USGS
science for a changing world

Prepared in cooperation with the U.S. Army Corps of Engineers

Development of Regional Skew Coefficients for Selected Flood Durations in the Columbia River Basin, Northwestern United States and British Columbia, Canada

Scientific Investigations Report 2020–5073
Version 1.1, October 2020

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

Development of Regional Skew Coefficients for Selected Flood Durations in the Columbia River Basin, Northwestern United States and British Columbia, Canada

By Greg D. Lind, Jonathan R. Lamontagne, and Adam J. Stonewall

Prepared in cooperation with the U.S. Army Corps of Engineers

Scientific Investigations Report 2020–5073
Version 1.1, October 2020

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

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

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

U.S. Geological Survey, Reston, Virginia: 2020
First release: 2020
Revised: October 2020 (ver. 1.1)

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/.

Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Suggested citation:
Lind, G.D., Lamontagne, J.R., and Stonewall, A.J., 2020, Development of regional skew coefficients for selected flood durations in the Columbia River Basin, northwestern United States and British Columbia, Canada (ver. 1.1, October 2020): U.S. Geological Survey Scientific Investigations Report 2020–5073, 48 p., https://doi.org/10.3133/sir20205073.

ISSN 2328-0328 (online)

Contents

Abstract 1
Introduction 1
Purpose and Scope 2
Study Area Description 3
Data Methods 8
Streamflow 7
Flood-Duration Computation 8
Basin and Climatic Characteristics 8
Cross-Correlation Model of Concurrent Flood Durations 11
Flood-Frequency Analysis 14
Flood Frequency Based on Log-Pearson Type III Distribution 14
Expected Moments Algorithm 14
Potentially Influential Small Floods 14
Correlations Between Skew Coefficients and Basin Characteristics 17
Regional Duration—Skew Analysis 21
Standard Bayesian Generalized Least Squares 22
Hybrid Bayesian Weighted Least Squares/Generalized Least Squares Procedure 23
Step 1—Ordinary Least Squares Analysis 23
Step 2—Bayesian Weighted Least Squares 24
Step 3—Bayesian Generalized Least Squares 24
Data Analysis of Duration-Skews for the Columbia River Basin 24
Redundancy Screening 24
Exploratory Data Analysis 26
Initial Screening with Only U.S. Geological Survey Sites 27
Screening with All Non-Redundant Basins 28
Precipitation Models of Regional Skew Coefficients 28
Final Duration-Skew Models 29
Use of Regional Skew Models 32
Summary 35
References Cited 36
Appendix 1. Columbia River Basin Characteristics 41
Appendix 2. Ancillary Tables for Regional Skew Study in the Columbia River Basin 43
Appendix 3. Diagnostic Statistics 45

Figures

1. Map showing Columbia River Basin and 13 subbasins including major rivers, in the northwestern United States and British Columbia, Canada 4
2. Map showing physiographic features of the Columbia River Basin, northwestern United States and British Columbia, Canada 5
3. Map showing mean annual precipitation for the Columbia River Basin, northwestern United States and British Columbia, Canada 6

4. Graph showing normalized daily streamflow, using water years 1928–2008, for major rivers in the Columbia River Basin, northwestern United States and British Columbia, Canada 7
5. Map showing 403 sites selected for analysis in this study in the Columbia River Basin, northwestern United States and British Columbia, Canada....9
6. Scatterplot showing Fisher transformation (Z) model of cross-correlation between concurrent annual 10-day flood durations and the distance between basin centroids, for 257 sites in the Columbia River Basin, northwestern United States and British Columbia, Canada, with at least 50 years of concurrent records 13
7. Graph showing cross-correlation models for selected annual N-day flood durations and distance between basin centroids in the Columbia River Basin, northwestern United States and British Columbia, Canada....13
8. Flood-frequency curve for maximum 3-day duration flows with censored data for the Luckiamute at Suver, Oregon streamgage (U.S. Geological Survey streamgage 14190500)....15
9. Flood-frequency curve for maximum 3-day duration flows with no censored data for the Luckiamute at Suver, Oregon streamgage (U.S. Geological Survey streamgage 14190500)....16
10. Flood-frequency curve for maximum 1-day duration flows for the Metolius River near Grandview, Oregon, streamgage (U.S. Geological Survey streamgage 14091500)....17
11. Flood-frequency curve for maximum 1-day duration flows for the Payette River near Horseshoe Bend, Idaho, streamgage (U.S. Geological Survey streamgage 13247500) 18
12. Graph showing skew coefficients and drainage areas for the 1-day flood durations in the Columbia River Basin, northwestern United States and British Columbia, Canada....18
13. Graphs showing skew coefficients and drainage areas for the selected flood durations in the Columbia River Basin, northwestern United States and British Columbia, Canada....19
14. Graphs showing regressions for skew coefficients and selected basin characteristics of the Columbia River Basin, northwestern United States and British Columbia, Canada 20
15. Boxplot showing skew coefficient distributions for select flood durations in the Columbia River Basin, northwestern United States and British Columbia, Canada....21
16. Graph showing continuous non-linear functions of mean annual precipitation for two non-linear models used in the study....29
17. Scatterplots showing fitted regional skew models for the 1-day, 3-day, 7-day, and 10-day duration floods, Columbia River Basin, northwestern United States and British Columbia, Canada 31
18. Scatterplots showing fitted regional skew models for the 15-, 30-, and 60-day duration floods, Columbia River Basin, northwestern United States and British Columbia, Canada....32
19. Graph showing nonlinear final fitted models of regional skew for all durations in the Columbia River Basin, Columbia River Basin, northwestern United States and British Columbia, Canada 33

Tables

1. Streamflow record information for 403 sites selected for analysis in this study, Columbia River Basin, northwestern United States, and British Columbia, Canada.........9
2. Characteristics of Columbia River Basin used in exploratory analysis for this study, northwestern United States and British Columbia, Canada 10
3. Ordinary least squares regression model coefficients (a, b and c) in equation 2a of cross-correlation of concurrent annual flood durations...12
4. Number and percentage of sites with censored floods for multiple ranges from 403 sites selected for analysis in this study, Columbia River Basin, northwestern United States and British Columbia, Canada..16
5. Coefficients of determination for the relation between drainage area and skew coefficients for selected flood durations in the Columbia River Basin, northwestern United States and British Columbia, Canada..19
6. Relation between standardized distance threshold and the number of identified redundant pairs for the Columbia River Basin, northwestern United States and British Columbia, Canada 25
7. Results of the mixed integer linear programming redundancy screening 26
8. Final fitted models for the seven study durations and several diagnostic statistics used to evaluate model performance 30
9. Pseudo analysis of variance (ANOVA) for the final nonlinear regional skew model for all N-day flood durations for the Columbia River Basin, northwestern United States and British Columbia, Canada 31
10. Regional skew coefficients and variance of prediction for seven durations as a function of mean annual basin precipitation, Columbia River Basin, northwestern United States and British Columbia, Canada..34

Conversion Factors

U.S. customary units to International System of Units

Multiply	By	To obtain
	Length	
inch (in.)	2.54	centimeter (cm)
inch (in.)	25.4	millimeter (mm)
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)

International System of Units to U.S. customary units

Multiply	By	To obtain
	Length	
centimeter (cm)	0.3937	inch (in.)
meter (m)	3.281	foot (ft)
kilometer (km)	0.6214	mile (mi)
meter (m)	1.094	yard (yd)
	Area	
square kilometer (km^2)	0.3861	square mile (mi^2)

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

°F = (1.8 × °C) + 32.

Datum

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

Abbreviations

AEP	annual exceedance probability
ANOVA	analysis of variance
ASEV	average sampling error variance
B-GLS	Bayesian generalized least squares
B-WLS	Bayesian weighted least squares
B-WLS/GLS	Bayesian weighted least squares/generalized least squares
EMA	expected moments algorithm
EVR	error variance ratio
GIS	geographic information system
GLS	generalized least squares
LP3	log-Pearson type III
MBV*	misrepresentation of beta variance
MGBT	Multiple Grubbs-Beck Test
MILP	multiple integer linear programming
MSE	mean square error
NHD	National Hydrography Dataset
NRNI	no-regulation no-irrigation
OLS	ordinary least squares
PILF	potentially influential low flow
PRISM	Parameter Elevation Regression on Independent Slopes Model
Reclamation	Bureau of Reclamation
TDA	The Dalles
USACE	U.S. Army Corps of Engineers
USGS	U.S. Geological Survey
VP	variance of prediction

This page intentionally left blank.

Development of Regional Skew Coefficients for Selected Flood Durations in the Columbia River Basin, Northwestern United States and British Columbia, Canada

By Greg D. Lind,[1] Jonathan R. Lamontagne,[2] and Adam J. Stonewall[1]

Abstract

Flood-frequency (hereinafter frequency) estimates provide information used to design, operate, and maintain hydraulic structures such as bridges and dams. Failures of these structures could cause catastrophic loss of property, life, or both. In addition to frequency estimates that use annual peak streamflow, frequency estimates of flood durations are required to safely and effectively operate the numerous dams in the Columbia River Basin of the northwestern United States, and British Columbia, Canada. Frequency studies rely on U.S. Geological Survey Guidelines for Determining Flood Flow Frequency (Bulletin 17C, published in 2018). A major consideration in estimating frequencies is the use of skew coefficients, which measure the asymmetry of flood flow distributions. Large uncertainties are associated with estimating the at-site skew coefficients directly from streamflow records, which are limited in length. Skew also is sensitive to extreme events for limited record lengths. Bulletin 17C recommends using regional skew coefficients to weight with the at-site skew estimate for more reliable frequency estimates. In this study, streamflow records from 313 unregulated U.S. Geological Survey streamgage sites and 97 regulated sites with naturalized streamflow records provided by the U.S. Army Corps of Engineers were used to develop regional skew models for the Columbia River Basin. The naturalized streamflow records were synthesized by removing regulatory components such as withdrawals and reservoir storage. Skew models were developed for 1-, 3-, 7-, 10-, 15-, 30-, and 60-day flood durations and used to estimate regional skew coefficients for the Columbia River Basin.

This report used Bayesian statistical regression methods to develop and analyze regional skew models based on hydrologically important basin characteristics. After examining a suite of available basin characteristics, mean annual precipitation had the strongest correlation to skew across the flood durations. Regional skew regression models were fit using mean annual precipitation for selected subbasins in the Columbia River Basin.

Introduction

Flood-frequency (hereinafter frequency) estimates provide information used to design, maintain, and operate structures that convey or retain large volumes of streamflow. Local, State, and Federal authorities use this information to avoid potential destruction of property and loss of life. For structures such as bridges and culverts, the frequency estimates of most interest are based on annual peak streamflow records from unregulated streams. Annual peaks, the maximum instantaneous streamflow that occurs in a water year (which starts October 1 of the previous calendar year and extends to September 30 of the named water year), are commonly used in frequency studies to estimate annual exceedance probabilities (AEPs). The 1-percent AEP flood event is identical to the 100-year flood event.

Many public agencies and private companies operating and maintaining water-retention structures, such as dams and levees, need AEP estimates of flood durations rather than annual peaks because the duration of an event can be more damaging to structures than magnitude alone. Flood durations are running averages of daily streamflow over selected time periods, often described as “N-days,” where “N” refers to the number of days during which flooding occurs. N-day duration frequency estimates are required to effectively and safely operate dams and reservoirs, especially if their primary purpose is flood control.

Flood Control Acts of 1936 and 1944 (Public Laws 78–534 and 74-738, respectively), initiated the construction and operation of dams for managing flood risks by the U.S. Army Corps of Engineers (USACE) (Zellmer, 2004; Klein and Zellmer, 2007). In 1948, Vanport, the second largest city in Oregon at the time, was destroyed by catastrophic flooding along the Columbia River (Rantz and Riggs, 1949) and prompted the construction of multiple dams for flood control in the Columbia River Basin of the northwestern United States

[1] U.S. Geological Survey.

[2] Tufts University.

and British Columbia, Canada (McKenzie, 2013). In 1964, the Columbia River Treaty was signed by the Canadian and United States governments, leading to the development of a multiple reservoir system on the Columbia River within both countries, intended to prevent major flooding in the Columbia River Basin (McKenzie, 2013).

The USACE owns, maintains, and operates numerous dams within the Columbia River Basin. Because many of these dams are aging, the USACE began a comprehensive dam-safety evaluation program for the region. Evaluating the effects of flooding on a dam and associated structures requires reliable and accurate estimates of flood magnitude and duration.

Federal agencies conducting frequency studies are recommended to follow the guidelines described in Bulletin 17C (England and others, 2018). Bulletin 17C recommends fitting a log-Pearson type III (LP3) probability distribution to annual flood flow data. Fitting a LP3 distribution involves using three sample moments: mean, standard deviation, and skew. Skew is a measure of asymmetry in a distribution and can significantly affect the magnitudes of frequency estimates, especially for the largest flood events (Veilleux and others, 2019).

To improve frequency estimates, Bulletin 17C recommends weighting the at-site skew coefficient with a regional skew coefficient. A regional skew coefficient can be estimated from a statistical regression of the skew coefficients of streamflow records in an area and their hydrologically significant basin characteristics. For this study, a set of skew regression models was developed to estimate regional N-day duration skew coefficients.

An important consideration in frequency studies is whether a site is affected by flow regulations such as those caused by dams, diversions, and (or) basin condition changes. For this study, “site” refers to an unregulated USGS streamgage or the location of a dam. Heavily regulated streamflow results in a departure from the natural flow regime. For example, a regulated stream may have truncated or attenuated streamflow peaks caused by water storage in an upstream reservoir during flood events, or conversely, during dry periods the streamflow may be augmented by dam releases. The use of streamflow records from regulated streams would violate two assumptions of frequency analysis: (1) that flood events are random, and (2) that basin conditions are relatively constant for the period of record used (England and others, 2003). One way to handle this potential issue is to reconstruct the regulated streamflow record to one that more closely represents the natural streamflow record, or the streamflow record that would exist without regulations. The USACE, the Bureau of Reclamation (Reclamation), and the Bonneville Power Administration provided reconstructed, or naturalized, daily no-regulation no-irrigation (NRNI) streamflow records for sites in this study affected by regulation (Bonneville Power Administration, 2011; U.S. Army Corps of Engineers, 2014; K. Duffy, U.S. Army Corps of Engineers, written commun., 2017).

This study is the result of a joint effort between the USGS, USACE, and Tufts University. To the best of the authors’ knowledge, no recent frequency studies have been done for the entire Columbia River Basin. A study on the magnitude and frequency of floods in the Columbia River Basin was completed by Rantz and Riggs (1949) and a similar study by Hulsing and Kallio (1964) focused on the lower Columbia River Basin. A regional skew study was completed for the Pacific Northwest, which included the states of Idaho, Oregon, and Washington along with part of western Montana (Wood and others, 2016). Frequency studies also have been completed for parts of the Columbia River Basin in Idaho (Wood and others, 2016), Washington (Mastin and others, 2016), and Oregon (Cooper, 2005, 2006). These recent frequency studies focused on annual peak flow records rather than flood durations.

Purpose and Scope

The primary goal of this study is to develop 1-, 3-, 7-, 10-, 15-, 30-, and 60-day regional skew models for the Columbia River Basin. Currently (2020), no flood-duration regional skew models have been developed for the entire basin. The regional skew coefficients computed from these models can be used to estimate AEPs of selected flood durations in the Columbia River Basin. This report presents flood-duration frequency statistics for estimates corresponding to the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent AEPs at selected sites in the Columbia River Basin for 1-, 3-, 7-, 10-, 15-, 30-, and 60-day durations. These estimated flood-duration frequencies will be used by the USACE; other Federal, State, and local agencies; and utilities to safely and effectively operate and maintain dams and reservoirs along the Columbia River system.

The objectives of this study include the following:

1. Compile and review long-term naturalized daily streamflow records at regulated sites in the Columbia River Basin.

2. Compile measured daily streamflow records at available unregulated sites in the Columbia River Basin.

3. Create regional skew models for 1-, 3-, 7-, 10-, 15-, 30-, and 60-day flood durations at regulated and unregulated sites in the Columbia River Basin.

4. Estimate 1-, 3-, 7-, 10-, 15-, 30-, and 60-day flood-duration frequency statistics at regulated and unregulated sites in the Columbia River Basin using regional skew coefficients developed during this study.

Study Area Description

The Columbia River Basin covers 259,000 square miles including part of the Canadian province of British Columbia and parts of seven U.S. states: Washington, Idaho, Montana, Oregon, Wyoming, Nevada, and Utah (fig. 1). The Columbia River Basin can be divided into 13 subbasins:

1. Upper Columbia,
2. Kootenai,
3. Pend Oreille,
4. Spokane,
5. Middle Columbia,
6. Yakima,
7. Upper Snake,
8. Middle Snake,
9. Lower Snake,
10. Deschutes,
11. Lower Columbia,
12. Willamette, and
13. Main-stem Columbia.

The headwaters of the upper Columbia River Basin are situated between the Rocky Mountains to the east and the Columbia and Purcell Mountains to the west in British Columbia, Canada (fig. 2). The Columbia River flows for about 1,240 miles from its headwaters in Canada to its mouth at the Pacific Ocean between the border of Washington and Oregon.

The topography varies considerably throughout the Columbia River Basin because of numerous mountain ranges and plateaus (fig. 2). The most prominent mountain ranges are the Rocky Mountains, which border the Columbia River Basin to the east, and the Cascade Range to the west. Other notable mountains are the Columbia Mountains in Canada, which form the northern border; the Bitterroot Range between Idaho and Montana; and the Blue Mountains in eastern Oregon. The Columbia Plateau in eastern Washington, the Snake River Plain in Idaho, the Willamette Valley in western Oregon, and the Yakima Valley in central Washington are predominant flat terrain areas in the Columbia River Basin.

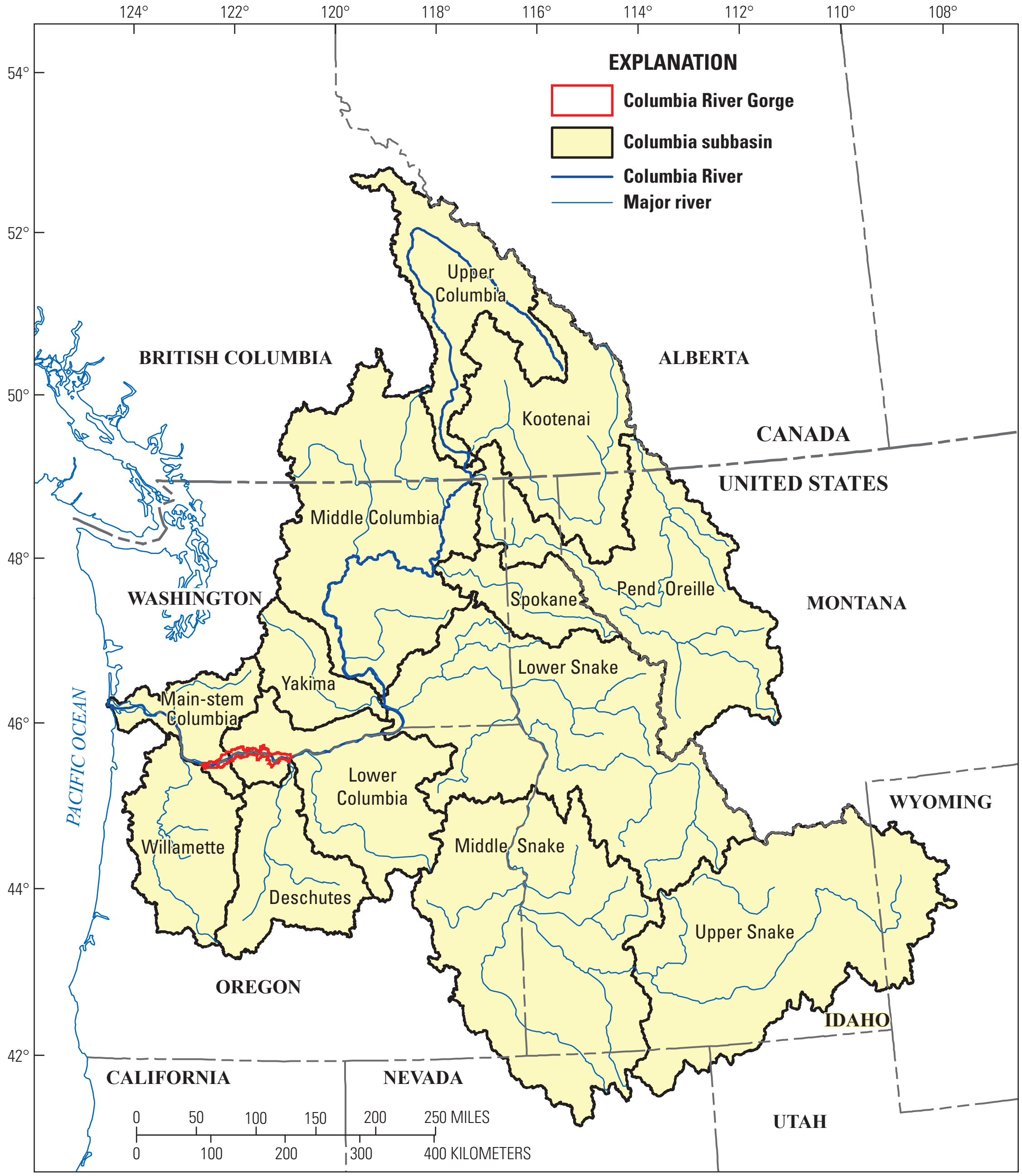
Precipitation and climate vary across the Columbia River Basin (fig. 3). The climate in the Columbia River Basin is the result of three types of air masses (Ferguson, 1999): (1) moist marine air coming off the Pacific Ocean, which helps to moderate temperatures (2) dry continental air from the south and the east; and (3) dry artic air from the north. Marine air masses affect the basin west of the Cascade Range. Warm moist subtropical air can flow into the Pacific Northwest bringing large amounts of rain that can last for multiple days and are termed atmospheric rivers (Dettinger and others, 2011; Rutz and Steenburgh, 2012). These types of events usually take place from October through April.

The Cascade Range blocks marine air masses coming from the west except when winds are strong enough to push these air masses over the mountains, which often happens during the winter (Ferguson, 1999). The Cascade Range causes a rain shadow effect; the east side of the range is drier than the west side. The Columbia River Gorge provides a corridor allowing Pacific air masses to flow from the Cascade Range into the Snake River subbasin. These Pacific air masses can affect the climate in these valleys into the spring (Ferguson, 1999). In winter, much of the precipitation in higher elevations on both sides of the Cascade Range falls as snow.

Continental air masses from the east and south dominate the middle to southern parts of the Columbia River Basin east of the Cascade Range. These air masses are cold and dry in the winter and hot and dry in the summer. Convective thunderstorms can produce intense rain events with short durations (usually <1 day), especially in the spring and summer (Cooper, 2006). The northern part of the Columbia River Basin east of the Cascade Range is largely influenced by arctic air masses from the north. Arctic air masses tend to be dry and bring cold air from the north in winter and keep temperatures relatively moderate in the summer.

The largest floods along the Columbia River have been the result of snowmelt and large rain-on-snow events. The largest gaged flood ever recorded on the Columbia River, occurring in 1894, and the 1948 flood on the main-stem Columbia River, which destroyed the city of Vanport, Oregon (Rantz and Riggs, 1949), were both caused by rain on snow and snowmelt. Small subbasins may have floods resulting from convective thunderstorms in the summer, especially for subbasins east of the Cascade Range. East of the Cascade Range, most widespread flooding is caused by snowmelt. Large rainstorms are the most common cause of flooding west of the Cascade Range, which may be combined with snowmelt. Longer duration rainfall events (30 and 60 days) are likely caused by a combination of more than one hydrologic event.

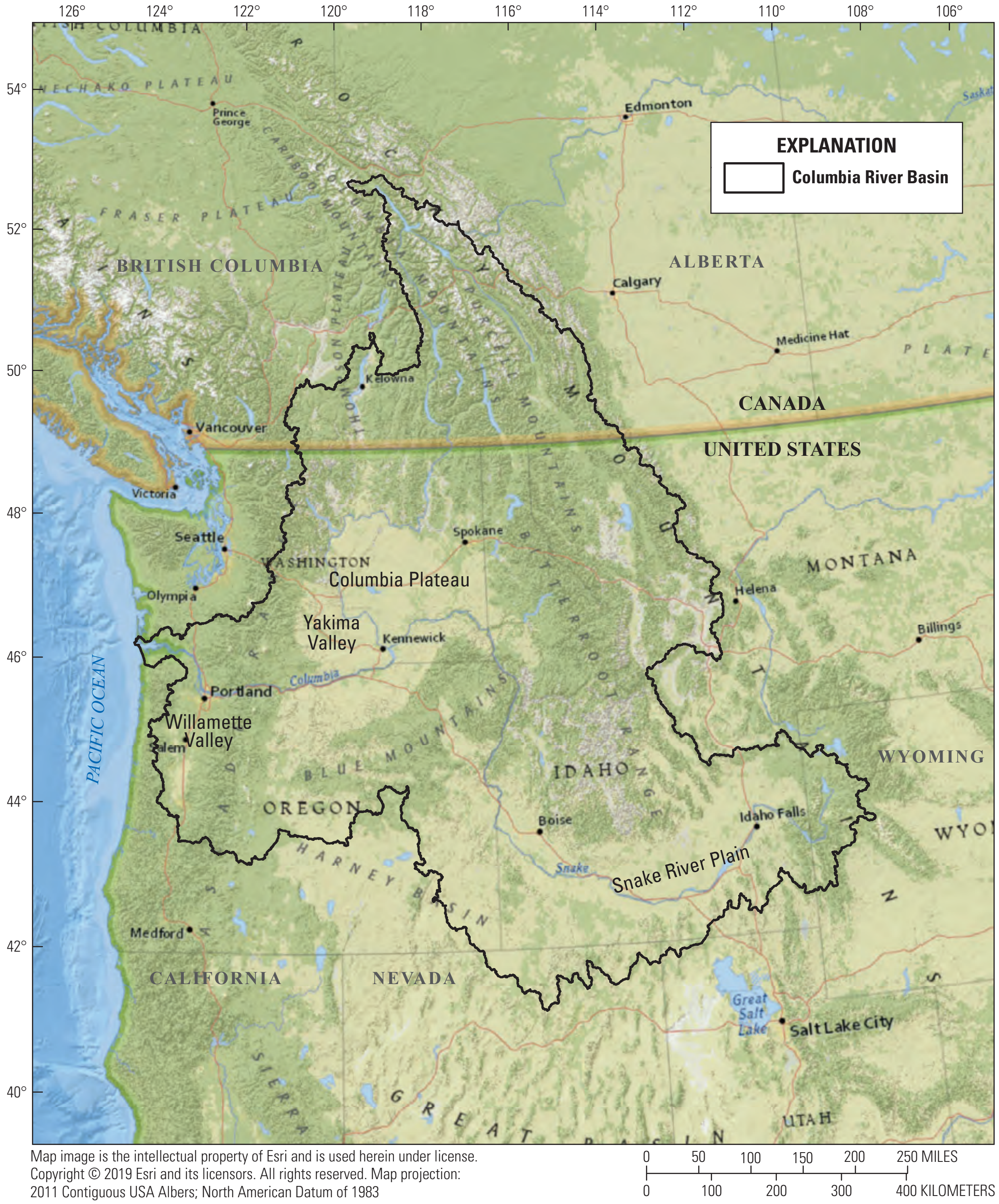
Floods in the western United States can be classified as snowmelt-dominated, rain-dominated, or transient (meaning a combination of rain and snowmelt; Hamlet and Lettenmaier, 2007). Basins that are transient have a mixed population of floods caused by rain, snowmelt, or rain-on-snow events (the combination of both). Whereas much of the Columbia River Basin has flooding from snowmelt, some areas are rain-dominated or transient. An analysis was done to investigate areas in the Columbia River Basin that may have mixed populations of floods by cyclically plotting daily streamflow data over all 12 months (fig. 4). Most high streamflow occurs during the months of May and June, indicating that snowmelt is likely the cause of these flows. Outside of these months, most basins have relatively low streamflow with the exceptions of the Spokane, Yakima and Willamette subbasins.



Base from U.S. Geological Survey, North American Atlas Political Boundaries dataset, 2010, 1:10,000,000; Map projection: 2011 Contiguous USA Albers; North American Datum of 1983

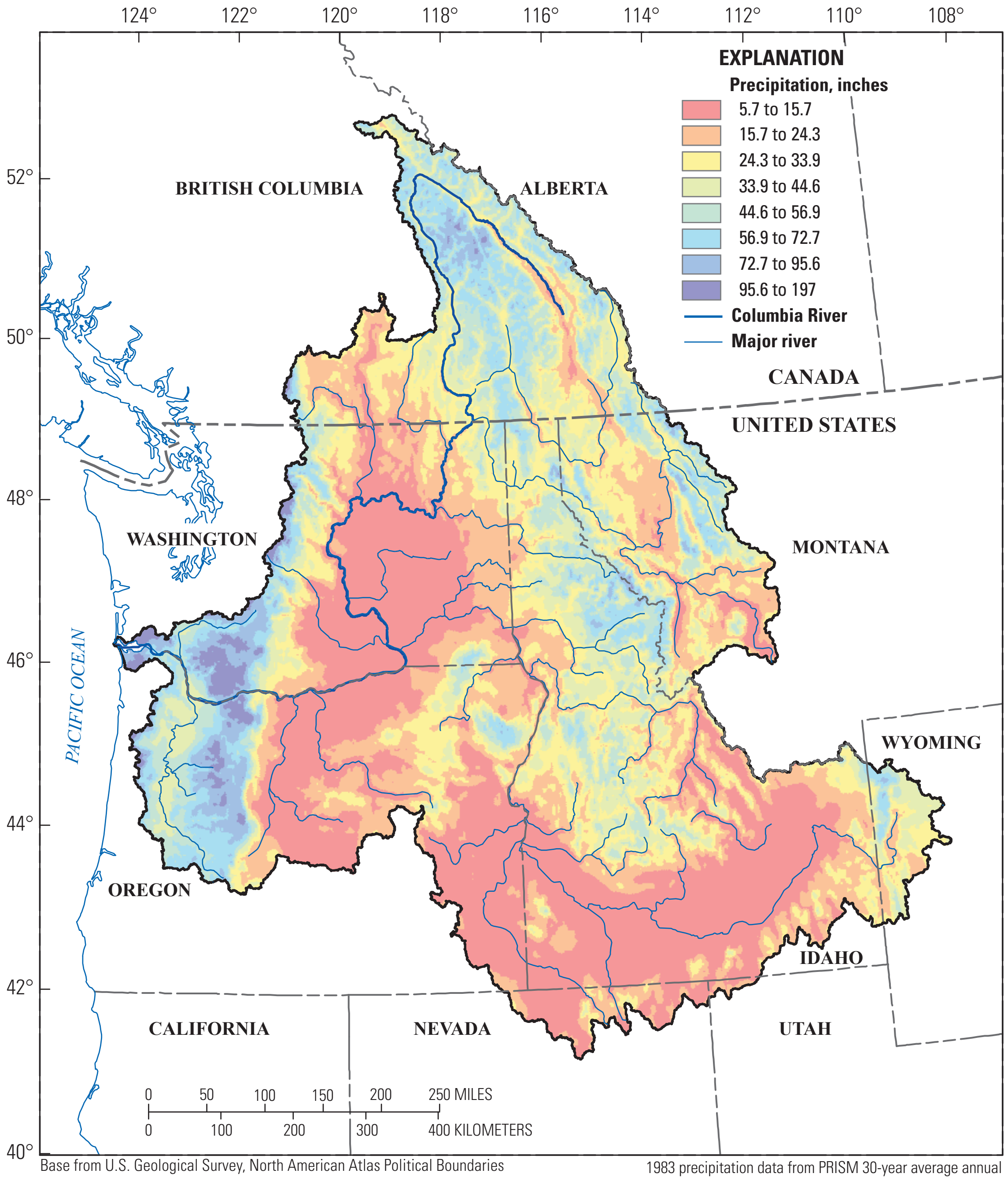
Subbasins from U.S. Geological Survey National Watershed Boundary dataset, 2013

Figure 1. Columbia River Basin and 13 subbasins including major rivers, in the northwestern United States and British Columbia, Canada.



Map image is the intellectual property of Esri and is used herein under license. Copyright © 2019 Esri and its licensors. All rights reserved. Map projection: 2011 Contiguous USA Albers; North American Datum of 1983

Figure 2. Physiographic features of the Columbia River Basin, northwestern United States and British Columbia, Canada.



Base from U.S. Geological Survey, North American Atlas Political Boundaries dataset, 2010, 1:10,000,000; Map projection: 2011 Contiguous USA Albers; North American Datum of 1983

1983 precipitation data from PRISM 30-year average annual precipitation, 1971–2000, Hamman and others, 2013

Figure 3. Mean annual precipitation for the Columbia River Basin, northwestern United States and British Columbia, Canada.

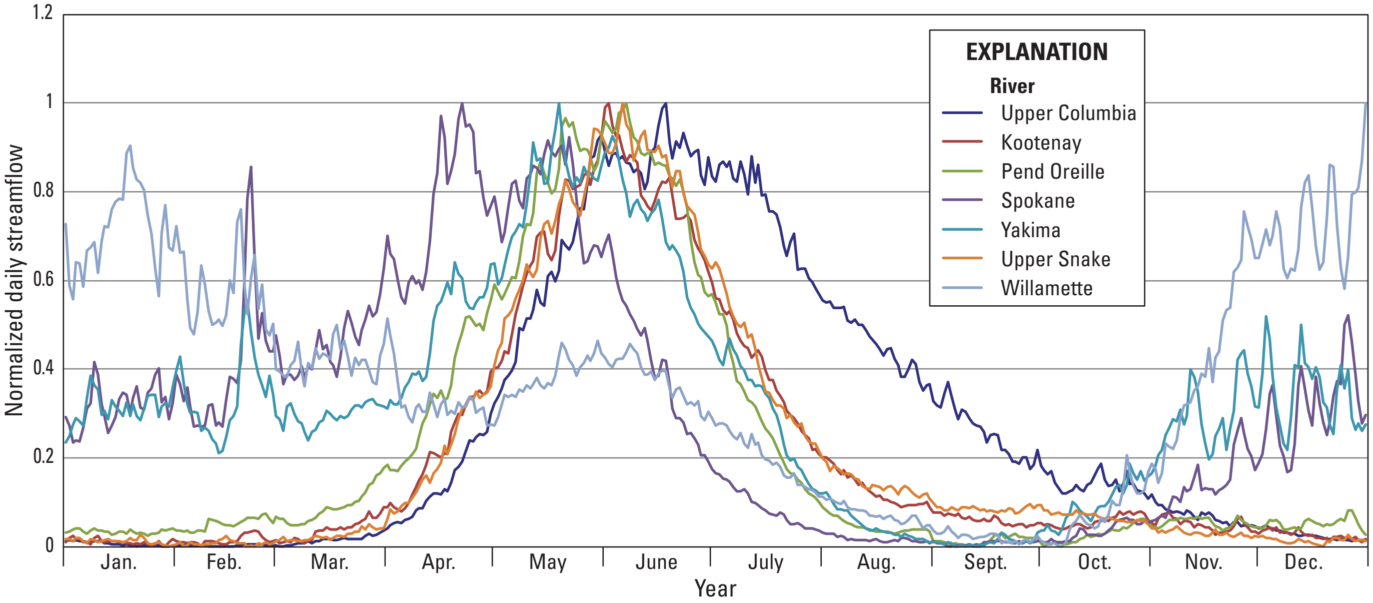


Figure 4. Normalized daily streamflow, using water years 1928–2008, for major rivers in the Columbia River Basin, northwestern United States and British Columbia, Canada. A value of 1 in the y-axis corresponds to the maximum daily streamflow.

The Spokane subbasin receives most of its annual precipitation from October through April as both rain and snow. High streamflow occurs from December through June, with the highest streamflow usually occurring in May. Significant snowmelt occurs from April through June and the streamflow exceeds the amount of precipitation received during those months (Fu and others, 2007). Marine air from the Pacific Ocean and continental air masses affect the Spokane subbasin climate (Northwest Power Planning Council, 2000).

Most of the annual precipitation for the Yakima subbasin occurs during the months of October through March (Bureau of Reclamation, 2002), and in the higher elevations much of the precipitation falls as snow. High streamflow occurs from April through June caused by snowmelt. The highest streamflows are caused by snowmelt and rain-on-snow events. Marine air passing over the Cascade Range can cause rain-on-snow events in the Yakima subbasin (Rinella and others, 1992).

The Willamette subbasin receives most precipitation from October through April. In the western part of the subbasin, the climate is influenced by marine air from the Pacific Ocean and winter rain events that produce the highest streamflow (Cooper, 2005). In the eastern part of the subbasin, the highest sustained flows are the result of spring snowmelt from the western Cascades and the largest peak flows occur in the winter as a result of rain-on-snow events. Atmospheric rivers are the dominant cause of floods west of the Cascade Range.

After this initial analysis, we thought that it may be beneficial to fit regional skew models to individual subbasins. During subsequent analyses, we decided that there were not enough sites in the subbasins to do adequate regression analyses (see section, “Initial Screening with only USGS sites”).

Data Methods

Streamflow

The reconstructed NRNI daily streamflow records were reviewed and rated based on the quality of each record. The NRNI streamflow records are a modification of the 2010 level modified streamflow records (Bonneville Power Administration, 2011). Both datasets and their accompanying reports are available at . NRNI streamflow records account for reservoir regulation and irrigation withdrawals. Regulatory effects, such as diversions or reservoir storage, were removed during the reconstruction process to produce streamflow records that are more representative of unaltered streamflow conditions. Calculations made to datasets during the process of naturalizing the records were verified for accuracy using the equations provided by the USACE (Bonneville Power Administration, 2011; U.S. Army Corps of Engineers, 2014). The 2010 evaporation and water-use data from outside sources, such as the National Oceanic and Atmospheric Administration and the Oregon Water Resources Department, were used to compare with NRNI components for evaporation and depletion. If the source of data was a USGS streamgage, its accuracy was checked by comparing data downloaded from the National Water Information System (U.S. Geological Survey, 2017) for the streamgage corresponding to the NRNI streamflow data. Changes in reservoir levels and storage capacity tables were compared with values used to synthesize the streamflow records. As a check for consistency, NRNI datasets were plotted along with the measured datasets of regulated streamflow and visually inspected for consistency. After a record was reviewed, a determination was made as to whether it was usable or not, based on the quality of the reconstructed streamflow data. All NRNI streamflow records included in this study were deemed usable following review.

The remainder of sites used in this study were unregulated USGS streamgages. These streamgages are from the USGS Geospatial Attributes of Gages for Evaluating Streamflow, version II report database (GAGES II; Falcone, 2011). The GAGES II study analyzed and classified streamgages operated by the USGS while providing a geospatial database of basin characteristics. Streamgages were analyzed and classified based on the amount of regulation occurring upstream from the streamgage and the extent of anthropogenically altered hydrology in their basins. Only streamgages that were considered unregulated with minimally altered hydrology were used for this study.

Four hundred ten sites within the Columbia River Basin were considered for this study; 97 NRNI sites and 313 unregulated USGS streamgages. Three pairs of NRNI sites—(1) Detroit Dam (DET) and Big Cliff Dam (BCL), (2) Dexter Lake (DEX) and Lookout Point Dam (LOP), and (3) Heise, Idaho (HEI), and Lorenzo, Idaho (LOR)—had similar drainage areas and streamflow records. One out of each pair of these highly correlated sites was retained; BCL, LOP, and HEI were retained, whereas the other three sites were dropped from the analysis. Four of the unregulated USGS sites also were dropped: (1) USGS streamgage 12350250, Bitterroot River at Bell Crossing near Victor, Montana; (2) USGS streamgage 12387450, Valley Creek near Arlee, Montana; (3) USGS streamgage 12433542, Blue Creek above Midnite Mine Drainage near Wellpinit, Washington; and (4) USGS streamgage 12512500, Providence Coulee at Cunningham, Washington. These sites were dropped because their records contained gaps during potentially high streamflow periods and, therefore, some floods were potentially absent from their records. For the 403 sites selected for further analysis, 309 sites were unregulated USGS sites and the other 94 sites were NRNI sites. For this study, NRNI headwater sites have no dam or reservoir upstream from them, whereas NRNI sites with at least one dam upstream were defined as local sites. Of the 94 NRNI sites, 33 were headwater and 61 were local (fig. 5). Record reconstructions of local sites were more complicated, with at least one more point of regulation than headwater sites to account for. Information about the records at the 403 sites selected for analysis is presented in table 1. Almost all of the NRNI streamflow records were 80 years in length (1928–2008), whereas the USGS streamflow records ranged from 16 to 110 years in length..

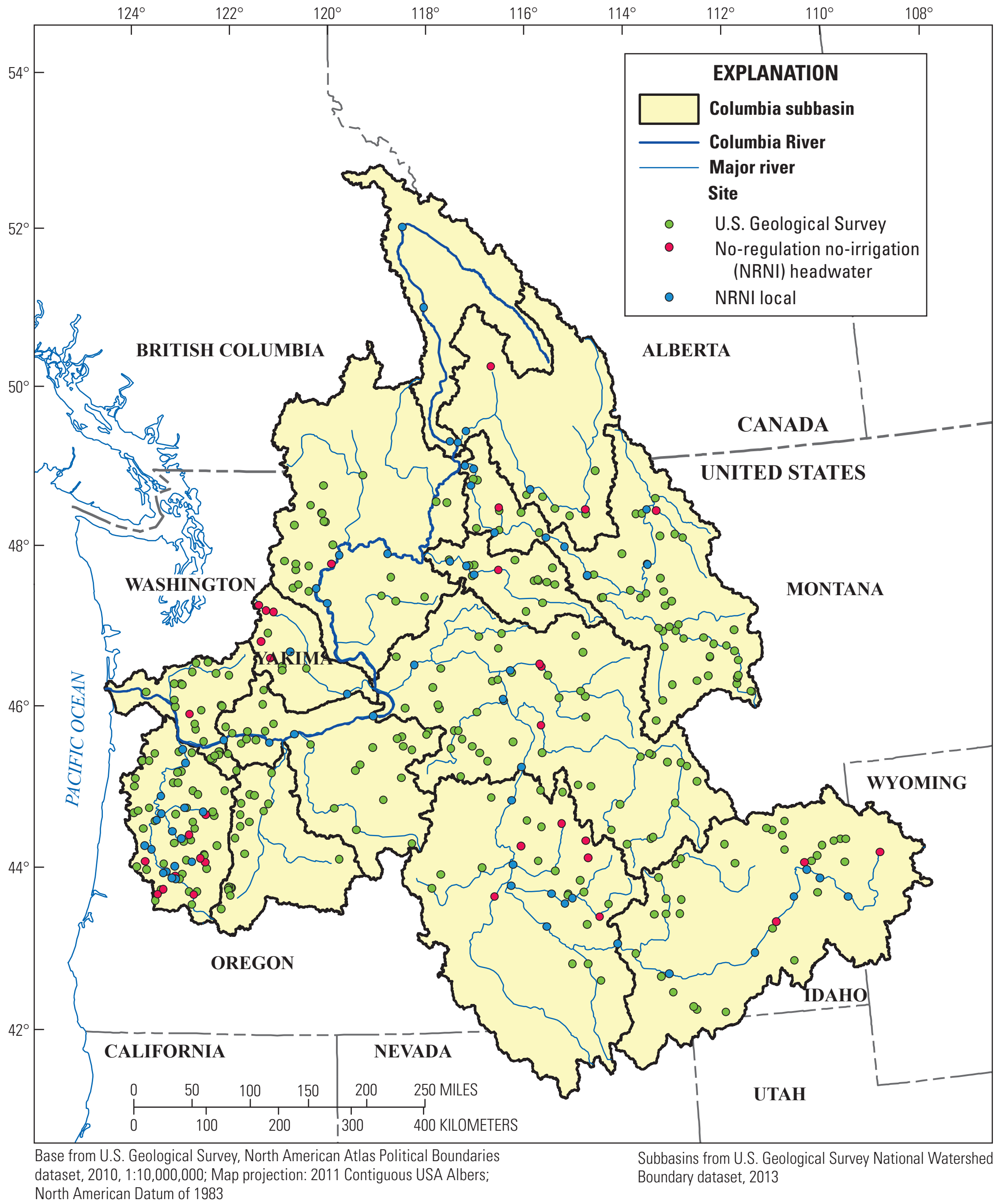
Flood-Duration Computation

Selected flood durations were computed from the daily streamflow records using three methods. GW Toolbox is a USGS-developed software package that allows users the option to compute selected N-day flood durations after inputting a daily mean streamflow record (Barlow and others, 2014). The program will output the maximum selected N-day flood duration for each complete water year from the daily streamflow record. With previously reviewed and approved NRNI daily streamflow records furnished by the USACE and the Reclamation, along with the USGS unregulated streamflow records, GW Toolbox was used to compute the N-day flood durations. A Microsoft Excel template and a script written in the R programming language (R Core Team, 2017) also were created to compute N-day flood durations as checks against the GW Toolbox output files. The three methods produced identical results, and no further data manipulation was deemed necessary.

Basin and Climatic Characteristics

Some of the basin characteristics thought to have a strong relation with skew were precipitation, elevation, temperature, percentage of precipitation falling as snow, aspect, basin compactness, and drainage areas. Mean annual precipitation and mean monthly precipitation for all 12 months were used in this study. Mean annual temperature and mean monthly temperature for all 12 months also were used. Maximum, minimum, and mean basin elevations were included in the analyses as well.

Most of the basin characteristics used in this study were from the GAGES II database (Falcone, 2011). These datasets came from various sources including the Parameter Elevation Regression on Independent Slopes Model (PRISM; PRISM Climate Group, 2017) statistical mapping system, the USGS digital elevation model (U.S. Geological Survey, 2018), and the National Hydrography Dataset (NHD; Horizon Systems Corporation, 2017). The basin characteristics used in this study and their explanations and sources are listed in table 2.



Base from U.S. Geological Survey, North American Atlas Political Boundaries dataset, 2010, 1:10,000,000; Map projection: 2011 Contiguous USA Albers; North American Datum of 1983

Subbasins from U.S. Geological Survey National Watershed Boundary dataset, 2013

Figure 5. The 403 sites selected for analysis in this study in the Columbia River Basin, northwestern United States and British Columbia, Canada.

Table 1. Streamflow record information for 403 sites selected for analysis in this study, Columbia River Basin, northwestern United States, and British Columbia, Canada.

Table 1 is a .csv file available for download at https://doi.org/10.3133/sir20205073.

Table 2. Characteristics of Columbia River Basin used in exploratory analysis for this study, northwestern United States and British Columbia, Canada.

[**Source:** NAWQA, National Water-Quality Assessment; NWIS, National Water Information System; NHDPlus, National Hydrography Dataset Plus; PRISM, Parameter Elevation Regression on Independent Slopes Model USGS, 30m in DEM, 30 meters in digital elevation models. **Abbreviations:** GAGES II, USGS Geospatial Attributes of Gages for Evaluating Streamflow, version II report database; NRNI, no-regulation no-irrigation; USGS, U.S. Geological Survey]

Name	Description	Source
	Only for USGS sites from GAGES II database	
Aspect North	Aspect "northness," ranging from -1 to 1. A value of 1 means that the basin faces north	USGS
Aspect East	Aspect "eastness," ranging from -1 to 1. A value of 1 means that the basin faces east	USGS
El. Median	Median watershed elevation (meters) from 100-meter National Elevation Dataset	USGS
El. Std. Dev.	Standard deviation of elevation (meters) across the watershed	USGS
El. Site	Elevation at gage location (meters) from 100-meter National Elevation Dataset	USGS
RR Mean	Dimensionless elevation - relief ratio, (mean elevation - minimum elevation)/ Relief	USGS
RR Median	Dimensionless elevation - relief ratio, (Median elevation-minimum elevation/ Relief	USGS
Ppt1	Mean monthly precipitation (centimeters) received in the basin for each month (Ppt1 is for Jan., Ppt 12 is for Dec.)	PRISM (1971–2000)
Temp1	Mean monthly temperature (degrees Celsius) of the basin for each month (Temp1 is for Jan. Temp2 is for Dec.)	PRISM (1971–2000)
	For USGS sites and most NRNI sites	
Aspect	Mean basin aspect, degrees (degrees of the compass, 0–360)	USGS
Slope Percent	Mean basin slope, percent, from 100-meter National Elevation Dataset	USGS
BFI	Base Flow Index; the ratio of base flow to total streamflow	USGS
Stream Density	Kilometers of streams per basin square kilometer	NHDPlus
Impervious	Percentage of total basin area that is considered impervious	USGS NLCD06
	For all sites (USGS and NRNI)	
Basin Area	Delineated basin area upstream from the site (square kilometers)	From 30 m DEM
Lat	Latitude of the site location in decimal degrees	USGS
Long	Longitude of the site location in decimal degrees	USGS
Lat. Cen.	Latitude of basin centroid location	From 30 m DEM
Long. Cen.	Longitude of basin centroid location	From 30 m DEM
El. Mean	Mean elevation of the basin (meters), NAD 83	USGS
El. Max	Maximum elevation of the basin (meters) NAD 83	USGS
El. Min	Minimum elevation of the basin (meters) NAD 83	USGS
Relief	Maximum elevation minus the minimum elevation (meters)	USGS
Basin Compact.	Basin compactness ratio, = basin area/basin perimeter^2 * 100; higher number = more compact shape	USGS NWIS and NAWQA
Annual Ppt.	Mean annual precipitation (centimeters) for the basin	PRISM (1971–2000)
Annual Temp.	Mean annual temperature (degrees Celsius) for the basin	PRISM (1971–2000)
Percent Snow	Percentage of total mean annual precipitation that falls in the form of snow	PRISM (1971–2000)

Several sites were missing some of the basin characteristics and their basins had to be delineated before computing the missing characteristics. One method used to delineate basins in the United States involved the USGS web-based software application, StreamStats (Ries and others, 2017). StreamStats has a delineation tool that automatically delineates basins in States for which StreamStats has been fully implemented. Sites with basins in Montana required a different delineation method because StreamStats is not fully implemented for that State. For basins in Montana, NHD flow lines were used along with the BasinDelineator Tool (Horizon Systems Corporation, 2017). The BasinDelineator Tool returns a shapefile of the delineated basin for a user-specified outlet point in the NHD network. For Canadian basins, a geodatabase of flowlines delineated by basins was provided by a geographic information system (GIS) specialist from Environment Canada. After delineating these basins, their missing characteristics were computed from downloaded digital elevation models (DEM) and PRISM climate datasets using ArcMap™ GIS tools such as Conversion, Data Management, and Spatial Analyst Tools (Esri, 2016). In appendix 1, table 1.1 presents basin characteristics available for USGS sites, table 1.2 presents basin characteristics available for USGS and some NRNI sites, and table 1.3 presents basin characteristics for all the NRNI sites.

Cross-Correlation Model of Concurrent Flood Durations

An important consideration in flood-frequency studies is whether cross-correlations between streamflow records exist. If streamflow records from multiple sites are highly correlated, they likely share the same hydrologic events and are influenced by similar factors of those events. In this case, these sites do not represent truly independent samples and should not be treated as such. A suitable model of cross-correlation of annual maximum N-day flood-duration flows at different sites is needed for regional skew studies. Such a cross-correlation model can be used to estimate the cross-correlation between the skew coefficients at individual sites. An accurate model of regional cross-correlation is needed to determine model uncertainty, which is a function of record length. Increasing amounts of regional cross-correlation reduces the effective record length for a region and results in greater model uncertainty, as the individual flood-duration samples are less independent. The effective record length is the at-site record length needed to calculate a skew coefficient with the same variance as the variance of prediction for the regional skew model (Lamontagne and others, 2012) and is a measure of the reduced reliability of an estimator due to serial correlation (Tasker, 1983). Bulletin 17C recommends that effective record length concepts be used to correct uncertainty estimates when serial correlation exists.

Basins that are spatially close, nested within other watersheds, and (or) within similar meteorological storm patterns typically have the same storm systems. This results in similar hydrologic conditions and, consequently, a relatively high degree of cross-correlation among concurrent N-day flood durations. Conversely, basins that are geographically farther apart are more likely to have different meteorological events and less cross-correlation among concurrent N-day flood durations. Previous studies have modeled cross-correlation between annual peaks or flood durations as a function of distance between basin centroids (Gruber and Stedinger, 2008; Parrett and others, 2011; Lamontagne and others, 2012).

In this study, an assumption was made that cross-correlation would be greater between sites for long-duration floods than for short-duration floods. Shorter-duration floods (1 or 3 days, for example) are likely to vary more in intensity, cover smaller geographic areas, and occur at more sporadic intervals than longer duration floods. Conversely, longer-duration floods (such as 30 or 60 days) are likely to vary less in intensity, cover greater geographic areas, and occur at more regular intervals than shorter-duration floods. Even if long- and short-duration floods occur in the same basin for a given water year from the same event, averaging runoff over a longer duration often should result in a dampening of spatial and temporal variability (Lamontagne and others, 2012).

A cross-correlation model was developed for each N-day duration flood using the methodology outlined in Lamontagne and others (2012). Only 257 sites with 50 or more years of concurrent record were used for the analysis. The sampled cross-correlations of annual N-day flood durations between site pairs (i,j), denoted as r_{ij}, were transformed from the [−1, +1] to the (−∞, +∞) range with a logit model using the Fisher Z Transform (Rodgers and Nicewander, 1988; fig. 6):

$$Z_{ij} \;=\; 0.5\, ln\left[\frac{1+r_{ij}}{1-r_{ij}}\right]. \qquad (1)$$

The symbols used in this report for reference are shown in appendix 2, table 2.1. The transformed variable was described by a model with one explanatory variable—the distance between watershed centroids, denoted as d_{ij}:

$$Z_{ij} \;=\; a+\exp\left(b-\frac{c\times d_{ij}}{100}\right)\,, \qquad (2a)$$

where

d_{ij} is the distance between centroids of the watersheds i and j; and a (no units), b (no units), and c (inverse to the units of d) are model coefficients used for best fit.

The cross-correlation between annual N-day flood durations at any two sites, denoted as ρ_{ij}, can then be calculated using the reverse transform:

$$\rho_{ij} = \frac{\exp(2\,Z_{ij})-1}{\exp(2\,Z_{ij})+1}\,. \qquad (2b)$$

This model for cross-correlation is used in section, "Regional Duration—Skew Analysis," to develop models for skew coefficients over a region, as a function of basin characteristics. It is similar to those used in previous California and southeastern U.S. annual peak and flood-duration studies (Gotvald and others, 2009; Parrett and others, 2011; Lamontagne and others, 2012). The cross-correlation model for each duration was fit using ordinary non-linear least squares regression. The parameters for the 1-, 3-, 7-, 15-, 30- and 60-day flood-duration models are presented in table 3. The fitted Fisher Z transformed cross-correlation model and the distance between basin centroids for the 458 site pairs from a total of 257 streamflow sites for the 10-day flood-duration flows are shown in figure 6. Fischer transformation values decrease quickly with distance between zero and 200 km, then level out at greater distances.

For each of the N-day duration floods, cross-correlation tends to decrease with distance (fig. 7). Additionally, longer N-day duration floods had more cross-correlation than shorter N-day duration floods. Both results were similar to those reported in Lamontagne and others (2012).

Table 3. Ordinary least squares regression model coefficients (a, b and c) in equation 2a of cross-correlation of concurrent annual flood durations.

[Coefficients a and b are unitless. Coefficient c has the unit km^{-1} or 1 ÷ kilometers]

Duration (days)	Coefficients		
	a	b	c (km^{-1})
1	0.264383	0.079488	0.56175
3	0.318009	0.145113	0.627384
7	0.38618	0.258339	0.833429
10	0.421767	0.309753	0.957896
15	0.448968	0.380756	1.08721
30	0.466667	0.45446	1.11916
60	0.52461	0.407531	1.02586

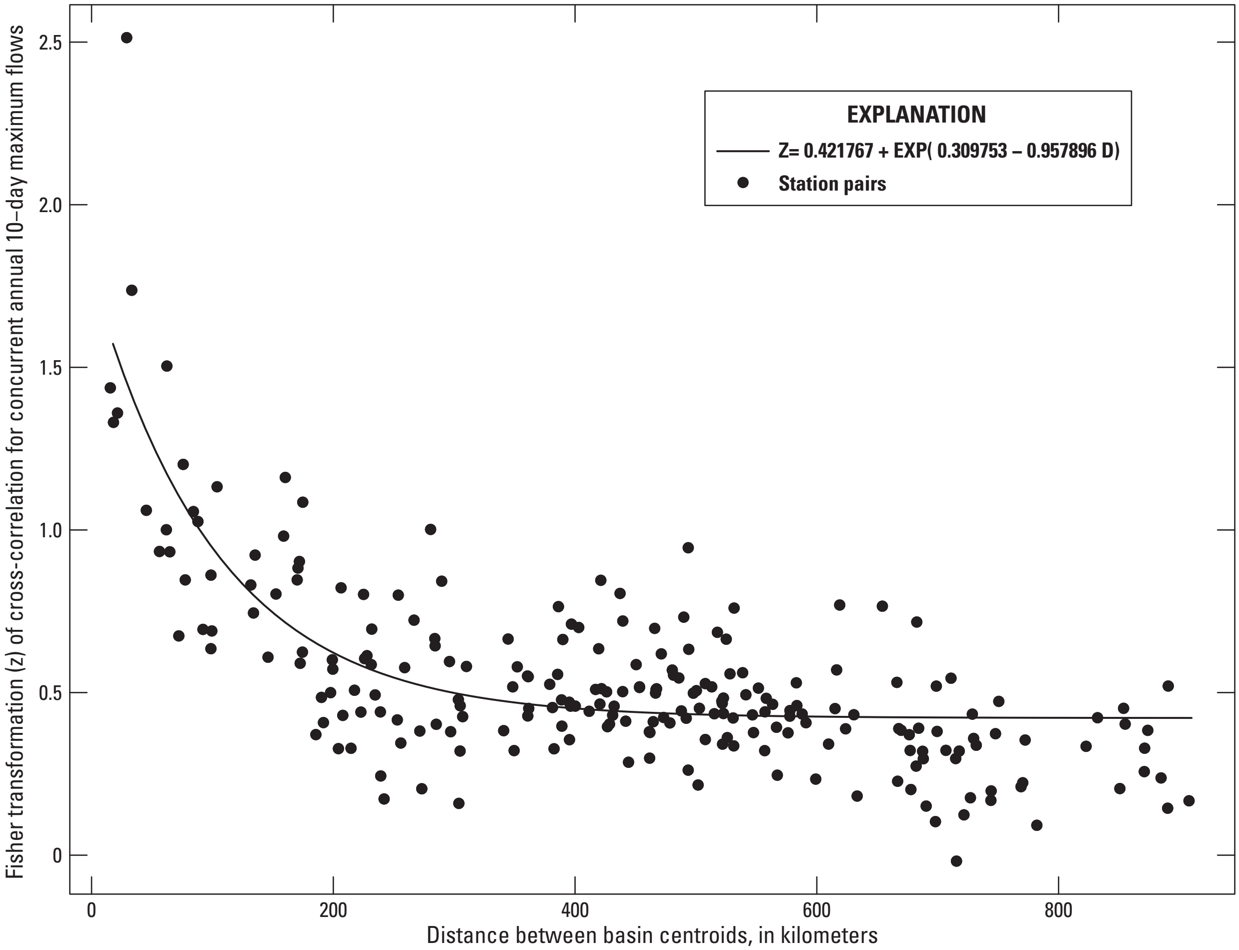


Figure 6. Fisher transformation (Z) model of cross-correlation between concurrent annual 10-day flood durations and the distance between basin centroids, for 257 sites in the Columbia River Basin, northwestern United States and British Columbia, Canada, with at least 50 years of concurrent records.

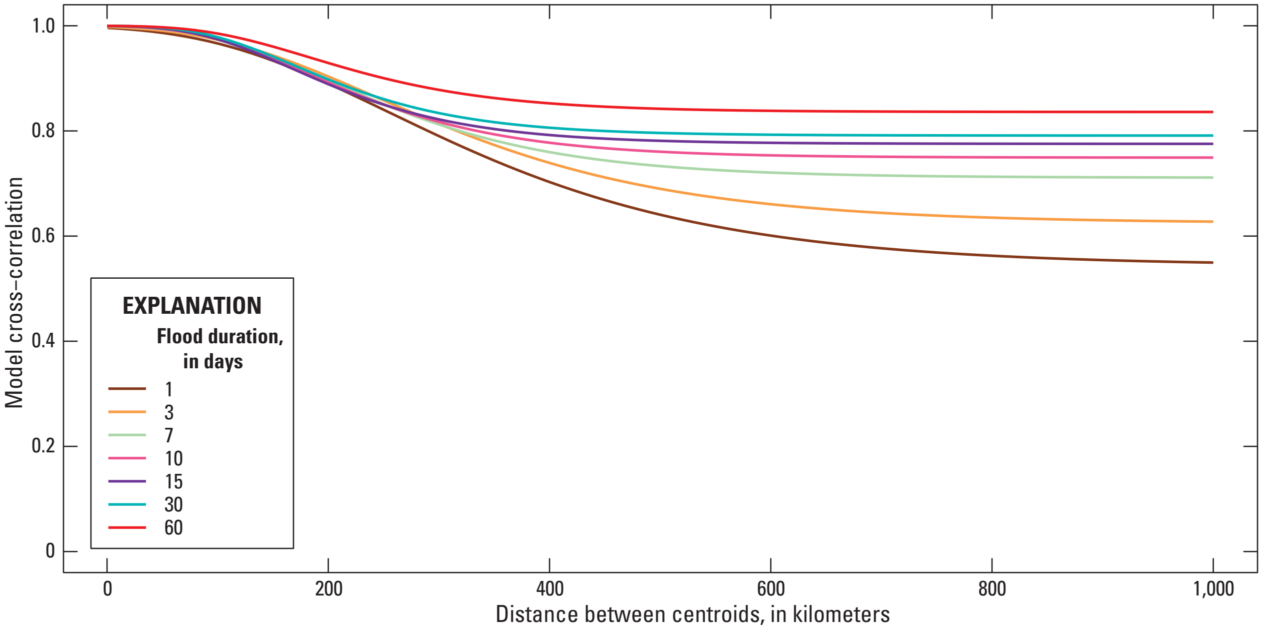


Figure 7. Graph showing cross-correlation models for selected annual N-day flood durations and distance between basin centroids in the Columbia River Basin, northwestern United States and British Columbia, Canada.

Flood-Frequency Analysis

A standard flood-frequency analysis is performed in part by fitting a continuous probability distribution to a series of discharges, typically annual maxima, allowing estimates of flood quantiles to be computed. Flood-frequency quantiles typically are reported as a T-year streamflow, where 1/T is the probability of a given magnitude of streamflow being equaled or exceeded for any 1 year. For example, a 100-year flood magnitude has a 1/100 (0.01 or 1.0 percent) chance of being equaled or exceeded for any given year. This value also is known as an annual exceedance probability (AEP). The recommended USGS approach for conducting a flood-frequency analysis is described in Bulletin 17C (England and others, 2018). Frequency analysis in this study was performed using the LP3, as recommended in the guidelines for Bulletin 17C.

Flood Frequency Based on Log-Pearson Type III Distribution

Based on guidance provided by Bulletin 17C, maximum N-day flood durations were fitted to the LP3 distribution using the method-of-moments estimators of the mean, standard deviation, and skew coefficient of the logarithms of the flood durations. Using these three parameters, various flood quantiles can be calculated:

$$\log Q_T = \overline{X} + K_T S, \tag{3}$$

where

Q_T is the flood quantile, in cubic feet per second, with a recurrence interval *T* in years;

$\overline{X}$ is the estimated mean of the logarithms of the annual N-day flood durations;

KT is a frequency factor (based on the skew coefficient and recurrence interval, T) that can be estimated from available algorithms (England and others, 2018); and

S is the estimated standard deviation of the logarithms of the annual N-day flood durations.

The skew coefficient was calculated using the station data (no regional weights were applied).

Expected Moments Algorithm

This study followed the Bulletin 17C recommendation of using the expected moments algorithm (EMA) for fitting the LP3 distribution (Cohn and others, 1997; Cohn and others, 2001; England and others, 2003; Griffis and others, 2004; Parrett and others, 2011). The EMA is a generalized method of moments procedure for fitting the LP3 curve. For flood records that include historical, paleoflood, or censored flows, the EMA method allows for a more robust and efficient method of calculating LP3 moment estimators—when streamflow records contain gaps or low outliers and censored flows (Cohn and others, 1997) —compared to the methods described in the previous standard guidance for flood frequency analysis, Guidelines for Determining Flood Flow Frequency (Bulletin 17B; Interagency Advisory Committee on Water Data, 1982). Another advantage of using the EMA method to fit the LP3 curve is its ability to estimate the mean-square error of the at-site skew coefficient. The mean-square errors of the at-site and regional skew coefficients are used when weighting both types of skew coefficients as recommended by Bulletin 17C.

Potentially Influential Small Floods

Bulletin 17C recommends the use of the Multiple Grubbs-Beck Test (MGBT; Grubbs and Beck, 1972; Cohn and others, 2013) for the detection of smaller events, that is, potentially influential low floods (PILF). A PILF is a flood of relatively low magnitude that can have a relatively large influence when fitting the flood frequency distribution and, in turn, can adversely affect the estimation of greater magnitude floods (low AEP). Put another way, a PILF can have a relatively large degree of leverage by resulting in a more negative skew coefficient, thus affecting the overall fit of the frequency distribution more than other flood magnitudes. Flood magnitudes at the low end of the distribution are undesirable (the smaller the magnitude, the greater the AEP values) because they can underestimate the rare and large flood flows (low AEP).

The flood-frequency curves for the annual maximum 3-day streamflow for USGS streamgage 14190500, Luckiamute at Suver, Oregon, are an example of a streamflow record in which one low observation can affect the calculation of low AEP values. In this instance, the 0.01 AEP estimate is about 35,000 ft^3/s with the low outlier removed (fig. 8). If the low outlier were to be included in the frequency analysis, it would result in a lower skew coefficient (-0.082 instead of 0.043), and an 0.01 AEP value of about 33,800 ft^3/s (fig. 9).

Censoring primarily was accomplished by using the MGBT without modification; in other words, without manually entering a different threshold to change the censoring. However, in some instances, better fits to the upper end of the distribution were developed by additional censoring or by decreasing the amount of censoring. An account of censoring at each site, including all censoring done outside the default Bulletin 17C format, is detailed in appendix 2, tables 2.2 and 2.3.

Because large floods are the primary focus of this study, adequate censoring of smaller floods was necessary to allow for the correct fitting of larger floods. The number of censored floods varied according to the duration of floods considered (table 4). Generally, the longer the flood duration being analyzed, the more censoring occurred.

About one-half of the 403 study sites (46–57 percent, depending on the duration of interest) included no PILFs identified using MGBT. The remaining sites (43–54 percent,

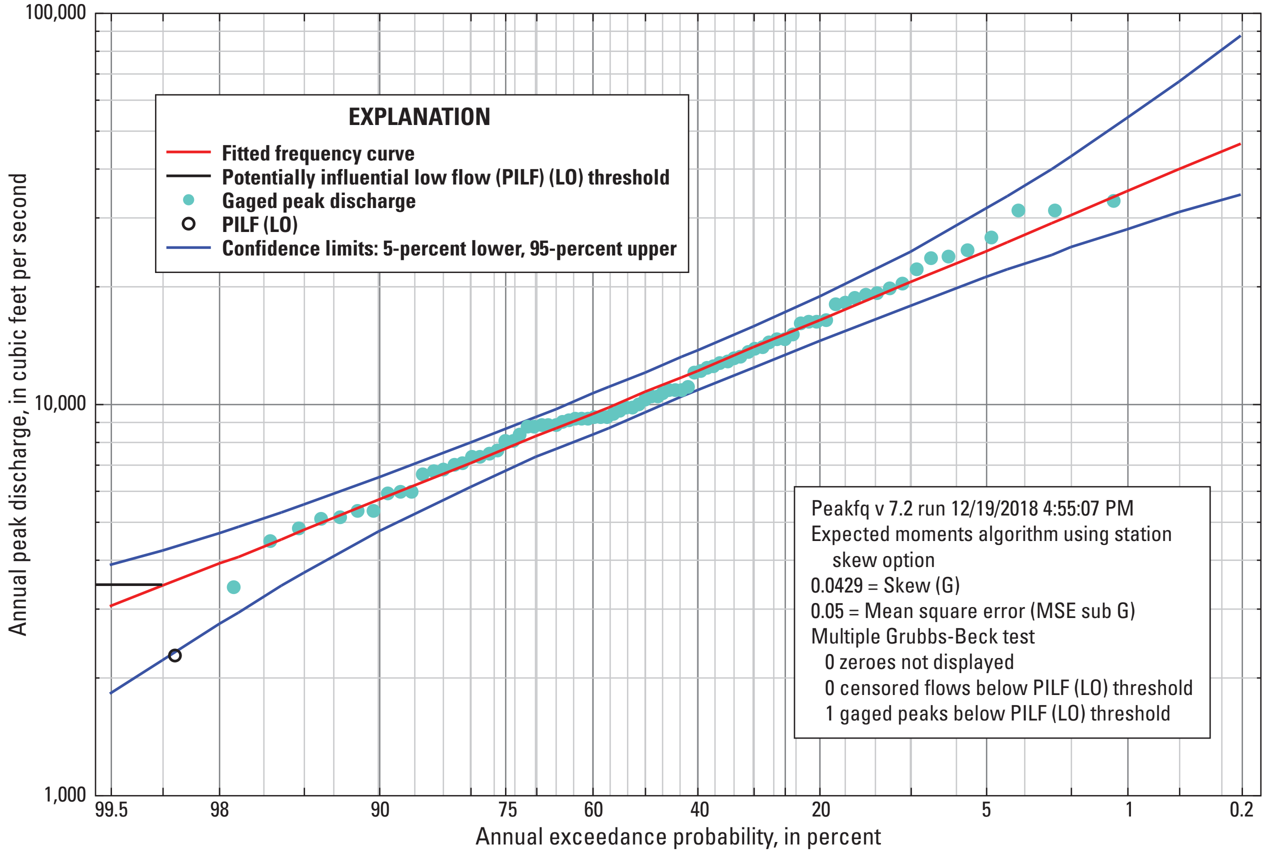


Figure 8. Flood-frequency curve for maximum 3-day duration flows with censored data for the Luckiamute at Suver, Oregon streamgage (U.S. Geological Survey streamgage 14190500).

depending on duration of interest) included in this study had one or more outliers less than the PILF criterion and were censored.

Two examples of censoring are shown in figures 10 and 11. No censoring occurred at USGS streamgage 14091500, the Metolius River near Grandview, Oregon, for the 1-day duration flood frequency analysis, as no low outliers were identified by the MGBT. Conversely, at USGS streamgage 13247500, Payette River near Horseshoe Bend, Idaho, 40 of the 107 years of record were less than the PILF threshold according to the MGBT. Censoring increases the mean square error (MSE) of the at-site skew-coefficient estimate. Additionally, because the weight given to the estimated skew coefficient at each site is weighted using the inverse of its MSE, extensive censoring such as that which was observed at station 13247500 may critically affect the weight placed on the station skew coefficient in a regional skew analysis.

In rare cases, additional censoring beyond what was prescribed by the MGBT was incorporated. This typically occurred with short streamflow records and (or) with records that appeared to contain some degree of multimodality. Multimodality can occur when there are mixed populations of flood distributions resulting from more than one cause of the flood events. An example of multimodality would be a basin that has flooding from rainfall and snowmelt. Any additional censoring that took place in this study is presented in appendix 2, table 2.2.

Censored floods were not always consistent across the various N-day durations. Consideration was given to adding consistency by removing an annual maximum streamflow that was identified as a PILF for any given N-day duration from all analyses (each of the other N-day durations). However, an assumption was made that the identification of a particular PILF for a 1-day flood-duration time series should not automatically be censored from time series with much longer durations (30 or 60 days), as the flood events often are caused by much different processes depending on the duration. The opposite also is true, as removing PILFs from a 60-day flood-duration time series should not influence censoring for short-duration time series. Additionally, censoring was relatively consistent through incremental changes in duration, which suggests that there would not be large changes in the sites used between their durations of similar length. Using these censoring methods, no crossovers of the frequency curves for the various durations occurred; in other words, longer-duration frequency curves did not exceed shorter-duration frequency curves for any of the sites used in this study.

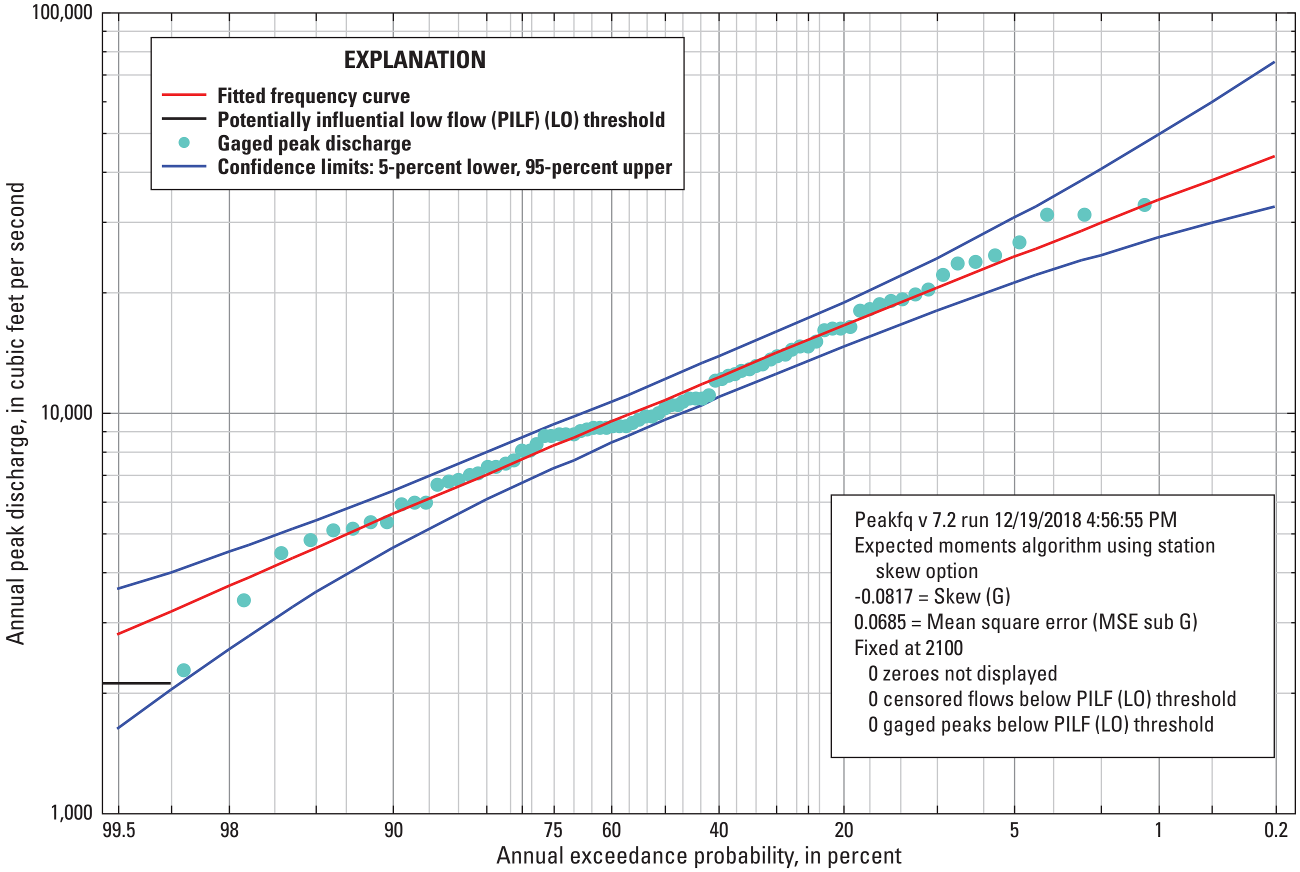


Figure 9. Flood-frequency curve for maximum 3-day duration flows with no censored data for the Luckiamute at Suver, Oregon streamgage (U.S. Geological Survey streamgage 14190500).

Table 4. Number and percentage of sites with censored floods for multiple ranges from 403 sites selected for analysis in this study, Columbia River Basin, northwestern United States and British Columbia, Canada.

[Flood duration is a running average of daily streamflow over a selected time period (N-1 is a 1-day time period, N-3 is a 3-day time period, etc.) **N-day duration:** Ranges start at 0 percent, meaning no floods were censored; 0–10% (percent) means that from 0 to 10 percent of floods were censored; and > (greater than) 50 percent means that at least one-half of the floods were censored]

N-day duration	N-1	N-3	N-7	N-10	N-15	N-30	N-60
	Number of sites and percentage of sites (in parentheses) for ranges of censored floods						
0%	234 (57)	235 (57)	208 (51)	199 (49)	199 (49)	189 (46)	190 (46)
0–10%	96 (23)	98 (24)	108 (26)	108 (26)	96 (24)	111 (27)	96 (23)
10–25%	36 (9)	34 (8)	41 (10)	44 (11)	48 (12)	52 (13)	48 (12)
25–50%	42 (10)	41 (10)	51 (12)	55 (13)	64 (16)	56 (14)	73 (18)
>50%	1 (0.2)	1 (0.2)	1 (0.2)	2 (0.5)	1 (0.2)	1 (0.2)	2 (0.5)

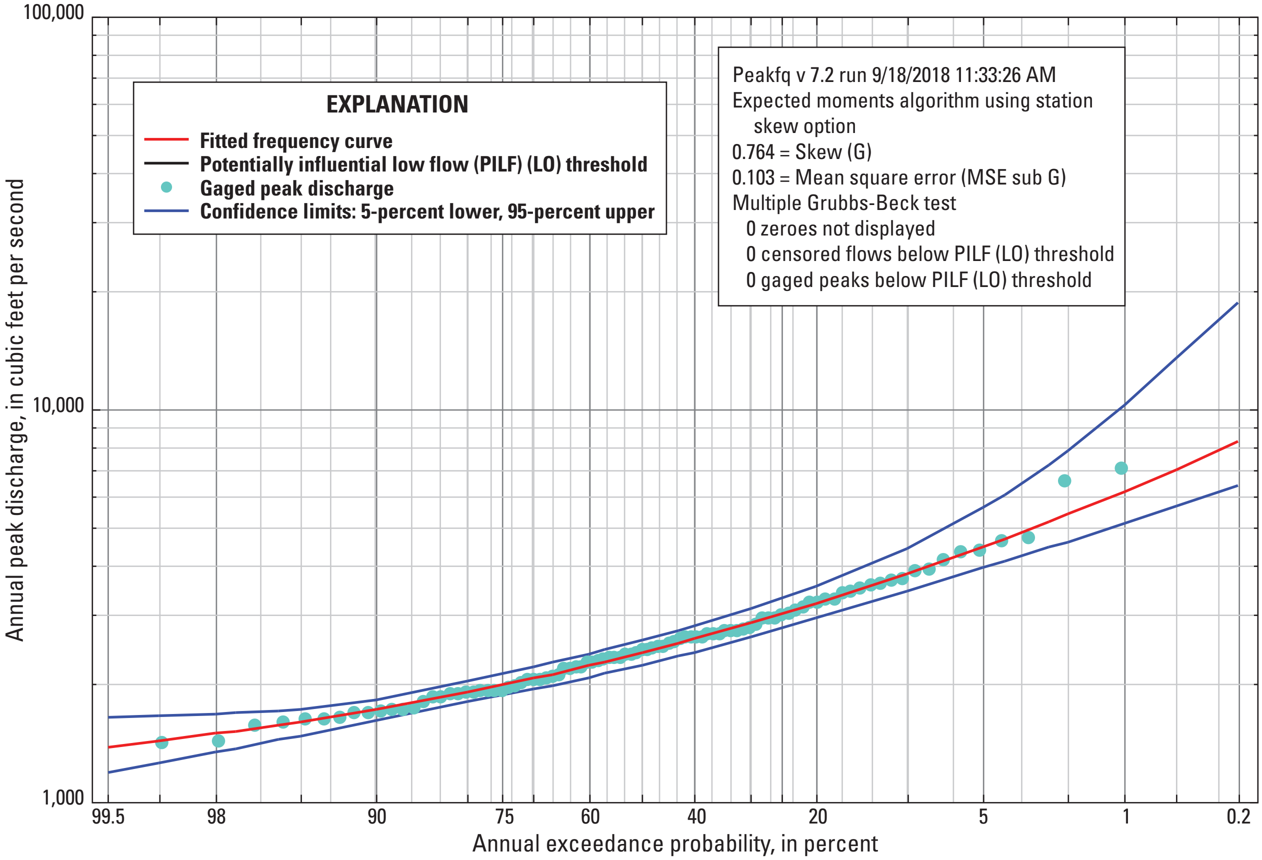


Figure 10. Flood-frequency curve for maximum 1-day duration flows for the Metolius River near Grandview, Oregon, streamgage (U.S. Geological Survey streamgage 14091500).

Correlations Between Skew Coefficients and Basin Characteristics

Once the at-site skew coefficients were computed using Bulletin 17C methodology, basin characteristics could be investigated for use in regional skew regression models. The magnitude of correlations between calculated N-day skew coefficients and select basin characteristics was evaluated. For example, the relation between drainage area and the at-site skew coefficient for the 1-day duration flow for 403 sites in the Columbia River Basin is shown in figure 12. Although the stations with the highest drainage areas tend to have smaller absolute values of skew coefficients than stations with smaller drainage areas, no strong relation is apparent between the two variables. The coefficient of determination (R^2) for this relation is 0.0174 (table 5). The low magnitude of correlation between calculated N-day skew coefficients and drainage area also is evident for other flood durations (fig. 13).

Other basin characteristics also were evaluated for correlation with skew coefficients using the same techniques previously described. Most basin characteristics, such as mean basin elevation and the percentage of precipitation that falls as snow (fig. 14), showed little or no relation with skew. Skew generally increased with mean annual precipitation values (fig. 14). Consequently, mean annual precipitation was chosen for further analysis (see section, "Exploratory Data Analysis").

The skew coefficients also were evaluated for consistency over the durations of interest. There was a general tendency for the median of skew coefficients to decrease as duration increased (fig. 15). The median skew coefficient ranged from -0.136 (1-day duration) to -0.336 (60-day duration). The skew coefficients with the greatest variability tended to occur at shorter-duration floods. These results were expected, as shorter-duration floods are more likely to deviate further from the average because of the varying landscape and storm event conditions, whereas longer-duration floods tend to dampen the spatial and temporal effects and show less variance (Lamontagne and others, 2012).

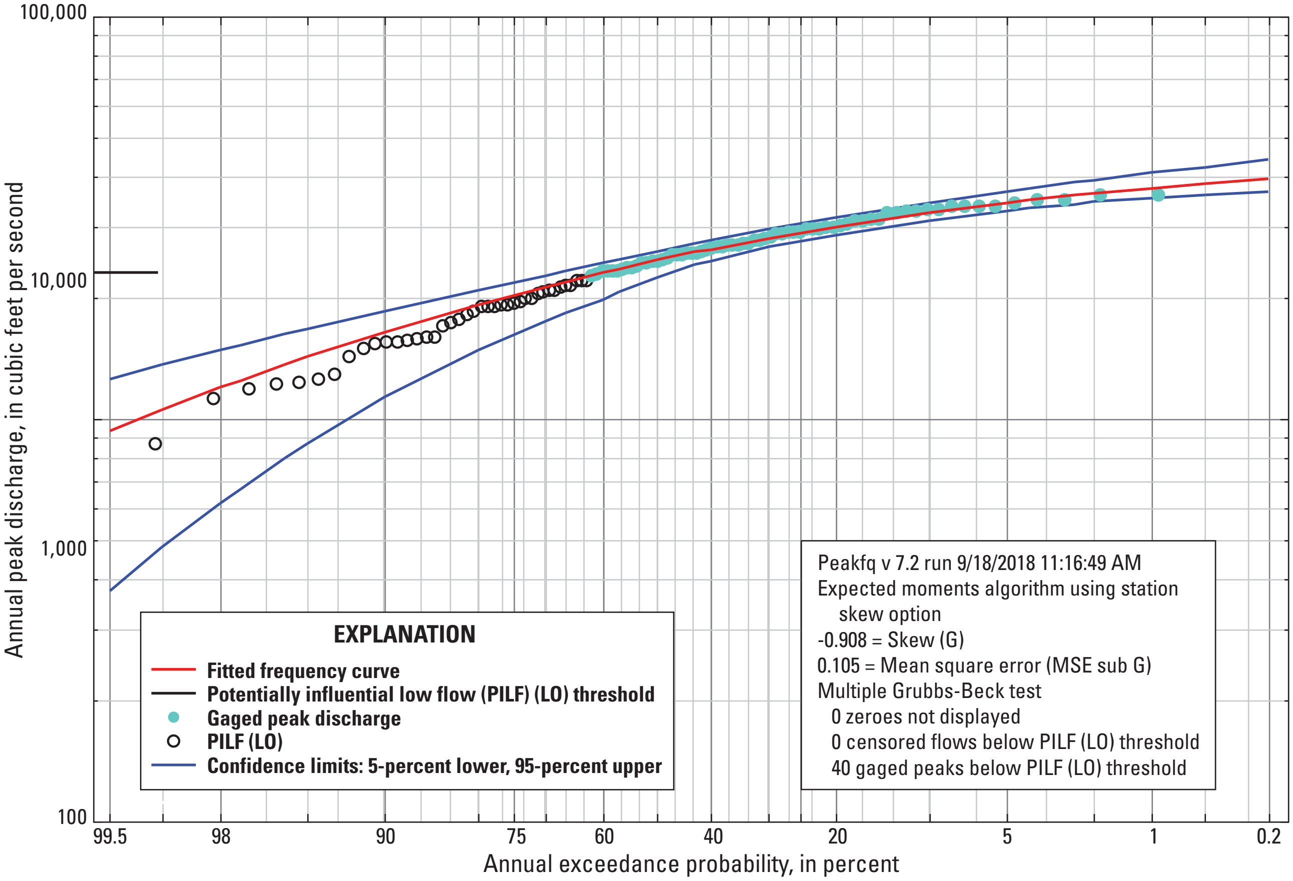


Figure 11. Flood-frequency curve for maximum 1-day duration flows for the Payette River near Horseshoe Bend, Idaho, streamgage (U.S. Geological Survey streamgage 13247500).

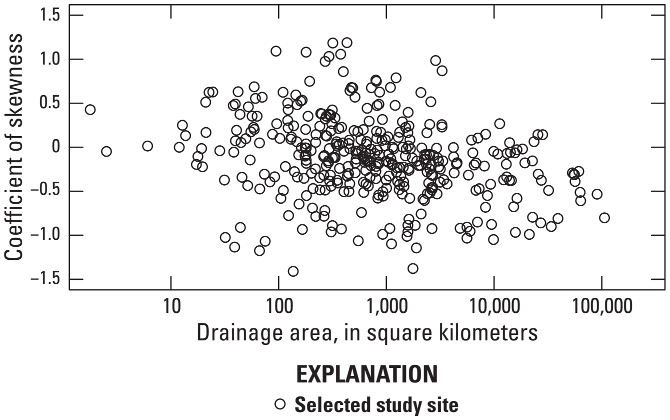


Figure 12. Skew coefficients and drainage areas for the 1-day flood durations in the Columbia River Basin, northwestern United States and British Columbia, Canada.

Table 5. Coefficients of determination for the relation between drainage area and skew coefficients for selected flood durations in the Columbia River Basin, northwestern United States and British Columbia, Canada.

[**Abbreviation:** R^2, coefficient of determination]

Duration (days)	R_2
1	0.0174
3	0.0358
7	0.0308
10	0.0195
15	0.0154
30	0.0147
60	0.0323

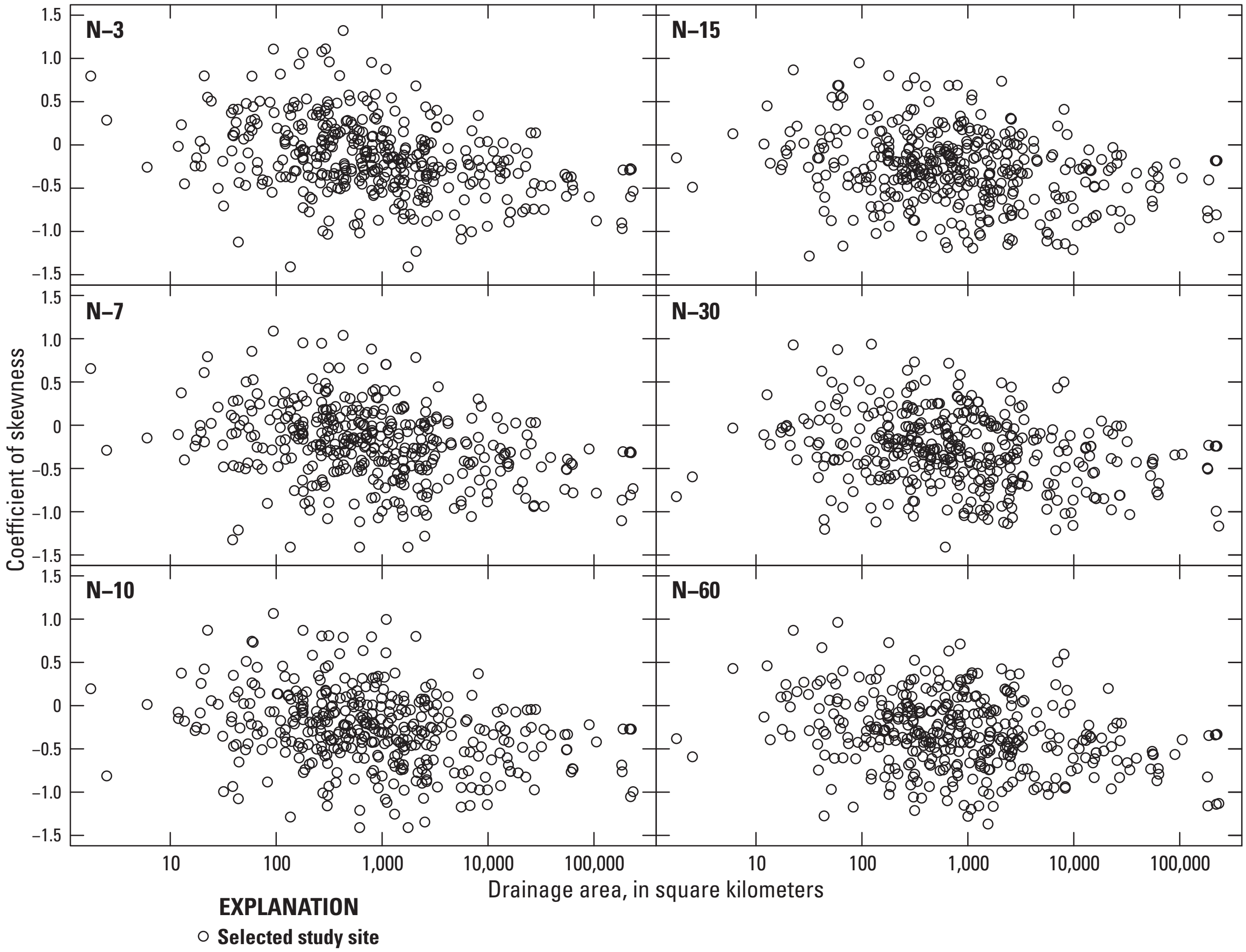


Figure 13. Skew coefficients and drainage areas for the selected flood durations in the Columbia River Basin, northwestern United States and British Columbia, Canada. N-3 is the 3-day flood duration, N-7 is the 7-day flood duration, etc.

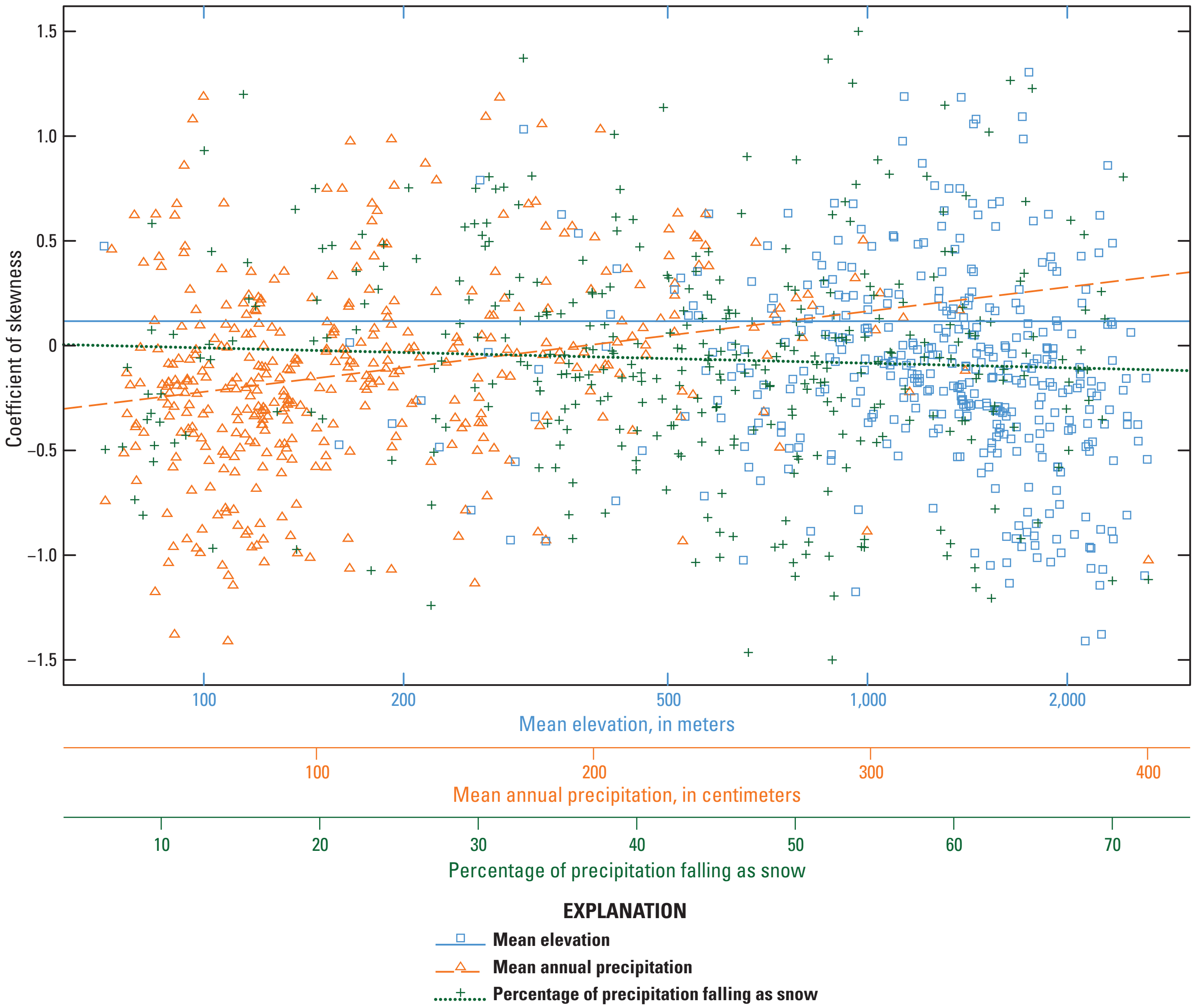


Figure 14. Regressions for skew coefficients and selected basin characteristics of the Columbia River Basin, northwestern United States and British Columbia, Canada.

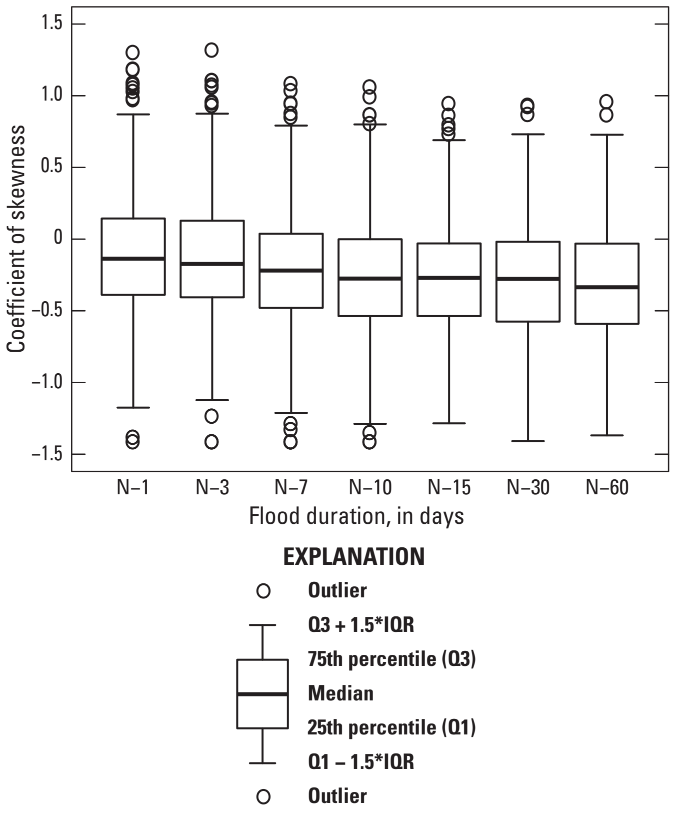


Figure 15. Skew coefficient distributions for select flood durations in the Columbia River Basin, northwestern United States and British Columbia, Canada. Interquartile range (IQR) is the range from the 25th to the 75th percentiles. The solid vertical lines and whiskers extend 1.5 times the IQR from the top or bottom of the box.

Regional Duration—Skew Analysis

The skew coefficient is a measure of asymmetry in a dataset and plays an important role in determining extreme flood quantiles. The skew coefficient is difficult to estimate directly from flood records of limited duration (Griffis and Stedinger, 2009), so Bulletins 17B and 17C recommend the use of a generalized (that is, regional) skew coefficient. Bulletin 17B included several recommendations for estimating generalized skew coefficients in addition to a national map that reported regional skew coefficients as a function of basin location. One of the recommendations included using regional skew coefficient prediction equations based on basin characteristics. To this end, Tasker and Stedinger (1986) proposed an operational method-of-moments weighted least squares (WLS) model to predict the skew coefficients as a linear function of basin characteristics. The WLS analysis is a marked improvement over ordinary least squares (OLS), as it recognizes that skew coefficients at individual basins (“sample” skews) are estimated with error and places more weight on sites with less sampling error (that is, longer record lengths). Additionally, Tasker and Stedinger’s approach decomposed errors into two sources: sampling error due to limited data and model error due to the use of an imperfect model.

Stedinger and Tasker (1985) and Tasker and Stedinger (1989) proposed operational generalized least squares (GLS) regression models for flood quantile estimation. GLS regression is an improvement over WLS in that it explicitly accounts for cross-correlation between flood quantile estimates. Such cross-correlation arises because nearby and similar sites have similar climatological forcings each year; for example, the peak flood in adjacent basins in a given year may be due to the same rainstorm. This correlation in annual maxima leads to a correlation between the log-space skew coefficients (Martins and Stedinger, 2002). Failure to account for this cross-correlation can affect fitted models and substantially misrepresent model precision.

Reis and others (2005) introduced a Bayesian generalized least squares (B-GLS) procedure for fitting linear skew models. The primary advantage of the Bayesian approach over the earlier method-of-moments procedure described in Stedinger and Tasker (1985) is that it provides the full posterior distribution of the model error variance. If sampling errors are sufficiently large, the approach in Stedinger and Tasker (1985) can estimate a model error variance of zero, implying a perfect model that clearly is unreasonable. The Bayesian approach incorporates new information about the model parameters so that they do not have to be estimated. Since 2009, B-GLS and its derivatives have served as the primary tool used by the USGS in regional skew studies across the United States (Feaster and others, 2009; Gotvald and others, 2009; Weaver and others, 2009; Parrett and others, 2011; Lamontagne and others, 2012). A summary of the Bayesian approach used for a regional regression skew analysis for annual peak floods in the Pacific Northwest is available in Mastin and others (2016, appendix A).

Standard Bayesian Generalized Least Squares

The B-GLS model proposed by Reis and others (2005) assumes that log-space skew coefficients can be predicted as a linear function of K available basin and (or) climatic characteristics (possible explanatory variables) with additive error (sampling and model error). The model is fitted to a dataset including N non-redundant sites. Non-redundant sites were screened to ensure that their basins were not nested or highly correlated spatially. In matrix notation, the standard model is:

$$\hat{\gamma} = X\beta + \varepsilon, \tag{4}$$

where

$\hat{\gamma}$ is an (N×1) vector of unbiased at-site skew coefficient for each site;

X is an [N×K] matrix of basin characteristics for each site;

β is a (K×1) vector of GLS regression coefficients; and

ε is an (N×1) vector of total errors representing the sum of the regional regression model error, δ, and the sampling error in the at-site sample skew coefficient estimate for each site.

The ith element of $\hat{\gamma}$ is the unbiased at-site skew coefficient for site i calculated using the Tasker and Stedinger (1986) bias correction factor:

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

where

G_i is the traditional biased at-site skew coefficient,

$P_{RL,i}$ is the pseudo-record length for site i as calculated using equation 11, and

$\hat{\gamma}_i$ is the unbiased at-site skew coefficient for site i.

For the GLS model, $E[\varepsilon]$=0, and the covariance matrix of ε is Λ=$E[\varepsilon\varepsilon^T]$. The error covariance matrix Λ is given by the equation:

$$\Lambda = \sigma_\delta^2 I + \Sigma(\hat{\gamma}), \tag{6}$$

where

σ_δ^2 is the model error variance;

I is an [N×N] identity matrix; and

$\Sigma(\hat{\gamma})$ is the [N×N] covariance matrix of sampling errors, composed of the at-site skew coefficient sampling variance on the diagonal (eq. 9), and sampling co-variance on the off-diagonal (computed with eqs. 7 and 9).

WLS analysis arises as a special case of GLS, where cross-correlation between the at-site skewness estimators is neglected so that the off-diagonal elements of $\Sigma(\hat{\gamma})$ are zero. Through extensive Monte Carlo simulation, Martins and Stedinger (2002) developed the following relation between the cross-correlation of annual maximums at two sites, i and j, and the cross-correlation between at-site skew coefficient estimators. That relation is used to estimate the cross-correlation between at-site skew estimators for the various durations used in this study:

$$\hat{\rho}(\hat{\gamma}_i, \hat{\gamma}_j) = sin(\rho_{ij}) cf_{ij} |\rho_{ij}|^{\kappa}, \tag{7}$$

where

$\hat{\rho}(\hat{\gamma}_i, \hat{\gamma}_j)$ is the estimated cross-correlation between the at-site skew coefficient at sites i and j;

ρ_{ij} is the cross-correlation between concurrent annual maximum at sites i and j;

κ is a constant between 2.8 and 3.3; and

cf_{ij} accounts for the difference in record length at the two sites relative to the concurrent record between at the two sites, and is defined as follows:

$$cf_{ij} = \frac{N_{ij}}{\sqrt{(N_{ij} + N_i)(N_{ij} + N_j)}}, \tag{8}$$

where

N_i and N_j are the non-concurrent observations corresponding to sites i and j, and

N_{ij} is the number of years of concurrent systematic record between sites i and j.

Here, ρ_{ij} is drawn from the cross-correlation models for each duration detailed in section, "Cross-Correlation Model of Concurrent Flood Durations."

The variance of the unbiased at-site skew coefficient estimators (the diagonal elements of $\Sigma(\hat{\gamma})$) is calculated using a modification of Tasker and Stedinger (1986) correction factor:

$$Var[\hat{\gamma}_i] = \left[1 + \tfrac{6}{P_{RL,i}}\right]^2 Var[G_i], \tag{9}$$

where

$Var[G_i]$ is the variance of the biased at-site skew coefficient computed using Griffis and Stedinger (2009):

$$Var[G_i] = \left[\tfrac{6}{P_{RL,i}} + a(P_{RL,i})\right] \times \left[1 + \left(\tfrac{9}{6} + b(P_{RL,i})\right) G_i^2 + \left(\tfrac{15}{48} + c(P_{RL,i})\right)\right], \tag{10}$$

where

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

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

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

The EMA adopted in Bulletin 17C and described in section, ”Expected Moments Algorithm,” explicitly considers censored data. In this study, 43–54 percent of records (depending on duration) contained PILFs that were censored. Censored observations provide valuable information in frequency analysis, but they provide less information than systematic peaks (that is, non-censored observed flows) (Cohn and others, 1997). As such, an adjustment must be applied to the record length when computing the skew coefficient variance (eq. 7). The pseudo-record length equation used by Mastin and others (2016) was adopted:

$$P_{RL,i} = P_{s,i} \times \tfrac{MSE(G_{s,i})}{MSE(G_{c,i})}, \tag{11}$$

where

$P_{s,i}$ is the number of non-censored systematic peaks in the record for site i;

$MSE(G_{s,i})$ is the estimated mean square error (MSE) of the at-site skew coefficient for site i when only the non-censored systematic peaks are analyzed; and

$MSE(G_{c,i})$ is the estimated MSE of the at-site skew coefficient for site i when all data, including censored observations, are analyzed.

Here, $P_{RL,i}$ is defined in terms of the number of years of systematic non-censored observations needed to yield the MSE computed using all data (including censored peaks), $MSE(G_{c,i})$. The pseudo-record length, $P_{RL,i}$, is constrained to be less than or equal to N_i, the number of observations for site i, as censored observations contain less information than non-censored systematic ones for continuous records of equal length (Mastin and others, 2016). $P_{RL,i}$ also is constrained to be greater than or equal to $P_{s,i}$, as censored observations in a continuous record contain more information than no observations would in an intermittent record over the same period.

Once $\Sigma(\hat{\gamma})$ is estimated using equations 7–11, the B-GLS methodology described in Veilleux and others (2011) and Veilleux (2011) is used to compute the model parameters and diagnostic statistics.

Hybrid Bayesian Weighted Least Squares/ Generalized Least Squares Procedure

The B-GLS procedure developed by Reis and others (2005) generally is not used for regional skew regression in the United States, as Lamontagne and others (2012) reported that high cross-correlation between annual maximum flows caused problems with B-GLS parameter estimation. B-GLS does not recognize that there is error in the estimation of the cross-correlation (off-diagonal elements of $\Sigma(\hat{\gamma})$) and seeks to exploit those cross-correlations through unjustifiably complex regression weights. To avoid this issue, Lamontagne and others (2012) and Veilleux and others (2011) developed a hybrid Bayesian weighted least squares/generalized least squares (B-WLS/B-GLS) procedure that ignores cross-correlations when computing the regression parameters, then evaluates the precision of those parameters by considering the cross-correlation of the sampling error. This study uses a slightly refined B-WLS/GLS approach, as described by Veilleux and others (2012), which has been implemented in USGS regional skew analyses since that report was published. An overview and explanation of steps used for these analyses is described in the following sections.

Step 1—Ordinary Least Squares Analysis

The high variability of skew coefficients, G_i, is a concern in using them directly to estimate their sampling variance in equation 5, which is the justification for using a regional skew coefficient in the first place. To limit the influence of sampling error on the estimate of sampling error variance, the hybrid B-WLS/B-GLS approach starts by performing an ordinary least squares (OLS) regression for the unbiased regional skew coefficient. OLS regression differs from GLS regression because cross-correlation is ignored and each site is given the same regression weight (regardless of $P_{RL,i}$). OLS regression returns stable and unbiased estimates of regional skew coefficients for all sites:

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

where

$\tilde{y}_{OLS}$ is an $(N{\times}1)$ vector of regional skew estimates from an OLS analysis,

X is an $[N{\times}K]$ matrix of basin characteristics for each site, and

$\hat{\beta}_{OLS}$ is a $(K{\times}1)$ vector of OLS model coefficients.

The OLS regional skew estimates contained in $\tilde{y}_{OLS}$ are used to estimate the at-site skew variances by replacing the biased at-site skew coefficients, G_i in equation 5. The at-site skew coefficients are difficult to estimate from limited record lengths, which could make the Bayesian weighted least squares (B-WLS) and B-GLS regression weights in the subsequent steps unstable. A benefit the employment of $\tilde{y}_{OLS}$ provides is it ensures that the subsequent regression weights are relatively independent of at-site estimates of the skew coefficients by using the OLS regional skew variances instead of the at-site skew variances.

Step 2—Bayesian Weighted Least Squares

In step 2, a B-WLS analysis is conducted to compute regression coefficients for each regional skew model, $\hat{\beta}_{WLS}$. The B-WLS model, a specialized version of the B-GLS model, also is given by equation 4, where $\Sigma(\hat{y})$ in equation 6 is a diagonal matrix with the variances of the at-site skew coefficients estimated in section, "Step 1—Ordinary Least Squares Analysis." Here, the cross-correlations are ignored to avoid the parameter estimation problems noted by Veilleux and others (2011) and Lamontagne and others (2012). The B-WLS analysis accounts for record lengths while ignoring cross-correlations that can be problematic in B-GLS analyses used to estimate precisions of the fitted model and regional regression coefficients from the B-WLS model.

Step 3—Bayesian Generalized Least Squares

The precision of the fitted model and $\hat{\beta}_{WLS}$ regression coefficients are evaluated using a B-GLS analysis that explicitly considers the covariance of the at-site skew coefficients (Veilleux and others, 2011). Diagnostic metrics include the standard error of $\hat{\beta}_{WLS}$, $SE\left(\hat{\beta}_{WLS}\right)$, the model error-variance $\sigma^2_{\delta,GLS}$, the pseudo coefficient of variation, pseudo-R^2_{δ}, and the average variance of prediction for a site not used to develop the regional model, AVP_{new}. Veilleux and others (2011) provided a derivation of these metrics. Importantly, the model derived using B-WLS is evaluated using the error covariance matrix (eq. 7) from a B-GLS analysis, which can be used to estimate the precisions of the B-WLS model.

Data Analysis of Duration-Skews for the Columbia River Basin

In this study, regional skew models were generated for seven flood durations (1-, 3-, 7-, 10-, 15-, 30-, and 60-day) for the Columbia River Basin (fig. 2).

Previous work generated a regional skew model with a similar spatial extent for the instantaneous annual maximum flows (Mastin and others, 2016), although that work did not consider the 97 NRNI sites addressed in this study. One-day duration skew coefficients are expected to be similar to instantaneous annual maxima but longer-duration skew coefficients might differ because the hydrologic mechanisms driving 30- or 60-day floods likely differ greatly from the instantaneous maxima (Lamontagne and others, 2012).

For this study, we computed annual maximum N-day flood duration at 410 sites in the Columbia River Basin as described in section, "Streamflow." After screening out 4 USGS sites for intermittent records and 3 NRNI sites that had nearly identical streamflow records and basin characteristics (see section, "Streamflow"), we considered a total of 403 sites in the analysis. The average record length for USGS sites is 50 years, whereas almost all the NRNI sites have 80 years of reconstructed record. The pseudo record length, $P_{RL,i}$, varies for each site by flood duration. Forty-nine physical and hydrometeorological basin characteristics were available for the USGS sites, of which 18 sites also were available for the NRNI sites. These basin characteristics are reported in table 2.

Redundancy Screening

Before the regional skew regression analysis could proceed, identification and removal of redundant sites was necessary because they do not represent independent samples. In this context, redundancy occurs when one basin is nested within another and the two basins have similar drainage areas. For example, the Columbia River at McNary Dam subbasin is nested within the Columbia River at Bonneville Dam subbasin, and they have similar drainage areas (214,000 and 240,00 mi^2, respectively). This is an issue in regional skew regression because the sites do not provide two spatially independent observations of how drainage characteristics relate to skew coefficients because the two basins overlap. This can cause an incorrect analysis of the data (Gruber and Stedinger, 2008). If redundant sites are retained in the analysis, then model errors could show correlation, which could undermine the GLS diagnostic statistics that are important for subsequent frequency analyses. In particular, failing to remove redundant basins can cause model error, δ, to be correlated, violating the error partition assumptions underpinning the Stedinger and Tasker (1985) regional skew regression approach (that is, eq. 5) that all subsequent USGS regional skew analyses have used.

Previous USGS studies used screening metrics to identify potentially redundant basins. For two basins to be redundant (1) one basin is nested within another, and (2) their drainage areas are similar. Past USGS GLS analyses have identified two redundancy screening metrics (Veilleux, 2009; Lamontagne, 2014):

$$SD_{i,j} = \frac{Dist_{i,j}}{\sqrt{0.5 \times A_i \times A_j}} \tag{13}$$

$$DAR_{i,j} = \exp\left(\left|\ln\left(\frac{A_i}{A_j}\right)\right|\right), \tag{14}$$

where

$SD_{i,j}$ the standardized distance between sites *i* and *j*;

$Dist_{i,j}$ is the distance between the centroids of sites *i* and *j*

A_i and A_j are the drainage areas for sites *i* and *j*, respectively; and

$DAR_{i,j}$ is the drainage area ratio for sites *i* and *j*.

Those studies have set thresholds of $SD_{i,j} \leq 0.5$ and $DAR_{i,j} \leq 5.0$. If both computed metric values are less than these thresholds, then sites *i* and *j* potentially are redundant and a closer examination is warranted to determine whether the sites are nested and of similar size. If the sites are nested and of similar size, they are declared redundant and at least one should be removed to avoid errors in the subsequent regression analyses.

When these screening criteria were applied to the Columbia River Basin, 293 potential redundant pairs were identified. Most of these pairings include at least one of the USACE NRNI sites, which are mostly sites along the main stem of the Columbia, Snake, and Willamette Rivers or their major tributaries. Upon closer examination, we determined that the application of the standard screening thresholds to the Columbia River Basin resulted in many false negatives, where actual redundancies existed but were not identified by the standard metric values. As the threshold for standardized distance is increased, the number of potentially redundant pairings also increases, as reported in table 6.

Based on a manual analysis of obviously redundant sites in the Columbia River Basin, the $SD_{i,j}$ threshold of 0.7 was selected, as it substantially decreased the number of false negatives while not substantially increasing the number of false positives as the $DAR_{i,j}$ threshold was held at 5. With this more stringent criterion, 442 potential redundant pairings were identified. These 442 potentially redundant pairs then were manually evaluated, and 39 pairs were identified as not redundant. This amounts to a false positive rate of 39÷442=9 percent.

Once redundant pairs have been identified, the analyst must next remove one of the redundant sites. In past studies, for a single paring, the site with the longer record (that is, the smaller variance and higher regression weight) would be selected. If two basins were nested within a larger basin but were not redundant with each other, then the two smaller basins generally would be retained. The complicated streamgage network in this study was more challenging, as the drainage areas of the various sites ranged over several orders of magnitude. Thus, NRNI sites with large drainage areas had as many as 10 redundant sites, each of which had its own nested and redundant sites that were not necessarily redundant with the original large NRNI site. These tertiary sites also had their own redundancy issues. This was further complicated by the fact that the smaller basins generally had shorter records, and thus larger variances and smaller regression weights.

Table 6. Relation between standardized distance threshold and the number of identified redundant pairs for the Columbia River Basin, northwestern United States and British Columbia, Canada.

Standardized distance threshold	Redundant pairings
0.5	293
0.6	375
0.7	442
0.8	512
0.9	603

The choice to remove or retain one site could have effects that propagate across dozens of other redundant pairings, presenting a difficult combinatorial problem. To tackle this problem, the station selection process was posed as a mixed integer linear programming (MILP) problem, and the MATLAB (The MathWorks, Inc., 2016) built-in MILP solver was used to determine the "optimal" redundancy screening.

Our objective function is:

$$\min - w^T x,$$

such that

$$Ax \leq 1$$

$$0 \leq x \leq 1$$

$$x \in \mathrm{Z},$$

where

x is a [403×1] vector of indicator decision variables that can take either a value of 1 if that site is retained or 0 if the site is not retained,

w is a [403×1] vector where the *i*th element is equal to the inverse of the *i*th site's variance (approximately proportional to the inverse record length),

T is the transpose operator, and

A is a [442×403] constraint matrix where the *j*th row corresponds to the *j*th redundant pairing.

If sites 1 and 2 are redundant, then the corresponding constraint would be:

$$x_1 + x_2 \leq 1,$$

indicating that at most only one of the two sites can be included in the final dataset. The *A* matrix simply states this in terms of linear algebra for all 442 site pairings.

The MILP could be conducted independently for every duration under consideration, as the *w* vector also is a function of the skew coefficient, which will change across the durations. However, we determined that similar or identical answers were obtained for different durations, probably because the variance of the skew coefficient is so closely linked to the record length, which is not changing across duration. It was deemed more advantageous to have the same dataset for all durations, so the redundancy screening results for the median (10-day) duration were selected as representative for all durations. The results of the optimization-driven screening are summarized in table 7.

Initially, a no-preference criteria for either NRNI sites or unregulated USGS sites was used in the optimization. This no-preference criteria resulted in the retention of 274÷403 sites, or about 68 percent of study sites, including 35÷95 (or about 37 percent) of the NRNI sites. The mean drainage area decreased for the smaller screened sample compared to the unscreened sample. This decrease in mean drainage area is because many NRNI sites, particularly in the lower Columbia Basin, have drainage areas several orders of magnitude greater than the rest of the sites in the study and were removed. However, our subsequent data analysis indicated that drainage area is not a significant explanatory variable for skew coefficients. The actual objective value, the sum of the inverse of the at-site skew variances, is not an intuitive metric, so cumulative retained period of record (a close analogue) is reported instead. Optimization theory dictates that imposing binding constraints on the optimization will degrade the quality of the solution. By imposing the redundancy constraints, 15,218 years of the total 22,987 cumulative years of record are retained.

Too few NRNI sites were retained in the initial screening. Their large size (on average) and successive location along the main stem of the Columbia River and its tributaries meant that they were likely to be redundant (that is, they contained very little unique information). To retain more NRNI sites, a two-step optimization scheme was devised: the NRNI favoring screening. In this case, an initial MILP was used that included only NRNI sites. In this initial step, no NRNI site would be removed from the analysis in favor of a USGS site; it would only be dropped if it was redundant with another NRNI site. The result of this initial MILP was converted to a set of equality constraints to force a larger, second MILP to include or exclude non-redundant or redundant NRNI sites. As expected, the result was the inclusion of additional NRNI sites (eight sites) in the analysis, bringing the total to 45 percent of the NRNI sites. By enforcing these additional constraints, the optimization quality degraded slightly, now including only 262 sites, and having a cumulative period of record of 14,449 of the possible 22,987 years. The reduced cumulative period of record from the initial screening (no NRNI favoring) was the result of fewer sites being retained overall. In consultation with the USACE, this result was preferred, as the degradation in performance was minimal, but the number of NRNI sites increased. Counterintuitively, the mean drainage area was reduced slightly under the NRNI favoring screening approach. This reduction in mean drainage area occurred because several smaller NRNI sites in the headwaters were preferred to larger unregulated USGS sites.

During the subsequent regression exploratory analysis, three additional sites were removed. Two USGS sites (USGS streamgage 12323250, Silver Bow Creek below Blacktail Creek at Butte, Montana; and USGS streamgage 14211500, Johnson Creek at Sycamore, Oregon) were removed because their impervious cover exceeded 5 percent, the threshold for inclusion in the instantaneous annual maximum study (Mastin and others, 2016). Another USGS site (USGS streamgage 14219800, Speelyai Creek near Cougar, Washington) was removed because its annual precipitation (in excess of 150 in. per year) and its at-site skew coefficients were atypical of the other sites in the study, indicating that it was not representative of the Columbia River Basin sites of interest.

Table 7. Results of the mixed integer linear programming redundancy screening.

[Number and percentage (in parentheses) of sites is given: (1) prior to screening (all sites), (2) after the initial screening with no preference, and (3) after the final NRNI-favoring screening with a preference for retaining more NRNI sites. **Abbreviations:** NRNI, no-regulation no-irrigation; POR, period of record; DA, drainage area; km^2, square kilometers]

	All sites	Initial screening	NRNI favoring screening
Total sites	403	274	262
	(100)	(68)	(65)
NRNI	95	35	43
	(100)	(37)	(45)
POR	22,987	15.218	14,449
	(100)	(37)	(45)
Mean DA (km^2)	15,692	5,129	5,000

Exploratory Data Analysis

Mastin and others (2016) performed a regional skew analysis for the instantaneous annual maximum flows in the Columbia River Basin (excluding the Canadian part). They reported that no basin characteristic was able to describe an appreciable amount of true variability in the skew coefficient, so they selected a parsimonious constant model for the entire basin. The increase in model complexity could not be justified by a sufficient increase in model precision. Furthermore, a constant model has been the preferred model in most studies over the last decade (Paretti and others, 2014; Southard and Veilleux, 2014; Kennedy and others, 2015; Curran and others, 2016; Wood and others, 2016). A similar result was expected for duration-floods, particularly for shorter durations. An additional concern in this study was consistency across durations (Lamontagne and others, 2012). If the skew models are substantially different for two durations, the resulting flood-duration frequency curves could cross (for example, the 1-percent AEP 7-day flood could exceed the 1-percent AEP 3-day flood).

Two primary concerns arise when assessing potential models: the quality of fit for a specific duration (considering $\sigma^2_{\delta,GLS}$ and AVP_{new}) and consistency across durations. A convenient statistic to evaluate the fraction of the true variability in the skew coefficient described by the model is the pseudo-R^2_δ (Gruber and others, 2007):

$$R^2_\delta \;=\; 1 - \frac{E\left[\sigma^2_\delta(k)\right]}{E\left[\sigma^2_\delta(0)\right]}, \qquad (15)$$

where

$\sigma^2_\delta(k)$ is the model error variance obtained with a model using k explanatory variables,

$\sigma^2_\delta(0)$ is the model error variance for the constant model, and

E is the expected value or expectation operator.

When $R^2_\delta > 0$, some of the true variability in the skew coefficient is being explained by the model. Even if $R^2_\delta > 0$, it is important to determine whether the k model coefficients are statistically different than zero. Furthermore, it is no guarantee that a model with statistically significant parameters has much true explanatory power, so R^2_δ might be closer to zero. Slightly negative values of R^2_δ can occur when $E\left[\sigma^2_\delta(k)\right] > E\left[\sigma^2_\delta(0)\right]$, because both $E\left[\sigma^2_\delta(0)\right]$ and $E\left[\sigma^2_\delta(k)\right]$ are statistics computed from data with error.

Initial Screening with Only U.S. Geological Survey Sites

A challenge in this analysis was that the full suite of 49 basin characteristics were available for the USGS sites, while only 18 basin characteristics were available for the NRNI sites. To determine whether any of the additional 24 hydrometeorological basin characteristics should be computed for the NRNI sites, an extensive screening exercise was conducted exploring all basin characteristics using just the USGS sites. The MILP redundancy screening was repeated to include only USGS sites, resulting in a total of 271 sites included in the analysis. Every possible univariate model (including a constant) was fit to the 271 retained USGS sites.

When using the USGS sites to fit models, no basin characteristic described more than 7 percent of the true variability in the skew coefficient for any duration (that is, $R^2_\delta \le 7$ percent). Significant basin characteristics were the basin compactness (area divided by squared perimeter distance) (1-, 3-, 7-, 10-, 15-, and 60-day durations; $R^2_\delta \le 4$ percent in all cases) and drainage area (3-, 7-, and 10-day durations; $R^2_\delta \le 7$ percent in all cases). This result confirmed, at least at a basin-scale, that none of the 30 hydrometeorological variables were significant descriptors of skew coefficients. Furthermore, it seems that no basin characteristic is consistently able to explain more than 7 percent of true variability in the skews, so it is unclear that a non-constant model is justified.

There is substantial hydrologic variability across the Columbia River Basin (see section, "Study Area Description"). The Columbia River Basin contains semi-arid basins (in the Snake River subbasin), as well as wet basins (in the Willamette River subbasin) (see fig. 3). The most distinct partition point in the basin is the Cascade Range because basins east of the Cascades tend to receive significantly less annual precipitation than those to the west. A second exploratory analysis was conducted where univariate models (including a constant intercept) were fit separately for each of the 49 explanatory variables for all durations to USGS sites to the east and west of the Cascade Range. East of the Cascade Range, the Snake River subbasin was deemed unique compared to the other subbasins of the mid- and upper-Columbia River Basin, so the eastern sites again were subdivided into two groups: the Snake River subbasins and non-Snake eastern subbasins (see fig. 1). Re-running of the MILP screening algorithm was unnecessary for each sub-region because there were few incidents of cross-regional redundancy; most main-stem sites on the Columbia River that would cause cross-region redundancies were NRNI sites rather than USGS sites.

West of the Cascade Range, some inconsistency was present between durations. For short-term durations (1-, 3-, and 7-, and 10-day), the average annual temperature was promising, with R^2_δ ranging from 12 to 35 percent. For mid-term durations (7- and 10-day), elevation-related terms (mean, minimum, and maximum) were significant at the 5-percent level with R^2_δ ranging from 20 to 25 percent. For longer term durations (15-, 30-, and 60-day), no statistically significant explanatory variables were able to explain any of the true variability in the skew coefficients (that is, $R^2_\delta \le 0$). This result was somewhat surprising, as most subbasins west of the Cascade Range are in the Willamette River subbasin, which has some hydrologic similarities to the Central Valley of California (Dettinger and others, 2011), where Lamontagne and others (2012) noted that elevation was a significant explanatory variable of skew coefficients. Importantly, this result confirms that none of the 24 hydrometeorological statistics were significant across all the durations west of the Cascade Range.

For the 1-day duration floods in non-Snake River eastern subbasins, October precipitation is statistically significant, although it explains less than 4 percent of the true variability in the skew coefficients (that is, $R^2_\delta \le 4$ percent). At the 3-day duration, October–-March precipitation is significant, although it has low R^2_δ, ranging from 5 to 7 percent. At all other durations, no explanatory variable is significant with an $R^2_\delta > 0$.

For Snake River eastern subbasins, basin compactness (area divided by squared perimeter distance, table 2) is significant at the 5-percent level for all durations but 1-day and 60-day, with R^2_δ ranging from 14 and 33 percent. However, AVP_{new} values were exceptionally high for the model because (1) there are not many sites located in the Snake River subbasin; and (2) they tend to be highly correlated, so the effective number of independent sites is low. The AVP_{new} statistic (1) describes the average variance of prediction for a site not included in developing the regional skew model and is used in Bulletin 17C to weigh against the station skew coefficient, and (2) is derived for the hybrid B-WLS/B-GLS analysis used here and in appendix 3. The AVP_{new} values for the best fitted models ranged from 32 to 104 percent larger for the Snake River eastern subbasins than for the best fitted models for the entire Columbia River Basin (fitted to the USGS sites). Thus, insufficient sites were deemed to be located within the Snake River subbasin to justify a Snake River subbasin-specific model; more precise and useful models could be derived by either considering a basin-wide model or grouping the Snake River sites with the rest of the Eastern USGS sites.

Screening with All Non-Redundant Basins

All the previously detailed analyses confirmed that the 24 monthly hydrometeorological variables that were available for the USGS sites but not for the NRNI sites were unlikely to be predictive of regional skew coefficients when applied to the full USGS/NRNI dataset. A screening analysis tested univariate models (with a constant intercept term) fit to all non-redundant sites (NRNI and USGS) for each of the 20 basin characteristics available at all sites. Initially, that screening analysis considered separately sites east (Snake and non-Snake subbasins) and west of the Cascade Range.

West of the Cascade Range, no basin characteristic was statistically significant at the 5-percent level for any duration, except the minimum basin elevation at the 10-day duration, which had a value of R^2_δ=17 percent. A goal of this analysis was to ensure consistency across durations. It was deemed unreasonable for the 10-day duration to have a regional skew model completely different from any other duration, especially if that model explained such a small fraction of the true variability in the skew coefficient. East of the Cascade Range, many basin characteristics—including longitude, drainage area, elevation-related metrics, basin compactness, precipitation, and air temperature—were significant at the 5-percent level for the 1-day duration, although most of these characteristics explained less than 3 percent of the true variability in the skew coefficients. The exceptions were basin compactness and drainage area, which explained 9 and 5 percent of the true variability, respectively. Basin compactness also was significant at all other durations (13 percent$\leq R^2_\delta \leq$9 percent), and drainage area was significant at all other durations except 60 days (4 percent$\leq R^2_\delta \leq$9 percent). These explanatory variables were not deemed to improve the precision of the models sufficiently to justify the added complexity. For example, at the 15-day duration, the AVP_{new} for the constant model is 0.13, whereas the AVP_{new} for the model with drainage area as an explanatory variable is 0.12, despite the fact that R^2_δ=9 percent. The model error-variance upon which AVP_{new} and R^2_δ are based, $\sigma^2_{\delta,GLS}$, is itself estimated with error. In the case of the 15-day duration drainage area model for the sites east of the Cascade Range, the standard error of $\sigma^2_{\delta,LS}$ is 0.019, so it is unclear that $\sigma^2_{\delta,GLS}$ for the drainage area model is truly smaller than that of the constant model.

The conclusion of this analysis is that no explanatory variable is significant across all the N-day durations, nor can it explain a sufficient amount of the true variation in the skew coefficients, either east or west of the Cascade Range, to justify the inclusion of explanatory variables. However, there was a difference between the average skew coefficients on the east and west sides of the Cascade Range. This would lead to a simple discontinuous model, wherein one constant is applied east of the Cascade Range and another constant is applied west of the Cascade Range. The hydrologic reasoning for having a discontinuous model with two constants is that eastern sites generally are more arid than the western sites. However, this type of discontinuous model likely is too broad a screening for the Columbia River Basin, as many high-elevation sites east of the Cascade Range receive significantly more precipitation than other eastern sites (fig. 3). Screening by elevation was not viable, as there are many high-elevation arid basins in the Snake River Plateau. So, four additional avenues of exploratory analysis were pursued.

Models were fit to individual subbasins (Willamette, Yakima, Upper Snake, Kootenai, etc.) as shown in figure 1. The major issue that arose during the fitting of these models is a lack of sites in individual subbasins. By narrowing the geographic focus, very few sites were located in each subbasin and those sites were highly correlated, making the effective number of independent observations too small to conduct credible regression analyses. A second screening analysis was performed wherein a Columbia River Basin wide model was fit where 4-digit hydrologic unit code (HUC4) regions (Seaber and others, 1987) were grouped, roughly mirroring the quantile regression regions by Mastin and others (2016), and each group was assigned its own indicator variable. This analysis again was compromised by narrowing the geographic scope, making it difficult to estimate important group constant models with the exception of the Willamette subbasin. Clustering analysis also was performed using k-means clustering to group sites with similar elevation, drainage area, centroid longitude, mean annual precipitation, and snow percent. However, we determined that the improvement in model precision did not sufficiently justify the increased complexity of the analysis, particularly for the analyst to determine group membership. The final exploratory analysis evaluated different non-linear functions of precipitation to model skew, as described in the next section.

Precipitation Models of Regional Skew Coefficients

Skew was determined to generally increase with mean annual precipitation (see section, “Correlations Between Skew Coefficients and Basin Characteristics”), which was selected for further analysis based on the strong correlation to skew. Average skew coefficients in the Columbia River Basin generally become less negative as mean annual precipitation increases, and that the variability of the skew coefficients tends to decrease with increasing mean annual precipitation. The first precipitation-based model tested was a discontinuous model:

$$\hat{y}_{i,d} = \hat{\beta}_{0,d} + \hat{\beta}_{1,d} I_i(c_1) , \qquad (16)$$

where

$\hat{y}_{i,d}$ is the regional skew coefficient for site i and duration d,

$\hat{\beta}_{0,d}$ and $\hat{\beta}_{1,d}$ are fitted model parameters for duration d, and

$I_i(c_1)$ is a binary indicator variable that takes a value of 0 if the precipitation for site i is less than c_1 inches and 1 if the precipitation is greater than c_1 inches.

Values of c_1 from 25 to 55 were tested in increments of 5. The model for c_1=40 in. performed well most consistently across durations. However, this discontinuous model version was disregarded because it could lead to inconsistent flood frequency analyses where two nearby sites on the same river could have very different regional skew coefficients if they happened to straddle the discontinuity point at around 40 in. of mean annual precipitation. To address this concern, two continuous non-linear functions were tested (see fig. 16). The first non-linear model has functional form:

$$NL_{1,i} = \max\left(0, 1 - \exp\left(\tfrac{-1\times(P_i - c_1)}{c_2}\right)\right), \qquad (17)$$

where

P_i is the mean annual precipitation at site i,

c_1 and c_2 are parameters of the non-linear model, and

NL_1 is fixed to zero below the threshold, c_1, and converges to 1 for large values of P_i.

Parameter c_2 controls the rate of transition from 0 to 1. The resulting regional skew model is:

$$\hat{y}_{i,d} = \hat{\beta}_{0,d} + \hat{\beta}_{1,d} NL_{1,i} \ . \qquad (18)$$

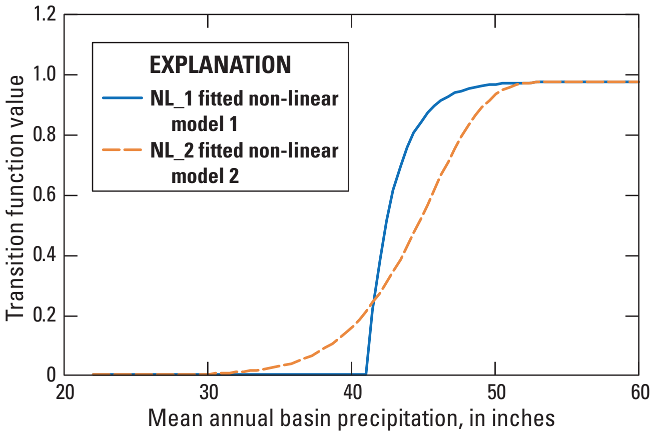


Figure 16. Continuous non-linear functions of mean annual precipitation for two non-linear models used in the study.

Models were fit for each flood duration for every combination of $c_1 \in \{35, 40, 45\}$ and $c_2 \in \{1, 2, 4, 8\}$. Additional tests also evaluated c_1 values of 50 and 55 in. We determined that c_1=40 and c_1=2 performed well most consistently across durations.

A second non-linear function of precipitation, inspired by the California duration skew analysis (Lamontagne and others, 2012), also was tested. Whereas the California non-linear model used basin elevation as a predictor, the model in this study uses precipitation:

$$NL_{2,i} = 1 - \exp\left(-\left(\tfrac{P_i}{c_3}\right)^{c_4}\right), \qquad (19)$$

where

c_3 and c_4 are model parameters;

$NL_{2,i}$ varies from 0 for low values of P_i to 1 for high values of P_i; and

c_3 controls the center of the transition from 0 to 1.

The resulting regional skew model is:

$$\hat{y}_{i,d} = \hat{\beta}_{0,d} + \hat{\beta}_{1,d} NL_{2,i}. \qquad (20)$$

The two nonlinear functions under consideration are plotted in figure 16. A range of c_3 and c_4 values were tested. The $NL_{1,i}$ model generally outperformed $NL_{2,i}$, based on the metrics used to evaluate previous models described above, so N $L_{2,i}$ was disregarded in favor of the $NL_{1,i}$ model with $c_1 = 40$ and $c_2 = 2$.

Final Duration-Skew Models

After the extensive exploratory analysis, the most consistently well-performing model across durations was determined to be the non-linear function of precipitation in equation 17, with $c_1 = 40$ and $c_2 = 2$:

$$\hat{y}_{i,d} = \hat{\beta}_{0,d} + \hat{\beta}_{1,d} \max\left(0, 1 - \exp\left(\tfrac{-1*(P_i - 40)}{2}\right)\right). \qquad (21)$$

For sites with less than 40 in. of mean annual basin precipitation, the final model in equation 21 returns a regional skew coefficient of $\hat{\beta}_{0,d}$ for duration d. For sites with mean annual basin precipitation more than about 51 in. per year, the final model in equation 21 returns a regional skew coefficient of about $\hat{\beta}_{0,d} + \hat{\beta}_{1,d}$ for duration d. For basins with mean annual precipitation from 40 to 51 in., the regional skew coefficient will be $\hat{\beta}_{0,d} \le \hat{y}_{i,d} \le \hat{\beta}_{0,d} + \hat{\beta}_{1,d}$.

The final fitted models and several diagnostic statistics used to evaluate model performance are reported in table 8. No R^2_δ is reported because the magnitude is less than 6 percent for all durations. The R^2_δ values are low because the fitted models are essentially constant models (with a step). The average sampling error variance (ASEV) describes the contribution of the sampling error in the model parameters to AV

P_{new}. For the 3-day model, the estimated AVP_{new} actually was slightly higher for the nonlinear precipitation model than for the constant model, as the increase in the ASEV from fitting an additional model parameter was not entirely offset by a decrease in the model error variance $\sigma^2_{\delta,B-GLS}$ (see appendix 3, eq. 3-1). However, $\sigma^2_{\delta,B-GLS}$ is estimated for both the constant and nonlinear precipitation models with some error, so very small increases in $\sigma^2_{\delta,B-GLS}$ or AVP_{new} are not necessarily points of concern if the models make hydrologic sense. The AVP_{new} for the fitted nonlinear precipitation models ranged from 0.07 to 0.11 depending on duration, compared to 0.18 in Mastin and others (2016) and 0.303 for the Bulletin 17B national skew map.

The results of the pseudo analysis of variance (ANOVA) for the final nonlinear precipitation model for the seven study durations are reported in table 9. The analysis divides the variability in the data into three sources: (1) variability described by the model, (2) variability in the true skew coefficient that the model cannot describe (model error), and (3) variability due to sampling errors. The model describes little variability in the true skew coefficient (the magnitude is less than 1 in all cases), as the final models at all durations are essentially constant models with a rapid step function. The negative values for model variability indicate a small increase in the model error variance for the nonlinear precipitation over the constant model. Both model error variances must be estimated and based on their standard errors; the difference is insignificant. The model error describes the precision with which the model parameters can be estimated. For all durations, the model error is less than one-half of the sampling error, indicating that the sampling error was the major source of variability.

The error variance ratio (EVR, appendix 3, eq. 3-8) is the ratio of the sampling error variance and the model error variance. This statistic is used to determine whether an OLS analysis is sufficient or if a more complicated WLS or GLS analysis should be used. A large EVR (exceeding 0.2) indicates that sampling errors are significant and that a WLS or GLS analysis should be used. The EVR in table 9 ranges from 2.3 to 3.6 depending on duration, indicating that a WLS or GLS analysis is necessary. The misrepresentation of beta variance (MBV, appendix 3, eq. 3-9) is the ratio of variance ascribed to the WLS regression constant by a GLS analysis and a WLS analysis. If MBV exceeds 1, this indicates that a GLS error analysis is needed. MBV ranges from 12.4 to 21.7 depending on duration, indicating that a GLS analysis of errors is needed.

The at-site sample skew coefficients (unbiased using eq. 5) against mean annual basin precipitation are shown in figures 17 and 18, as well as the fitted final regional skew models for the 1-, 3-, 7-, 10-, 15-, 30-, and 60-day duration floods. Significant scatter is present in the plotted points, largely due to sampling error. The sampling errors of the at-site skew coefficients are strongly correlated (see section, "Cross-Correlation Model of Concurrent Flood Durations"), and the sampling errors varied widely across sites due to differences in pseudo-record length (P_{RL}). Thus, not every point in figures 17 and 18 carries the same information in the analysis, and some points carry little information; for example, most of the non-Snake River subbasin sites east of the Cascade Range that have very negative skew coefficients around mean annual precipitation of 60 in. have short P_{RL} and small influence on the regression. The main bodies of points in figures 17 and 18 show an increase in at-site skew coefficients

Table 8. Final fitted models for the seven study durations and several diagnostic statistics used to evaluate model performance.

[Model parameters in **bold** are statistically significant at the 5-percent level. $\hat{\beta}_{0,d}$: Fitted model parameter. $\hat{\beta}_{1,d}$: Fitted model parameter. $\sigma^2_{\delta,GLS}$: Model error variance. **ASEV:** Average sampling error variance. AVP_{new}: Variance of prediction for a site not used to fit the model]

Duration	Model type	$\hat{\beta}_{0,d}$	$\hat{\beta}_{1,d}$	$\sigma^2_{\delta,GLS}$	ASEV	AVP_{new}
1 day	Constant	-0.10		0.08	0.01	0.09
	Nonlinear precipitation	**-0.28**	**0.38**	0.07	0.01	0.09
3 days	Constant	-0.13		0.05	0.01	0.06
	Nonlinear precipitation	**-0.32**	**0.39**	0.05	0.02	0.07
7 days	Constant	-0.24		0.07	0.01	0.09
	Nonlinear precipitation	**-0.40**	**0.33**	0.07	0.02	0.09
10 day	Constant	**-0.31**		0.07	0.02	0.09
	Nonlinear precipitation	**-0.47**	**0.34**	0.07	0.02	0.09
15 days	Constant	**-0.32**		0.07	0.02	0.09
	Nonlinear precipitation	**-0.48**	**0.32**	0.07	0.02	0.09
30 days	Constant	**-0.34**		0.06	0.02	0.08
	Nonlinear precipitation	**-0.48**	**0.28**	0.06	0.02	0.08
60 days	Constant	**-0.37**		0.09	0.02	0.11
	Nonlinear precipitation	**-0.47**	0.20	0.09	0.03	0.11

Table 9. Pseudo analysis of variance (ANOVA) for the final nonlinear regional skew model for all N-day flood durations for the Columbia River Basin, northwestern United States and British Columbia, Canada.

[**Source:** EVR, error variance ratio; MBV, misrepresentation of beta variance statistic. **Abbreviation:** N-day flood, flood whose duration is the running average of daily streamflow over a selected time period]

Source	Duration						
	1 day	3 days	7 days	10 days	15 days	30 days	60 days
Model	1	1	0	0	-1	-1	0
Model error	19	13	19	18	18	16	22
Sampling error	45	45	47	48	49	49	51
Total	65	59	66	66	66	64	73
EVR	2.4	3.6	2.5	2.6	2.7	3.1	2.3
MBV	12.4	16.0	15.6	16.7	17.6	19.7	21.7

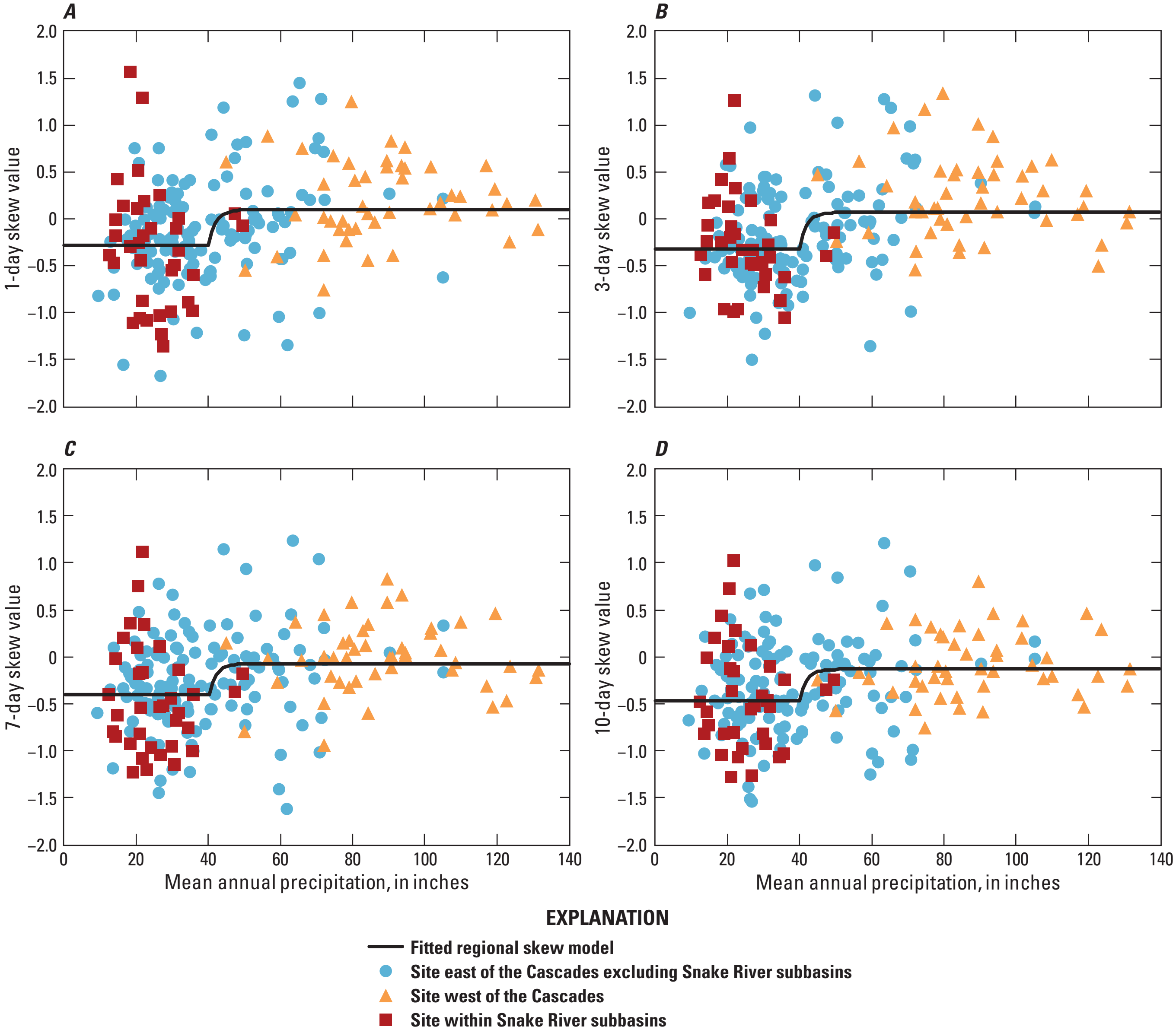


Figure 17. Fitted regional skew models for the (*A*) 1-day, (*B*) 3-day, (*C*) 7-day, and (*D*) 10-day duration floods, Columbia River Basin, northwestern United States and British Columbia, Canada. Line in each scatterplot is the fitted regional skew model.

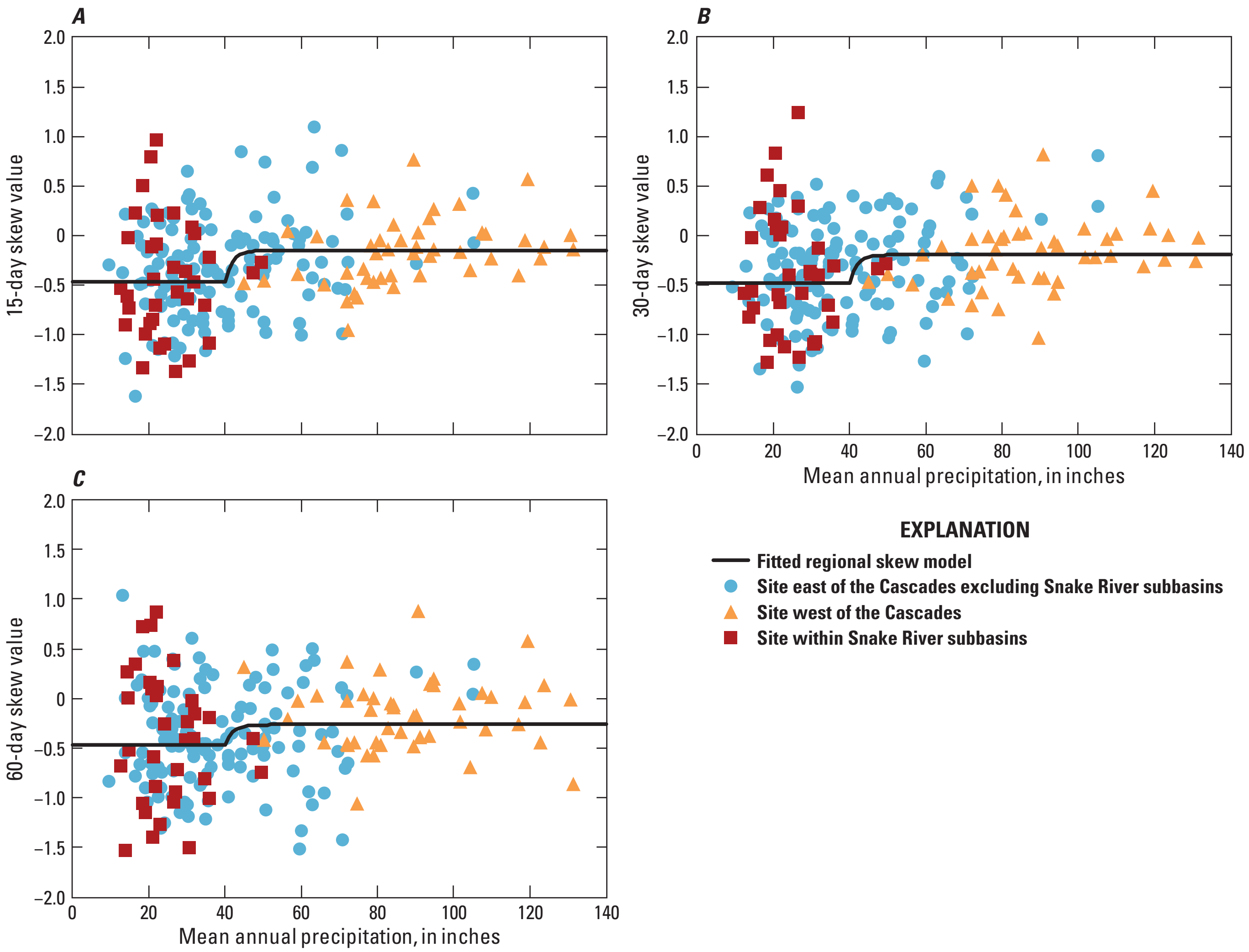


Figure 18. Fitted regional skew models for the (*A*) 15-day, (*B*) 30-day, and (*C*) 60-day duration floods, Columbia River Basin, northwestern United States and British Columbia, Canada. Line in each scatterplot is the fitted regional skew model.

at about 40 in. of precipitation per year, including all sites west of the Cascade Range as well as some sites east of the Cascade Range. The at-site log-space skew coefficients for the Snake River subbasin show significant variability.

The fitted regional skew models for all seven durations are shown in figure 19. Overall, the regional skew coefficients range from -0.48 for 15- and 30-day duration floods at sites with less than 40 in. of mean annual precipitation, to 0.09 for 1-day duration floods at sites with more than 51 in. of mean annual precipitation (see table 10). Regional skew coefficients generally decrease with increasing duration, particularly at higher levels of mean annual precipitation. The difference between regional skew coefficients for high and low precipitation sites also tends to decrease with increasing duration. The statistical significance (based on p-values) of $\hat{\beta}_{1,d}$ decreased with increasing duration. The trend in high compared to low precipitation differences could indicate that flood characteristics are more homogenous across the study basins for longer durations than shorter ones. The dampening effect of averaging flood flows over longer periods of time likely plays some role in the increasing uniformity (Lamontagne and others, 2012).

Use of Regional Skew Models

In Bulletin 17C, the at-site skew coefficient is combined with a regional skew coefficient in a weighted average where the weights depend on the relative precision of the at-site and regional skew coefficients (England and others, 2018). The skew coefficient is difficult to estimate, even with relatively long records, and might be affected by very large or small floods (potentially influential low flows; see Lamontagne and others [2012]). To estimate the regional skew coefficient at a site, the mean annual precipitation must be known. Figure 19 or equation 21 and table 8 can then be used to compute the regional skew coefficient for a given duration. For sites used in this analysis, the regional skew coefficients are reported

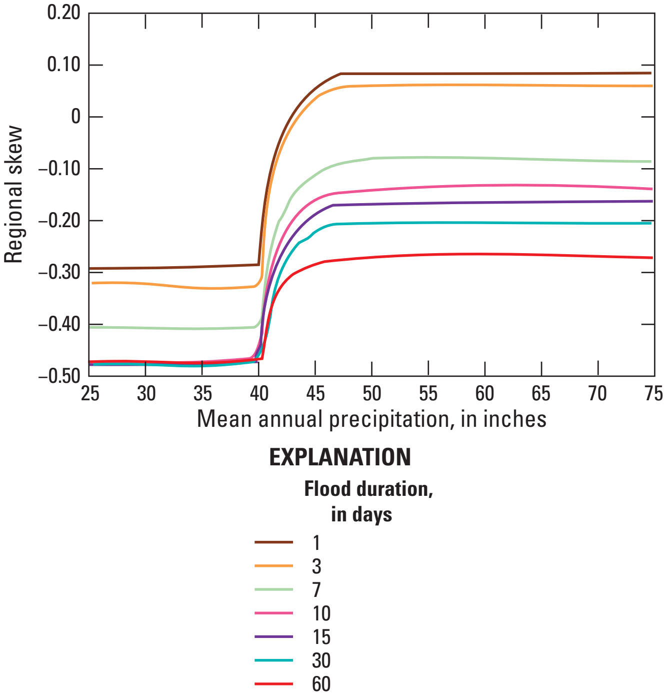


Figure 19. Nonlinear final fitted models of regional skew for all durations in the Columbia River Basin, Columbia River Basin, northwestern United States and British Columbia, Canada.

in table 2.3 of appendix 2. Alternatively, table 10 reports the regional skew coefficients directly (in 1-in. precipitation increments).

The variances of prediction (VPs) reported in this study are equivalent to the MSE from the National Skew Map in Bulletin 17B and should be used to weigh the at-site and regional skew coefficients. Because the nonlinear model in equation 21 is a function of precipitation, the variance of prediction also changes with precipitation. Furthermore, the VP for sites included in the study, VP_{old}, is less than a site with the same precipitation that was not used in the study, VP_{new} (see appendix 3). For a new site not used to compute the regional skew coefficients in this study, table 10 reports the appropriate variance of prediction as a function of flood duration and mean annual basin precipitation. Table 2.4 in appendix 2 reports the appropriate variance of prediction for sites used in the study.

To use the regional model developed in this report, the analyst should follow these steps:

1. Compute the regional skew coefficient for a site for the desired duration using (A) equation 21 and table 8, or (B) figure 18, or (C) table 10 (to the nearest inch).
2. Determine whether the site was used to derive the regional models, by checking table 2.4 in appendix 2.
3. If the site was used to develop the regional skew model, select the site VP_{old} statistic for the appropriate duration; otherwise use table 10 to obtain VP_{new}.
4. Use the regional skew coefficient from step (1) and the appropriate variance of prediction from step (3) in the EMA procedure as implemented in PeakFQ, a USGS software application used to compute AEPs (Flynn and others, 2006; Veilleux and others, 2014).

As an example, consider a 7-day regional skew estimation for the Columbia River at the NRNI site The Dalles (TDA). The mean annual basin precipitation for TDA is about 26 in. Using equation 21, figure 18, or table 10, the 7-day regional skew coefficient is -0.40. An examination of table 2.4 in appendix 2 indicates that TDA was not used to derive the regional skew models, so the 7-day VP_{new} from table 10 should be used: 0.09. PeakFQ could then be used with this regional skew coefficient and VP to estimate the desired 7-day AEP flood.

As another example, consider 7-day regional skew estimation for Charlton Creek above Crane Prairie Reservoir near La Pine, Oregon (USGS streamgage 14053000). The mean annual basin precipitation for 14053000 is about 61.9 in. Using equation 21, figure 19, or table 10, the 7-day regional skew coefficient is -0.08. An examination of table 2.4 in appendix 2 indicates that 14054500 was used to derive the regional skew models, so the 7-day VP_{old} from table 2.4 in appendix 2 should be used: 0.10. PeakFQ could then be used with this regional skew coefficient and VP to estimate the desired 7-day AEP flood. Appendix 2, table 2.5 presents skew coefficients and VP_{new} values for NRNI sites not used to fit the regional skew models.

Table 10. Regional skew coefficients and variance of prediction for seven durations as a function of mean annual basin precipitation, Columbia River Basin, northwestern United States and British Columbia, Canada.

[VP_{new}: Variance of prediction for a site not used in fitting the models]

Basin precipitation (inches)	Duration													
	1 day		3 days		7 days		10 days		15 days		30 days		60 days	
	Skew	VP_{new}	Skew	VP_{new}	Skew	VP_{new}	Skew	VP_{new}	Skew	VP_{new}	Skew	VP_{new}	Skew	VP_{new}
≤40	−0.28	0.09	−0.32	0.06	−0.4	0.09	−0.47	0.09	−0.48	0.09	−0.48	0.08	−0.47	0.11
41	−0.14	0.08	−0.17	0.06	−0.4	0.09	−0.47	0.09	−0.48	0.09	−0.48	0.08	−0.47	0.11
42	−0.05	0.09	−0.07	0.06	−0.27	0.09	−0.34	0.09	−0.35	0.09	−0.37	0.08	−0.39	0.11
43	0.01	0.09	−0.02	0.06	−0.2	0.09	−0.26	0.09	−0.27	0.09	−0.3	0.08	−0.34	0.11
44	0.04	0.09	0.02	0.07	−0.15	0.09	−0.21	0.09	−0.23	0.09	−0.26	0.08	−0.31	0.11
45	0.06	0.09	0.04	0.07	−0.12	0.09	−0.18	0.09	−0.2	0.09	−0.24	0.08	−0.29	0.11
46	0.07	0.09	0.05	0.07	−0.1	0.09	−0.16	0.09	−0.18	0.09	−0.22	0.09	−0.28	0.12
47	0.08	0.09	0.06	0.07	−0.09	0.1	−0.15	0.09	−0.17	0.09	−0.21	0.09	−0.28	0.12
48	0.09	0.09	0.06	0.07	−0.09	0.1	−0.14	0.09	−0.17	0.09	−0.21	0.09	−0.27	0.12
49	0.09	0.09	0.06	0.07	−0.08	0.1	−0.14	0.09	−0.16	0.09	−0.2	0.09	−0.27	0.12
50	0.09	0.09	0.07	0.07	−0.08	0.1	−0.14	0.09	−0.16	0.09	−0.2	0.09	−0.27	0.12
≥51	0.09	0.09	0.07	0.07	−0.08	0.1	−0.14	0.1	−0.16	0.09	−0.2	0.09	−0.27	0.12

Summary

Flood-frequency estimates are essential to effectively design, operate, and maintain hydraulic structures such as bridges and dams, especially considering that many of them have aged beyond their designed life spans. Failures of these structures could cause catastrophic loss of property, life, or both. People responsible for designing and maintaining bridges rely on frequency estimates of flood peaks, which are instantaneous values. For operators of dams, additional flood-frequency estimates of interest are flood durations, which are running averages, instead of instantaneous values, usually described by multiple days. This report provides regional skew estimates for 1-, 3-, 7-, 10-, 15-, 30-, and 60-day flood durations.

U.S. Geological Survey Guidelines for Determining Flood Flow Frequency (Bulletin 17C) provided guidance to government agencies involved with flood-frequency studies since its publication in 2018 (U.S. Geological Survey, 2018). The skew coefficient used in flood-frequency estimates is influential in predicting large floods, and methods used to estimate skew coefficients, as recommended by Bulletin 17C, have progressed for flood-frequency studies. In this study, over 300 unregulated U.S. Geological Survey (USGS) streamgage records and almost 100 naturalized streamflow records from regulated sites (with a dam and or reservoir upstream) were used in a regional skew study for the Columbia River Basin of the northwestern United States and British Columbia, Canada. The naturalized streamflow records, provided by the U.S. Army Corps of Engineers, Bureau of Reclamation, and Bonneville Power Administration, were reviewed for consistency and reasonableness by USGS hydrologists.

This report used Bayesian statistical regression methods, which have been used for numerous flood-frequency studies, to develop and analyze regional skew models based on hydrologically significant basin characteristics. Various exercises applying regression analyses and modern statistical tools were conducted while exploring potentially important variables. As expected, basin characteristics such as elevation and climate proved to explain some variability in skew coefficients for the Columbia River Basin sites. After examining the suite of available basin characteristics, precipitation seemed to show the most significance across all flood durations. Using incremental steps of mean annual precipitation while developing skew models, 40 in. of annual precipitation seemed to be a natural breakpoint for the relations between basins and their skew coefficients. As such, a nonlinear regression model was fitted for all N-day durations to precipitation with a sigmoidal function used to smoothly transition the boundary of 40 in. of precipitation a year. The skew coefficient is constant when the mean annual precipitation is below 40 in. and above 51 in. The skew coefficient varies for the transition between constants, from 40 to 51 in. of mean annual precipitation.

As flood-frequency studies progress, additional basin characteristics, and (or) their combinations, may lead to improved regional skew models for the Columbia River Basin. This report is intended to facilitate future studies by providing a summary of trials and results while investigating skew models based on basin characteristics using modern techniques. Reliable flood-frequency estimates can be challenging, especially for a river basin as large and diverse as the Columbia River. The findings of this study provide a foundation for future studies to build upon.

References Cited

Bonneville Power Administration, 2011, 2010 level modified streamflow 1928–2008: Portland, Oregon, Bonneville Power Administration DOE/BP–4352, 581 p.

Barlow, P.M., Cunningham, W.L., Zhai, T., and Gray, M., 2014, U.S. Geological Survey Groundwater toolbox, a graphical and mapping interface for analysis of hydrologic data (version 1.0)—User guide for estimation of base flow, runoff, and groundwater recharge from streamflow data: U.S. Geological Survey Techniques and Methods, book 3, chap. B10, 27 p., https://doi.org/10.3133/tm3B10.

Bureau of Reclamation, 2002, Interim comprehensive basin operating plan for the Yakima Project, Washington: Yakima, Washington, Bureau of Reclamation, [variously paged].

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. https://doi.org/10.1002/wrcr.20392.

Cohn, T.A., Lane, W.L., and Baier, W.G., 1997, An algorithm, for computing moments-based flood quantile estimates when historical flood information is available: Water Resources, v. 33, no. 9, p. 2089–2096. https://doi.org/10.1029/97WR01640.

Cohn, T.A., Lane, W.L., and Stedinger, J.R., 2001, Confidence intervals for expected moments algorithm flood quantile estimates: Water Resources Research, v. 37, no. 6, p. 1695–1706. https://doi.org/10.1029/2001WR900016.

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

Cooper, R.M., 2005, Estimation of peak discharges for rural, unregulated streams in western Oregon: U.S. Geological Survey Scientific Investigations Report 2005–5116, 134 p., https://pubs.usgs.gov/sir/2005/5116/ https://doi.org/10.3133/sir20055116

Cooper, R.M., 2006, Estimation of peak discharges for rural, unregulated streams in eastern Oregon: State of Oregon Water Resources Department Open File Report SW 06-001, 86 p. plus appendixes.

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

Dettinger, M.D., Ralph, F.M., Das, T., Neiman, P.J., and Cayan, D.R., 2011, Atmospheric rivers, floods and the water resources of California: Water (Basel), v. 3, no. 2, p. 445–478 https://doi.org/10.3390/w3020445.

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., https://doi.org/10.3133/tm4B5.

England, J.F., Jr., Salas, J.D., and Jarrett, R.D., 2003, Comparisons of two moments-based estimators that utilize historical and paleoflood data for the log Pearson type III distribution: Water Resources Research, v. 39, no. 9, p. 16, https://doi.org/10.1029/2002WR001791.

Esri, 2016, ArcGIS™ desktop—Release 10.4: Redlands, California, Environmental Systems Research Institute software application.

Falcone, J.A., 2011, GAGES-II—Geospatial attributes of gages for evaluation streamflow: U.S. Geological Survey digital spatial dataset, accessed July 17, 2017, at https://pubs.er.usgs.gov/publication/70046617.

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., https://doi.org/10.3133/sir20095156.

Ferguson, S.A., 1999, Climatology of the interior Columbia River Basin: U.S. Forest Service General Technical Report PNW-GTR-445, 40 p. https://doi.org/10.2737/PNW-GTR-445

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

Fu, G.B., Barber, M.E., and Chen, S., 2007, The impacts of climate change on regional hydrological regimes in the Spokane River watershed: Journal of Hydrologic Engineering, v. 12, no. 5, p. 452–461, https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(452).

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., https://doi.org/10.3133/sir20095043.

Griffis, V.W., 2006, Flood frequency analysis—Bulletin 17, regional information, and climate change: Ithaca, New York, 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, https://doi.org/10.1061/(ASCE)1084-0699(2009)14:2(121).

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, no. 7, W07503 https://doi.org/10.1029/2003WR002697.

Grubbs, F.E., and Beck, G., 1972, Extension of sample sizes and percentage points for significance tests of outlying observations: Technometrics, v. 14, no. 4, p. 847–854 https://doi.org/10.1080/00401706.1972.10488981.

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 2007, May 15–18, 2007, Tampa, Florida: American Society of Civil Engineers, 10 p.

Gruber, A.M., and Stedinger, J.R., 2008, Models of LP3 regional skew, data selection, and Bayesian GLS regression, *in* Babcock, R.W., Jr., and Walton, R., eds., Ahupua‘a—Proceedings of the World Environmental and Water Resources Congress 2008, May 12–16, 2008, Honolulu, Hawaii: American Society of Civil Engineers, paper 596, 10 p. https://doi.org/10.1061/40976(316)563

Hamlet, A., and Lettenmaier, D., 2007, Effects of 20th century warming and climate variability on flood risk in the western U.S: Water Resources Research, v. 43, no. 6, https://doi.org/10.1029/2006WR005099.

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.

Horizon Systems Corporation, 2017, NHDPlus: Horizon Systems web page, accessed April 10, 2017, at http://www.horizon-systems.com/NHDPlus/index.php

Hulsing, H., and Kallio, N.A., 1964, Magnitude and frequency of floods in the United States—Part 14, Pacific Slope basins in Oregon and lower Columbia River Basin: U.S. Geological Survey Water-Supply Paper 1689, 320 p.

Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood-flow frequency: U.S. Geological Survey Bulletin 17B, 183 p. [Also available at https://water.usgs.gov/osw/bulletin17b/dl_flow.pdf.]

Kennedy, J.R., Paretti, N.V., and Veilleux, A.G., 2015, Methods for estimating magnitude and frequency of 1-, 3-, 7-, 15-, and 30-day flood-duration flows in Arizona (version 1.1, April 2015): U.S. Geological Survey Scientific Investigations Report 2014–5109, 35 p., 10.3133/sir20145109.

Klein, C.A., and Zellmer, S.B., 2007, Mississippi River Stories—Lessons from a Century of Unnatural Disasters: Southern Methodist University Law Review, v. 60, p. 1471–1537, https://scholar.smu.edu/smulr/vol60/iss4/7. https://doi.org/10.2139/ssrn.1010611.

Lamontagne, J.R., 2014, Development of regional skew models for rainfall floods in California using Bayesian least squares regression: Ithaca, New York, Cornell University, Master’s thesis, 226 p.

Lamontagne, J.R., and 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. https://doi.org/10.3133/sir20125130

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, https://doi.org/10.1029/2002WR001589.

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

McKenzie, S., 2013, A river runs through it—The future of the Columbia River Treaty, water rights, development, and climate change: Georgia State University Law Review, v. 29, no. 4, 38 p.

Northwest Power Planning Council, 2000, Draft Spokane River subbasin summary: Portland, Oregon, Northwest Power Planning Council Portland, 45 p.

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

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., https://doi.org/10.3133/sir20105260.

PRISM Climate Group, 2017, 30-year normals: Northwest Alliance for Computational Science and Engineering, Oregon State University database, accessed June 4, 2015, at https://prism.oregonstate.edu/normals.

R Core Team, 2017, R—A language and environment for statistical computing, version 3.4.1: Vienna, Austria, R Foundation for Statistical Computing, accessed June 7, 2020, at https://www.R-project.org/

Rantz, S.E., and Riggs, H.C., 1949, Floods of May–June 1948 in Columbia River Basin: U.S. Geological Survey Water-Supply Paper 1080, 476 p.

Reis, D.S., Jr., Stedinger, J.R., and Martins, E.S., 2005, Bayesian generalized least squares regression with application to the log Pearson type III regional skew estimation: Water Resources Research, v. 41, no. 10, W10419 https://doi.org/10.1029/2004WR003445.

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., accessed July 10, 2020, at https://doi.org/10.3133/fs20173046. [Supersedes U.S. Geological Survey Fact Sheet 2008–3067.]

Rinella, J.F., McKenzie, S.W., and Fuhrer, G.J., 1992, Executive summary—Surface-water-quality assessment of the Yakima River Basin, Washington—Analysis of available water quality data through 1985 water year: U.S. Geological Survey Open File Report 91–453, 15 p., accessed July 10, 2020, at https://doi.org/10.3133/ofr91454.

Rodgers, J.L., and Nicewander, W.A., 1988, Thirteen ways to look at the correlation coeffcient: The American Statistician, v. 42, no. 1, p. 59–66, accessed July 10, 2020, at https://doi.org/10.2307/2685263.

Rutz, J.J., and Steenburgh, W.J., 2012, W. J. Quantifying the role of atmospheric rivers in the interior western United States: Atmospheric Science Letters, v. 13, no. 4, p. 257–261, accessed July 10, 2020, at https://doi.org/10.1002/asl.392.

Seaber, P.R., Kapinos, F.P., and Knapp, G.L., 1987, Hydrologic unit maps: U.S. Geological Survey Water-Supply Paper 2294, 63 p.

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., https://doi.org/10.3133/sir20145165.

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 [with correction, Water Resources Research, 1986, v. 22, no. 5, p. 844], https://doi.org/10.1029/WR021i009p01421.

Tasker, G.D., 1983, Effective record length for the T-year event: Journal of Hydrology (Amsterdam), v. 64, no. 1-4, p. 39–47 https://doi.org/10.1016/0022-1694(83)90059-8.

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 https://doi.org/10.1061/(ASCE)0733-9496(1986)112:2(225).

Tasker, G.D., and Stedinger, J.R., 1989, An operational GLS model for hydrologic regression: Journal of Hydrology (Amsterdam), v. 111, no. 1–4, p. 361–375 https://doi.org/10.1016/0022-1694(89)90268-0.

MathWorks, Inc., 2016, MATLAB release 2016a: Natick, Massachusetts, Mathworks, Inc. software application.

U.S. Army Corps of Engineers, 2014, Draft NRNI flows—Columbia River Basin development of NRNI flows using 2010 modified flow results: U.S. Army Corps of Engineers, 12 p.

U.S. Geological Survey, 2017, USGS water data for the nation: U.S. Geological Survey National Water Information System data release, accessed July 3, 2017, at https://doi.org/10.5066/F7P55KJN.

U.S. Geological Survey, 2018, National elevation dataset: U.S. Geological Survey dataset, accessed October 2018 at https://catalog.data.gov/dataset/usgs-national-elevation-dataset-ned

Veilleux, A.G., 2009, Bayesian GLS regression for regionalization of hydrologic statistics, floods and Bulletin 17 skew: Ithaca, New York, Cornell University, Master's thesis, 155 p.

Veilleux, A.G., 2011, Bayesian GLS regression, leverage and influence for regionalization of hydrologic statistics: Ithaca, New York, 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., https://doi.org/10.3133/fs20133108.

Veilleux, A.G., Stedinger, J.R., and Eash, D.A., 2012, Bayesian WLS/GLS regression for regional skewness analysis for regions with large crest stage gage networks, *in* Loucks, E.D., ed., Proceedings of the World Environmental and Water Resources Congress 2012—Crossing boundaries, May 20–24, 2012, Albuquerque, New Mexico: American Society of Civil Engineers, p. 2253–2263, accessed March 5, 2018, at https://doi.org/10.1061/9780784412312.227.

Veilleux, A.G., Stedinger, J.R., and Lamontagne, J.R., 2011, Bayesian WLS/GLS regression for regional skewness analysis for regions with large cross-correlations among flood flows, in Proceedings of the World Environmental and Water Resources Congress 2011—Bearing knowledge for sustainability, May 22–26, 2011, Palm Springs, California: American Society of Civil Engineers, paper 1303, p. 3103–3112, https://doi.org/10.1061/41173(414)324.

Veilleux, A.G., Zarriello, P.J., Hodgkins, G.A., Ahearn, E.A., Olson, S.A., and Cohn, T.A., 2019, Methods for estimating regional coefficient of skewness for unregulated streams in New England, based on data through water year 2011: U.S. Geological Survey Scientific Investigations Report 2017–5037, 29 p., https://doi.org/10.3133/sir20175037.

Weaver, J.C., Feaster, T.D., and Gotvald, A.J., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 2, North Carolina: U.S. Geological Survey Scientific Investigations Report 2009–5158, 113 p., https://doi.org/10.3133/sir20095158.

Wood, M.S., Fosness, R.L., Skinner, K.D., and Veilleux, A.G., 2016, Estimating peak-flow frequency statistics for selected gaged and ungaged sites in naturally flowing streams and rivers in Idaho (version 1.1, April 2017): U.S. Geological Survey Scientific Investigations Report 2016–5083, 56 p., https://doi.org/10.3133/sir20165083.

Zellmer, S.B., 2004, A new corps of discovery for Missouri River management: Nebraska Law Review, v. 83, no. 2, article 4, p. 305–361, accessed June 2019 at https://digitalcommons.unl.edu/nlr/vol83/iss2/4.

This page intentionally left blank.

Appendix 1. Columbia River Basin Characteristics

Appendix 1 provides tables presenting climatological and physiographical basin characteristics for the subbasins used in this study. Table 1.1 presents basin characteristics available for USGS sites, table 1.2 presents basin characteristics available for USGS and some NRNI sites, and table 1.3 presents basin characteristics available for all the NRNI sites.

Table 1.1. Monthly climate and basin characteristics for U.S. Geological Survey sites from Geospatial Attributes of Gages for Evaluating Streamflow, version II report database (GAGES II; Falcone, 2011), Columbia River Basin, northwestern United States and British Columbia, Canada.

Table 1.1 is a .csv file showing U.S. Geological Survey (USGS) monthly climate and basin characteristics that is available for download at https://doi.org/10.3133/sir20205073.

Table 1.2. Characteristics for U.S. Geological Survey (USGS) and no-regulation no-irrigation (NRNI) sites in the Columbia River Basin, northwestern United States and British Columbia, Canada.

Table 1.2 is a .csv Microsoft Excel file showing basin characteristics for USGS and most no-regulation no-irrigation (NRNI) sites that is available for download at https://doi.org/10.3133/sir20205073.

Table 1.3. Characteristics for no-regulation no-irrigation (NRNI) sites in the Columbia River Basin, northwestern United States and British Columbia, Canada.

Table 1.3 is a .csv file showing basin characteristics for all NRNI sites that is available for download at https://doi.org/10.3133/sir20205073.

This page intentionally left blank.

Appendix 2. Ancillary Tables for Regional Skew Study in the Columbia River Basin

Appendix 2 provides tables presenting supplemental information. Table 2.1 presents the symbols used in the equations and their explanations. Table 2.2 presents potentially influential low floods that were censored for all sites. Table 2.3 presents manual overrides of the Multiple Grubbs-Beck censoring. Table 2.4 presents the skew and variance of prediction values for all the sites used to fit the regional skew models. Table 2.5 presents the skew and variance of prediction values for NRNI sites that were not used to fit the regional skew models.

Table 2.1. Table of symbols used for equations in this report.

Table 2.1 is a .csv file showing symbols used for the equations in this report that is available for download at https://doi.org/10.3133/sir20205073.

Table 2.2. Total number of potentially influential low floods (PILFs) censored for all sites, Columbia River Basin, northwestern United States and British Columbia, Canada.

Table 2.2 is a .csv file showing total number of potentially influential low floods (PILFs) censored for all sites. This file is available for download at https://doi.org/10.3133/sir20205073.

Table 2.3. Manual override of Multiple Grubbs-Beck Test censoring.

[**Site ID:** Site identifier. **Number censored MBBT:** Number censored Multiple Grubbs-Beck Test]

Site ID	Duration (days)	Number censored MGBT	Number censored manually
14171000	1	17	19
12325500	3	31	32
12332000	3	26	27
12340500	3	14	15
14044000	3	8	9
BLU	3	0	2
12325500	7	28	0
14167000	7	6	8
14181500	7	3	4
14162200	10	3	4
10406500	15	13	16
12457000	15	14	15
13344500	15	0	27
MER	60	9	4

Table 2.4. Skew and variance of prediction (VPold) values for sites used in regional skew, northwestern United States and British Columbia, Canada.

Table 2.4 is a .csv file showing skew and variance of prediction (VP_{old}) values for sites used to fit the regional skew models. This file is available for download at https://doi.org./10.3133/sir20205073.

Table 2.5. Skew and variance of prediction (VPnew) values for no-regulation no-irrigation sites not used in regional skew, northwestern United States and British Columbia, Canada.

Table 2.5 is a .csv file showing skew and variance of prediction (VP_{new}) values for sites not used to fit the regional skew models. This file is available for download at https://doi.org./10.3133/sir20205073.

Appendix 3. Diagnostic Statistics

This section describes the diagnostic statistics used to evaluate the goodness of fit and precision of the regional skew regression models developed in section, “Regional Duration—Skew Analysis.” These statistics include the variance of prediction, which is used to weight the regional skew with the at-site skew in the Bulletin 17C analysis (England and others, 2018), and leverage and influence, which measure the potential of an observation and actual effect on a regression analysis, respectively.

Variance of Prediction

The variance of prediction is a common measure when selecting among several models; lower values of variance of prediction indicate more precise estimates of the dependent variable. The variance of prediction depends on whether or not a site i was used to develop the regional skew model. For a new site, not included in developing the model, the variance of prediction is:

$$VP_{new}(i) = E\left[\sigma^2_{\delta,B-GLS}\right] + x_i\,Var\left[\hat{\beta}_{WLS}\right] x_i^T, \tag{3-1}$$

where

$\sigma^2_{\delta,B-GLS}$ is the estimated model error variance from the Bayesian generalized least squares (B-GLS) analysis described in section, “Step 3—Bayesian Generalized Least Squares”;

$\hat{\beta}_{WLS}$ is a $[K \times 1]$ vector of fitted model parameters from the Bayesian weighted least squares (B-WLS) analysis described in section, “Step 2—Bayesian Weighted Least Squares”; and

x_i is a $[1 \times K]$ vector of basin characteristics for site i, which was not used to fit the model.

Here, $\sigma^2_{\delta,B-GLS}$ indicates the underlying model error arising from using an imperfect model, and $x_i\,Var\left[\hat{\beta}_{WLS}\right] x_i^T$ indicates the precision of the fitted model parameters and the expected error when predicting the skew for a site with basin characteristics x_i. However, if site i was used to fit the regional model, the variance of prediction will be lower:

$$VP_{old}(i) = VP_{new} - 2E\left[\sigma^2_{\delta,B-GLS}\right] x_i W e_i, \tag{3-2}$$

where

e_i is a $[N{\times}1]$ vector with 1 in the ith row and zero otherwise; and

W is a $[K{\times}N]$ matrix of regression weights from the analysis described in section, “Step 2—Bayesian Weighted Least Squares,” which is computed as:

$$W = \left[X^T \Lambda_{WLS}{}^{-1} X\right]^{-1} X^T \Lambda_{WLS}^{-1}, \tag{3-3}$$

where

Λ_{WLS} is the covariance matrix from the B-WLS analysis described in section, “Step 2—Bayesian Weighted Least Squares.”

Average variance of prediction for a new site, AVP_{new}, commonly is reported in regression diagnostics (Mastin and others, 2016). AVP_{new} is computed taking the average of the computed $VP_{new}(i)$ for each site used to generate the regional model:

$$AVP_{new} = \tfrac{1}{n}\sum_{i=1}^{n} VP_{new}(i). \tag{3-4}$$

Leverage and Influence

Leverage identifies sites whose observation and explanatory variables are unusual and thus have the potential to exert significant influence on the fitted regression model (Hoaglin and Welsch, 1978). Sites are said to have high leverage if the value exceeds $2k/n$ (Stedinger and Tasker, 1985), which is 0.016 in this study. Leverage values do not vary substantially between durations because the matrix of basin characteristics and sample sizes are roughly the same for all durations. Leverage for the hybrid Bayesian weighted least squares/generalized least squares (B-WLS/B-GLS) analysis used in this study is derived from Veilleux and others (2011) and Lamontagne and others (2012), and is computed as:

$$H^{*}_{WLS} = XW. \tag{3-5}$$

Influence tries to identify sites that actually exert significant influence on the fitted regression model (Cook and Weisberg, 1982; Tasker and Stedinger, 1989). Sites with influence exceeding 4/n, where n is the total number of sites, are considered to have high influence; the value is 0.016 in this study. Because influence metrics depend partly on the residual from the fitted model, influence can vary significantly between durations. Influence for site i for the hybrid B-WLS/B-GLS analysis used in this study is derived from Veilleux and others (2011) and Lamontagne and others (2012), and is computed as:

$$D_i^{WG} = \left(\frac{1}{k}\right)\left(\frac{h^{*}_{WLS,ii}}{1-h^{*}_{WLS,ii}}\right)\left(\frac{\varepsilon_i^2}{Var[\hat{\varepsilon}|GLS\ model]}\right), \tag{3-6}$$

where

$h^{*}_{WLS,ii}$ is the i^{th} diagonal element of H^{*}_{WLS}, and

$$\begin{aligned} &Var[\hat{\varepsilon}|GLS\ model] = \\ &\Lambda_{GLS} - (H^{*}_{WLS})\Lambda_{GLS} - \Lambda_{GLS}(H^{*}_{WLS})^{T} \\ &+(H^{*}_{WLS})\Lambda_{GLS}(H^{*}_{WLS})^{T} \end{aligned} \tag{3-7}$$

where

Λ_{GLS} is the covariance matrix used in the B-GLS analysis used in section, “Step 3—Bayesian Generalized Least Squares.”

Error Variance Ratio and Misrepresentation of Beta Variance

The error variance ratio (EVR) is a regression diagnostic designed to determine whether a weighted least squares (WLS) or generalized least squares (GLS) analysis is needed or whether a simpler ordinary least squares (OLS) regression is sufficient. The statistic measures the relative magnitudes of the sampling error and model error. If model error is much larger than sampling error and OLS analysis is sufficient, but if sampling errors are large, a WLS or GLS analysis is warranted. EVR is computed as

$$EVR = \frac{\sum_{i=1}^{n} Var(\hat{\gamma}_i)}{n\,\sigma^2_{\delta,GLS}(k)}, \tag{3-8}$$

where

$Var(\hat{\gamma}_i)$ is the variance of the unbiased skew coefficient estimator for site i (eq. 9),

$\sigma^2_{\delta,GLS}(k)$ is the model error variance from the hybrid B-WLS/B-GLS analysis (eq. 6), and

n is the number of sites used in the regression.

Values of $EVR > 0.2$ usually are considered indicative of the need for a more complex analysis.

The misrepresentation of beta variance (MBV) is a statistic used to determine if a GLS analysis is needed or if a WLS analysis is sufficient (Griffis, 2006; Veilleux, 2011). Stedinger and Tasker (1985) noted that correlation between at-site station skews has the largest effect on the estimated precision of the constant term in a regression. Specifically, if correlations between station skews are large, the estimated variance of the regression constant from a WLS will be smaller than from a GLS analysis. If the cross-correlations between station skews are insignificant, the variance of the regression constant from a WLS analysis should be similar to that from a GLS analysis. Veilleux and others (2011) recast the misrepresentation of beta variance statistic for the hybrid B-WLS/B-GLS analysis as follows:

$$MBV = \frac{Var\left[b_0^{WLS}|GLS\ analysis\right]}{Var\left[b_0^{WLS}|WLS\ analysis\right]} = \frac{w^{T}\Lambda w}{\sum_{i=1}^{n} w_i}, \tag{3-9}$$

where $w_i = \frac{1}{\sqrt{\Lambda_{ii}}}$. Values of MBV greater than 1 indicate that GLS analysis should be used.

Leverage and Influence for the Columbia River Basin

The number of sites exerting unusual influence on the fitted regression varied depending on the duration. The leverage and influence for the 20 sites with the highest influence in each duration is shown in figure 3.1. On each graph, the threshold for "high influence" is indicated by a dark black line (set at 0.016). Each site in each duration having high influence was examined for potential issues in the underlying frequency or some other irregularity. The most notable site is 14091500 (Metolius River near Grandview, Oregon), which had the highest influence for the 1-, 3-, 7-, 10-, and 15-Day duration, and second highest influence for the 30-Day duration. This site has high leverage because it had an unusually positive skew (mean annual precipitation of 50.4) and a very long record length ($P_{RL} = 95$ for all durations). To evaluate the potential effects of site 14091500 on the analysis, the final model was fit for all durations without that site. Dropping site 14091500 had limited effects on the fitted model and would not have changed the model selection determination.

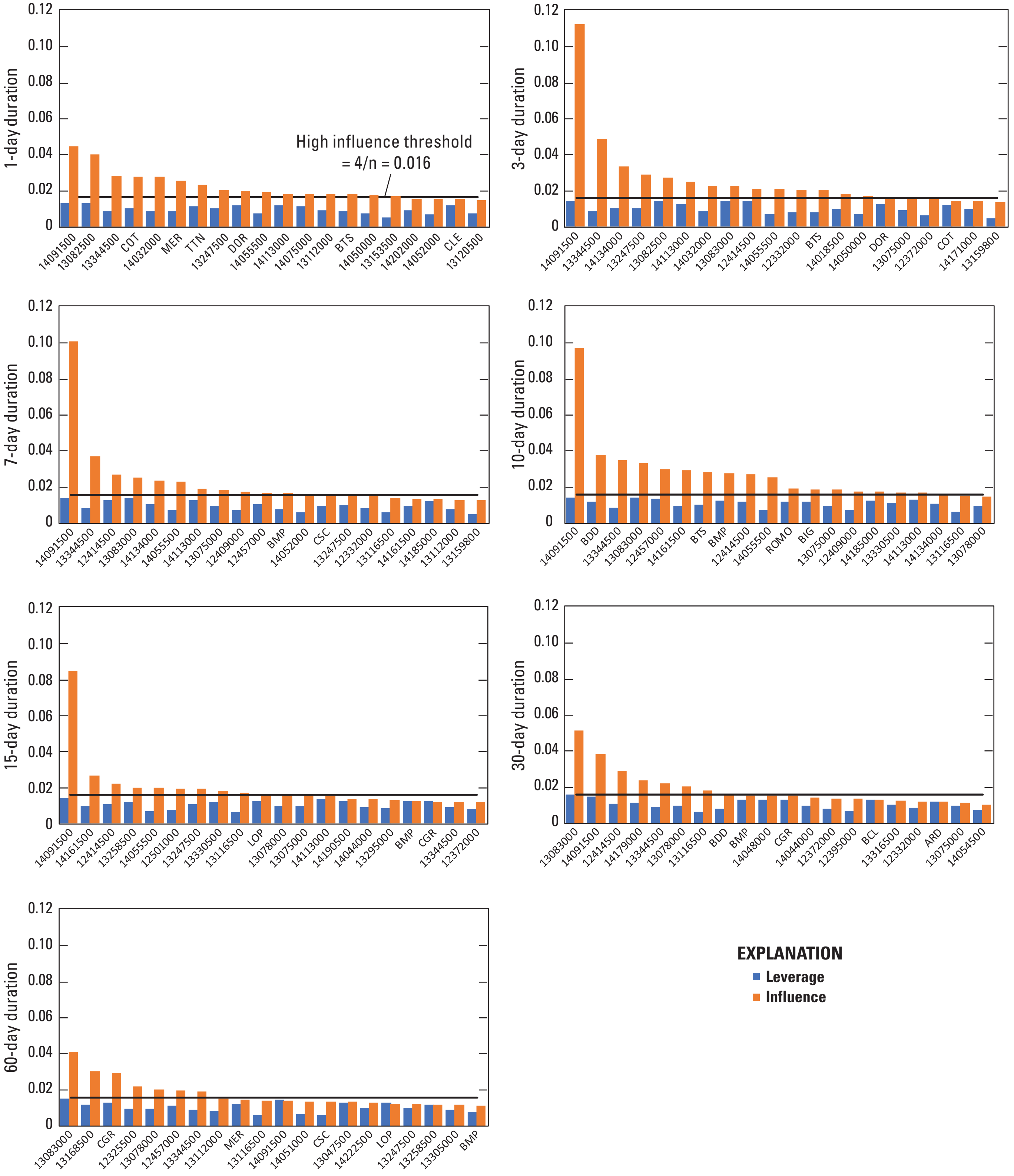


Figure 3.1. Leverage and influence for the 20 sites with the highest influence for each of seven flood durations, Columbia River Basin, northwestern United States and British Columbia, Canada. n, number of study sites.

Publishing support provided by the U.S. Geological Survey
Science Publishing Network, Tacoma Publishing Service Center

For more information concerning the research in this report, contact the
Director, Oregon Water Science Center
U.S. Geological Survey
2130 SW 5th Avenue
Portland, Oregon 97201
https://www.usgs.gov/centers/or-water

Lind and others—**Development of Regional Skew Coefficients for Selected Flood Durations, Columbia River Basin, United States and Canada**—SIR 2020–5073, ver. 1.1

ISSN 2328-0328 (online)
https://doi.org/10.3133/sir20205073