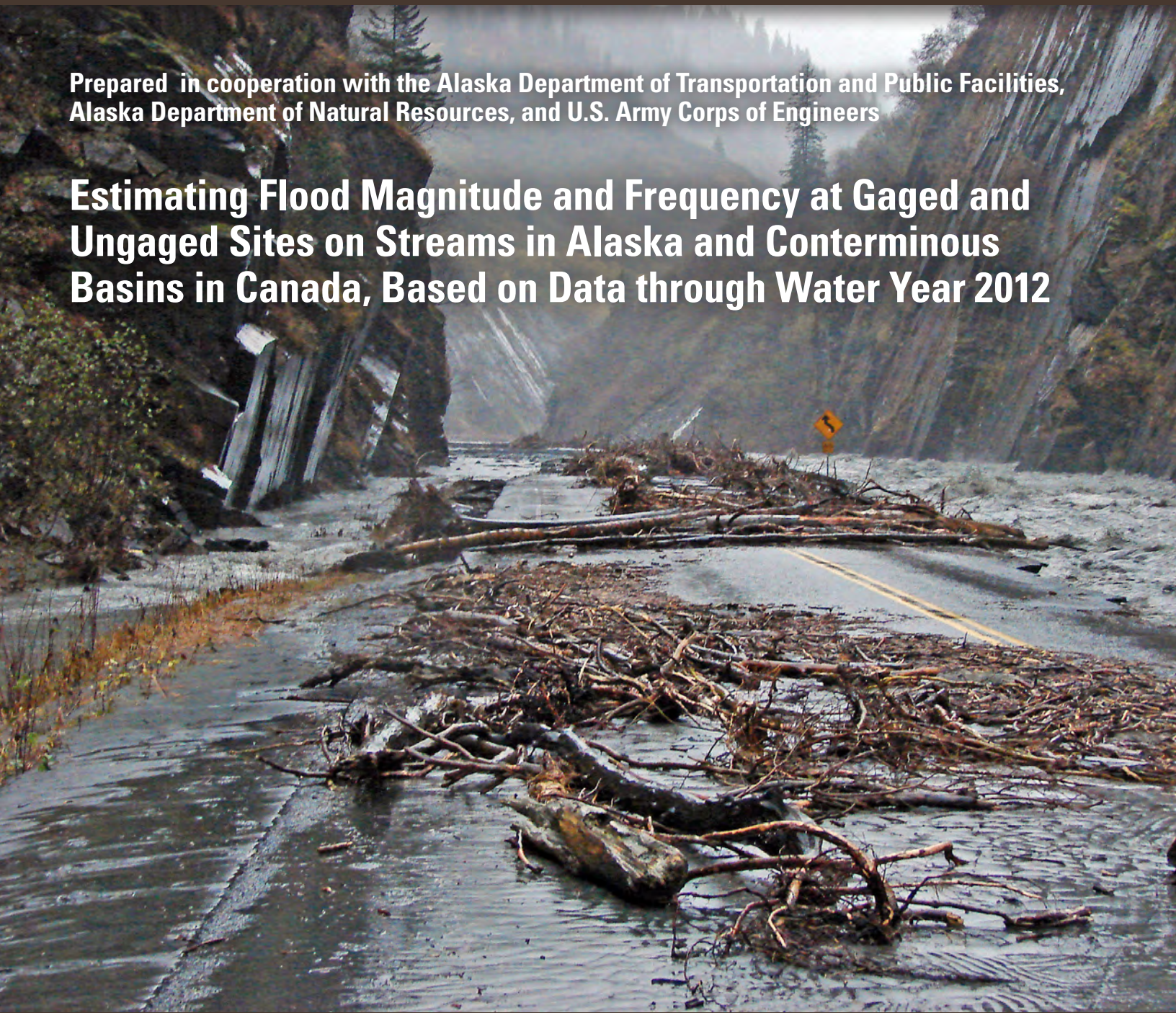


Prepared in cooperation with the Alaska Department of Transportation and Public Facilities, Alaska Department of Natural Resources, and U.S. Army Corps of Engineers

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



Scientific Investigations Report 2016–5024

Cover: Richardson Highway in Keystone Canyon during a flood of the Lowe River in October 2006. (Photograph by Mike Isaacs, Alaska Department of Transportation and Public Facilities, October 10, 2006.)

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

By Janet H. Curran, Nancy A. Barth, Andrea G. Veilleux, and Robert T. Ourso

Prepared in cooperation with the Alaska Department of Transportation and Public Facilities, Alaska Department of Natural Resources, and U.S. Army Corps of Engineers

Scientific Investigations Report 2016–5024

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

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

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

U.S. Geological Survey, Reston, Virginia: 2016

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 <http://www.usgs.gov> or call 1–888–ASK–USGS.

For an overview of USGS information products, including maps, imagery, and publications, visit <http://www.usgs.gov/pubprod/>.

This report was funded in part with qualified outer continental shelf oil and gas revenues by the Coastal Impact Assistance Program, Fish and Wildlife Service, U.S. Department of the Interior. The views and conclusions contained in this document should not be interpreted as representing the opinions or policies of the U.S. Department of Interior, Fish and Wildlife Service.

Although this information product, for the most part, is in the public domain, it also may contain copyrighted materials as noted in the text. Permission to reproduce copyrighted items must be secured from the copyright owner.

Suggested citation:

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., <http://dx.doi.org/10.3133/sir20165024>.

ISSN 2328-0328 (online)

Contents

Abstract.....	1
Introduction.....	2
Purpose and Scope	2
Previous Studies	2
Description of Study Area	3
Data Compilation.....	7
Peak-Flow Data.....	7
Trend Analysis.....	8
Physical and Climatic Basin Characteristics	8
Flood Magnitude and Frequency at Gaged Sites	11
Log-Pearson Type III Frequency Analysis	12
Expected Moments Algorithm	12
Systematic Record and Nonstandard Censored Flow	13
Historical Flood Information.....	13
Multiple Grubbs-Beck Test for Detecting Multiple Potentially Influential Low Floods	15
Statistical Analysis of Regional Skew.....	16
Estimating Flood Magnitude and Frequency at Ungaged Sites.....	18
Elimination of Redundant and Other Non-Eligible Sites from Regression Analysis	18
Exploratory Regression Analysis	19
Regionalization of Flood-Frequency Estimates.....	20
Regional Regression Equations.....	21
Accuracy and Limitations.....	23
Variance of Prediction	24
Standard Error of Prediction	24
Pseudo Coefficient of Determination	24
Prediction Intervals	25
Limitations.....	26
Application of Methods for Estimating Flood Magnitude and Frequency	27
Regression Estimate and Prediction Interval.....	27
Weighted Estimate for a Gaged Site.....	27
Estimate for an Ungaged Site near a Streamgage.....	28
Estimate for a Site on the Yukon River	29
Cook Inlet Basin Streamstats	30
Summary and Conclusions.....	31
Acknowledgments	32
References Cited.....	32
Appendix A. Basin Characteristics for Selected Streams in Alaska and Conterminous Basins in Canada	37
Appendix B. Regional Skewness Regression Analysis.....	39

Figures

1. Maps showing physical features and streamgages used in regional skew and regression analyses for Alaska and conterminous basins in Canada	4
2. Graph showing example of annual peak-flow series containing peaks collected outside the systematic gaging record, U.S. Geological Survey streamgage 15208000, Tonsina River at Tonsina, Alaska.....	14
3. Example of a Log-Pearson Type III flood frequency curve for U.S. Geological Survey streamgage 15208000, Tonsina River at Tonsina, Alaska, showing the Expected Moments Algorithm (EMA) with a multiple Grubbs-Beck (MGB) test representation for two peaks collected outside the systematic gaging record, water years 1995 and 2006	14
4. Log-Pearson Type III fit when multiple Grubbs-Beck (MGB) test found multiple potentially influential low floods (PILFs) for U.S. Geological Survey streamgage 15476400, Dry Creek near Dot Lake, Alaska	15
5. Map showing regional skew areas for Alaska and conterminous basins in Canada.....	16
6. Map showing streamflow analysis regions previously presented for Alaska and conterminous basins in Canada	20
7. Graph showing relation of discharge to drainage area for selected annual exceedance probabilities for the Yukon River at and downstream of Carmacks, Yukon.....	23
8. Graph showing relation between 1-percent annual exceedance probability discharges computed for observed streamflow and predicted from regression equations for streamgages in Alaska and conterminous basins in Canada	25

Tables

1. Description of streamgages used in flood frequency analysis and considered for use in regional skew and regional regression analysis for Alaska and conterminous basins in Canada.....	7
2. Basin characteristics considered in regional skew analysis and peak flow regression analysis for Alaska and conterminous basins in Canada.....	9
3. <i>P</i> -percent annual exceedance probabilities and corresponding <i>T</i> -year recurrence intervals for flood frequency flow estimates	11
4. Flood frequency statistics for streamgages in Alaska and conterminous basins in Canada with at least 10 years of record through water year 2012.....	12
5. Definition of regional skew areas for Alaska and conterminous basins in Canada, by hydrologic unit code (HUC).....	17
6. Regional skew and summary statistics for regions in Alaska and conterminous basins in Canada.....	18
7. Regional regression equations for estimating annual exceedance-probability discharges for unregulated streams in Alaska and conterminous basins in Canada	22
8. Values used to determine prediction intervals for the regional flood frequency regression equations for Alaska and conterminous basins in Canada.....	26
9. Variance estimates for station, regression, and weighted estimates of flood frequency statistics for streamgages in Alaska and conterminous basins in Canada....	28

Conversion Factors

Inch/Pound to International System of Units

Multiply	By	To obtain
Length		
inch (in.)	2.54	centimeter (cm)
inch (in.)	25.4	millimeter (mm)
foot (ft)	0.3048	meter (m)
mile (mi)	1.609	kilometer (km)
Area		
square mile (mi ²)	259.0	hectare (ha)
square mile (mi ²)	2.590	square kilometer (km ²)
Volume		
cubic foot (ft ³)	28.32	cubic decimeter (dm ³)
cubic foot (ft ³)	0.02832	cubic meter (m ³)
Flow rate		
cubic foot per second (ft ³ /s)	0.02832	cubic meter per second (m ³ /s)

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

$$^{\circ}\text{C}=(^{\circ}\text{F}-32)/1.8.$$

Datums

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

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

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

Acronyms

AEP	annual exceedance probability
ASC	Alaska Science Center
AVP	average variance of prediction
EMA	expected moments algorithm
GIS	geographic information system
GLS	generalized least squares
HUC	hydrologic unit code
MGB	multiple Grubbs-Beck
MSE	mean square error
NHD	National Hydrography Dataset
NWIS	National Water Information System
OLS	ordinary least squares
PRISM	Parameter-Elevation Regressions on Independent Slopes Model
USGS	U.S. Geological Survey
WBD	Watershed Boundary Dataset
WLS	weighted least squares
WREG	weighted-multiple-linear regression

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

By Janet H. Curran, Nancy A. Barth, Andrea G. Veilleux, and Robert T. Ourso

Abstract

Estimates of the magnitude and frequency of floods are needed across Alaska for engineering design of transportation and water-conveyance structures, flood-insurance studies, flood-plain management, and other water-resource purposes. This report updates methods for estimating flood magnitude and frequency in Alaska and conterminous basins in Canada. Annual peak-flow data through water year 2012 were compiled from 387 streamgages on unregulated streams with at least 10 years of record. Flood-frequency estimates were computed for each streamgage using the Expected Moments Algorithm to fit a Pearson Type III distribution to the logarithms of annual peak flows. A multiple Grubbs-Beck test was used to identify potentially influential low floods in the time series of peak flows for censoring in the flood frequency analysis.

For two new regional skew areas, flood-frequency estimates using station skew were computed for stations with at least 25 years of record for use in a Bayesian least-squares regression analysis to determine a regional skew value. The consideration of basin characteristics as explanatory variables for regional skew resulted in improvements in precision too small to warrant the additional model complexity, and a constant model was adopted. Regional Skew Area 1 in eastern-central Alaska had a regional skew of 0.54 and an average variance of prediction of 0.45, corresponding to an effective record length of 22 years. Regional Skew Area 2, encompassing coastal areas bordering the Gulf of Alaska, had a regional skew of 0.18 and an average variance of prediction of 0.12, corresponding to an effective record length of 59 years. Station flood-frequency estimates for study sites in regional skew areas were then recomputed using a weighted skew incorporating the station skew and regional skew. In a new regional skew exclusion area outside the regional skew areas, the density of long-record streamgages was too sparse for regional analysis and station skew was used for all estimates. Final station flood frequency estimates for all study streamgages are presented for the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probabilities.

Regional multiple-regression analysis was used to produce equations for estimating flood frequency statistics from explanatory basin characteristics. Basin characteristics, including physical and climatic variables, were updated for all study streamgages using a geographical information system and geospatial source data. Screening for similar-sized nested basins eliminated hydrologically redundant sites, and screening for eligibility for analysis of explanatory variables eliminated regulated peaks, outburst peaks, and sites with indeterminate basin characteristics. An ordinary least-squares regression used flood-frequency statistics and basin characteristics for 341 streamgages (284 in Alaska and 57 in Canada) to determine the most suitable combination of basin characteristics for a flood-frequency regression model and to explore regional grouping of streamgages for explaining variability in flood-frequency statistics across the study area. The most suitable model for explaining flood frequency used drainage area and mean annual precipitation as explanatory variables for the entire study area as a region. Final regression equations for estimating the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probability discharge in Alaska and conterminous basins in Canada were developed using a generalized least-squares regression. The average standard error of prediction for the regression equations for the various annual exceedance probabilities ranged from 69 to 82 percent, and the pseudo-coefficient of determination (pseudo- R^2) ranged from 85 to 91 percent.

The regional regression equations from this study were incorporated into the U.S. Geological Survey StreamStats program for a limited area of the State—the Cook Inlet Basin. StreamStats is a national web-based geographic information system application that facilitates retrieval of streamflow statistics and associated information. StreamStats retrieves published data for gaged sites and, for user-selected ungaged sites, delineates drainage areas from topographic and hydrographic data, computes basin characteristics, and computes flood frequency estimates using the regional regression equations.

Introduction

Flooding in Alaska has caused millions of dollars of damage to towns and villages and has disrupted major transportation links. Additionally, riverine infrastructure inadequately designed for flood flows has resulted in impaired aquatic biota and aquatic habitat. To minimize the damage from floods, protect human health and safety, and conserve wildlife habitat, reliable estimates of flood frequency and magnitude are essential. Federal, State, regional, and local agencies rely on these estimates to effectively plan and manage land use and water resources, protect lives and property, administer flood insurance programs, and conserve habitat. Streamflow statistics compiled from records maintained at streamgages provide a basis for the design of infrastructure, such as roads and bridges; the management of flood risk; the protection of aquatic species; and other engineering and environmental analyses. Flood frequency statistics can be computed directly for a particular streamgage with a suitable length of record. For short-record streamgages or ungaged sites, regression equations developed from flood frequency statistics and basin characteristics for a regional group of streamgages can provide estimates of flood frequency statistics.

The U.S. Geological Survey (USGS) has published reports that provide methods for estimating streamflow statistics at gaged and ungaged sites in States across the United States, including Alaska. These studies benefit from periodic updates to incorporate new streamflow information, which improve how well the equations represent the hydrology of the region, and improved measurements of basin characteristics or improved computational techniques, which can strengthen the statistical quality of the equations. Updates are of particular interest for Alaska, where the gaging network is spatially sparse, ranking near the bottom of all major hydrologic regions of the United States for percentage of area gaged except for very large basins. Individual streamgage record lengths in Alaska also are relatively short. Alaska reference-quality streamgages have a median record length of 28 years, the lowest among the 50 States (Kiang and others, 2013). The most recent streamflow statistics reports for Alaska consist of a flood frequency update (Curran and others, 2003) and a high-flow/low-flow statistics study (Wiley and Curran, 2003), both based on streamflow data through water year 1999. This present update of methods for estimating flood magnitude and frequency was undertaken to incorporate 13 additional years of streamflow data, comprehensively updated basin characteristics using new digital geospatial datasets, and new computational techniques including Bayesian weighted least-squares regression for analysis of generalized (regional) skew and Expected Moments Algorithm (EMA) with a multiple Grubbs-Beck (MGB) test for estimation of flood frequency statistics. The update also provides a pilot of a StreamStats model for Alaska, facilitating the delivery of streamflow statistics for the Cook Inlet Basin.

Purpose and Scope

This report presents results of a study conducted in cooperation with the Alaska Department of Transportation and Public Facilities, the Alaska Department of Natural Resources, and the U.S. Army Corps of Engineers to update methods for estimating the magnitude and frequency of floods in Alaska and conterminous basins in Canada. The selected areas of Canada were included to facilitate estimates in trans-boundary basins and improve estimates for near-border basins.

This report describes methods for determining the magnitude of floods that have annual exceedance probabilities (AEP) of 50, 20, 10, 4, 2, 1, 0.5, and 0.2 percent, and presents discharge estimates for these AEP for 387 streamgages in the study area. The report describes methods for revising regional skew and presents an updated regional skew for selected areas in Alaska and a new regional skew exclusion area for sparsely gaged areas. For estimating flood magnitude and frequency at ungaged basins, the report presents methods for developing regional regression equations using AEP discharge estimates at 341 streamgages and updated basin characteristics and describes the accuracy and limitations of these equations. Example applications for estimating flood frequency and magnitude in the study area are provided. This report also documents the development of a USGS StreamStats Web-based application for obtaining basin characteristics and estimating streamflow statistics, including flood magnitude and frequency, for a selected area of the State—the Cook Inlet Basin. StreamStats is intended to streamline procedures for estimating streamflow statistics by providing data and computation tools to minimize user error and the need for specialized software for managing large datasets. This report supersedes previous reports, most recently Curran and others (2003), that describe methods for estimating flood frequency and magnitude for Alaska.

Previous Studies

Early analyses of annual peak-flow statistics for Alaska include Berwick and others (1964), Childers (1970), Lamke (1978), and Parks and Madison (1985), all of which maximized use of the State's small but growing network of streamgages by including some or all streamgages with at least 5 years of peak-flow record. Beginning with Childers (1970), all studies applied a log-Pearson Type III analysis to annual peaks and adopted a multiple-regression technique using basin characteristics as independent, or explanatory, variables for estimating peak-flow statistics. All studies used a regional approach, resulting in a suite of equations for estimating streamflow magnitude and frequency in ungaged watersheds in each region. Regions varied by study, but commonly identified some part of the southern coastal areas of Alaska that border the Gulf of Alaska as a distinct hydrologic region. Lamke (1978) expressed a distinction between areas with autumn and winter rains and areas more commonly with spring and summer floods.

Subsequent peak-flow studies gradually increased the number of years of record required for inclusion in the study. Jones and Fahl (1994) included streamgages with at least 8 years of record instead of the recommended 10 years of record in an effort to include small streams, and Curran and others (2003) included streamgages with at least 10 years of record except for streamgages used in Jones and Fahl (1994) for which no additional records were available. The number of streamgages included in the regression analysis increased from 200 in Alaska in the 1985 study (Parks and Madison, 1985) to 355 in Alaska and conterminous basins in Canada in the 2003 study (Curran and others, 2003).

Parks and Madison (1985) presented a statewide regression equation for estimating flood frequency statistics and introduced six hydrologic regions in Alaska, presenting reliable regression equations for three of the regions. All regression equations contained drainage area and mean annual precipitation as explanatory variables. Jones and Fahl (1994) used three to six variables in regression equations for five regions in Alaska and conterminous basins in Canada, always including drainage area, mean annual precipitation, and area of lakes and ponds, and selectively including mean minimum January temperature, mean basin elevation, or area of forests to the equations. Curran and others (2003) updated peak-flow statistics with data through water year 1999 and slightly modified the region boundaries established by Jones and Fahl (1994) by splitting the northern part of the State into an interior and a far-north region. The 2003 study presented equations using one to four variables in six regions in Alaska and conterminous basins in Canada, always including drainage area and selectively including area of lakes and ponds, mean annual precipitation, area of forest, mean minimum January temperature, and elevation. Data included in previous reports included maps of climate characteristics that provide the required datasets of basin characteristics for regression analysis (Lamke, 1978; Jones and Fahl, 1994) and envelope curves developed from maximum known floods, most recently presented in Jones and Fahl (1994).

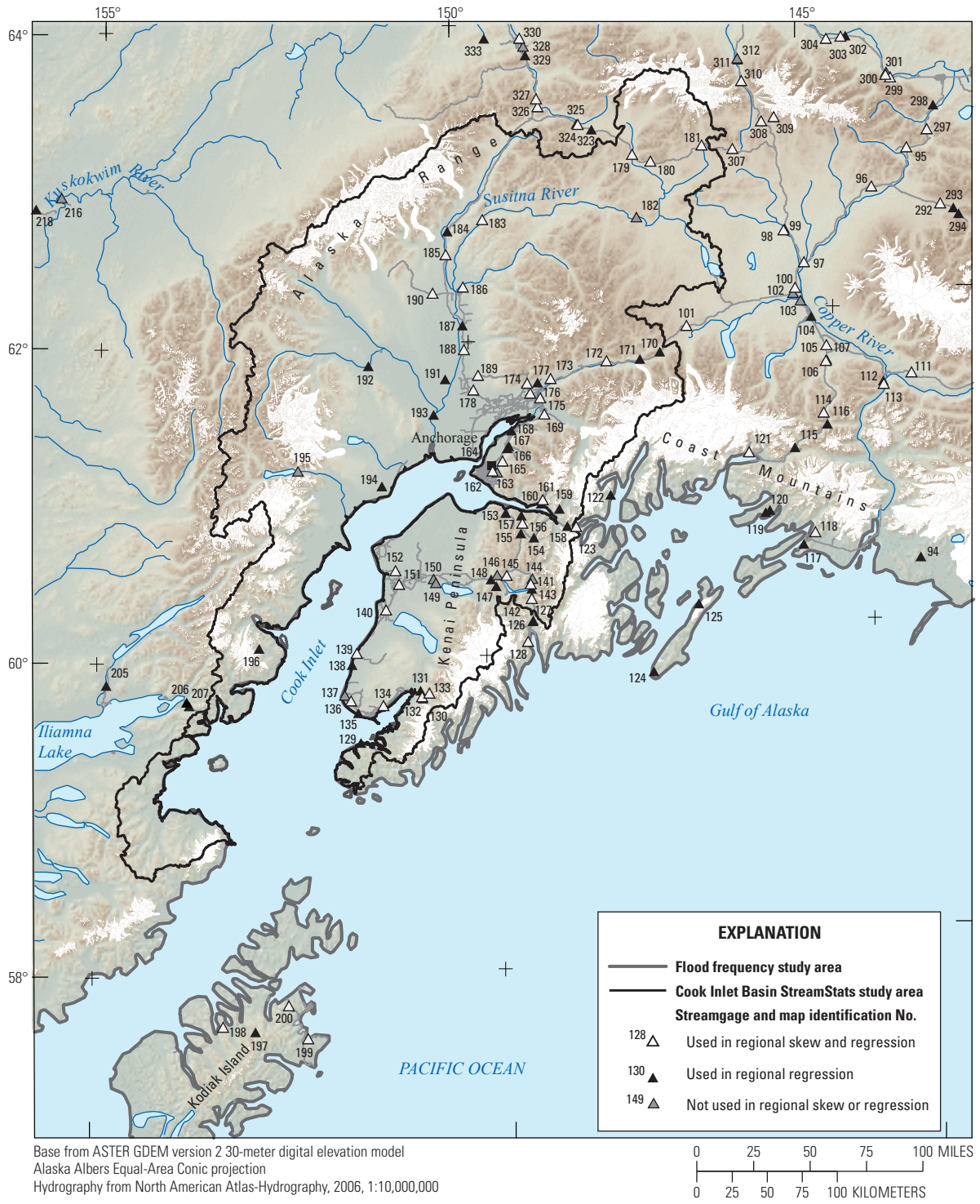
Description of Study Area

The area considered for flood frequency analysis in this study, referred to in this report as the “study area,” consists of the State of Alaska and Canadian basins in Yukon and British Columbia that drain to Alaska, an area encompassing 750,000 mi² (fig. 1). The study area excludes the Aleutian Islands and several other islands off the west coast of Alaska that lack sufficient streamgages to characterize streamflow for regression analysis. A small part of Alaska, the Cook Inlet Basin, which encompasses a land area of 39,000 mi² in a horseshoe shape around Cook Inlet (fig. 1 inset A), was included in the StreamStats application developed for this study.

The study area encompasses diverse physical and climatic settings. Mountainous areas arc in two major bands across the study area and include 23 summits exceeding 4,000 m (13,176 ft) in elevation. An extensive high-elevation mountain system consisting of the Coast Mountains, the Alaska Range, and the Aleutian Range spans the southern part of the study area. The Brooks Range, an extension of the Rocky Mountain system, extends across the study area near the Arctic Circle. The intermontane plateau separating these systems includes hills and gentler mountains in Interior Alaska and the expansive, low-relief, and lake-rich Yukon-Kuskokwim River delta. The cold, dry Arctic coastal plains extend north from the Brooks Range to the Arctic Ocean.

Environmental Protection Agency Level III Ecoregions (Gallant and others, 2010) and climate divisions (Bieniek and others, 2012) of Alaska define various regional divisions but generally recognize some part of temperate, wet south-east Alaska and cold, dry Arctic Alaska, respectively, as markedly different from the rest of the State, and generally divide the remaining area into interior, western, and south-central regions. Mean annual precipitation in the study area ranges from more than 300 in. in south-east Alaska to less than 9 in. on the Arctic coastal plain (Spatial Climate Analysis Service at Oregon State University, 2002; Gibson, 2009a). Average annual temperatures generally are correlated with latitude and range from 44 °F in southern areas to 10 °F in northern areas (Spatial Climate Analysis Service at Oregon State University, 2002; Gibson, 2009b). As a result of cold winter temperatures, winter precipitation can fall as snow onto frozen ground at high elevations throughout the study area and as low as sea level at many latitudes.

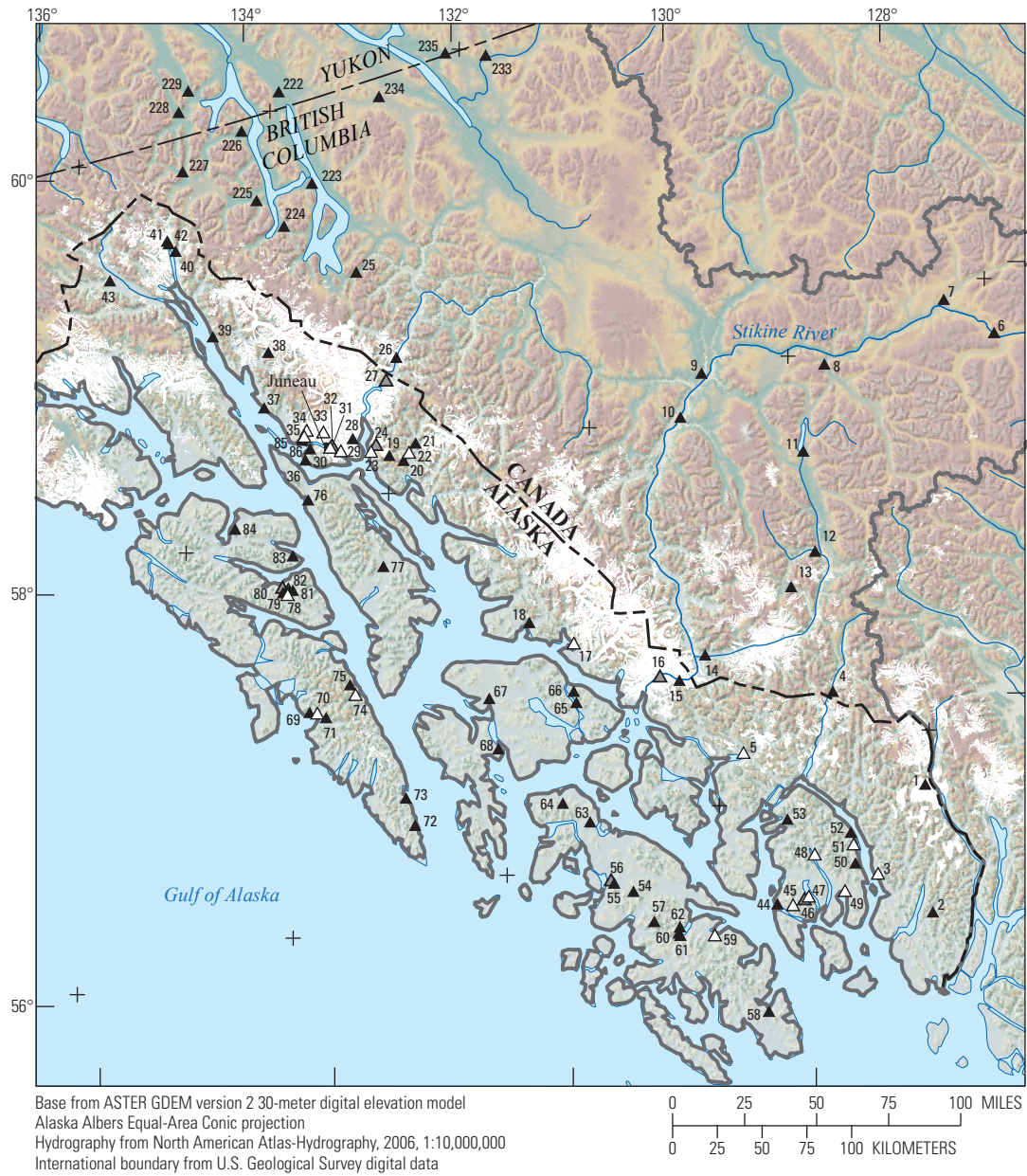
Processes generating peak flows in the study area can be inferred from inspection of the dates of occurrence of the peak flow and from previous studies by Wiley and Curran (2003) and Curran (2012). These processes include snowmelt, rainfall, and glacier-related melt. Although considered a single population for the purpose of flood frequency analysis, floods from these various processes respond to different forcing mechanisms. In most study streams, with the exception of those in basins bordering the Gulf of Alaska, spring snowmelt produces a prominent increase in mean daily discharge (see examples in Wiley and Curran, 2003, and Curran, 2012). This pulse of increased flow also can produce the annual peak flow, particularly in Interior and northern Alaska streams. In basins covered by a considerable area of glaciers, mean daily streamflow often continues to increase into summer as glacier and high-elevation snowmelt and glacier melt augment streamflow. Streamflow in glacierized basins can peak from these glacier-related melt processes, from spring snowmelt, or from rainfall. The glacier-related melt contribution to flow is a separate process from glacier outbursts, in which water impounded by or stored within the glacier is released suddenly. Glacier outbursts, which occur annually in some basins, can exceed floods from non-outburst processes.



INSET A, South-central Alaska

Figure 1.—Continued

6 Estimating Flood Magnitude and Frequency on Streams in Alaska and Conterminous Basins in Canada



INSET B, South-east Alaska





EXPLANATION	
	Flood frequency study area
Streamgage and map identification No.	
59 	Used in regional skew and regression
62 	Used in regional regression
24 	Not used in regional skew or regression

Figure 1.—Continued

Rainfall-related flooding is most common in autumn in most of the study area, or in autumn and winter in basins bordering the Gulf of Alaska, where winter precipitation can fall as rain instead of snow or as rain on snow. Atmospheric rivers, or narrow corridors of moisture in the lower atmosphere (Ralph and Dettinger, 2011), often form over the Pacific Ocean and flow towards the Pacific coast of North America and can bring heavy rainfall that generates notable regional flooding in Alaska.

Ice-jam flooding, which occurs when large blocks of winter stream ice cover break free during spring melt and partially dam streams, can create local high water levels from backwater generally not associated with large magnitudes of discharge. Although these floods generally do not constitute the maximum discharge of the year and are therefore not considered in this statistical study, ice-jam flooding is an important consideration for public safety.

Notable non ice-jam floods since 1949, when widespread streamgaging began in Alaska, include floods in Interior Alaska in 1964, near Fairbanks in 1967 (Childers and others, 1972), in south-central Alaska in 1971 (Lamke, 1972) and 1986 (Lamke and Bigelow, 1988), along the Copper River in 1981 (Brabets, 1997), along the middle Koyukuk River in 1994 (Meyer, 1995), and in south-central Alaska and the Kenai Peninsula in 1995 (Stauffer, 2010; National Weather Service Alaska-Pacific River Forecast Center, 2015), 2006 (Joling, 2006; Stauffer, 2010), and 2012 (Homer Tribune, 2012), and on the Kenai Peninsula in 2002 (Eash and Rickman, 2004). A list of maximum known floods, which can include floods other than those considered for flood frequency analysis, for selected stations in Alaska and conterminous basins of Canada was presented by Jones and Fahl (1994) and was not updated for this study.

Data Compilation

Peak-flow records from the USGS and from the Water Survey of Canada, the branch of Environment Canada that is responsible for streamflow monitoring for Canada, were screened for suitability for flood frequency analysis and eligibility for regression analysis. Selection considerations for flood frequency analysis included record length and the effect of any streamflow regulation or diversion, urbanization, or natural damming and release of water on peak flow. Peak-flow data for the selected sites were then reviewed to assure the quality of the records and tested for homogeneity or presence of trends over time, which could invalidate the assumptions of the analyses. For use in development of regression equations, basin characteristics were determined for all streamgages in the study.

Peak-Flow Data

Records considered for analysis in this study consisted of annual series of peak-flow data at least 10 years in length from generally unregulated and non-urbanized USGS streamgages in Alaska or Water Survey of Canada streamgages in conterminous basins in Canada. The streamgages considered included 387 streamgages in Alaska and Canada that were selected for analysis (table 1). Alaskan peak-flow data were obtained from the USGS National Water Information System (NWIS) peak-flow file (U.S. Geological Survey, 2015) and Canadian peak-flow data were obtained from the HYDAT database via the Environment Canada Data Explorer (Environment Canada, 2015). Streamgages used in this study included continuous-record streamgages that documented streamflow on a daily basis and crest-stage partial-record streamgages that documented only the annual peak flow. No seasonal partial-record streamgages, which operate for only part of a year, were used in this study. Regardless of the type of streamgage, the annual peak flow consisted of the maximum instantaneous discharge for the water year (October 1 through September 30) and maximum daily discharge for selected streamgages, when the maximum instantaneous discharge was not available.

Table 1. Description of streamgages used in flood frequency analysis and considered for use in regional skew and regional regression analysis for Alaska and conterminous basins in Canada.

[Table 1 is a Microsoft® Excel file and can be downloaded at <http://dx.doi.org/10.3133/sir20165024>]

All analyses used data compiled by water year. Although USGS data are published on a water-year basis, Canadian data are published on a calendar-year basis. Conversion of Canadian data to a water-year basis occasionally encountered two calendar-year peaks in a single water year, requiring omission of the smaller peak and an accompanying reduction of the available record length. The water years for peaks omitted from the analysis during the conversion of Canadian data are shown in table 1. For years when maximum instantaneous discharge was not available for part of the Canadian record but annual maximum daily mean discharge was available, an assessment of comparability was made using years when these two measures were available. For records where maximum daily mean discharge was within 5–10 percent of the maximum instantaneous peak discharge for other years in the record, the annual maximum daily mean discharge was used as a surrogate for maximum instantaneous peak discharge. This bias toward smaller discharge for selected stations is expected to be minor relative to other errors in the analysis. The Canadian records as modified for this study are available in the NWIS peak-flow file (U.S. Geological Survey, 2015).

8 Estimating Flood Magnitude and Frequency on Streams in Alaska and Conterminous Basins in Canada

The Alaska peak-flow records for streamgages in this study were inspected for anomalous values, and qualification codes associated with the peaks were reviewed for accuracy. Revisions were applied in NWIS prior to a final extraction of streamflow data for this study. Peak flows for which a discharge qualification code was published in the peak-flow file were included or omitted in the flood frequency analysis according to default procedures described for USGS flood frequency analysis program PeakFQ (Flynn and others, 2006) except for selected cases described here. Published USGS records containing a large number of instances of use of the annual maximum daily mean discharge (peak-flow discharge code 1) as a surrogate for the maximum instantaneous peak discharge were reviewed to ensure that daily and peak values were similar to each other for years when both were available. Where the daily and peak values differed by more than 5–10 percent, the annual maximum daily mean discharges were omitted from the record for analysis, as noted in [table 1](#). Regulated or diverted sites (peak-flow discharge codes 5 or 6 in the peak-flow file) were reviewed individually to determine the effect of the unknown (code 5) or known (code 6) degree of regulation or diversion on peak flows and whether the peak-flow data were suitable for flood frequency analysis. Similarly, sites having a peak-flow discharge code 3, when the basin was known to be subject to glacier outburst floods, were reviewed individually and included in flood frequency analysis when appropriate. The water year and condition for all instances where a published peak was omitted from flood frequency analysis is shown in [table 1](#).

Trend Analysis

Records used in flood frequency analysis are assumed to contain a series of independent discharges free of trends from changes in basin or climate conditions. Kendall's *tau*, a statistic in a nonparametric test for whether values increase or decrease over time, can test the strength and significance of trends in streamflow data (Helsel and Hirsch, 2002). For the purposes of determining suitability for flood frequency analysis, the Kendall *tau* test was applied to each record in the study, regardless of length and regardless of long gaps in the record. Statistically significant trends in peak flow can occur for many reasons, but a trend indicating a changing basin condition, such as a change in land cover or urbanization, can violate the assumptions of flood frequency analysis. When the condition creating the trend is understood, it might be appropriate to adjust the record for trend and estimate flood frequency using the adjusted record. Detection of trends in peak flow across a region or the entire State poses a different question that would require selection of a suitable subset of these records to evaluate and is beyond the scope of this report.

Kendall's *tau* depends on the rank rather than the magnitude of the values, making it effective for identifying trends in streamflow where extreme values and skewness (skew) are present. The test is conducted by comparing the rank of each peak-flow value to the rank of all other values in the record. A positive *tau* value indicates that the number of pairs increasing in value exceeds the number of pairs decreasing in value, and a negative *tau* value indicates the opposite. *Tau* ranges from -1 to 1 and approaches 0 when no trend is likely to exist. For this study, a trend was considered to be significant if the probability (*p*) value, or probability that a true null hypothesis of no trend is erroneously rejected, was less than or equal to 0.05.

Statistically significant trends were detected at 43 of the 387 streamgages in the study ([table 1](#)). Of the streamgages with significant trends, 22 streamgages showed upward trends, and 21 streamgages showed downward trends. No underlying cause of any trend was obvious when considering spatial distribution, regulation, land-use changes, and urbanization. Although a cursory consideration of climate as a variable in peak-flow trends suggested no obvious patterns, a thorough assessment of any correlation of significant peak-flow trends at individual sites to temporal changes in climate was beyond the scope of this report. Because adjustments to compensate flood frequency estimates for trends or exclude records showing trends from analysis might result in a spurious correction when the underlying cause is not well understood, all records having statistically significant trends were included in the analysis without adjustment.

Physical and Climatic Basin Characteristics

Peak flow in unregulated basins generally varies as a function of basin size and the physical and climatic characteristics of the basin that govern the quantity and rate of water delivery to the stream. Basin characteristics that can be correlated to peak flow in gaged basins provide convenient explanatory variables for estimating peak flows in ungaged basins. Basin characteristics tested in the regional skew and regional regression analyses as potential explanatory variables are listed in [table 2](#). Values for basin characteristics for streamgages used in the analyses or presented in Cook Inlet Basin StreamStats are shown in [appendix A](#) and in a geodatabase (.zip file can be downloaded at <http://dx.doi.org/10.3133/sir20165024>). The basin characteristics included topographic variables; land cover characteristics including various vegetation cover categories and glacier and permafrost cover; hydrologic characteristics such as coverage by lakes; and climate variables summarizing temperature and precipitation.

Values for all basin characteristics were updated from the most recent geospatial datasets available as of 2012. For topographic, climatic, and glacier coverage variables, suitable datasets were available that spanned both the United States and Canadian parts of the study area.

Table 2. Basin characteristics considered in regional skew analysis and peak flow regression analysis for Alaska and conterminous basins in Canada.

[Data source: ASTER GDEM, Advanced Spaceborne Thermal Emission and Reflection Radiometer Global Digital Elevation Map; NHD, National Hydrography Dataset; NLCD, National Land Cover Dataset; NOAA, National Oceanic and Atmospheric Administration; PRISM, parameter-elevation regressions on independent slopes model; USGS, U.S. Geological Survey; WBD, Watershed Boundary Dataset. **Data source citation and link:** NALCMS, North American Land Change Monitoring System; NASA, National Aeronautics and Space Administration; NHD, National Hydrography Dataset; SCAS/OSU, Spatial Climate Analysis Service, Oregon State University. –, not applicable]

Abbreviation compatible with StreamStats version 3	Description	Unit	Data source	Data source citation and link	Geographic coverage for datasets not available for full study area
DRNAREA	Area that drains to a point on a stream	square miles	USGS 1:63,360 topographic maps and WBD (Alaska); Natural Resources Canada 1:50,000 topographic maps (Canada)	The National Map (http://nationalmap.gov) (Alaska); GeoGratis (http://www.nrcan.gc.ca/earth-sciences/geography/topographic-information) (Canada)	–
LAT_CENT	Latitude of the basin centroid	degrees	USGS 1:63,360 topographic maps and WBD (Alaska); Natural Resources Canada 1:50,000 topographic maps (Canada)	The National Map (http://nationalmap.gov) (Alaska); GeoGratis (http://www.nrcan.gc.ca/earth-sciences/geography/topographic-information) (Canada)	–
ELEV	Mean basin elevation	feet	ASTER GDEM Version 2 (30 m)	NASA (2011) (http://reverb.echo.nasa.gov/reverb/)	–
PRECPRIS00	Basin average mean annual precipitation for 1971–2000 from the PRISM climate dataset	inches	PRISM: 1971–2000 (Alaska), 1961–1990 (Canada)	Gibson (2009a) (Alaska) (https://irma.nps.gov/App/portal/Home); SCAS/OSU (2002) (Canada) (http://www.climate-source.com)	–
JANMINTMP	Mean minimum January basin temperature	degrees Fahrenheit	PRISM: 1971–2000 (Alaska), 1961–1990 (Canada)	Gibson (2009b) (Alaska) (https://irma.nps.gov/App/portal/Home); SCAS/OSU (2002) (Canada) (http://www.climate-source.com)	–
LAKEAREA	Percentage of basin covered by lakes and ponds	percent	NHD (Alaska)	NHD (http://nhd.usgs.gov)	Alaska
GLACIER	Percentage of basin covered by glaciers	percent	Randolph Glacier Inventory Version 2.0	Arendt and others (2012) (http://www.glims.org/RGI/index.html)	–
LC01FOREST	Percentage of basin covered by forest	percent	NLCD 2001, classes 41–43	Homer (2007) (http://www.mrlc.gov/nlcd2001.php)	Alaska
BROADLEAF ¹	Percentage of basin covered by broadleaf vegetation	percent	2005 Land Cover of North America, classes 1–6 and 15–20	NALCMS (2013) (http://www.cec.org/naatlases)	–
PERMAFROST_LOWUP ¹	Percentage of basin in upland and lowland areas covered by permafrost	percent	Ferrians (1998), classes Perm 21–25	Ferrians (1998) (http://nsidc.org/data/ggd320.html)	Alaska

Table 2. Basin characteristics considered in regional skew analysis and peak flow regression analysis for Alaska and conterminous basins in Canada.—Continued

Abbreviation compatible with StreamStats version 3	Description	Unit	Data source	Data source citation and link	Geographic coverage for datasets not available for full study area
	Alternate				
LC01WETLND	Percentage of basin covered by wetlands	percent	NLCD 2001, classes 90 and 95	Homer (2007) (http://www.mrlc.gov/nlcd2001.php)	Alaska
LC01SNOIC ¹	Percentage of basin covered by snow and ice	percent	NLCD 2001, class 12	Homer (2007) (http://www.mrlc.gov/nlcd2001.php)	Alaska
I24H2Y	24-hour, 2 year precipitation, or maximum 24-hour precipitation that occurs on average once in 2 years	inches	NOAA atlas	Perica and others (2012) (http://hdsc.nws.noaa.gov/hdsc/pfds/pfds_gis.html)	Alaska
LC01SCRUB ¹	Percentage of basin covered by shrub or scrub	percent	NLCD 2001, classes 51 and 52	Homer (2007) (http://www.mrlc.gov/nlcd2001.php)	Alaska
LC01BARE	Percentage of basin covered by barren land	percent	NLCD 2001, class 31	Homer (2007) (http://www.mrlc.gov/nlcd2001.php)	Alaska
PERMAFROST_MTN	Percentage of basin in mountainous areas covered by permafrost	percent	Ferrians (1998), classes Perm 11–13	Ferrians (1998) (http://nsidc.org/data/ggd320.html)	Alaska
NEEDLELEAF ¹	Percentage of basin covered by needleleaf vegetation	percent	2005 Land Cover of North America, classes 7–12 and 21–26	NALCMS (2013) (http://www.cec.org/naatlas)	–
MIXBROADNEEDLE ¹	Percentage of basin covered by mixed broadleaf and needleleaf vegetation	percent	2005 Land Cover of North America, classes 13, 14, 27, and 28	NALCMS (2013) (http://www.cec.org/naatlas)	–

¹Informal name; basin characteristic is not available in StreamStats.

Selected hydrographic and land-cover datasets available for Alaska that had no Canadian counterpart having comparable resolution or categorical structure are shown in table 2. Although variables obtained from these Alaska-only datasets could be used for analysis of Alaska regions, alternate variables, generally obtained from study-wide databases with lower resolution, were required for use in analysis of Canadian regions or the full study area.

A few potential basin characteristics were of insufficient quality for this study. Minor artifacts in the digital elevation model (DEM) used for topographic variables artificially inflated values for extreme measures of topography, preventing use of maximum elevation or relief. Inconsistent or low resolution digital topographic and hydrographic data across the study area precluded use of basin characteristics such as channel length and basin slope.

A drainage area boundary, or delineation of land area draining to a streamgage location, was used to clip digital basin characteristics datasets to obtain basin characteristics for each streamgage. For the Alaska sites in this study, drainage area boundaries previously prepared by the USGS Alaska Science Center (ASC) were synchronized with the Watershed Boundary Dataset (WBD) for Alaska (available within the USGS National Hydrography Dataset [NHD] at <http://nhd.usgs.gov/>). Where ASC boundaries could improve the WBD, edits were reconciled with the Alaska WBD steward. New drainage area boundaries required for this study were delineated from the streamgage location to the nearest WBD boundary on the basis of digital topographic maps, and then followed the WBD for the upstream part of the basin. Final drainage area boundaries matched the WBD at the time of final analysis; however, the WBD is continually edited and future revisions will not be reflected in the study data. For the Canadian sites in this study, draft drainage area boundaries provided by Environment Canada (Judy Kwan, Meteorological Service of Canada, written commun., 2013) or from previous USGS studies were reviewed for general agreement with digital topographic maps and the WBD (for areas in Alaska). Because the Canadian drainage area boundaries were not being prepared for publication with this study, edits were made only where desired changes appeared to potentially affect drainage area by about 5 percent or more.

Drainage area and basin centroid latitude and longitude were computed in a Geographic Information System (GIS) from the drainage area boundary alone. Drainage area computed for this study was used to update the value published in NWIS for the streamgage. For all other basin characteristics, summary statistics computed in a GIS for the respective clipped digital datasets produced the value of the basin characteristic.

Flood Magnitude and Frequency at Gaged Sites

The frequency analysis of annual peak-flow data collected at a streamgage provides an estimate of the flood magnitude and frequency for that particular stream site. Previously, flood-frequency estimates commonly were described as the “T-year” floods based on the recurrence interval for the flood statistics (for example, the “100-year flood”). The use of the recurrence-interval terminology is shifting to annual-exceedance-probability terminology to avoid the common misinterpretation of relating recurrence interval to a set length of time between floods of a particular magnitude (Holmes and Dinicola, 2010). Flood frequency estimates relate the probability of a flood of a given magnitude being equaled or exceeded in any given year. For example, a 1-percent AEP (formerly known as the “100-year” flood) corresponds to the flow magnitude that has a 0.01 (1/T-year recurrence interval) probability or 1-percent chance of being equaled or exceeded in any given year. The P-percent AEPs and the corresponding T-year recurrence intervals for flood frequency estimates reported in this study are shown in table 3.

Flood-frequency estimates in this report were computed using the USGS program PeakFQ, version 7.1 (Veilleux and others, 2014), which performs statistical flood-frequency analyses of annual peak flows following procedures recommended in Bulletin 17B of the Interagency Advisory Committee on Water Data (1982) and modified by the Advisory Committee on Water Information, Subcommittee on Hydrology, Hydrologic Frequency Analysis Work Group (HFAWG) (http://acwi.gov/hydrology/Frequency/b17_swfaq/EMAFQA.html).

Table 3. P-percent annual exceedance probabilities and corresponding T-year recurrence intervals for flood frequency flow estimates.

P-percent annual exceedance probability	T-year recurrence interval
50	2
20	5
10	10
4	25
2	50
1	100
0.5	200
0.2	500

The PeakFQ program and documentation are available at <http://water.usgs.gov/software/PeakFQ/>. The resulting flood frequency statistics for streamgages in Alaska and conterminous basins in Canada with at least 10 years of record through water year 2012 are shown in table 4.

Table 4. Flood frequency statistics for streamgages in Alaska and conterminous basins in Canada with at least 10 years of record through water year 2012.

[Table 4 is a Microsoft® Excel file and can be downloaded at <http://dx.doi.org/10.3133/sir20165024>]

Log-Pearson Type III Frequency Analysis

Flood-frequency estimates for streamgages are computed by fitting the base-10 logarithms of the series of annual peak flows to a known statistical distribution. The flood magnitude and frequency estimates for this study were computed using the log-Pearson Type III (LP3) distribution as recommended in Bulletin 17B (Interagency Advisory Committee on Water Data, 1982). The fitting of this distribution requires calculating the three statistics—the mean, standard deviation, and skew of the logs of annual peak flows, which describe the midpoint, slope, and curvature of the peak-flow frequency curve, respectively. The estimates of the P -percent AEP flows are computed by inserting the three statistics into the following equation:

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

where

- Q_p is the P -percent AEP flow, in cubic feet per second;
- \bar{X} is the mean of the logarithms of annual peak flow;
- K_p is a factor based on the skew coefficient and the AEP as obtained from Bulletin 17B (Interagency Advisory Committee on Water Data, 1982, appendix 3); and
- S is the standard deviation of the logarithms of annual peak flow.

The skew coefficient is reflected in the curvature of the flood-frequency curve. A positively skewed distribution curves convexly such that floods are unbounded. A negatively skewed distribution has a concave curve such that floods have some upper bound. The skew coefficient can be estimated from the series of annual peak flows (considered the sample data), but tends to be an unreliable estimator of the population for streamgages with short periods of record. Guidelines from Bulletin 17B of the Interagency Advisory Committee on Water Data (1982) recommend weighting the station skew

with a generalized, or regional, skew to improve the accuracy of the station skew estimator. The regional skew is assumed to be an unbiased and independent estimate, which improves the uncertainty of the skew estimate when weighted with the station skew. The station skew coefficients for streamgages within new regional skew areas, which are discussed in section, “[Statistical Analysis of Regional Skew](#),” were weighted with the updated regional skew developed for this study when computing the LP3 distribution.

Expected Moments Algorithm

The 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent AEP flows for each streamgage in the study (table 4) were computed using the Expected Moments Algorithm (EMA), as recommended by HFAWG (http://acwi.gov/hydrology/Frequency/b17_swfaq/EMAFaq.html). The EMA method applies the LP3 distribution for estimating streamflow frequency recommended in Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) but more efficiently incorporates censored flows and historical flows (Cohn and others, 1997, 2001). Censored data includes flows that are below or above a monitoring threshold, potentially influential low floods (PILFs, formerly referred to as low outliers in Bulletin 17B), zero flows, and uncertain (non-point discharge) observations; historical data include non-exceedance flood information (Gotvald and others, 2012). For sites that have no PILFs, no historical information, and no other censored information, flood frequency estimates are identical to those values obtained when using the conventional method of moments described in Bulletin 17B.

An EMA flood frequency analysis requires two types of information to describe the flows for every year in the period of record—(1) flow intervals that describe the magnitude of the annual peak flow and (2) measurement capabilities of the streamgaging techniques, known as perception thresholds. Flow intervals are defined by a lower bound, Q_L , and an upper bound, Q_U . Perception thresholds are defined by the minimum flow to the maximum flow that a streamgage could record, T_L to T_U , respectively. A recorded point discharge value of 300 ft³/s from a standard continuous streamgage, for example, is described in EMA by a flow interval from 300 ft³/s (Q_L) to 300 ft³/s (Q_U), with perception thresholds set from 0 ft³/s (T_L) to infinity (T_U).

Some streamflow records span a longer historical period than just the peaks documented as part of a regular streamgaging program, or systematic peaks. Additional (historic) peaks and associated historical information can be used to extend the period of record beyond the period of systematic record or fill in gaps in the period of systematic record. Because EMA uses a more general description of flood information, all types of flood data can be incorporated in the analysis.

Systematic Record and Nonstandard Censored Flow

Systematic peaks are those documented routinely, regardless of magnitude, as part of a regular streamgaging program. Like most USGS science centers, the USGS Alaska Science Center (ASC) streamgaging program has focused on the routine collection of streamflow data at continuous and crest-stage streamgages since it began widespread streamgaging in 1949. The resulting systematic records include many intermittent and short records because of fluctuations in funding, and because of streamgaging strategies that balanced funding long-term sites against funding many short-term sites in order to bolster Alaska's relatively sparse streamgaging network. The Water Survey of Canada records used in this study consisted only of standard systematic streamflow data, which could be continuous or intermittent.

Within the USGS systematic streamflow record, data can include nonstandard, censored flow data that describes a recorded interval. For example, a crest-stage gage (CSG) partial-record site records the highest water level above a minimum recordable elevation (gage base) since the station was last visited. In some cases, an annual peak flow is known to not have exceeded the gage base at the CSG. This below-gage-base peak flow is properly described in EMA as a flow interval with $(Q_L) = 0 \text{ ft}^3/\text{s}$ and $Q_U = Q_{\text{Gage base}}$ and perception thresholds from $T_{Q_{\text{Gage base}}}$ to infinity. In some cases, the opposite censored flow conditions could arise. For example, it might be known that an annual peak flow overtopped a streamgage and left no visible high water marks to estimate the magnitude of that flow event. The user can set a flow interval in EMA from the maximum value recordable at the gage, $Q_{\text{Gage max}}$, (Q_L) to infinity (Q_U) , respectively, with the perception thresholds set from $0 \text{ ft}^3/\text{s}$ to $T_{Q_{\text{Gage max}}}$. EMA allows the information from these nonstandard, censored flow data to be applied in the analysis rather than ignored or crudely estimated as would be the case for a conventional analysis using Bulletin 17B of the Interagency Advisory Committee on Water Data (1982) such as Curran and others (2003).

Historical Flood Information

At many locations, flood plain observation over an extended period or preserved evidence of floods can be used to inform a longer historical period about large floods and about peak flows that were not recorded but were known to be below some value. The longer historical period consists of the envelope of all years for which streamflow information is known, including the systematic period of record, the years for which peak flows were collected outside the systematic record, and non-gaged years for which it was known no peak equaled or exceeded some value. Particularly large floods can be recorded within the systematic record or, in the case of USGS streamflow records, as annual peaks collected before, after, or in breaks in the systematic record. Collected

for some purpose other than documenting the annual peak regardless of its magnitude, the non-systematic peaks belong to a separate, biased population. Because the occurrence of particularly large floods led to the collection of many of these non-systematic peaks, they have been conventionally referred to as historic peaks in the USGS peak-flow file. The term historic can be misleading because it infers that the peak must be particularly large if collected outside the systematic record; in fact, as found in many Alaska streamgage records, peak flow data collected outside the systematic streamgaging record can be any magnitude. For example, as part of routine collection of streamflow data associated with major regional rainstorms, ASC opportunistically recorded flood data at inactive (unfunded for regular operation) streamgages as requested by a funding agency, particularly if damage to infrastructure occurred. The streamgaging resulted in annual peaks collected outside the systematic record that may or may not have been particularly large, depending on factors such as infrastructure damage disproportionate to flood magnitude. Because the peak-flow file does not designate the motivation for collecting non-systematic peaks, all peaks coded as historic in the peak-flow file for this study were first reviewed to determine if the peak was large enough to potentially inform periods of missing record. Peaks coded as historic but omitted from analysis because they could not be established as part of the population of non-random, large peaks that could inform missing periods of record were maintained as part of the peak-flow file to provide important information for other studies. For records containing particularly large historic peaks or particularly large systematic peaks, historical information was then researched to ascertain, where possible, that the record contained the full population of peaks above some value in the longer historical period, and that non-recorded peak flows were known to not have equaled or exceeded that value. Historical flood information to support extending the historical period included published reports or historical accounts, observer interviews, high water marks, damage to infrastructure, and other marks or observations from the particular site in question.

Flow data for historic peaks used in the analysis was included in EMA as $Q_{\text{Historic}} (Q_L)$ to $Q_{\text{Historic}} (Q_U)$. Flow data for missing record in the longer historical period was represented as $0 \text{ ft}^3/\text{s} (Q_L)$ to Q_{Historic} or $Q_{\text{Large systematic}} (Q_U)$ for records having a historic peak or large systematic peak, respectively. The perception threshold was set from $T_{Q_{\text{Historic}}}$ or $T_{Q_{\text{Large systematic}}}$ (T_L) , as appropriate, to infinity (T_U) , for both the water year of the historic peak, if present, and the missing period in the longer historical period. As an example, [figure 2](#) shows the annual peak-flow series for streamgage 15208000 Tonsina River at Tonsina, Alaska, which included two peaks collected outside the systematic gaged record. The 7,000 ft^3/s water year 1995 peak, which was exceeded during the period of gaged record, was deemed too small to inform periods of missing record and thus was not part of a population appropriate for flood frequency analysis. The 14,000 ft^3/s water year 2006 peak was determined from highway damage reports and

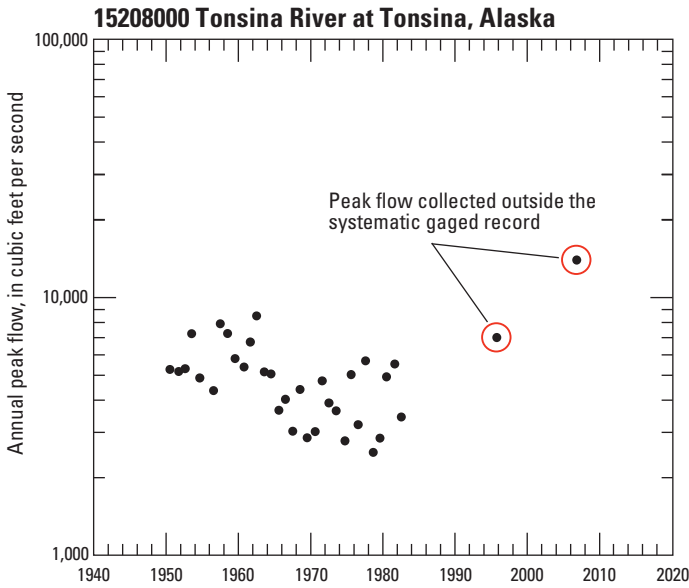


Figure 2. Example of annual peak-flow series containing peaks collected outside the systematic gaging record, U.S. Geological Survey streamgage 15208000, Tonsina River at Tonsina, Alaska.

reports from local residents to represent a non-exceedance threshold that implied that all floods greater than or equal to that magnitude would be known for the historical period. In this case, the historical period was extended forward to 2012, the latest water year for all data in the study, and perception thresholds for water years 1983 to 2012 were set as 14,000 (T_U) to infinity (T_V). The corresponding at-site flood frequency LP3 curve is shown in figure 3.

For each streamgage in the study, table 1 lists the water years for the period of record used, the historic period length (number of water years in the period of record used), the water years omitted from the frequency analysis, the number of peaks in the record (as reported in output from the PeakFQ software), the perception thresholds set for every year in the period of record (if different from the default 0 ft³/s-to-infinity for the systematic gaged record), and interval discharge ranges (Q_{INT}) for censored flows. The value in the “number of peaks in record” column is equal to the number of peaks in the NWIS peak-flow file minus the number of peaks listed in the table as omitted that fall outside the years listed in “period of record used.” The number of peaks that are actually used in the analysis equals the value given for “number of peaks in record” minus the number of peaks listed in the table as omitted that fall within the years listed in “period of record used.”

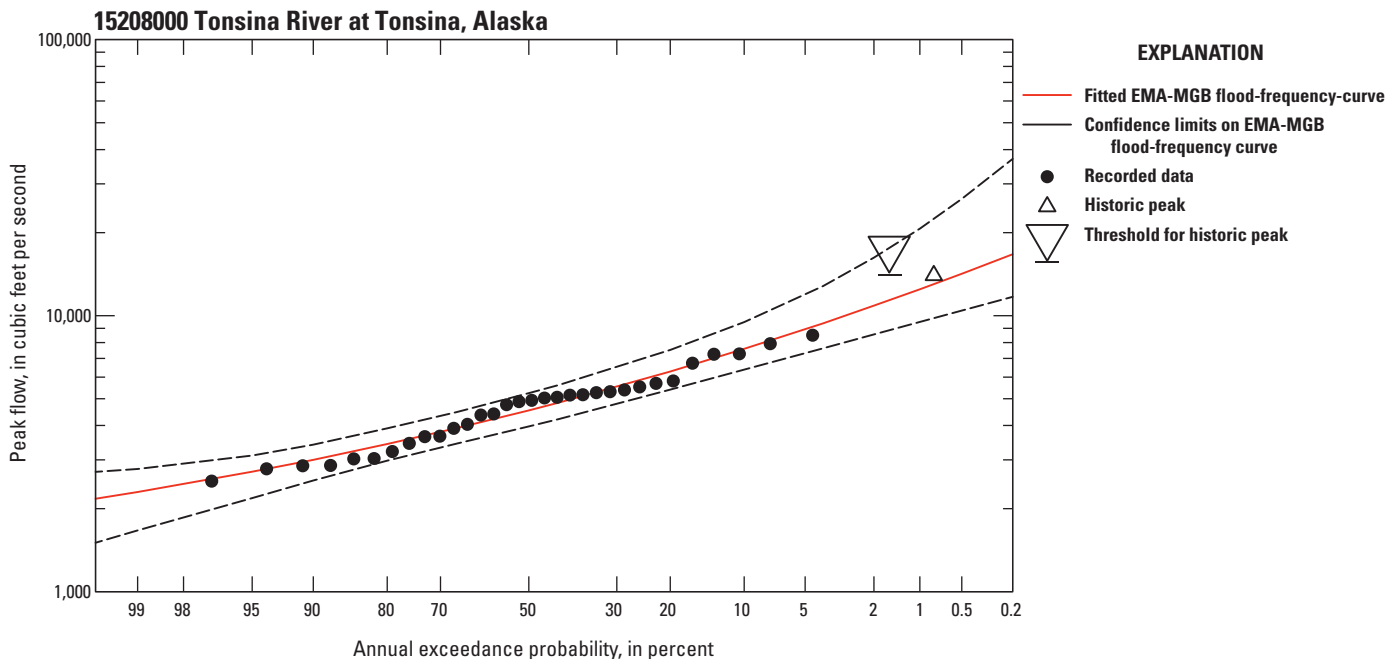


Figure 3. Example of a Log-Pearson Type III flood frequency curve for U.S. Geological Survey streamgage 15208000, Tonsina River at Tonsina, Alaska, showing the Expected Moments Algorithm (EMA) with a multiple Grubbs-Beck (MGB) test representation for two peaks collected outside the systematic gaging record, water years 1995 and 2006. The water year 1995 peak was omitted from the analysis on the basis of a lack of at-site information to establish the value of the peak as a non-exceedance threshold. At-site information established the value of the water year 2006 peak as a non-exceedance threshold for the missing period of record through the end of the study period.

Multiple Grubbs-Beck Test for Detecting Multiple Potentially Influential Low Floods

To prevent zero-flow and very small annual peaks from having undue influence on the fit of the distribution for large annual peaks, Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) recommends the use of the Grubbs-Beck test (Grubbs and Beck, 1972) to detect low outliers in flood series that deviate from the trend of the data. The Grubbs-Beck test provides an objective identification of low outliers but does not consider the condition of multiple outliers. As described by Cohn and others (2013), the multiple Grubbs-Beck (MGB) test is a generalization of the Grubbs-Beck method that allows for a standard procedure for identifying multiple PILFs. In flood-frequency analysis, PILFs are annual peaks that meet three criteria—(1) their magnitude is much smaller than the flood statistic of interest; (2) they occur below a statistically significant break in the flood-frequency data plot; and (3) they have excessive influence on the estimated frequency of large floods (Veilleux and others, 2014). PILFs may constitute one-half or more of the observations and are believed to arise from physical processes that are not relevant to the processes associated with

large floods. Because the magnitudes of PILFs are expected to explain little about the upper right-hand tail of the frequency distribution (largest floods), the best estimates for the typical suite of AEP flows can be obtained by censoring the PILFs (see Subcommittee on Hydrology, Hydrologic Frequency Analysis Work Group Frequently Asked Questions at <http://acwi.gov/hydrology/Frequency/B17bFAQ.html#low>). When an observation is identified as a PILF, all values smaller than that flood magnitude also are categorized as PILFs in the EMA analysis. Identifying PILFs and recoding them and all smaller peaks as censored peaks can greatly improve estimator robustness with little or no loss of efficiency. Thus, the use of the MGB test can improve the fit of the small annual exceedance probabilities (right hand tail of the distribution), while minimizing lack-of-fit due to unimportant PILFs in an annual peak series (Cohn and others, 2013). An example of a flood-frequency curve for a streamgage with multiple PILFs is shown in figure 4. Sites for which the default MGB test was overridden by a visual inspection, the PILF threshold below which all flows are recoded as censored values in the EMA analysis, and the corresponding number of PILFs that were censored are shown in table 1.

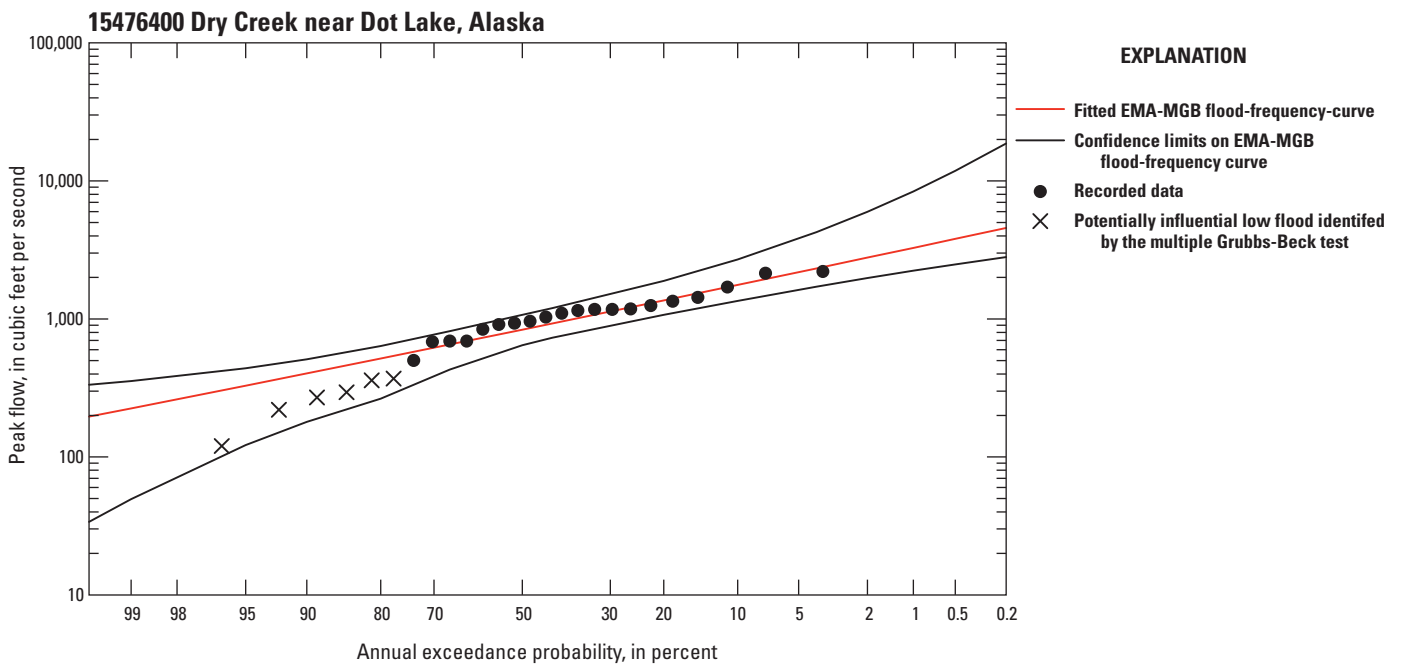


Figure 4. Log-Pearson Type III fit when multiple Grubbs-Beck (MGB) test found multiple potentially influential low floods (PILFs) for U.S. Geological Survey streamgage 15476400, Dry Creek near Dot Lake, Alaska.

Statistical Analysis of Regional Skew

Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) guidelines recommend using a weighted average of the station skew and a regional skew to improve the accuracy of the skew estimator used to estimate the AEP discharges. As presented in Bulletin 17B, this is computed from the station and regional skew weighted in inverse proportion to their mean square errors (MSE):

$$G_w = \frac{MSE_{G_r} (G_s) + MSE_{G_s} (G_r)}{MSE_{G_r} + MSE_{G_s}} \quad (2)$$

where

- G_w is the weighted skew;
- G_s is the station skew; and
- G_r is the regional skew; and
- MSE_{G_r} and MSE_{G_s} are the mean square error of the regional and station skew, respectively.

Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) supplies a national map of regional skew but also encourages hydrologists to develop more specific

local relations. Curran and others (2003) present regional skew coefficients for various streamflow analysis regions across the study area determined as an average of the station skew for long-record streamgages. More recently, Reis and others (2005), Gruber and others (2007), and Gruber and Stedinger (2008) developed a Bayesian generalized least-squares (GLS) regression model for regional skew analyses. The Bayesian methodology allows for the computation of a posterior distribution of both the regression parameters and the model error variance. Due to complications introduced by the use of the EMA with MGB censoring of low outliers (Cohn and others, 1997) and large cross-correlations between annual peak discharges at pairs of streamgages, the Bayesian weighted least-squares/Bayesian generalized least-squares (B-WLS/B-GLS) regression framework was developed to provide both stable and defensible results for regional skew (Veilleux, 2011; Lamontange and others, 2012; Veilleux and others, 2012).

For this study, the B-WLS/B-GLS analysis was applied only to areas where gaging density provided an adequate dataset of long-record streamgages. The regional skew analysis used streamgages with at least 25 years of pseudo record length, a measure computed from the record length and mean square error of the historical and systematic parts of the record. Two areas of Alaska contained a density of streamgages meeting these requirements (fig. 5).

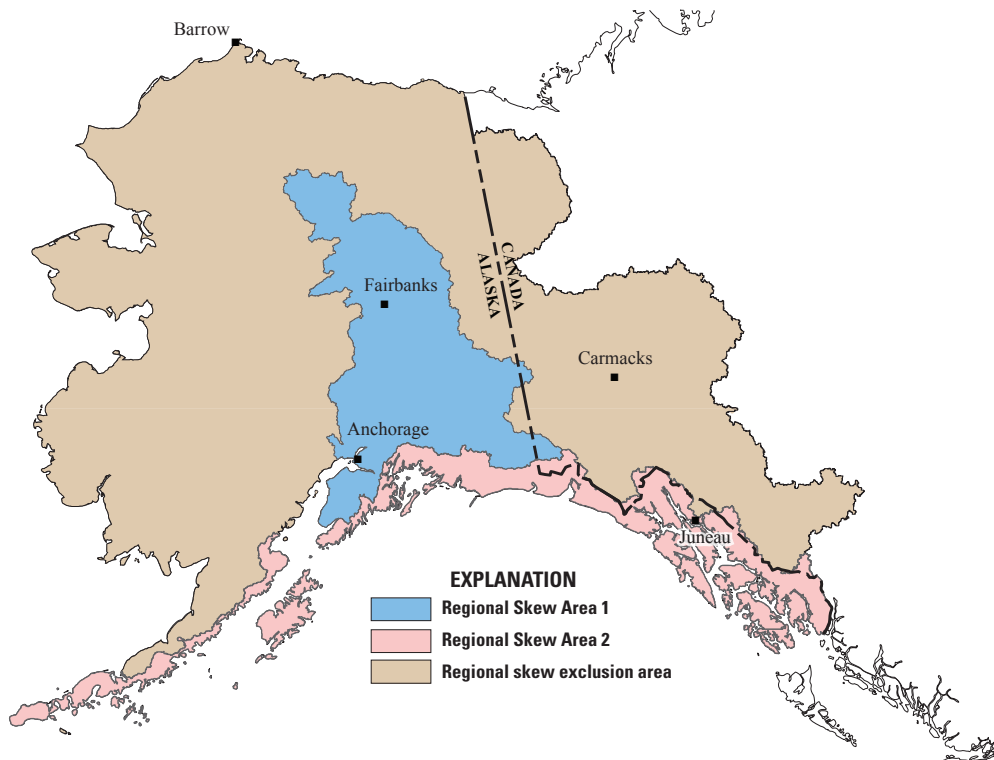


Figure 5. Regional skew areas for Alaska and conterminous basins in Canada.

Regional Skew Area (RSA) 1 in Interior Alaska contains the eastern part of Curran and others (2003) Streamflow Analysis Regions 4 and 6. The boundary for RSA 1 was delineated from the WBD as the hydrologic unit code-8s (HUC-8s) containing sites used in regional skew analysis, plus adjacent HUC-8s needed to make a continuous area. In several places, the HUC-8s were split to omit non-applicable areas. Regional Skew Area 2 contains 2003 Streamflow Analysis Regions 1 and 3 and was modified from Curran and others (2003) to match current HUC-8 boundaries. The HUC-8s, and HUC-10s where partial HUC-8s were used, contained in each RSA, are shown in table 5. The areas outside of the RSAs can be

considered regional skew exclusion areas, where low gage density limited the ability to establish regional patterns and only station skew applies to flood frequency analysis.

For a representative assessment of regional skew, sites used in the analysis should have reasonably similar basin characteristics. For RSA1, sites on the Yukon River, which are considerably larger in drainage area than any other sites, were excluded. Regional skew should only be applied to weighting with station skew for sites physically located within a regional skew area and within the range of drainage area for the sites that were used for analysis (table 6).

Table 5. Definition of regional skew areas for Alaska and conterminous basins in Canada, by hydrologic unit code (HUC).

HUC-8		HUC-10s within the selected HUC-8 that are included in Regional Skew Area		HUC-8		HUC-10s within the selected HUC-8 that are included in Regional Skew Area	
		Regional Skew Area 1				Regional Skew Area 2	
19020101	All			19010102	All		
19020102	All			19010103	All		
19020103	All			19010104	All		
19020104	1902010401, 1902010402, 1902010403, 1902010404			19010105	All		
19020301	1902030101, 1902030102, 1902030103, 1902030104, 1902030105, 1902030106, 1902030107, 1902030108, 1902030109, 1902030110			19010106	1901010601, 1901010602, 1901010603		
19020302	All			19010107	1901010701		
19020401	All			19010204	All		
19020402	All			19010206	All		
19020501	All			19010207	1901020701, 1901020702, 1901020703		
19020502	All			19010208	All		
19020503	All			19010209	All		
19020505	All			19010210	All		
19080301	All			19010211	All		
19080302	All			19010212	All		
19080303	All			19010301	All		
19080304	All			19010302	All		
19080305	All			19010303	All		
19080306	All			19010304	1901030404		
19080307	All			19010402	All		
19080308	All			19010404	1901040402, 1901040403, 1901040404, 1901040405		
19080309	All			19010405	All		
19080401	All			19010406	All		
19080402	All			19020104	1902010405, 1902010406, 1902010407, 1902010408, 1902010409, 1902010410, 1902010411, 1902010412, 1902010413, 1902010414, 1902010415, 1902010416, 1902010417, 19020104218		
19080403	All						
19080404	All			19020201	All		
19090101	All			19020202	All		
19090102	All			19020203	All		
				19020301	1902030111, 1902030112		
				19020602	1902060208, 1902060209, 1902060210, 1902060211, 1902060212, 1902060213		
				19020701	All		
				19020702	All		
				19030101	All		
				19030102	1903010201, 1903010202, 1903010203, 1903010204, 1903010206		

Table 6. Regional skew and summary statistics for regions in Alaska and conterminous basins in Canada.[Abbreviation: mi², square mile]

Location	Number of stations with at least 25 non-censored peaks	Regional skew	Average variance of prediction (for a site not in the analysis)	Standard error of the regional skew	Applicable range of drainage area (mi ²)
¹ Regional Skew Area 1	75	0.54	0.45	0.67	1.2–25,560
¹ Regional Skew Area 2	28	0.18	0.12	0.34	1.7–123
Sites outside regional skew areas, and sites in regional skew areas but not within the applicable range of drainage area		No regional skew available. Use station skew.			

¹Where basin drainage area is within applicable range of drainage area.

The updated regional skew values and associated statistics and performance metrics for RSA1 and RSA2 are summarized in [table 6](#). For both RSAs, no basin characteristics were found that improved model precision substantially and warranted the added model complexity, resulting in the adoption of a constant model for regional skew. The average variance of prediction (AVP) for a site not in the analysis can be substituted for MSE_G in equation 2. The AVP for RSA1 and RSA2 corresponded to an effective record length of 22 and 59 years, respectively. The new regional skews and their associated standard errors were used in the flood frequency analysis for all sites in the study that were located in the new RSAs and met drainage-area requirements. Additional details of the methods and results of the regional skew analysis are provided in [appendix B](#).

Estimating Flood Magnitude and Frequency at Ungaged Sites

A regional regression analysis was used to develop a set of equations for estimating the magnitude and frequency of floods for ungaged sites in Alaska and conterminous basins in Canada. These equations relate the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent AEP flows computed from peak-flow records for streamgages ([table 4](#)) to measured basin characteristics of the associated drainage basins. After elimination of streamgages ineligible for regression because of the occurrence of annual glacier outburst floods, the lack of a defined basin, or a close similarity to other streamgages in the same basin, 341 streamgages (284 in Alaska and 57 in Canada) were eligible for regional regression. The regression process consisted of exploratory analysis to determine the most suitable explanatory variables and their transformations, consideration of whether the equations could be improved by grouping sites into regions, and the use of more robust regression methods to develop the final equations and the uncertainty of their estimates.

Elimination of Redundant and Other Non-Eligible Sites from Regression Analysis

Regional regression analysis requires that streamflow data for all sites be correlated to basin characteristics in a comparable manner and that streamflow at a site be reasonably independent of streamflow at other sites. Review of streamflow records for these criteria resulted in exclusion of a number of sites eligible for flood frequency analysis but considered ineligible for the purpose of developing predictive regression equations from basin characteristics. These include sites having non-eligible streamflow populations or basin conditions, sites on the Yukon River with very large drainage areas, and sites considered hydrologically redundant.

Four streamgages included in this study had streamflow records that were suitable for fitting to a frequency distribution but included peaks controlled by glacier outbursts (releases of water from glacier-controlled storage), when those outbursts occur annually. In certain glacierized basins, annual glacier outburst peaks are routinely larger in magnitude than annual peaks from similarly sized basins not subject to outbursts. These large glacier outbursts are considered a separate population from snowmelt, rainfall, or glacier-affected melt peaks and cannot be used to develop regression equations for non-outburst floods. Peak qualification codes in the NWIS peak-flow file provided an initial screening tool for identifying basins subject to outbursts. Indeterminate drainage areas, which result in indeterminate relations to basin characteristics, included one site (USGS streamgage 15485500) where a portion of the stream bypassed the streamgage and one discontinued site (USGS streamgage 15201900) having ambiguous documentation of streamgage location. On the Yukon River, drainage areas for sites at and downstream of Carmacks, Yukon, are larger than any other gaged locations in Alaska. The 10 Yukon River streamgages at and downstream of Carmacks (drainage areas greater than 31,100 mi²) were excluded from the development of regression equations and considered separately in section, “[Estimate for a Site on the Yukon River.](#)”

Flood frequency estimates for records excluded from regression analysis on the basis of non-eligible streamflow populations or basin conditions did not use weighted skew and were not weighted with regression estimates. A usage designation that distinguishes these station-only sites from sites used in regional regressions is included in [table 1](#).

Nested basins having substantially similar size and basin characteristics are considered redundant for regression analysis. Including redundant basin pairs would unduly weight strongly related flows (Gruber and Stedinger, 2008). A screening tool developed for identifying potentially redundant basins (Veilleux, 2009) uses the drainage area ratio (DAR) between the basins and the normalized distance (ND) between the basin centroids. For this study, basin pairs with DAR less than 5 and ND less than 0.5 were considered redundant unless a visual inspection noted that they were not nested or had substantially different characteristics likely to affect streamflow, such as a large lake in one basin. Within each redundant pair, the site with the longest record generally was retained for use in regressions. Where record lengths were similar, one site was selected randomly for retention. For redundant sites in Alaska, station skew was weighted with regional skew as appropriate, but the resulting flood frequency estimates were not used to develop the regression equations. Final flood frequency estimates for redundant sites in Alaska were prepared from the station estimate weighted with the regression estimate as for non-redundant sites. The usage column in [table 1](#) shows the 30 sites included in the flood frequency analysis but omitted from development of regression equations as redundant. Sites located in Canada that were initially considered for use but omitted from the study as redundant are not shown in the table; these were the streamgages with USGS site identification Nos. 15024098, 15024300, 15024640, 15024695, 15120730, 15305030, 15305100, 15305260, 15305400, 15305406, 15305418, 15305520, 15305582, 15305590, 15305620, 15355000, and 15388944.

Exploratory Regression Analysis

Multiple-linear regression analysis is commonly used to develop equations that relate two or more physical or climatic basin characteristics to a specific streamflow statistic. After developing these empirical equations for an area from eligible streamgages and a particular basin characteristics dataset, the equations can be applied to estimation of flow at ungaged sites in that area using basin characteristics for the ungaged site obtained from the same or comparable basin characteristics dataset. Multiple-linear regression models the relation between multiple independent, or explanatory, variables (basin characteristics, such as drainage area) and a single dependent variable (a streamflow statistic, such as $Q_{1\%}$) by fitting a linear equation to the data.

Ordinary least-squares (OLS) regression is a simple form of multiple-linear regression that assumes that the peak-flow values at streamgages are independent and that each streamflow record has similar variance, which is influenced by the length of the record. OLS is a useful technique for identifying the most statistically important explanatory variables, testing different combinations of variables, and determining the general form of the equations. R, a programming language and environment for statistical computing and graphics, and USGS functions in the R package USGSwsStats (renamed smwrStats in 2015) were used for OLS regression analysis for this study.

To achieve a linear relation and improve the spread of data between the explanatory variables and the dependent variable, many variables associated with streamflow analysis require a data transformation. A log transformation commonly improves the linearity of most streamflow statistics and many basin characteristics, and was applied to all variables in this study except mean basin elevation and mean minimum January temperature. The logarithm of zero or negative values is undefined, but datasets containing these values can be first transformed by the addition of a constant, then log transformed. For this study, all variables expressed as a percentage of basin area were transformed by adding a constant 1 percent to all values, and mean minimum January temperature was transformed by adding a constant 32 °F to all values.

Highly correlated explanatory variables create a condition called multicollinearity, which restricts the ability of regression analysis to evaluate the importance of the respective variables and increases the variance of the regression coefficient for the explanatory variables. Obvious multicollinearity is generated by redundant metrics for a similar basin property, such as the percentage of the basin covered by glaciers from the Randolph Glacier Inventory and the percentage of the basin covered by snow and ice from the National Land Cover Dataset dataset. For this study, redundant variables included many land cover variables, metrics for amount and intensity of precipitation, and monthly versus annual mean values of climate variables. For these cases, a preferred dataset was chosen for initial analysis on the basis of dataset quality, extent of coverage of the study area, and hydrologic judgment of the strength of the explanatory nature of the variable ([table 2](#)). Selected alternate variables to these primary variables ([table 2](#)) were tested in multiple regression analysis by first removing the primary variable from the multiple regression. Correlation coefficients and inspection of correlation plots helped eliminate other highly correlated explanatory variables, such as mean annual precipitation and mean minimum January temperature, or either of those variables coupled with latitude of the basin centroid, from consideration as potential variables together in a single equation.

Selection of the most suitable explanatory variables for all study sites together and for each of six streamflow analysis regions designated by Curran and others (2003) and shown in figure 6—Regions 1 and 3 together, and Regions 2, 4, 5, 6, and 7—was based on all-possible-subsets regressions. Explanatory variables selected for final models were statistically significant, improved the coefficient of determination (R^2), and minimized the standard error of the estimate. Other considerations included whether the sign and magnitude of the regression coefficient was hydrologically reasonable, the ease of computation of the explanatory variable, and the effect of adding the variable on the Mallows's C_p statistic and the variance inflation factor (VIF). For all study sites together and for all 2003 regions, logDRNAREA formed the strongest explanatory variable. The variable logPRECPRIS00, which was statistically significant, considerably improved the fit and error of the equation for all study sites together and for some, but not all, 2003 regions. For certain AEPs in some regions, a third variable, such as ELEV or logGLACIER+1, was statistically significant and did not appreciably increase the VIF, but also did not considerably improve the fit or error of the equation.

Regionalization of Flood-Frequency Estimates

The streamflow analysis regions from Curran and others (2003) (fig. 6) provided an initial suite of regions for testing geographic groupings of sites and were supplemented by (1) proposed combinations of the 2003 regions and (2) a single study-wide region. The 2003 analysis presented seven regions for various streamflow analyses, two of which (Regions 1 and 3) were combined for the purposes of peak-flow analysis. The 2003 regions followed general climate patterns and major hydrologic basins, including a southern coastal area (Regions 1 and 3 combined) where precipitation is relatively high and temperatures moderate, a band of regions to the north (Region 4, including south-central Alaska, and Region 2, encompassing mostly large streams in Canada that drain to southeast Alaska), regions splitting the Yukon River Basin (Region 5 in the mostly Canadian upper Yukon River Basin, and Region 6 in Interior Alaska), and a region spanning the colder and drier northern and northwestern part of the State (Region 7). Consideration of streamgauge density (average streamgages per square mile, spatial clustering of streamgages, and total streamgages in a region) was a primary factor in seeking

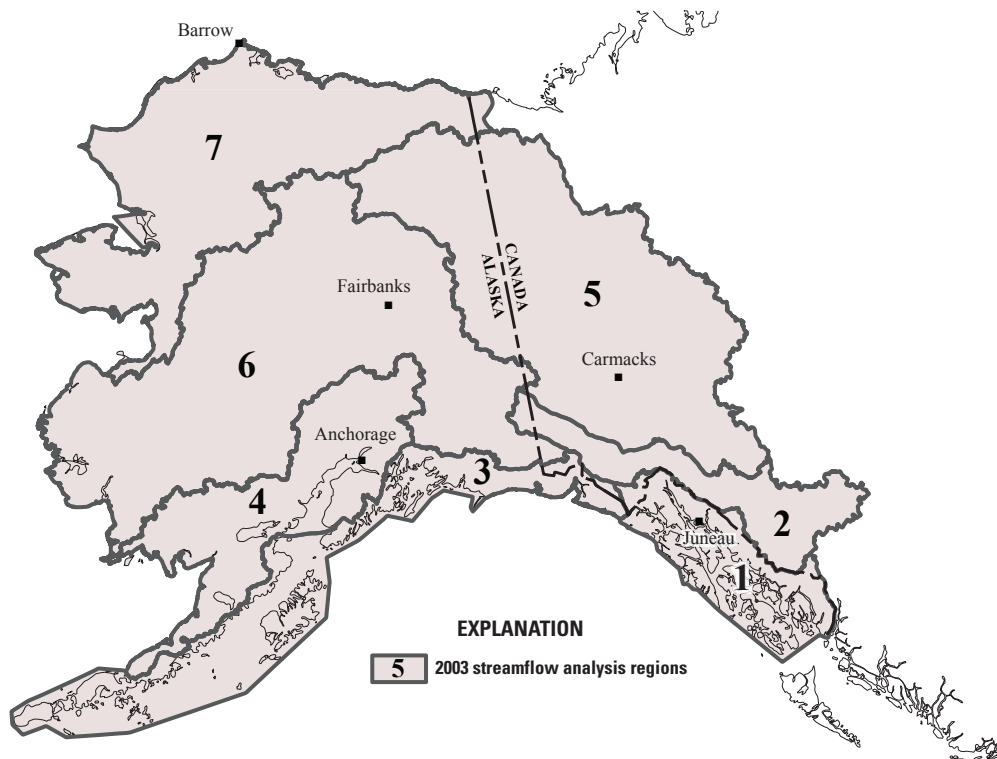


Figure 6. Streamflow analysis regions previously presented for Alaska and conterminous basins in Canada (from Curran and others, 2003, fig. 1).

alternatives to the 2003 regions. In particular, Regions 2 and 7 contained only 20 and 29 streamgages eligible for regression in this study, respectively, limiting the statistical strength of the equations and the representation of large land areas. Various combinations of Regions 4, 6, and 7, and of Regions 2 and 5, were considered, as was a Cook Inlet Basin region to match the boundary of the companion StreamStats project. A single study-wide region containing all streamgages eligible for regression analysis also was proposed to overcome the issues of streamgage density in selected areas of the study.

Evaluation of the 2003 regions, various combinations of the 2003 regions, and the study-wide region consisted of inspection of performance metrics for the best all-possible-subsets OLS regressions for each region followed by an analysis of covariance. The $Q_{1\%}$ and $Q_{50\%}$ were chosen as representative flow statistics for assessing potential regions. Results from a single-variable model using logDRNAREA or a two-variable model using logDRNAREA and logPRECPRI00, where both were statistically significant, showed that grouping sites by regions or combinations of regions produced a range of fit and associated error that straddled the fit and error produced by grouping all study sites together. Performance metrics were within generally acceptable limits for all regions and combinations tested.

Analysis of covariance (Helsel and Hirsch, 2002) using an indicator variable can be applied to test the statistical significance of the difference between any two regions. The indicator variable is set to 1 for all streamgages in a particular region and 0 for all other streamgages in the other region or regions being compared, and is then included in the regression as an explanatory variable. Statistical significance for the indicator variable indicates a difference in the regression intercept between streamgages in that region and the other streamgages in the test. An OLS regression was conducted using logDRNAREA and the indicator variable, and logDRNAREA, logPRECPRI00, and the variable, for the $Q_{1\%}$ flow for the 2003 regions and various alternate combinations of the 2003 regions. Although this analysis of covariance confirmed the statistical significance of the pairing of Regions 1 and 3, for example, the independence of many other regions varied depending on how regions were combined, suggesting ambiguity in the delineation of the 2003 region boundaries. The 2003 regions and various combinations of the 2003 regions, although not invalidated as suitable streamflow analysis regions, were not strongly supported by the analysis of variance. The comparable performance metrics of the group of all study sites together relative to the performance metrics of various combinations of the 2003 regions, coupled with ambiguous results of the analysis of covariance between regions, led to the selection of the entire study area except sites on the Yukon River at and downstream from Carmacks, Yukon as the single region for analysis for this study.

Use of a single study-wide region equivalent to the study area increased the statistical strength of the regression equations as compared to use of the 2003 regions in part by substantially increasing the number of sites used to develop the equation. Replacing multiple sets of equations in regions based largely on climate with a statistically stronger, single set of equations based in part on a climate variable provided comparable flow estimates having a more direct relation to basin characteristics and greater simplicity in application. The fit of a single set of equations for the study area suggests that this relation of streamflow statistics to basin characteristics is applicable over a wide range of basins, reducing, although not eliminating, the concern about the applicability of the regression equations over the large areas of the study area where no streamgages are present.

Regional Regression Equations

Streamflow data are naturally correlated spatially and temporally, making the assumptions of OLS regression incompletely satisfied. A more sophisticated technique, generalized least-squares (GLS) analysis, improves the equations by accounting for time-sampling error, which is a function of record length, and cross-correlation of annual peak flows between streamgages (Stedinger and Tasker, 1985). If two streamgages are in close proximity and flooding is caused by regional rainstorms or other basin climate conditions, the annual series of peak flow will be largely similar (correlated) at both streamgages and cannot be considered independent information for the purposes of the regression. GLS assigns different weights to each observation on the basis of its contribution to total variance. The USGS weighted-multiple-linear-regression (WREG) computer program, version 1.05 (Eng and others, 2009) incorporates GLS techniques and metrics developed by Stedinger and Tasker (1985), Tasker and Stedinger (1989), Martins and Stedinger (2002), Griffis and Stedinger (2007, 2009), and was used for GLS analysis of the final equations for this study.

Following the OLS regression analysis to develop a preferred multiple-linear regression model for relating each flood frequency statistic to basin characteristics and regionalization, which established the entire study area as the preferred regional structure, a GLS regression analysis was used to produce final equations. The resulting equations were evaluated using performance metrics and diagnostic tools in the WREG program. The significance of regression coefficients for each basin characteristic was checked to confirm the results of the OLS regression. Diagnostic plots of all residuals (difference of the predicted and observed values for a streamgage) against the predicted flows and against each explanatory variable in the equation were examined to ensure points were generally randomly distributed around zero, the assumed condition for linear regression (Helsel and Hirsch, 2002).

WREG calculates leverage and influence statistics for the GLS analysis (Eng and others, 2009), which serve as a regression diagnostic for the effects of an individual streamgage on the regional regression models. Leverage is a measure of how much the values of explanatory variables at a streamgage vary from the values of those variables at all other streamgages. Because unusual values for a streamgage, that is, values with high leverage, might or might not have a significant impact on the regression equation, the influence metric is computed to indicate how strongly the values for a streamgage influenced the estimated regression parameters. A streamgage can have high leverage, indicating that its independent variables are substantially different from those at all other streamgages, but not have a large influence on the fitted regression relation. Conversely, a streamgage with high influence might not have high leverage. Streamgages that had leverage or influence metrics that exceeded the thresholds calculated by WREG, especially those that had both high leverage and high influence, were evaluated for potential erroneous data reporting or conditions that would make the streamgage ineligible for regression. If no such errors existed, high leverage or influence metrics alone were insufficient justification for removing the streamgage from the regression analysis, and these streamgages were kept in the model.

The final regional regression equations for the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent AEP flows for a single region encompassing Alaska basins, except the Aleutian Islands and other islands off the western coast of Alaska (fig. 1) and sites on the Yukon River with very large basins, together with those major basins in Canada that drain to

Alaska are shown in table 7. The combination of drainage area and mean annual precipitation as explanatory variables produced the best fit and lowest error while minimizing the disadvantages of adding variables to the equation. The streamgage-specific values of drainage area and mean annual precipitation for the 341 streamgages used in the analysis are given in table 1, and ranges of these values are presented with the regional regression equations in table 7.

Sites along the Yukon River downstream of USGS streamgage 15305350 in Carmacks, Yukon have drainage areas that exceed that of all other rivers in the study. These large basins (drainage areas greater than 31,100 mi²) might not respond to basin characteristics as do other, smaller basins and were not included in the regression analysis. Regression analysis would not be well-suited for this small collection of streamgages along a single river. Instead, a method for simple drainage-area-based interpolation of flow statistics between long-term sites is presented. The graphical relation of peak flow to drainage area for Yukon River sites with at least 20 years of record is shown in figure 7 and may be used to estimate flood frequency statistics for sites along the Yukon River in Alaska and Canada with very large basins. USGS streamgage 15305700 had about one-half the record length of nearby USGS streamgage 15356000 and was removed to improve the continuity of the curve. The estimate of flood frequency flows may be obtained mathematically by a linear interpolation between the logarithms of the flood frequency flows in table 4 for the nearest designated long-term upstream and downstream Yukon River estimator streamgages shown in figure 7 using the logarithms of the drainage areas.

Table 7. Regional regression equations for estimating annual exceedance-probability discharges for unregulated streams in Alaska and conterminous basins in Canada.

[Regional regression equation: DRNAREA, drainage area, in square miles; PRECPRIS00, basin average mean annual precipitation, in inches, for 1971 to 2000 from the PRISM climate dataset. AVP: Average variance of prediction. SEP: Average standard error of prediction. R^2_{pseudo} : pseudo coefficient of determination]

Percent annual exceedance probability	Regional regression equation for estimating annual exceedance probability discharge, in cubic feet per second ^{1,2}	AVP (log units)	SEP (percent)	R^2_{pseudo} (percent)
50	0.944 (DRNAREA) ^{0.836} (PRECPRIS00) ^{1.023}	0.077	70.8	91.1
20	2.47 (DRNAREA) ^{0.795} (PRECPRIS00) ^{0.916}	0.074	69.1	90.6
10	4.01 (DRNAREA) ^{0.775} (PRECPRIS00) ^{0.865}	0.074	69.2	90.0
4	6.53 (DRNAREA) ^{0.755} (PRECPRIS00) ^{0.816}	0.077	71.2	89.0
2	8.79 (DRNAREA) ^{0.743} (PRECPRIS00) ^{0.787}	0.080	72.8	88.2
1	11.4 (DRNAREA) ^{0.732} (PRECPRIS00) ^{0.764}	0.083	74.6	87.4
0.5	14.3 (DRNAREA) ^{0.723} (PRECPRIS00) ^{0.744}	0.089	77.4	86.3
0.2	18.7 (DRNAREA) ^{0.712} (PRECPRIS00) ^{0.721}	0.097	81.9	84.7

¹Equations are valid for DRNAREA between 0.4 and 1,000 mi² with PRECPRIS00 between 8 and 280 in. and for DRNAREA greater than 1,000 and less than 31,100 mi² with PRECPRIS00 between 10 and 111 in.

²Equations are not suitable for use in the Aleutian Islands and other islands outside the study area.

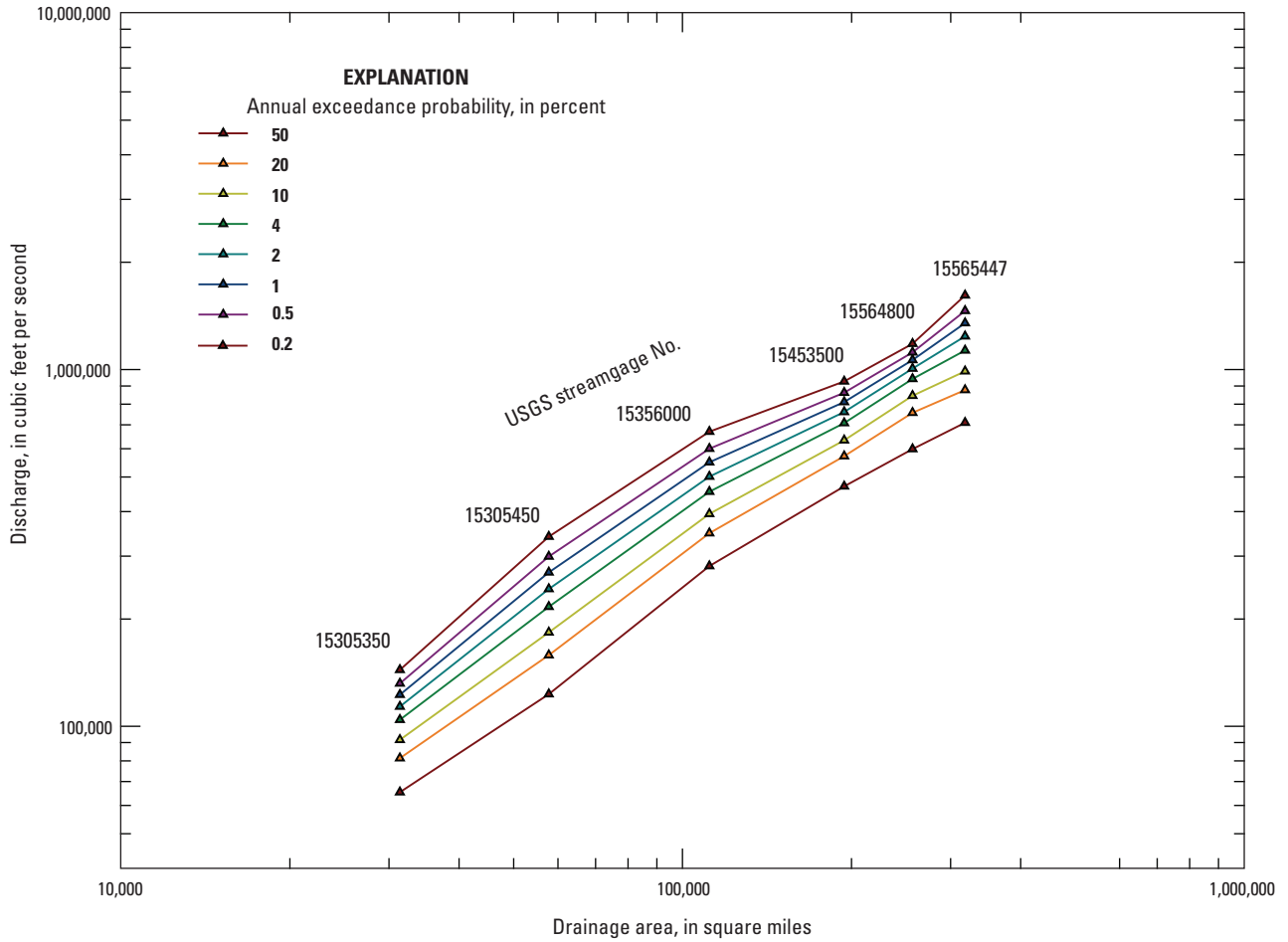


Figure 7. Relation of discharge to drainage area for selected annual exceedance probabilities for the Yukon River at and downstream of Carmacks, Yukon.

Accuracy and Limitations

The regression equations presented are empirical models that relate AEP flood flows to a particular dataset of physical and climatic basin attributes. These statistical relations must be interpreted and applied within the limits of the data and with the understanding that the results are best-fit estimates with an associated variance. As with all models, the regression equations have an associated measure of quality indicating

how well the predicted values represent the true values, and a reported uncertainty. For an individual regression equation estimate for a site, the variance of prediction, the standard error of prediction, and prediction intervals describe the accuracy and uncertainty of the estimate. The regression equations can be evaluated by the average variance of prediction and average standard error of prediction for the sites used to develop the equations, and by the pseudo coefficient of determination.

Variance of Prediction

For a site i , the variance of prediction, $V_{reg,i}$ can be thought of as a measure of the uncertainty of the regression model predictions. The $V_{reg,i}$ is the sum of the model error variance and the sampling error variance and can be computed as:

$$V_{reg,i} = \sigma_{\delta}^2 + \sigma_{\eta,i}^2, \tag{3}$$

where

σ_{δ}^2 is the model error variance; and
 $\sigma_{\eta,i}^2$ is the sampling mean square error for site i .

The variance of prediction is computed in WREG for each site used in the regression from the regression covariance matrix and the site-specific basin characteristics. For a regression equation, assuming that the explanatory variables for the streamgages used to develop the equation are representative of all sites in the region, the average accuracy of prediction for a regression equation can be determined by computing the average variance of prediction, AVP, for n number of streamgages:

$$AVP = \sigma_{\delta}^2 + \left(\frac{1}{n}\right) \sum_{i=1}^n \sigma_{\eta,i}^2. \tag{4}$$

Standard Error of Prediction

The standard error of prediction, SEP, is an alternate way to express the accuracy of the regression equations. The SEP is simply the square root of the V_{reg} , transformed from log units to percent. For a regression equation, the average standard error of prediction, SEP_{avg} , as a percentage of the respective AEP flow, can be computed from the AVP in log units using the following transformation:

$$SEP_{avg} = 100 \left[e^{(\ln 10)^2 AVP} - 1 \right]^{1/2} \tag{5}$$

About two-thirds of the estimates obtained from a regression equation for ungaged sites will have errors less than the SEP_{avg} for the equation (Helsel and Hirsch, 2002).

Pseudo Coefficient of Determination

A measure of the percentage of the variability in the dependent variable (AEP flow) explained by the explanatory variables in OLS regressions is the coefficient of determination, R^2 . For GLS regression, a more appropriate performance metric is the pseudo coefficient of determination, or R_{pseudo}^2 (Griffis and Stedinger, 2007). The R_{pseudo}^2 is a measure of the variability in the dependent variable explained by the regression after removing the effect of the time sampling error and is computed as:

$$R_{pseudo}^2 = 1 - \frac{\sigma_{\delta}^2(k)}{\sigma_{\delta}^2(0)} \tag{6}$$

where

$\sigma_{\delta}^2(k)$ is the model error variance from a GLS regression with k explanatory variables; and
 $\sigma_{\delta}^2(0)$ is the model error variance from a GLS regression with no explanatory variables.

The average variance of prediction, average standard error of prediction, and R_{pseudo}^2 for the final regression equations are given in table 7. The average standard error of prediction ranged from 69 to 82 percent for equations for the various AEPs, and the R_{pseudo}^2 ranged from 85 to 91 percent. These study performance metrics indicate that the flood frequency regression equations explained a fair amount of the variation in the AEP flows. In similar studies in other States, regression equations for many regions had lower errors and higher R_{pseudo}^2 . However, other States have noted that considerably greater errors and lower R_{pseudo}^2 can exist in areas of extreme flow variability (Gotvald and others, 2012; Paretti and others, 2014). The relation between observed and predicted flood discharges for the 1-percent AEP is shown as an example in figure 8. A slight bias appears in the distribution of points at the upper end of the plot, where the regression equation tends to under-predict the 1-percent AEP for streamflows on the order of 100,000 ft³/s or more. The uncertainty of the regression estimates can be seen graphically as a greater scatter of plotted observed to predicted points along the 1:1 line.

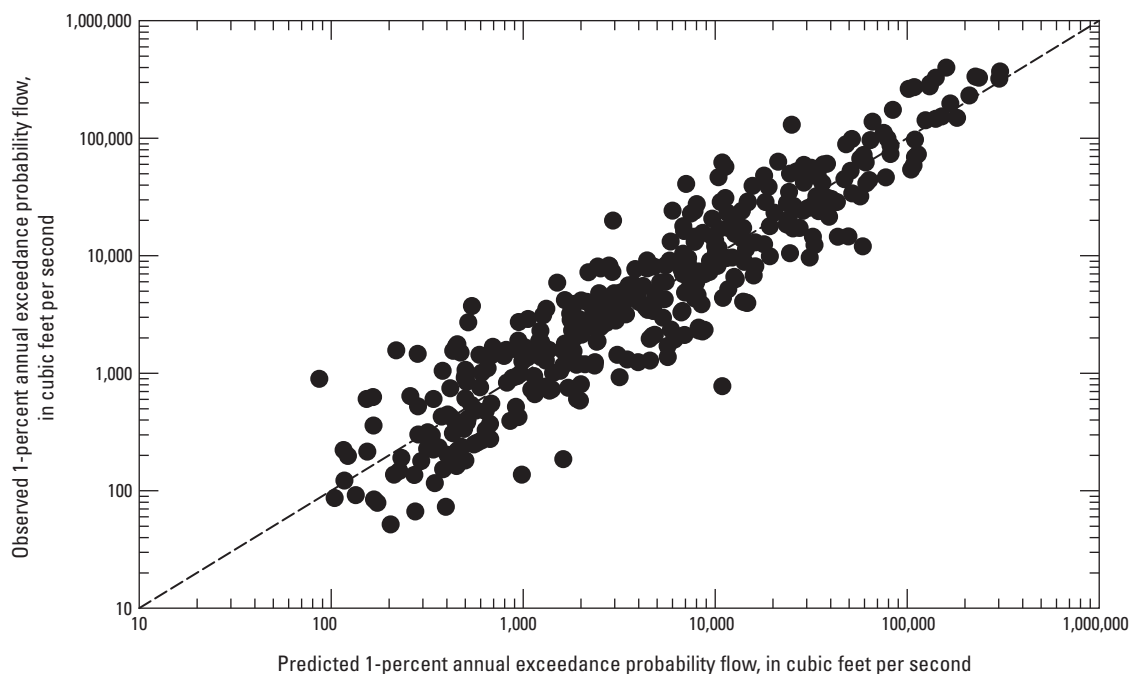


Figure 8. Relation between 1-percent annual exceedance probability discharges computed for observed streamflow and predicted from regression equations for streamgages in Alaska and conterminous basins in Canada.

Prediction Intervals

For a particular P -percent AEP flow at a streamgage, a site-specific uncertainty can be expressed as the confidence interval of a prediction, or prediction interval. The prediction interval is the range in values of an estimated dependent variable over which the true value of the dependent variable occurs with some stated probability. For example, for a 90-percent prediction interval for an estimated flow value, the probability that the true flow value lies within that interval is 90 percent. Tasker and Driver (1988) determined that a 100 $(1-\alpha)$ prediction interval for a streamflow statistic estimated at an ungaged site from a regression equation can be computed as follows:

$$\frac{Q}{C} < Q < QC \quad (7)$$

where

Q is the P -percent AEP flow computed from the regional regression equation; and
 C is computed as:

$$C = 10^{\left[t_{(\alpha/2, n-p)} \text{SEP}_i \right]}, \quad (8)$$

where

$t_{(\alpha/2, n-p)}$

is the critical value from the Student's t -distribution at a particular alpha-level and degrees of freedom $(n-p)$ where n is the number of streamgages included in the regression analysis and p is the number of parameters in the equation, including the intercept coefficient, and is equal to 1.65 for an α of 0.10, corresponding to a 90-percent prediction interval, and degrees of freedom computed from 341 streamgages and 3 regression parameters; and

SEP_i is the standard error of prediction for site i and is computed as:

$$\text{SEP}_i = \left[\sigma_s^2 + \mathbf{X}_i (\mathbf{X}^T \mathbf{\Lambda}^{-1} \mathbf{X})^{-1} \mathbf{X}_i^T \right]^{0.5} \quad (9)$$

where

σ_s^2 is the model error variance (computed using WREG and presented in table 8);
 \mathbf{X}_i is a row vector of the explanatory variables for site i , augmented by 1 as the first element;

$(\mathbf{X}^T \mathbf{\Lambda}^{-1} \mathbf{X})^{-1}$ is the covariance matrix for the regression coefficients (computed using WREG and presented in table 8); and

\mathbf{X}_i^T is the transpose of \mathbf{X}_i (Ludwig and Tasker, 1993).

Table 8. Values used to determine prediction intervals for the regional flood frequency regression equations for Alaska and conterminous basins in Canada.

[σ_{δ}^2 , the regression model error variance; $(\mathbf{X}^T \Lambda^{-1} \mathbf{X})^{-1}$, the covariance matrix; Intercept, y-axis intercept of regression equation; DRNAREA, drainage area, in square miles; PRECPRIS00, basin average mean annual precipitation, in inches, for 1971 to 2000 from the PRISM climate dataset]

P-percent annual exceedance probability	σ_{δ}^2	$(\mathbf{X}^T \Lambda^{-1} \mathbf{X})^{-1}$			
			Intercept	DRNAREA	PRECPRIS00
50	0.076	Intercept	7.26E-3	-7.56E-4	-3.42E-3
		DRNAREA	-7.56E-4	2.22E-4	2.12E-4
		PRECPRIS00	-3.42E-3	2.12E-4	1.86E-3
20	0.073	Intercept	7.42E-3	-7.70E-4	-3.47E-3
		DRNAREA	-7.70E-4	2.20E-4	2.18E-4
		PRECPRIS00	-3.47E-3	2.18E-4	1.88E-3
10	0.073	Intercept	7.82E-3	-8.09E-4	-3.64E-3
		DRNAREA	-8.09E-4	2.27E-4	2.30E-4
		PRECPRIS00	-3.64E-3	2.30E-4	1.96E-3
4	0.076	Intercept	8.64E-3	-8.91E-4	-4.01E-3
		DRNAREA	-8.91E-4	2.45E-4	2.55E-4
		PRECPRIS00	-4.01E-3	2.55E-4	2.15E-3
2	0.079	Intercept	9.29E-3	-9.56E-4	-4.30E-3
		DRNAREA	-9.56E-4	2.60E-4	2.74E-4
		PRECPRIS00	-4.30E-3	2.74E-4	2.30E-3
1	0.082	Intercept	9.95E-3	-1.02E-3	-4.60E-3
		DRNAREA	-1.02E-3	2.76E-4	2.94E-4
		PRECPRIS00	-4.60E-3	2.94E-4	2.46E-3
0.5	0.087	Intercept	1.08E-2	-1.11E-3	-4.98E-3
		DRNAREA	-1.11E-3	2.97E-4	3.19E-4
		PRECPRIS00	-4.98E-3	3.19E-4	2.66E-3
0.2	0.095	Intercept	1.20E-2	-1.23E-3	-5.54E-3
		DRNAREA	-1.23E-3	3.29E-4	3.55E-4
		PRECPRIS00	-5.54E-3	3.55E-4	2.96E-3

An example computation is provided in section, “Application of Methods for Estimating Flood Magnitude and Frequency.” A variety of tools are available online and within common spreadsheet software for computation of matrix algebra.

Limitations

The regression equations are valid for use in unregulated basins in Alaska and conterminous basins in Canada except for the Aleutian Islands and other islands off the west coast of Alaska (fig. 1) and for locations along the Yukon River at and downstream of Carmacks, Yukon (Map ID 242, fig. 1). No estimates are available for the Aleutian Islands, where insufficient streamgages are available for analysis. Separate methods are available for estimating flood frequency statistics

for sites on the Yukon River with particularly large drainage areas, considered to be those at and downstream of Carmacks, Yukon, where drainage area exceeds that of any other river in the study. Users should be cautioned that regression estimates are not exact and should be accompanied with estimates of uncertainty.

The regression equations are intended for use with basin characteristics obtained using the methods and datasets described in this report. Substituting basin characteristics obtained through alternate methods or datasets might have unpredictable results. Although the drainage area might vary only slightly depending on the topographic data available for a basin, the mean annual precipitation is likely to be sensitive to the method used to obtain it. In particular, users are cautioned to consider potential future changes in the processes that generate streamflow when exploring application of estimates

of future precipitation from climate models in these regression equations. Similarly, although site-specific data collection can improve the estimate of mean annual precipitation for a particular basin, the equations were developed from a regional-scale dataset that did not include such values. For example, regional precipitation datasets such as the PRISM model data used in this study have a known scarcity of data from high elevations, where few weather stations exist. Using a more accurate basin-specific mean annual precipitation developed from site-specific data in the regression will generate an estimate that has unknown error because the PRISM model did not consider such data.

The regression equations are intended for use for basins having values for drainage area and precipitation within the ranges shown in table 7. Applying the equations to sites in basins having values of explanatory variables outside the ranges of those used to develop the equations could result in prediction errors considerably greater than those indicated in table 7. The regression equations are not applicable for sites where peak-flow magnitudes are affected substantially by flow regulation, urbanization (as determined by impervious area), or glacier outbursts.

Application of Methods for Estimating Flood Magnitude and Frequency

Within the limitations previously described, the equations and streamflow statistics in this report can be used to estimate the 50-, 20, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent AEP flows (see table 3 for corresponding recurrence intervals) for gaged and ungaged streams in the study area. The best estimates of flood frequency statistics for a site typically are obtained by properly weighting independent estimates produced by more than one method (Interagency Advisory Committee on Water Data, 1982). Estimates from an EMA with MGB analysis for a gaged site can be weighted with estimates from the regression equations in this report in inverse proportion to their variance for an improved, weighted estimate (Cohn and others, 2012). Estimates at ungaged sites near a gaged site on the same stream can be improved by weighting with estimates for the nearby gaged site.

Regression Estimate and Prediction Interval

The regression estimate of flood frequency statistics, $Q_{P\%(reg)}$, for a gaged or ungaged site can be computed from the equations in table 7 by inserting the values of the basin characteristics for the site. The following example demonstrates the use of the regression equations to obtain an estimate of the $Q_{1\%(reg)}$ and the use of error equations to obtain the corresponding 90-percent prediction interval and standard error of prediction for USGS streamgage 15129500 Situk River near Yakutat, Alaska.

1. Using GIS tools, obtain drainage area for the basin and determine mean annual basin precipitation from the Gibson (2009a) PRISM dataset (DRNAREA=35.5 mi², and PRECPRIS00=170 in.);
2. If the computed basin characteristics are within the range of values listed in table 7, apply the regression equation in table 7 for estimating the flood frequency statistic ($Q_{1\%(reg)} = 11.4 \times 35.5^{0.732} \times 170^{0.764} = 7,870 \text{ ft}^3/\text{s}$);
3. Compute the standard error of prediction using matrix algebra to solve equation 9 where the model error σ_8^2 is retrieved from table 8, the \mathbf{X}_i vector is in the format $\mathbf{X}_i = \{1, \log_{10} \text{DRNAREA}, \log_{10} \text{PRECPRIS00}\}$ and the covariance matrix $(\mathbf{X}^T \mathbf{\Lambda}^{-1} \mathbf{X})^{-1}$ is retrieved from table 8, ($\sigma_8^2 = 0.082$, $\mathbf{X}_{15129500} = \{1, \log_{10}(35.5), \log_{10}(170)\}$, $\text{SEP}_{15129500} = (0.082 + 0.001202)^{0.5} = 0.2884$);
4. Compute C from equation 8 ($C = 10^{1.65 \times 0.2884} = 2.992$);
5. Compute the 90-percent prediction interval from equation 7 ($Q/C < Q_{1\%(reg)} < QC$, or $2,630 \text{ ft}^3/\text{s} < Q_{1\%(reg)} < 23,500 \text{ ft}^3/\text{s}$, meaning that one can be 90 percent confident that the true value of the estimate for site 15129500 lies between 2,630 and 23,500 ft³/s).

The regression estimate, prediction interval, and standard error of prediction also may be computed using the application tool provided as a Microsoft® Excel file at <http://dx.doi.org/10.3133/sir20165024>. For an ungaged site, the regression estimate as described above will likely be the final estimate unless it is near a long-term gaged site (see section, “Estimate for an Ungaged Site near a Streamgage” for this procedure) and should be reported as an estimate with the associated prediction interval and standard error of prediction for a particular AEP. Estimates using the regression equations in table 7 may vary slightly from the estimates shown in table 4, which were computed using WREG, because of rounding differences in equation parameters. For a gaged site, the regression estimate as described above is weighted with a station estimate from the station annual peak data as described in the following section.

Weighted Estimate for a Gaged Site

Flood frequency estimates at streamgages can be improved by computing a weighted average of the streamgage estimate obtained by log-Person Type III analysis of peak flows, here referred to as the station estimate, and the estimate from the regression equation (Interagency Advisory Committee on Water Data, 1982). Optimal weighted flow estimates can be obtained if the variance for each of the two estimates is known or can be estimated accurately. If the two flow estimates can be assumed to be independent and are weighted in inverse proportion to their associated variances, the variance of the weighted estimate will be less than the variance of either of the independent estimates.

For streamgages meeting eligibility requirements addressed in section, “**Limitations**,” the station and regression estimates can be computed using the methods described in this report, and then weighted using their respective variances using the following equation for a particular P -percent AEP:

$$\log Q_{wtd} = \frac{V_{reg} \log Q_{sta} + V_{sta} \log Q_{reg}}{V_{sta} + V_{reg}} \quad (10)$$

where

- Q_{wtd} is the weighted estimate of the peak flow for a streamgage, in cubic feet per second;
- V_{reg} is the variance of prediction for the regression estimate of peak flow for the streamgage, in log units (from [table 9](#) for sites in the study or from the average variance of prediction for sites not in this study);
- Q_{sta} is the station estimate of peak flow for the streamgage, in cubic feet per second (from [table 4](#) for sites in this study or from software such as PeakFQ for a site not in this study);
- V_{sta} is the variance of the station estimate of peak flow, in log units (from [table 9](#) for sites in the study or from software such as PeakFQ for a site not in this study); and
- Q_{reg} is the regression estimate of peak flow for the streamgage, in cubic feet per second (from the equations in [table 7](#)).

Weighted flow estimates computed for streamgages in the study are presented in [table 4](#). The variance associated with the weighted estimate, V_{wtd} is computed as:

$$V_{wtd} = \frac{V_{sta} V_{reg}}{V_{sta} + V_{reg}} \quad (11)$$

The variance of the station, regression, and weighted estimates of the P -percent AEP flows are presented in [table 9](#).

Table 9. Variance estimates for station, regression, and weighted estimates of flood frequency statistics for streamgages in Alaska and conterminous basins in Canada.

[Table 9 is a Microsoft® Excel file and can be downloaded at <http://dx.doi.org/10.3133/sir20165024>]

An example of the application of the procedure for obtaining a weighted estimate of a streamflow statistic for a gaged site is the following computation of the weighted 50-percent AEP flow for USGS streamgage 15297475, Red Cloud River tributary near Kodiak, Alaska. Slight differences between values listed below and in the respective tables are the result of rounding.

1. Obtain the station estimate of the 50-percent AEP flow at the streamgage using log-Pearson type III techniques and weighting the station skew with the appropriate regional skew if the site is located in a regional skew area ([table 5](#), [fig. 5](#)). For streamgages in the study, the station estimate was computed in PeakFQ and is listed in [table 4](#) ($Q_{50\%(sta)} = 386 \text{ ft}^3/\text{s}$);
2. Obtain explanatory variables drainage area and basin average mean annual precipitation for 1971–2000 from the PRISM climate dataset. [Table 1](#) contains values for these basin characteristics for sites in this study and [table 2](#) provides links to data sources. (DRNAREA = 1.7 mi², PRECPRIS00 = 85 in.);
3. Compute the regression estimate by inserting the explanatory variables in the equation for the 50-percent AEP from [table 7](#) ($Q_{50\%(reg)} = 0.944 (1.7^{0.836} \times 85^{1.023}) = 138 \text{ ft}^3/\text{s}$);
4. Obtain the variance for the station estimate from output from software such as PeakFQ, or from [table 9](#) for streamgages in this study ($V_{50\%(sta)} = 0.001$);
5. Obtain the variance of prediction for the regression estimate from [table 9](#) for gages used to develop the regression equation or from the AVP for other gages ($V_{50\%(reg)} = 0.077$);
6. Compute the weighted 50-percent AEP flow for the streamgage using equation 10 ($\log Q_{50\%(wtd)} = (0.077(\log(386)) + 0.001(\log(138)))/(0.001+0.077) = 2.5786$ and $Q_{50\%(wtd)} = 381 \text{ ft}^3/\text{s}$);
7. Compute the weighted variance of prediction for the weighted flow estimate for the streamgage using equation 11 ($V_{50\%(wtd)} = (0.001 \times 0.077)/(0.001+0.077) = 0.001$).

Estimate for an Ungaged Site near a Streamgage

For ungaged sites near a gaged site on the same stream, an improved estimate can be obtained from the regression estimate for the ungaged site weighted with an estimate based on the weighted estimate for the gaged site and a drainage-area based multiplier (Sauer, 1974; Ries, 2007). The sites are considered near if the drainage area of the ungaged site is within 50–150 percent of the drainage area of a gaged site. To obtain a weighted flood frequency estimate for the ungaged site, $Q_{(u)wtd}$, the weighted flow estimate for a particular AEP for an upstream or downstream gage, $Q_{(g)wtd}$, must first be computed from the equation given in the previous section for gaged sites. The estimate for the gaged site is then scaled by

a drainage-area-ratio based multiplier to obtain a gage-based estimate for the ungaged site, $Q_{(u)g}$, from the equation:

$$Q_{(u)g} = \left[\frac{A_{(u)}}{A_{(g)}} \right]^b Q_{(g)wtd} \tag{12}$$

where

- $A_{(u)}$ is the drainage area of the ungaged site;
- $A_{(g)}$ is the drainage area of the gaged site; and
- b is the exponent of the drainage area variable in the regional regression equation (DRNAREA, table 7).

The regression estimate for the ungaged site, $Q_{(u)reg}$, is computed from the regional regression equations as described in a previous section and then weighted with the gage-based estimate for the ungaged site in the following equation:

$$Q_{(u)wtd} = \frac{2\Delta A}{A_g} Q_{(u)reg} + \left[1 - \frac{2\Delta A}{A_g} \right] Q_{(u)g} \tag{13}$$

where

- ΔA is the absolute value of the difference between the drainage areas of the gaged and ungaged site, $|A_g - A_u|$.

The weighting procedure gives full weight to the regression estimates when the drainage area for the gaging station is less than 0.5 or greater than 1.5 times the drainage area for the ungaged site and increasing weight to the gage-based estimates as the drainage area ratio approaches 1. This procedure for weighting a regression-based and gage-based estimate does not account for the length of the streamgage record. For gaged sites with short records having a large difference in drainage area from the ungaged site, the regional regression equations may produce a better estimate.

An example of the application of the procedure for obtaining an estimate of an AEP discharge for an ungaged site near a gaged site on the same stream is the following computation of the weighted 10-percent AEP flow for USGS streamgage 15291700 Susitna River above Tsusena Creek near Chulitna, Alaska, which had only 3 years of record at the time of this report and will be treated here as an ungaged site. USGS streamgage 15292000 Susitna River at Gold Creek, Alaska is a nearby streamgage that is included in this study and is used in this example to provide a weighted estimate between the gaged and ungaged estimates for the ungaged site.

1. Obtain the value of $Q_{(g)wtd}$ for the desired AEP (from table 4 for the 10-percent AEP, $Q_{(g)wtd} = 66,900 \text{ ft}^3/\text{s}$);
2. Obtain the drainage areas and mean annual precipitation for the gaged and ungaged sites (from table 1, $\text{DRNAREA}_g = 6,130 \text{ mi}^2$ and $\text{PRECPRIS00}_g = 25 \text{ in.}$; from GIS computations using a drainage area boundary

and the mean annual precipitation dataset (Gibson, 2009a), $\text{DRNAREA}_u = 5,160 \text{ mi}^2$, $\text{PRECPRIS00}_u = 24 \text{ in.}$);

3. Determine the ratio of the ungaged site drainage area to the gaged site drainage area and confirm that it lies between 0.5 and 1.5 ($A_u/A_g = 5,160/6,130 = 0.842$, indicating the gaged site is near the ungaged site);
4. Compute $Q_{(u)g}$ using the results of steps 1 and 3 and table 7 in equation 12 (for the 10-percent AEP, $b = 0.775$; $Q_{(u)g} = (0.842)^{0.775} \times 66,900 = 58,552 \text{ ft}^3/\text{s}$);
5. Compute $Q_{(u)reg}$ from the drainage area and mean annual precipitation for the ungaged site and the regression equation in table 7 for the desired AEP (for the 10-percent AEP, $Q_{(u)reg} = 4.01 \times \text{DRNAREA}_u^{0.775} \times \text{PRECPRIS00}_u^{0.865} = 4.01 \times 5,160^{0.775} \times 24^{0.865} = 47,242 \text{ ft}^3/\text{s}$);
6. Compute the absolute value of the difference in drainage areas ($\Delta A = 6,130 - 5,160 = 970 \text{ mi}^2$); and
7. Compute the weighted estimate for the desired AEP for the ungaged site, $Q_{(u)wtd}$, using equation 13,

$$\begin{aligned} Q_{(u)wtd} &= \frac{2 \times 970}{6130} 47,242 + \left[1 - \frac{2 \times 970}{6130} \right] 58,552 \\ &= 0.316(47,242) + 0.684(58,552) \\ &= 55,000 \text{ ft}^3/\text{s} \end{aligned}$$

In this case, the estimate for the ungaged site obtained from the data for gaged site was weighted more heavily than the estimate obtained from the regression equation.

Estimate for a Site on the Yukon River

Gaged or ungaged sites on the Yukon River upstream of Carmacks, Yukon that meet the eligibility requirements addressed in section, “Limitations” are eligible for computation of the regression estimate of flood frequency statistics as in section, “Regression Estimate and Prediction Interval.” Gaged sites on the Yukon River upstream of Carmacks, Yukon also are eligible for computation of a weighted station and regression estimate. Downstream of Carmacks, Yukon, however, drainage areas for sites on the Yukon River exceed 31,100 mi², the maximum size eligible for application of the regression equations. For gaged sites exceeding this drainage area limit, station estimates can be computed using the methods described in this report but may not be weighted with regression estimates. For ungaged sites with drainage areas greater than 31,100 mi², flood frequency statistics may be estimated by assuming a linear relation between the nearest upstream

and downstream streamgages. The estimate can be obtained graphically from [figure 7](#) or more accurately by linear interpolation from the logarithms of the drainage areas and flood frequency flows for the sites shown in [figure 7](#). The estimate for an ungaged large Yukon River site, $Q_{P\%lgYukon}$ can be expressed in terms of the drainage area and $Q_{P\%}$ for the nearest upstream (*up*) and downstream (*down*) streamgages

$$\begin{aligned} \log Q_{P\%lgYukon} = & \log(Q_{P\%up}) \\ & + \frac{\log(DRNAREA_{site}) - \log(DRNAREA_{up})}{\log(DRNAREA_{down}) - \log(DRNAREA_{up})} \\ & \times (\log(Q_{P\%down}) - \log(Q_{P\%up})). \end{aligned} \quad (14)$$

An example of a linear interpolation to estimate the 2-percent AEP flow for a hypothetical ungaged site near Fort Yukon, Alaska, with a drainage area of 177,000 mi² is shown below.

1. Determine the nearest eligible upstream and downstream long-term USGS streamgages from [figure 7](#) or from [table 1](#) (15356000 is the nearest upstream gage, and 15453500 is the nearest downstream gage);
2. Obtain the streamgage drainage areas from [table 1](#) ($DRNAREA_{up} = 111,600$ mi², $DRNAREA_{down} = 194,000$ mi²) and the streamgage $Q_{2\%}$ from [table 4](#) ($Q_{2\%up} = 501,000$ ft³/s, $Q_{2\%down} = 760,000$ ft³/s);
3. Estimate the $Q_{2\%}$ by linear interpolation using equation 14 ($\log Q_{2\%} = \log(501,000) + ((\log(177,000) - \log(111,600)) \times (\log(760,000) - \log(501,000)) / (\log(194,000) - \log(111,600))) = 5.8508$ and $Q_{2\%} = 709,000$ ft³/s).

Cook Inlet Basin Streamstats

StreamStats (<http://water.usgs.gov/osw/streamstats/index.html>) is a web-based GIS application developed by the USGS and ESRI, Inc. that facilitates retrieval of streamflow statistics and associated information. StreamStats allows users to easily obtain selected streamflow-statistic estimates, upstream drainage-basin characteristics, and other information for a user-selected point on a stream. Using a GIS-based interactive map, the user can select a point on a stream and StreamStats will delineate the drainage area boundary upstream of the selected point. StreamStats will then use the boundary to obtain basin characteristics and solve streamflow-statistics estimation equations. The user also can select USGS streamgages to obtain selected streamflow statistics. Ries and others (2008) provide a detailed description of the StreamStats application. Although designed to eventually be a national application, StreamStats is being implemented on a state-by-state basis, typically through

cooperative funding agreements between the USGS and local partners. A StreamStats application for the Cook Inlet Basin of Alaska was implemented as the initial application for Alaska to test the functionality and utility of the relatively low resolution geospatial data available for Alaska at the time of this report for use in StreamStats.

StreamStats delineates drainage areas for user-selected sites from topographic and hydrographic data within limits enforced by pre-defined major watershed boundaries. Three GIS-data layers were processed to produce the Cook Inlet Basin StreamStats data layers: The 1:63,360-scale USGS National Hydrography Dataset (NHD) (<http://nhd.usgs.gov/>; Simley and Carswell, 2009) for Alaska, the 1:63,360-scale Watershed Boundary Dataset (WBD) (stored within the NHD geodatabase), and a digital elevation model (DEM) modified from the 60-m (197-ft) USGS National Elevation Dataset (NED) (<http://ned.usgs.gov/>; Gesch and others, 2009). The NHD is a digital vector dataset representing the locations of streams, lakes, and other surface-water features organized into a streamflow network. The WBD is a digital nested hierarchy of hydrologic units defining the areal extent of surface water drainage to a point. Hydrologic units in the WBD are delineated such that a 2-digit hydrologic code, or HUC-2, represents a major geographic region in the United States, and is then subsequently broken down into smaller units (HUC-4s, 8s, 10s, and 12s), each defining a smaller area (U.S. Geological Survey and U.S. Department of Agriculture, Natural Resources Conservation Service, 2009). The Cook Inlet Basin StreamStats incorporated HUC-12s as interior divisions within the eleven HUC-8s in the Basin. StreamStats for Cook Inlet Basin used a DEM built from the 60-m NED, and then resampled to 30 m (98 ft), as part of the National Water Quality Assessment Program for the Cook Inlet Basin (Brabets and others, 1999).

Several preprocessing steps were needed for each of the three data layers to facilitate rapid determination of the drainage area to a selected point and subsequent computation of basin characteristics. Preprocessing of the NHD included removing flowline paths disconnected from the stream network and selection of the primary flow path in those areas where the NHD indicated split flow (such as might happen when flow diverges around an island in a river or with a braided channel). Preprocessing also verified that any stream from the NHD only crossed the HUC boundary at the inlet and outlet of the HUC and that HUC outlets aligned exactly to the confluences of the streams. Streamgage locations were adjusted, or snapped, from their published coordinates to a new location on the corresponding NHD flowline for the purposes of network functionality. Snapped locations were reviewed to ensure their placement on the correct stream and edited manually as required. A hydro-corrected DEM was developed by filling depressions or sinks using the basin boundaries from the WBD to conserve known drainage divides and using the streams from the preprocessed NHD to create well-defined flow paths through the elevation data.

ArcHydro Tools, version 1.3, a set of utilities developed to operate in the ArcGIS v. 9.3.1 environment (Environmental Systems Research Institute, 2015) was used to process StreamStats data layers for Cook Inlet. As the project progressed, the migration to ArcGIS version 10.2.2 and ArcHydro Tools version 2.0 was adopted as a programmatically sound transition that would not adversely affect work already accomplished.

Basin characteristics for the purpose of developing flood frequency regression equations were obtained external to StreamStats. Mean annual precipitation from the PRISM precipitation dataset, developed by the PRISM Climate Group and published for Alaska by Gibson (2009a), was selected as a variable in flood frequency regression equations in this study and was built into StreamStats as the basis for obtaining basin characteristics. The regression equations for estimating flood frequency statistics published in the report will be available online in Cook Inlet Basin StreamStats immediately following publication of this report. Estimates of flood frequency statistics obtained using StreamStats to determine basin characteristics should be very close to estimates obtained by GIS methods external to StreamStats.

The NHD, WBD, and NED are national datasets that are subject to change as improvements are made. Edits to the NHD or WBD required for this study were derived from interpretations of various sources such as topographic maps, aerial photography, and satellite imagery. Edits to the WBD were reconciled with the Alaska WBD steward prior to implementation of StreamStats. Edits to the NHD were tracked for submission to the Alaska NHD steward. Any edits to the NHD or WBD subsequent to July 2011 or December 2012, respectively, will not be reflected in StreamStats for the Cook Inlet Basin.

Summary and Conclusions

This report, prepared by the U.S. Geological Survey in cooperation with the Alaska Department of Transportation and Public Facilities, the Alaska Department of Natural Resources, and the U.S. Army Corps of Engineers, updates methods for determining flood magnitude and frequency at streamgages and ungaged sites in Alaska and conterminous basins in Canada. An annual series of peak-flow data through water year 2012 was compiled for each of 387 streamgages that were not substantially affected by regulation or urbanization and that had at least 10 years of record. Flood-frequency estimates were computed for each streamgage using the Expected Moments Algorithm with a multiple Grubbs-Beck test for detecting multiple potentially influential low floods to fit a log-Pearson type III distribution to the logarithms of the annual peak flows. Evaluation of historical and censored flood information facilitated incorporation of these types of non-standard streamgage data into the flood

frequency analysis where appropriate. A Bayesian least-squares regression regional skew analysis using station skew coefficients for streamgages having at least 25 years of record produced a new constant model for regional skew in two new regional skew areas covering parts of Alaska. For streamgages in the regional skew areas, the station skew coefficients were weighted with the new regional skew for computation of final station estimates of the magnitude of the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probability flows. For streamgages in locations outside the regional skew areas, considered regional skew exclusion areas, station skew coefficients were used for computation of final station estimates of these annual exceedance probability flows.

Basin characteristics consisting of physical and hydrologic properties of the streamgage basins were obtained using GIS methods and used as explanatory variables in an exploratory all-subsets ordinary least-squares regression. Regression analysis used 341 streamgages that did not have an ineligible relation to basin characteristics, such as glacier outbursts or indeterminate basin, were not considered redundant with a nested basin, and were not a site on the Yukon River with a drainage area greater than 31,100 mi². A separate method for streamgages on the Yukon River with very large drainage areas was defined. Regression analysis for the eligible streamgages included testing various combinations of streamflow analysis regions and basin characteristics to determine the simplest model of statistically significant and hydrologically reasonable variables that best explained the variation in the flood frequency flows. Drainage area and mean annual precipitation together in a single study-wide set of equations performed as well as the best sets of variables within previously published streamflow analysis regions, collectively, and eliminated concerns associated with the small number of streamgages within a region. Final regional regression equations were developed using generalized least-squares techniques to account for cross correlation between streamgage locations and concurrent records. Average standard errors of prediction for the regression equations for the various annual exceedance probabilities ranged from 69 to 82 percent, and the pseudo coefficient of determination ranged from 85 to 91 percent.

Techniques for applying the regional regression equations for gaged and ungaged sites and for weighting the station and regression estimates for gaged sites are presented along with final estimates for streamgages used in the study. Simultaneously with this study, a StreamStats pilot application for the state of Alaska was developed for the Cook Inlet Basin that incorporates the new regional regression equations and facilitates access to published flood frequency and basin characteristic statistics and estimates for ungaged sites. StreamStats is a web-based application that provides published data for streamgages and delineates basins, computes basin characteristics, and obtains flood frequency estimates for ungaged sites.

Acknowledgments

Data and assistance provided by Environment Canada, particularly Lynne Campo (Water Survey of Canada) and Judy Kwan (Meteorological Service of Canada); by the National Weather Service Alaska-Pacific River Forecast Center; and by streamside residents and other river observers is gratefully acknowledged. This study benefitted from discussions of concepts and methods of flood frequency analysis with Timothy Cohn and Julie Kiang of the USGS Office of Surface Water and discussions of preliminary results and interpretations with many Alaska hydrologists. Sheryl Boyack of the USGS coordinated the WBD edits necessary for preparation of drainage area boundaries, and Donna Knifong of the USGS provided a computational tool for obtaining basin characteristics. The authors also thank Katharine Kolb, Alan Rea, Peter Steeves, and Lovina Turney of the USGS for their assistance with StreamStats.

References Cited

- Arendt, A., Bliss, A., Bolch, T., Cogley, J.G., Gardner, A.S., Hagen, J.-O., Hock, R., Huss, M., Kaser, G., Kienholz, C., Pfeffer, W.T., Moholdt, G., Paul, F., Radić, V., Andreassen, L., Bajracharya, S., Barrand, N., Beedle, M., Berthier, E., Bhambri, R., Brown, I., Burgess, E., Burgess, D., Cawkwell, F., Chinn, T., Copland, L., Davies, B., De Angelis, H., Dolgova, E., Filbert, K., Forester, R., Fountain, A., Frey, H., Giffen, B., Glasser, N., Gurney, S., Hagg, W., Hall, D., Haritashya, U.K., Hartmann, G., Helm, C., Herreid, S., Howat, I., Kapustin, G., Khromova, T., König, M., Kohler, J., Kriegel, D., Kutuzov, S., Lavrentiev, I., LeBris, R., Lund, J., Manley, W., Mayer, C., Miles, E.S., Li, X., Menounos, B., Mercer, A., Mölg, N., Mool, P., Nosenko, G., Negrete, A., Nuth, C., Pettersson, R., Racoviteanu, A., Ranzi, R., Rastner, P., Rau, F., Raup, B., Rich, J., Rott, H., Schneider, C., Seliverstov, Y., Sharp, M., Sigurðsson, O., Stokes, C., Wheate, R., Winsvold, S., Wolken, G., Wyatt, F., and Zheltyhina, N., 2012, Randolph glacier inventory—A dataset of global glacier outlines, version 2.0: Boulder, Colorado, Global Land Ice Measurements from Space, accessed June 18, 2012, at <http://www.glims.org/RGI/index.html>.
- Belsley, D.A., Kuh, E., and Welsch, R.E., 1980, Regression diagnostics—Identifying influential data and sources of collinearity: New York, John Wiley and Sons, Inc., p. 6–84.
- Berwick, V.K., Childers, J.M., and Kuentzel, M.A., 1964, Magnitude and frequency of floods in Alaska, south of the Yukon River: U.S. Geological Survey Circular 493, 15 p.
- Bieniek, P.A., Bhatt, U.S., Thoman, R.L., Angeloff, H., Partain, J., Papineau, J., Fritsch, F., Holloway, E., Walsh, J.E., Daly, C., Shulski, M., Hufford, G., Hill, D.F., Calos, S., and Gens, R., 2012, Climate divisions for Alaska based on objective methods: *Journal of Applied Meteorology and Climatology*, v. 51, no. 7, p. 1276–1289.
- Brabets, T.P., 1997, Geomorphology of the Lower Copper River, Alaska: U.S. Geological Survey Professional Paper 1581, 89 p.
- Brabets, T.P., Nelson, G.L., Dorava, J.M., and Milner, A.M., 1999, Water-quality assessment of the Cook Inlet basin, Alaska—Environmental setting: U.S. Geological Survey Water-Resources Investigations Report 99-4025, 65 p.
- Childers, J.M., 1970, Flood frequency in Alaska: U.S. Geological Survey Open-File Report, 30 p.
- Childers, J.M., Meckel, J.P., and Anderson, G.S., 1972, Floods of August 1967 in east-central Alaska: U.S. Geological Survey Water-Supply Paper 1880-A, 77 p.
- Cohn, T.A., Lane, W.L., and Baier, W.G., 1997, An algorithm for computing moments-based flood quantile estimates when historical flood information is available: *Water Resources Research*, v. 33, no. 9, p. 2089–2096.
- Cohn, T.A., Lane, W.L., and Stedinger, J.R., 2001, Confidence intervals for expected moments algorithm flood quantile estimates: *Water Resources Research*, v. 37, no. 6, p. 1695–1706.
- Cohn, T.A., Berenbrock, Charles, Kiang, J.E., and Mason, R.R., Jr., 2012, Calculating weighted estimates of peak streamflow statistics: U.S. Geological Survey Fact Sheet 2012–3038, 4 p., available at <http://pubs.usgs.gov/fs/2012/3038/>.
- 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.
- Cook, R.D., and Weisberg, S., 1982, Residuals and influence in regression: New York, Chapman and Hall, 230 p.
- Curran, J.H., 2012, Streamflow record extension for selected streams in the Susitna River Basin, Alaska: U.S. Geological Survey Scientific Investigations Report 2012-5210, 36 p.
- Curran, J.H., Meyer, D.F., and Tasker, G.D., 2003, Estimating the magnitude and frequency of peak streamflows for ungaged sites on streams in Alaska and conterminous basins in Canada: U.S. Geological Survey Water-Resources Investigations Report 03-4188, 101 p., available at <http://pubs.usgs.gov/wri/wri034188/>.

- Eash, J.D., and Rickman, R.L., 2004, Floods on the Kenai Peninsula, Alaska, October and November 2002: U.S. Geological Survey, Fact Sheet 2004-3023, 4 p.
- Eash, D.A., Barnes, K.K., and Veilleux, A.G., 2013, Methods for estimating annual exceedance-probability discharges for streams in Iowa, based on data through water year 2010: U.S. Geological Survey Scientific Investigations Report 2013-5086, 63 p. plus appendixes.
- Eng, K., Chen, Y.-Y., and Kiang, J., 2009, User's guide to the weighted-multiple-linear-regression program (WREG version 1.0): U.S. Geological Survey Techniques and Methods, book 4, chapter A8, 21 p., accessed February 2, 2015, at <http://pubs.usgs.gov/tm/tm4a8/pdf/TM4-A8.pdf>.
- Environment Canada, 2015, Environment Canada Data Explorer: Web site, accessed September 2013, at <http://www.ec.gc.ca/rhc-wsc/default.asp?lang=En&n=0A47D72F-1>.
- Environmental Systems Research Institute, Inc., 2015, ArcGIS resources arc hydro overview: Environmental Systems Research Institute, accessed July 1, 2015, at <http://resources.arcgis.com/en/communities/hydro/01vn0000000s000000.htm>.
- 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.
- Ferrians, O.J., 1998, Permafrost map of Alaska, USA: Boulder, Colorado, National Snow and Ice Data Center, accessed May 16, 2012, at <http://nsidc.org/data/ggd320>.
- Flynn, K.M., Kirby, W.H., and Hummel, P.R., 2006, User's Manual for Program PeakFQ Annual Flood-Frequency Analysis Using Bulletin 17B Guidelines: U.S. Geological Survey, Techniques and Methods Book 4, Chapter B4; 42 p.
- Gallant, A.L., Binnian, E.F., Omernik, J.M., and Shasby, M.B., 2010, Level III ecoregions of Alaska: U.S. Environmental Protection Agency, scale 1:5,000,000.
- Gesch, D., Evans, G., Mauck, J., Hutchinson, J., Carswell, W.J., Jr., 2009, The national map—Elevation: U.S. Geological Survey Fact Sheet 2009-3053, 4 p.
- Gibson, W., 2009a, Mean precipitation for Alaska 1971–2000: National Park Service, Alaska Regional Office, Geospatial Dataset-2170508, accessed February 1, 2010, at <https://irma.nps.gov/App/portal/Home>.
- Gibson, W., 2009b, Mean minimum temperature for Alaska 1971–2000: National Park Service, Alaska Regional Office, Geospatial Dataset-2170519, accessed February 1, 2010, at <https://irma.nps.gov/App/portal/Home>.
- 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.
- Gotvald, A.J., Barth, N.A., Veilleux, A.G., and Parrett, Charles, 2012, Methods for determining magnitude and frequency of floods in California, based on data through water year 2006: U.S. Geological Survey Scientific Investigations Report 2012-5113, 38 p., 1 pl., accessed July 24, 2012, at <http://pubs.usgs.gov/sir/2012/5113/>.
- 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., 2007, The use of GLS regression in regional hydrologic analyses: *Journal of Hydrology*, v. 344, p. 82–95.
- Griffis, V.W., and Stedinger, J.R., 2009, Log-Pearson type 3 distribution and its application in flood frequency analysis, III—Sample skew and weighted skew estimators: *Journal of Hydrologic Engineering*, v. 14, no. 2, p. 121–130.
- Griffis, V.W., Stedinger, J.R., and Cohn, T.A., 2004, Log Pearson type 3 Quantile estimators with regional skew information and low outlier adjustments: *Water Resources Research*, v. 40, W07503, 17 p.
- 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.
- Gruber, A.M., and Stedinger, J.R., 2008, Models of LP3 regional skew, data selection and Bayesian GLS regression, Paper 596, in Babcock, R.W., and Watson, R., eds., *World Environmental and Water Resources Congress—Ahupua'a*, Honolulu, Hawai'i, May 12–16, 2008: Environmental and Water Resources Institute, p. 1–10, accessed December 2, 2015, at [http://dx.doi.org/10.1061/40976\(316\)563](http://dx.doi.org/10.1061/40976(316)563).
- 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 2007 World Environmental and Water Resources Congress: American Society of Civil Engineers*, 10 p.
- Helsel, D.R., and Hirsch, R.M., 2002, Statistical methods in water resources: U.S. Geological Survey Techniques of Water Resources Investigations, book 4, chap. A3., 522 p.
- Hoaglin, D.C., and Welsch, R.E., 1978, The Hat Matrix in regression and ANOVA: *The American Statistician*, v. 32, no. 1, p. 17–22.

- Holmes, R.R., Jr., and Dinicola, K., 2010, 100-Year flood—It's all about chance: U.S. Geological Survey General Information Product 106, 1 p.
- Homer, C., Dewitz, J., Fry, J., Coan, M., Hossain, N., Larson, C., Herold, N., McKerrow, A., VanDriel, J.N., and Wickham, J., 2007, Completion of the 2001 National Land Cover Database for the conterminous United States: Photogrammetric Engineering and Remote Sensing, v. 73, no. 4, p. 337–341.
- Homer Tribune, 2012. Kenai Peninsula flooding disaster declared: Homer Tribune, September 26, 2012, accessed December 2, 2015, at <http://homertribune.com/2012/09/kenai-peninsula-flooding-disaster-declared/>.
- Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood flow frequency: Hydrology Subcommittee Bulletin 17B, 28 p., 14 appendixes, 1 pl.
- Joling, D., 2006, Flood waters begin to recede in southcentral Alaska: Juneau Empire, August 21, 2006, accessed December 2, 2015, at http://juneauempire.com/stories/082106/sta_20060821006.shtml#.Vlbqyt8nxz.
- Jones, S.H., and Fahl, C.B., 1994, Magnitude and frequency of floods in Alaska and conterminous basins of Canada: U.S. Geological Survey Water-Resources Investigations Report 93-4179, 122 p.
- Kiang, J.E., Stewart, D.W., Archfield, S.A., Osborne, E.B., and Eng, K., 2013, A national streamflow network gap analysis: U.S. Geological Survey Scientific Investigations Report 2013–5013, 79 p. plus 1 app. as a separate file, accessed July 1, 2015, at <http://pubs.usgs.gov/sir/2013/5013/>.
- Lamke, R.D., 1972, Floods of the summer of 1971 in South-Central Alaska: U.S. Geological Survey Open-File Report, 88 p.
- Lamke, R.D., 1978, Flood characteristics of Alaskan streams: U.S. Geological Survey Water-Resources Investigations Report 78-129, 61 p.
- Lamke, R.D., and Bigelow, B.B., 1988, Floods of October 1986 in south-central Alaska: U.S. Geological Survey Open-File Report 87-391, 31 p.
- Lamontagne, J.R., Stedinger, J.R., Berenbrock, Charles, Veilleux, A.G., Ferris, J.C., and Knifong, D.L., 2012, Development of regional skews for selected flood durations for the Central Valley Region, California, based on data through water year 2008: U.S. Geological Survey Scientific Investigations Report 2012-5130, 60 p.
- Ludwig, A.H., and Tasker, G.D., 1993, Regionalization of low-flow characteristics of Arkansas streams: U.S. Geological Survey Water-Resources Investigations Report 93-4013, 19 p.
- Martins, E.S., and Stedinger, J.R., 2002, Cross-correlation among estimators of shape: Water Resources Research, v. 38, no. 11, p. 34-1–34-7.
- Meyer, D.F., 1995, Flooding in the Middle Koyukuk River Basin, Alaska, August 1994: U.S. Geological Survey Water-Resources Investigations Report 95-4118, 8 p, 2 pls.
- National Weather Service Alaska-Pacific River Forecast Center, 2015, 1995 Floods in South Central Alaska: National Weather Service Alaska-Pacific River Forecast Center Web page, accessed December 2, 2015, at <http://aprfc.arh.noaa.gov/general/flood95.html>.
- National Aeronautics and Space Administration Land Processes Distributed Archive Center, 2011, ASTER global digital elevation model version 2: U.S. Geological Survey/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota.
- North American Land Change Monitoring System, 2013, 2005 Land Cover of North America at 250 meters, version 2.0: Natural Resources Canada, Instituto Nacional de Estadística y Geografía, and the U.S. Geological Survey, accessed March 5, 2013, at <http://www.cec.org/naatlas>.
- Paretti, N.V., Kennedy, J.R., Turney, L.A., and Veilleux, A.G., 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., accessed February 10, 2015, at <http://dx.doi.org/10.3133/sir20145211>.
- Parks, B., and Madison, R.J., 1985, Estimation of selected flow and water-quality characteristics of Alaskan streams: U.S. Geological Survey Water-Resources Investigations Report 84-4247, 64 p.
- 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.
- Perica, S., Kane, D., Dietz, S., Maitaria, K., Martin, D., Pavlovic, S., Roy, I., Stuefer, S., Tidwell, A., Trypaluk, C., Unruh, D., Yekta, M., Betts, E., Bonnin, G., Heim, S., Hiner, L., Lilly, E., Narayanan, J., Fenglin, Y., and Zhao, T., 2012, NOAA Atlas Volume 7 Version 2.0, Precipitation-frequency atlas of the United States, Alaska: National Oceanic and Atmospheric Administration, National Weather Service, Silver Spring, Maryland.
- Ralph, F.M., and Dettinger, M.D., 2011, Storms, floods, and the science of atmospheric rivers: Eos, Transactions American Geophysical Union, v. 92, no. 32, p. 265–266.

- Reis, D.S., Jr., Stedinger, J.R., and Martins, E.S., 2005, Bayesian generalized least squares regression with application to the log Pearson type III regional skew estimation: *Water Resources Research*, v. 41, W10419, doi:10.1029/2004WR003445, 14 p.
- Ries III, K.G., 2007, The national streamflow statistics program—A computer program for estimating streamflow statistics for ungaged sites: U.S. Geological Survey Techniques and Methods 4-A6, 37 p.
- Ries III, K.G., Guthrie, J.D., Rea, A.H., Steeves, P.A., and Stewart, D.W., 2008, StreamStats—A water resources web application: U.S. Geological Survey Fact Sheet 2008-3067, 6 p.
- Sauer, V.B., 1974, Flood characteristics of Oklahoma streams—Techniques for calculating magnitude and frequency of floods in Oklahoma, with compilations of flood data through 1971: U.S. Geological Survey Water-Resources Investigations Report 73-52, 307 p.
- Simley, J.D., and Carswell, W.J., Jr., 2009, The National Map—Hydrography: U.S. Geological Survey Fact Sheet 2009-3054, 4 p.
- Stauffer, C., 2010, Learning to live with water: A history of flooding in Seward, Alaska, 1903–2009: Seward/Bear Creek Flood Service Area, 10 p., accessed December 2, 2015, at http://www.kpb.us/images/KPB/Service_Areas/sbcfsa/documents/Seward_Flood_History_Complete.pdf.
- Southard, R.E., and Veilleux, A.G., 2014, Methods for estimating annual exceedance-probability discharges and largest recorded floods for unregulated streams in rural Missouri: U.S. Geological Survey Scientific Investigations Report 2014–5165, 39 p., accessed June 10, 2015, at <http://dx.doi.org/10.3133/sir20145165>.
- Spatial Climate Analysis Service, 2002, Western Canada average monthly or annual precipitation, 1961–1990: Corvallis, Oregon State University, accessed December 1, 2011, at www.climatesource.com.
- Stedinger, J.R., and Cohn, T.A., 1986, Flood frequency analysis with historical and paleoflood information: *Water Resources Research*, v. 22, no. 5, p. 785–793.
- Stedinger, J., and Griffis, V., 2008, Flood frequency analysis in the United States—Time to update: *Journal of Hydrologic Engineering*, v. 13, no. 4, p. 199–204.
- Stedinger, J.R., and Tasker, G.D., 1985, Regional hydrologic analysis, 1, ordinary, weighted, and generalized least squares compared: *Water Resources Research*, v. 21, no. 9, p. 1421–1432.
- Tasker, G.D., and Driver, N.E., 1988, Nationwide regression models for predicting urban runoff water quality at unmonitored sites: *Journal of the American Water Resources Association*, v. 24, no. 5, p. 1091–1101.
- Tasker, G.D., and Stedinger, J.R., 1986, Regional skew with weighted LS regression: *Journal of Water Resources Planning and Management, ASCE*, v. 112, no. 2, p. 225–237.
- Tasker, G.D., and Stedinger, J.R., 1989, An operational GLS model for hydrologic regression: *Journal of Hydrology*, v. 111, nos. 1–4, p. 361–375.
- U.S. Geological Survey, 2015, Peak streamflow for Alaska: U.S. Geological Survey National Water Information System, accessed various dates from November 2014 through February 2015, at <http://nwis.waterdata.usgs.gov/ak/nwis/peak>.
- U.S. Geological Survey and U.S. Department of Agriculture, Natural Resources Conservation Service, 2009, Federal guidelines, requirements, and procedures for the national Watershed Boundary Dataset: U.S. Geological Survey Techniques and Methods 11–A3, 55 p., accessed July 1, 2015, at <http://pubs.usgs.gov/tm/11/a3/>.
- Veilleux, A.G., 2009, Bayesian GLS regression for regionalization of hydrologic statistics, floods, and Bulletin 17B skew: Ithaca, New York, Cornell University, M.S. Thesis, 155 p.
- Veilleux, A.G., 2011, Bayesian GLS regression, leverage and influence for regionalization of hydrologic statistics: Cornell, New York, Cornell University, Ph.D. dissertation, 184 p.
- Veilleux, A.G., Stedinger, J.R., and Lamontagne, J.R., 2011, Bayesian WLS/GLS regression for regional skewness analysis for regions with large cross-correlations among flood flows, in Beighley, R.E., and Kilgore, M.W., eds., *World Environmental and Water Resources Congress 2011—Bearing knowledge for sustainability*, Proceedings: Palm Springs, California, May 22–26, 2011, American Society of Civil Engineers, p. 3103–3112.
- 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.E., ed., *World Environmental and Water Resources Congress 2012—Crossing Boundaries*, Proceedings: Albuquerque, New Mexico, May 20–24, 2012, American Society of Civil Engineers, Paper 227, p. 2253–2263.

- Veilleux, A.G., Cohn, T.A., Flynn, K.M., Mason, R.R., Jr., and Hummel, P.R., 2014, Estimating magnitude and frequency of floods using the PeakFQ 7.0 program: U.S. Geological Survey Fact Sheet 2013–3108, 2 p., accessed October 23, 2014, at <http://dx.doi.org/10.3133/fs20133108>.
- Wiley, J.B., and Curran, J.H., 2003, Estimating annual high-flow statistics and monthly and seasonal low-flow statistics for ungaged sites on streams in Alaska and conterminous basins in Canada: U.S. Geological Survey Water-Resources Investigations Report 03-4114, 61 p.
- Weaver, J.C., Feaster, T.D., and Gotvald, A.J., 2009, Magnitude and frequency of rural floods in the Southeastern United States, through 2006—Volume 2, North Carolina: U.S. Geological Survey Scientific Investigations Report 2009–5158, 111 p., accessed March 20, 2014, at <http://pubs.usgs.gov/sir/2009/5158/>.

Appendix A. Basin Characteristics for Selected Streams in Alaska and Conterminous Basins in Canada

Appendix A is a Microsoft® Excel file and can be downloaded at <http://dx.doi.org/10.3133/sir20165024>.

This page left intentionally blank

Appendix B. Regional Skewness Regression Analysis

By Andrea G. Veilleux

Introduction to Statistical Analysis of Regional Skew

For the log-transformation of annual peak discharges, Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) recommends using a weighted average of the station skew coefficient and a regional skew coefficient to help improve estimates of annual exceedance probability (AEP) discharges (eq. 2). Bulletin 17B supplies a national map but also encourages hydrologists to develop more specific local relations. Since the first map was published in 1976, some 40 years of additional information has accumulated and better spatial estimation procedures have been developed (Stedinger and Griffis, 2008).

Tasker and Stedinger (1986) developed a weighted least-squares (WLS) procedure for estimating regional skew coefficients based on sample skew coefficients for the logarithms of annual peak-discharge data. Their method of regional analysis of skew estimators accounts for the precision of the skew-coefficient estimate for each streamgauge or station, which depends on the length of record for each streamgauge and the accuracy of an ordinary least-squares (OLS) regional mean skew. More recently, Reis and others (2005), Gruber and others (2007), and Gruber and Stedinger (2008) developed a Bayesian generalized least-squares (GLS) regression model for regional skew analyses. The Bayesian methodology allows for the computation of a posterior distribution of both the regression parameters and the model error variance. As shown in Reis and others (2005), for cases in which the model error variance is small compared to the sampling error of the station estimates, the Bayesian posterior distribution provides a more reasonable description of the model error variance than both the GLS method-of-moments and maximum likelihood point estimates (Veilleux, 2011). Although WLS regression accounts for the precision of the regional model and the effect of the record length on the variance of skew-coefficient estimators, GLS regression also

considers the cross-correlations among the skew-coefficient estimators. In some studies, the cross-correlations have had a large impact on the precision attributed to different parameter estimates (Feaster and others, 2009; Gotvald and others, 2009; Weaver and others, 2009; Parrett and others, 2011).

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

The Alaska regional skew study described here used the Expected Moments Algorithm to estimate the station skew and its mean square error. Because EMA allows for the censoring of potentially influential low floods (PILFs), as well as the use of estimated interval discharges for missing, censored, and historic data, it complicates the calculations of effective record length (and effective concurrent record length) used to describe the precision of sample estimators because the peak discharges are no longer solely represented by single values. To properly account for these complications, the new B-WLS/B-GLS procedure was used.

Methodology for Regional Skew Model

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

Ordinary Least-Squares Analysis

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

$$\tilde{y}_{OLS} = \mathbf{X}\hat{\beta}_{OLS} \quad (\text{B1})$$

where

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

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

Weighted Least-Squares Analysis

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

Generalized Least-Squares Analysis

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

Data Analysis

This regional skew study is based on annual peak-discharge data from 103 streamgages in Alaska (table 1). Streamgages included in the skew study were screened as described in the main section of the report to omit streamgages on the same river close enough to be considered hydrologically redundant. The annual peak-discharge data were downloaded from the USGS National Water Information System (NWIS) database. In addition to the peak-discharge data, 11 basin characteristics for each of the 103 sites were available as explanatory variables in the regional study. The basin characteristics available included hydrologic regions (Curran and others, 2003), as well as the more standard morphometric parameters (basin centroid latitude and longitude, drainage area, and mean basin elevation), climate variables (mean annual precipitation and mean minimum January temperature), and land cover variables (percentage of basin forested, percentage of basin covered by lakes and ponds, percentage of basin covered by glaciers, and percentage of basin covered by permafrost).

Due to the sparsity of streamgages in parts of Alaska and conterminous basins in Canada, a regional skew analysis could not be performed for the entire study area. Instead, two regions of Alaska were identified which each contained a sufficient density of gages from which to build regional skew models. These regions are shown in figure 5 of the report. Regional Skew Area 1 (RSA1) (75 streamgages) covers 108,000 mi² and encompasses a swath along the road corridors near the more populated areas of the interior part of the State. Regional Skew Area 2 (RSA2) (28 streamgages) covers 72,800 mi² and encompasses the coastal areas of the State bordering the Gulf of Alaska. The boundaries of these regions are generally defined by HUC 8 borders, split where necessary to omit non-applicable areas.

RSA1 is situated in the Interior of Alaska, which experiences a continental climate having warmer summers, colder winters, and less precipitation than RSA2. The median value for mean annual precipitation for the study basins in RSA1 is 25 in., and winter precipitation generally falls as snow. Mean basin elevations generally are higher than elevations in coastal basins. Annual peak flow can be generated by snowmelt or rainfall, or by glacier-related melt in glacierized basins.

The temperate, moist climate of RSA2 reflects the maritime influence of the Gulf of Alaska. Mean annual precipitation is much higher than for RSA1; the median value of the mean annual precipitation for study basins is 145 in. for RSA2. Flooding is more commonly generated by rainfall, particularly in autumn and winter. Most of the gaged basins in RSA2 are small relative to the median study basin size, which in part mirrors the physiography of short drainages extending from the mountain ranges surrounding the Gulf of Alaska to the coast.

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

Station Skew Estimators

To estimate the station logarithm base10 (log) skew coefficient, G , and its mean square error, MSE_G , the skew study used the results of the EMA analysis described in section, “Flood Magnitude and Frequency at Gaged Sites” (Cohn and others, 1997; Griffis and others, 2004). EMA provides a straightforward and efficient method for the incorporation of historical information and censored data, such as those from a CSG, contained in the record of annual peak discharges for a streamgage. For this analysis, PeakFQ (Veilleux and others, 2014), which combines EMA with the multiple Grubbs-Beck test for PILFs, was used. Use of the most current PeakFQ version available at the time resulted in a small subset of study sites being affected by computational discrepancies in version 7.0 before version 7.1, which corrected processing of selected conditions, became available. Documentation for PeakFQ is available at <http://water.usgs.gov/software/PeakFQ/>. PeakFQ was used to generate the station log estimates of G and the corresponding MSE_G , assuming a log-Pearson Type III distribution and generally using a multiple Grubbs-Beck test for PILF screening. EMA estimates, based on annual peak-discharge data through September 30, 2012, of G and MSE_G are listed in table 4 of the report for the 103 streamgages evaluated for the Alaska regional skew study.

Pseudo Record Length

Because the dataset includes censored data and historical information, the effective record length used to compute the precision of the skew estimators is no longer simply the number of annual peak discharges at a streamgage. Instead, a more complex calculation was used to take into account the availability of historical information and censored values. Although historical information and censored peaks provide valuable information, they often provide less information than an equal number of years with systematically recorded peaks (Stedinger and Cohn, 1986). The following calculations provide a pseudo record length, P_{RL} , associated with skew, which appropriately accounts for all peak-discharge data types available for a site.

The P_{RL} is defined in terms of the number of years of systematic record that would be required to yield the same mean square error of the skew ($MSE(\hat{G})$) as the combination of historical and systematic record actually available at a streamgage. Thus, the P_{RL} of the skew is a ratio of the MSE of the at-site skew when only the systematic record is analyzed ($MSE(\hat{G}_S)$) versus the MSE of the at-site skew when all data, including historical and censored data, are analyzed ($MSE(\hat{G}_C)$).

$$P_{RL} = \frac{P_s * MSE(\hat{G}_S)}{MSE(\hat{G}_C)} \tag{B2}$$

where

- P_{RL} is the pseudo record length for the entire record at the streamgage;
- P_s is the number of systematic peaks in the record;
- $MSE(\hat{G}_S)$ is the estimated MSE of the skew when only the systematic record is analyzed; and
- $MSE(\hat{G}_C)$ is the estimated MSE of the skew when all data, including historical and censored data, is analyzed.

As the P_{RL} is an estimate, the following conditions also must be met to ensure a valid approximation. P_{RL} must be non-negative. If P_{RL} is greater than P_H (length of the historical period), then P_{RL} should be set to equal P_H . Also, if P_{RL} is less than P_s , then P_{RL} is set to P_s . This ensures that the pseudo record length will not be larger than the complete historical period or less than the number of systematic peaks.

For the Alaska skew, only sites with P_{RL} greater than or equal to 25 years of record were used in the analysis.

Unbiasing the Station Estimators

The station skew estimates were unbiased by using the correction factor developed by Tasker and Stedinger (1986) and employed in Reis and others (2005). The unbiased station skew estimator using the pseudo record length is

$$\hat{\gamma}_i = \left[1 + \frac{6}{P_{RL,i}} \right] G_i \tag{B3}$$

where

- $\hat{\gamma}_i$ is the unbiased station sample skew estimate for site i ,
- $P_{RL,i}$ is the pseudo record length for site i as calculated in equation B2; and
- G_i is the traditional biased station skew estimator for site i from EMA.

42 Estimating Flood Magnitude and Frequency on Streams in Alaska and Conterminous Basins in Canada

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

$$Var[\hat{\gamma}_i] = \left[1 + \frac{6}{P_{RL,i}} \right]^2 Var[G_i] \quad (B4)$$

where

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

$$Var(\hat{G}) = \left[\frac{6}{P_{RL}} + a(P_{RL}) \right] * \left[1 + \left(\frac{9}{6} + b(P_{RL}) \right) \hat{G}^2 + \left(\frac{15}{48} + c(P_{RL}) \right) \hat{G}^4 \right] \quad (B5)$$

where

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

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

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

Estimating the Mean Square Error of the Skew Estimator

There are several possible ways to estimate MSE_{G_s} . The approach used by EMA (Cohn and others, 2001, eq.55) generates a first-order estimate of the MSE_{G_s} , which should perform well when interval data are present. Another option is to use the Griffis and Stedinger (2009) formula in equation B5 (variance is equated to the MSE), using either the systematic record length or the length of the whole historical period. However, this method does not account for censored data, and thus can lead to inaccurate and underestimated MSE_{G_s} . This issue has been addressed by using the pseudo record length instead of the length of the historical period; the pseudo record length reflects the impact of the censored data and the number of recorded systematic peaks. Thus, the unbiased Griffis and Stedinger (2009) MSE_{G_s} was used in the regional skew model because it is more stable and relatively independent of the station skew estimator. This methodology was used in previous regional skew studies (Eash and others, 2013; Southard and Veilleux, 2014).

Cross-Correlation Models

A critical step for a GLS analysis is estimation of the cross-correlation of the skew coefficient estimators. Martins and Stedinger (2002) used Monte Carlo experiments to derive a relation between the cross-correlation of the skew estimators

at two stations i and j as a function of the cross-correlation of concurrent annual maximum flows, ρ_{ij} :

$$\hat{\rho}(\hat{\gamma}_i, \hat{\gamma}_j) = Sign(\hat{\rho}_{ij}) cf_{ij} |\hat{\rho}_{ij}|^k \quad (B6)$$

where

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

$$cf_{ij} = CY_{ij} / \sqrt{(P_{RL,i})(P_{RL,j})} \quad (B7)$$

where

CY_{ij} is the pseudo record length of the period of concurrent record; and,
 $P_{RL,i}$ and $P_{RL,j}$ are the pseudo record length corresponding to sites i and j , respectively (see equation B2).

Pseudo Concurrent Record Length

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

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

$$CY_{ij} = (YE_{ij} - YB_{ij} + 1) \left(\frac{P_{RL,i}}{P_{H,i}} \right) \left(\frac{P_{RL,j}}{P_{H,j}} \right). \quad (B8)$$

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

Alaska Study Area Cross-Correlation Models of Concurrent Annual Peak Discharge

Cross-correlation models for the log annual peak discharges were developed for both RSA1 and RSA2 in Alaska. The cross-correlation model for RSA1 was developed using 27 sites with at least 35 years of concurrent systematic peaks (zero flows not included). Similarly, the cross-correlation model for RSA2 was developed using 28 sites with at least 20 years of concurrent systematic peaks (zero flows

not included). Various models relating the cross-correlation of the concurrent annual peak discharge at two sites, ρ_{ij} , to various basin characteristics were considered. A logit model, termed the Fisher Z Transformation ($Z = \log[(1+r)/(1-r)]$), provided a convenient transformation of the sample correlations r_{ij} from the $(-1, +1)$ range to the $(-\infty + \infty)$ range.

The adopted models for estimating the cross-correlations of concurrent annual peak discharge at two stations, which used the distance between basin centroids, D_{ij} , as the only explanatory variable, are

$$\rho_{ij} = \frac{\exp(2Z_{ij}) - 1}{\exp(2Z_{ij}) + 1} \quad (B9)$$

where

$$Z(\text{RSA1})_{ij} = \exp(0.48 - 0.0082 * D_{ij}), \text{ and}$$

$$Z(\text{RSA2})_{ij} = \exp(0.39 - 0.0078 * D_{ij}).$$

For RSA1, an OLS regression analysis based on 252 station-pairs indicated that this model is as accurate as having 67 years of concurrent annual peaks from which to calculate cross-correlation. For RSA2, an OLS regression analysis based on 157 station-pairs indicated that this model is as accurate as having 900 years of concurrent annual peaks from which to calculate cross-correlation. The fitted relation between Z and distance between basin centroids together with the plotted sample data for RSA1 and RSA2 is shown in figure B1. The functional relation between the untransformed cross correlation and distance between basin centroids together with the plotted sample data for RSA1 and RSA2 is shown in figure B2. The cross correlation models were used to estimate site-to-site cross correlations for concurrent annual peak discharges at all pairs of sites in RSA1 and RSA2.

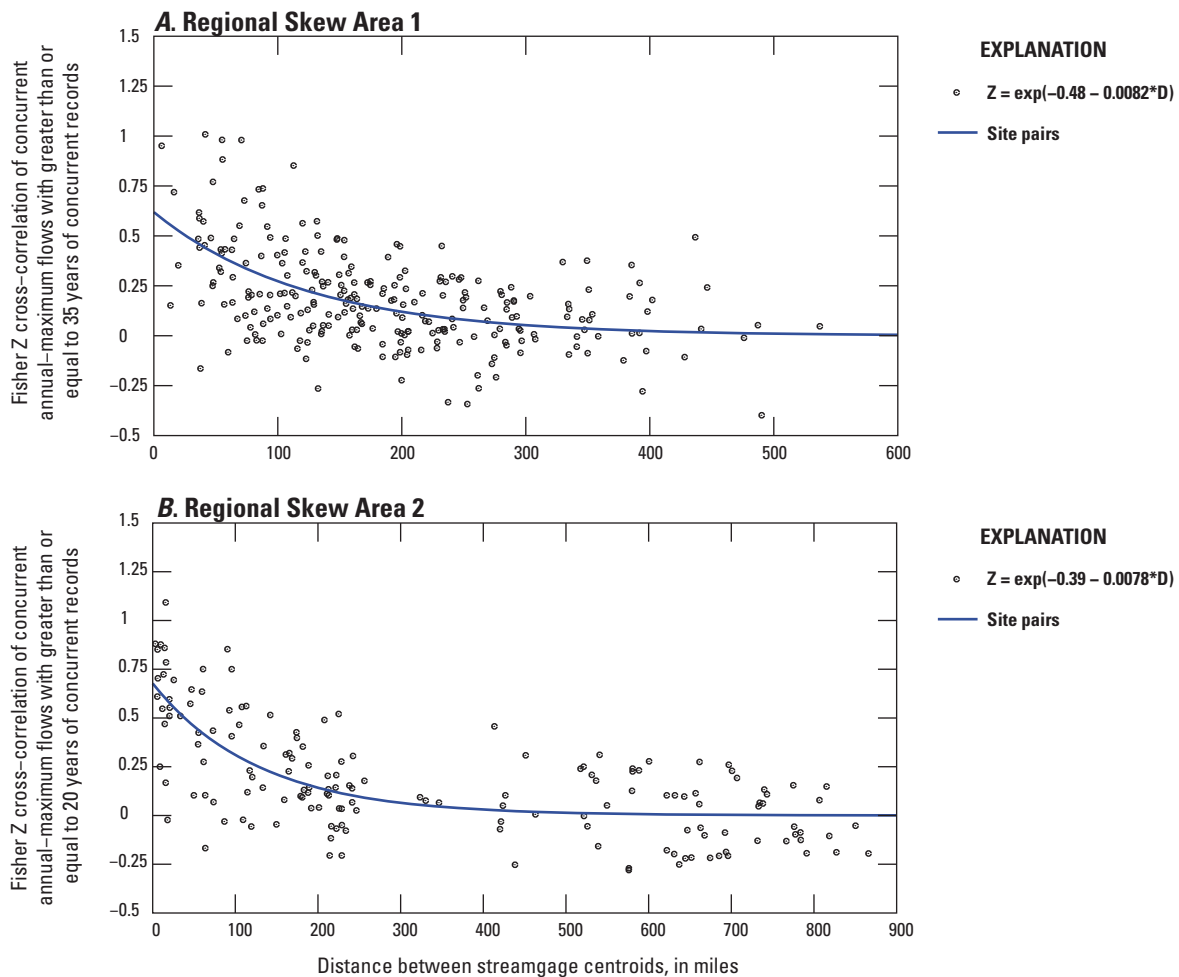


Figure B1. Relation between Fisher Z transformed cross-correlation of logs of annual peak discharge and distance between basin centroids for (A) Regional Skew Area 1 and (B) Regional Skew Area 2, Alaska and conterminous basins in Canada.

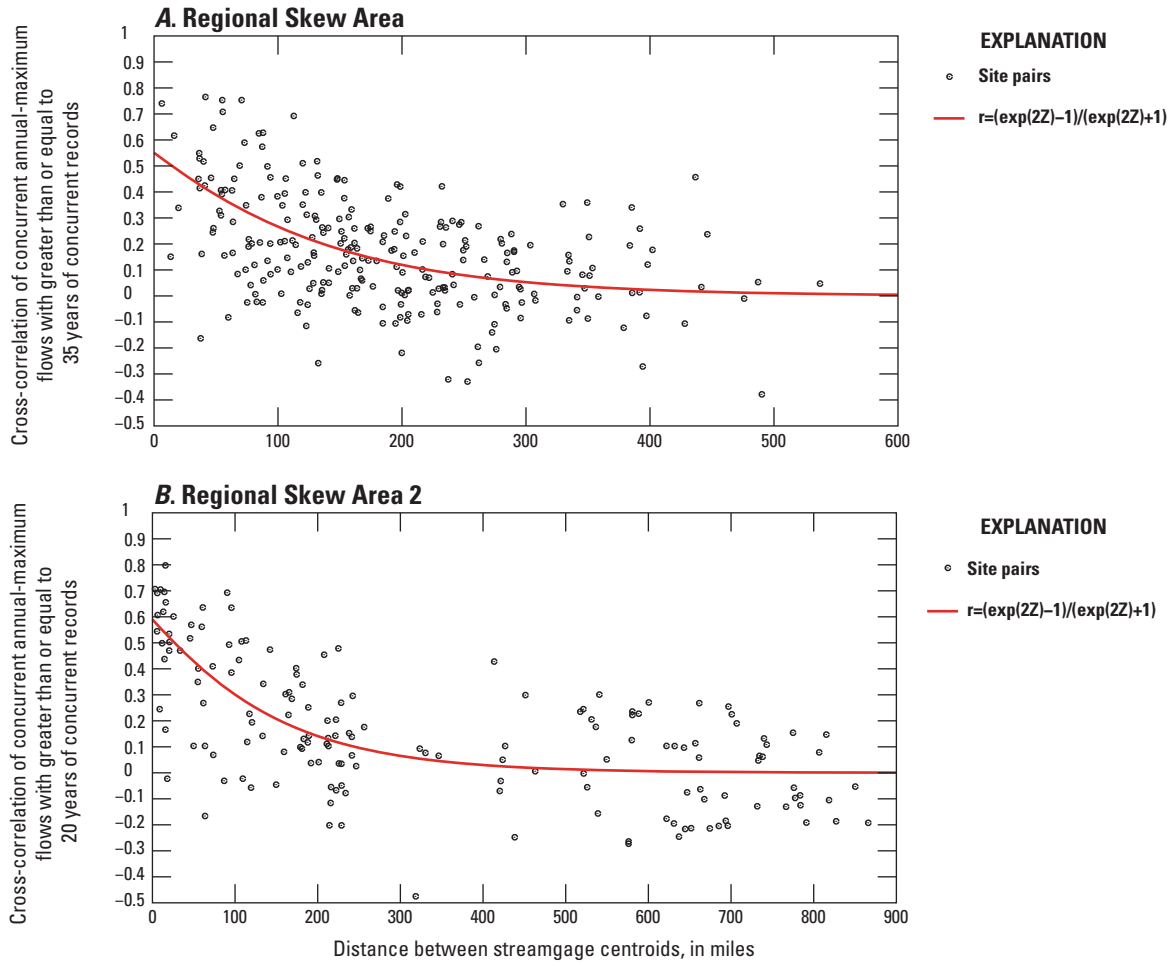


Figure B2. Relation between untransformed cross-correlation of logs of annual peak discharge and distance between basin centroids for (A) Regional Skew Area 1 and (B) Regional Skew Area 2, Alaska and conterminous basins in Canada.

Alaska Regional Skew Study Results

The results of the Alaska regional skew study using the B-WLS/B-GLS regression methodology are provided in the following sections for RSA1 and RSA2.

Regional Skew Area 1

All available basin characteristics were initially considered as explanatory variables in the regression analysis for regional skew in region RSA1. The basin characteristics that were statistically significant in explaining the streamgauge-to-streamgauge variability in skew were the mean annual precipitation (PRECPRI00), drainage area (DRNAREA), percent glacier, and mean minimum January temperature.

The best regional skew model is classified as having the smallest model error variance, σ_{δ}^2 , and largest pseudo R_{δ}^2 while minimizing the number of variables, or model complexity. The

results for the two regional skew models that best met these criteria—the constant skew model denoted “CONSTANT” and the mean annual precipitation and drainage area model “PRECPRI00+DRNAREA” are shown in table B1.

The pseudo R_{δ}^2 describes the estimated fraction of the variability in the true skew from streamgauge-to-streamgauge explained by each model (Gruber and others, 2007; Parrett and others, 2011). A constant model does not explain any variability, so the pseudo R_{δ}^2 equals 0 for that model. The addition of basin characteristics in the PRECPRI00+DRNAREA model produced a pseudo R_{δ}^2 equal to 23 percent. This indicates that the inclusion of the PRECPRI00 and DRNAREA basin characteristics as explanatory variables in the regression only help to explain 23 percent of the streamgauge-to-streamgauge variability in the true skew. The addition of the PRECPRI00 and DRNAREA basin characteristics to the RSA1 regional skew model was not warranted as it resulted in a small improvement in model precision while increasing model complexity.

Thus, the CONSTANT model is chosen as the best regional skew model for the Alaska RSA1 study area. The posterior mean of the model error variance, σ_{δ}^2 , for the CONSTANT model is $\sigma_{\delta}^2 = 0.44$. The average sampling error variance (ASEV) in table B1 is the average error in the regional skew estimator at the sites in the dataset. The average variance of prediction at a new site (AVP_{new}) corresponds to the mean square error (MSE) used in Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) to describe the precision of the generalized skew. The CONSTANT model has an AVP_{new} equal to 0.45, which corresponds to an effective record length of 22 years.

Drainage areas for the 75 long-term streamgages that were used to develop the CONSTANT regional skew model for RSA1 ranged from 1.2 to 25,560 mi². The CONSTANT regional skew model is only applicable if the basin drainage area in RSA1 is greater than or equal to 1.2 mi² and less than or equal to 25,560 mi².

Regional Skew Area 2

All available basin characteristics were initially considered as explanatory variables in the regression analysis for regional skew in region RSA2. None of the basin characteristics were statistically significant in explaining the streamgage-to-streamgage variability in skew. Thus, the best model, as classified by having the smallest model error variance, σ_{δ}^2 , and largest pseudo R_{δ}^2 , is the constant model. The final results for the constant skew model denoted ‘‘CONSTANT’’ are shown in table B2.

The CONSTANT model is chosen as the best regional skew model for the Alaska RSA2 study area as none of the available basin characteristics were statistically significant in a model. The posterior mean of the model error variance, σ_{δ}^2 , for the CONSTANT model is $\sigma_{\delta}^2 = 0.10$. The CONSTANT model has an AVP_{new} equal to 0.12, which corresponds to an effective record length of 59 years.

As the 28 long-term streamgages used to develop the CONSTANT regional skew model for RSA2 had drainage areas ranging from 1.7 to 123 mi², the CONSTANT regional skew model is only applicable if the basin drainage area in RSA2 is greater than or equal to 1.7 mi² and less than or equal to 123 mi².

Table B1. Regional skewness models for Alaska Regional Skew Area 1 study area.

[Standard deviations are in parentheses. Bayesian plausibility (in percent) are in square brackets. σ_{δ}^2 , model error variance; ASEV, average sampling error variance; AVP_{new} , average variance of prediction for a new site; $pseudo-R_{\delta}^2$, fraction of the variability in the true skews explained by each model (Gruber and others, 2007)]

Model	Regression parameters			σ_{δ}^2	ASEV	AVP_{new}	$Pseudo-R_{\delta}^2$ (percent)
	β_1	β_2	β_3				
CONSTANT: $\hat{\gamma} = \beta_1$	0.54 (0.11)			0.44 (0.11)	0.013	0.45	0
PRECPRI00+DRNAREA: $\hat{\gamma} = \beta_1 + \beta_2 [\log_{10}(PRECPRI00)] + \beta_3 [\log_{10}(DRNAREA)]$	-1.8 (0.68)	1.4 (0.49) [0]	0.22 (0.09) [2]	0.33 (0.09)	0.031	0.36	23

Table B2. Regional skewness models for Alaska Regional Skew Area 2 study area.

[Standard deviations are in parentheses. σ_{δ}^2 , model error variance; ASEV, average sampling error variance; AVP_{new} , average variance of prediction for a new site; $pseudo-R_{\delta}^2$, fraction of the variability in the true skews explained by each model (Gruber and others, 2007)]

Model	Regression parameter	σ_{δ}^2	ASEV	AVP_{new}	$Pseudo-R_{\delta}^2$ (percent)
	β_1				
CONSTANT: $\hat{\gamma} = \beta_1$	0.18 (0.12)	0.10 (0.07)	0.014	0.12	0

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

To determine if a model is a good representation of the data and which regression parameters, if any, should be included in a regression model, diagnostic statistics have been developed to evaluate how well a model fits a regional hydrologic data set (Griffis, 2006; Gruber and others, 2008). In this study, the goal was to determine the set of possible explanatory variables that best fit annual peak discharges for the Alaska skew study areas RSA1 and RSA2 affording the most accurate skew predictions while also keeping the model as simple as possible. This section presents the diagnostic statistics for a B-WLS/B-GLS analysis, and discusses the specific values obtained for the Alaska RSA1 and RSA2 regional skew studies.

A Pseudo Analysis of Variance (Pseudo ANOVA) table for the Alaska RSA1 and RSA 2 regional skew analysis is shown in table B3, respectively. The table contains regression diagnostics/goodness of fit statistics. In particular, the table describes how much of the variation in the observations can be attributed to the regional model, and how much of the residual variation can be attributed to model error and sampling error,

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

For a model with no parameters other than the mean (that is, the constant skew model), the estimated model error variance $\sigma_\delta^2(0)$ describes all anticipated variation in $\gamma_i = \mu + \delta$, where μ is the mean of the estimated station sample skews. Thus, the total expected sum of squares variation due to model error δ_i and due to sampling error $\eta_i = \hat{\gamma}_i - \gamma_i$ in expectation should equal $n\sigma_\delta^2(0) + \sum_{i=1}^n Var(\hat{\gamma}_i)$. Therefore, the expected sum of squares attributed to a regional skew model with k parameters equals $n[\sigma_\delta^2(0) - \sigma_\delta^2(k)]$, because the sum of the model error variance $n\sigma_\delta^2(k)$ and the variance explained by the model must sum to $n\sigma_\delta^2(0)$. Table B3 considers a model with $k = 0$ (a constant model).

Table B3. Pseudo ANOVA table for the Alaska Regional Skew Area 1 CONSTANT model regional skew and the Alaska Regional Skew Area 2 CONSTANT model regional skew.

[EVR, error variance ratio; MBV*, misrepresentation of the beta variance, $pseudo-R_\delta^2$, fraction of variability in the true skews explained by the model]

Source	Degrees of freedom			Sum of squares		
	Equations	RSA1 CONSTANT model	RSA2 CONSTANT model	Equations	RSA1 CONSTANT model	RSA2 CONSTANT model
Model	k	0	0	$n[\sigma_\delta^2(0) - \sigma_\delta^2(k)]$	0	0
Model error	$n - k - 1$	74	27	$n[\sigma_\delta^2(k)]$	33	3
Sampling error	n	75	28	$\sum_{i=1}^n Var(\hat{\gamma}_i)$	22	6
Total	$2n - 1$	149	55	$n[\sigma_\delta^2(k)] \sum_{i=1}^n Var(\hat{\gamma}_i)$	55	9
Model diagnostics						
EVR					0.7	2.3
MBV*					1.3	1.2
$Pseudo-R_\delta^2$ (percent)					0	0

This division of the variation in the observations is referred to as a Pseudo ANOVA because the contributions of the three sources of error are estimated or constructed, rather than being determined from the computed residual errors and the observed model predictions, while also ignoring the impact of correlation among the sampling errors.

Table B3 contains the Pseudo ANOVA results for the RSA 1 CONSTANT model and the RSA2 CONSTANT model. The CONSTANT models do not have any explanatory variables, thus the variation attributed to the models is 0.

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

$$\text{EVR} = \frac{\text{SS}(\text{sampling error})}{\text{SS}(\text{model error})} = \frac{\sum_{i=1}^n \text{Var}(\hat{\gamma}_i)}{n\sigma_{\delta}^2(k)} \quad (\text{B10})$$

For the Alaska regional skew study areas, EVR had a value of 0.7 for the RSA1 CONSTANT model and 2.3 for the RSA2 CONSTANT model. The sampling variability in the sample skew estimators was larger than the error in the regional model. Thus an OLS model that neglects sampling error in the station skew estimators may not provide a statistically reliable analysis of the data. Given the variation of record lengths from site-to-site, it is important to use a WLS or GLS analysis to evaluate the final precision of the model, rather than a simpler OLS analysis.

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

$$\text{MBV}^* = \frac{\text{Var}[b_0^{\text{WLS}} | \text{GLS analysis}]}{\text{Var}[b_0^{\text{WLS}} | \text{WLS analysis}]} = \frac{w^T \Lambda w}{\sum_{i=1}^n w_i} \quad (\text{B11})$$

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

For the Alaska regional skew study areas, MBV* had a value of 1.3 for the RSA1 CONSTANT model and 1.2 for the RSA2 CONSTANT model. For both RSA1 and RSA2, the MBV* were larger than 1, indicating that the cross-correlation among the skew estimators had an impact on the precision with which the regional average skew coefficient can be estimated; if a WLS precision analysis were used for the estimated constant parameter in the CONSTANT model, the variance would be underestimated by a factor of 0.3 in the case of RSA1 and 0.2 in the case of RSA2. Thus, a WLS analysis would misrepresent the variance of the constant in the CONSTANT model in both RSA1 and RSA2. Moreover, a WLS model would have resulted in underestimation of the variance of prediction, given that the sampling error in the constant term in both models was sufficiently large enough to make an appreciable contribution to the average variance of prediction.

Leverage and Influence

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

For the B-WLS/B-GLS CONSTANT regional skew models for Alaska RSA1 and RSA2, no sites had high leverage. The differences in leverage values for the constant models reflect the variation in record lengths among sites.

In RSA1, six sites in the study (streamgages 15281000, 15212000, 15546200, 15291100, 15476400, and 15564887) have high influence, and thus have an unusual impact on the fitted regression relation. These six sites also have the six largest magnitude residuals in the study.

In RSA2, only one site in the study (streamgage 15098000) has high influence, and thus has an unusual impact on the fitted regression relation. Streamgage 15098000 has the highest influence value due to its large residual, the largest magnitude residual in the study; its unbiased station skew of 1.8 is the largest in the study.

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, Alaska Science Center
U.S. Geological Survey
4210 University Drive
Anchorage, Alaska 99508-4560
<http://alaska.usgs.gov>

