance levels for 91 streamgages on the main island of Puerto Rico by using annual peak-flow data through 2017. Since the previous flood frequency study in 1994, the U.S. Geological Survey has collected additional peak flows at additional streamgages, and Puerto Rico has experienced numerous flood events. This updated study was performed using longer annual peak-flow datasets from more stations to provide more representative equations to predict flood flows. Screening criteria for these streamgages included 10 or more years of annual peak-flow data, unregulated flow, and less than 10 percent impervious drainage area.

The magnitude and frequency of floods at selected streamgages in Puerto Rico were estimated using updated methods outlined in Bulletin 17C. The new procedures include a regional skew analysis that incorporates Bayesian regression techniques, the Expected Moments Algorithm to better represent missing record and estimate parameters of the log-Pearson Type III distribution, and the Multiple Grubbs-Beck test for low outlier detection.

Regional regression equations were developed to estimate peak-flow statistics at ungaged locations by using selected basin and climatic characteristics as explanatory variables. These variables were determined from digital spatial datasets and geographic information systems by using the most recent data available. Ordinary least-squares regression techniques were used to filter the basin characteristics and determine two separate regions, region 1 (west) and region 2 (east), based on residuals. A generalized least-squares procedure was used to account for cross-correlation of sites and develop the final set of equations that have drainage area as the only explanatory variable. The average standard errors of prediction ranged from 18.7 to 46.7 percent in region 1 and 33.4 to 57.6 percent in region 2 for all annual exceedance probabilities (AEPs) examined. The updated statistics showed a greater accuracy of prediction when compared to those from the previous study using drainage area as the only explanatory variable for all AEPs examined in region 1 and the 0.01 and 0.002 AEP flows for region 2. When compared to equations developed in the previous study that have drainage area, mean annual rainfall, and (or) depth-to-rock as explanatory variables, the updated statistics show a greater accuracy of prediction in region 1 at AEP flows of 0.02 and lower (that is, higher flows). Those developed for region 2 do not show a greater accuracy of prediction for any AEP flows when compared to the equations having multiple explanatory variables in the previous study.

The calculated regression equations, basin characteristics, and at-site statistics will be incorporated into the U.S. Geological Survey web application, StreamStats (https://streamstats.usgs.gov/ss/). This application allows users to select a location on a stream, whether gaged or ungaged, to obtain estimates of basin characteristics and flow statistics.

Introduction

Many infrastructure projects, such as those for roads, bridges, and water control structures, must remain unaltered and safe after a specified flood. The design criteria of these structures depend on the magnitude and frequency of floods within a watershed, so reliable estimates of these values are essential. Federal, territory, and local officials also use flood frequency estimates to manage land and water resources and update flood zones and maps for the safety of residents and dwellings. With a 2010 population of 3.7 million residents, Puerto Rico has an average of over 1,000 people per square mile and ranks third in population density among the States

and territories polled in the 2010 census (U.S. Census Bureau, 2019). Storms causing major floods documented by the National Weather Service (NWS) include Hurricane Maria in 2017; Hurricane Georges in 1998; Hurricane Hortense in 1996; Three Kings Flash Floods in 1992; floods of October 6–7, 1985; Mameyes Landslide in 1985; floods of October 5–10, 1970; and floods of September 6, 1960 (NWS, 2019).

Flood frequency studies are typically performed every 10 years; however, the last one for Puerto Rico used streamflow data through 1994 (Ramos-Gines, 1999). This update utilizes data through 2017, incorporating 23 additional years of peak-flow data from U.S. Geological Survey (USGS) streamgages, updated basin characteristics from newer digital geospatial datasets, and the latest statistical procedures in flood frequency analysis using Expected Moments Algorithm (EMA) with a Multiple Grubbs-Beck test (MGBT). Puerto Rico and its streamflow monitoring program have undergone many changes since 1994, and data from new streamgages, as well as additional record at the previous stations, provide improved estimates of flood flows in Puerto Rico.

Purpose and Scope

This report presents updated methods for estimating magnitude and frequency of floods for rural, unregulated streams in Puerto Rico by using streamflow data through September 2017 following procedures outlined in Bulletin 17C (England and others, 2018). Rural conditions are defined in this study as those with less than 10 percent of impervious area included within the drainage area of the streamgage. For purposes of this report, the term "Puerto Rico" refers to the main island of Puerto Rico and does not include any barrier islands, such as Isla de Vieques or Isla de Culebra. Statistical analysis was completed using streamgages with at least 10 years of streamflow record unaffected by tidal fluctuations or regulation at medium to high flows.

This report updates (1) estimates of regional skew using streamgages with 25 or more years of peak-flow record; (2) estimates of the magnitude of floods at the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent chance annual exceedance levels for 91 streamgages in Puerto Rico; (3) basin characteristics computed at gaged locations using geographic information systems (GIS) tools; and (4) regional flood frequency equations (regional regression equations) to predict flood flows using multiple-regression techniques. The resultant equations allow for flow computation at similar ungaged locations at specified annual exceedance levels by using the selected basin characteristics. Results of this study are also incorporated into the USGS StreamStats application (https://streamstats.usgs.gov/ss/), which is a web-based tool that utilizes the statistics, basin characteristics, and predictive equations for streamgages allowing the user to calculate the magnitude and frequency of specified floods at ungaged locations (Ries and others, 2017).

Previous Studies

Lopez and Fields (1970) published the earliest documented USGS investigation of flood frequency for streams in Puerto Rico. They presented techniques to estimate floods of 50-, 20-, 10-, 4-, and 2-percent chance annual exceedance levels and developed equations for ungaged basins by using multiple-regression techniques. Drainage area was used in the equations for all exceedance levels, and mean annual rainfall was included in the 50-percent annual exceedance equation.

Lopez and others (1979) computed flood frequencies at 37 unregulated streamgages with 10 or more years of record at 50-, 10-, 4-, 2-, and 1-percent chance annual exceedance levels. Methods recommended in the U.S. Water Resources Council Bulletin 17A were used, but there was not enough record to develop a regional skew value (U.S. Water Resources Council, 1977). Drainage area and mean annual rainfall were statistically significant basin characteristics at a 95-percent confidence level in the predictive equations for computation of flood flows at ungaged locations.

Ramos-Gines (1999) performed the most recent flood frequency analysis for Puerto Rico using data through September 1994 from 57 rural, unregulated streamgages. The study followed Bulletin 17B procedures by the Interagency Advisory Committee on Water Data (1982) and estimated magnitude and frequency for 50-, 20-, 10-, 4-, 2-, 1-, and 0.2-percent chance annual exceedance levels. Regional regression analysis showed the use of the entire island as one region yielded lower standard errors than those obtained from the use of separate regions. The resultant predictive equations for ungaged locations used the same variables as Lopez and others (1979) for the 50-percent exceedance level, and depth-to-rock was added to drainage area and mean annual rainfall for all other exceedance levels.

Oki and others (2010) computed the magnitude and frequency of floods in Hawai'i using data from 2008 and earlier. Hawai'i is similar to Puerto Rico in that both are mountainous islands affected by warm climates and consist of areas that receive large amounts of rainfall and others that remain relatively dry. Also, Hawai'i has similar stream characteristics, including many streams with small drainage areas. The Hawai'i study divided each island into two regions, and regional regression equations were developed using one or more of the following basin characteristics: drainage area, mean annual rainfall, and maximum 48-hour rainfall that occurs, on average, once in 5 years.

Description of Study Area

The study area includes the main island of Puerto Rico, the easternmost island of the Greater Antilles (fig. 1). The island is bordered by the Atlantic Ocean to the north and the Caribbean Sea to the south. The climate is tropical maritime with average temperatures near 80 °F. The topography of Puerto Rico is very diverse, including mountains in the center of the island, flat lowlands near the coastlines, and karst areas

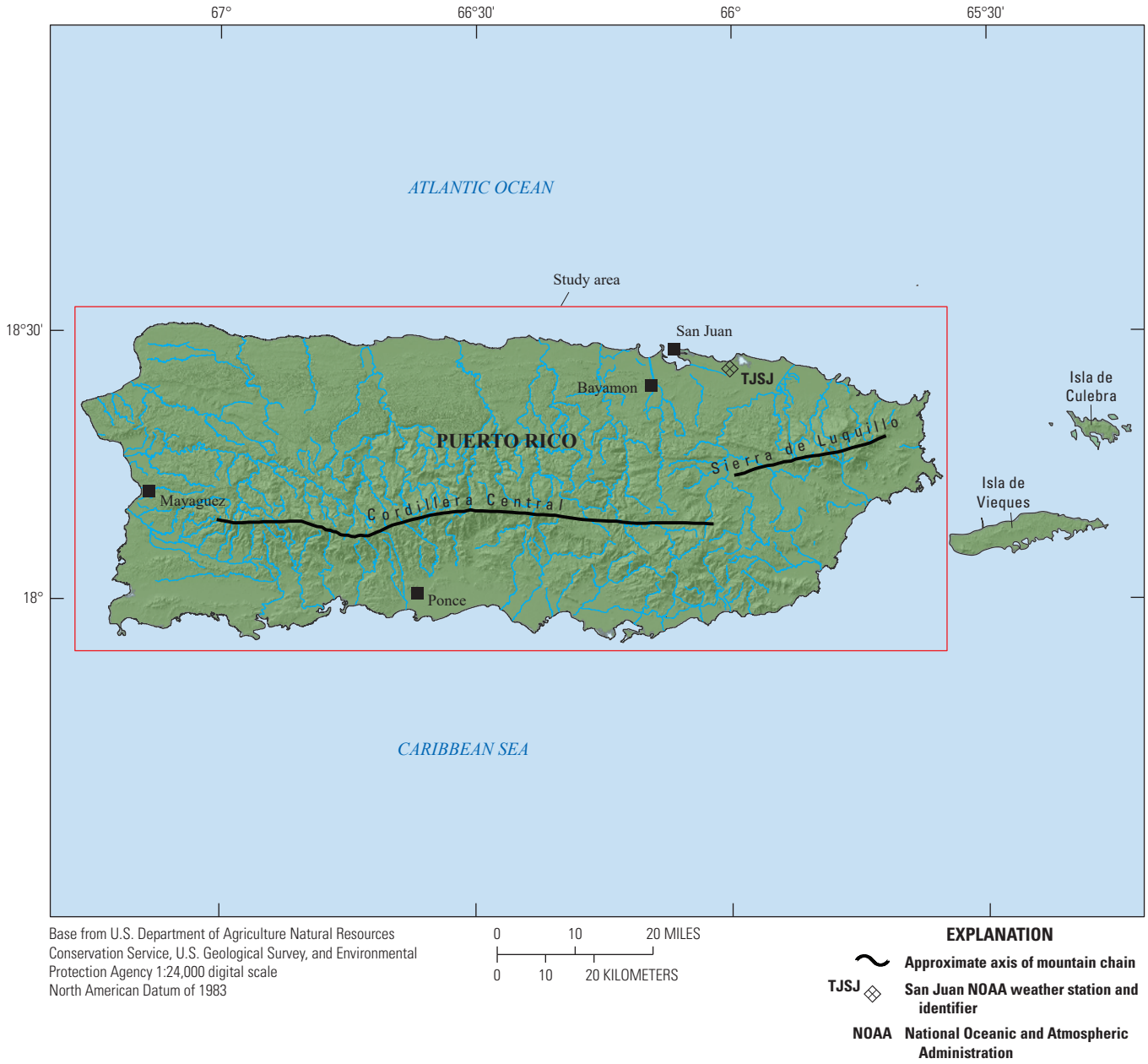

Figure 1. Study area of Puerto Rico.

spread throughout the northern and northwestern parts of the island. Three major physiographic divisions include upland, northern karst, and coastal plains (Monroe, 1976). The primary mountains on the island are those of Cordillera Central, which span east-west near the center of the island and range in peak elevation from approximately 3,000 to 4,000 feet above sea level. This range divides the drainage on the island into two main flow regions, one covering the northern two-thirds and another covering the southern one-third of the island. Streams originating in the steep mountains are carved into deeply incised channels as the water flows downhill into valleys toward the coast. The coastal lowlands are relatively flat areas that extend inland 8–12 miles in the north and 2–8 miles in the south. Considerable overbank flow occurs in these areas during floods because of the gentle slopes and meandering channels. Streams on the southern coast are susceptible to flash flooding because of the short and steep drainages of the Cordillera Central. Tropical rain forests in northeastern Puerto Rico, in the Sierra de Luquillo mountain range (1) have streams that are typically short and steep, (2) lack developed valleys, and (3) have small drainage areas. All of these factors make the rainforests prone to flash flooding.

A detailed map and GIS layer of karst features in Puerto Rico was produced by Alemán González (2010). This karst topography is mainly in the northern and northwestern parts of the main island, where numerous sinks, caves, and underground rivers are located. One of the largest rivers with karst influence flows underground for 6 kilometers until it emerges into a narrow, meandering gorge with nearly vertical walls 100 meters high (Monroe, 1976).

Rainfall in Puerto Rico varies throughout the island because of many factors, some of which include the mountainous terrain, prevailing winds, and hurricanes. Consistent easterly trade winds help create the humid climate on the north side of the island and semi-arid climate on the south side. The average annual rainfall for 1965–2018 in San Juan is 56 inches, as measured at the National Oceanic and Atmospheric Administration (NOAA) weather station on the northern side of the island (NOAA, 2019). In Puerto Rico, the wettest area is the Sierra de Luquillo Range, which receives about 180 inches per year, and the driest area is typically the southwestern corner of the island, which receives about 30–35 inches per year (NOAA, 2019). A map of the mean annual rainfall from Parameter-elevation Regressions on Independent Slopes Model (PRISM) data used for this study is shown in figure 2. The typical dry season is from December 1 to May 31; however, a cold front mixed with tropical moisture can cause considerable rainfall, even in the winter months. The typical wet season, as well as hurricane season, is from June 1 to November 30. Most major flooding during this time results from hurricanes or other tropical systems, but localized rainfall from smaller storms can also cause flooding because of the geographic features discussed earlier.

Data Compilation

The computation of flood-frequency estimates for rural streams requires the selection of relevant streamgages with 10 or more years of annual peak-flow record, excluding those gages affected by urbanization, regulation, or trends. The peak-flow data are reviewed for quality assurance and quality control (QA/QC) to ensure the analysis uses accurate data. After the screening process, the physical and climatic basin characteristics for each streamgage are then compiled and used in the regression analysis.

Peak-Flow Data

Annual peak discharge, also referred to herein as peak flow, is the largest calculated instantaneous streamflow recorded each year at streamgage. The earliest streamflow data collection in Puerto Rico began in 1907 by the Puerto Rico Water Resources Authority but consisted of mostly short-term records (Lopez and Fields, 1970). The USGS constructed a streamgage network in 1958 that expanded from 10 initial stations with peak-flow record in 1960 to 63 in early 1970 and to 119 active and discontinued stations by 2017. Two types of streamgages, continuous-record stations and crest-stage gage (CSG) stations, were used to capture the annual peak flow in Puerto Rico. Discharge is computed continuously at specified intervals at continuous-record stations, whereas only the maximum peak between site visits is documented at crest-stage stations. Because of the nature of a CSG station, it is possible to not record a peak flow if the highest water level between site visits does not reach the gage and remains below the minimum-recordable elevation. The annual peak flow is the maximum instantaneous discharge, or in some cases the maximum daily discharge, during a given water year (October 1 through September 30). Both types of recording stations are susceptible to missing the annual peak flow because of vandalism, flooding or overtopping, and damage from hurricanes, among other reasons. Peak flows that occur during periods of continuous streamgage data collection are called "systematic," and floods that are quantified on the basis of their magnitude during times of no data collection are termed "historic."

All peak-flow data used in this analysis were produced by the USGS, and the peak-flow records are available from the USGS National Water Information System (NWIS) database (U.S. Geological Survey, 2020a). The peak-flow records were reviewed for QA/QC to ensure the analysis was performed on representative data. Puerto Rico streamgages having 10 or more years of annual peak-flow record through the 2017 water year were initially selected for use in this flood frequency analysis. A minimum of 10 years of peak-flow record is recommended for statistical analysis of flood frequency data per Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) and remains the standard for flood frequency investigations per the current guidelines of Bulletin 17C (England and others, 2018). The stations were then screened further to eliminate those whose record was affected by trends, urbanization, or regulation. Streamgages whose record was affected by flow regulation or diversion, having a code 5 or 6 in the peak-flow record (U.S. Geological Survey, 2020b), were evaluated individually to determine the degree of regulation or diversion. Stations whose streams and rivers were regulated at medium and high flows were omitted from this analysis. Some streams and rivers in Puerto Rico are only regulated in a small portion of the drainage area or include small withdrawals, such as for a municipal water supply, that would only affect low flow. Streamgages with such drainage areas were used in the flood frequency analysis because the regulation was not substantial relative to the size of the annual peak flows. For stations whose streams and rivers were regulated at all flows, only the record prior to regulation was used to ensure a homogeneous sample. Streamgages having 10 percent or more of impervious drainage area, which was computed using GISs, were considered urban and omitted from this investigation unless the effects of urbanization were minimal, as indicated by the trend analysis discussed later. Regulation or flow diversion typically attenuates flood flows, and increased urbanization causes a change in the flow regime, both of which invalidate the assumption of random, homogeneous events required to use the Pearson Type III distribution with logarithmic transformation (log-Pearson Type III) of the peak flows.

Hurricanes and tropical storms can alter typical rainfall patterns and influence peak flows at any location on the island. The peak flows affected by hurricanes (code 9 in the peak-flow file) were not separated in this analysis because of

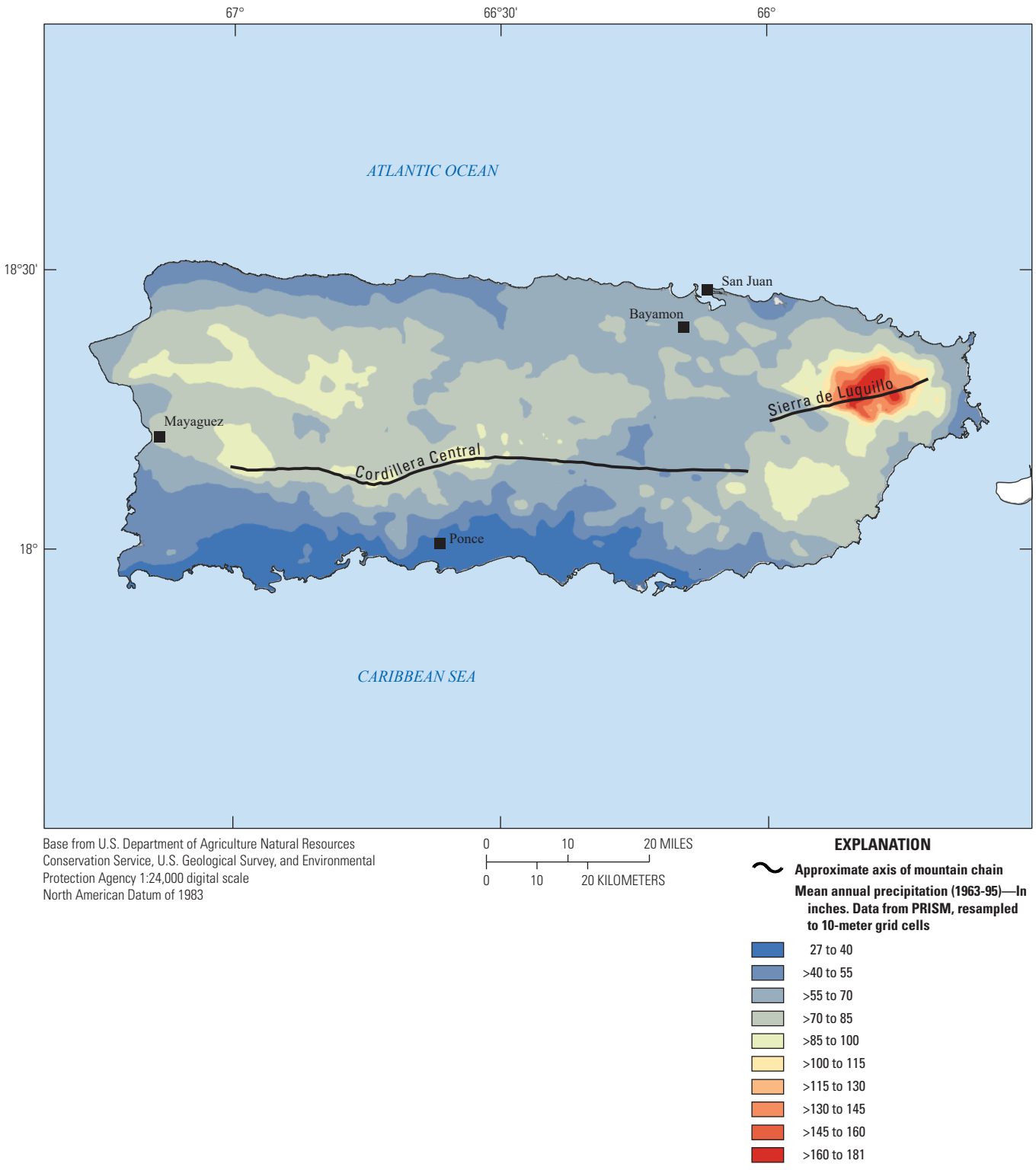

Figure 2. Map showing 1963–95 mean annual rainfall for Puerto Rico. (Data from PRISM Climate Group, Oregon State University, http://prism.oregonstate.edu, created January 1, 2003).

the small number of peaks influenced by these conditions for each streamgage. Depending on the drainage area of a given streamgage, normal rainfall events can cause peak flows even greater than those noted to be influenced by hurricanes.

The significance of trends was analyzed using visual inspection and the Kendall's tau statistic (Helsel and Hirsch, 2002). A significance level of 95 percent (p-value < 0.05) was used to evaluate whether a trend in the data existed. The Kendall's tau statistic is positive for an upward trend and negative for a downward trend, with a better correlation of peak streamflow with time as the statistic approaches ±1. Record length should also be considered, as the presence of a trend may be temporary because of a brief period of higher or lower flow. In such cases, an actual change in the system or conditions cannot be confirmed over a longer period when no data were collected. If a strong trend existed in the annual peak-flow record, that station was excluded from the flood frequency analysis.

From the list of streamgages, data from 91 were initially selected for analysis in this report (fig. 3). Of these, 46 stations have more than 25 years of peak-flow record, and 15 stations have 50 or more years of peak-flow record (appendix 1).

Physical and Climatic Basin Characteristics

The calculation of flood frequency information was performed for streamgages having a minimum of 10 years of annual peak-flow record. However, it was necessary to calculate flood flows for specified annual exceedance probabilities (AEPs) at locations where streamgages were not present. This was accomplished by using regression analysis with physical and climatic basin characteristics as explanatory variables to develop a relation with streamflow statistics. This relation allows the estimation of flood flows in areas that lacked streamgages but had the appropriate basin characteristics within the respective drainage basin boundaries. The basin characteristics and these equations, known as regional regression equations, are stored and accessed within the StreamStats application. StreamStats is a publicly available, map-based web application developed by the USGS that provides estimates of streamflow statistics at user defined locations (Ries and others, 2017).

The basin and climatic characteristics explored in this regional analysis were selected based on a previous flood frequency study for Puerto Rico (Ramos-Gines, 1999) and other factors that typically influence flood flows. Source data for over 30 characteristics for Puerto Rico were compiled from available data for possible use in regression equations. However, only the characteristics deemed relevant from this study through 2017, the previous flood frequency study through 1994 (Ramos-Gines, 1999), and the Puerto Rico low-flow study by Williams-Sether (2021) are currently available for use in StreamStats version 4 (Kolb and Ryan, 2021). The StreamStats abbreviations, measurement units, definitions, methods, and source data for peak-flow basin and climatic characteristics are shown in table 1.

Drainage basin boundaries were generated at gaged locations using ArcGIS version 10.3 with the high definition National Hydrography Dataset and the National Elevation Dataset at 10-meter resolution (Dixon and others, 2021). QA/QC was performed on the original National Hydrography Dataset, where flow direction, connection, and sink discrepancies were resolved and updated within ArcGIS. Basin boundaries were checked by visual assessment and drainage area comparison with those previously calculated and published in NWIS. U.S. Geological Survey (2012) suggests no revision to the NWIS-published drainage area is required if the newly calculated area agrees within 2 percent. Any discrepancies that exceeded these criteria were analyzed individually, if possible. In most cases in Puerto Rico, the NWIS-published drainage areas were calculated manually with a planimeter from older topographic maps and subject to human error. Another source of error in the manual drainage basin calculations was the numerous sinks and karst landforms located on the island, which were digitized in 2010 using GIS (Alemán-González, 2010). Higher resolution data and GIS methods were used to revise and update 27 NWIS-published drainage areas. The drainage basin calculations performed by StreamStats are nearly identical to those used with GIS methods, and no drainage area differences between the two were observed for any streamgages analyzed for this study. The drainage areas for streamgages used in this study range in size from 0.06 to 208 square miles (mi^2).

The same basin characteristics investigated in the previous flood frequency study for Puerto Rico by Ramos-Gines (1999) were also considered in this analysis except soil permeability. Available soil permeability data were spatially incomplete, with large areas of missing data scattered throughout the island, which limited the applicability of StreamStats and the use of the regional regression equations in ungaged basins. Saturated hydraulic conductivity, which has island-wide coverage, was investigated instead. The most recent datasets available were used for all basin characteristics to represent present-day conditions.

Analysis of Flow at Gaged Locations

The AEP is the probability of a given-magnitude flood being equaled or exceeded in any given year. For example, an AEP of 0.01 has a 1-percent chance of being equaled or exceeded in any given year. Historically, these were reported as recurrence intervals, which, in this example, would have been called a "100-year flood." This terminology means there is a 1 in 100 chance of that flood occurring in any given year but is somewhat misleading, and some may falsely suggest it means that such a flood will occur once every 100 years. Floods are random events, however, and two floods of this magnitude could occur within a 2-year timeframe. Table 2 lists the AEPs investigated in this study and the respective recurrence intervals for comparison.

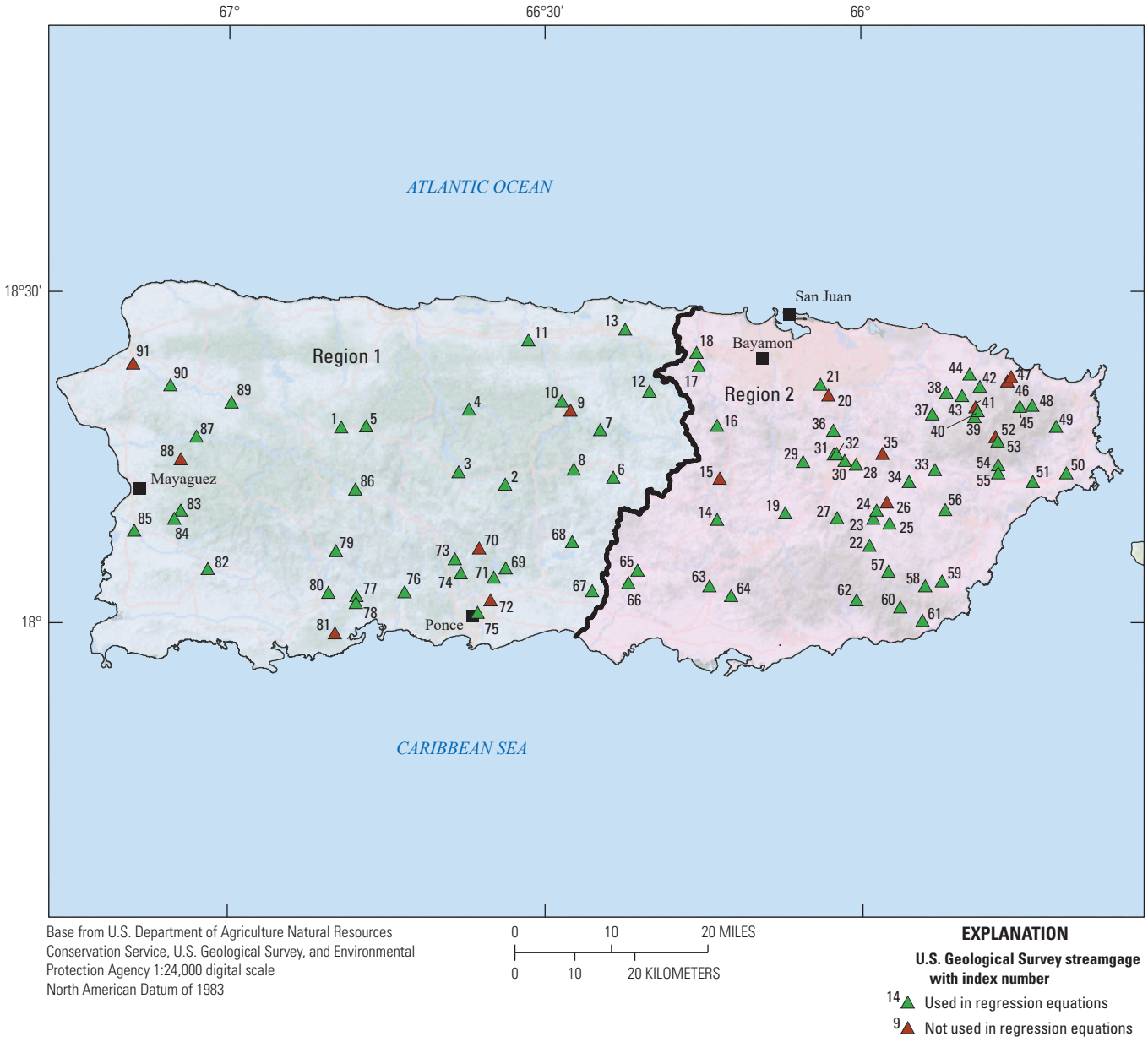

Figure 3. Study area, U.S. Geological Survey (USGS) streamgages initially considered for analysis, and regions used for the regression equations. Index numbers for streamgages are cross-referenced to USGS station identifiers in appendix 1.

The magnitude and frequency of floods at gaged locations are estimated by fitting the annual peak-flow data to a known statistical distribution. The annual peak-flow data observed over the years of data collection, typically by means of a streamgage, is the sample, which is a portion of the population of every peak flow that has ever occurred at a river or stream. Bulletin 17C recommends the use of the log-Pearson Type III distribution for analysis of annual peak-flow data (England and others, 2018). The Pearson Type III distribution is fit to the logarithms of peak-flow record by estimating the three moments of mean, standard deviation, and skew coefficient of the population. The equation used to fit the log-Pearson Type III distribution to annual peak-flow data is

$$log Q_P = \bar{X} + KS, \quad (1)$$

where

Q_P is the P-percent annual exceedance probability flow, in cubic feet per second;

$\bar{X}$ is the mean of the logarithms of the annual peak flows;

K is a factor based on the skew coefficient and the annual exceedance probability; and

S is the standard deviation of the logarithms of the annual peak flows.

Table 1. Basin and climatic characteristics considered for use in the regional regression analysis.

[ALOS, advanced land observing satellite; PALSAR, Phased array type L-band synthetic aperture radar; DEM, digital elevation model; NHD, National Hydrography Dataset; NOAA, National Oceanic and Atmospheric Administration; NLCD, National Land Cover Dataset; NRCS, Natural Resources Conservation Service; PRISM, Parameter-elevation Regressions on Independent Slopes Model]

StreamStats abbreviation	Basin characteristic	Units	Definition	Method	Source(s)
ALPA10BARE	Bare ground	Percent	Percentage of bare ground land-cover within the watershed boundary	Area-weighted average	2010 ALOS/PALSAR
ALPA10FLWD	Flat woodland	Percent	Percentage of flat woodland land-cover within the watershed boundary	Area-weighted average	2010 ALOS/PALSAR
ALPA10FRST	Forest	Percent	Percentage of forest within the watershed boundary	Area-weighted average	2010 ALOS PALSAR Land Cover Dataset
ALPA10FRWT	Forested wetland	Percent	Percentage of forested wetland within the watershed boundary	Area-weighted average	2010 ALOS/PALSAR
ALPA10HERB	Herbaceous	Percent	Percentage of herbaceous land-cover within the watershed boundary	Area-weighted average	2010 ALOS/PALSAR
ALPA10MTWD	Mountain woodland	Percent	Percentage of mountain woodland land-cover within the watershed boundary	Area-weighted average	2010 ALOS/PALSAR
ALPA10URBA	Urban land use	Percent	Percentage of urban land-use within the watershed boundary	Area-weighted average	2010 ALOS/PALSAR
ALPA10WATR	Water	Percent	Percentage of water within the watershed boundary	Area-weighted average	2010 ALOS/PALSAR
BASINPERIM	Basin perimeter	Miles	Perimeter of the drainage basin	Total length	Watershed boundaries as defined in the drainage area method section of this table
BSLDEM10M	Basin slope	Percent	Mean basin slope, based on slope percent grid	Area-weighted average	DEM
CSL10_85fm	Main channel slope	Feet per mile	Change in elevation between points 10 and 85 percent of length along main channel to basin divide, divided by length between points	Single point	DEM
DRNAREA	Drainage area	Square miles	Area that drains to a point on a stream	Sum	StreamStats, version 4
ELEV	Mean basin elevation	Feet	Mean basin elevation	Area-weighted average	DEM
ELEVMAX	Maximum basin elevation	Feet	Maximum elevation value within the watershed boundary	Single point	DEM
I24H(i)Y	24-hour, 1-, 2-, 5-, 10-, 25-, 50-, 100-, or 500-year maximum precipitation	Inches	Maximum 24-hour precipitation that occurs on average once in the specified (i) number of years	Area-weighted average	NOAA Atlas 14 Volume 3
LC01FOREST	Forest	Percent	Percentage of forest within the watershed boundary	Area-weighted average	2001 NLCD classes 41-43

Table 1. Basin and climatic characteristics considered for use in the regional regression analysis.—Continued

[ALOS, advanced land observing satellite; PALSAR, Phased array type L-band synthetic aperture radar; DEM, digital elevation model; NHD, National Hydrography Dataset; NOAA, National Oceanic and Atmospheric Administration; NLCD, National Land Cover Dataset; NRCS, Natural Resources Conservation Service; PRISM, Parameter-elevation Regressions on Independent Slopes Model]

StreamStats abbreviation	Basin characteristic	Units	Definition	Method	Source(s)
LAKESNWI	Lakes and freshwater ponds	Percent	Percentage of lakes and ponds within the watershed boundary	Area-weighted average	2001 National Wetlands Inventory
LC01IMP	Impervious surface area	Percent	Average percentage of impervious land cover within the watershed boundary	Area-weighted average	2001 NLCD Impervious Surface
LFPLENGTH	Main channel length	Miles	Length of longest flow path in a drainage area	Single point	DEM data used to create the watershed boundaries as defined in the drainage area section of this table
MINBELEV	Minimum basin elevation	Feet	Minimum elevation value within the watershed boundary	Single point	DEM
PRECIP	Mean annual precipitation	Inches	Basinwide mean annual precipitation	Area-weighted average	StreamStats, version 4
RCN	Runoff curve number	Dimensionless	Runoff curve number as defined by NRCS (NRCS, 1986)	Area-weighted average	NRCS
RELIEF	Relief	Feet	Maximum − minimum elevation	Single point	DEM
RELIEFRAT	Relief ratio	Dimensionless	(Mean − Min) / (Max − Min)	Single point	DEM
ROCKDEP	Depth to rock	Feet	Minimum depth to bedrock	Area-weighted average	StreamStats, version 4
RUGGED	Ruggedness number	Feet per mile	Stream density times basin relief	Area-weighted average	NHD and DEM
SLOPERAT	Slope ratio	Dimensionless	Main channel slope divided by basin slope	Single point	DEM
SSURGODEP	Soil depth	Feet	Average soil depth	Area-weighted average	NRCS
SSURGO(i)	Hydrologic soil group	Percent	Percentage of soil group A, B, C, or D within the watershed boundary	Single point	NRCS
SSURGOKSAT	Saturated hydraulic conductivity	Micrometers per second	Saturated hydraulic conductivity	Area-weighted average	NRCS
STRMTOT	Total stream length	Miles	Total length of mapped streams in basin	Sum	DEM data used to create the watershed boundaries as defined in the drainage area section of this table
STRDEN	Stream density	Miles per square mile	Total length of all streams in a basin divided by drainage area	Area-weighted average	NHD

Table 2. Annual exceedance probabilities with corresponding recurrence intervals.

Annual exceedance probability	P-percent chance exceedance	T-year recurrence interval (years)	Probability of occurrence in any given year
0.5	50	2	1 in 2
0.2	20	5	1 in 5
0.1	10	10	1 in 10
0.04	4	25	1 in 25
0.02	2	50	1 in 50
0.01	1	100	1 in 100
0.005	0.5	200	1 in 200
0.002	0.2	500	1 in 500

The USGS peak-flow analysis program (PeakFQ) version 7.3 was used in this analysis to estimate the moments of the log-Pearson type III distribution, screen low flow outliers using the MGBT, and compute AEPs (Veilleux and others, 2014). The estimations of the parameters of the log-Pearson type III distribution are performed using the EMA, which is a generalized method of moments procedure (England and others, 2018). PeakFQ allows the use of flow intervals and perception thresholds to incorporate the most information possible about the peak-flow record.

Flow Intervals and Perception Thresholds

Flow intervals and perception thresholds are defined in PeakFQ for every year of peak-flow record. The guidelines outlined in Bulletin 17C using the EMA procedure improves over those previously used by accommodating interval peak-flow data that are based on observations, written records, or physical evidence. The use of intervals more accurately describes estimated historical floods, censored observations, and estimates having large uncertainty (England and others, 2018). For most peak flows within the systematic period of record, the default lower and upper bounds of the 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. The lower and upper bounds of the interval may be set to a range of flows to account for increased uncertainty in the peak-flow value. These intervals may be determined from any relevant NWIS code present (U.S. Geological Survey, 2020b), other annual peak flows, or any streamflow measurement information available.

Perception thresholds (lower and upper) show the range of flows a streamgage could possibly record if they occurred. For a typical continuously recording streamgage, the perception threshold is usually zero to infinity. At other streamgages such as a CSG, however, no information is recorded unless water reaches a minimum level. The lower perception threshold for this streamgage would be that minimum recordable streamflow. For this study, the minimum recordable streamflow was used as the lower perception threshold when available. If no information was available about this level, then the lowest flow associated with the annual peak was used, if no major changes (such as CSG movement or datum changes) were noted, and the upper threshold of infinity was used.

Some peak-flow records include historic peaks that were measured outside of years with systematic record because of the magnitude of the flood. If ungaged years are present between these historic peaks and the systematic record, a value that would be measured if it had occurred (typically equal to the historic peak-flow value) is used for the lower perception threshold, and infinity is used for the upper perception threshold. The flow interval, perception threshold, and peak-flow code information for all sites used in this study can be accessed in appendix 1. If a flow interval was not specified, the default interval assigned by PeakFQ was used.

Potentially Influential Low Flows

In a flood frequency analysis, it is important to have a good fit of the log-Pearson type III distribution to substantial flood and near-flood events (low AEP flows), but sometimes the fit at these higher peak flows departs from that of low-magnitude peak flows. The MGBT identifies these multiple low outliers, referred to as potentially influential low flows (PILFs), which can have a large impact on the fitted frequency curve if used in the analysis (Cohn and others, 2013). These low outliers may be caused by different processes than those that produce larger floods (England and others, 2018). The use of this test provides the ability to censor these PILFs, which improves the fit of the frequency distribution to the higher peak flows. Bulletin 17C recommends the use of MGBT for flood frequency studies with investigation into PILFs and low outlier thresholds based on hydrologic considerations, knowledge of the watershed, and site characteristics (England and others, 2018). Results of the MGBT along with the number of censored peaks are presented in appendix 1.

Test for Redundancy

The streamgages were screened to evaluate those having drainage areas nested inside each other, referred to as redundant stations. This results in cross correlation, which denies the model assumption of independent observations because multiple stations record the same peak flows. Two streamgages are considered redundant if the standardized distance (*SD*) is less than or equal to 0.5 and the drainage area ratio (*DAR*) is less than or equal to 5 (Veilleux, 2011), as calculated in equations (2) and (3):

$$SD = \frac{D_{ij}}{\sqrt{0.5*\left(DA_i + DA_j\right)}} \text{ , and} \quad (2)$$

$$DAR = MAX\left[\frac{DA_i}{DA_j}, \frac{DA_j}{DA_i}\right] , \quad (3)$$

where

SD is the standardized distance between two basins;
D_{ij} is the distance between centroids of basin *i* and basin *j*;
DA_i is the drainage area of streamgage *i*;
DA_j is the drainage area of streamgage *j*; and
DAR is the maximum (MAX) ratio of two drainage basins.

This screening procedure was performed for both the regional skew analysis and the regional regression analysis. Ten out of the 91 possible streamgages were removed from consideration in the regression because of redundancy.

Regional Skew Analysis

This section presents a general overview of skew and criteria used for selection of sites in the regional skew analysis. More details about the regional skew regression analysis, including the methodology and calculations, are located in appendix 2. The skew coefficient is a measure of the asymmetry of the probability distribution of a set of annual peak-flow values. High outliers generally produce positive skew coefficients, and low outliers generally produce negative skew coefficients. A skew coefficient is first calculated by using annual peak-flow data from each of the gaged locations, typically referred to as the station skew or at-site skew. Station skew is sensitive to extreme events and may not be representative of a true population skew for short peak-flow record lengths, so a more detailed analysis is warranted (Griffis and Stedinger, 2007).

The accuracy of the skew coefficient estimate can be improved by using the regional skew to weight with the station skew. A generalized (or regional) skew is produced by using the station skew from stations with longer peak-flow record lengths of typically 25 years or more. Bulletin 17C recommends using a Bayesian generalized least-squares regression model, which removes cross correlation, to calculate a generalized skew coefficient (England and others, 2018). After a generalized skew coefficient is computed, a weighted average of the station skew and generalized skew is calculated and used for the station in the log-Pearson type III analysis.

For the current study, a generalized skew was calculated using the filtered gaged locations having at least 25 years of peak-flow data. These locations were screened further by the removal of redundant streamgages, as decided by means of equations 2 and 3. Using these criteria, four sites were removed from consideration in the calculation of generalized skew. The decision for removal was based on record length and date range of the period of record. Record overlap occurred in all four nested basins, and the range of years in the peak-flow record for the sites used covered at least the same range as the redundant streamgage. The Bayesian model was used with the remaining 42 sites to calculate a constant skew of 0.28 for the entire island, with a mean squared error (MSE) of 0.20 (appendix 2). Two separate regions were investigated as found by Ramos-Gines (1999), but the addition of the region variable to the constant model offered little explanation of the variance in the at-site skew and was not used.

At-Site Flood Frequency Statistics

The final flood-flow estimates for all AEPs considered incorporate the EMA, the MGBT, and the new generalized skew for each streamgage using annual peak-flow data through 2017. The perception thresholds, flow intervals, and MGBT low thresholds and outliers are included in appendix 1. The input and output files from PeakFQ used to calculate the estimates are available in a USGS data release by Ryan and Hazelbaker (2021). The resultant at-site flood-flow estimates, as well as prediction intervals at the 95-percent confidence limit, are available in appendix 3.

Estimating Flood Frequency Statistics at Ungaged Locations

Data collected at streamgages quantify streamflow during the period of record, which is used to estimate a range of flood magnitudes and frequencies. Although the number of streamgages in Puerto Rico is limited, discharge magnitude and frequency can still be estimated for ungaged streams on the island. Regionalization techniques use streamflow data and basin characteristics from a selection of streamgages within a hydrologic region to estimate the magnitude and frequency of floods at other areas within that region where a streamgage is not present.

The magnitude and frequency of floods at ungaged locations were estimated by using regional regression equations computed from a relationship between physical and climatic basin characteristics and the previously calculated peak flows of specified AEPs at gaged locations. Multiple basin characteristics were considered in this analysis, including those investigated in the previous flood frequency study of Puerto Rico by Ramos-Gines (1999). Calculation of these equations required exploratory data analysis, definition of regions, and final analysis using weighted regression techniques.

Exploratory Data Analysis

The first part of the regional regression analysis was performed using ordinary least-squares (OLS) regression techniques to delineate the hydrologic regions and remove correlated explanatory variables from consideration. Selection of the explanatory variables to explore further with OLS was based on all-possible-subsets regression methods (Neter and others, 1985). The goal of a least-squares regression is to minimize the sum of the squared residuals while providing a consistent, reproduceable analysis. Major assumptions of the OLS regression are (1) a linear relation exists between the response (estimated peak-flow statistics at specified AEPs) and explanatory variables (basin characteristics); (2) homoscedasticity (variance of the residuals is constant over the range of explanatory variables); and (3) independent and normally distributed residuals. All observations are weighted equally in the OLS regression, as it assumes an equal uncertainty associated with each of the observations (Farmer and others, 2019).

In the exploratory analysis, peak flows and basin characteristics were log-transformed where applicable to achieve a linear relation. Where some of the basin characteristic values computed as zero, the characteristic was not log-transformed to avoid an error when taking the logarithm of zero. The analysis was performed using the R programming language (R Core Team, 2020) and screening procedures described by Helsel and Hirsch (2002), which produced statistical diagnostics, comparisons, and multiple plots of peak flows versus basin characteristics. These scatterplots were examined to evaluate the linear fit of explanatory variables to the response variable. The selection of relevant explanatory variables was made on the basis of several measures from the OLS regressions, including the coefficient of determination (R^2), adjusted coefficient of determination (Radj^2), standard error of the estimate, mean squared error (MSE), root mean squared error, residuals, multicollinearity, Cook's D statistic, and professional judgement of the coefficient (Farmer and others, 2019).

Multicollinearity occurs when one explanatory variable is closely related to one or more other explanatory variables, and hence, they are not independent variables. This was investigated using the output plots of each variable and a calculated variance inflation factor (VIF), as shown in equation 4.

$$VIF_j = 1/(1 - R_j^2) \quad (4)$$

where

VIF_j is the variance inflation factor of the jth explanatory variable; and

R_j^2 is the R^2 from a regression of the jth explanatory variable on all other explanatory variables.

There was no exact value of VIF above which the data were automatically excluded, but values below 2 were preferred, and values above 10 were heavily scrutinized, as serious problems can occur when using those explanatory variables in the regression (Helsel and Hirsch, 2002). After removal of highly correlated variables via visual inspection of the plots, all calculated VIFs were below 2.

A point with high leverage has the capability to highly influence the regression line because of the large distance between that point and the mean of the other values (Helsel and Hirsch, 2002). Data plots were examined to see whether points with high leverage also highly influenced the regression line, compared to what would be predicted if the point were absent. Cook's D statistic was also used as a measure of influence, and highly influential values were scrutinized to find the cause of influence.

The exploratory analysis was originally performed with 81 streamgages (10 stations were removed from equation model development because of redundancy) at 10- and 1-percent chance exceedance flows. Two streamgages were excluded after showing on multiple plots as outliers having large residuals with high leverage and high influence. Upon investigation, both drainage areas were extremely small (less than 0.3 mi^2), which made the equations inaccurate for flow prediction because of the low number of sites with extremely small drainage areas in this analysis. Depth-to-rock, which was chosen as one of the explanatory variables in the previous study by Ramos-Gines (1999), did not appear significant in the regression analysis (p-value < 0.05) after the two sites with small drainage areas were removed.

The percentage of soil type C, forested wetland, water, and lakes and freshwater ponds each showed a minimal increase in predictive ability as a potential variable as other variables were removed, but they were removed from further consideration for various reasons. The classification of soil type C has a moderately high runoff potential and slow infiltration rate, which should increase streamflow, but the negative coefficient in the equation produces the opposite result. The percentage of forested wetland also has a negative coefficient and a very small range of values (0.00–1.76 percent) for the sites considered, with all but one site having values below 0.3 percent. The percentages of water land use and lakes and freshwater ponds also have small ranges of values and multiple sites at 0.00 percent (49 for water land use and 20 for lakes and freshwater ponds). A small range of values and large number of zero values hinder the ability of the variable to explain a large range of resultant peak flows when used in the equation, and the small increase in performance metrics does not warrant their inclusion in the regional regression analysis.

Of the multiple rainfall parameters investigated, the 24-hour, 5-year intensity was chosen because it appeared to be significant (p-value < 0.05) for both AEP flows with the lowest standard error when compared to the other intensities investigated. When used separately, however, the 24-hour, 10-, 25-, and 100- year rainfall intensities each appeared significant with a small increase in standard error. The mean annual rainfall also appeared to be significant, but it was not chosen as a possible explanatory variable because it did not improve the standard error over the 5-year intensity, which has better data resolution in Puerto Rico. A common recommendation of regionalization studies is to use only 1 basin characteristic per 10 sites (Farmer and others, 2019). Thus, results of the exploratory OLS analysis reduced the regression model to a maximum of three explanatory variables in which drainage area, 24-hour, 5-year rainfall intensity (both variables log-transformed), and Natural Resources Conservation Service (NRCS) runoff curve number were selected as significant explanatory variables for all AEPs for the remaining 79 sites, which were subsequently used in the regionalization analysis.

Definition of Regions

Regional regression analysis in the previous flood frequency study by Ramos-Gines (1999) concluded the best results for the regional regression equations were obtained using the entire island of Puerto Rico as a single region. This approach was investigated for the current study, and residuals from the OLS regression for the 10- and 1-percent chance exceedance flow estimates using all combinations of drainage area, 24-hour, 5-year rainfall intensity (both variables log-transformed), and NRCS runoff curve number as explanatory variables were plotted at the centroid of the drainage area of each streamgage. This resulted in a cluster of positive residuals for streamgages in the eastern portion of the island and mostly negative residuals in the western portion for all combinations of explanatory variables used. No U.S. Environmental Protection Agency ecoregion information was available for this study, and the trend in residuals did not follow mountain ranges or the divides identified by the three physiographic divisions on the island (Monroe, 1976).

The island was divided into two separate regions to minimize and remove trends in regression residuals. The region division that resulted in lower and more balanced residuals runs primarily north-south near the center of the island, as shown in figure 3, mostly along an 8-digit hydrologic unit code (HUC8) boundary. In the center of the island, this boundary closely follows that of the eastern and western interior NWS climate divisions in Puerto Rico (NWS, 2020). The division line runs through a HUC8 polygon on the southern end of the island, but care was taken to include entire watersheds and consideration was given where hydrologic and physiographic properties differed. Figure 4 shows all 91 streamgages initially considered for the study, the 3 physiographic regions, NWS climate divisions, and HUC8 boundaries. This region division resulted in 33 sites in the western region and 46 sites in the eastern region that were used for initial regression equation development. The use of two regions resulted in more geographically balanced residuals when compared to the one region model, even after a repeated analysis using drainage area as the only explanatory variable and the removal of two more sites from the final regression analysis, which is discussed later.

Generalized Least-Squares Regression Analysis

Generalized least-squares (GLS) techniques were used with the results from the exploratory data analysis and definition of regions to calculate final regression equations. Regressions for each region that use the refined list of basin characteristics containing drainage area, rainfall intensity, and NRCS runoff curve number were further analyzed by using version 3.0 of the USGS Weighted Multiple Linear Regression (WREG) package (Farmer, 2019) written in the R statistical language (R Core Team, 2020). The major advantages of the GLS approach over OLS are (1) assigned weights based on uncertainty of the observations (that is, record length and variance); and (2) correlated streamflows and time-sampling errors are accounted for (Farmer and others, 2019).

Cross-correlation occurs when streamgages are located near each other and have overlapping periods of record. Time-sampling errors result from the variation in record length among stations. Parameters of a correlation smoothing function (alpha, α, and theta, θ) are estimated from visual inspection of a plot that relates the correlation between streamflow at two sites to the geographic distance between them. More information about the smoothing function and estimation of the parameters, the previous stand-alone version of WREG, inputs, equations, and diagnostic outputs and performance metrics is available by Eng and others (2009). The GLS technique is described in more detail in Stedinger and Tasker (1985) and Farmer and others (2019).

Metrics used to evaluate the regression model fit are the same as those explained in the exploratory data analysis with the addition of pseudo coefficient of determination (pseudo-R^2), average variance of prediction (AVP, in log units), average standard error of prediction (SEP, in percent), model error variance (MEV, in log units), and standard model error variance (S_{MEV}, in percent). The pseudo-R^2 differs from R^2 and R_{adj}^2 in that the effect of the time-sampling error has been removed (Eng and others, 2009). The pseudo-R^2 shows how well the regression model fits the peak flows from gaged locations used to create the model, with a better fit as the number approaches 1. The AVP is the arithmetic average of all variances of prediction for each streamgage used in the regression. The SEP is the square root of the AVP and is represented in WREG as a percentage of the total error for the regression model. The SEP provides an estimate of reliability of the predicted peak flows. The AVP and SEP were used to evaluate how well each regression model predicted peak flows at ungaged locations.

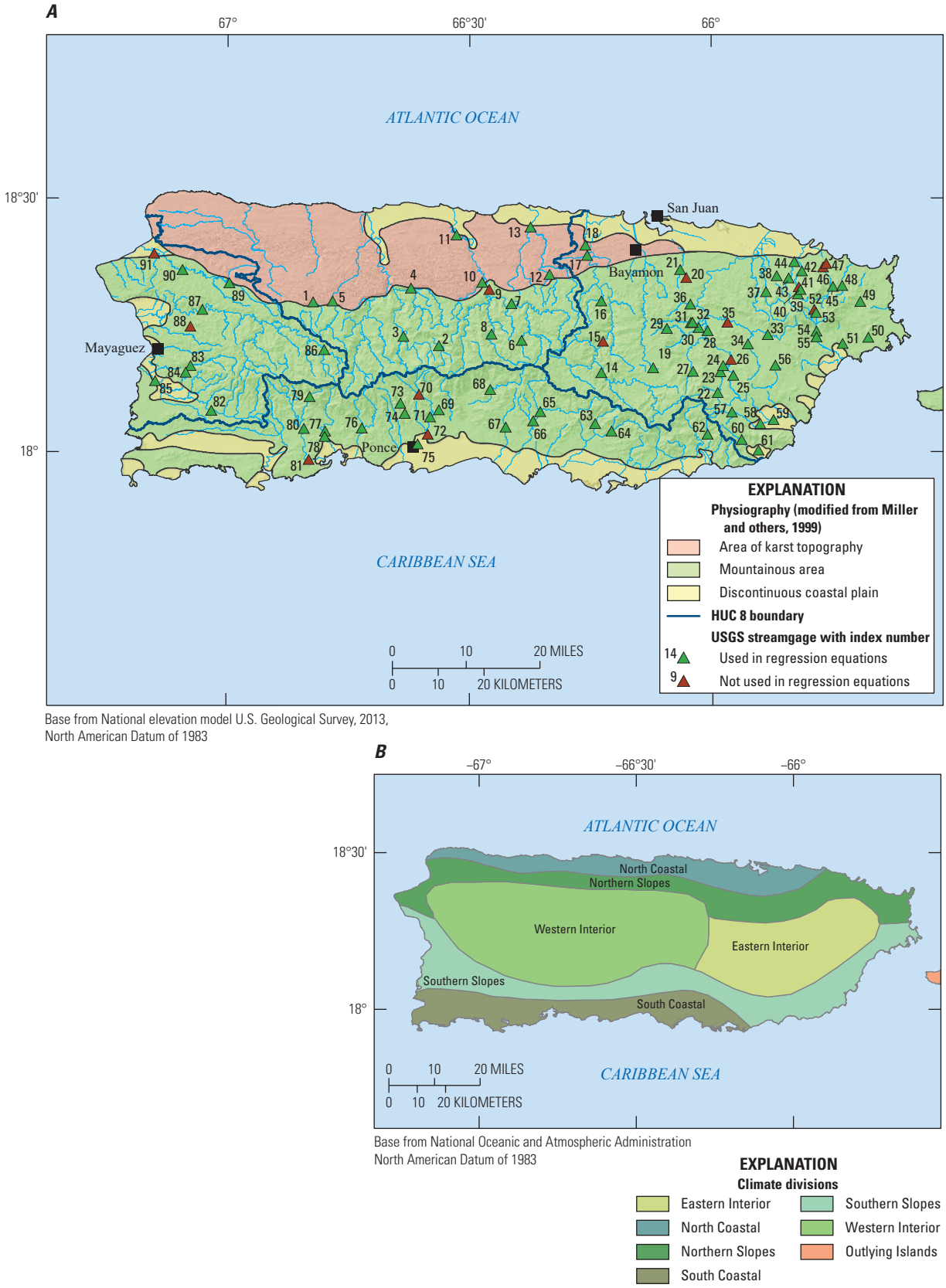

Figure 4. *A*, physiographic regions, U.S. Geological Survey (USGS) streamgages, and 8-digit hydrologic unit code (HUC8) boundaries, and *B*, climate divisions of Puerto Rico (Miller and others, 1999). Index numbers for streamgages are cross-referenced to USGS station identifiers in appendix 1.

The resultant regression model takes the form of equation 5, and transformation out of log space takes the form of equation 6:

$$\log Q_{i,AEP} = \log K + a_1 \log x_1 + a_2 \log x_2 + \ldots a_P \log x_P \text{, and} \quad (5)$$

$$Q_{AEP} = K' x_1^{a_1} x_2^{a_2} \ldots x_p^{a_p}, \quad (6)$$

where

$Q_{i,AEP}$ is the peak flow, in cubic feet per second, of region *i* with a specified AEP;
K is the regression constant;
a_1 through a_P are regression coefficients;
x_1 through x_P are values of the explanatory variables (basin characteristics);
p is the total number of explanatory variables (basin characteristics); and
K' is the antilog (10^K) of the regression constant.

The GLS analysis using WREG was initially performed on the remaining 79 sites (33 in region 1 and 46 in region 2, as shown in fig. 2) by using drainage area, NRCS runoff curve number, and 24-hour, 5-, 10-, 25-, 50-, and 100-year rainfall intensities; only one rainfall intensity was selected per model run. Multiple rainfall intensities were considered again in WREG to investigate the model fit for a different intensity by using the weighting techniques compared to the 24-hour, 5-year rainfall intensity chosen in the exploratory analysis, because there was minimal increase in the standard error based on the intensity used. No default or previous values were found for Puerto Rico for parameters of the correlation smoothing function used in the GLS analysis, so they were selected on the basis of visual interpretation of the plot of sample correlation versus geographic distance for each region. The plots showed correlation, which confirmed the use of GLS techniques, with final values for region 1 of α equal to 0.006 and θ equal to 0.96 and region 2 of α equal to 0.004 and θ equal to 0.94. The NRCS runoff curve number added no benefit to the regression for either region and thus was removed from further consideration.

Region 1 GLS

The region 1 analysis in WREG began with 33 sites by using drainage area and each of the selected rainfall intensities in separate model runs for all AEPs. In all cases, the 24-hour, 5-year rainfall intensity produced the lowest AVP and MEV and the highest R_{adj}^2 and pseudo-R^2 when compared to the results obtained using drainage area and each of the other rainfall intensities considered. Leverage and influence plots showed extremely high values at one site, 50145395 (index no. 88, fig. 2 and appendix 3). Upon further investigation, the peak-flow record was found to be only 10 years long, and the peak streamflows had little to no verification, so the site was removed from the regression analysis because of the short and unreliable record. Removal of this site increased the fit of the model for all metrics calculated in the analysis for all AEPs.

The final analysis of region 1 included 32 sites and used only drainage area as the explanatory variable. The drainage areas ranged from 3.42 to 165 mi^2. The 24-hour, 5-year rainfall intensity was statistically significant (p-value < 0.05) for AEPs of 0.02 and less (2-percent chance exceedance and larger peak flows). However, the addition of intensity did not reduce the SEP by more than 3 percent when compared to the SEP calculated by using only drainage area for any AEP and showed a minimal decrease in S_{MEV}. Therefore, the rainfall intensity was not used to calculate peak flows for region 1. The coefficients for all AEPs are shown in table 3, and a plot of peak flows calculated from the regression equation to the at-site log-Pearson Type III flows for the 0.01 AEP is shown in figure 5. The plot shows a good fit of the peak flows to the one-to-one line with no noticeable pattern and low residuals overall.

Region 2 GLS

The region 2 analysis in WREG began with 46 sites by using drainage area and each of the selected rainfall intensities in separate model runs for all AEPs. In all cases, the 24-hour, 5- and 25-year rainfall intensity produced the same or very similar results, including the lowest AVP and MEV and the highest R_{adj}^2 and pseudo-R^2 when compared to the results obtained using drainage area and each of the other rainfall intensities considered. For consistency with region 1, the 24-hour, 5-year rainfall intensity was investigated as a second explanatory variable. Performance metrics showed one streamgage, 50048690 (index no. 20, fig. 2 and appendix 3), had high leverage and much higher influence than all other sites in this region. Further investigation shows this location is downstream from a fixed outflow on Lago Las Curias, which suppresses the peak flows. Therefore, this site was removed from the regression analysis. Removal of this site increased the fit of the model for all metrics calculated in the analysis for all AEPs.

Three USGS streamgages used in the analysis had an impervious area greater than the maximum 10-percent criteria, namely 50048770, 50055225, and 50064700 (index nos. 21, 30, 44, respectively, fig. 2 and appendix 1). The analysis was repeated with these stations removed, and it was determined that these stations help improve model fit, so they were retained for this study. Streamgages 50048770 (14 percent impervious area) and 50055225 (12 percent impervious area) were just above the criteria, and 50064700 (33 percent impervious area) only included peak-flow data through 1982, so it is likely that the impervious development in the relatively small drainage basin occurred after the period of record ended. The log-Pearson Type III distribution plots look reasonable, and the residuals for each of these three sites were not indicative of being affected by urbanization.

Table 3. Regression coefficients from weighted multiple linear regression analysis for specified annual exceedance probabilities (AEPs) in Puerto Rico.

[AVP, average variance of prediction; SEP, average standard error of prediction; MEV, model error variance; SEM, standard error of model; R^2, coefficient of determination; $t_{(\alpha/2, n-p)}$, 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; DA, drainage area of basin; mi^2, square mile ; <, less than]

AEP	Regional regression equation	AVP (log units)	SEP (percent)	MEV (log units)	SEM (percent)	R^2 (percent)	Pseudo-R^2 (percent)	$t_{(\alpha/2, n-p)}$
Region 1; n = 32; 3.42 mi^2 < DA < 165 mi^2								
0.5	$Q_{0.5} = 10^{2.87}\ DA^{0.59}$	0.0373	46.7	0.0340	44.4	60.8	62.7	2.042
0.2	$Q_{0.2} = 10^{3.09}\ DA^{0.66}$	0.0217	34.9	0.0189	32.5	74.7	78.8	2.042
0.1	$Q_{0.1} = 10^{3.21}\ DA^{0.71}$	0.0160	29.8	0.0131	26.8	79.2	85.5	2.042
0.04	$Q_{0.04} = 10^{3.33}\ DA^{0.77}$	0.0120	25.7	0.0087	21.7	81.2	90.8	2.042
0.02	$Q_{0.02} = 10^{3.40}\ DA^{0.80}$	0.0105	24.0	0.0066	18.9	81.1	93.2	2.042
0.01	$Q_{0.01} = 10^{3.46}\ DA^{0.83}$	0.0090	22.2	0.0045	15.6	79.8	95.5	2.042
0.005	$Q_{0.005} = 10^{3.52}\ DA^{0.85}$	0.0076	20.2	0.0023	11.1	77.3	97.8	2.042
0.002	$Q_{0.002} = 10^{3.60}\ DA^{0.88}$	0.0065	18.7	0.0001	0.1	72.7	99.9	2.042
Region 2; n = 45; 0.42 mi^2 < DA < 208 mi^2								
0.5	$Q_{0.5} = 10^{3.02}\ DA^{0.55}$	0.0539	57.5	0.0506	55.5	55.6	60.0	2.017
0.2	$Q_{0.2} = 10^{3.23}\ DA^{0.64}$	0.0255	38.1	0.0232	36.1	74.2	81.2	2.017
0.1	$Q_{0.1} = 10^{3.34}\ DA^{0.69}$	0.0202	33.6	0.0177	31.4	79.0	86.5	2.017
0.04	$Q_{0.04} = 10^{3.47}\ DA^{0.74}$	0.0199	33.4	0.0168	30.5	79.9	88.3	2.017
0.02	$Q_{0.02} = 10^{3.57}\ DA^{0.76}$	0.0212	34.5	0.0175	31.2	78.6	88.7	2.017
0.01	$Q_{0.01} = 10^{3.65}\ DA^{0.78}$	0.0230	36.0	0.0186	32.2	76.5	88.8	2.017
0.005	$Q_{0.005} = 10^{3.73}\ DA^{0.80}$	0.0249	37.5	0.0196	33.1	74.0	88.9	2.017
0.002	$Q_{0.002} = 10^{3.83}\ DA^{0.82}$	0.0270	39.3	0.0207	34.0	70.6	89.1	2.017

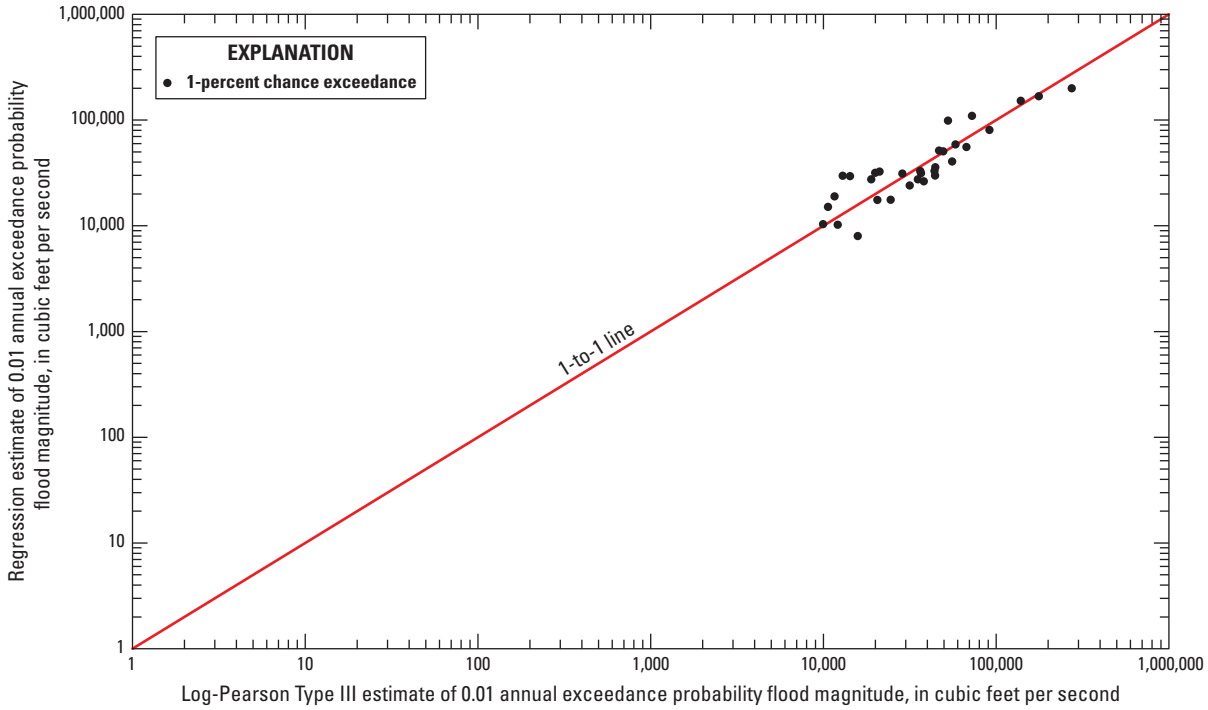

Figure 5. Region 1 plot showing 1-percent chance exceedance flows calculated from regression equations versus the at-site flows from log-Pearson type III distribution.

The final analysis of region 2 included 45 sites and used only drainage area as the explanatory variable. The drainage areas ranged from 0.42 to 208 mi^2. The 24-hour, 5-year rainfall intensity was only statistically significant (p-value < 0.05) at large AEPs (low peak flows) of 0.1 and higher. Among the AEPs where the rainfall intensity appeared significant, only the AEPs of 0.2 and 0.5 showed a reduction in SEP of more than 3 percent when the intensity variable was added to the drainage area regression. These flows are relatively low compared to the flood flows for which these equations are primarily intended, so only drainage area was used as an explanatory variable for region 2 to retain the same model form for all statistics in the region. The coefficients for all AEPs are shown in table 3, and a plot of peak flows calculated from the regression equation to the at-site log-Pearson Type III flows for the 0.01 AEP is shown in figure 6. The plot shows more scatter around the one-to-one fit line than the region 1 plot (fig. 4) but also encompasses a larger range of peak flows than that of region 1 and provides a good fit overall.

Example Calculation of Peak Flow Using a Regression Equation

Example 1. Calculate the 1-percent AEP peak flow for USGS streamgage 50056400, Rio Valenciano near Juncos, Puerto Rico (index no. 34, fig. 2 and appendix 3), using the appropriate regional regression equation. The streamgage is located in the eastern portion of the island at lat 18°12′58” N., long 65°55′34” W. The drainage area is 16.5 mi^2, and the basin is considered rural and unaffected by substantial regulation, diversion, or karst influence.

1. Using figure 3 and the latitude and longitude, the streamgage is located in region 2.
2. Using table 3, the 1-percent AEP regional regression equation for region 2 is $Q_{0.01} = 10^{3.65}\ DA^{0.78}$
3. Substituting the drainage area into the region 2 equation for the 1-percent AEP yields

$$Q_{0.01} = 10^{3.65}\ (16.5)^{0.78}$$

$$Q_{0.01} = 4{,}467 * 8.90$$

$$Q_{0.01} = 39{,}800 \text{ cubic feet per second}$$

(39,758 ft^3/s; rounded to three significant figures).

Limitations of Regional Regression Equations

Regional regression equation models are used to compute flood flows at specified AEPs using selected basin characteristics. These statistical models were developed using the most recent basin characteristic data available for Puerto Rico, accessible via StreamStats (Kolb and Ryan, 2021). Use of these equations to compute flood flow statistics outside of

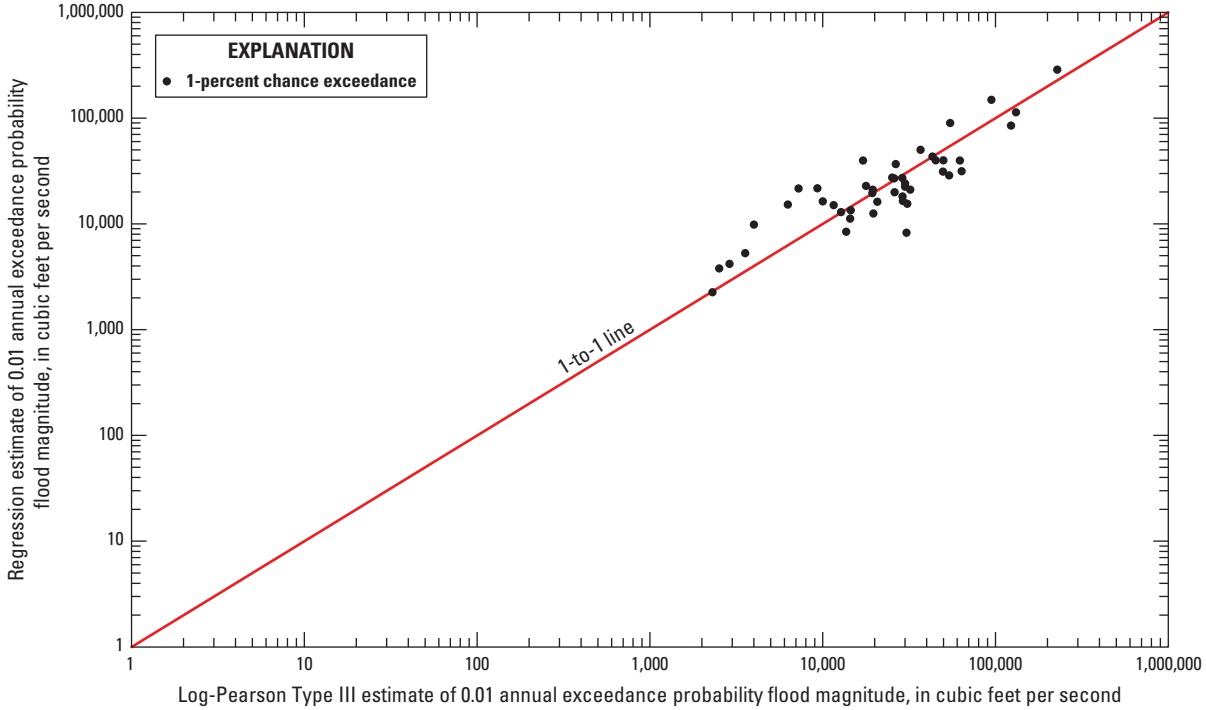

Figure 6. Region 2 plot showing 1-percent chance exceedance flows calculated from regression equations versus the at-site flows from log-Pearson Type III distribution.

the conditions from which they were developed is extremely discouraged, as the limits of statistical confidence are unknown.

The only basin characteristic used in the final regression equation for either region was the contributing drainage area of the basin. The drainage areas for streamgages used in region 1 ranged from 3.42 to 165 mi^2 and in region 2 from 0.42 to 208 mi^2. Streamgages considerably affected by regulation at medium and high flows were not used in model development, nor were any stations used that were affected by tidal influence or substantial urbanization. Therefore, the regression equations should only be applied to naturally flowing, relatively unregulated streams and river systems. The criterion to denote a rural station was set at 10 percent of impervious area within the drainage basin, as denoted by the basin characteristic using data from 2001 National Land Cover Dataset for impervious surfaces (table 1). Three stations used in the regression analysis, all in region 2, did exceed this criterion but showed a reasonable fit to the log-Pearson type III distribution and yielded no major flags in the performance metrics output of WREG. No relation was established between peak flows and impervious area, so the actual impact of urban development on flow statistics is unknown. The effects of karst influence on the peak flows in some areas are unknown, and the drainage areas for some streamgages in Puerto Rico could not be determined because of this uncertainty. The regression equations should not be used in these areas where the effects between groundwater and surface water interactions are unknown. However, some drainage areas within the karst boundaries outlined in figure 4 have been defined and verified, and these stations were used in the regression analysis, if applicable.

Uncertainty of Regional Regression Equations

The resultant regional regression equations, although developed using advanced weighting techniques, have uncertainty because they are statistical models. A meaningful way to display this uncertainty is to specify a desired level of confidence, then calculate a range of peak-flow values within which the actual peak should occur. These prediction intervals describe both the uncertainty in the placement of the regression line and the uncertainty associated with the residuals (Farmer and others, 2019). The intervals should contain approximately $(1 - \alpha)*100$ percent of the data points within them and have $(\alpha/2)*100$ percent of the data both above and below the intervals (Helsel and Hirsch, 2002). For example, a prediction interval computed at the 95-percent significance level ($\alpha = 0.05$) for a given AEP has a 95-percent chance of including the true value of the peak flow within that interval. The interval is computed using the calculated flow, AVP, and Student's t-distribution as shown in equations 7 and 8.

$$PI_{i,\,lower} = 10^{\left[X_{reg\,i} - t_{\left(\frac{\alpha}{2}\right),(n-p)} * (AVP_{reg})^{0.5}\right]}, \text{ and} \quad (7)$$

$$PI_{i,\,upper} = 10^{\left[X_{reg\,i} + t_{\left(\frac{\alpha}{2}\right),(n-p)} * (AVP_{reg})^{0.5}\right]}, \quad (8)$$

where

$PI_{i,lower}$ is the lower prediction interval limit at site *i*;

$PI_{i,upper}$ is the upper prediction interval limit at site *i*;

$X_{reg\,i}$ is the peak-flow estimate (in log base 10 units) using the regional regression equation for a given AEP at site *i*;

$t_{\left(\frac{\alpha}{2}\right),(n-p)}$ is the critical value from the Student's t distribution 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 (values are listed in table 3); and

AVP_{reg} is the AVP (in log base 10 units) from the regression for the specified region and AEP.

The arithmetic AVP is used for the corresponding region when computing intervals at ungaged locations or at sites that were not used to develop the regression equations. The actual variance of prediction for each site is used when calculating the intervals for sites that were used in development of the regression equations, and the values are shown in appendix 3.

The performance metrics from the GLS regression show a better regression model fit and more predictability for region 1 than region 2 for nearly every metric and AEP, as shown in table 3. Region 2 used a larger number of streamgages but also had a larger range in drainage area extent compared to those used from region 1. The SEP for all AEPs ranged from 18.7 to 46.7 percent in region 1 and from 33.4 to 57.5 percent in region 2. For only AEPs of 0.01 and lower (that is, high flows), the SEP ranged from 18.7 to 22.2 percent in region 1 and from 36.0 to 39.3 percent in region 2. Although the model for region 1 showed a better fit than that of region 2, the overall application of both models should provide sound estimates of peak-flow statistics. These SEP values show the expected average accuracy when using the specified regression model to predict peak flows at ungaged stations in the corresponding region (table 3).

Results Comparison With Previous Studies

The previous regionalization study for Puerto Rico by Ramos-Gines (1999) provided two regions for regional skew and two sets of regional regression equations to estimate flood flows. The regional skew in the north region was −0.47 with a standard error of 0.56, and the south region skew was 0.39 with a standard error of 0.71. The regional skew computed in the current study using the Bayesian model and the entire island as one region was 0.28 with a standard error of 0.45. In Ramos-Gines (1999), the preferred equations used drainage area, mean annual rainfall, and depth to rock for all AEPs except 50 percent, which only used drainage area and

mean annual rainfall. For simplicity and data availability, another set of equations that use only drainage area were calculated for all peak flows studied. The comparison of results from those equations to these provided in the current study are shown in table 4 for all AEPs except 0.5-percent, which was not considered in the previous study. The new region 1 equations have smaller SEP and MEV for all AEPs when compared to the previous one-variable model and smaller SEP and MEV at the 2-percent chance exceedance and lower AEP flows (that is, highest flows) when compared to the 2-variable model. The metric parameters for the new region 2 equations were larger for all AEPs when compared to both previous models except the 1-percent chance exceedance and lower AEP flows (that is, highest flows) had a smaller SEP and MEV than the previous 1-variable model.

A comparison plot of the drainage-area-only equation from the previous study to those developed for regions 1 and 2 in the current study for 1-percent chance exceedance flows is shown in figure 7. The equations for the new study show the regression differences and higher peak flow predicted for region 1 for drainage areas greater than 10.8 mi^2 and for region 2 over the entire range of drainage areas used in the equation calculations.

Weighting Streamflow Estimates at Gaging Stations

The estimated at-site peak-flow statistics at unregulated, rural streamgages can be weighted with those calculated using the regional regression equations to reduce uncertainty in the final estimate. Procedures used to weight these estimates are outlined in U.S. Geological Survey (2010) and England and others (2018). The weighting method uses the at-site variance (from PeakFQ) and the variance of the regression model (from WREG; AVP if the site was not used in the regression) using log-transformed values (in log base 10 units) as shown in equations 9 and 10:

$$X_{wtd_{AEP,i}} = \frac{\left(X_{site_i} * V_{reg_i}\right) + \left(X_{reg_i} * V_{site_i}\right)}{V_{site_i} + V_{reg_i}}, \quad (9)$$

$$V_{wtd_{AEP,i}} = \frac{V_{site_i} * V_{reg_i}}{V_{site_i} + V_{reg_i}}, \quad (10)$$

where, in log base 10 units,

$X_{wtd_{AEP,i}}$ is the weighted peak-flow estimate for a given AEP at site *i*;

X_{site_i} is the observed at-site streamflow 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*;

X_{reg_i} is the peak-flow estimate using the regional regression equation for a given AEP at site *i*;

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

V_{wtd_i} is the weighted variance for a given AEP at site i.

Transformation of the peak-flow values back to arithmetic space, in cubic feet per second, is shown in equation 11:

$$Q_{wtd_{AEP,i}} = 10^{X_{wtd_{AEP,i}}}, \quad (11)$$

where

$Q_{wtd_{AEP,i}}$ is the weighted peak-flow estimate for a given AEP at site *i*.

The computation of weighted prediction intervals is similar to those discussed previously and shown in equations 7 and 8, except the weighted peak flow (X_{wtd_i}) is used in place of the regional peak-flow estimate (X_{reg_i}), and weighted variance (V_{wtd_i}) is used in place of the AVP from the regression (AVP_{reg}).

Table 4. Comparison of standard error of prediction (SEP) and model error variance (MEV) between this study and a previous study by Ramos-Gines (1999).

[The 0.0005 annual exceedance probability was not calculated by Ramos-Gines (1999) and is not included in this table. %, percent]

Annual exceedance probability	SEP (%)				MEV (log units)			
	Ramos-Gines (1999)[1]	Ramos-Gines (1999)[2]	Current study, region 1	Current study, region 2	Ramos-Gines (1999)[1]	Ramos-Gines (1999)[2]	Current study, region 1	Current study, region 2
0.5	47.2	41.4	46.7	57.5	0.0356	0.0272	0.034	0.0506
0.2	35.3	31.8	34.9	38.1	0.0203	0.0154	0.0189	0.0232
0.1	30.9	26.8	29.8	33.6	0.0153	0.0104	0.0131	0.0177
0.04	30.7	24.4	25.7	33.4	0.0149	0.008	0.0087	0.0168
0.02	33.1	24.7	24.0	34.5	0.0172	0.008	0.0066	0.0175
0.01	36.7	26.3	22.2	36.0	0.021	0.0089	0.0045	0.0186
0.002	47.7	32.2	18.7	39.3	0.0348	0.0136	0.0001	0.0207

[1]Equations developed by Ramos-Gines (1999) that use drainage area as the only explanatory variable.

[2]Equations developed by Ramos-Gines (1999) that use drainage area, mean annual rainfall, and (or) depth-to-rock as explanatory variables.

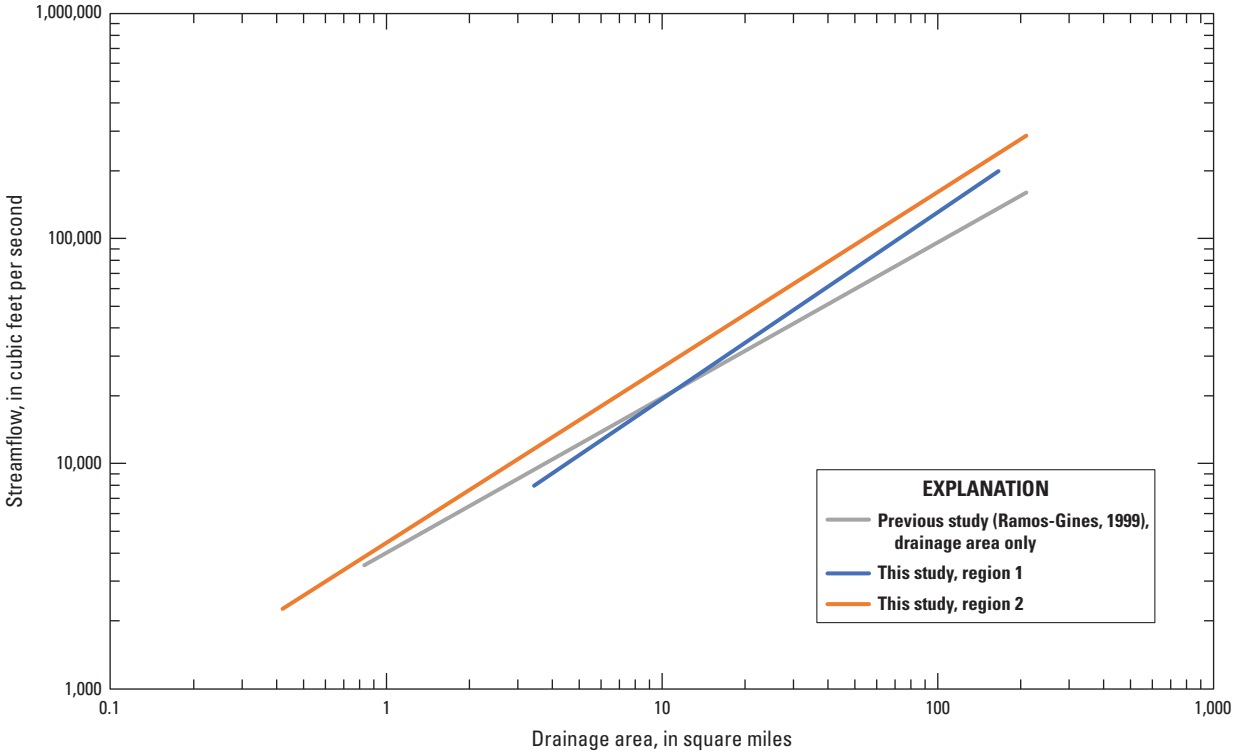

Figure 7. Comparison of 1-percent chance exceedance flows from previously published and new regional regression equations that use drainage area as the only explanatory variable.

Example Computation of Weighted Peak-Flow Estimates

Example 2. Calculate the weighted peak flow (Q_{wtd}) for a 1-percent chance exceedance flow (0.01 AEP) at USGS streamgage 50056400, Rio Valenciano near Juncos, Puerto Rico. The streamgage is located in the eastern portion of the island at lat 18°12′58" N., long 65°55′34" W. The drainage area is 16.5 mi^2, and the basin is considered rural and unaffected by substantial regulation, diversion, or karst influence.

1. From example 1, a predicted 1-percent AEP peak-flow estimate ($Q_{reg,0.01}$) of 39,800 ft^3/s (39,758 ft^3/s rounded to three significant figures) was computed using the appropriate regional regression equation for region 2.
2. This peak-flow estimate, in log base 10 units, (X_{reg}, 0.01) is log10(39,800) = 4.600 ft^3/s.
3. From appendix 3, the predicted variance for the regression estimate (V_{reg}, 0.01) is 0.0227 (log base 10 units).
4. From appendix 3, the observed at-site peak-flow estimate (Q_{site}, 0.01) is 50,000 ft^3/s (49,960 ft^3/s rounded to 3 significant figures). Transforming to log units, the value (X_{site}, 0.01) is 4.699 ft^3/s.
5. From appendix 3, the observed at-site variance (V_{site}, 0.01) is 0.0116, in log base 10 units.
6. Using equation 9, a weighted peak-flow estimate is computed by

$$X_{wtd_{0.01,50056400}} = \frac{(4.699*0.0227)+(4.600*0.0116)}{0.0116+0.0227}$$

$$X_{wtd_{0.01,50056400}} = 4.666 \text{ft}^3/\text{s}, \text{ in log units}$$

7. Using equation 11 to transform to arithmetic units, the weighted peak-flow estimate, rounded to 3 significant figures, is

$$Q_{wtd_{0.01,50056400}} = 10^{4.666}$$

$$Q_{wtd_{0.01,50056400}} = 46{,}300 \text{ ft}^3/\text{s}$$

Example Computation of Prediction Intervals

Example 3. Calculate the 95-percent prediction interval limits ($PI_{i,lower}$, $PI_{i,upper}$) for a 1-percent AEP peak flow at USGS streamgage 50056400, Rio Valenciano near Juncos, Puerto Rico.

1. Calculate the weighted variance using equation 10. From appendix 3, the at-site variance (V_{site}, 0.01) is 0.0116, and the predicted variance for the regression estimate (V_{reg}, 0.01) is 0.0227, both in log base 10 units. The weighted variance is calculated by

$$V_{wtd_{0.01,\,50056400}} = \frac{0.0116*0.0227}{0.0116+0.0227}$$

$$V_{wtd_{0.01,\,50056400}} = 0.0077$$

2. From table 3, the critical value from the Student's t distribution for region 2 is 2.017.

3. From example 2, the weighted peak-flow estimate in log units ($X_{wtd_{0.01,50056400}}$) was calculated as 4.666 ft^3/s.

4. Using equation 7 and substituting the weighted peak-flow estimate for the regional regression flow estimate (X_{reg_i}) and the computed weighted variance ($V_{wtd_{0.01,50056400}}$) calculated in step 1 for AVP_{reg}, a lower prediction level ($PI_{i,lower}$) is computed by

$$PI_{\,50056400,\,lower} = 10^{\left[X_{wtd0.01,50056400} - t_{\left(\frac{\alpha}{2}\right),(n-p)}*\left(V_{wtd0.01,50056400}\right)^{0.5}\right]}$$

$$PI_{\,50056400,\,lower} = 10^{\left[4.666-2.017*(0.0077)^{0.5}\right]}$$

$$PI_{\,50056400,\,lower} = 10^{(4.489)}$$

$$PI_{\,50056400,\,lower} = 30{,}800\ \text{ft}^3/\text{s},$$

rounded to 3 significant figures.

5. Using the same procedure with equation 8, the upper prediction level is computed by

$$PI_{\,50056400,\,upper} = 10^{\left[X_{wtd0.01,50056400} + t_{\left(\frac{\alpha}{2}\right),(n-p)}*\left(V_{wtd0.01,50056400}\right)^{0.5}\right]}$$

$$PI_{\,50056400,\,upper} = 10^{\left[4.666+2.017*(0.0077)^{0.5}\right]}$$

$$PI_{\,50056400,\,upper} = 69{,}700\ \text{ft}^3/\text{s},$$

rounded to 3 significant figures.

From this example, there is a 95-percent chance that the actual 1-percent chance exceedance peak flow at streamgage 50056400 is between 30,800 ft^3/s and 69,700 ft^3/s. As a result of rounding issues, the resultant computational values provided in these examples may not exactly match the values reported in appendix 3.

Estimation of Peak-Flow Statistics at Ungaged Sites Near Streamgages

A more accurate estimate of peak flow than using only the regression equations is possible from an ungaged location near an existing unregulated, rural streamgage with 10 or more years of record by adjusting the weighted estimate at the gage with a *DAR* if certain criteria are met. This alternative method calculates a weighted estimate at the ungaged site ($Q_{(u)wtd}$) by using equation 12 and is applicable if (1) the ungaged site is located on the same stream as the streamgage, and (2) the drainage area of the ungaged site is between 0.5 and 1.5 times the drainage area of the streamgage.

$$Q_{(u)wtd_{AEP}} = \left[\frac{2\Delta DA}{DA_{(g)}} + \left(1 - \frac{2\Delta DA}{DA_{(g)}}\right)\left(\frac{Q_{(g)wtd_{AEP,\bar{i}}}}{Q_{(g)reg_{AEP,\bar{i}}}}\right)\right] Q_{(u)reg\,AEP} \qquad (12)$$

where

$Q_{(u)wtd_{AEP}}$ is the weighted peak-flow estimate for a given AEP at the ungaged site;

$DA_{(g)}$ is the drainage area of the streamgage;

ΔDA is the absolute value of the difference between the drainage areas of the streamgage and the ungaged site;

$Q_{(g)wtd_{AEP,i}}$ is the weighted peak-flow estimate for a given AEP at the streamgage *i*;

$Q_{(g)reg_{AEP,i}}$ is the peak-flow estimate calculated by the regression equation for a given AEP at streamgage *i*; and

$Q_{(u)reg_{AEP}}$ is the peak-flow estimate calculated by the regression equation for a given AEP at the ungaged site.

As mentioned in Ries and others (2007), 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 streamgage and increasing weight to the streamgage estimates as the *DAR* approaches 1. It should not be used for *DAR* less than 0.5 or greater than 1.5. This method may also prove more

effective than the regression equations when predicting estimates for a location that has a drainage area outside the range of those used to develop the regional regression equations (table 3).

Example 4. Calculate the peak-flow estimate for the 1-percent chance exceedance flow (0.01 AEP) at an ungaged site located in region 2 upstream from USGS streamgage 50056400, Rio Valenciano near Juncos, Puerto Rico. The drainage area of the streamgage is 16.5 mi^2, and the drainage area of the ungaged site is 12.8 mi^2. The basin of the ungaged site is considered rural and unaffected by substantial regulation, diversion, or karst influence.

1. Calculate the absolute value of the difference in drainage areas (ΔDA) between that of the streamgage and ungaged site.

$$\Delta DA = |16.5\text{-}12.8| = 3.7 \text{ mi}^2$$

2. From example 2 (final step 7), the weighted peak-flow estimate at the streamgage ($Q_{(g)wtd_{0.01,50056400}}$) is 46,300 ft^3/s

3. The peak-flow estimate for the 1-percent AEP at the streamgage using the region 2 regression equation from table 3 is

$$Q_{(g)reg_{0.01,50056400}} = 10^{3.65}\,(16.5)^{0.78}$$

$$= 39{,}800 \text{ ft}^3/\text{s}$$

4. The peak-flow estimate for the 1-percent AEP at the ungaged site using the region 2 regression equation from table 3 is

$$Q_{(u)reg_{0.01}} = 10^{3.65}\,(12.8)^{0.78} = 32{,}600 \text{ ft}^3/\text{s}$$

5. Using these values and equation 12, the weighted streamflow for the 1-percent AEP at the specified ungaged site is

$$Q_{(u)wtd_{0.01}} = \left[\frac{2\left(3.7 \text{ mi}^2\right)}{16.5 \text{ mi}^2} + \left(1 - \frac{2\left(3.7 \text{ mi}^2\right)}{16.5 \text{ mi}^2}\right)\left(\frac{46{,}300 \text{ ft}^3/\text{s}}{39{,}800 \text{ ft}^3/\text{s}}\right)\right] * 32{,}600 \text{ ft}^3/\text{s}$$

$$Q_{(u)wtd_{0.01}} = \left[0.448 + \left(0.552\right)\left(1.163\right)\right] * 32{,}600 \text{ ft}^3/\text{s}$$

$$Q_{(u)wtd_{0.01}} = 35{,}500 \text{ ft}^3/\text{s}$$

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

Multiple options are available to estimate the magnitude and frequency of rural, unregulated peak flows. The preferred computation method with the highest accuracy depends on the data available for the location of interest. The decisions should be made according to the following, in order of preference:

1. If the location of interest is at a streamgage with 10 or more years of record, techniques described in Bulletin 17C (England and others, 2018) should be used to compute at-site estimates. These estimates should be weighted by variance with the appropriate regression equation using procedures outlined in U.S. Geological Survey (2010), as shown in equations 9–11 and example 2, to produce a better estimate.

2. If the location 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, a weighted discharge using the DAR should be calculated, as shown in example 4 and equation 12.

3. If the location of interest is ungaged and is not within 0.5 to 1.5 times the drainage area of a nearby streamgage on the same unregulated stream, then the appropriate regional regression equation in table 3 should be used, as shown in example 1.

Estimates for the magnitude and frequency of floods in Puerto Rico can be calculated manually by using the methods mentioned previously or by using the USGS web application, StreamStats. StreamStats provides the at-site information for gaged locations and incorporates the regional regression equations, as well as basin characteristics, from this study to calculate specified streamflow at ungaged locations. More information on StreamStats can be found in Ries and others (2017).

Summary

The U.S. Geological Survey (USGS) revised at-site flood flow statistics and computed regional regression equations on rural, unregulated streams in Puerto Rico using annual peak-flow data through 2017. Flood-flow statistics with 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent chance annual exceedance levels were estimated for 91 streamgages having at least 10 or more years of annual peak-flow record. The stations considered were also not affected by regulation at moderate to high flows, tidal influence, or by urban conditions, namely 10 percent or less of impervious area in the watershed.

At-site flood frequency estimates were computed using the most recent methods and procedures outlined in Bulletin 17C. A generalized skew analysis was performed with the Bayesian generalized least squares model and used data from 42 streamgages having 25 or more years of record. An average skew value of 0.28, with a mean squared error of 0.20, was computed for the entire island and used in the at-site frequency analysis. The at-site analysis was performed by using the USGS program PeakFQ version 7.3, which uses the Expected Moments Algorithm to fit the three parameters of the log-Pearson Type III distribution and the Multiple Grubbs-Beck test for low outlier detection.

By using the at-site frequency estimates, regression models were built to predict peak flows based on known basin characteristics. Sites were removed from use in regression equation development if they were determined to be redundant based on location, drainage area, and record length. Ordinary least-squares regression techniques were used initially to select relevant basin characteristics and divide the island into two regions, roughly east and west, based on residuals. Generalized least-squares regression techniques were used to develop the final model equations for each of the two regions using drainage area as the only explanatory variable. The sites in each region were examined using metrics for high influence, leverage, trends, overall fit, and accuracy of the model. The final analysis included 32 streamgages in region 1 and 45 streamgages in region 2. The average standard errors of prediction ranged from 18.7 to 46.7 percent in region 1 and from 33.4 to 57.5 percent in region 2 for all annual exceedance probabilities examined. At exceedance levels of 1-percent and lower (that is, higher flows), the standard errors of prediction ranged from 18.7 to 22.2 percent in region 1 and from 36 to 39.3 percent in region 2.

These regression equations and basin characteristics are also accessible in the USGS web application StreamStats. This application allows the user to delineate watersheds and estimate peak flows at specified frequencies at both gaged and ungaged locations.

Acknowledgments

This work was prepared with Hurricane Maria supplemental funding.

The authors would like to acknowledge staff from the U.S. Geological Survey (USGS) who provided assistance in recent years with data collection and review efforts to publish the best streamflow data possible. The published peak streamflow data were used to perform the analysis included in this study.

The authors also acknowledge USGS employee Joann Dixon for her assistance and production of figures and basin characteristics using geographic information systems.

Special thanks are given to USGS employee Andrea G. Veilleux for the development of the generalized skew that was used for the study.

References Cited

Alemán-González, W.B., 2010, Karst map of Puerto Rico: U.S. Geological Survey Open-File Report 2010–1104, 1 sheet, scale 1:140,000, accessed June 6, 2019, at https://pubs.usgs.gov/of/2010/1104/.

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, accessed June 27, 2017, at https://doi.org/10.1002/wrcr.20392.

Dixon, J.F., Kolb, K.R., and Ryan, P.J., 2021, Geospatial datasets for watershed delineation used in the development of the USGS StreamStats application for Puerto Rico: U.S. Geological Survey data release, https://doi.org/10.5066/P9XFTY06.

Eng, K., Chen, Y., and Kiang, J.E., 2009, User's guide to the weighted-multiple-linear-regression program (WREG version 1.0): U.S. Geological Survey Techniques and Methods, book 4, chap. A8, 21 p. [Also available at 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., Jr., 2018, Guidelines for determining flood flow frequency—Bulletin 17C: U.S. Geological Survey Techniques and Methods, book 4, chap. B5, 148 p., accessed May 1, 2019, at https://doi.org/10.3133/tm4B5.

Farmer, W.H., 2019, WREG—U.S. Geological Survey WREG R package version 3.00: U.S. Geological Survey Official Source Code Archive website, accessed July 20, 2020, at https://code.usgs.gov/wfarmer/WREG.

Farmer, W.H., Kiang, J.E., Feaster, T.D., and Eng, K., 2019, Regionalization of surface-water statistics using multiple linear regression: U.S. Geological Survey Techniques and Methods, book 4, chap. A12, 40 p., accessed August 4, 2020, at https://doi.org/10.3133/tm4A12.

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

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. [Also available at https://pubs.usgs.gov/twri/twri4a3/.]

Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood flow frequency: Bulletin 17B. [Interagency Advisory Committee on Water Data, Hydrology Subcommittee Technical Report.].

Kolb, K.R., and Ryan, P.J., 2021, Basin Characteristics Rasters for Puerto Rico StreamStats, 2021: U.S. Geological Survey data release, https://doi.org/10.5066/P9HK9SSQ.

Lopez, M.A., Colon-Dieppa, E., and Cobb, E.D., 1979, Floods in Puerto Rico, magnitude and frequency: U.S. Geological Survey Water-Resources Investigations Report 78–141, 69 p.

Lopez, M.A., and Fields, F.K., 1970, A proposed streamflow-data program for Puerto Rico: U.S. Geological Survey Open-File Report 70–199, 35 p. [Also available at https://doi.org/10.3133/ofr70199.]

Miller, J.A., Whitehead, R.L., Oki, D.S., Gingerich, S.B., and Olcott, P.G., 1999, Ground water atlas of the United States—Segment 13, Alaska, Hawaii, Puerto Rico, and the U.S. Virgin Islands: U.S. Geological Survey Hydrologic Atlas 730–N, accessed August 10, 2020, at https://pubs.er.usgs.gov/publication/ha730N.

Monroe, W.H., 1976, The karst landforms of Puerto Rico: U.S. Geological Survey Professional Paper 899, 69 p., accessed June 6, 2019, at https://pubs.usgs.gov/pp/0899/report.pdf.

National Resources Conservation Service [NRCS], 1986, Urban Hydrology for Small Watersheds: TR-55, accessed May 5, 2021, at https://www.nrcs.usda.gov/Internet/FSE_DOCUMENTS/stelprdb1044171.pdf.

National Oceanic and Atmospheric Administration [NOAA], 2019, 2018 local climatological data annual summary with comparative data, San Juan, Puerto Rico (TJSJ): NOAA National Centers for Environmental Information database, accessed June 7, 2019, at https://www.ncdc.noaa.gov/IPS/lcd/lcd.html?_page=1&state=PR&stationID=11641&_target2=Next+%3E.

National Weather Service [NWS], 2019, Flooding in Puerto Rico and U.S. Virgin Islands: NWS web page, accessed June 6, 2019, at https://www.weather.gov/safety/flood-states-pr.

National Weather Service [NWS], 2020, Puerto Rico and U.S. Virgin Islands Climate Divisions: NWS web page, accessed August 5, 2020, at https://www.weather.gov/images/sju/climo/monthlymaps/temps/ClimateZones.png.

Neter, J., Wasserman, W., and Kutner, M.H., 1985, Applied linear statistical models (2d ed): Homewood, Ill., Richard D. Irwin Inc., 1,127 p.

Oki, D.S., Rosa, S.N., and Yeung, C.W., 2010, Flood-frequency estimates for streams on Kaua'i, O'ahu, Moloka'i, Maui, and Hawai'i, State of Hawai'i: U.S. Geological Survey Scientific Investigations Report 2010–5035, 121 p.

R Core Team, 2020, The R Project for statistical computing: The R Foundation, Vienna, Austria, accessed March 16, 2020, at https://www.R-project.org/.

Ramos-Gines, O., 1999, Estimation of magnitude and frequency of floods for streams in Puerto Rico—New empirical models: U.S. Geological Survey Water-Resources Investigations Report 99–4142, 42 p.

Ries, K.G., III, Atkins, J.B., Hummel, P.R., Gray, M., Dusenbury, R., Jennings, M.E., Kirby, W.H., Riggs, H.C., Sauer, V.B., and Thomas, W.O., Jr., 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. [Also available at https://doi.org/10.3133/tm4A6.]

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. [Supersedes USGS Fact Sheet 2008–3067.].

Ryan, P.J., and Hazelbaker, C.L., 2021, Data files for the development of regression equations for estimation of the magnitude and frequency of floods at rural, unregulated gaged and ungaged streams in Puerto Rico through water year 2017: U.S. Geological Survey data release, https://doi.org/10.5066/P91XT14B.

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

U.S. Census Bureau, 2019, QuickFacts—Puerto Rico: U.S. Census Bureau web page accessed June 6, 2019, at https://www.census.gov/quickfacts/geo/chart/PR/POP060210.

U.S. Geological Survey, 2010, Weighted estimates of peak-flow frequency statistics: U.S. Geological Survey Office of Surface Water Technical Memorandum No. 2010.05, 3 p., accessed September 16, 2020, at https://water.usgs.gov/water-resources/memos/.

U.S. Geological Survey, 2012, Guidance on determination and revision of watershed drainage areas: U.S. Geological Survey Office of Surface Water Technical Memorandum No. 12.07, 2 p., accessed October 2, 2019, at https://water.usgs.gov/water-resources/memos/.

U.S. Geological Survey, 2020a, USGS water data for the Nation: U.S. Geological Survey National Water Information System database, accessed February 1, 2020, at https://doi.org/10.5066/F7P55KJN. [Peak-flow data directly accessible at https://water.usgs.gov/usa/nwis/peak.]

U.S. Geological Survey, 2020b, USGS peak streamflow for the Nation: U.S. Geological Survey National Water Information System web interface, accessed October 26, 2020, at https://nwis.waterdata.usgs.gov/nwis/peak?help.

U.S. Water Resources Council, 1977, Guidelines for determining flood flow frequency, Bulletin No. 17A: U.S. Water Resources Council, Subcommittee on Hydrology, Washington, D.C.

Veilleux, A.G., 2011, Bayesian GLS regression, leverage and influence for regionalization of hydrologic statistics: Ithaca, N.Y., Cornell University Ph.D. dissertation, 184 p. [Also available at https://hdl.handle.net/1813/30607.]

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., accessed December 17, 2019, at https://doi.org/10.3133/fs20133108.

Williams-Sether, T., 2021, Estimating low-flow frequency and duration statistics for selected streams and the implementation a StreamStats web-based tool in Puerto Rico: U.S. Geological Survey Scientific Investigations Report 2021–5054, 18 p., https://doi.org/10.3133/sir20215054.

Appendix 1. Streamgages Considered for Development of Regional Regression Equations in Puerto Rico and Details of At-Site Statistic Inputs

The spreadsheet is available for download in .xls and .csv format at https://doi.org/10.3133/sir20215062.

Appendix 2. Regional Skew Regression Analysis for Puerto Rico

By Andrea G. Veilleux and Daniel M. Wagner

Introduction to Statistical Analysis of Regional Skew

To help improve estimates of annual exceedance probability discharges, current guidance for flood-frequency analysis by Federal agencies in Bulletin 17C (England and others, 2018) recommends using a weighted average of the station skewness coefficient and a regional skewness coefficient. Previous guidance (Bulletin 17B, Interagency Advisory Committee on Water Data, 1982) supplied a national map but encouraged hydrologists to develop more specific local relations. Since Bulletin 17B was published, nearly 40 years of additional annual peak-flow data have been collected, and better spatial estimation procedures have been developed (Stedinger and Griffis, 2008).

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

Because of complications introduced by the use of the Expected Moments Algorithm (EMA) with Multiple Grubbs-Beck test (MGBT) for potentially influential low flows (PILFs; Cohn and others, 1997) and large cross correlations between annual peak flows at pairs of streamgages, an alternate regression procedure was developed to provide stable and defensible results for regional skew (Veilleux, 2011; Lamontagne and others, 2012; Veilleux and others, 2012). This procedure is referred to as the Bayesian WLS/Bayesian GLS (B–WLS/B–GLS) regression framework (Veilleux, 2011; Veilleux and others, 2011, 2012). The B–WLS/B–GLS framework uses OLS regression to fit an initial model of regional skew that is used to generate a stable estimate of regional skew for each streamgage. This estimate is the basis for computing the variance of each estimate of station skew used in the B–WLS analysis. B–WLS is then used to generate estimators of the regional skew model parameters; finally, B-GLS is used to estimate the precision of those estimators, the MEV and its precision, and compute various diagnostic statistics.

In this study, EMA with MGBT was used to estimate the station skew and its mean squared error. Because EMA with MGBT allows for the censoring of PILFs as well as the use of flow intervals to describe missing, censored, and historic data, it complicates the calculations of effective record length (and effective concurrent record length) used to describe the precision of skew estimates because the annual peak flows are no longer represented by single values. To properly account for these complications, the new B–WLS/B–GLS procedure was used.

Methodology for Developing the Regional Skew Model

This section provides a brief description of the B–WLS /B–GLS methodology as it appears in Veilleux and others (2012). More detailed descriptions can be found in Veilleux (2011) and Veilleux and others (2011).

Ordinary Least-Squares Analysis

The first step in the B–WLS/B–GLS regional skew analysis is the estimation of a regional skew model using OLS regression. The OLS regression yields coefficients ($\hat{\boldsymbol{\beta}}_{OLS}$) and a model that can be used to generate unbiased and relatively stable regional estimates of skew for all streamgages:

$$\tilde{\mathbf{y}}_{OLS} = \mathbf{X}\hat{\boldsymbol{\beta}}_{OLS}, \quad (2\text{–}1)$$

where

$\mathbf{X}$ is an (n x k) matrix of basin characteristics;
$\tilde{\mathbf{y}}_{OLS}$ are the estimated regional skew values;
n is the number of streamgages; and
k is the number of basin characteristics, including a column of ones to estimate the constant.

These estimated station-regional skew values, $\tilde{\mathbf{y}}_{OLS}$, are then used to calculate unbiased station-regional skew variances using the equations reported in Griffis and Stedinger (2009). These station-regional skew variances are based on the OLS estimator of the skew instead of the station skew, thus making the weights in the subsequent steps relatively independent of the station skew.

Weighted Least-Squares Analysis

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

Generalized Least-Squares Analysis

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

Data Analysis

Station Skew

To estimate the station skew, G, and its mean squared error, MSEG, results of the EMA–MGBT analysis described in the body of this report were used (Cohn and others, 1997; Griffis and others, 2004). The EMA–MGBT provides a straightforward and efficient method for the incorporation of historical information and censored data, such as those from a crest-stage gage (CSG). For this analysis, version 7.3 of U.S. Geological Survey (USGS) PeakFQ software (Veilleux and others, 2014), which incorporates EMA–MGBT, was used to generate G and its corresponding MSEG for the 42 streamgages used in the study, assuming a log-Pearson Type III distribution and generally employing MGBT for screening of PILFs (table 2.1; available for download in .xls and .csv format at https://doi.org/10.3133/sir20215062; see section "Potentially Influential Low Flows" in the main part of this report for more detail regarding EMA and MGBT).

Pseudo Record Length

Because the annual peak-flow records of the streamgages include historic information and censored data, the effective record length is used to compute the precision of the skew estimates and takes into account the availability of historic information and censored values. Although historic information and censored peaks are valuable information, they often provide less information than an equal number of years of gaged peaks (Stedinger and Cohn, 1986). The following calculations provide a pseudo record length, P_{RL}, which appropriately accounts for all types of data available for a streamgage.

The P_{RL} is defined in terms of the number of years of gaged record that would be required to yield the same mean squared error of the skew ($MSE\left(\tilde{G}\right)$) as the combination of historic and gaged record available at a streamgage; thus, the P_{RL} of the skew is a ratio of the MSE of the station skew when only the gaged record is analyzed ($MSE\left(\hat{G}_S\right)$) to the MSE of the station skew when the all of the data, including historic and censored data, are analyzed ($MSE\left(\hat{G}_C\right)$).

$$P_{RL} = \frac{P_s * MSE\left(\hat{G}_S\right)}{MSE\left(\hat{G}_C\right)}, \qquad (2\text{–}2)$$

where

Ps is the number of gaged peaks in the record.

Because the P_{RL} is an estimate, the following conditions must also be met to ensure a valid approximation: (1) the P_{RL} must be nonnegative; if the P_{RL} is greater than PH (the length of the historical period), then P_{RL} should be set to equal P_H; and (2) if the P_{RL} is less than P_S, then the P_{RL} is set to P_S. This ensures that the P_{RL} will not be larger than P_H or less than P_S.

The estimate of station skew is sensitive to extreme events, and more accurate estimates can be obtained from longer records (England and others, 2018). Therefore, streamgages that have less than 35 years of P_{RL} are normally not used for regional skew analysis; however, because of the limited number of streamgages in Puerto Rico that were available, none were removed based on the P_{RL}. The minimum P_{RL} used in the study was 27 years, and the maximum was 72 years (table 2.1).

Redundant Streamgages

Redundancy results when the drainage basins of two streamgages are nested, meaning that one basin is contained inside the other and the two basins are of similar size. Instead of representing two independent spatial observations that depict how characteristics of the drainage basins are related to annual peak flows or skew, these two basins will have a similar hydrologic response to a given storm and thus represent only one spatial observation. When streamgages are redundant, a statistical analysis using both streamgages incorrectly represents the information in the regional dataset (Gruber and Stedinger, 2008). To determine if two streamgages are redundant

and thus represent the same hydrologic conditions, two types of information are considered: (1) whether their basins are nested, and (2) the ratio of the drainage areas of the basins.

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

$$SD_{ij} = \frac{D_{ij}}{\sqrt{0.5\left(DRNAREA_i + DRNAREA_j\right)}}, \quad (2–3)$$

where

$D_{i,j}$ is the distance between centroids of basin *i* and basin *j*, in miles; and

$DRNAREA_i$ is the drainage area at site *i*, in square miles; and

$DRNAREA_j$ is the drainage area at site *j*, in square miles.

The drainage area ratio (*DAR*) is used to determine if two nested basins are sufficiently similar in size to conclude that they are, or are at least in large part, the same basin for the purposes of developing a regional hydrologic model. The *DAR* is defined as follows (Veilleux and others, 2011):

$$DAR = Max\left[\frac{DRNAREA_i}{DRNAREA_j}, \frac{DRNAREA_j}{DRNAREA_i}\right], \quad (2–4)$$

where

Max is the maximum of the two values in brackets;

$DRNAREA_i$ is the drainage area at site *i*, in square miles; and

$DRNAREA_j$ is the drainage area at site *j*, in square miles.

Two basins might be redundant if they are similar in size and their basins are nested. Previous studies suggest that streamgage pairs having *SD* less than or equal to 0.50 and *DAR* less than or equal to 5 were likely to be redundant for purposes of determining regional skew. If *DAR* is large enough, even if the streamgage pairs are nested, they will reflect different hydrologic responses because storms of different sizes and durations will affect each streamgage differently. All possible combinations of streamgage pairs from 46 candidate streamgages were considered in the redundancy analysis. All streamgage pairs identified as redundant were then investigated to determine if, in fact, one streamgage of the pair was nested inside the other. For streamgage pairs that were nested, one streamgage from the pair was removed from the regional skew analysis. For this analysis, four streamgages were removed for redundancy, leaving 42 streamgages for use in the regional skew analysis: USGS 50035000, Rio Grande De Manati at Ciales, Puerto Rico; USGS 50043800, Rio De La Plata at Comerio, Puerto Rico; USGS 50051800, Rio Grande De Loiza at Highway 183, San Lorenzo, Puerto Rico; and USGS 50057000, Rio Gurabo at Gurabo, Puerto Rico. The centroids of streamgages used for analysis and possible regional skew division used by Ramos-Gines (1999) are shown in figure 2.1.

Unbiasing the Station Skew

The station skew estimates were unbiased by using the correction factor developed by Tasker and Stedinger (1986) and employed by Reis and others (2005). The unbiased station skew estimate using the PRL is

$$\hat{\gamma}_i = \left[1 + \frac{6}{P_{RL,i}}\right] G_i, \quad (2–5)$$

where

$\hat{\gamma}_i$ is the unbiased station skew estimate for streamgage *i*;

$P_{RL,i}$ is the pseudo record length, in years, for streamgage *i*, as calculated in equation 2–2; and

G_i is the biased estimate of station skew for streamgage *i* from the flood frequency analysis.

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

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

where

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

$$Var\left(\hat{G}\right) = \left[\frac{6}{P_{RL}} + a\left(P_{RL}\right)\right] * \left[1 + \left(\frac{9}{6} + b\left(P_{RL}\right)\right)\hat{G}^2 + \left(\frac{15}{48} + c\left(P_{RL}\right)\right)\hat{G}^4\right], \quad (2–7)$$

where

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

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

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

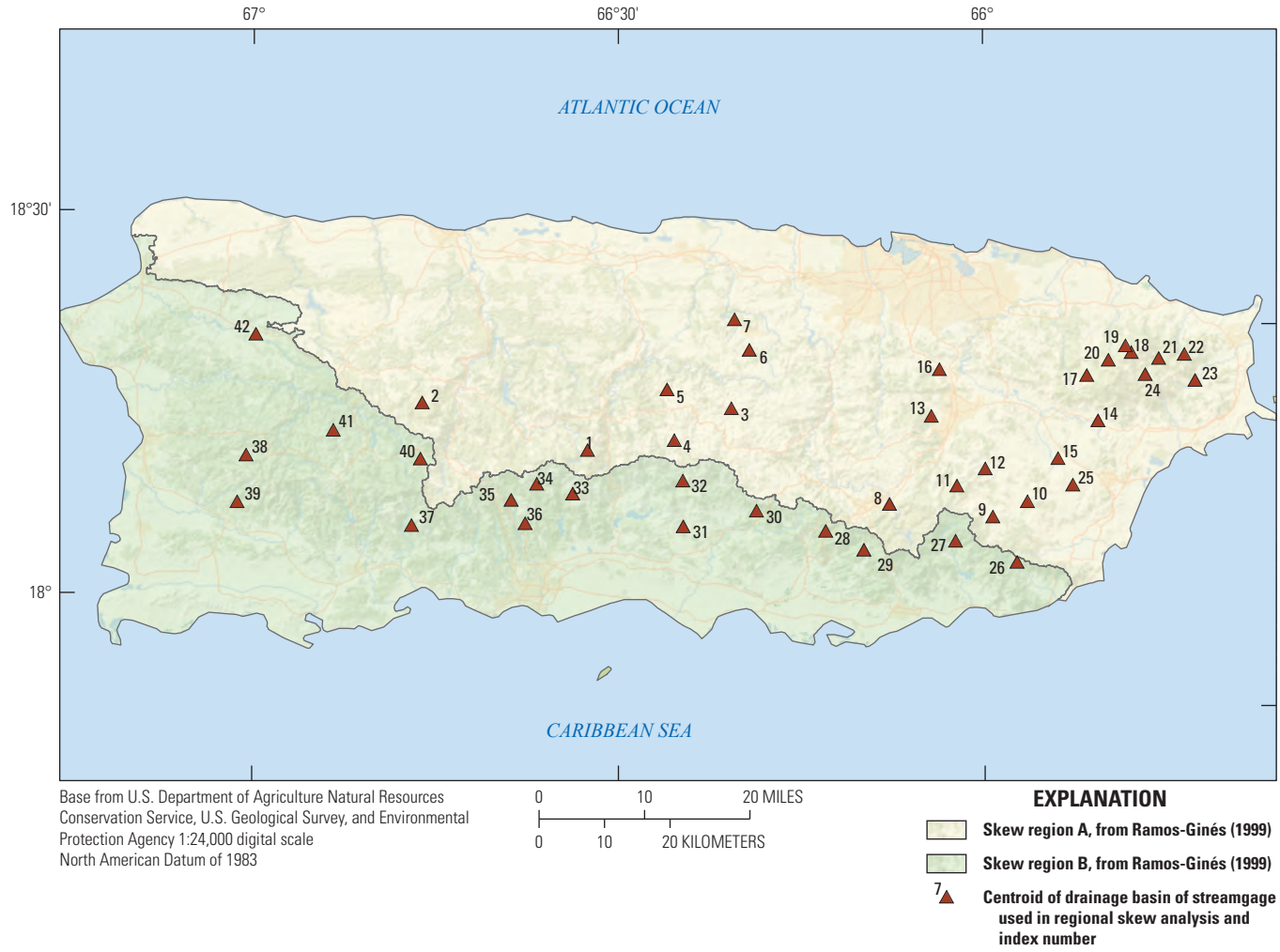

Figure 2.1. Map showing (1) the centroid of drainage basin of U.S. Geological Survey streamgages in Puerto Rico that were used for regional skew analysis and (2) the two skew regions investigated.

Estimating the Mean Squared Error of the Skew

There are several possible ways to estimate the MSE_G. The approach used by the EMA (see equation 55 in Cohn and others, 2001) generates a first-order estimate of the MSE_G, which should perform well when interval data are present. Another option is to use the formula in equation 2–7 (the variance is equated to the MSE), employing either the length of the gaged record or the length of the historical period (H_p); however, this method does not account for censored data and can lead to an inaccurate and underestimated MSE_G. This issue has been addressed by using the PRL instead of H_p; the P_{RL} reflects the impact of the censored data and the number of gaged peaks. Thus, the unbiased MSE_G, computed using the formula from Griffis and Stedinger (2009), was used in the regional skew model because it is more stable and relatively independent of the station skew. This methodology was used in previous regional skew studies (Eash and others, 2013; Southard and Veilleux, 2014).

Cross-Correlation Model

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

$$\hat{\rho}\left(\hat{\gamma}_i, \hat{\gamma}_j\right) = Sign\left(\hat{\rho}_{ij}\right) cf_{ij} \left|\hat{\rho}_{ij}\right|^k , \quad (2–8)$$

where

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

κ is a constant between 2.8 and 3.3, and

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

$$cf_{ij} = CY_{ij} / \sqrt{\left(P_{RL,i}\right)\left(P_{RL,j}\right)} , \quad (2–9)$$

where

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

$P_{RL,i}$, and $P_{RL,j}$ are the pseudo record lengths corresponding to streamgages *i* and *j*, respectively (see equation 2–2).

After calculating the P_{RL} for each streamgage in the study, the pseudo concurrent record length between pairs of streamgages can be calculated. Because of the use of censored data and historic data, calculation of the effective concurrent record length is more complex than determining which years the two streamgages both have recorded systematic peaks. First, the number of years of historical record in common between the two streamgages is determined. Next, for the years in common, with beginning year YB_{ij} and ending year YE_{ij}, the following equation is used to calculate the concurrent years of record between site *i* and site *j*:

$$CY_{ij} = \left(YE_{ij} - YB_{ij} + 1\right)\left(\frac{P_{RL,i}}{H_{p,i}}\right)\left(\frac{P_{RL,j}}{H_{p,j}}\right) , \quad (2–10)$$

The computed pseudo concurrent record length depends upon the years of historical record in common between the two streamgages, as well as the ratios of the P_{RL} to the H_p for each of the streamgages.

To relate the concurrent annual peak flows at two streamgages, ρ_{ij}, to explanatory variables, a cross correlation model using 21 streamgages having at least 40 years of concurrent gaged peaks (zero flows not included) was considered. A logit model, termed the Fisher Z-Transformation ($Z = \log[(1 + r)/(1 – r)]$), provided a convenient transformation of the sample correlations, r_{ij}, from the (−1, +1) range to the (-∞, + ∞) range. The model used to estimate the cross correlations of concurrent annual peak flows at two streamgages, which incorporated the distance between basin centroids, D_{ij}, as the only explanatory variable, is

$$\rho_{ij} = \frac{\exp(2Z_{ij}) + 1}{\exp(2Z_{ij}) - 1} , \quad (2–11)$$

where

$$Z_{ij} = exp\left(0.47 - 0.054\left(\frac{D_{ij}^{0.58} - 1}{0.58}\right)\right) .$$

An OLS regression analysis, based on 169 streamgage pairs from 21 sites, indicated this model is as accurate as having 40 years of concurrent gaged peaks from which to calculate cross correlation. Figure 2.2 shows the fitted relation between Z and distance between basin centroids together with the plotted sample data from the 169 streamgage pairs. Figure 2.3 shows the functional relation between the untransformed cross correlation and distance between basin centroids together with the plotted sample data from the 169 streamgage pairs. The cross-correlation model was used to estimate streamgage-to-streamgage cross correlation of concurrent annual peak flows for all streamgage pairs.

Regional Skew Model for Puerto Rico

This study used annual peak-flow data from 42 streamgages operated by the USGS in Puerto Rico (fig. 2.1). Records ending in the 2017 water year (September 30, 2017), if available, were downloaded from the USGS National Water Information System database (https://waterdata.usgs.gov/nwis).

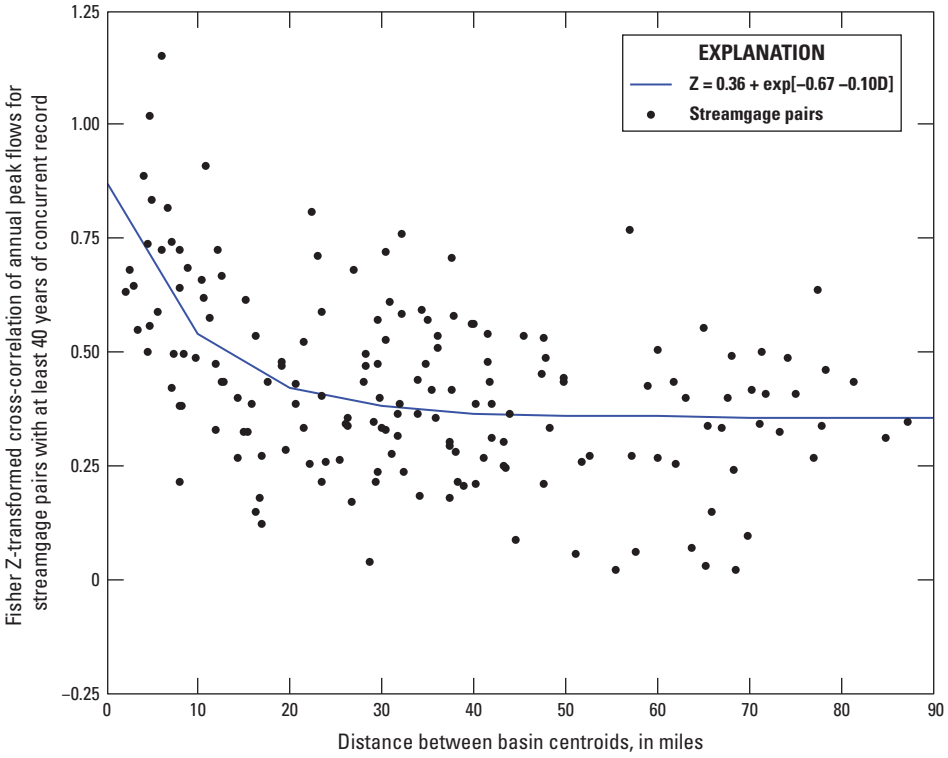

Figure 2.2. Relation between Fisher Z-transformed cross correlation of logs of annual peak flow and distance between basin centroids for Puerto Rico regional skew study. [Z, Fisher Z-transformation; exp, natural exponential function; D, distance between basin centroids, in miles.]

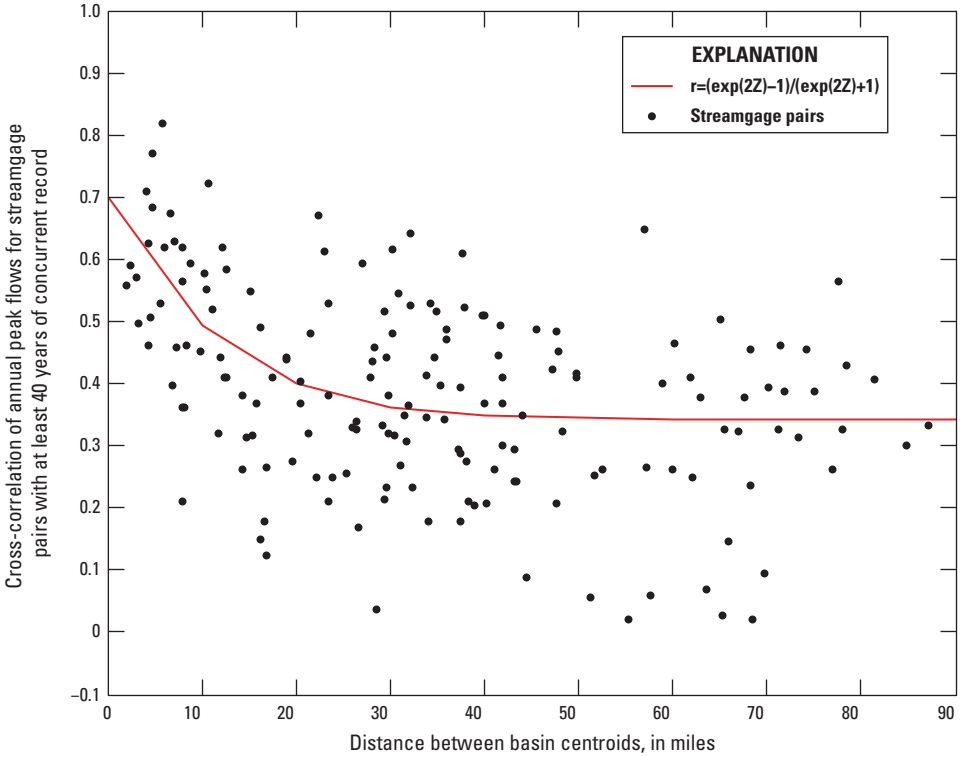

Figure 2.3. Relation between untransformed cross correlation of logs of annual peak discharge and distance between basin centroids for Puerto Rico regional skew study. [r, sample correlations; exp, natural exponential function; Z, Fisher Z-transformation.]

B-WLS / B-GLS models of regional skew were developed for all of Puerto Rico and the skew regions from the previous flood-frequency study for Puerto Rico (Ramos-Gines, 1999; table 2.1, fig. 2.1); however, the small numbers of streamgages and years of annual peak-flow record in each region warranted using the model of regional skew for all of Puerto Rico. No basin characteristics were available for testing because explanatory variables and resources were not available to compute characteristics at the time of this regional skew analysis; therefore, a constant B–WLS / B–GLS model of regional skew for all of Puerto Rico was selected (table 2.2).

The best regional skew model is classified as having the smallest model error variance, σ_δ^2, and largest pseudo- R_δ^2. A constant model does not explain variability in the true skews, so the pseudo- R_δ^2, which describes the estimated fraction of the variability in the true skew from streamgage-to-streamgage explained by each model (Gruber and others, 2007; Parrett and others, 2010), is zero. The posterior mean of σ_δ^2 is 0.18. The average sampling error variance, *ASEV*, is 0.021 and represents the average error in the regional skew for the streamgages in the dataset. The average variance of prediction at a new site, AVP_{new}, is 0.20, which corresponds to an effective record length of 38.7 years and is equivalent to the MSE used in Bulletin 17B to describe the precision of the generalized skew map.

Diagnostic Statistics for Bayesian Weighted Least-Squares/Bayesian Generalized Least-Squares Regression

To evaluate how well a regression model fits a regional hydrologic dataset, diagnostic statistics have been developed (Griffis, 2006; Gruber and others, 2007). A pseudo analysis of variance (pseudo ANOVA) was conducted for the constant model of regional skew in Puerto Rico (table 2.3). The pseudo ANOVA shows how much of the variation in the observed skews can be explained by the regional model, and how much of the variation in residuals can be attributed to model error and sampling error, respectively. Difficulties arise in determining these quantities. The model errors cannot be resolved, because the values of the sampling errors, ηi, for each site, i, are not known; however, the total sampling error sum of squares can be described by its mean value, $\sum_{i=1}^{n} Var[\hat{\gamma}_i]$. Because there are n equations, the total variation because of the model error δ for a model with k parameters has a mean equal to $n\sigma_\delta^2(k)$; thus, the residual variation attributed to the sampling error is $\sum_{i=1}^{n} Var[\hat{\gamma}_i]$, and the residual variation attributed to the model error is $n\sigma_\delta^2(k)$. This division of the variation in the observations is referred to as a pseudo ANOVA, because the contributions of the three sources of error are estimated or constructed, rather than being determined from the residuals and the model predictions, while also ignoring the effect of correlation among the sampling errors.

For a model with no parameters other than the mean (a constant skew model), the estimated model error variance, $\sigma_\delta^2(0)$, describes all of the anticipated variation in $\gamma_i = \mu + \delta_i$, where μ is the mean of the estimated station sample skews; thus, the total expected sum of squares variation because of model error, δ_i, and because of sampling error, $\eta_i = \hat{\gamma}_i - \gamma_i$, in expectation should equal $n\sigma_\delta^2(0) + \sum_{i=1}^{n} Var(\hat{\gamma}_i)$. The expected sum of squares attributed to a regional skew model with k parameters should then equal $n[\sigma_\delta^2(0) - \sigma_\delta^2(k)]$, because the sum of the model error variance $n\sigma_\delta^2(k)$ and the variance explained by the model must sum to $n\sigma_\delta^2(0)$. Table 2.3 considers a model with k = 0 (a constant model). The constant model does not have any explanatory variables; thus, the variation attributed to the models is 0.

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

$$EVR = \frac{\mathrm{SS(sampling\ error)}}{\mathrm{SS(model\ error)}} = \frac{\sum_{i=1}^{n} Var(\hat{\gamma}_i)}{n\sigma_\delta^2(k)} \quad (2\text{–}12)$$

.

The EVR for the constant model is 1.3 (table 2.3). The sampling variability in the station skew was larger than the error in the regional model; thus, an OLS model that neglects sampling error in the station skew might not have provided a statistically reliable analysis of the data. Given the variation in record length among streamgages, it was important to use a WLS or GLS analysis to evaluate the final precision of the model, rather than a simpler OLS analysis.

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

$$MBV^* = \frac{Var\left[b_0^{WLS} \mid GLS\ analysis\right]}{Var\left[b_0^{WLS} \mid WLS\ analysis\right]} = \frac{w^T \Lambda w}{\sum_{i=1}^{n} w_i}, \quad (2\text{–}13)$$

where

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

Table 2.2. Regional skew model for Puerto Rico.

[Standard deviations are in parentheses. σ_δ^2, model error variance; *ASEV*, average sampling error variance; AVP_{new}, average variance of prediction for a new site; pseudo- R_δ^2, fraction of the variability in the true skews explained by each model (Gruber and others, 2007); NA, not applicable]

Model	Regression constant	σ_δ^2	ASEV	AVP_{new}	Pseudo R_δ^2
Constant	0.28	0.18	0.021	0.20	0%
Model	(0.15)	(0.01)	NA	NA	NA

Table 2.3. Pseudo analysis of variance (pseudo ANOVA) of the regional skew model for Puerto Rico.

[*k*, number of estimated regression parameters not including the constant; *n*, number of observations (streamgages) used in regression; $\sigma_\delta^2(0)$ model error variance of a constant model; $\sigma_\delta^2(k)$,model error variance of a model with k regression parameters and a constant ; NA, not applicable; $Var(\hat{\gamma}_i)$ variance of the estimated sample skew at streamgage *i*; *EVR*, error variance ratio; *MBV**, misrepresentation of the beta variance; b_0^{B-WLS} , regression constant from B–WLS analysis; W^T, the transformation of *W*; *Λ,* covariance matrix*; W,* the (k × n) matrix of weights determined by B–WLS analysis; $W_i = 1/\Lambda_{ii}$; B–WLS, Bayesian weighted least-squares; B–GLS, Bayesian generalized least-squares; pseudo R_δ^2 , fraction of variability in the true skews explained by each model (Gruber and others, 2007); %, percent]

Source	Degrees of freedom	Equations	Sum of squares	Result
Model	k = 0	$n\left[\sigma_\delta^2(0)-\sigma_\delta^2(k)\right]$	0	NA
Model error	$n-k-1=41$	$n\left[\sigma_\delta^2(k)\right]$	7.6	NA
Sampling error	$n=42$	$\sum_{i=1}^{n} Var(\hat{\gamma}_i)$	9.8	NA
Total	$2n-1=83$	$n\left[\sigma_\delta^2(k)\right]+\sum_{i=1}^{n} Var(\hat{\gamma}_i)$	17.4	NA
EVR	NA	$\frac{\sum_{i=1}^{n} Var(\hat{\gamma}_i)}{n\left[\sigma_\delta^2(k)\right]}$	NA	1.3
MBV*	NA	$\frac{Var\left[b_0^{B-WLS} \mid B-GLS\ analysis\right]}{Var\left[b_0^{B-WLS} \mid B-WLS\ analysis\right]}=\frac{W^T\Lambda W}{\sum_{i=1}^{n} W_i}$	NA	2.3
Pseudo R_δ^2	NA	$R_\delta^2=1-\frac{\sigma_\delta^2(k)}{\sigma_\delta^2(0)}$	NA	0%

*MBV** for the constant model was 2.3 (table 2.3). This is a large value, indicating the cross correlation among the station skew estimates affected the precision with which the regional skew could be estimated. If a WLS analysis were used to estimate the precision of the constant, the variance would be underestimated by a factor of 2.3; moreover, a WLS model would underestimate the variance of prediction, given that the sampling error in the constant term was sufficiently large to make an appreciable contribution to the average variance of prediction.

Leverage and Influence

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

No streamgages in the regional skew analysis for Puerto Rico exhibited high leverage (greater than 0.0476). The differences in leverage values for the constant model reflect the variation in record length among streamgages. Four streamgages exhibited high influence (Cook's D greater than 0.0952) and thus had an unusual impact on the fitted regression: USGS 50038100, 50038320, 50112500, and 50144000 (table 2.1). However, the impact on the regression was not large enough to warrant exclusion from the skew calculation, and they were included in the analysis.

References Cited

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

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.

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

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

Eash, D.A., Barnes, K.K., and Veilleux, A.G., 2013, Methods for estimating annual exceedance-probability discharges for streams in Iowa, based on data through water year 2010: U.S. Geological Survey Scientific Investigations Report 2013–5086, 63 p., 1 app.

England, J.F., Jr., Cohn, T.A., Faber, B.A., Stedinger, J.R., Thomas, W.O., Jr., Veilleux, A.G., Kiang, J.E., and Mason, R.R., Jr., 2018, Guidelines for determining flood flow frequency—Bulletin 17C: U.S. Geological Survey Techniques and Methods, book 4, chap. B5, 148 p., accessed May 1, 2019, at https://doi.org/10.3133/tm4B5.

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the Southeastern United States, 2006—Volume 3, South Carolina: U.S. Geological Survey Scientific Investigations Report 2009–5156, 226 p.

Gotvald, A.J., Feaster, T.D., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 1, Georgia: U.S. Geological Survey Scientific Investigations Report 2009–5043, 120 p.

Griffis, V.W., 2006, Flood frequency analysis—Bulletin 17, regional information, and climate change: Ithaca, N.Y., Cornell University, Ph.D. dissertation, 246 p.

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

Griffis, V.W., Stedinger, J.R., and Cohn, T.A., 2004, Log Pearson type 3 quantile estimators with regional skew information and low outlier adjustments: Water Resources Research, v. 40, article W07503, 17 p., accessed November 10, 2020, at https://doi.org/10.1029/2003WR002697.

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

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

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

Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood flow frequency: Hydrology Subcommittee Bulletin 17B, 28 p., 14 apps., 1 pl.

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

Martins, E.S., and Stedinger, J.R., 2002, Cross-correlation among estimators of shape: Water Resources Research, v. 38, no. 11, p. 34-1–34-7, accessed November 10, 2020, at https://doi.org/10.1029/2002WR001589.

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

Ramos-Gines, O., 1999, Estimation of magnitude and frequency of floods for streams in Puerto Rico—New empirical models: U.S. Geological Survey Water-Resources Investigations Report 99–4142, 41 p., app.

Reis, D.S., Jr., Stedinger, J.R., and Martins, E.S., 2005, Bayesian generalized least squares regression with application to the log Pearson type III regional skew estimation: Water Resources Research, v. 41, article W10419, 14 p., accessed November 10, 2020, at https://doi.org/10.1029/2004WR003445.

Southard, R.E., and Veilleux, A.G., 2014, Methods for estimating annual exceedance-probability discharges and largest recorded floods for unregulated streams in rural Missouri: U.S. Geological Survey Scientific Investigations Report 2014–5165, 39 p., accessed April 30, 2015, at https://doi.org/10.3133/sir20145165.

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

Stedinger, J., and Griffis, V., 2008, Flood frequency analysis in the United States—Time to update: Journal of Hydrologic Engineering, v. 13, no. 4, p. 199–204.

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

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

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

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

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., accessed October 22, 2015, at https://doi.org/10.3133/fs20133108.

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

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

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

Appendix 3. At-Site, Regression Equation, and Weighted Magnitude, Variance, and Prediction Intervals of Annual Exceedance Probability Floods for Select Unregulated Streamgages in Puerto Rico

The spreadsheet is available for download in .xls and .csv format at https://doi.org/10.3133/sir20215062.

For more information about this publication, contact

Director, Caribbean-Florida Water Science Center
U.S. Geological Survey
4446 Pet Lane, Suite 108
Lutz, FL 33559

For additional information visit
https://www.usgs.gov/centers/car-fl-water

Publishing support provided by
Lafayette Publishing Service Center

ISSN 2328-0328 (online)
https://doi.org/10.3133/sir20215062