USGS
science for a changing world

Prepared in cooperation with the Federal Emergency Management Agency

# Methods for Determining Magnitude and Frequency of Floods in California, Based on Data through Water Year 2006


Scientific Investigations Report 2012–5113

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

**Cover.** Flooding on the Russian River near Hopland, California, on December 31, 2005. Photo taken on Highway 101 south of Hopland looking north. Photograph by Ken Markham, USGS (retired).

# Methods for Determining Magnitude and Frequency of Floods in California, Based on Data through Water Year 2006

By Anthony J. Gotvald, Nancy A. Barth, Andrea G. Veilleux, and Charles Parrett

Prepared in cooperation with the Federal Emergency Management Agency

Scientific Investigations Report 2012–5113

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

**U.S. Department of the Interior**
KEN SALAZAR, Secretary

**U.S. Geological Survey**
Marcia K. McNutt, Director

U.S. Geological Survey, Reston, Virginia: 2012



Suggested citation:
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., available online only at *http://pubs.usgs.gov/sir/2012/5113/*.

# Contents



# Plate

## Figures



## Tables

# Conversion Factors and Datums

Inch/Pound to SI

| Multiply | By | To obtain |
|---|---|---|
| | Length | |
| inch | 2.54 | centimeter (cm) |
| inch | 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 | |
| foot per mile (ft/mi) | 0.1894 | meter per kilometer (m/km) |
| cubic foot per second ($ft^3/s$) | 0.02832 | cubic meter per second ($m^3/s$) |

SI to Inch/Pound

| Multiply | By | To obtain |
|---|---|---|
| | Length | |
| centimeter (cm) | 0.3937 | inch |
| millimeter (mm) | 0.03937 | inch |
| meter (m) | 3.281 | foot (ft) |
| kilometer (km) | 0.6214 | mile (mi) |
| | Area | |
| square kilometer ($km^2$) | 0.3861 | square mile ($mi^2$) |
| | Volume | |
| cubic meter ($m^3$) | 35.31 | cubic foot ($ft^3$) |
| cubic meter ($m^3$) | 1.308 | cubic yard ($yd^3$) |
| cubic kilometer ($km^3$) | 0.2399 | cubic mile ($mi^3$) |
| | Flow rate | |
| cubic meter per second ($m^3/s$) | 35.31 | cubic foot per second ($ft^3/s$) |

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

$$°F = (1.8 \times °C) + 32$$

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

$$°C = (°F - 32) / 1.8$$

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 refers to distance above or below NAVD 88.

Water year is the 12-month period October 1 through September 30 and is designated by the calendar year in which the period ends. Thus, the water year ending September 30, 2001, is called “water year 2001.”

# Acronyms

| | |
|---|---|
| AEP | annual exceedance probability |
| APS | all possible subsets |
| EMA | expected moments algorithm |
| EVR | error variance ratio |
| FEMA | Federal Emergency Management Agency |
| GIS | geographic information system |
| GLS | generalized least squares |
| LP3 | log-Pearson Type 3 |
| MBV* | misrepresentation of the beta variance statistic |
| MSE | mean square error |
| NHDPlus | National Hydrologic Dataset |
| NLCD | National Land Cover Dataset |
| NWIS | National Water Information System |
| OLS | ordinary least squares |
| PRISM | Parameter-Elevation Regressions on Independent Slopes Model |
| USACE | U.S. Army Corps of Engineers |
| USGS | U.S. Geological Survey |
| WIE | weighted independent estimates |
| WLS | weighted least squares |
| WREG | weighted-multiple-linear regression |

# Methods for Determining Magnitude and Frequency of Floods in California, Based on Data through Water Year 2006

By Anthony J. Gotvald, Nancy A. Barth, Andrea G. Veilleux, and Charles Parrett

## Abstract

Methods for estimating the magnitude and frequency of floods in California that are not substantially affected by regulation or diversions have been updated. Annual peak-flow data through water year 2006 were analyzed for 771 streamflow-gaging stations (streamgages) in California having 10 or more years of data. Flood-frequency estimates were computed for the streamgages by using the expected moments algorithm to fit a Pearson Type III distribution to logarithms of annual peak flows for each streamgage. Low-outlier and historic information were incorporated into the flood-frequency analysis, and a generalized Grubbs-Beck test was used to detect multiple potentially influential low outliers. Special methods for fitting the distribution were developed for streamgages in the desert region in southeastern California. Additionally, basin characteristics for the streamgages were computed by using a geographical information system.

Regional regression analysis, using generalized least squares regression, was used to develop a set of equations for estimating flows with 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probabilities for ungaged basins in California that are outside of the southeastern desert region. Flood-frequency estimates and basin characteristics for 630 streamgages were combined to form the final database used in the regional regression analysis. Five hydrologic regions were developed for the area of California outside of the desert region. The final regional regression equations are functions of drainage area and mean annual precipitation for four of the five regions. In one region, the Sierra Nevada region, the final equations are functions of drainage area, mean basin elevation, and mean annual precipitation. Average standard errors of prediction for the regression equations in all five regions range from 42.7 to 161.9 percent.

For the desert region of California, an analysis of 33 streamgages was used to develop regional estimates of all three parameters (mean, standard deviation, and skew) of the log-Pearson Type III distribution. The regional estimates were then used to develop a set of equations for estimating flows with 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probabilities for ungaged basins. The final regional regression equations are functions of drainage area. Average standard errors of prediction for these regression equations range from 214.2 to 856.2 percent.

Annual peak-flow data through water year 2006 were analyzed for eight streamgages in California having 10 or more years of data considered to be affected by urbanization. Flood-frequency estimates were computed for the urban streamgages by fitting a Pearson Type III distribution to logarithms of annual peak flows for each streamgage. Regression analysis could not be used to develop flood-frequency estimation equations for urban streams because of the limited number of sites. Flood-frequency estimates for the eight urban sites were graphically compared to flood-frequency estimates for 630 non-urban sites.

The regression equations developed from this study will be incorporated into the U.S. Geological Survey (USGS) StreamStats program. The StreamStats program is a Web-based application that provides streamflow statistics and basin characteristics for USGS streamgages and ungaged sites of interest. StreamStats can also compute basin characteristics and provide estimates of streamflow statistics for ungaged sites when users select the location of a site along any stream in California.

# Introduction

Reliable estimates of the magnitude and frequency of floods are essential for flood insurance studies, flood-plain management, and the design of transportation and water-conveyance structures, such as roads, bridges, culverts, dams, and levees. Federal, State, regional, and local officials rely on these estimates to effectively plan and manage land use and water resources, protect lives and property in flood-prone areas, and determine flood-insurance rates. Griffis and Stedinger (2007a) determined that estimates of magnitude and frequency of floods using streamflow-gaging stations, hereafter referred to as streamgages, with shorter records of annual peak-flow data have higher standard errors or uncertainties when compared to estimates using streamgages with longer annual peak-flow records. Thus, long-term data collection at streamgages is important in the determination of reliable estimates of the magnitude and frequency of floods.

Estimates of the magnitude and frequency of floods are needed not only at locations where streamflow is monitored but also at ungaged basins where streamflow is not recorded. Therefore, other methods, such as regionalization, must be used to estimate the magnitude and frequency of floods at ungaged sites. Regionalization uses regression analysis to develop equations that relate flood-frequency information determined for a group of streamgages within a hydrologic region to various basin characteristics for the same streamgages. The resultant equations then can be used to estimate flood magnitude and frequency for ungaged sites within the hydrologic region.

## Purpose and Scope

The purpose of this report, prepared in cooperation with the Federal Emergency Management Agency (FEMA), is to present methods for estimating the magnitude and frequency of floods for streams in California. The report (1) describes the general statistical methods used to estimate the magnitude and frequency of floods for streamgages in California; (2) describes special methods used to analyze flood frequency for 33 streamgages in the desert region of southeastern California (specifically defined later in the report); (3) presents estimates of the magnitude of floods for the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probabilities determined for 769 streamgages in California (including 339 streamgages for which data were previously reported by Parrett and others (2011)); (4) describes methods used to develop regression equations to estimate the magnitude and frequency of floods for ungaged sites in California; (5) describes the accuracy and limitations of the equations; (6) shows example applications of the methods; (7) describes an analysis of the magnitude and frequency of floods for 8 streamgages in California that are affected by urbanization; and (8) describes the StreamStats Web application for automatically measuring required basin-characteristics data and solving the regression equations so that flood estimates can be quickly and easily obtained.

## Previous Studies

The earliest study of flood frequency of streams in California was done by Cruff and Rantz (1965), which compared methods used in flood-frequency studies for coastal basins in California. A series of reports entitled “Magnitude and Frequency of Floods in the United States” was published by the U.S. Geological Survey (USGS) as Water-Supply Papers. These reports provided summaries of flood data and presented methods for determining flood magnitude and frequency at ungaged sites. Data and methods used for the Great Basin are given in USGS Water-Supply Paper 1684 (Butler and others, 1966). Data for the Pacific slope basins are presented in two parts in Water-Supply Papers 1685 and 1686 (Young and Cruff, 1967; Young, 1967). Crippen and Beall (1971) developed methods for estimating various streamflow characteristics in California, including flood-frequency characteristics. The analysis included 385 streamgages, using data through 1967. Methods for estimating flood frequency for a small area in the San Bernardino and San Gabriel Mountains were described in a report by Busby and Hirashima (1972). Flood-frequency information for streamgages and methods for estimation of flood frequency at ungaged sites throughout California were developed and described in a report by Waananen and Crippen (1977). In addition, methods for estimating flood frequency in the desert regions of California were described in a report by Thomas and others (1997). The methods described by Thomas and others (1997) subsequently were updated for use in desert regions of California by Teal and Gusman (2007).

The regionalization methods described by Waananen and Crippen (1977) for use throughout California were based on data only through 1975 and thus may be unreliable given the 30 years of additional data now available. In addition, improved regionalization techniques have become available since the completion of previous reports. A study by Parrett and others (2011) began the process of updating flood-frequency estimates in California by describing the development of a method for estimating regional skew, a key component in the statistical analysis of gaged data. The method for determination of regional skew was used to update flood-frequency information for 364 streamgages, 206 of which are in the Sacramento–San Joaquin River Basin.

## Description of the Study Area

California is a state of widely varying topography and climate and consequently experiences a wide range of flood conditions. The State borders about 800 miles of the Pacific Ocean, and the seasonal variation in Pacific moisture gives California two distinct seasons—a wet winter and a dry summer. In addition, much of California is rugged and mountainous, with several major mountain ranges (Klamath Mountains, Cascade and Sierra Nevada, Coast Ranges, Transverse Range, and San Gabriel/San Bernardino Ranges) most of which roughly parallel the coastline and can disrupt the flow

of atmospheric moisture moving inland (fig. 1). As a result of generally differing atmospheric circulation patterns, the Pacific Ocean annually delivers more moisture into northern California than it does into southern California. Much of the southern part of the Great Basin physiographic region, together with the Salton Trough and Sonoran Desert physiographic regions, can be considered desert, largely as a result of smaller amounts of incoming Pacific moisture coupled with mountain barriers to the west that intercept much of the reduced precipitation source. Consequently, a specific desert region for flood-frequency analysis was delineated on the basis of physiographic regions shown in figure 1 and on desert regions previously delineated by Thomas and others (1997) and Teal and Gusman (2007).

As a result of the rugged and variable topography and differences in atmospheric moisture from the Pacific Ocean, mean annual precipitation in California ranges from about 3 inches in the desert region to more than 120 inches in the coastal mountains near the Oregon border. Large floods in California most often occur during the winter rainy season, although snowmelt floods commonly occur in the spring on larger streams draining the mountains. Convective rainstorms in the summer occasionally produce flooding on small streams throughout California.

# Data Compilation

The first step in the regionalization of flood-frequency estimates for streams is the compilation of streamgages with 10 or more years of annual peak-flow record. It is important that the peak-flow data are reviewed to assure quality of the records and homogeneity or absence of trends, which implies relatively constant watershed and climatic conditions during the period of record. Once peak-flow records are compiled and reviewed, then basin characteristics must be determined for each of the streamgages.

## Peak-Flow Data

Streamgages record the water-surface elevation, or stage, of a stream at various intervals, typically every 15 minutes, throughout the course of a water year. Streamflow, or discharge, is periodically measured throughout the range of recorded stages, and a relation between stage and discharge is developed for the streamgage. Using this stage-discharge relation, or rating, discharges for all recorded stages at the streamgage are determined. The largest discharge that occurs during a water year is the annual peak flow for the year, and the compilation of annual peak flows is the annual peak-flow record. The peak-flow records for streamgages are available from the USGS National Water Information System (NWIS) database at *http://nwis.waterdata.usgs.gov/usa/nwis/peak*.

Hundreds of streamgages in California were investigated for possible use in this study. Streamgages were only used in the analysis if 10 or more years of annual peak-flow data were available and if peak flows were not affected substantially by diversions or urbanization. The peak-flow record for streamgages that meet these criteria were then compiled and reviewed by using the PFReports computer program described by Ryberg (2008).

Parrett and others (2011) performed a monotonic analysis of 69 long-term peak-flow records outside the desert region of California using Kendall's tau, a non-parametric test for trends described by Helsel and Hirsch (1992). Trends are generally considered to be significant when the p-value is less than or equal to 0.05. A p-value of 0.05 indicates that there is a 5 percent probability that the test will identify a trend when no actual trend is present. Parrett and others (2011) determined that monotonic trends in peak-flow record are not considered to be a factor anywhere in California outside the desert region. For this study, six long-term streamgages in the desert region that had complete annual peak-flow records from 1967 to 2006 (40 years) were analyzed using Kendall's tau test and also were found to have no significant trends in annual peak flow. The 6 streamgages are representative of all 33 streamgages used for flood-frequency analysis in the desert region of California.

For the streamgages with regulation, if 10 or more years of pre-regulation peak-flow record were available, then the pre-regulation portion of the record was considered for this study. Also, 14 streamgages below dams selected by the U.S. Army Corps of Engineers (USACE) that are detailed in Parrett and others (2011) were considered for use in this study. The unregulated peak-flow record for these streamgages were estimated using methods described in Parrett and others (2011). The peak-flow record review resulted in the selection of 858 streamgages that were considered for use in this study.

## Physical and Climatic Basin Characteristics

Peak-flow information can be estimated at ungaged sites through a multiple regression analysis that develops a relation between peak-flow characteristics (such as 1-percent annual exceedance probability flow) and selected physical and climatic basin characteristics for gaged drainage basins. Selected basin characteristics for each of the 858 streamgages considered for use in this study were derived from various national geo-spatial datasets, including the National Hydrologic Dataset (NHDPlus), the 2001 National Land Cover Dataset (NLCD), and the Parameter-Elevation Regressions on Independent Slopes Model (PRISM) climatic dataset, which is based on data from 1971 to 2000. Basin-characteristic names, descriptions, units, and sources of information considered for this study are given in table 1. At most of the streamgages, the drainage area determined from the NHDPlus dataset closely matched the drainage area manually determined from topographic maps and reported in the NWIS peak-flow database. The NHDPlus dataset is based on relatively coarse digital elevation data (30 meter), however, and may not always provide accurate basin delineations, particularly for small basins in flat areas with little topographic relief. In addition, the NHDPlus dataset

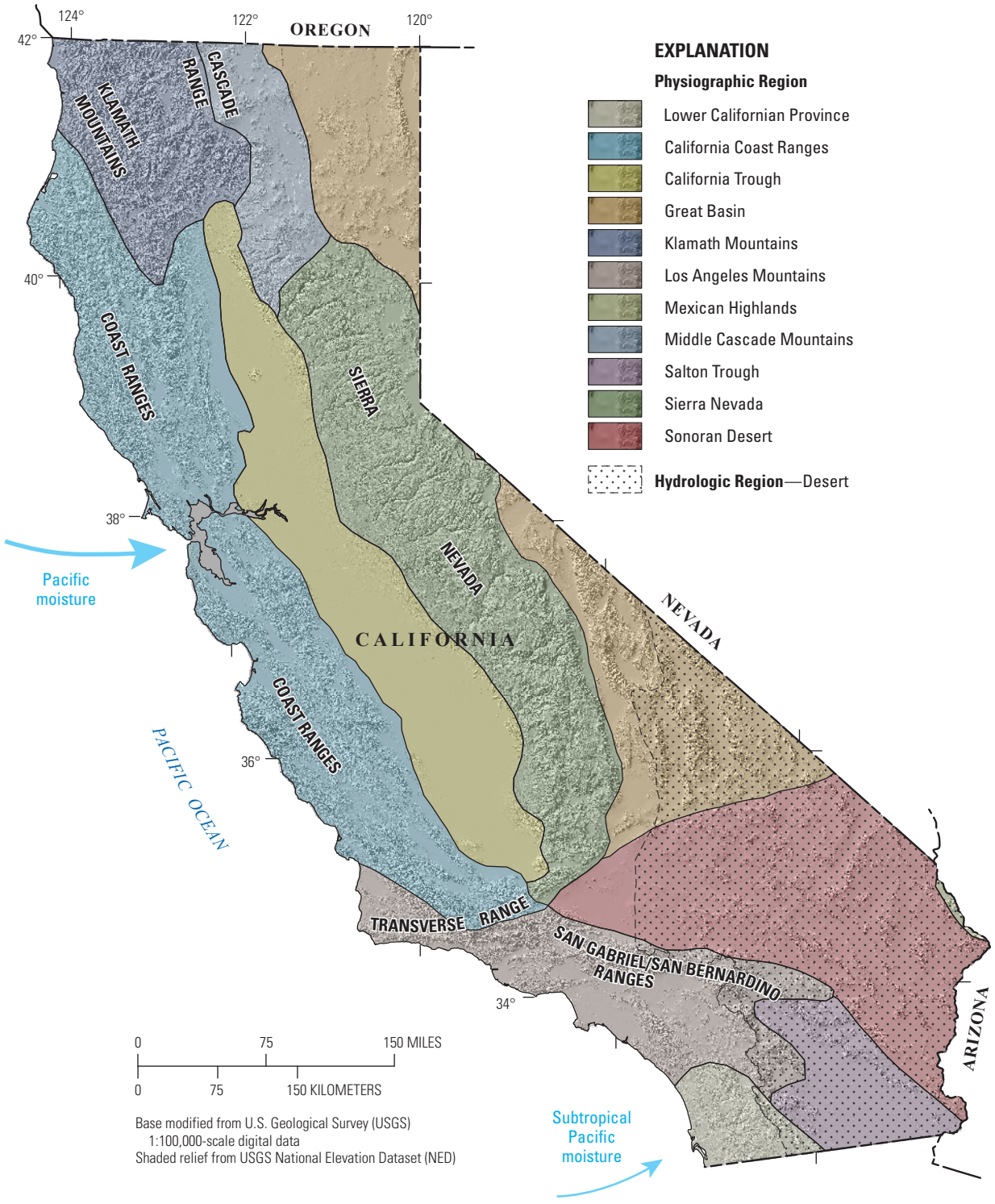


**Figure 1.** Physiographic regions in California (Fenneman and Johnson, 1946).

**Table 1.** Basin characteristics considered for use in regional regression analysis for California.

[DEM, digital elevation model; NHDPlus, National Hydrography Dataset Plus; PRISM, Parameter-elevation Regressions on Independent Slopes Model]

| Name | Description | Unit | Data source |
|---|---|---|---|
| DRNAREA | Drainage area of the basin | Square miles | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| BASINPERIM | Perimeter of the basin | Miles | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| RELIEF | Difference between maximum and minimum elevations in the basin | Feet | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| ELEVMAX | Maximum elevation in the basin | Feet | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| ELEVMIN | Minimum elevation in the basin | Feet | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| LAKEAREA | Percentage of basin area covered by lakes and ponds | Percent | 2001 National Land Cover Database (NLCD)—Land Cover *http://www.mrlc.gov/nlcd2001.php* |
| EL6000 | Percentage of basin area above an elevation of 6,000 feet | Percent | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| OUTLETELEV | Elevation at the outlet of the basin | Feet | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| RELRELF | Basin relief divided by the basin perimeter | Feet per mile | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| DIST2COAST | Distance from basin centroid to coast along a line perpendicular to eastern California border | Miles | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| ELEV | Average basin elevation | Feet | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| BSLDEM30M | Average basin slope | Percent | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| FOREST | Percentage of basin area covered by forest | Percent | 2001 National Land Cover Database (NLCD)—Percent Canopy *http://www.mrlc.gov/nlcd2001.php* |
| IMPERV | Percentage of basin area covered by impervious surface | Percent | 2001 National Land Cover Database (NLCD)—Percent Impervious *http://www.mrlc.gov/nlcd2001.php* |
| PRECIP | Mean annual precipitation | Inches | 800-meter resolution PRISM 1971–2000 data *http://www.prism.oregonstate.edu/products/* |
| JANMAX | Average maximum January temperature in the basin | Degrees Fahrenheit | 800-meter resolution PRISM 1971–2000 data *http://www.prism.oregonstate.edu/products/* |
| JANMIN | Average minimum January temperature in the basin | Degrees Fahrenheit | 800-meter resolution PRISM 1971–2000 data *http://www.prism.oregonstate.edu/products/* |
| LONG_CENT | Longitude of the basin centroid | Degrees | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |
| LAT_CENT | Latitude of the basin centroid | Degrees | 30-meter DEM, NHDPlus elev_cm grid *http://www.horizon-systems.com/NHDPlus/* |

was found to have errors in the stream network that resulted in drainage basin errors in portions of the upper Sacramento River Basin, including some parts of the Pit and Feather River drainages. The basin characteristics computed using a geographic information system (GIS) were used for all streamgages in the study, but some streamgages in the upper Sacramento River Basin were eliminated from further analysis because of potential inaccuracies in the computed basin characteristics. Streamgages were eliminated from further analysis if the GIS-calculated drainage areas differed substantially from those published in the NWIS peak-flow database. Differences in drainage area were considered substantial if they exceeded 50 percent for drainage basins smaller than 0.1 square mile, 20 percent for drainage basins between 0.1 and 1 square mile in size, and 10 percent for basins greater than 1 square mile in size. These criteria resulted in 771 streamgages that were considered for further use in the study (pl. 1; table 2). Streamgages that were excluded from analysis because they did not meet the criteria for drainage area differences included 23 streamgages for which flood-frequency characteristics were previously reported by Parrett and others (2011). Because of the limitations of the NHDPlus dataset, users are cautioned to verify basin boundaries and drainage areas computed at ungaged sites to ensure that results are reasonable.

**Table 2.** Summary of streamgages in California that were considered for use in the regional regression analysis, 2006.

[Table 2 is available in a Microsoft© Excel spreadsheet and can be accessed and downloaded at *http://pubs.usgs.gov/sir/2012/5113/*]

# Flood Magnitude and Frequency at Streamgages

A frequency analysis of annual peak-flow data collected at a streamgage provides an estimate of the flood magnitude and frequency at that specific stream site. Flood-frequency flows were described in previous USGS reports as T-year floods based on the recurrence interval for the flood quantile (for example, the "100-year flood"). The use of recurrence-interval terminology is now discouraged because it can be confusing to the general public. The term has been interpreted to imply a set time interval between floods of a particular magnitude, when in fact floods are random processes that are best understood by using probabilistic terms. While the T-year recurrence interval flood is statistically expected to occur or be exceeded, on average, once during the T-year period, it may be equaled or exceeded multiple times during the period or not at all.

Terminology associated with flood-frequency estimates is shifting away from the T-year recurrence interval flood to the P-percent annual exceedance probability (AEP) flood. The use of percent AEP flood is now preferred because it conveys the probability, or odds, of a flood of a given magnitude being equaled or exceeded in any given year. For example, a 1-percent AEP flood (formerly known as the "100-year flood") corresponds to the flow magnitude that has a 0.01 probability of being equaled or exceeded in any given year. The P-percent is computed as the reciprocal of the recurrence interval "T" multiplied by 100 (for example, $1/100 \times 100 = 1$ percent). T-year recurrence intervals with corresponding percent AEPs are listed in table 3 (Gotvald and others, 2009).

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

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

Flood-frequency estimates for streamgages are computed by fitting a known statistical distribution to the series of annual peak flows. The statistical distribution commonly used in the United States is the log-Pearson Type III distribution (hereafter referred to as the LP3 distribution). Guidelines and computational methods for using the LP3 distribution are described in Bulletin 17B of the Hydrology Subcommittee of the Interagency Advisory Committee on Water Data (1982). General procedures for fitting the LP3 distribution, the expected moments algorithm (EMA), a new method for statistically detecting multiple potentially influential low outliers when fitting the LP3 distribution, a special application of the LP3 method developed for the desert region of California, and a trial application of a mixed-population flood-frequency analysis for high-elevation streamgages in the California mountains are described in the following sections of the report. The final flood-frequency estimates from the LP3 analysis for the 771 streamgages in California considered in this study are given in table 4. Flood-frequency estimates could not be computed for station 11067000 Day Creek near Etiwanda, Calif. (map identification number 122), because of the uncertainty of debris flow effects on some of the higher annual peak flows. Also, estimates could not be computed for station 11142500 Arroyo De La Cruz near San Simeon, Calif. (map identification number 219), because of the uncertainty of the indirect-discharge measurements used to determine the larger annual peak flows.

**Table 4.** Flood-frequency statistics for streamgages in California that were considered for use in the regression equations, 2006.

[Table 4 is available in a Microsoft© Excel spreadsheet and can be accessed and downloaded at *http://pubs.usgs.gov/sir/2012/5113/*]

## General Log-Pearson Type III Frequency Analysis

Flood-frequency estimates for the streamgages outside of the desert region were computed by fitting the LP3 distribution to the logarithms (base 10) of the annual peak flows as described in Bulletin 17B (Interagency Advisory Committee on Water Data, 1982). Fitting the distribution requires calculating the mean, standard deviation, and skew coefficient of the logarithms of the annual peak-flow record, which describe the mid-point, slope, and curvature of the peak-flow frequency curve, respectively. Estimates of the P-percent AEP flows are computed by inserting the three statistics of the frequency distribution into the equation:

$$\log Q_P = \overline{X} + K_P S \,, \tag{1}$$

where

$Q_P$ is the P-percent annual exceedance probability flow, in cubic feet per second ($ft^3/s$);

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

$K_P$ is a factor based on the skew coefficient and the given percent annual exceedance probability and is obtained from appendix 3 in Bulletin 17B; and

$S$ is the standard deviation of the logarithms of the annual peak flows, which is a measure of the degree of variation of the annual values about the mean value.

The mean, standard deviation, and skew coefficient can be estimated from the available sample data (recorded annual-peak flows), but a skew coefficient calculated from small samples tends to be an unreliable estimator of the population skew coefficient. Accordingly, the guidelines in Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) indicate that the skew coefficient calculated from at-site sample data (station skew) needs to be weighted with a generalized, or regional, skew determined from an analysis of selected long-term streamgages in the study region. The value of the skew coefficient used in equation 1 is the weighted skew that is based on station skew and regional skew. The station skew coefficients for the streamgages outside of the desert region were weighted with the generalized skew coefficients developed by Parrett and others (2011).

A series of annual peak flows at a streamgage may include statistically determined outliers, which are annual peak flows that are substantially lower or higher than other peak flows in the series. The peak-flow record also may include information about peak flows that occurred outside of the period of collected data, also called systematic record. These peak flows are known as historical peak flows and usually are considered to have been the largest peak flows during an extended period of time that is longer than the systematic record. Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) provides guidelines for detecting outliers and interpreting historical data points and provides computational methods for appropriate corrections to the distribution to account for the outliers and historical information. While these adjustments generally improve flood-frequency estimates, the EMA method incorporates censored flows (high and low outliers) and historical flows more efficiently (Cohn and others, 1997) than the methods outlined in Bulletin 17B (Interagency Advisory Committee on Water Data, 1982).

## Expected Moments Algorithm (EMA)

The EMA method was used for all sites in this study to determine LP3 at-site frequency estimates. For sites that have systematic annual peak-discharge records for complete periods, no low outliers, and no historical flood information, the EMA method calculates identical values of the LP3 parameters (mean log, standard deviation log, and station skew) as the conventional method of moments described in Bulletin 17B. The EMA method, however, can incorporate censored and interval peak-discharge data into the analysis. Censored data may be expressed in terms of discharge perception thresholds that are most often used during historical periods outside the period of systematic data collection. For example, a site may have historical information that indicates that a recorded peak discharge, $Q_{hist}$, was the largest since 1900, before systematic data collection was started in 1930. Each annual peak from 1900 to 1929 can be characterized as a censored discharge for which the value is known not to have exceeded the perception threshold, $Q_{hist}$, and estimates of those bounded discharges between 0 and $Q_{hist}$ can be used in the LP3 flood-frequency analysis. The EMA method also allows use of interval discharges to characterize peak flows that are known to be greater or less than some specific value or that can only be reliably estimated within a specific range in discharge. Interval discharges commonly are used by the EMA method to characterize missing data during periods of systematic data collection. For example, if a peak discharge was not determined because the water level did not reach the bottom of the gage, the missing peak can be characterized as an interval discharge with a range that is bounded by zero and the discharge associated with the elevation of the bottom of the gage. Missing peaks during periods of systematic data collection typically are ignored when the conventional LP3 method is used (Parrett and others, 2011).

## Multiple Grubbs-Beck Test for Detecting Low Outliers

The Grubbs-Beck test is recommended in Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) for detecting low outliers that can be subsequently censored in EMA so they do not have a large influence on the fitting of the upper tail (that is, larger flows with smaller AEPs) of the LP3 distribution. The Grubbs-Beck test uses the at-site logarithms of the peak-flow data to calculate a one-sided, 10-percent significance-level critical value for a normally distributed sample. Although more than one recorded peak flow for a streamgage may be smaller than the Grubbs-Beck critical value, usually only one non-zero recorded peak flow is identified from the test as being a low outlier. As described by Parrett and others (2011), many streamgages in California have several annual peak flows that are substantially smaller than most of the recorded annual peak flows, and the Grubbs-Beck test identifies only one or even no non-zero low outliers for these streamgages. Consequently, visual inspections of plotted flood-frequency curves were used by Parrett and others (2011) to identify all small peak flows that had an unduly large influence on the upper tail of the fitted curve. Selection of a low-outlier censoring threshold that eliminated the small peak flows typically resulted in a significantly better frequency curve fit to the larger peak flows. In some instances, these user-selected low-outlier thresholds resulted in 50 percent of the recorded annual peak flows being considered as low outliers.

Since completion of the report by Parrett and others (2011), a method for statistically detecting multiple potentially influential low outliers using a generalized Grubbs-Beck test was developed (T.A. Cohn, U.S. Geological Survey, written commun., February 2011). The multiple Grubbs-Beck test is also based on a one-sided, 10-percent significance-level critical value for a normally distributed sample, but the test is constructed so that groups of ordered data are examined (for example, the eight smallest values) and excluded from the dataset when the critical value is calculated. If the critical value is greater than the eighth smallest value in the example, then all eight values are considered to be low outliers. As described by Cohn (T.A. Cohn, U.S. Geological Survey, written commun., February 2011), the low outliers identified by the multiple Grubbs-Beck test closely match user-selected low-outlier thresholds determined from plotted flood-frequency curves. The multiple Grubbs-Beck test was used for this study, but user-selected low-outlier thresholds determined in the previous flood-frequency study for California (Parrett and others, 2011) were not changed. The streamgages that had multiple low outliers determined from a user-selected threshold or the new multiple Grubbs-Beck test are noted in table 2. An example of a flood-frequency curve for a streamgage with the complete lower tail of the distribution (50 percent of all recorded annual peak flows) identified and subsequently censored as low outliers is shown in figure 2. The shape of the resultant LP3 curve would have been significantly different if all low outliers had not been censored.

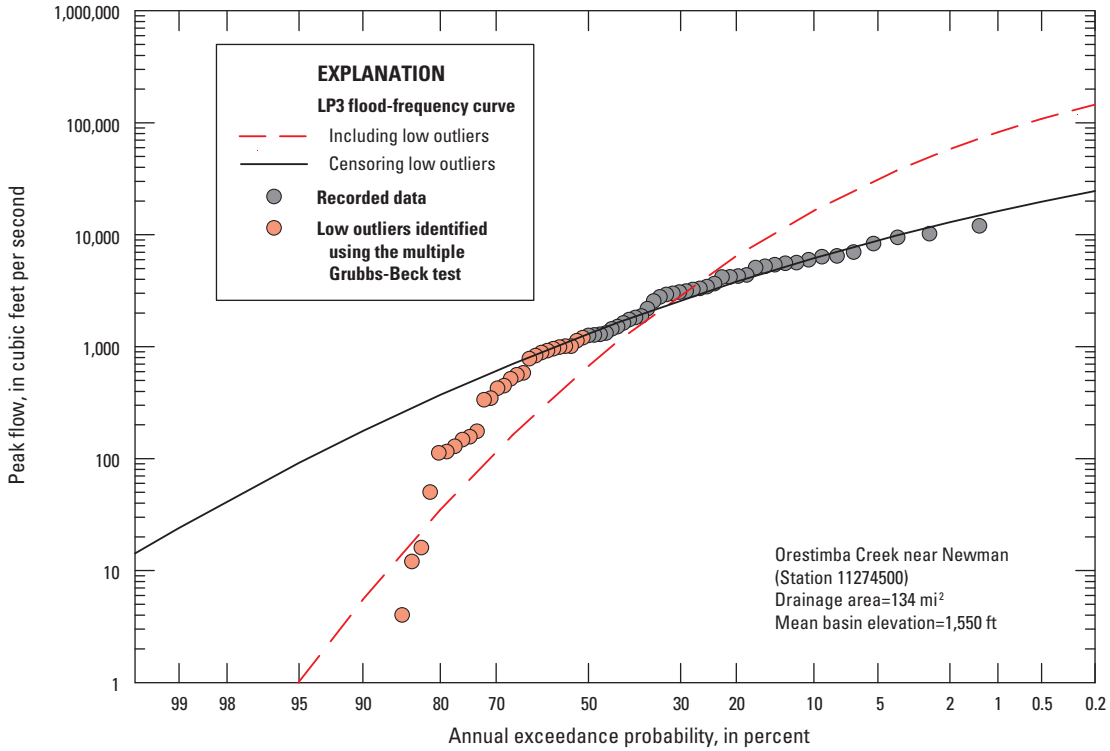


**Figure 2.** Flood-frequency curves for Orestimba Creek near Newman, California (station 11274500), showing the effects of including or censoring potentially influential low outliers identified from the multiple Grubbs-Beck test.

## Parameter Estimation Method for Frequency Analysis in the Desert Region

Flood-frequency analysis in the California desert is complicated because of short annual peak-flow records (usually less than 20 years) and numerous zero flows and (or) low outliers for many streamgages. Estimates of the three parameters (mean, standard deviation, and skew) required for fitting the LP3 distribution are likely to be highly unreliable based on the limited and heavily censored at-site data. Although the LP3 distribution was previously used to determine at-site flood frequency in the California desert (Waananen and Crippen, 1977), two more recent studies (Thomas and others, 1997; Teal and Gusman, 2007) used a hybrid method based on pooled at-site peak-flow data from similar sized basins and a plotting-position method for determining flood frequency.

For this study, a generalization of the recommendations in Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) was used to develop a method based on the use of the LP3 distribution and regional estimates for all three parameters (mean, standard deviation, and skew) to determine flood-frequency estimates in the desert. As described in Bulletin 17B (Interagency Advisory Committee on Water Data, 1982), flood-frequency estimates are improved by weighting at-site skew with more robust estimates of regional skew. Because of the at-site data limitations in the desert, flood-frequency estimates are believed to be more robust and reliable if the at-site mean and standard deviation also are weighted with regional estimates of those parameters. Consequently, regional regression models for the mean and standard deviation developed using weighted least squares (WLS) regression, together with a previously developed model for regional skew (Thomas and others, 1997), and the appropriate model error metrics for weighting purposes were used to compute the at-site flood-frequency estimates for the streamgages in the desert region. The EMA program (PeakfqSA, version 0.972) was modified to enable the weighting of at-site mean and standard deviation in a similar fashion to the weighting of at-site and regional skew (T.A. Cohn, U.S. Geological Survey, written commun., October 2011).

Thirty-three streamgages in the desert region had 10 or more years of recorded annual peak-flow data that were essentially unregulated and acceptable for flood-frequency analysis. Figure 3 shows the number of zero and non-zero annual peak flows for the 33 desert streamgages and indicates that about half the streamgages had record lengths of less than 20 years. In addition, figure 3 indicates that many of the desert streamgages had one or more zero peak flows in

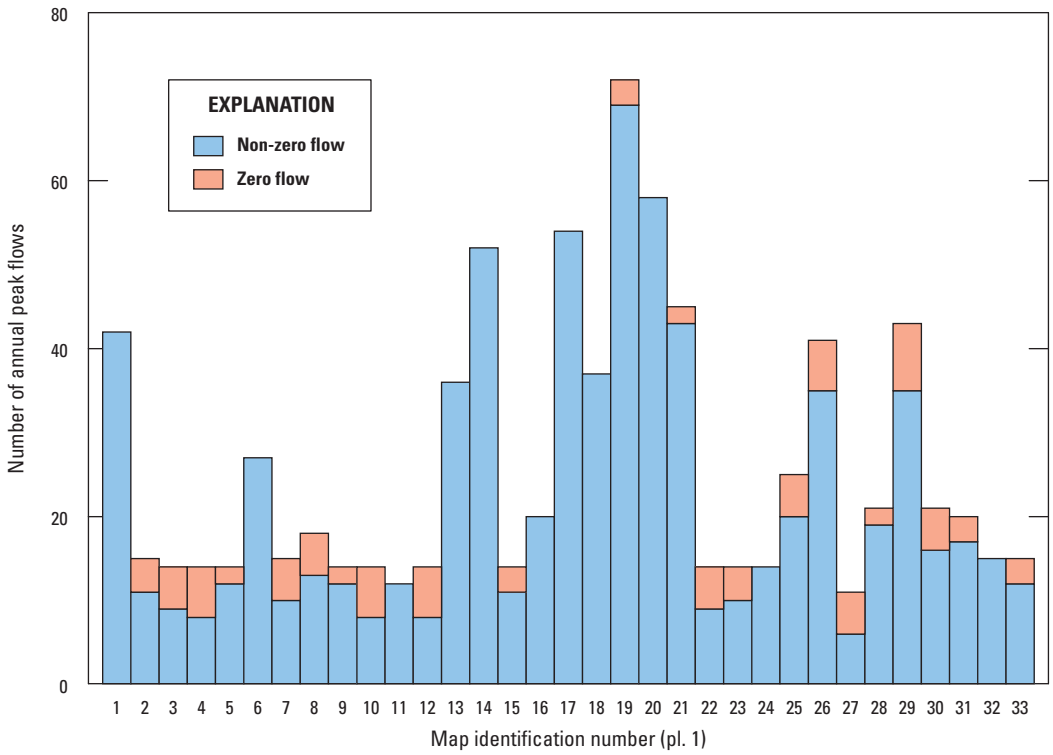


**Figure 3.** Number of zero and non-zero annual peak flows for streamgages used in the regional regression analysis in the California desert region.

the recorded data. Use of the EMA method with the multiple Grubbs-Beck test identified potentially influential low outliers at many desert streamgages in addition to the zero peak flows. Figure 4 shows the number of potentially influential low outliers that were identified and subsequently censored as low outliers for each streamgage, and figure 5 shows the percentage of annual peak-flow data identified as potentially influential low outliers per streamgage. This extensive censoring of potentially influential low-outlier data has a large effect on the initial at-site values for the mean, standard deviation, and skew for the desert streamgages. These initial at-site values were used to develop regional estimates of the mean and standard deviation using regression and were also subsequently weighted with those regional values to calculate the final values of at-site LP3 parameters. The details of the procedures and required mathematics for determining the final at-site flood-frequency estimates for the 33 streamgages in the desert are described in the appendix. The final flood-frequency estimates from the modified Bulletin 17B analysis for the 33 streamgages in the desert region are given in table 4.

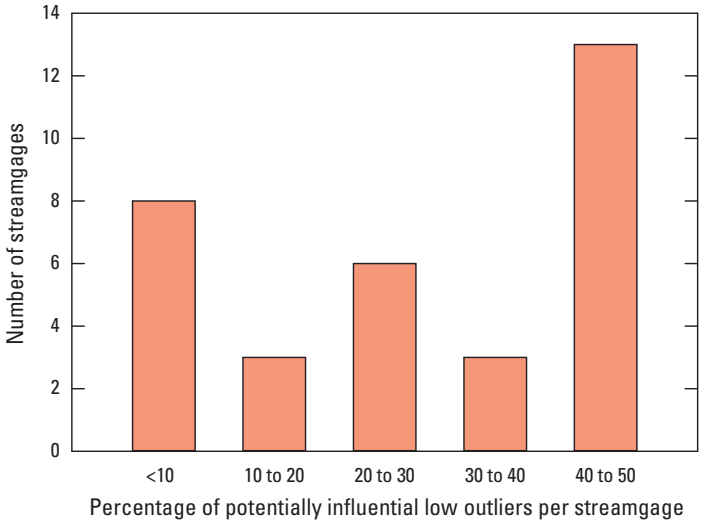


**Figure 5.** Distribution of the percentage of annual peak-flow data identified as potentially influential low outliers for each streamgage in the California desert region.

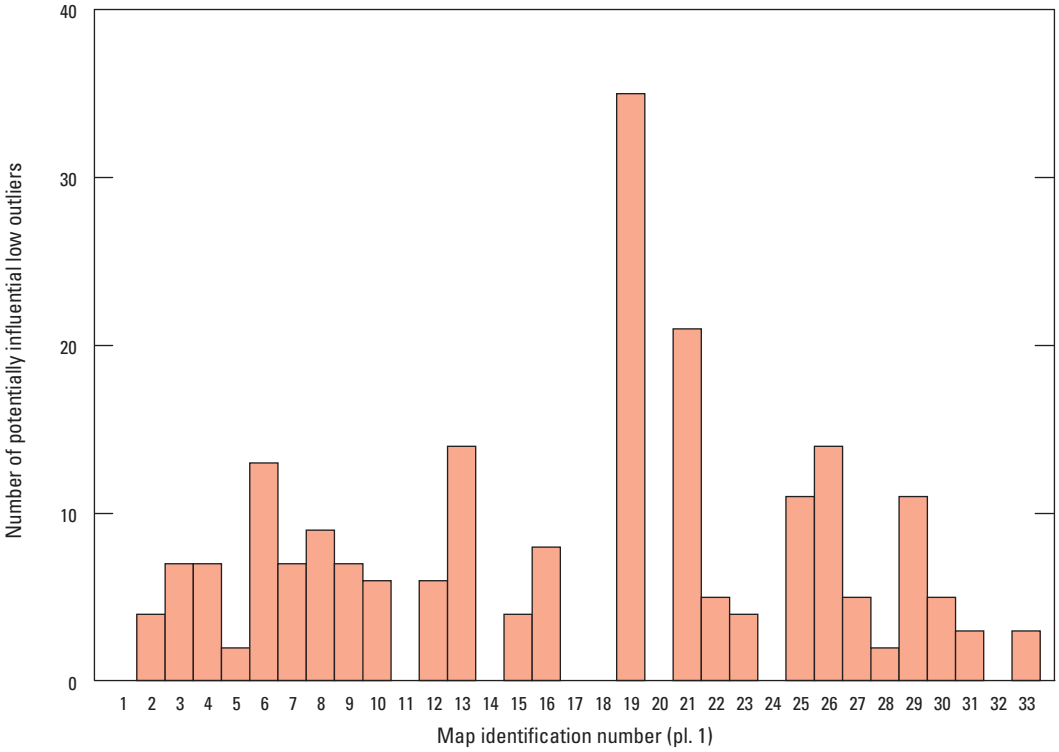


**Figure 4.** Number of potentially influential low outliers identified by the multiple Grubbs-Beck test for the streamgages in the California desert region.

## Trial Mixed-Population Frequency Analysis

As described by Parrett and others (2011), annual peak flows at streamgages in mountainous areas may be caused by winter rainstorms, springtime snowmelt runoff, or some combination of snowmelt mixed with rainstorm runoff. The number of annual peak flows that are predominantly caused by snowmelt runoff tends to increase with increasing mean basin elevation. At most mountainous streamgages in California, a single LP3 flood-frequency distribution applied to all the annual peak flows provides a reasonable fit to the data. At 10 streamgages in the Sierra Nevada region, however, the LP3 fit to all recorded peak flows deviated substantially from recorded peak flows in the upper tail (large annual peak flows), and a mixed-population flood-frequency analysis was considered for those streamgages. Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) also described the Sierra Nevada region of California as an example region where mixed-population analyses of rain floods and snowmelt floods might be warranted.

At each of the 10 streamgages, the recorded annual peak flows were separated into two groups: those assumed to be predominantly caused by rain and those assumed to be predominantly caused by snowmelt. Separate LP3 curves were developed for the rain-caused peak flows and the snowmelt-caused peak flows, and the separate curves were statistically combined using conditional probability calculations. Unfortunately, six of the streamgages had only a small number (10 or less) of rain-caused peak flows, and the LP3 curves for the rain-caused peak flows at those streamgages were considered unreliable. At the four streamgages having 10 or more rain-caused peak flows, the calculated at-site skews for the rain-caused floods varied from –1.1 to 0.5, and the resultant combined LP3 curves for those streamgages also were considered to be unreliable. The rain-caused peak flows at all 10 streamgages were made dimensionless and were pooled together in an attempt to develop a regional LP3 curve for rain-caused peak flows that could be applied to all 10 streamgages, but that also produced inconsistent combined flood-frequency curves at several streamgages. The LP3 curves based on all recorded annual peak-flow data, therefore, were considered to be at least as reliable as the curves based on the mixed-population trial approach and were used for all streamgages in the study. The resultant LP3 curve for one of the trial mixed-population sites is shown in figure 6. The LP3 curve for this site represents the poorest fit to the recorded data for all 10 trial mixed-population sites. Despite the relatively poor fit to some of the recorded peak flows, the LP3 curve is considered to provide the most reliable estimates of flood frequency at this site given the general uncertainty about mixed populations and the analysis with the limited at-site data available.

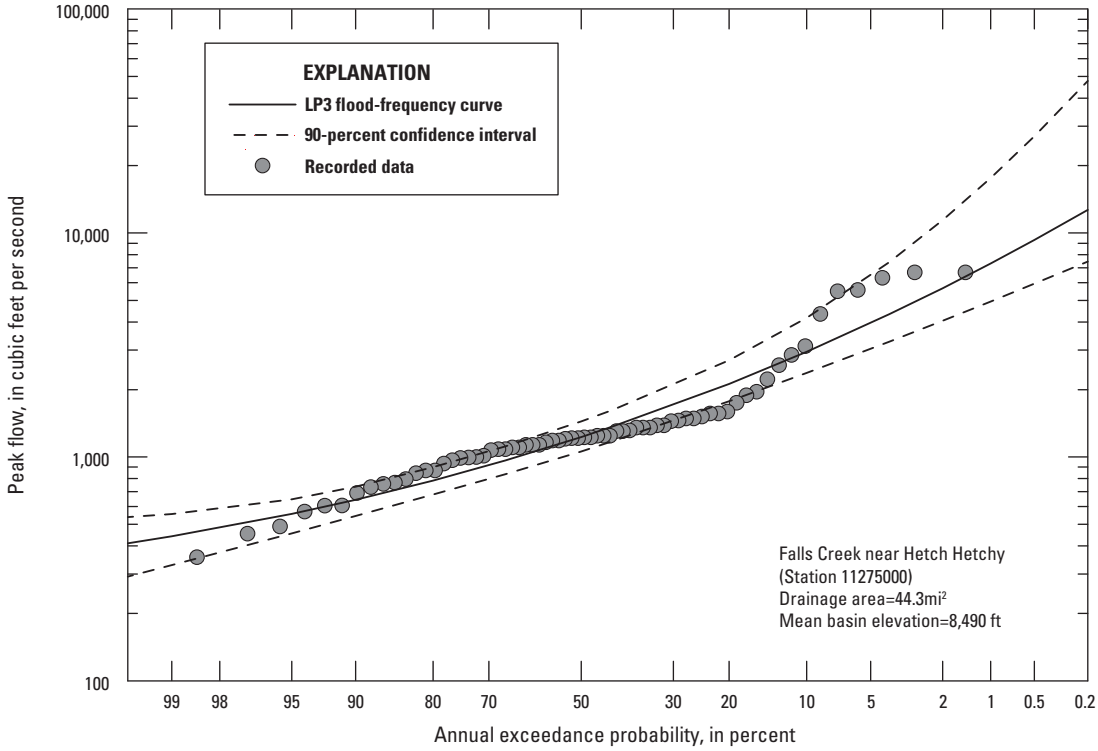


**Figure 6.** Flood-frequency curve for Falls Creek near Hetch Hetchy, California (station 11275000).

# Estimation of 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 at ungaged sites in California. 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 to measured basin characteristics of the associated drainage basins. All 769 streamgages for which flood-frequency and basin characteristics had been determined were considered for use in the regional regression analysis (table 4).

## Regression Analysis

Ordinary least squares (OLS) regression techniques were used in the exploratory analysis to determine the best candidate regression models for all combinations of basin characteristics and the development of hydrologic regions with differing flood-frequency characteristics. Because an OLS regression uses a linear relation between the explanatory (basin characteristics) and response variables (P-percent AEP flows), variables may have to be transformed in order to create linear relations. For example, the relation between the arithmetic values of basin drainage area and P-percent AEP flow typically is curvilinear; however, the relation between the logarithms of basin drainage area and the logarithms of P-percent AEP flow typically is linear. Homoscedasticity (a constant variance in the response variable over the range of the explanatory variables) and normality of the residuals also are requirements for an OLS regression. The logarithmic transformation of the P-percent AEP flow and explanatory variables enhances the homoscedasticity of the data. Homoscedasticity and normality of residuals were examined graphically.

Selection of the explanatory variables for each hydrologic region was based on all-possible-subsets (APS) regression methods (Neter and others, 1985). The final selection of explanatory variables for inclusion into each model for each hydrologic region was based on several factors, including standard error of the estimate, Mallow's $C_p$ statistic, statistical significance of the explanatory variables, coefficient of determination ($R^2$), and ease of computing the basin characteristics. Multicollinearity (correlation among the candidate explanatory variables) also was assessed by the variance inflation factor (*VIF*).

In all regions except the desert region, generalized least square (GLS) regression methods, as described by Stedinger and Tasker (1985), were used to determine the final regional P-percent AEP flow regression equations with the use of the weighted-multiple-linear regression (WREG) program, version 1.03 (U.S. Geological Survey, 2010). Details on this computer program are described by Eng and others (2009). Stedinger and Tasker (1985) found that GLS regression equations are more accurate and provide a better estimate of the regression accuracy than the simpler OLS regression equations when annual peak-flow records at streamgages are of different and widely varying lengths and when concurrent flows at different streamgages are correlated. The GLS regression techniques give less weight to streamgages that have shorter periods of record than to streamgages with longer periods of record. Less weight is also given to streamgages where concurrent peak flows are correlated because of the geographic proximity to other streamgages (Hodgkins, 1999). For the desert region of California, however, regression analysis was not used to relate P-percent AEP flows to basin characteristics; rather, a WLS regression analysis (Tasker, 1980) was used to develop regional estimates of the mean and standard deviation. These regional estimates, together with a regional estimate of skew from a previous report (Thomas and others, 1997) were used in the basic LP3 flood-frequency equation (eq. 1) to develop estimation equations for P-percent AEP flows. Details on the regional regression analysis for the desert region are provided in the appendix.

Regression analysis requires that data be as spatially independent as possible. Redundancy results when the drainage basins of two streamgages are nested, meaning that one is contained inside the other, and the sizes of the two basins are similar. Then, instead of providing two independent spatial observations depicting how basin characteristics are related to AEP flows, these two basins will likely have the same hydrologic response to a given storm and thus represent only one spatial observation. A statistical analysis using redundant streamgages misrepresents the information in the regional dataset (Gruber and Stedinger, 2008). In order to remove the errors associated with nested streamgages for the regional regression analysis, the methods detailed in Veilleux (2009) and Parrett and others (2011) were used to determine the redundant streamgages for this study. Of the 769 streamgages, 104 were omitted from the regional regression analysis because of redundant record, leaving a total of 665 streamgages for further regional analysis.

## Regionalization of Flood-Frequency Estimates

Because the streamgages in the desert region of California required special at-site flood-frequency analysis due to the extreme flow variability and large number of censored low annual peak flows (usually zero), a hydrologic region consisting of the desert streamgages was developed using the desert regions from Waananen and Crippen (1977) and Thomas and others (1997). An OLS regression analysis was performed on the 632 streamgages outside of the desert region to determine whether additional hydrologic regions needed to be determined for California. All response and non-zero explanatory variables were transformed to logarithms (base 10) prior to the regression analyses to (1) obtain linear relations between the response variables and the explanatory

variables and (2) achieve homoscedasticity. The standard errors of estimate using varying combinations of explanatory variables ranged from 86.0 to 99.3 percent for the 1-percent AEP flow estimate when using only one hydrologic region outside the desert region. Regression residuals for the 1-percent AEP flows were plotted at the centroid of the respective drainage basin in order to determine geographical patterns of bias. Large errors of estimate and geographic bias of the regression residuals indicated that California needed to be subdivided into hydrologic regions. The physiographic regions (fig. 1) and the hydrologic regions from the Waananen and Crippen (1977) study were used together with the observed patterns of regression residuals to develop the hydrologic regions for the area of California outside of the desert region. A total of six hydrologic regions, including the desert region, were developed for California (pl. 1).

In addition to the six hydrologic regions determined suitable for development of regression equations, a region including only two gages was delineated as an indeterminate region for flood-frequency estimation. This indeterminate region that includes Mono Lake and the upper Owens River valley is generally high in elevation but also generally dry. This region is outside the area for which regional skew was determined in California (Parrett and others, 2011), and it is within the general desert region identified by Thomas and others (1997) for which regional skew was determined to be zero. Annual peak flows from the two streamgages in this region are smaller than those from comparably sized drainages in any of the adjoining hydrologic regions. Consequently, expanding the adjoining hydrologic regions to include portions of this indeterminate region was considered likely to result in equations that would overpredict peak flow in this unique region. Using a rainfall-runoff model, calibrated with streamflow data from the two usable gages, might provide reliable estimates of flood frequency in this indeterminate region.

The APS regression methods were conducted on each of the five groups of streamgages outside the California desert to determine the candidate explanatory variables for each hydrologic region. The results of the APS analyses indicated that drainage area was the most significant variable for all exceedance probabilities, while the addition of mean annual precipitation reduced the standard error of estimate more than any of the other explanatory variables. In the Sierra Nevada region, adding mean basin elevation helped remove a bias in regression residuals that was pronounced at both low elevations and high elevations. Adding other variables to drainage area and mean annual precipitation did not significantly improve prediction equations in the other four regions. Thus, drainage area, mean basin elevation, and mean annual precipitation were selected as the only basin characteristics for further analysis in the Sierra Nevada region, and only drainage area and mean annual precipitation were selected in the other four regions outside the California desert. An OLS regression analysis was performed for the Sierra Nevada region using the following regression model:

$$Q_P = a_0 (DRNAREA)^{b_0} (ELEV)^{c_0} (PRECIP)^{d_0}, \quad (2)$$

where

$Q_P$ is the P-percent annual exceedance probability flow, in cubic feet per second;
*DRNAREA* is the drainage area, in square miles;
*ELEV* is the mean basin elevation, in feet;
*PRECIP* is the mean annual precipitation, in inches; and
$a_0$, $b_0$, $c_0$, and $d_0$ are the regression coefficients.

The regression model was logarithmically transformed to the following linear form:

$$\log Q_p = \log a_0 + b_0 (\log DRNAREA) + c_0 (\log ELEV) + d_0 (\log PRECIP) \quad . \quad (3)$$

For the other four regions outside the California desert, equations 2 and 3 included only the variables *DRNAREA* and *PRECIP* and only regression coefficients $a_0$, $b_0$, and $c_0$.

The residuals from the OLS analysis were plotted for each region in order to determine the need for dividing the regions into subregions. The residuals showed no geographical bias in the proposed hydrologic regions; therefore, the five hydrologic regions outside the desert region were used for the final GLS analysis (pl. 1).

## Regional Regression Equations

A GLS analysis was run on the final 630 streamgages outside of the desert that were considered for the regional regression analysis by using the WREG program. The multiple performance metrics from the WREG program were used to identify possible problem streamgages used in the regression. Residuals randomly distributed around zero are preferred. The leverage metric is used to measure how unusual the values of independent variables at one streamgage are compared to the values of the same variables at all other streamgages. The influence metric indicates whether the data at a streamgage had a large influence on the estimated regression parameter values (Eng and others, 2009). A streamgage may have a large leverage metric, indicating that its independent variables are substantially different from those at all other streamgages, but the same streamgage may not have a large influence on the regression parameters. Conversely, a streamgage with a large influence may not have a large leverage metric. Measurement or typographic errors in reported values of some independent variables may produce large leverage or influence metrics, and streamgages with such errors may need to be excluded. Streamgages that were identified by the WREG program as having large influence or leverage in this study were not excluded because no known errors were associated with the basin characteristic data, and a reasonable hydrologic justification for excluding the data could not be identified.

Combinations of independent explanatory variables that do not have multicollinearity and provide the lowest estimation error for each AEP were selected for inclusion in the final regression equations. Drainage area, mean basin elevation, and mean annual precipitation were the most appropriate basin characteristics used to estimate peak-streamflow frequency for ungaged sites in the Sierra Nevada region, and drainage area and mean annual precipitation were the basin characteristics used in the other regions outside of the desert region of California. The final regional regression equations for the 50- through 0.2-percent AEP flows for the five hydrologic regions are given in table 5. The values of drainage area, mean annual precipitation, and mean basin elevation for the 630 streamgages used in the regression analysis are given in table 2.

As previously described, regression analysis relating P-percent AEP flows to basin characteristics was not used to develop estimation equations for ungaged sites in the desert region. A WLS regression was used to determine regional models for the standard deviation and mean. The best model for the standard deviation was a constant model with a value of 0.91 log units. The best model for the mean was a linear model relating the mean to the log of drainage area. The best model for skew was previously determined by Thomas and others (1997) to be a constant value of zero. Placing these regional values of LP3 parameters into the basic LP3 equation (eq. 1) provided final equations for estimating P-percent AEP flows using drainage area (*DRNAREA*) as the only explanatory variable. Details on the development of the equations for estimating P-percent AEP flows in the desert region are given in the appendix. The final regional equations for the 50- through 0.2-percent AEP flows for the desert region are given in table 5. The values of drainage area for the 33 streamgages used in the regression analysis are given in table 2.

**Table 5.** Regional flood-frequency equations for rural ungaged streams in California.

[mi$^2$, square miles; *DRNAREA*, drainage area, in mi$^2$; *PRECIP*, mean annual precipitation, in inches; *ELEV*, mean basin elevation, in feet]

| Percent annual exceedance probability | Hydrologic region (shown in pl. 1) | | |
|---|---|---|---|
| | North Coast (Region 1) | Lahontan (Region 2) | Sierra Nevada (Region 3) |
| 50 | $1.82(DRNAREA)^{0.904}(PRECIP)^{0.983}$ | $0.0865(DRNAREA)^{0.736}(PRECIP)^{1.59}$ | $2.43(DRNAREA)^{0.924}(ELEV)^{-0.646}(PRECIP)^{2.06}$ |
| 20 | $8.11(DRNAREA)^{0.887}(PRECIP)^{0.772}$ | $0.182(DRNAREA)^{0.733}(PRECIP)^{1.58}$ | $11.6(DRNAREA)^{0.907}(ELEV)^{-0.566}(PRECIP)^{1.70}$ |
| 10 | $14.8(DRNAREA)^{0.880}(PRECIP)^{0.696}$ | $0.260(DRNAREA)^{0.734}(PRECIP)^{1.59}$ | $17.2(DRNAREA)^{0.896}(ELEV)^{-0.486}(PRECIP)^{1.54}$ |
| 4 | $26.0(DRNAREA)^{0.874}(PRECIP)^{0.628}$ | $0.394(DRNAREA)^{0.733}(PRECIP)^{1.58}$ | $20.7(DRNAREA)^{0.885}(ELEV)^{-0.386}(PRECIP)^{1.39}$ |
| 2 | $36.3(DRNAREA)^{0.870}(PRECIP)^{0.589}$ | $0.532(DRNAREA)^{0.733}(PRECIP)^{1.58}$ | $21.1(DRNAREA)^{0.879}(ELEV)^{-0.316}(PRECIP)^{1.31}$ |
| 1 | $48.5(DRNAREA)^{0.866}(PRECIP)^{0.556}$ | $0.713(DRNAREA)^{0.731}(PRECIP)^{1.56}$ | $20.6(DRNAREA)^{0.874}(ELEV)^{-0.250}(PRECIP)^{1.24}$ |
| 0.5 | $61.0(DRNAREA)^{0.863}(PRECIP)^{0.531}$ | $0.944(DRNAREA)^{0.729}(PRECIP)^{1.55}$ | $19.4(DRNAREA)^{0.870}(ELEV)^{-0.188}(PRECIP)^{1.18}$ |
| 0.2 | $79.3(DRNAREA)^{0.860}(PRECIP)^{0.503}$ | $1.35(DRNAREA)^{0.727}(PRECIP)^{1.52}$ | $17.4(DRNAREA)^{0.865}(ELEV)^{-0.110}(PRECIP)^{1.11}$ |

| Percent annual exceedance probability | Hydrologic region (shown in pl. 1) | | |
|---|---|---|---|
| | Central Coast (Region 4) | South Coast (Region 5) | Desert (Region 6) |
| 50 | $0.00459(DRNAREA)^{0.856}(PRECIP)^{2.58}$ | $3.60(DRNAREA)^{0.672}(PRECIP)^{0.753}$ | $10.3(DRNAREA)^{0.506}$ |
| 20 | $0.0984(DRNAREA)^{0.852}(PRECIP)^{1.97}$ | $7.43(DRNAREA)^{0.739}(PRECIP)^{0.872}$ | $60.0(DRNAREA)^{0.506}$ |
| 10 | $0.460(DRNAREA)^{0.846}(PRECIP)^{1.66}$ | $6.56(DRNAREA)^{0.783}(PRECIP)^{1.07}$ | $151(DRNAREA)^{0.506}$ |
| 4 | $2.13(DRNAREA)^{0.842}(PRECIP)^{1.34}$ | $4.71(DRNAREA)^{0.832}(PRECIP)^{1.32}$ | $403(DRNAREA)^{0.506}$ |
| 2 | $5.32(DRNAREA)^{0.840}(PRECIP)^{1.15}$ | $3.84(DRNAREA)^{0.864}(PRECIP)^{1.47}$ | $760(DRNAREA)^{0.506}$ |
| 1 | $11.0(DRNAREA)^{0.840}(PRECIP)^{0.994}$ | $3.28(DRNAREA)^{0.891}(PRECIP)^{1.59}$ | $1{,}350(DRNAREA)^{0.506}$ |
| 0.5 | $20.3(DRNAREA)^{0.840}(PRECIP)^{0.865}$ | $2.84(DRNAREA)^{0.915}(PRECIP)^{1.70}$ | $2{,}270(DRNAREA)^{0.506}$ |
| 0.2 | $39.0(DRNAREA)^{0.842}(PRECIP)^{0.729}$ | $2.31(DRNAREA)^{0.943}(PRECIP)^{1.83}$ | $4{,}280(DRNAREA)^{0.506}$ |

## Accuracy and Limitations

When applying regression equations, users are advised against interpreting the empirical results as exact. Regression equations are statistical models that 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 scatter or variance. The development and use of a regression equation raises questions about how well the predicted values represent true values. Differences between predicted and observed values at streamgages can be used to describe the accuracy of a regression equation, which depends on both the model and sampling error. Model error measures the ability of a set of explanatory variables to estimate the values of peak-flow characteristics calculated from the streamgage records that were used to develop the equation. The model error depends on the number and predictive power of the explanatory variables in a regression equation. Sampling error measures the ability of a finite number of streamgages with a finite number of recorded annual peak flows to describe the true peak-flow characteristics for a streamgage. The sampling error depends on the number of streamgages and record length of streamgages used in the analysis and decreases as either the number of streamgages or length of record increases.

A measure of the uncertainty in a regression equation estimate for a site, *i*, is the variance of prediction, $VP_i$. The $VP_i$ is the sum of the model error variance and sampling error variance and is computed using the following equation:

$$VP_i = \sigma_\delta^2 + \sigma_{\eta,i}^2, \quad (4)$$

where

$\sigma_\delta^2$ is the model error variance; and

$\sigma_{\eta,i}^2$ is the sampling mean square error for site *i*.

Assuming that the explanatory variables for the streamgages in a regression analysis are representative of all streamgages 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:

$$AVP = \sigma_\delta^2 + \left(\frac{1}{n}\right)\sum_{i=1}^{n}\sigma_{\eta,i}^2, \quad (5)$$

A more traditional measure of the accuracy of P-percent AEP flow regression equations is the standard error of prediction, $SE_P$, which is simply the square root of the variance of prediction. The average standard error of prediction for a regression equation can be computed in error percentage by using *AVP*, in log units, and the following transformation formula:

$$SE_{P,ave} = 100\left[10^{2.3026(AVP)} - 1\right]^{0.5}, \quad (6)$$

where

$SE_{P,ave}$ is the average standard error of prediction, in percent.

Approximately two-thirds of the estimates obtained from a regression equation for ungaged sites will have errors less than the standard error of prediction (Helsel and Hirsch, 1992).

A measure of the proportion of the variation in the dependent variable explained by the independent variables in OLS regressions is the coefficient of determination, $R^2$ (Montgomery and others, 2001). For WLS and GLS regressions, a more appropriate performance metric than $R^2$ is $R_{pseudo}^2$ described by Griffis and Stedinger (2007b). Unlike the $R^2$ metric, $R_{pseudo}^2$ is based on the variability in the dependent variable explained by the regression after removing the effect of the time-sampling error. The $R_{pseudo}^2$ is computed by using the following formula:

$$R_{pseudo}^2 = 1 - \frac{\sigma_\delta^2(k)}{\sigma_\delta^2(0)}, \quad (7)$$

where

$\sigma_\delta^2(k)$ is the model error variance from a GLS regression with *k* independent variables; and

$\sigma_\delta^2(0)$ is the model error variance from a GLS regression with no independent variables.

The average variance of prediction, average standard error of prediction, and $R_{pseudo}^2$ for the final set of regional regression equations are given in table 6. The $R_{pseudo}^2$ values cannot be computed for the desert region, because a regional regression on the P-percent AEP flows (flood quantiles) was never performed.

The results in table 6 indicate that the average standard errors of prediction are smallest for all AEP flows in the North Coast region (hydrologic region 1), with a range in values from 42.7 percent for the 4-percent and 2-percent AEP flow to 58.6 percent for the 50-percent AEP flow. Conversely, the average standard errors of prediction are largest for all AEP flows in the desert region (hydrologic region 6), where the values range from 214.2 percent for the 50-percent AEP flow to 856.2 percent for the 0.2-percent AEP flow, indicating the difficulty in accurately predicting flood flows in this region of extreme flow variability.

The standard errors of estimate for hydrologic regions 1 through 5 are less than the standard errors of estimate for the similar regions in Waananen and Crippen (1977; table 7); however, there are substantial differences in some of the regional boundaries used in the two studies, especially for regions 2 and 3 in this study. Likewise, results from the report by Busby and Hirashima (1972) cannot be compared to results from this study because the study areas are markedly different. Also, the standard errors of estimate for the California desert region shown by Waananen and Crippen (1977) cannot be compared to the standard errors of estimate for the desert region (hydrologic region 6) from this study, primarily because the region defined by Waananen and Crippen (1977) included many sites draining mountainous areas and did not include many sites with peak-flow records having numerous zero flows and other low outliers. However, the average standard errors of prediction in the desert region given in table 6 are

**Table 6.** Average variance of prediction, average standard error of prediction, and pseudo coefficient of determination for the regional regression equations.

[*AVP*, average variance of prediction; $SE_{p,ave}$, average standard error of prediction; $R^2_{pseudo}$, pseudo coefficient of determination; —, not applicable]

| Percent annual exceedance probability | Hydrologic region (shown in pl. 1) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | North Coast (Region 1) | | | Lahontan (Region 2) | | | Sierra Nevada (Region 3) | | |
| | *AVP* (log units) | $SE_{p,ave}$ (percent) | $R^2_{pseudo}$ (percent) | *AVP* (log units) | $SE_{p,ave}$ (percent) | $R^2_{pseudo}$ (percent) | *AVP* (log units) | $SE_{p,ave}$ (percent) | $R^2_{pseudo}$ (percent) |
| 50 | 0.056 | 58.6 | 93.7 | 0.126 | 97.5 | 82.6 | 0.083 | 74.4 | 88.3 |
| 20 | 0.038 | 47.4 | 95.4 | 0.098 | 82.7 | 85.6 | 0.049 | 54.4 | 92.0 |
| 10 | 0.034 | 44.2 | 95.9 | 0.089 | 77.7 | 86.6 | 0.044 | 51.5 | 92.3 |
| 4 | 0.032 | 42.7 | 96.0 | 0.086 | 76.1 | 86.7 | 0.046 | 52.3 | 91.6 |
| 2 | 0.032 | 42.7 | 96.0 | 0.085 | 75.7 | 86.6 | 0.049 | 54.6 | 90.8 |
| 1 | 0.034 | 44.3 | 95.6 | 0.088 | 77.2 | 86.0 | 0.055 | 58.0 | 89.6 |
| 0.5 | 0.034 | 44.4 | 95.6 | 0.091 | 78.9 | 85.3 | 0.060 | 61.5 | 88.5 |
| 0.2 | 0.036 | 46.0 | 95.2 | 0.099 | 82.9 | 84.0 | 0.070 | 67.3 | 86.7 |

| Percent annual exceedance probability | Hydrologic region (shown in pl. 1) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Central Coast (Region 4) | | | South Coast (Region 5) | | | Desert (Region 6) | | |
| | *AVP* (log units) | $SE_{p,ave}$ (percent) | $R^2_{pseudo}$ (percent) | *AVP* (log units) | $SE_{p,ave}$ (percent) | $R^2_{pseudo}$ (percent) | *AVP* (log units) | $SE_{p,ave}$ (percent) | $R^2_{pseudo}$ (percent) |
| 50 | 0.243 | 161.9 | 76.9 | 0.194 | 134.2 | 60.6 | 0.325 | 214.2 | — |
| 20 | 0.125 | 97.0 | 86.3 | 0.099 | 83.1 | 78.8 | 0.342 | 226.2 | — |
| 10 | 0.092 | 79.4 | 89.2 | 0.065 | 64.0 | 86.4 | 0.371 | 248.1 | — |
| 4 | 0.075 | 69.9 | 90.7 | 0.044 | 51.5 | 91.4 | 0.432 | 297.6 | — |
| 2 | 0.069 | 66.2 | 91.4 | 0.038 | 47.6 | 93.1 | 0.494 | 356.9 | — |
| 1 | 0.070 | 66.9 | 91.0 | 0.038 | 47.2 | 93.7 | 0.572 | 444.3 | — |
| 0.5 | 0.071 | 67.6 | 90.8 | 0.039 | 47.7 | 93.9 | 0.665 | 574.5 | — |
| 0.2 | 0.078 | 71.5 | 89.8 | 0.045 | 52.0 | 93.3 | 0.813 | 856.2 | — |

**Table 7.** Standard errors of estimate from this investigation and from Waananen and Crippen (1977).

| Percent annual exceedance probability | Hydrologic region (shown in pl. 1) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | North Coast (Region 1) | | Lahontan (Region 2) | | Sierra Nevada (Region 3) | | Central Coast (Region 4) | | South Coast (Region 5) | |
| | Standard error of estimate ($SE_e$) (log units) | | | | | | | | | |
| | This investigation (2012) | Waananen and Crippen (1977) | This investigation (2012) | Waananen and Crippen (1977) | This investigation (2012) | Waananen and Crippen (1977) | This investigation (2012) | Waananen and Crippen (1977) | This investigation (2012) | Waananen and Crippen (1977) |
| 50 | 0.23 | 0.26 | 0.34 | 0.46 | 0.33 | 0.34 | 0.48 | 0.47 | 0.43 | 0.47 |
| 20 | 0.19 | 0.24 | 0.30 | 0.38 | 0.26 | 0.32 | 0.34 | 0.39 | 0.30 | 0.37 |
| 10 | 0.18 | 0.24 | 0.28 | 0.38 | 0.23 | 0.27 | 0.29 | 0.35 | 0.24 | 0.33 |
| 4 | 0.17 | 0.24 | 0.27 | 0.40 | 0.22 | 0.30 | 0.25 | 0.35 | 0.19 | 0.32 |
| 2 | 0.17 | 0.25 | 0.27 | 0.42 | 0.22 | 0.34 | 0.24 | 0.38 | 0.17 | 0.35 |
| 1 | 0.18 | 0.26 | 0.27 | 0.45 | 0.23 | 0.37 | 0.24 | 0.41 | 0.17 | 0.39 |

smaller than comparable errors reported for the prediction equations for the Southwestern United States by Thomas and others (1997). Although Teal and Gusman (2007) used the same method as Thomas and others (1997) to develop prediction equations for the desert region, their reported errors did not account for the extreme variability and uncertainty of the flood-frequency estimates at the gaged sites and thus cannot be compared to the errors listed in table 6.

Users of the regression models may be interested in a measure of uncertainty for a flow estimate at a particular site as opposed to the average uncertainty based on all streamgage data used to generate the regression models. One such measure of uncertainty at a particular ungaged site is the confidence interval of a prediction, or prediction interval. A prediction interval is the range in values of an estimated response variable over which the true value of the response variable occurs with some stated probability. For example, the 90-percent prediction interval for an estimated flow value means that 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:

$$Q / C < Q < CQ, \tag{8}$$

where

$Q$ is the streamflow characteristic for the ungaged site; and

$C$ is computed as:

$$C = 10^{t_{(\alpha/2,\, n-p)}\, SE_{P,i}}, \tag{9}$$

where

$t_{(\alpha/2,\, n-p)}$ is the critical value from the student's $t$-distribution at a particular alpha-level ($\alpha$) and degrees of freedom ($n-p$) and is equal to 1.653, 1.675, 1.653, 1.662, and 1.661 for hydrologic regions 1, 2, 3, 4, and 5, respectively, for a prediction interval of 90 percent ($\alpha = 0.1$); and

$SE_{P,i}$ is the standard error of prediction for site $i$ and is computed as

$$SE_{P,i} = [\sigma_\delta^2 + \mathbf{X}_i \mathbf{U} \mathbf{X}_i^T]^{0.5}, \tag{10}$$

where

$\sigma_\delta^2$ is the model error variance;

$\mathbf{X}_i$ is a row vector of the explanatory variables for site $i$, augmented by a 1 as the first element;

$\mathbf{U}$ is the covariance matrix for the regression coefficients; and

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

The values for $\sigma_\delta^2$ and $\mathbf{U}$ are presented in table 8. Prediction intervals cannot be computed for the desert region (hydrologic region 6) because of the non-standard methods used to develop the equations.

The procedure required to obtain the prediction intervals for P-percent AEP flow estimates is explained in the following example computation of the 2-percent AEP flow for a hypothetical ungaged site on Indian Creek near Big City, Calif., in hydrologic region 1. The results are rounded to three significant figures.

1. Obtain the drainage area and mean annual precipitation for the ungaged site (*DRNAREA*= 5.00 mi$^2$, *PRECIP* = 30.0 inches);

2. Compute $Q_{2\%}$ using the equation in table 5 for hydrologic region 1 ($Q_{2\%} = 36.3 \times (5.00^{0.870} \times 30.0)^{0.589} = 1{,}090$ ft$^3$/s);

3. Determine the $\mathbf{X}_i$ vector ($\mathbf{X}_i = \{1, \log_{10}(5.00), \log_{10}(30.0)\}$);

4. Compute the standard error of prediction using equation 10 with $\sigma_\delta^2$ and $\mathbf{U}$ for the 2-percent AEP flow from table 8; $SE_{P,i} = (0.0291 + 0.003339)^{0.5} = 0.1801$;

5. Compute $C$ using equation 9; $C = 10^{(1.653 \times 0.1801)} = 1.985$; and

6. Compute the 90-percent prediction interval using equation 8; $(1{,}090/1.985) < Q_{2\%} < (1{,}090 \times 1.985)$ or 549 ft$^3$/s $< Q_{2\%} <$ 2,160 ft$^3$/s.

**Table 8.** Values used to determine prediction intervals for the regional regression equations.

[$\sigma_\delta^2$, the regression model error variance used in equation 10; **U**, the covariance matrix used in equation 10; —, not applicable]

| Percent annual exceedance probability | Hydrologic region (shown in pl. 1) | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | North Coast (Region 1) | | | | Lahontan (Region 2) | | | | Sierra Nevada (Region 3) | | | | |
| | $\sigma_\delta^2$ | U | | | $\sigma_\delta^2$ | U | | | $\sigma_\delta^2$ | U | | | |
| 50 | 0.053 | 4.82E-2 | –5.42E-4 | –2.61E-2 | 0.116 | 1.93E-1 | –3.72E-3 | –1.15E-1 | 0.079 | 7.91E-2 | –1.81E-3 | –7.28E-3 | –2.81E-2 |
| | | –5.42E-4 | 4.81E-4 | –1.80E-4 | | –3.72E-3 | 3.36E-3 | –5.86E-4 | | –1.81E-3 | 8.95E-4 | –4.81E-4 | 1.14E-3 |
| | | –2.61E-2 | –1.80E-4 | 1.50E-2 | | –1.15E-1 | –5.86E-4 | 7.29E-2 | | –7.28E-3 | –4.81E-4 | 6.44E-3 | –9.12E-3 |
| | | | | | | | | | | –2.81E-2 | 1.14E-3 | –9.12E-3 | 3.51E-2 |
| 20 | 0.036 | 3.86E-2 | –4.49E-4 | –2.07E-2 | 0.089 | 1.67E-1 | –4.22E-3 | –9.77E-2 | 0.046 | 5.55E-2 | –1.10E-3 | –5.32E-3 | –1.92E-2 |
| | | –4.49E-4 | 3.62E-4 | –1.24E-4 | | –4.22E-3 | 2.84E-3 | 3.42E-5 | | –1.10E-3 | 6.26E-4 | –4.76E-4 | 9.60E-4 |
| | | –2.07E-2 | –1.24E-4 | 1.18E-2 | | –9.77E-2 | 3.42E-5 | 6.10E-2 | | –5.32E-3 | –4.76E-4 | 5.46E-3 | –7.88E-3 |
| | | | | | | | | | | –1.92E-2 | 9.60E-4 | –7.88E-3 | 2.69E-2 |
| 10 | 0.032 | 3.93E-2 | –4.76E-4 | –2.09E-2 | 0.080 | 1.66E-1 | –4.86E-3 | –9.57E-2 | 0.041 | 5.86E-2 | –1.04E-3 | –6.20E-3 | –1.92E-2 |
| | | –4.76E-4 | 3.46E-4 | –1.01E-4 | | –4.86E-3 | 2.76E-3 | 3.96E-4 | | –1.04E-3 | 6.17E-4 | –5.27E-4 | 1.02E-3 |
| | | –2.09E-2 | –1.01E-4 | 1.19E-2 | | –9.57E-2 | 3.96E-4 | 5.93E-2 | | –6.20E-3 | –5.27E-4 | 6.19E-3 | –8.75E-3 |
| | | | | | | | | | | –1.92E-2 | 1.02E-3 | –8.75E-3 | 2.85E-2 |
| 4 | 0.029 | 4.28E-2 | –5.35E-4 | –2.26E-2 | 0.075 | 1.77E-1 | –5.78E-3 | –1.01E-1 | 0.041 | 6.90E-2 | –1.11E-3 | –7.85E-3 | –2.16E-2 |
| | | –5.35E-4 | 3.54E-4 | –8.46E-5 | | –5.78E-3 | 2.88E-3 | 7.36E-4 | | –1.11E-3 | 6.80E-4 | –6.23E-4 | 1.18E-3 |
| | | –2.26E-2 | –8.46E-5 | 1.27E-2 | | –1.01E-1 | 7.36E-4 | 6.20E-2 | | –7.85E-3 | –6.23E-4 | 7.60E-3 | –1.06E-2 |
| | | | | | | | | | | –2.16E-2 | 1.18E-3 | –1.06E-2 | 3.34E-2 |
| 2 | 0.029 | 4.63E-2 | –5.86E-4 | –2.44E-2 | 0.074 | 1.86E-1 | –6.46E-3 | –1.06E-1 | 0.044 | 7.85E-2 | –1.21E-3 | –9.22E-3 | –2.41E-2 |
| | | –5.86E-4 | 3.72E-4 | –8.01E-5 | | –6.46E-3 | 3.00E-3 | 9.51E-4 | | –1.21E-3 | 7.55E-4 | –7.09E-4 | 1.33E-3 |
| | | –2.44E-2 | –8.01E-5 | 1.37E-2 | | –1.06E-1 | 9.51E-4 | 6.47E-2 | | –9.22E-3 | –7.09E-4 | 8.84E-3 | –1.22E-2 |
| | | | | | | | | | | –2.41E-2 | 1.33E-3 | –1.22E-2 | 3.81E-2 |
| 1 | 0.031 | 5.11E-2 | –6.46E-4 | –2.69E-2 | 0.075 | 2.01E-1 | –7.19E-3 | –1.13E-1 | 0.049 | 8.92E-2 | –1.33E-3 | –1.07E-2 | –2.71E-2 |
| | | –6.46E-4 | 4.06E-4 | –8.52E-5 | | –7.19E-3 | 3.21E-3 | 1.13E-3 | | –1.33E-3 | 8.51E-4 | –8.11E-4 | 1.50E-3 |
| | | –2.69E-2 | –8.52E-5 | 1.51E-2 | | –1.13E-1 | 1.13E-3 | 6.94E-2 | | –1.07E-2 | –8.11E-4 | 1.02E-2 | –1.41E-2 |
| | | | | | | | | | | –2.71E-2 | 1.50E-3 | –1.41E-2 | 4.36E-2 |
| 0.5 | 0.031 | 5.42E-2 | –6.90E-4 | –2.84E-2 | 0.077 | 2.16E-1 | –7.91E-3 | –1.21E-1 | 0.054 | 9.99E-2 | –1.46E-3 | –1.21E-2 | –3.02E-2 |
| | | –6.90E-4 | 4.23E-4 | –8.23E-5 | | –7.91E-3 | 3.43E-3 | 1.29E-3 | | –1.46E-3 | 9.52E-4 | –9.15E-4 | 1.67E-3 |
| | | –2.84E-2 | –8.23E-5 | 1.59E-2 | | –1.21E-1 | 1.29E-3 | 7.43E-2 | | –1.21E-2 | –9.15E-4 | 1.16E-2 | –1.60E-2 |
| | | | | | | | | | | –3.02E-2 | 1.67E-3 | –1.60E-2 | 4.92E-2 |
| 0.2 | 0.033 | 5.95E-2 | –7.56E-4 | –3.12E-2 | 0.082 | 2.41E-1 | –8.95E-3 | –1.35E-1 | 0.063 | 1.15E-1 | –1.64E-3 | –1.41E-2 | –3.47E-2 |
| | | –7.56E-4 | 4.61E-4 | –8.87E-5 | | –8.95E-3 | 3.81E-3 | 1.48E-3 | | –1.64E-3 | 1.11E-3 | –1.07E-3 | 1.93E-3 |
| | | –3.12E-2 | –8.87E-5 | 1.75E-2 | | –1.35E-1 | 1.48E-3 | 8.25E-2 | | –1.41E-2 | –1.07E-3 | 1.37E-2 | –1.89E-2 |
| | | | | | | | | | | –3.47E-2 | 1.93E-3 | –1.89E-2 | 5.76E-2 |

**Table 8.** Values used to determine prediction intervals for the regional regression equations.—Continued

[$\sigma_\delta^2$, the regression model error variance used in equation 10; **U**, the covariance matrix used in equation 10; —, not applicable]

| Percent annual exceedance probability | Hydrologic region (shown in pl. 1) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Central Coast (Region 4) | | | | South Coast (Region 5) | | | | Desert (Region 6) | |
| | $\sigma_\delta^2$ | U | | | $\sigma_\delta^2$ | U | | | $\sigma_\delta^2$ | U |
| 50 | 0.228 | 3.36E-1 | –1.01E-2 | –2.29E-1 | 0.184 | 3.28E-1 | –1.06E-2 | –2.22E-1 | — | — |
| | | –1.01E-2 | 3.90E-3 | 2.52E-3 | | –1.06E-2 | 3.72E-3 | 4.24E-3 | | |
| | | –2.29E-1 | 2.52E-3 | 1.65E-1 | | –2.22E-1 | 4.24E-3 | 1.56E-1 | | |
| 20 | 0.115 | 2.13E-1 | –7.16E-3 | –1.41E-1 | 0.091 | 1.95E-1 | –6.30E-3 | –1.32E-1 | — | — |
| | | –7.16E-3 | 2.37E-3 | 2.16E-3 | | –6.30E-3 | 2.17E-3 | 2.66E-3 | | |
| | | –1.41E-1 | 2.16E-3 | 9.96E-2 | | –1.32E-1 | 2.66E-3 | 9.32E-2 | | |
| 10 | 0.083 | 1.89E-1 | –6.67E-3 | –1.23E-1 | 0.058 | 1.57E-1 | –5.00E-3 | –1.06E-1 | — | — |
| | | –6.67E-3 | 2.02E-3 | 2.16E-3 | | –5.00E-3 | 1.66E-3 | 2.26E-3 | | |
| | | –1.23E-1 | 2.16E-3 | 8.50E-2 | | –1.06E-1 | 2.26E-3 | 7.47E-2 | | |
| 4 | 0.065 | 1.85E-1 | –6.67E-3 | –1.18E-1 | 0.037 | 1.41E-1 | –4.36E-3 | –9.48E-2 | — | — |
| | | –6.67E-3 | 1.91E-3 | 2.25E-3 | | –4.36E-3 | 1.40E-3 | 2.11E-3 | | |
| | | –1.18E-1 | 2.25E-3 | 8.08E-2 | | –9.48E-2 | 2.11E-3 | 6.71E-2 | | |
| 2 | 0.058 | 1.89E-1 | –6.83E-3 | –1.20E-1 | 0.030 | 1.43E-1 | –4.33E-3 | –9.62E-2 | — | — |
| | | –6.83E-3 | 1.90E-3 | 2.34E-3 | | –4.33E-3 | 1.38E-3 | 2.17E-3 | | |
| | | –1.20E-1 | 2.34E-3 | 8.10E-2 | | –9.62E-2 | 2.17E-3 | 6.82E-2 | | |
| 1 | 0.058 | 2.02E-1 | –7.31E-3 | –1.28E-1 | 0.029 | 1.55E-1 | –4.61E-3 | –1.04E-1 | — | — |
| | | –7.31E-3 | 2.02E-3 | 2.52E-3 | | –4.61E-3 | 1.46E-3 | 2.36E-3 | | |
| | | –1.28E-1 | 2.52E-3 | 8.62E-2 | | –1.04E-1 | 2.36E-3 | 7.37E-2 | | |
| 0.5 | 0.058 | 2.15E-1 | –7.77E-3 | –1.36E-1 | 0.029 | 1.68E-1 | –4.98E-3 | –1.13E-1 | — | — |
| | | –7.77E-3 | 2.13E-3 | 2.70E-3 | | –4.98E-3 | 1.56E-3 | 2.58E-3 | | |
| | | –1.36E-1 | 2.70E-3 | 9.13E-2 | | –1.13E-1 | 2.58E-3 | 8.05E-2 | | |
| 0.2 | 0.064 | 2.40E-1 | –8.63E-3 | –1.51E-1 | 0.033 | 1.97E-1 | –5.83E-3 | –1.33E-1 | — | — |
| | | –8.63E-3 | 2.36E-3 | 3.00E-3 | | –5.83E-3 | 1.82E-3 | 3.02E-3 | | |
| | | –1.51E-1 | 3.00E-3 | 1.02E-1 | | –1.33E-1 | 3.02E-3 | 9.48E-2 | | |

The following limitations need to be recognized when using the final regional regression equations:

1. The ranges of explanatory variables used to develop the regional regression equations are given in table 9. Applying the equations to sites on streams having explanatory variables outside the ranges of those used in this study may result in prediction errors that are considerably greater than those indicated by the standard error of prediction percentages listed in table 6.
2. The methods are not appropriate (or applicable) for sites where the peak-flow magnitudes are affected substantially by flow regulation.
3. The methods are not appropriate (or applicable) for streams in urban areas (impervious area greater than 10 percent) unless the effects of urbanization are insignificant.
4. The methods may not be applicable at some high-elevation locations where mixed populations of snow-melt and rainfall flood events might not be adequately described by a single LP3 distribution.

**Table 9.** Ranges of explanatory variables used to develop the regional regression equations for California.

[Min, minimum; Max, maximum; $mi^2$, square miles; —, not applicable]

| Basin characteristic | Hydrologic region (shown in pl. 1) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | North Coast (Region 1) | | Lahontan (Region 2) | | Sierra Nevada (Region 3) | | Central Coast (Region 4) | | South Coast (Region 5) | | Desert (Region 6) | |
| | Min | Max | Min | Max | Min | Max | Min | Max | Min | Max | Min | Max |
| Drainage area ($mi^2$) | 0.04 | 3,200 | 0.45 | 1,500 | 0.07 | 2,000 | 0.11 | 4,600 | 0.04 | 850 | 0.04 | 173 |
| Mean annual precipitation (inches) | 20.0 | 125.0 | 13.0 | 85.0 | 15.0 | 100.0 | 7.0 | 46.0 | 10.0 | 45.0 | — | — |
| Mean basin elevation (feet) | — | — | — | — | 90 | 11,000 | — | — | — | — | — | — |

# Application of Methods

The best estimates of flood frequencies for a site typically are obtained through a weighted combination of estimates produced from more than one method. Tasker (1975) demonstrated that if two independent estimates of a streamflow statistic are available, a properly weighted average of the independent estimates will provide an estimate that is more accurate than either of the independent estimates. Improved flood-frequency estimates can be determined for California streamgages outside of the desert region by weighting estimates determined from the Bulletin 17B analysis with estimates obtained from the regression equations provided in this report. Weighting of estimates in the desert region is not appropriate, because the at-site estimates determined using the special LP3 analysis previously described almost exclusively depend on regional estimates of the three LP3 parameters. Thus, the at-site estimates for flows with various AEP and estimates from regional equations are not independent. Flow estimates at ungaged sites on the same stream as gaged sites in California can be improved by weighting the estimates obtained from the regression equations with estimates that were determined on the basis of flow at an upstream or downstream streamgage. The following sections describe the weighting process for streamgages and ungaged sites in more detail and provide example calculations. The results are rounded to three significant figures.

## Estimation for a Streamgage

The Interagency Advisory Committee on Water Data (1982) recommends that better estimates of flood-frequency statistics for a streamgage can be obtained by combining (weighting) at-site flow estimates determined from the LP3 analysis of the annual peak flows with flow estimates obtained for the streamgage from regression equations. Optimal weighted flow estimates can be obtained if the variance of prediction for each of the two estimates is known or can be estimated accurately. The variance of prediction can be thought of as a measure of the uncertainty in either the at-site estimate or the regional regression results. If the two estimates can be assumed to be independent and are weighted inversely proportional to their associated variances, the variance of the weighted estimate will be less than the variance of either of the independent estimates. Optimal weighted flow estimates were computed for the study using the Weighted Independent Estimates (WIE) computer program available at *http://water.usgs.gov/usgs/osw/swstats/freq.html*. Details on this computer program are described by Cohn and others (2012).

The variance of prediction corresponding to the at-site flow estimate from the LP3 analysis is computed using the asymptotic formula given in Cohn and others (2001) with the addition of the mean-squared error of generalized skew (Griffis and others, 2004). This variance varies as a function of the length of record, the fitted LP3 distribution parameters (mean, standard deviation, and weighted skew), and the accuracy of the method used to determine the generalized skew component of the weighted skew. The variance of prediction for the at-site estimate generally decreases with increasing length of record and generally improving quality of the LP3 distribution fit. The variance of prediction values for the at-site flow estimates for the 734 streamgages located outside the desert region of California are given in table 10. The variance of prediction from the regional regression equations is a function of the regression equations and the values of the independent variables used to develop the flow estimate from the regression equations. This variance generally increases as the values of the independent variables move further from the mean values of the independent variables. The variance of prediction values for the regional regression equations used in this study are given in table 10.

**Table 10.** Variance of prediction values for streamgages in California that were weighted using the Bulletin 17B estimates and the regional regression estimates.

[Table 10 is available in a Microsoft© Excel spreadsheet and can be accessed and downloaded at *http://pubs.usgs.gov/sir/2012/5113/*]

Once the variances have been computed, the two independent flow estimates can be weighted using the following equation:

$$\log Q_{P(g)w} = \frac{VP_{P(g)r} \log Q_{P(g)s} + VP_{P(g)s} \log Q_{P(g)r}}{VP_{P(g)s} + VP_{P(g)r}}, \quad (11)$$

where

$Q_{P(g)w}$ is the weighted estimate of peak flow for any P-percent annual exceedance probability for a streamgage, *g*, in cubic feet per second;

$VP_{P(g)r}$ is the variance of prediction at the streamgage derived from the applicable regional regression equations for the selected P-percent annual exceedance probability (from table 6), in log units;

$Q_{P(g)s}$ is the estimate of peak flow at the streamgage from the log-Pearson Type III analysis for the selected P-percent annual exceedance probability, in cubic feet per second;

$VP_{P(g)s}$ is the variance of prediction at the streamgage from the log-Pearson Type III analysis for the selected P-percent annual exceedance probability (from table 10), in log units; and

$Q_{P(g)r}$ is the peak-flow estimate for the P-percent annual exceedance probability at the streamgage derived from the applicable regional regression equations in table 5, in cubic feet per second.

When the variance of prediction corresponding to one of the estimates is high, the uncertainty is also high, and the weight for that estimate is relatively small. Conversely, when the variance of prediction is low, the uncertainty is also low and the weight is correspondingly large. The variance of prediction associated with the weighted estimate, $VP_{P(g)w}$, is computed using the following equation:

$$VP_{P(g)w} = \frac{VP_{P(g)s} VP_{P(g)r}}{VP_{P(g)s} + VP_{P(g)r}} . \quad (12)$$

The weighted (best) flow estimates were computed using equation 11 along with the variance of prediction values from tables 6 and 10 for the 734 streamgages in California that are outside of the desert region. The weighted flow estimates for the 734 streamgages are listed in table 4. The variance of prediction values associated with the weighted estimates are given in table 10.

An example of the application of the procedure described above is the following computation of the weighted 1-percent AEP flow for the streamgage on Corralitos Creek at Freedom, Calif. (station number 11159200):

1. Obtain the at-site estimate of the 1-percent AEP flow at the streamgage based on the Bulletin 17B analysis from table 4 ($Q_{1\%(g)s}$ = 6,980 ft$^3$/s);
2. Obtain drainage area and mean annual precipitation from table 2 for the streamgage (*DRNAREA* = 27.8 mi$^2$, *PRECIP* = 32.7 inches);
3. Compute $Q_{1\%(g)r}$ using equation in table 5 for hydrologic region 1 ($Q_{1\%(g)r} = 48.5(27.8^{0.866} \times 32.7^{0.556})$ = 6,000 ft$^3$/s);
4. Obtain the variance of prediction for the at-site estimate for the 1-percent AEP from table 10 ($VP_{1\%(g)s}$ = 0.0127);
5. Obtain the variance of prediction for the 1-percent AEP flow regression equation from table 10 ($VP_{1\%(g)r}$ = 0.0338);
6. Compute the weighted 1-percent AEP flow for the streamgage using equation 11 (log $Q_{1\%(g)w}$ = ((0.0338)(log 6,980)+(0.0127)(log 6,000))/(0.0127+0.0338) = 3.826, and $Q_{1\%(g)w}$ = 6,700 ft$^3$/s); and
7. Compute the weighted variance for the streamgage using equation 12 ($VP_{1\%(g)w}$ = (((0.0127)(0.0338))/(0.0127+0.0338) = 0.0092).

Previous USGS flood-frequency reports used the equivalent years of record associated with the regression equations in addition to the length of record at the streamgage to weight the flow estimates obtained from the regional regression equation and the LP3 analysis. The length of record, however, often fails to account for the true variance of LP3 flood-frequency estimates. For example, although longer record lengths generally result in decreased variance, record length fails to account for the improvement in information content provided by the generalized skew or the addition of historic peak flows. Furthermore, flood-frequency distributions computed from two different streamgage records of the same length may not be of equal reliability, owing to differences in underlying variances of the peak-flow records. For example, smaller drainage areas may have flashier, more highly varied peak-flow records, or may be more difficult to accurately gage than a large basin, hence the LP3 distributions could be expected to have larger variances. More importantly, the equivalent year of record concept, while relatively easy to grasp, can sometimes misconstrue the relation between flood-frequency estimates and associated variances. Using variances provides a more accurate characterization of the underlying uncertainty of the various flow estimates.

## Estimation for an Ungaged Site Near a Streamgage

Sauer (1974) presented the following method to improve flood-frequency estimates for an ungaged site near a streamgage, on the same stream, having 10 or more years of peak-flow record. To obtain a weighted flow estimate $Q_{P(u)w}$ for P-percent AEP at the ungaged site, the weighted flow estimate for an upstream or downstream streamgage $Q_{P(g)w}$ must first be determined by using the equation provided in the previous section. The weighted flow estimate for the ungaged site ($Q_{P(u)w}$) is then computed using the following equation:

$$Q_{P(u)w} = \left[\left(\frac{2\Delta A}{A_{(g)}}\right) + \left(1 - \frac{2\Delta A}{A_{(g)}}\right)\left(\frac{Q_{P(g)w}}{Q_{P(g)r}}\right)\right] Q_{P(u)r} , \quad (13)$$

where

$Q_{P(u)w}$ is the weighted estimate of peak flow for the selected P-percent annual exceedance probability at the ungaged site, *u*, in cubic feet per second;

$\Delta A$ is the absolute value of the difference between the drainage areas of the streamgage and the ungaged site, in square miles;

$A_{(g)}$ is the drainage area for the streamgage, in square miles; and

$Q_{P(u)r}$ is the peak-flow estimate derived from the applicable regional equations in table 5 for the selected P-percent annual exceedance probability at the ungaged site, in cubic feet per second.

Use of equation 13 gives full weight to the regression equation estimates when the drainage area for the ungaged site is equal to 0.5 or 1.5 times the drainage area for the streamgage and increasing weight to the streamgage estimates as the drainage area ratio approaches 1. The weighting procedure is not applicable when the drainage area ratio for the ungaged site and streamgage is less than 0.5 or greater than 1.5.

An example application of the procedure described in this section is the following computation of the weighted 1-percent AEP flow for a hypothetical ungaged site on the Corralitos Creek located above the USGS streamgage at Freedom, Calif. (station number 11159200) cited in the previous section:

1. Calculate the value of $Q_{1\%(g)w}$ ($Q_{1\%(g)w}$ = 6,700 ft$^3$/s);
2. Obtain the drainage areas and mean annual precipitation for both the gaged and ungaged sites ($DRNAREA_g$ = 27.8 mi$^2$, $DRNAREA_u$ = 25.0 mi$^2$, $PRECIP_g$ = 32.7 inches, and $PRECIP_u$ = 32.5 inches);
3. Compute $Q_{1\%(u)r}$ for the ungaged site using the equation in table 5 for hydrologic region 1 ($Q_{1\%(u)r}$ = $48.5(25.0^{0.866} \times 32.5^{0.556})$ = 5,460 ft$^3$/s);
4. Compute $Q_{1\%(g)r}$ for the streamgage using the equation in table 5 for hydrologic region 1 ($Q_{1\%(g)r}$ = 6,000 ft$^3$/s);
5. Compute $\Delta A$, where $\Delta A$ = (27.8–25.0) = 2.8 mi$^2$; and
6. Compute the weighted estimate for the ungaged site, $Q_{1\%(u)w}$, using equation 13

$$\left(Q_{1\%(u)w} = \left[\left(\frac{2\times 2.8}{27.8}\right) + \left(1 - \frac{2\times 2.8}{27.8}\right)\left(\frac{6,700}{6,000}\right)\right] \times 5,460 = 5,970 \text{ ft}^3/\text{s}\right)$$

For an ungaged site that is located between two streamgages on the same stream, two flow estimates can be made using the methods and criteria outlined in this section. Some additional hydrologic judgment may be necessary to determine which of the two estimates (or some interpolation thereof) is most appropriate. Other factors that might be considered when evaluating the two estimates include differences in the length of record for the two streamgages and the hydrologic conditions that existed during the data-collection period for each streamgage (that is, does the time series represent a climatic period that was predominately wet or dry?).

## Estimation for an Ungaged Site Draining More Than One Hydrologic Region

For an ungaged site on a stream that crosses hydrologic regions, the regression equations for each region can be applied separately using basin characteristics for the entire drainage basin above the ungaged site. The individual results can then be weighted by the proportion of drainage area within each region and added to produce final estimates for the ungaged site. For example, if 40 percent of the drainage area at an ungaged site is in the upstream region and 60 percent is in the downstream region, the discharge estimate based on equations for the upstream region are multiplied by 0.40 and added to 0.60 times the regression estimate based on equations for the downstream region. The variance of prediction for such a weighted estimate can also be approximated by using the same weighting procedure based on proportional drainage areas.

# Effects of Urbanization on Floods

Protecting life and property in flood plains of urbanized basins requires an understanding of the effects of urbanization on peak flows. Urbanization changes the response of a basin to precipitation. The most common effects are reduced infiltration and decreased lag time, which can significantly increase peak flows (U.S. Department of Agriculture, 1986). Engineers and planners routinely consider the potential effects of urban development on peak flows in their design and planning efforts. Because urbanization can produce significant changes in flood-frequency characteristics of streams, flood-frequency relations developed for streams in rural areas are not always applicable to urban streams. This section describes an analysis that was done to try to determine the effects of urbanization on California streams.

Waananen and Crippen (1977) provided an evaluation of the effects of urbanization on California basins and presented methods for adjusting California stream peak flows for urbanization based on simulated rather than recorded data (Rantz, 1971). Feaster and Guimaraes (2004) found a significant difference between flood-frequency estimates determined using simulated peak-flow data and estimates using only recorded peak-flow data. Southard (2010) also showed that urban regression equations that were developed using simulated data predicted larger flood-frequency flows than urban regression equations that were developed using recorded data. Although the methods used by Rantz (1971) to simulate the data are not the same as the methods used for the simulated data analyzed by Feaster and Guimaraes (2004) and Southard (2010), the simulated data used by Rantz (1971) may also be biased high. Durbin (1974) found that the effects of urbanization on the larger floods (AEPs less than 2 percent) are not significant in the southern California river basins they analyzed, even with the use of simulated peak-flow data.

Urban streamgages in California were investigated for possible use in developing methods for estimating the magnitude and frequency of floods in ungaged urban basins in California. Because of the potential for bias in simulated peak-flow data, only recorded data were used. The percentage of impervious surface area has long been recognized as an effective indicator of the intensity of urban development and its potential effects on streamflow and the environment (Klein, 1979). The threshold of influence of impervious surface area on streamflow has been reported in previous studies to be between 5 and 21 percent (Brabec and others, 2002). For this study, a streamgage was considered urban if the impervious area within the drainage area was 10 percent or greater.

Urban streamgages were used in the analysis only if 10 or more years of homogeneous annual peak-flow data were available. Homogeneous, in this context, means that the recorded annual peak flows showed no significant trends, which further implies that the percentage of impervious area was stable throughout the period of record. The peak-flow records for the candidate urban streamgages were then

compiled and reviewed for quality by using the PFReports computer program (Ryberg, 2008). Kendall's tau was chosen to assess the significance of time trends in the peak-flow record attributed to the effects of urbanization (Helsel and Hirsch, 1992). If it was determined that a streamgage record was not homogeneous, the entire record for that streamgage was not considered. However, if a significant portion of the record was determined to be homogeneous, the homogeneous portion of the record was considered for this study if the basin characteristics were representative of that portion of record. Topographic maps and aerial photographs were used to help determine if the cause of a positive trend in flood-peak magnitude was a result of increasing urbanization during the gaged period of record in the basin. The data quality review and trend analysis resulted in the identification of only eight urban streamgages for use in this study (pl. 1; table 11).

Flood-frequency estimates were computed for the eight urban streamgages by using general Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) guidelines. Station skew rather than weighted skew was used to determine flood frequency at the urban streamgages because of uncertainty about the value of regional skew for urban sites. The final flood-frequency estimates from the Bulletin 17B analysis for the eight urban streamgages in California are given in table 12.

Although the number of urban streamgages in California is too small to develop estimation equations using regression analysis, flood-frequency data for the eight urban sites was graphically compared to flood-frequency data for all 630 rural streamgages used in the regional regression analysis (fig. 7). The actual P-percent AEP flows are those quantile estimates determined using the LP3 analysis of the annual peak flows, and the predicted P-percent AEP flows are the quantile estimates from the regional regression equations. Figure 7 shows that the regression equations developed in this study for the rural streams generally tend to underpredict the flood estimates for streamgages affected by urbanization in regions 1 and 5, and adjustments are needed to account for the effects of urbanization in these regions. Additional data from urban streamgages likely would provide a better understanding of the effects of urbanization and enable the development of urban flood-frequency estimation equations for California. Sauer and others (1983) incorporated peak-flow data from 10 California urban streamgages into a nationwide study of flood characteristics for urban sites. Thus, methods developed by Sauer and others (1983) can be used to assess the effects of urbanization on California streams until additional urban flood data become available in California.

## StreamStats

StreamStats is a Web-based GIS that provides users with access to an assortment of analytical tools that are useful for water-resources planning and management, and for engineering design applications, such as the design of bridges. StreamStats allows users to easily obtain streamflow statistics, basin characteristics, and other information for user-selected sites on streams. StreamStats users can choose locations of interest from an interactive map and obtain information for these locations. If a user selects the location of a USGS streamgage, the user will be provided with a list of previously published information for the station. If a user selects a location where no data are available (an ungaged site), StreamStats will delineate the drainage-basin boundary, measure basin characteristics, and estimate streamflow statistics for the site. StreamStats also allows users to identify stream reaches that are upstream or downstream from user-selected sites and to identify and obtain information for locations along the streams where streamflow may be affected by human activities. Ries and others (2008) provide a detailed description of the 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.

Complete instructions for using StreamStats are provided through links on the StreamStats Web site at *http://water.usgs.gov/osw/streamstats/index.html*. The Web site also provides links to (1) information about general limitations of the application, (2) other State applications, (3) user instructions, (4) definitions of terms, (5) answers to frequently asked questions, (6) downloadable presentations and other technical information about the application, (7) information that can be accessed only by USGS employees, and (8) contact information.

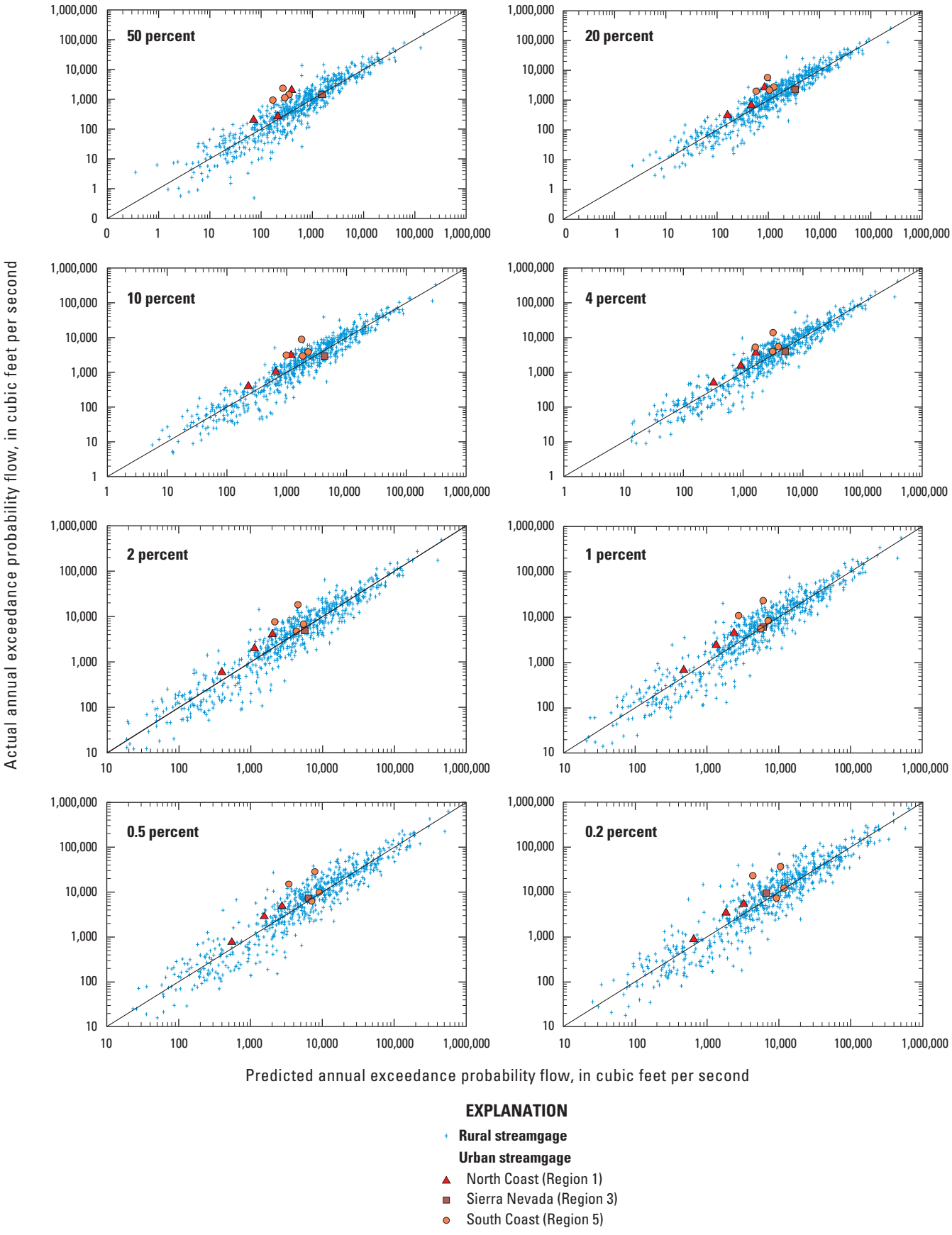


**Figure 7.** Actual and predicted annual exceedance probability flows for streamgages in California.

**Table 11.** Summary of streamgages with 10 or more years of record in urban areas of California, 2006.

[USGS, U.S. Geological Survey; GIS, geographic information system; $mi^2$, square miles]

| Map identification number (pl. 1) | USGS station number | Station name | Hydrologic region (pl. 1) | GIS derived drainage area ($mi^2$) | Mean annual precipitation (inches) |
|---|---|---|---|---|---|
| 772 | 11023330 | Los Penasquitos Creek below Poway Creek near Poway, CA | 5 | 31.3 | 15.7 |
| 773 | 11023340 | Los Penasquitos Creek near Poway, CA | 5 | 42.3 | 15.4 |
| 774 | 11047200 | Oso Creek at Crown Valley Pkwy near Mission Viejo, CA | 5 | 14.1 | 15.8 |
| 775 | 11120000 | Atascadero Creek near Goleta, CA | 5 | 19.0 | 21.9 |
| 776 | 11162720 | Colma Creek at South San Francisco, CA | 1 | 11.0 | 26.2 |
| 777 | 11162800 | Redwood Creek at Redwood City, CA | 1 | 1.77 | 24.8 |
| 778 | 11182500 | San Ramon Creek at San Ramon, CA | 1 | 5.89 | 24.8 |
| 779 | 11447360 | Arcade Creek near Del Paso Heights, CA | 3 | 31.7 | 23.8 |

| Map identification number (pl. 1) | Percentage of impervious area (percent) | Systematic record length (years) | Historical record length (years) | Period of historical record (water years) | Number of low outliers | Perception threshold discharges for missing peaks (cubic feet per second) |
|---|---|---|---|---|---|---|
| 772 | 17.1 | 12 | 12 | 1982–1993 | 0 | — |
| 773 | 20.0 | 25 | 25 | 1982–2006 | 0 | — |
| 774 | 31.9 | 11 | 11 | 1971–1981 | 0 | — |
| 775 | 10.3 | 16 | 16 | 1991–2006 | 0 | — |
| 776 | 37.3 | 14 | 15 | 1982–1996 | 0 | [5]7,120 |
| 777 | 14.7 | 38 | 38 | 1960–1997 | 9[u] | — |
| 778 | 39.2 | 54 | 54 | 1953–2006 | 0 | — |
| 779 | 42.3 | 12 | 12 | 1996–2006 | 0 | — |

[u] Low outliers censored using user defined low outlier threshold.

[5] Perception threshold set equal to twice the highest recorded discharge.

**Table 12.** Flood-frequency statistics for urban streamgages in California that were considered in the regression equations, 2006.

[USGS, U.S. Geological Survey]

| Map identification number (pl. 1) | USGS station number | Annual exceedance probability flow (cubic feet per second) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 50 percent | 20 percent | 10 percent | 4 percent | 2 percent | 1 percent | 0.5 percent | 0.2 percent |
| South Coast (Hydrologic Region 5) | | | | | | | | | |
| 772 | 11023330 | 1,130 | 2,180 | 2,960 | 3,980 | 4,750 | 5,510 | 6,270 | 7,260 |
| 773 | 11023340 | 1,430 | 2,770 | 3,870 | 5,480 | 6,830 | 8,300 | 9,890 | 12,200 |
| 774 | 11047200 | 940 | 1,990 | 3,120 | 5,290 | 7,630 | 10,800 | 15,100 | 23,000 |
| 775 | 11120000 | 2,350 | 5,750 | 8,900 | 13,900 | 18,200 | 23,100 | 28,600 | 36,700 |
| North Coast (Hydrologic Region 1) | | | | | | | | | |
| 776 | 11162720 | 2,220 | 2,910 | 3,360 | 3,920 | 4,330 | 4,730 | 5,140 | 5,690 |
| 777 | 11162800 | 223 | 344 | 430 | 542 | 629 | 717 | 808 | 931 |
| 778 | 11182500 | 292 | 733 | 1,110 | 1,660 | 2,100 | 2,560 | 3,020 | 3,650 |
| Sierra Nevada (Hydrologic Region 3) | | | | | | | | | |
| 779 | 11447360 | 1,460 | 2,260 | 2,940 | 3,990 | 4,940 | 6,050 | 7,350 | 9,400 |

# Summary and Conclusions

This report presents methods for determining flood magnitude and frequency at streamgages and ungaged sites in California. For this study, 771 streamgages in California were considered for use in the regional regression analysis. The streamgages used for this study have 10 or more years of peak-flow record that are not significantly affected by regulation or urbanization. Flood-frequency estimates were computed for the streamgages by using the expected moments algorithm to fit a Pearson Type III distribution to the logarithms of annual peak flows for each streamgage (Interagency Advisory Committee on Water Data, 1982). Low-outlier and historic information were incorporated in the analysis, and a generalized Grubbs-Beck test was used to detect multiple potentially influential low outliers. The station skew coefficients were weighted with the generalized skew coefficients developed by Parrett and others (2011) for the streamgages outside of the desert region of California. For the streamgages in the desert region, the station skew coefficients were weighted with the generalized skew coefficients developed by Thomas and others (1997). Because of short peak-flow record lengths and numerous zero flows/low outliers for many streamgages in the desert, the station mean and standard deviations also were weighted with a generalized mean and standard deviation developed using weighted least squares regression for the streamgages in the desert region.

Regional regression analysis, using generalized least squares regression, was used to develop a set of equations for estimating flows with 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probabilities for ungaged basins in California that are outside of the desert region. Flood-frequency estimates and basin characteristics for 630 rural streamgages were combined to form the final database used in the regional regression analysis. Five hydrologic regions were developed for the area of California outside of the desert region. The final equations are functions of drainage area, mean basin elevation, and mean annual precipitation for the Sierra Nevada region and functions of drainage area and mean annual precipitation for the other four regions. Average standard errors of prediction for these regression equations range from 42.7 to 161.9 percent.

For the desert region of California, a set of equations for estimating flows with 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent annual exceedance probabilities for ungaged basins was developed by directly combining the separate regional relations developed for the skew, standard deviation, and mean into the LP3 equation in Bulletin 17B (eq. 1; Interagency Advisory Committee on Water Data, 1982). The final equations are functions of drainage area. Average standard errors of prediction for these regression equations range from 214.2 to 856.2 percent. While the prediction errors are large in the desert region, they are smaller than comparable errors reported by Thomas and others (1997) for the Southwestern United States.

At 10 streamgages in the Sierra Nevada, the log-Pearson Type 3 (LP3) fit to all recorded peak flows deviated substantially from peak-flow data in the upper tail (larger annual peak flows), and a mixed-population flood-frequency analysis was considered for those streamgages. Nevertheless, the trial methods for fitting separate LP3 curves to annual peaks presumably caused by both rainfall and snowmelt and statistically combining the separate curves were not reliable. Thus, the LP3 curves based on the use of all at-site recorded peak-flow data were considered to provide the best flood-frequency information at the 10 gaged sites, given the general uncertainty about mixed populations and their analysis with the limited at-site data available.

Annual peak-flow data through water year 2006 were analyzed for eight streamgages in California having 10 or more years of data considered to be affected by urbanization. Flood-frequency estimates were computed for the urban streamgages by fitting a Pearson Type III distribution to logarithms of annual peak flows for each streamgage. Regression analysis could not be used to develop flood-frequency estimation equations for urban streams because of the limited number of sites. Flood-frequency estimates for the eight urban sites were graphically compared to flood-frequency estimates for 630 non-urban sites. Additional data from urban streamgages likely would provide a better understanding of the effects of urbanization and more accurate flood-frequency estimates for urban streams in California.

The regression equations developed in this study will be incorporated into the U.S. Geological Survey (USGS) StreamStats program. The StreamStats program is a Web-based application that provides streamflow statistics and basin characteristics for USGS streamgages and ungaged sites of interest. StreamStats can also compute basin characteristics and provide estimates of streamflow statistics for ungaged sites when users select the location of a site along any stream in California.

# References Cited


Brabec, E., Schulte, S., and Richards, P.L., 2002, Impervious surfaces and water quality—A review of current literature and its implications for watershed planning: Journal of Planning Literature, v. 16, no. 4, p. 499–514.

Busby, M.W., and Hirashima, G.T., 1972, Generalized streamflow relations of the San Bernardino and eastern San Gabriel Mountains, California: U.S. Geological Survey Open-File Report, 72 p.

Butler, E., Reid, J.K., and Berwick, V.K., 1966, Magnitude and frequency of floods in the United States, Part 10—The Great Basin: U.S. Geological Survey Water-Supply Paper 1684, 256 p.

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.

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.

Crippen, J.R., and Beall, R.M., 1971, A proposed streamflow-data program for California: U.S. Geological Survey Open-File Report 71–80, 46 p.

Cruff, R.W., and Rantz, S.E., 1965, A comparison of methods used in flood-frequency studies for coastal basins in California: U.S. Geological Survey Water-Supply Paper 1580–E, 56 p.

Durbin, T.J., 1974, Digital simulation of the effects of urbanization on runoff in the upper Santa Ana Valley, California: U.S. Geological Survey Water-Resources Investigations Report 41–73, 44 p.

Eng, Ken, Chen, Yin-Yu, and Kiang, J.E., 2009, User's guide to the weighted-multiple-linear-regression program (WREG version 1.0): U.S. Geological Survey Techniques and Methods, book 4, chap. A8, 21 p.

Feaster, T.D., and Guimaraes, W.B., 2004, Estimating the magnitude and frequency of floods in small urban streams in South Carolina, 2001: U.S. Geological Survey Scientific Investigations Report 2004–5030, 58 p.

Fenneman, N.M., and Johnson, D.W., 1946, Physiographic divisions of the conterminous U.S.: U.S. Geological Survey special map series, scale 1:7,000,000.

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

Griffis, V.W., and Stedinger, J.R., 2007a, The LP3 distribution and its application in flood frequency analysis, 2. Parameter estimation methods: Journal of Hydrologic Engineering, v. 12, no. 5, p. 492–500.

Griffis, V.W., and Stedinger, J.R., 2007b, The use of GLS regression in regional hydrologic analyses: Journal of Hydrology, v. 344, p. 82–95.

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.

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., World Environmental and Water Resources Congress 2007—Restoring our Natural Habitat: American Society of Civil Engineers, Tampa, Florida, May 15–19, 2007.

Gruber, A.M., and Stedinger, J.R., 2008, Models of LP3 regional skew, data selection, and Bayesian GLS regression, Paper 596, *in* Babcock, R., and Walton, R., eds., World Environmental and Water Resources Congress 2008—Ahupua'a: American Society of Civil Engineers, Honolulu, Hawaii, May 12–16, 2008.

Helsel, D.R., and Hirsch, R.M., 1992, Statistical methods in water resources: New York, Elsevier, 326 p.

Hodgkins, G.A., 1999, Estimating the magnitude of peak flows for streams in Maine for selected recurrence intervals: U.S. Geological Water-Resources Investigations Report 99–4008, 45 p.

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

Klein, R.D., 1979, Urbanization and stream quality impairment: Water Resources Bulletin, v. 15, no. 4, p. 948–963.

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.

Montgomery, D.C., Peck, E.A., and Vining, G.G., 2001, Introduction to linear regression analysis (3d ed.): New York, John Wiley and Sons, 641 p.

Neter, John, Wasserman, William, and Kutner, M.H., 1985, Applied linear statistical models: Homewood, Illinois, Richard D. Irwin, Inc., 1,127 p.

Parrett, Charles, Veilleux, Andrea, 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.

Rantz, S.E., 1971, Suggested criteria for hydrologic design of storm-drainage facilities in the San Francisco Bay region, California: U.S. Geological Survey Open-File Report 71–341, 69 p.

Reis, D.S., Jr., Stedinger, J.R., and Martins, E.S., 2005, Bayesian generalized least squares regression with application to the log-Pearson type 3 regional skew estimation: Water Resources Research, v. 41, W10419, doi:10.1029/2004WR003445, 14 p.

Ries, K.G., III, Guthrie, J.G., 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.

Ryberg, K.R., 2008, PFReports—A program for systematic checking of annual peaks in NWISWeb: U.S. Geological Survey Open-File Report 2008–1284, 17 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.

Sauer, V.B., Thomas, W.O., Jr., Stricker, V.A., and Wilson, K.V., 1983, Flood characteristics of urban watersheds in the United States: U.S. Geological Survey Water-Supply Paper 2207, 63 p.

Southard, R.E., 2010, Estimation of the magnitude and frequency of floods in urban basins in Missouri: U.S. Geological Survey Scientific Investigations Report 2010–5073, 27 p.

Stedinger, J.R., and Tasker, G.D., 1985, Regional hydrologic analysis 1. Ordinary, weighted, and generalized least squares compared: Water Resources Research, v. 21, no. 9, p. 1421–1432 [with correction, *in* Stedinger, J.R., and Tasker, G.D, 1986, Water Resources Research, v. 22, no. 5, p. 844].

Tasker, G.D., 1975, Combining estimates of low-flow characteristics of streams in Massachusetts and Rhode Island: U.S. Geological Survey Journal of Research, v. 3, no. 1, p. 107–112.

Tasker, G.D., 1980, Hydrologic regression with weighted least squares: Water Resources Research, v. 16, no. 6, p. 1107–1113.

Tasker, G.D., and Driver, N.E., 1988, Nationwide regression models for predicting urban runoff water quality at unmonitored sites: Water Resources Bulletin, 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, v.112, no. 2, p. 225–237.

Teal, M.J., and Gusman, A.J., 2007, Improved highway design methods for desert storms: California Department of Transportation, Division of Research and Innovation, 333 p.

Thomas, B.E., Hjalmarson, H.W., and Waltemeyer, S.D., 1997, Methods for estimating magnitude and frequency of floods in the Southwestern United States: U.S. Geological Survey Water-Supply Paper 2433, 205 p.

U.S. Department of Agriculture, 1986, Urban hydrology for small watersheds: Technical Release 55.

U.S. Geological Survey, 2010, WREG—Weighted-multiple-linear regression program: U.S. Geological Survey, accessed June 30, 2009, at *http://water.usgs.gov/software/WREG/*.

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, 170 p.

Waananen, A.O., and Crippen, J.R., 1977, Magnitude and frequency of floods in California: U.S. Geological Survey Water-Resources Investigations 77–21, 96 p.

Young, L.E., 1967, Magnitude and frequency of floods in the United States, Part II, Pacific slope basins in California, volume 2, Klamath and Smith River basins and Central Valley drainage from the east: U.S. Geological Survey Water-Supply Paper 1686, 308 p.

Young, L.E., and Cruff, R.W., 1967, Magnitude and frequency of floods in the United States, Part II, Pacific slope basins in California, volume 1, Coastal basins south of the Klamath River basin and Central Valley drainage from the west: U.S. Geological Survey Water-Supply Paper 1685, 272 p.

# Appendix. Parameter Estimation Method for the Desert Region of California

Flood-frequency analyses in the desert region of California are complicated because of short annual peak-flow records, numerous zero/low outliers, and highly variable peak-flow data for many streamgages. As recommended in Bulletin 17B (Interagency Advisory Committee on Water Data, 1982), flood-frequency estimates are improved by weighting the at-site skew with a more robust regional skew estimate. Because of the problems of the at-site peak-flow data, the at-site standard deviation and mean also were weighted with regional estimates of those parameters in the desert region. A description of the methods used to regionalize both the standard deviation and mean and their model error metrics for weighting purposes are described in this appendix.

Bulletin 17B (Interagency Advisory Committee on Water Data, 1982) recommends using the log-Pearson Type 3 (LP3) distribution to develop flood-frequency estimates. The LP3 distribution uses the method-of-moments to relate the mean, standard deviation, and skew to various flood quantiles. Discussion of the regionalization of all three parameters begins with the skew, where results from a previous study with some adjustment of the error term were used. The complex regional analysis of the standard deviation is discussed next, followed by a discussion of the regional analysis for the mean. Finally, this appendix describes the weighting procedures for combining at-site parameter estimates with the new regional values and new methods used to develop regional flood-frequency prediction equations and their regression diagnostics for ungaged basins in the desert region.

## Regional Skew Model

In a recent analysis of regional skew for California, Parrett and others (2011) found that regional skew could not be reliably determined in the hydrologically distinct desert region because of a lack of streamgages with sufficient long-term peak-flow records. In an earlier study of flood frequency in the Southwestern United States, Thomas and others (1997) used data from more than 1,000 streamgages in desert areas of several states to estimate regional skew and its variance for the multi-state study area. Thomas and others (1997) tried several methods for determining regional skew, including multiple regression analysis and kriging, and concluded that regional skew was a constant value of zero with an associated total error, or variance, of 0.31 log units. Although this regional skew analysis examined peak-flow records for many sites, the analysis did not account for the effects of different at-site record lengths and thus different reliabilities of at-site sample skew; nor did it account for cross correlation among at-site skew values. Newer Bayesian generalized least square (GLS) regression methods for estimating regional skew that were successfully applied in California outside the desert region (Parrett and others, 2011) and in the southeastern United States (Gotvald and others, 2009) have been shown to provide more reliable error metrics for regional skew than other methods, especially methods that ignore differences in at-site station skew reliability and cross correlations of station skews. Thus, for example, the mean square error (MSE) for the Bayesian GLS regression model for regional skew in California outside the desert region was 0.14 log units. In comparison, the MSE for the national skew map in Bulletin 17B is 0.303 log units. The error for the national skew map, like that reported by Thomas and others (1997) for the Southwestern United States, did not account for differences in at-site station skew reliability and cross correlations of station skews. On this basis, the MSE (0.31 log units) for the regional skew of zero reported by Thomas and others (1997) is believed to overestimate the true error for desert areas of the Southwestern United States, including the desert region of California. On the other hand, the MSE for skew (0.14 log units) reported by Parrett and others (2011) for the rest of California is probably too small for the desert region because of greater peak-flow variability in the desert region. An average of the two reported regional skew MSEs, rounded to one digit (0.2), was thus considered to be a reasonable representation of the true error for a constant skew model for the California desert region.

## Regional Regression Model for Standard Deviation

Tasker and Stedinger (1986) suggest using a weighted least squares (WLS) procedure for estimating regional standard deviation unless there is substantial cross correlation between at-site standard deviations, in which case a GLS regression should be used (Griffis and Stedinger, 2007b). The basic regression model for a regional standard deviation analysis for $n$ sites is:

$$\widetilde{\boldsymbol{\sigma}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1–1}$$

where

$\widetilde{\boldsymbol{\sigma}}$ is an ($n \times 1$) vector of the model estimates of at-site standard deviation for every station;

$\mathbf{X}$ is an ($n \times k$) matrix of $k$ basin characteristics with a column of ones corresponding to a constant in the model;

$\boldsymbol{\beta}$ is a ($k \times 1$) vector of model coefficients;

$\boldsymbol{\varepsilon}$ is the ($n \times 1$) vector of total errors, including both model and sampling errors where $E[\boldsymbol{\varepsilon}]=0$ and $\boldsymbol{\Lambda}$ is the covariance matrix that represents $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T]$.

The standard WLS or GLS estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\widehat{\boldsymbol{\sigma}} \tag{1–2}$$

where

$\boldsymbol{\Lambda}$ is the ($n \times n$) weighting covariance matrix; and

$\widehat{\boldsymbol{\sigma}}$ is the ($n \times 1$) vector of estimates of at-site standard deviation based on the sample data;

and all other terms are as previously described.

In WLS, which is a special case of GLS, the $\boldsymbol{\Lambda}$ matrix contains non-zero elements only on the diagonal, reflecting variability due to differences in record length (sampling variability). In GLS, the $\boldsymbol{\Lambda}$ matrix contains the same diagonal elements; however, the off-diagonal elements are also non-zero to reflect the cross correlation among the at-site standard deviation estimators $\hat{\sigma}_i$.

The matrix $\boldsymbol{\Lambda}$ can be decomposed into two covariance matrices (Reis and others, 2005), $\sigma_\delta^2\mathbf{I} + \boldsymbol{\Sigma}(\hat{\sigma})$, where $\sigma_\delta^2$ is the model error variance describing the precision with which the proposed model $\mathbf{X}\boldsymbol{\beta}$ (eq. 1–1) can predict the true standard deviations, which are denoted $\sigma_i$, and $\boldsymbol{\Sigma}(\hat{\sigma})$ represents the sampling variances and covariances of the standard deviation estimators $\hat{\sigma}_i$. The model error variance, $\sigma_\delta^2$, and $\hat{\boldsymbol{\beta}}$ values in equation 1–2 are jointly determined by iteratively searching for a positive-definite matrix $\boldsymbol{\Lambda}$ that satisfies the following equation (Stedinger and Tasker, 1985):

$$(\widehat{\boldsymbol{\sigma}} - \mathbf{X}\hat{\boldsymbol{\beta}})^T\,\boldsymbol{\Lambda}^{-1}\left(\widehat{\boldsymbol{\sigma}} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = n - \left(k+1\right) \tag{1–3}$$

When performing a least squares regional regression on standard deviation, the residuals, $\boldsymbol{\varepsilon}$, are equal to the at-site estimates, $\widehat{\boldsymbol{\sigma}}$, minus the regional (model) estimates, $\widetilde{\boldsymbol{\sigma}}$:

$$\boldsymbol{\varepsilon} = \widehat{\boldsymbol{\sigma}} - \widetilde{\boldsymbol{\sigma}} \tag{1–4}$$

Substituting equation 1–1 and 1–2 into equation 1–4 yields the following:

$$\boldsymbol{\varepsilon} = \widehat{\boldsymbol{\sigma}}\left\{\mathbf{I} - \mathbf{X}(\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\Lambda}^{-1}\right\} \tag{1–5}$$

As described previously, $\boldsymbol{\Lambda}$ is the covariance matrix, which is the sum of the model error variance and the sampling variance. The sampling variance is a function of the standard deviation, $\widehat{\boldsymbol{\sigma}}$, estimates and $\boldsymbol{\Lambda}$ for a WLS regression can be expressed as a function of $\widehat{\boldsymbol{\sigma}}$, together with some other terms, (see equations 1–9 and 1–10 for a detailed discussion of $\boldsymbol{\Lambda}$) as:

$$\boldsymbol{\Lambda}_{WLS} = \begin{bmatrix} f(\hat{\sigma}, etc) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f(\hat{\sigma}, etc) \end{bmatrix} \tag{1–6}$$

By substituting equation 1–6 into equation 1–5, the residuals can be expressed in terms of the at-site standard deviation estimates as:

$$\boldsymbol{\varepsilon} = \hat{\boldsymbol{\sigma}}\left\{\mathbf{I} - \mathbf{X}\left(\mathbf{X}^T\begin{bmatrix} f(\hat{\sigma},etc) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f(\hat{\sigma},etc) \end{bmatrix}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\begin{bmatrix} f(\hat{\sigma},etc) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & f(\hat{\sigma},etc) \end{bmatrix}^{-1}\right\} \quad (1\text{–}7)$$

Equation 1–7 violates a requirement of least squares regression that the residuals should not be correlated with the regressors, which in this case are the at-site standard deviation estimates $\hat{\boldsymbol{\sigma}}$. Thus, as recommended in Griffis and Stedinger, 2007b, a separate simple ordinary least squares (OLS) regression was used to estimate the at-site standard deviations by using the following equation:

$$\tilde{\boldsymbol{\sigma}}_{OLS} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad (1\text{–}8)$$

where

$\tilde{\boldsymbol{\sigma}}_{OLS}$ is an ($n \times 1$) vector of OLS model estimates of at-site standard deviation;

$\mathbf{X}\boldsymbol{\beta}$ is defined in equation 1–1; and

the error $\boldsymbol{\varepsilon}$ is normally distributed with zero mean and variance $\sigma_{\varepsilon}^2 = \sigma^2[10^{\ln(10)\sigma_{\delta}^2} - 1]$.

None of the basin characteristics in the desert region significantly described the variability in at-site standard deviations, $\hat{\sigma}_i$, using the OLS regression in equation 1–8. The best OLS model was a constant model with a value of 0.91 log units for $\tilde{\boldsymbol{\sigma}}_{OLS}$.

The values of $\hat{\Sigma}(\hat{\sigma})_{ii}$ and $\hat{\Sigma}(\hat{\sigma})_{ij}$ in $\boldsymbol{\Lambda}$ are determined by the length of record at each station, the regional standard deviation estimates from an OLS regression of at-site standard deviations, the regional skew, and the cross-correlation of the concurrent flows by the following equations (Griffis and Stedinger, 2007b):

$$\hat{\Sigma}(\hat{\sigma})_{ii} = Var(\hat{\sigma}) = \frac{1}{2}(1 + 0.75\tilde{\gamma}_i)\frac{\tilde{\sigma}_{OLS,i}^2}{m_i}, \text{ for } i = j \quad (1\text{–}9)$$

$$\hat{\Sigma}(\hat{\sigma})_{ij} = Cov[\hat{\sigma}_i, \hat{\sigma}_j] = \frac{1}{2}(\hat{\rho}_{ij} + 0.75\tilde{\gamma}_i\tilde{\gamma}_j)\hat{\rho}_{ij}\frac{m_{ij}\tilde{\sigma}_{OLS,i}\tilde{\sigma}_{OLS,j}}{m_i m_j}, \text{ for } i \neq j \quad (1\text{–}10)$$

where

$\tilde{\gamma}_i$ and $\tilde{\gamma}_j$ are the regional skew values at sites $i$ and $j$; the regional skew is assumed to be zero;

$\tilde{\sigma}_{OLS,i}$ and $\tilde{\sigma}_{OLS,j}$ are the estimated regional standard deviations from an OLS regression;

$m_i$ and $m_j$ are record lengths for sites $i$ and $j$;

$\hat{\rho}_{ij}$ is an estimated value for the cross correlation of the logs of concurrent peak flows at sites $i$ and $j$ (Tasker and Stedinger, 1986); and

$m_{ij}$ is the concurrent record length for sites $i$ and $j$.

Therefore, the covariance matrix $\boldsymbol{\Lambda}$ required for solving the general WLS or GLS equation 1–2 is estimated using:

$$\hat{\Lambda}_{\hat{\sigma},ii} = E[\tilde{\sigma}_{OLS,i}]^2[10^{\ln(10)\sigma_{\delta}^2} - 1] + \hat{\Sigma}(\hat{\sigma})_{ii} \quad \text{and} \quad \hat{\Lambda}_{\hat{\sigma},ij} = \hat{\Sigma}(\hat{\sigma})_{ij} \quad (1\text{–}11)$$

As described in Parrett and others (2011), the error variance ratio (EVR) is a modeling diagnostic used to evaluate whether a more sophisticated WLS or GLS analysis is appropriate to correctly interpret the data rather than a simple OLS regression. The EVR is the ratio of the average sampling error variance to the model error variance. An OLS regression is sufficient when the EVR is close to 20 percent. An EVR value of 550 percent was found for the constant WLS model, meaning the sampling variability of standard deviation estimators was considerably larger than model error variance of the regional model. Thus, given the wide variation in record lengths ranging from 11 to 72 years, use of a WLS or GLS analysis is important for evaluating the final precision of the model rather than using a simpler OLS model that neglects sampling error in the at-site standard deviation estimators.

The misrepresentation of the beta variance statistic, MBV*, is used to determine whether a WLS regression or a GLS regression is appropriate to determine the precision of the estimated regression parameters (Parrett and others, 2011). If the MBV* is substantially greater than 1, a GLS error analysis should be used. In this study, the MBV* value was 1, indicating that a WLS regression was sufficient and that cross correlation of the at-site standard deviation estimators was not significant.

The final WLS constant model and error metrics are given in table 1–1. Also given in table 1–1 for comparison purposes are the WLS model and metrics for the regression using *DRNAREA* as an explanatory variable. As WLS regressions use only the sampling error (eq. 1–9), the cross-correlation error is set to zero (eq. 1–10) in the $\boldsymbol{\Lambda}$ matrix. As indicated in equation 1–9, the sampling error is a function of the $\widetilde{\boldsymbol{\sigma}}_{OLS}$ estimates, which are constant, and the at-site record lengths. In essence, those sites with the longest record lengths had the most influence on the final regional regression model. The regional WLS constant model of 0.91 log units has a pseudo-$R_{\delta}^2$ of zero percent and a MSE ($AVP_{new}$) of 0.03 log units (table 1–1). The pseudo-$R_{\delta}^2$ is zero percent because a constant value of regional standard deviation does not describe any variability of at-site standard deviation estimators. The model that includes *DRNAREA* also has a pseudo-$R_{\delta}^2$ of zero percent, indicating that $\widehat{\boldsymbol{\sigma}}$ is not significantly correlated with *DRNAREA*. The MSE for the model with *DRNAREA* as an explanatory variable is 0.04 log units, which is greater than that for the constant model. Figure 1–1 shows the at-site standard deviation estimators for the 33 desert sites, the WLS regional regression model for a constant standard deviation (0.91 log units), and the WLS regional regression model relating standard deviation to *DRNAREA*.

**Table 1–1.** Regional standard deviation models for California.

[Constant is the linear regression model with a constant standard deviation; DRNAREA is the linear regression model relating standard deviation to *DRNAREA* (drainage area); $\beta_0$, regression model constant, $\beta_1$, regression model coefficient; $\sigma_{\delta}^2$, model error variance; ASEV, average sampling error variance; $AVP_{new}$, average variance of prediction for a new site; Pseudo-$R_{\delta}^2$, describes the fraction of the variability in the true standard deviations explained by each model (Gruber and other, 2007); standard deviations are in parentheses; —, not applicable]

| Model | Model equation | $\beta_0$ | $\beta_1$ | $\sigma_{\delta}^2$ | ASEV | $AVP_{new}$ | Pseudo-$R_{\delta}^2$ (percent) |
|---|---|---|---|---|---|---|---|
| Constant | $\widetilde{\sigma}_{WLS} = \beta_0$ | 0.91 (0.07) | — | 0.03 | 0.004 | 0.03 | 0 |
| DRNAREA | $\widetilde{\sigma}_{WLS} = \beta_0 + \beta_1[\log(DRNAREA)]$ | 0.89 (0.07) | 0.058 (0.06) | 0.03 | 0.009 | 0.04 | 0 |

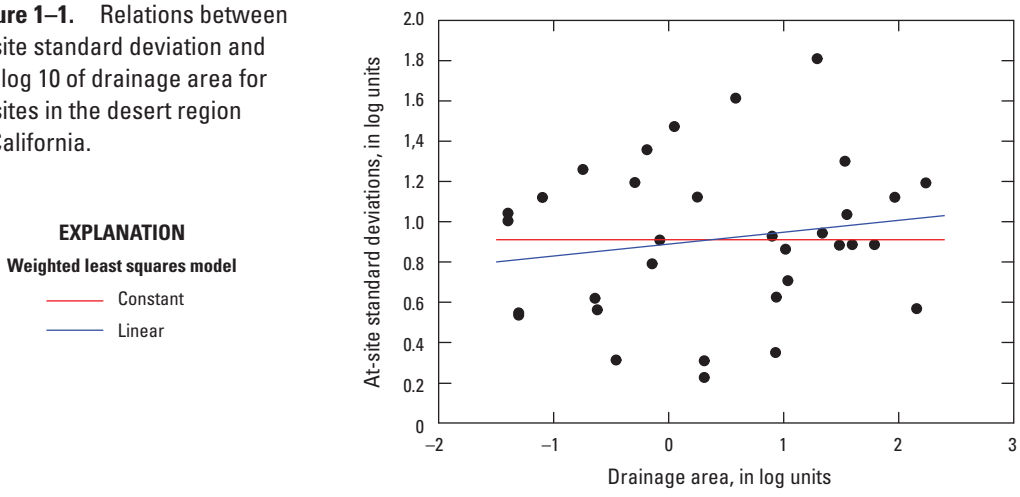


**Figure 1–1.** Relations between at-site standard deviation and the log 10 of drainage area for 33 sites in the desert region of California.

## Regional Regression Model for Mean

Traditionally, GLS regression is used to build a regional model for estimating a flood quantile, $\widetilde{\mathbf{Q}}_P$, with an exceedance probability $P$. If the quantile estimator is a function of drainage area, for example, then the regional quantile model for $\widetilde{\mathbf{Q}}_P$ would be

$$\log(\widetilde{Q}_P)_i = \beta_0 + \beta_1 \log(DRNAREA_i) + \varepsilon_i \tag{1–12}$$

where

$\beta_0$ and $\beta_1$ are the estimated regression parameters;
$DRNAREA_i$ is the drainage area at gage site $i$; and
$\varepsilon_i$ is the error associated with the model.

The generalized LP3 equation in Bulletin 17B for calculating at-site quantile is:

$$\log(\widetilde{Q}_P)_i = \hat{\mu}_i + K(\gamma_{W,i}, P) \times \hat{\sigma}_i \tag{1–13}$$

where

$\hat{\mu}_i$ is the at-site log-space mean;
$\hat{\sigma}_i$ is the at-site log-space standard deviation; and
$K(\gamma_{W,i}, P)$ is a factor that is a function of the weighted log-space skew coefficient $\gamma_{W,i}$ and the selected exceedance probability $P$.

Following guidelines in Bulletin 17B, the weighted log-space skew coefficient is determined by weighting the at-site skew $\hat{\gamma}_i$ and the regional skew $\tilde{\gamma}_i$ inversely proportional to their respective mean square errors by using the following equation:

$$\gamma_{W,i} = \frac{MSE_{\tilde{\gamma},i}(\hat{\gamma}_i) + MSE_{\hat{\gamma},i}(\tilde{\gamma}_i)}{MSE_{\tilde{\gamma},i} + MSE_{\hat{\gamma},i}} \tag{1–14}$$

where $MSE_{\hat{\gamma},i}$ and $MSE_{\tilde{\gamma},i}$ are the mean square errors of the at-site and regional skew, respectively. Thus, equations 1–12 and 1–13 can be combined to form

$$\log(\widetilde{Q}_P)_i = \hat{\mu}_i + K(\gamma_{W,i}, P) \times \hat{\sigma}_i = \beta_0 + \beta_1 \log(DRNAREA_i) + \varepsilon_i \tag{1–15}$$

As previously described, regional models for both skew $\tilde{\boldsymbol{\gamma}}$ and standard deviation $\tilde{\boldsymbol{\sigma}}$ were developed in the desert region of California, and both regional parameters need to be weighted with at-site parameters for determining more reliable flood quantile estimates. Both regional parameter estimates were determined to be constant. When weighting the at-site estimates of skew and standard deviation with the constant regional values using equation 1–14, the resulting weighted estimates were heavily influenced by the record length. In equation 1–14, the at-site MSEs for the skew and standard deviation are essentially inversely proportional to the record length. Therefore, three distinct types of weighted at-site estimates of the skew and standard deviation and their related MSEs were determined for sites with short, long, or intermediate record lengths. The resulting at-site quantile estimators for these three types of record lengths produced very different error structures than is typical.

For those sites with short record lengths, the weighted skew and standard deviations were more heavily influenced by the regional estimates ($\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\sigma}}$, respectively) than the at-site estimates ($\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\sigma}}$) and can generally be set equal to the regional skew and standard deviation ($\boldsymbol{\gamma}_W = \tilde{\boldsymbol{\gamma}}, \boldsymbol{\sigma}_W = \tilde{\boldsymbol{\sigma}}$). These weighted estimates can be inserted into equation 1–15 to produce

$$\log(\widetilde{Q}_P)_i = \hat{\mu}_i + K(\tilde{\gamma}, P) \times \tilde{\sigma} = \beta_0 + \beta_1 \log(DRNAREA_i) + \varepsilon_i \tag{1–16}$$

The term $[K(\tilde{\gamma},P)\times\tilde{\sigma}]$ is a constant for a given $P$ because $\tilde{\boldsymbol{\gamma}}$ and $\tilde{\boldsymbol{\sigma}}$ are constants. Therefore, equation 1–16 can be simplified to:

$$\hat{\mu}_i = \beta_0^{'} + \beta_1 \log(DRNAREA_i) + \varepsilon_i \tag{1–17}$$

where

$\beta_0^{'}$ is a new constant equal to $\beta_0$ minus $[K(\tilde{\gamma},P)\times\tilde{\sigma}]$; and

$\varepsilon_i$ is the sum of the model error for the mean $\sigma_{\delta,\hat{\mu}}$ plus the sample error for the mean $\sigma_{\eta,\hat{\mu}}$ $(\varepsilon=\sigma_{\delta,\hat{\mu}}+\sigma_{\eta,\hat{\mu}})$.

The variance of the error term is:

$$Var(\varepsilon_i) = \sigma^2_{\delta,\hat{\mu}} + (\sigma^2_{\eta,\hat{\mu}})_i \tag{1–18}$$

where

$\sigma^2_{\delta,\hat{\mu}}$ is the model error variance for equation 1–17; and

$(\sigma^2_{\eta,\hat{\mu}})_i$ is the sampling variance in the at-site mean estimator $\hat{\mu}_i$.

Equations 1–16 to 1–18 are applicable at short record sites where the regional estimates of skew and standard deviation have considerably more weight than the at-site estimates of standard deviation and skew, respectively. Conversely, for those sites with long-term streamgaging records, the at-site standard deviation and skew are more heavily weighted than the regional estimates. Therefore, in general, $\boldsymbol{\gamma}_W=\hat{\boldsymbol{\gamma}}, \boldsymbol{\sigma}_W=\hat{\boldsymbol{\sigma}}$, and these weighted estimates can be inserted into equation 1–15 to produce the following:

$$\log(\widetilde{Q}_P)_i = \hat{\mu}_i + K(\hat{\gamma}_i,P)\times\hat{\sigma}_i = \beta_0 + \beta_1\log(DRNAREA_i) + \varepsilon_i \tag{1–19}$$

where $\hat{\mu}_i, \hat{\sigma}_i$, and $\hat{\gamma}_i$ are at-site estimates of the mean, standard deviation, and skew, respectively. Because $[\hat{\mu}_i + K(\hat{\gamma}_i,P)\times\hat{\sigma}_i]$ is assumed to be an unbiased estimate of the true $Q_{P,i}$, $\varepsilon_i$ will be the error describing the difference between the regression equation $\log(\widetilde{Q}_P)_i$ and the true $\log\left(Q_P\right)_i$.

In this case, the estimated model error variance is

$$(\sigma^2_{\delta,Q_P})_i = \sigma^2_{\delta,\hat{\mu}_i} + E^2[K(\hat{\gamma}_i,P)]\times Var(\hat{\sigma}) + E^2[\hat{\sigma}_i^2]\times Var(\hat{\gamma})\times\left(\frac{dK(\hat{\gamma}_i,P)}{d\hat{\gamma}_i}\right)^2 \tag{1–20}$$

Thus, the $Var(\varepsilon_i)$

$$Var(\varepsilon_i) = (\sigma^2_{\delta,Q_P})_i + Var[\hat{\mu}_i + K(\hat{\gamma}_i\times\hat{\sigma}_i)] \tag{1–21}$$

The final possible scenario results when at-site record lengths are intermediate and the weighted skew and standard deviations are weighted averages of the at-site estimates and the regional estimates. In this case, $\log(\widetilde{Q}_P)_i$ is a weighted average of $[\hat{\mu}_i + K(\hat{\gamma}_i,P)\times\hat{\sigma}_i]$ and $[\hat{\mu}_i + K(\tilde{\gamma},P)\times\tilde{\sigma}]$, and $\varepsilon_i$ will have a variance between those indicated in equations 1–18 and 1–21.

Using GLS regression as described in the equations above is quite complicated. While GLS regression can accommodate the different sampling variances, it cannot easily accommodate multiple equations for estimating the model error variance. GLS regression can best estimate the model error variance when each equation has the same model error variance.

Thus, instead of regressing on $\log(\widetilde{Q}_P)$, it is possible to instead regress on the at-site mean $\hat{\mu}_i$. This allows the use of GLS to analyze the model

$$\tilde{\mu}_i = \beta_0 + \beta_1\log(DRNAREA_i) + \varepsilon_i \tag{1–22}$$

Now each streamgage in equation 1–22 has the same model error variance.

Similar regression techniques were used to regionalize the mean for California's desert region as were used for the standard deviation. The same basic regression model described in equation 1–1 was used. The **Λ** matrix is computed as the sum of two covariance matrices:

$$\hat{\Lambda}_{\hat{\mu},ii} = \sigma_{\delta}^2 + \frac{\hat{\sigma}_i^2}{m_i}, \qquad (i=j) \tag{1–23}$$

$$\hat{\Lambda}_{\hat{\mu},ij} = \frac{\hat{\rho}_{ij}\hat{\sigma}_i\hat{\sigma}_j m_{ij}}{m_i m_j}, \qquad (i\neq j) \tag{1–24}$$

(Eng and others, 2009) where the model error variance, $\sigma_{\delta}^2$, is described in equation 1–2, $\hat{\sigma}_i$ and $\hat{\sigma}_j$ are the at-site standard deviation estimators from EMA using the multiple Grubbs-Beck test, and $\hat{\rho}_{ij}, m_{ij}, m_i$, and $m_j$ are defined in equations 1–9 and 1–10.

As described in the regionalization of standard deviation, both WLS and GLS regression models were initially considered for regionalization of the mean. An EVR value of 117 percent and a MBV* value of 1.2 were determined. These two regression diagnostics indicate that WLS regression was required for evaluating the final precision of the model error rather than an OLS regression and that the GLS model was not needed because of a lack of cross correlation between the at-site mean estimators.

Unlike the constant regional skew and constant standard deviation models for the desert region, one basin characteristic, drainage area (*DRNAREA*), was determined to be statistically significant at explaining the site-to-site variability in mean. The constant WLS model for regional mean had a pseudo-$R_{\delta}^2$ value of zero percent (table 1–2), while the model based on *DRNAREA* had a pseudo-$R_{\delta}^2$ value of 51 percent. The model error variance, $\sigma_{\delta}^2$, was 0.61 log units for the constant model and only 0.30 log units for the model using *DRNAREA*. Finally, the MSE ($AVP_{new}$), which describes the precision of the regional mean, was 0.63 log units for the constant model and only 0.32 log units for the model using *DRNAREA*. Figure 1–2 shows the relation between the at-site mean estimators and the WLS constant model and the model using *DRNAREA*.

**Table 1–2.** Regional mean models for California.

[Constant is the linear regression model with a constant mean; DRNAREA is the linear regression model relating mean to *DRNAREA* (drainage area); $\beta_0$, regression model constant; $\beta_1$, regression model coefficient; $\sigma_{\delta}^2$, model error variance; ASEV, average sampling error variance; $AVP_{new}$, average variance of prediction for a new site; Pseudo-$R_{\delta}^2$, describes the fraction of the variability in the true mean explained by each model (Gruber and others, 2007); standard deviations are in parentheses; —, not applicable]

| Model | Model equation | $\beta_0$ | $\beta_1$ | $\sigma_{\delta}^2$ | ASEV | $AVP_{new}$ | Pseudo-$R_{\delta}^2$ (percent) |
|---|---|---|---|---|---|---|---|
| Constant | $\tilde{\mu}_{WLS} = \beta_0$ | 1.20 (–0.14) | — | 0.61 | 0.02 | 0.63 | 0 |
| DRNAREA | $\tilde{\mu}_{WLS} = \beta_0 + \beta_1[\log(DRNAREA)]$ | 1.01 (–0.11) | 0.51 (0.09) | 0.3 | 0.02 | 0.32 | 51 |

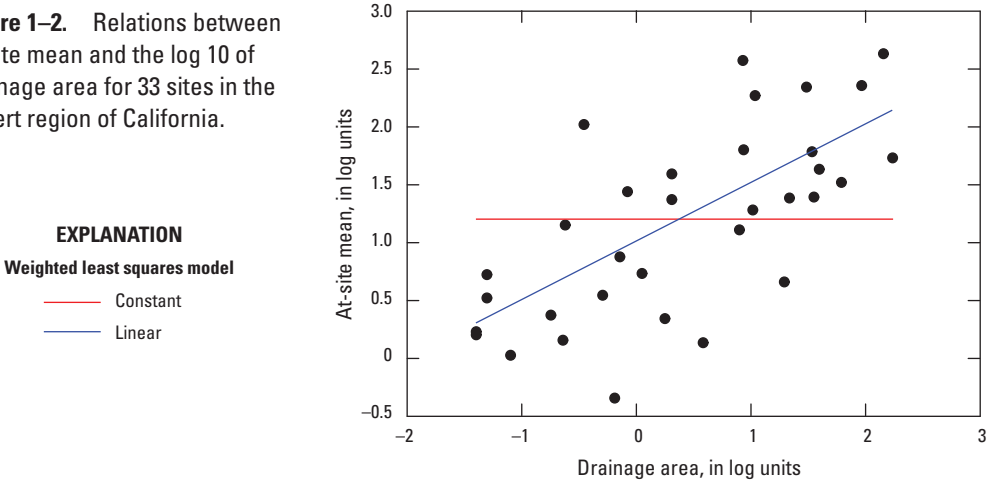


**Figure 1–2.** Relations between at-site mean and the log 10 of drainage area for 33 sites in the desert region of California.

## Equations for Estimating Flood Frequency at Ungaged Sites

As described previously in this appendix, equations for estimating flood frequency at ungaged sites are commonly developed from a regional regression analysis that relates at-site flood quantiles at gaged sites to basin characteristics at the gaged sites (eq. 1–13). For the desert region of California, however, the separate regional relations developed for the skew, standard deviation, and mean can be combined directly in a form that results in estimation equations for ungaged sites. Thus, the basic LP3 equation for calculating at-site flood frequency described in Bulletin 17B can be expressed in terms of the regional mean, standard deviation, and skew as:

$$\log(Q_P)_i = \tilde{\mu}_i + K(\gamma_{W,i}, P) \times \tilde{\sigma}_i, \tag{1–25}$$

where

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

$\tilde{\mu}_i$ is the WLS model of the regional mean: $\tilde{\mu}_{WLS,i} = 1.01 + 0.51 \times \log(DRNAREA)_i$;

$K(\gamma_{W,i}, P)$ is the $K$ factor obtained from appendix 3 in Bulletin 17B that relates a regional skew of zero to the probability of exceedance, $P$; and

$\tilde{\sigma}_i$ is the WLS constant model of the regional standard deviation, 0.91 log units.

The resulting regional equations for estimating annual peak flows with 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2- percent annual exceedance probabilities (AEP) in the desert region of California are given in table 5.

The MSE for the regional regression equations described in equation 1–25 can be expressed as follows:

$$MSE_{\log Q_P} = Var[\tilde{\mu}\{\log(DRNAREA)\}] + E^2[K(\tilde{\gamma}, P)] \times Var[\tilde{\sigma}] + E^2[\tilde{\sigma}] \times Var[K(\tilde{\gamma}, P)], \tag{1–26}$$

where

$Var[\tilde{\mu}\{\log(DRNAREA)\}]$ equals 0.32 log units;

$E[K(\tilde{\gamma}, P)]$ is the $K$ factor obtained from appendix 3 in Bulletin 17B that relates a regional skew of zero to the probability of exceedance, $P$;

$Var[\tilde{\sigma}]$ equals 0.03 log units;

$E[\tilde{\sigma}]$ equals 0.91 log units;

$Var[K(\tilde{\gamma}, P)]$ is approximated by: $Var[\tilde{\gamma}] \times \left\{ \frac{dK(\tilde{\gamma}, P)}{d\tilde{\gamma}} \right\}^2$;

$Var[\tilde{\gamma}]$ equals 0.20 log units; and

$\left\{ \frac{dK(\tilde{\gamma}, P)}{d\tilde{\gamma}} \right\}^2$ is approximated by: $\frac{\left(Z_P^2 - 1\right)}{6}$, when $\tilde{\gamma} = 0$ (Eng and others, 2009).

The values of the MSE for the regional regression equations for ungaged sites in California's desert region are listed in table 6.

Manuscript approved on June 5, 2012

Edited by Kay P. (Hedrick) Naugle
Illustrations and layout by Caryl J. Wipperfurth
Science Publishing Network, Raleigh PSC

For more information about this publication, contact:
USGS California Water Science Center
Placer Hall
6000 J Street
Sacramento, CA 95819
*http://ca.water.usgs.gov/*