science for a changing world

Methods for Estimating Regional Coefficient of Skewness for Unregulated Streams in New England, Based on Data Through Water Year 2011

Scientific Investigations Report 2017–5037

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

Methods for Estimating Regional Coefficient of Skewness for Unregulated Streams in New England, Based on Data Through Water Year 2011

By Andrea G. Veilleux, Phillip J. Zariello, Glenn A. Hodgkins, Elizabeth A. Ahearn, Scott A. Olson, and Timothy A. Cohn

Scientific Investigations Report 2017–5037

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

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

U.S. Geological Survey
James Reilly II, Director

U.S. Geological Survey, Reston, Virginia: 2019

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

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

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

ISSN 2328-0328 (online)

Contents

Figures

Tables

Conversion Factors

U.S. customary units to International System of Units

Multiply	By	To obtain
inch (in.)	2.54	centimeter (cm)
foot (ft)	0.3048	meter (m)
mile (mi)	1.609	kilometer (km)

Datum

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

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

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

Abbreviations

AEP	annual exceedance probability
B17B	Bulletin 17B (Interagency Advisory Committee on Water Data, 1982)
B–GLS	Bayesian generalized least squares
B–WLS	Bayesian weighted least squaresEMA expected moments algorithm
GAGES II	Geospatial Attributes of Gages for Evaluating Streamflow, version II
GLS	generalized least squares
LP3	log-Pearson type III [distribution]
MGB	multiple Grubbs-Beck [test]
MSE	mean squared error
OLS	ordinary least squares
PILF	potentially influential low flood
USGS	U.S. Geological Survey
WLS	weighted least squares

Methods for Estimating Regional Coefficient of Skewness for Unregulated Streams in New England, Based on Data Through Water Year 2011

By Andrea G. Veilleux, Phillip J. Zarriello, Glenn A. Hodgkins, Elizabeth A. Ahearn, Scott A. Olson, and Timothy A. Cohn

Abstract

The magnitude of annual exceedance probability floods is greatly affected by the coefficient of skewness (skew) of the annual peak flows at a streamgage. Standard flood frequency methods recommend weighting the station skew with a regional skew to better represent regional and stable conditions. This study presents an updated analysis of a regional skew for New England developed using a robust Bayesian weighted and generalized least squares regression model. Nineteen explanatory variables for 153 streamgages were tested in the regression analysis, but none were statistically significant and, as a result, a constant model was selected to define the regional skew for New England. The constant model for the New England region has, in log units, a skew of 0.37, a model error variance of 0.13, and an average variance of prediction at a new site of 0.14. An assessment of the selected regional skew model was conducted using a Monte Carlo analysis. The Monte Carlo simulations reveal that the perceived pattern in the station skews among the 153 streamgages is an artifact of the sample variability and the covariance structure of the errors.

Introduction

Bulletin 17B (B17B; Interagency Advisory Committee on Water Data, 1982) provides a consistent, objective, reproducible method for flood frequency analysis; B17B recommends the use of the log-Pearson type III (LP3) distribution to fit a series of annual maximum flood peaks to obtain estimates of annual exceedance probability (AEP) discharges at streamgages. This distribution, in the specific case of flood frequency analysis, is described by three moments: the mean, the standard deviation, and the skewness coefficient (skew) of the logarithms of the flow. This third moment, skew, is a measure of the asymmetry of the annual peak flow distribution and can have a large effect on the estimated magnitude of floods for a given AEP. The traditional sample estimator of skew is very sensitive to extreme events, such as large floods or unusually small values, as they cause a sample to be highly skewed, or asymmetrical. Thus, in flood frequency analysis, skew becomes important because interest is focused on the right-hand tail (largest flows) of the distribution. B17B recommends using a weighted average of the station skew and a regional skew; a weighted skew reduces the sensitivity of station skew to extreme events, particularly for streamgages with short records. In addition, B17B supplies a map (Interagency Advisory Committee on Water Data, 1982, pl. I) containing estimated generalized (or regional) skews for the Nation but recommends using additional methods for developing more accurate estimates of regional skew. Since the map was published in 1976, more than 35 years of additional streamflow data have been accumulated and spatial estimation procedures have been refined (Stedinger and Griffis, 2008).

Tasker and Stedinger (1986) developed a weighted least squares (WLS) procedure for estimating regional skews based on station skews for the logarithms of annual peak discharge data. This method of regional analysis of skew accounts for the precision of the estimated station skew for each streamgage, which depends on the length of the streamgage record and the accuracy of the ordinary least squares (OLS) regional mean skew. More recently, Reis and others (2005), Gruber and others (2007), and Gruber and Stedinger (2008) developed Bayesian generalized least squares (B–GLS) regression models for regional skew analyses. A Bayesian methodology allows for the computation of a posterior distribution of both the regression parameters and the model error variance. When the model error variance is small compared with the sampling error of the station estimates (Reis and others, 2005), the Bayesian posterior distribution provides a more robust description of the model error variance than both the generalized least squares (GLS) method-of-moments and maximum-likelihood point estimates (Veilleux, 2011). WLS regression accounts for the precision of the regional model and the effect of the length of the streamgage record on the variance of skew estimators, whereas B–GLS regression also considers the cross correlations among the estimated station skews. Cross correlations have had a large effect on the precision attributed to estimates of different parameters (Feaster and others, 2009; Gotvald and others, 2009; Weaver and others, 2009; Parrett and others, 2011).

The expected moments algorithm (EMA) with multiple Grubbs-Beck (MGB) censoring of low outliers (Cohn and others, 2013) was used to perform flood frequency analysis at streamgages; however, this methodology introduces a complication in the calculation of streamgage record length (and concurrent record length) used to describe the precision of sample estimators. In addition to censoring of low outliers, the EMA uses estimated interval discharges for missing, censored, and historic data because the peak discharges are no longer solely represented by single values. Further, large cross correlations between annual peak discharges at pairs of streamgages can introduce a bias if not properly accounted for. To account for these complications, an alternate regression procedure was developed to provide stable, unbiased results for determining regional skew (Veilleux, 2011; Lamontagne and others, 2012; Veilleux and others, 2012). This alternate procedure is referred to as the Bayesian WLS/Bayesian GLS (B–WLS/B–GLS) regression framework (Veilleux, 2011; Veilleux and others, 2011, 2012). The B–WLS/B–GLS procedure uses an OLS analysis to fit an initial regional skew model that is then used to generate a stable regional skew estimate for each streamgage. That stable regional estimate is the basis for computing the variance of each station skew employed in the WLS analysis. Next, the B–WLS procedure is used to generate estimates of the regional skew model parameters. Finally, the B–GLS procedure is used to estimate the precision of the WLS parameter values, to estimate the model error variance and its precision, and to compute various diagnostic statistics.

Purpose and Scope

The primary purpose of this report is to describe the methods and the results of a regional skew analysis for New England (Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, and Vermont). This report also includes details of the existing data for the 186 streamgages that were evaluated for use in the regional skew analysis, including annual peak discharges, as well as selected basin characteristics. The EMA with MGB censoring of potentially influential low floods was used to compute moments of the logarithms for the LP3 distribution to determine an at-site skew at each streamgage for use in the regional skew analysis. The B–WLS/B–GLS regression method is described in detail along with diagnostics for analyzing the B–WLS/B–GLS results.

Study Area

The study area for the New England regional skew analysis consists of the entire States of Connecticut, Maine, Massachusetts (excluding Cape Cod and the islands), New Hampshire, Rhode Island, Vermont, and the eastern most part of New York. The region spans about 500 miles from southeastern Connecticut to northern Maine. A total of 186 streamgages were evaluated for use in the analysis (fig. 1). Streamgage basin characteristics described below were obtained from the Geospatial Attributes of Gages for Evaluating Streamflow version II (GAGES II) database (Falcone, 2011).

The geomorphology of New England is described as hills and mountains that slope into valleys or coastal lowlands (Denny, 1982). Mountainous areas (elevations greater than 1,600 feet [ft]) are found in the Green Mountains of Vermont, the White Mountains of New Hampshire, and isolated peaks such as Mount Katahdin in Maine, all of which are part of the Northeastern Highlands level III ecoregion (U.S. Environmental Protection Agency, 2012, 2013); this ecoregion consists of most of Vermont, New Hampshire, western Maine, western Massachusetts, and northwestern Connecticut. The Northeastern Coastal Zone ecoregion consists of all of Rhode Island, most of Connecticut, central and eastern Massachusetts, southeastern New Hampshire, and far southeastern Maine, and central-eastern and northeastern Maine make up the Acadian Plains and Hills ecoregion. The Northeastern Coastal Zone and Acadian Plains and Hills have gently rolling hills with incised valleys. Overall, the mean basin elevation at the streamgages used in the analysis ranged from 25 to 1,005 ft (average 273 ft), relief ranged from 36 to 1,813 ft (average 436 ft), and basin slopes ranged from 1.2 to 35 percent (average 8.4 percent). The study area excludes southeasternmost Massachusetts and Cape Cod because they are in the Atlantic Coastal Pine Barrens ecoregion, which is characterized by relatively low relief with deep coarse-grained deposits and a water-table surface at depth. These characteristics cause streamflow to respond uniquely to storms from other parts of New England and, in addition, little streamflow information is available in this area to support inclusion into the New England regional skew model.

Climate in New England is characterized as being humid continental with temporal and spatial variations caused by the interchange of air masses from the South and the Midwest and from the north (National Oceanic and Atmospheric Administration, 2013). The exchange of air masses leads to frontal systems that, coupled with wet antecedent conditions, often produce the peak flows of the year. Isolated intense storms from these systems also produce flooding, particularly in small basins. Hurricanes, remnants of hurricanes, and tropical storms that never develop to hurricane strength are major causes of flooding in southern New England (National Oceanic and Atmospheric Administration, undated). These storms typically originate in the central Atlantic Ocean and follow a track along the eastern United States up through New England.

In general, precipitation is evenly distributed throughout the year. On the basis of data collected between 1981 and 2010 (Northeast Regional Climate Center, 2014), mean annual

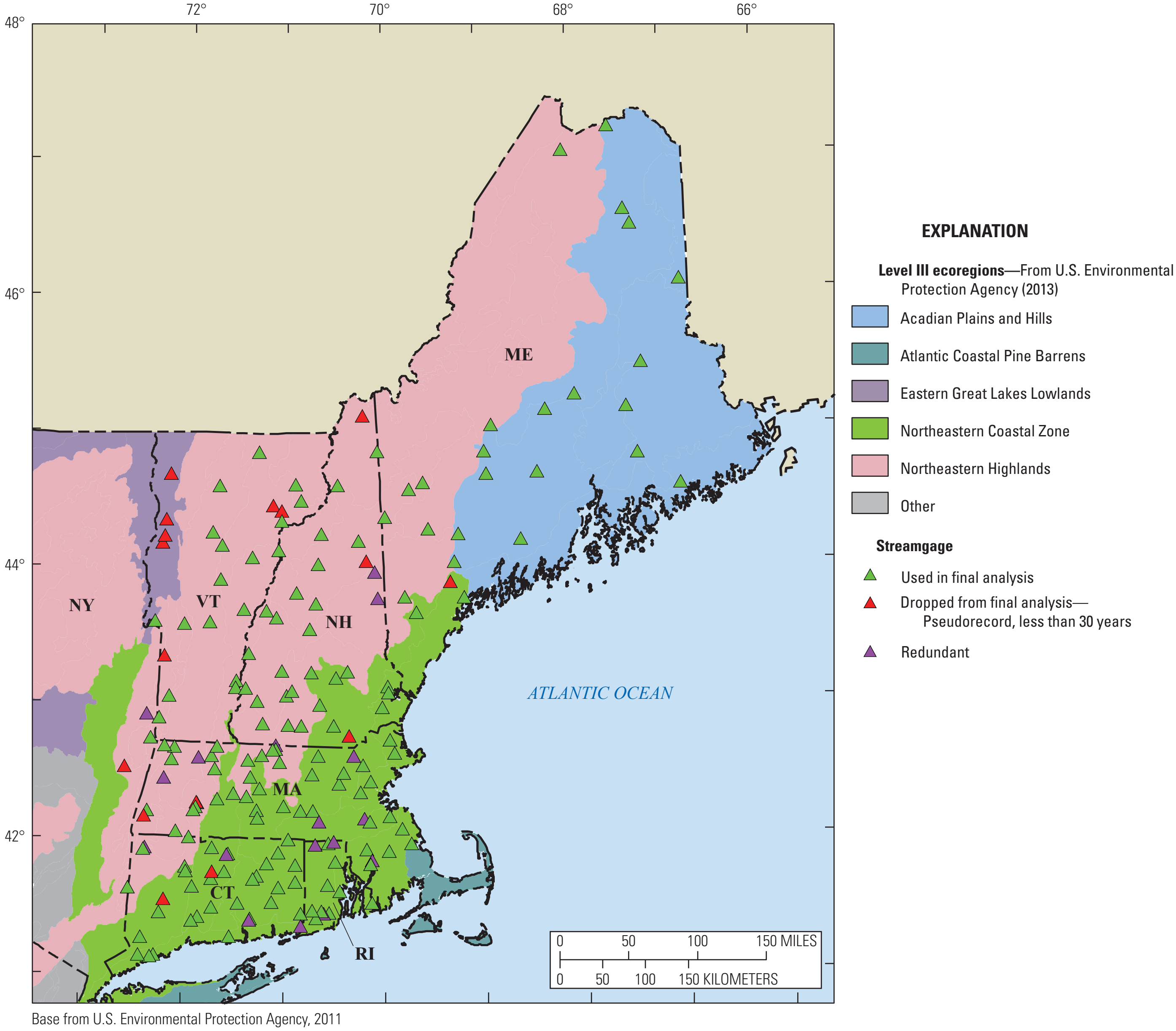


Figure 1. Streamgages used in the regional skew analysis for New England. Streamgages that were not used in the final analysis because of redundancy or insufficient pseudo record length are also shown.

precipitation ranged from 38 inches per year in northern Maine to 50 inches per year in southern New England. Mean annual temperatures ranged from 38 degrees Fahrenheit (°F) in northern Maine to 50 °F in southern New England. Because of the combination of weather conditions, more precipitation falls as snow and accumulates during the winter months in northern New England (about 20 to 120 inches [in.]). Snowmelt combined with late-winter or spring rainfall often yields the highest annual peak flows of the year in the northern part of New England. Mean precipitation for streamgage basins in the study area ranged from 37 to 74 in. (average 48 in.), and the percentage of precipitation that fell as snow ranged from 20 to 44 percent (average 31 percent).

Land use and land cover in New England range from highly developed in and near major metropolitan centers to predominantly forest. Most heavily developed areas are in eastern and southern New England; basins considered highly urbanized were excluded from the study. Land use in the streamgage basins used in the study ranged from 13 to 99 percent (average 69 percent) forested, from 0 to 12 percent (average 0.5 percent) high density urban, 0 to 37 percent (average 2.4 percent) medium-density urban, and 0 to 25 percent (average 3.9 percent) low-density urban. Basin area classified as open water and wetlands ranged from 0 to 7.7 percent (average 1.5 percent) and 0 to 24 percent (average 6.8 percent), respectively (Falcone, 2011).

Streamgage Data for Regional Skew Analysis

Data preparation for the regional skew analysis included selection of streamgages and corresponding basin characteristics as well analyses of AEPs to estimate the skew at each streamgage. The AEP analyses were performed using the EMA with the MGB test for potentially influential low floods.

Streamgage Selection and Basin Characteristics

The New England regional skew analysis used data for 186 U.S. Geological Survey (USGS) streamgages (table 1, in back of report; Wagner and Veilleux, 2019) that have essentially unregulated annual peak flow records through water year[1] 2011; in total, data were used for 36 streamgages in Connecticut, 26 streamgages in Maine, 51 streamgages in Massachusetts, 43 streamgages in New Hampshire, 13 streamgages in Rhode Island, 14 streamgages in Vermont, and 3 streamgages in New York. Annual peak flow data for these streamgages were obtained from the USGS National Water Information System (U.S. Geological Survey, 2014a). Only those streamgages included in the GAGES II database maintained by the USGS were considered for the study. The GAGES II database consists of a subset of USGS streamgages with at least 20 years of discharge record since 1950 or that were active as of water year 2009 and whose watersheds lie entirely within the United States (Falcone, 2011). Only streamgages that are included in the GAGES II database in the New England regional skew study were used because this subset provided a consistent set of basin characteristics across the region. The basin characteristics selected for use to potentially help explain the variation in station skew in the New England study region include morphometric (drainage area, mean basin slope, mean basin elevation, and basin compactness ratio), climatological (mean annual precipitation, average annual air temperature, and snow percentage of total precipitation), geologic, land cover (including percent open water, percent lakes/ponds/reservoirs, percent emergent herbaceous wetlands, percent woody wetlands, percent impervious surface; percent of watershed covered by dominant potential natural vegetation) , hydrologic (sinuosity of mainstem stream line, and base flow index), and other characteristics. In addition to basin characteristics obtained from the GAGES II database, the 50th percentile maximum seasonal snow depth was determined from Cember and Wilks (1993, map 56) and the flood flow peaks caused by hurricanes were identified from local records and knowledge.

[1]A water year is the 12-month period beginning October 1 and ending September 30. It is designated by the year in which it ends.

Redundancy Analysis

Redundancy results when the drainage basins of two streamgages are nested, meaning that one is contained inside the other, and also of similar size. Instead of providing two spatially independent observations that depict how drainage basin characteristics are related to skew (or the magnitude of the AEP flow), these two basins will have exceedingly similar hydrologic responses to a given storm and thus represent only one spatial observation. When redundant streamgages are both included, a statistical analysis incorrectly represents the value of the information contained in the regional dataset (Gruber and Stedinger, 2008). To determine if two streamgages are redundant and thus represent the same hydrologic conditions, two types of information are considered: whether the basins of the streamgages are nested and the ratio of the drainage areas for the two basins.

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

$$SD_{ij} = \frac{D_{ij}}{\sqrt{0.5\left(DRNAREA_i + DRNAREA_j\right)}} \tag{1}$$

where

D_{ij} is the distance between centroids of basin *i* and basin *j*,

$DRNAREA_i$ is the drainage area at streamgage *i*, and

$DRNAREA_j$ is the drainage area at streamgage *j*.

The drainage area ratio (*DAR*) is used to determine if two nested basins are sufficiently similar in size to conclude that they are essentially or are at least in large part the same watershed for the purposes of developing a regional hydrologic model. The *DAR* (Veilleux, 2009) is defined as follows:

$$DAR = Max\left[\frac{DRNAREA_i}{DRNAREA_j}, \frac{DRNAREA_j}{DRNAREA_i}\right] \tag{2}$$

where

Max is the maximum of the two values in brackets,

$DRNAREA_i$ is the drainage area at streamgage *i*, and

$DRNAREA_j$ is the drainage area at streamgage *j*.

Previous studies suggest that, for the purposes of determining regional skew, streamgage pairs having *SD* less than or equal to 0.50 and *DAR* less than or equal to 5 were likely to be redundant. However, if the drainage area ratio is large enough, even if the basins of the streamgages are nested, then they will reflect different hydrologic responses because storms of different sizes and duration will typically affect each site differently.

Table 1 shows the results of the redundant streamgage screening on the New England regional skew data. Note that the information in table 1 is also available as USGS data release (Wagner and Veilleux, 2019).

All possible combinations of streamgage pairs from the 186 streamgages were considered in the redundancy analysis. In order to be conservative, all streamgage pairs with basin characteristics of $SD < 0.75$ and $DAR < 8$ were identified as possible redundant streamgage pairs. All streamgages identified as possibly redundant were then investigated to determine if, in fact, the basin of one streamgage of the pair is nested inside the other. The procedure identified 37 possible redundant streamgage pairs; of these, 34 streamgage pairs were found to be redundant and 17 streamgages were removed from the analysis (table 1, in back of report; streamgages indicated as “No-R”). The other three streamgage pairs identified as potentially redundant were determined to be independent and kept in the analysis, leaving 169 streamgages for the regional skew study.

Analytical Methods To Generate Regional Skew

The skew is very sensitive to extreme events, such as large floods, because they can cause the position of a sample within the distribution to be highly skewed, or asymmetrical. Thus, in flood frequency analysis, the measurement of skew becomes significant and interest is focused on the right-tail of the distribution. However, the span of recorded flood data at a given streamgage is usually too short to provide a highly reliable estimate of the skew. To improve the precision of the skewness estimator, B17B (Interagency Advisory Committee on Water Data, 1982) advises combining a regional skew with the at-site skew (Beard, 1974; Hardison, 1975; Tasker, 1978; McCuen, 1979, 2001; Griffis and Stedinger, 2007b). Griffis and Stedinger (2009b) showed that the B17B weighted skewness estimator (inverse weighting using the mean squared error [MSE] of the skew) results in the estimator with the smallest MSE, provided that the regional skew is unbiased and independent of the at-site skew estimator. Griffis and Stedinger (2007a, 2009a) illustrate the value of a good regional skewness estimator in terms of the precision of flood quantile estimates (Veilleux, 2009).

To improve the regional skew estimate for New England, the at-site skew and its MSE for each streamgage to be used in the regional analysis were determined. For this analysis PeakFQ version 7.0 (Veilleux and others, 2014; U.S. Geological Survey, 2014b) was used to compute the flood frequency analysis because it combines the EMA with the MGB test as an efficient and robust means for estimating skew and MSE.

Expected Moments Algorithm (EMA) Analysis

The EMA with the MGB test was used to fit the LP3 and compute AEPs and correspondingly the at-site skew and its MSE for the streamgages included in the regional skew analysis. The EMA with the MGB test flood frequency analysis includes at-site estimates of AEP flood flows, which are not included in the report because the focus of this study is to update the regional skews for New England. Logically, however, the next step following a regional skew analysis would be to run the EMA with the MGB test analyses again, weighting the at-site skew with the revised regional skews.

The EMA addresses several methodological concerns identified in the procedures specified by B17B (Interagency Advisory Committee on Water Data, 1982) while retaining the essential structure and moments-based approach of the existing B17B procedures for determining flood frequency. The EMA can accommodate interval data, which simplifies the analysis of datasets containing censored observations, historic data, low outliers, and uncertain data points, while also providing enhanced confidence intervals on the estimated discharges (Veilleux and others, 2014). Unlike B17B, which recognizes two categories of data [systematic peaks (annual peaks observed in the course of the systematic operation of a streamgage) and historic peaks (records of floods that occurred outside the period of regular streamgage operation)], the EMA employs a more general description of flood information for the historical period that includes both systematic and historic peaks (Veilleux and others, 2014). This is accomplished through the use of flow intervals to describe the knowledge of the peak flow Q_Y in each year Y and through the use of perception thresholds to describe the range of measurable potential discharges in each year Y. It is important to note that for streamgages that have complete periods of record, no low outliers, no censored flood values, and no historic flood information, the EMA with the MGB test method provides identical estimates of the three LP3 moments (mean, standard deviation, and skew) as the standard LP3 method described in B17B (Interagency Advisory Committee on Water Data, 1982; Gotvald and others, 2012). A complete description and application of the algorithm is given in Cohn and others (1997, 2001).

Multiple Grubbs-Beck Test for Detecting Potentially Influential Low Outliers

B17B (Interagency Advisory Committee on Water Data, 1982) recommends the use of the Grubbs-Beck test (Grubbs and Beck, 1972) to detect low outliers in flood frequency analysis. As described by Cohn and others (2013), the MGB test is a generalization of the Grubbs-Beck method that allows for a standard procedure for identifying multiple potentially influential low floods. In flood frequency analysis, potentially influential low floods are annual peaks that meet three criteria: their magnitude is much smaller than the flood quantile of interest, they occur below a statistically significant break in the flood frequency plot, and they have excessive influence on the estimated frequency of large floods. When an observation is identified as a potentially influential low flood, all values smaller than that flood are also categorized as potentially influential low floods. Identifying potentially influential low floods and recording them as censored peaks can greatly improve estimator robustness with little or no loss of efficiency. Thus, the use of the MGB test can improve the fit of the small AEPs (larger magnitude floods) by minimizing lack-of-fit due to unimportant potentially influential low floods in an annual peak series (Cohn and others, 2013; Veilleux and others, 2014).

Data Analysis

Prior to performing a B–WLS/B–GLS regional skew analysis, three data analysis procedures must first be completed. This section describes the steps for the calculations for pseudorecord length for each streamgage given the number of censored observations and concurrent record lengths, the corrections for structural bias in the at-site estimate of skew and its MSE, and the development of a cross-correlation model of concurrent annual peak discharges between streamgages.

Pseudorecord Length

The record length of the annual peak series at each streamgage is used in the regional skew study in several steps, including unbiasing the at-site skew and its MSE, determining the concurrent record length between two streamgages, and computing the cross correlation of the at-site skews (see “Cross-Correlation Models” section). Because the dataset includes censored data and historic information, the effective record length used to compute the precision of the skewness estimators is no longer simply the number of annual peak discharges at a streamgage. Instead, a more complex calculation is used to take into account the availability of historic information and censored values. Although historic information and censored peaks provide valuable information, they often provide less information than an equal number of years with systematically recorded peaks (Stedinger and Cohn, 1986). The historic peak itself may be known imprecisely, and a number of years of record may simply be known to have peak values less than the historic peak. The following calculations provide a pseudorecord length (P_{RL}) that appropriately accounts for all peak discharge data types available for a streamgage. The pseudorecord length equals the systematic record length if such a complete record, with no interval data, is all that is available for a streamgage.

The first step in computing P_{RL} is to conduct an EMA analysis with all available information, including historic information and censored peaks (EMA_C). From this EMA_C, the at-site skew without regional information $\left(\hat{G}_C\right)$ and the MSE of that skewness estimator are extrapolated for each streamgage, as well as the year the historical period begins (YB_C), the year the historical period ends (YE_C), and the length of the historical period (H_C) from the following equation:

$$H_C = YE_C - YB_C + 1. \tag{3}$$

The second step is to run EMA again with only the systematic peaks (EMA_{sys}). From the EMA_{sys} analysis, the at-site skew without regional information ($\hat{G}_{sys}$) and the MSE of that skew ($MSE\left(\hat{G}_{sys}\right)$) are extracted, as well as the number of systematic peaks (P_{sys}). If no historical or censored data exist for the streamgage, then these values are the same as $\hat{G}_C$ and $MSE\left(\hat{G}_C\right)$.

The third step is to represent, from both EMA_C and EMA_{sys}, the precision of the skew as two record lengths (RL_C and RL_S, respectively) based upon the estimated skew and MSE. The corresponding record lengths are calculated using the following equation from Griffis and others (2004) and Griffis and Stedinger (2009b):

$$MSE\left(\hat{G}\right) = \left[\frac{6}{RL} + a\left(RL\right)\right] \times \left[1 + \left(\frac{9}{6} + b\left(RL\right)\right)\hat{G}^2 + \left(\frac{15}{48} + c\left(RL\right)\right)\hat{G}^4\right] \tag{4}$$

where

$MSE\left(\hat{G}_C\right)$	uses RL_C and $\hat{G}_C$,
$MSE\left(\hat{G}_{sys}\right)$	uses RL_S and $\hat{G}_{sys}$, and
$a(RL)$, $b(RL)$, and $c(RL)$	are calculated as follows:

$$a\left(RL\right) = -\frac{17.75}{RL^2} + \frac{50.06}{RL^6}, \tag{5}$$

$$b\left(RL\right) = \frac{3.93}{RL^{0.3}} - \frac{30.97}{RL^{0.6}} + \frac{37.1}{RL^{0.9}}, \text{ and} \tag{6}$$

$$c(RL) = -\frac{6.16}{RL^{0.56}} + \frac{36.83}{RL^{1.12}} - \frac{66.9}{RL^{1.68}} \ . \tag{7}$$

Next, the difference (RL_{diff}) between RL_C and RL_S is used as a measure of the extra information provided by the historic and censored information that was included in the EMA_C analysis but not in the EMA_{sys} analysis:

$$RL_{diff} = RL_C - RL_{sys} \ . \tag{8}$$

The P_{RL} for the entire record at the streamgage is calculated using RL_{diff} and the number of systematic peaks (P_{sys}) as follows:

$$P_{RL} = RL_{diff} + P_{sys} \ . \tag{9}$$

P_{RL} must be nonnegative. If no historic or censored data exist for the streamgage, P_{RL} equals P_{sys}. If P_{RL} is greater than H_C, then P_{RL} is set equal to H_C. Also, if P_{RL} is less than P_{sys}, then P_{RL} is set equal to P_{sys}. This ensures that the P_{RL} will not be greater than H_C or less than P_{sys}.

The at-site skew is sensitive to extreme events and more accurate estimates can be obtained from longer records (Interagency Advisory Committee on Water Data, 1982). Thus, after ensuring adequate spatial and hydrologic coverage, streamgages with P_{RL} less than 30 years were removed from the study. Of the 169 sites remaining after removing the 17 redundant sites, 16 were removed because their P_{RL} is less than 30 years (table 1, in back of report; indicated as “no-P”), leaving 153 streamgages from which a regional skewness model for New England was developed.

Unbiasing the At-Site Estimators

The at-site skew estimates become unbiased by using the correction factor developed by Tasker and Stedinger (1986) and employed in Reis and others (2005). The unbiased at-site skew for streamgage i ($\hat{\gamma}_i$) is computed from P_{RL} and the at-site skew ($G_{c,i}$) computed by EMA for streamgage i using the following equation:

$$\hat{\gamma}_i = \left[1 + \frac{6}{P_{RL,i}}\right] G_i, \tag{10}$$

where

$P_{RL,i}$ is the pseudorecord length for streamgage i (eq. 9), and

G_i is the biased at-site skew estimate for streamgage i; for the purposes of this equation, G_i equals $G_{c,i}$.

The variance of the unbiased at-site skew ($Var(\hat{\gamma}_i)$) includes the correction factor developed by Tasker and Stedinger (1986):

$$Var(\hat{\gamma}_i) = \left[1 + \frac{6}{P_{RL,i}}\right]^2 Var(G_i), \tag{11}$$

where $Var(G_i)$ is calculated using the following equation from Griffis and Stedinger (2009b):

$$Var(\hat{G}) = \left[\frac{6}{P_{RL}} + a(P_{RL})\right] \times \left[1 + \left(\frac{9}{6} + b(P_{RL})\right)\hat{G}^2 + \left(\frac{15}{48} + c(P_{RL})\right)\hat{G}^4\right], \tag{12}$$

where

$$a(P_{RL}) = -\frac{17.75}{P_{RL}^{\ 2}} + \frac{50.06}{P_{RL}^{\ 3}} \ , \tag{13}$$

$$b\left(P_{RL}\right)=\frac{3.92}{P_{RL}{}^{0.3}}-\frac{31.10}{P_{RL}{}^{0.6}}+\frac{34.86}{P_{RL}{}^{0.9}}\text{, and} \tag{14}$$

$$c\left(P_{RL}\right)=-\frac{7.31}{P_{RL}{}^{0.59}}+\frac{45.90}{P_{RL}{}^{1.18}}-\frac{86.50}{P_{RL}{}^{1.77}}\,. \tag{15}$$

Estimating the Mean Square Error of the Skew

There are several ways to estimate the MSE of the at-site skew (MSE_G). The approach used by EMA (from Cohn and others, 2001, eq. 55) generates a first-order estimate of the MSE_G, which should perform well when interval data are present. Another option is to use the approach described by Griffis and Stedinger (2009b) (the variance as shown in equation 12 is equated to the MSE), employing either the systematic record length or the length of the whole historical period. However, this method does not account for censored data and thus can lead to inaccurate and underestimated MSE_G. This issue has been addressed by using the P_{RL} instead of the length of the historical period because the P_{RL} reflects the effect of the censored data on the number of systematic peaks. Thus, the unbiased Griffis and Stedinger (2009b) MSE_G is used in the regional skew model because it is more stable than the approach by Cohn and others (2001) and relatively independent of the at-site skew. This methodology has been used in recently published regional skew studies (Parrett and others, 2011; Eash and others, 2013; Paretti and others, 2014; Southard and Veilleux, 2014).

Cross-Correlation Models

A critical step for a GLS analysis is estimation of the cross correlation among at-site skews. Martins and Stedinger (2002) used Monte Carlo simulations to derive a relation between the cross correlation of the skew at two streamgages (i and j) as a function of the cross correlation of concurrent annual peak flows (ρ_{ij}):

$$\hat{\rho}\left(\hat{\gamma}_i,\hat{\gamma}_j\right)=Sign\left(\hat{\rho}_{ij}\right)cf_{ij}\left|\hat{\rho}_{ij}\right|^{\kappa} \tag{16}$$

where

$\hat{\gamma}_i$ and $\hat{\gamma}_j$ are the unbiased at-site skew estimate for streamgages i and j, respectively;

$\hat{\rho}_{ij}$ is the estimated cross correlation of concurrent annual peak flow for streamgages i and j;

κ is a constant between 2.8 and 3.3; and

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

$$cf_{ij}=\frac{CY_{ij}}{\sqrt{\left(P_{RL,i}\right)\left(P_{RL,i}\right)}}\,. \tag{17}$$

where

CY_{ij} is the pseudorecord length of the period of concurrent record (pseudoconcurrent record length); and

$P_{RL,i}$ and $P_{RL,j}$ are the pseudorecord length corresponding to streamgages i and j, respectively, as calculated from equation 9.

Pseudoconcurrent Record Length

The CY_{ij} (eq. 17) is used to compute cross correlation between streamgages. Because both censored and historic data are used, the effective concurrent record length calculation is more complex than determining in which years the two streamgages both have recorded systematic peaks.

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

$$CY_{ij}=\left(YE_{ij}-YB_{ij}+1\right)\left(\frac{P_{RL,i}}{H_{C,i}}\right)\left(\frac{P_{RL,j}}{H_{C,j}}\right), \tag{18}$$

where

YB_{ij} is the beginning year of the common historical record for streamgages i and j;

YE_{ij} is the ending year of the common historical record for streamgages i and j;

$P_{RL,i}$ and $P_{RL,j}$ are the pseudorecord length for streamgages i and j, respectively; and

$H_{C,i}$ and $H_{C,j}$ are the length of the historical record for streamgages i and j, respectively.

The computed pseudoconcurrent record length depends upon the years of historical record in common between the two streamgages, as well as the ratios of the pseudorecord length to the historical record length for each streamgage.

New England Cross-Correlation Model

A cross-correlation model for the annual peaks in New England was developed using logarithms (base 10) of the annual peak flows from 34 streamgages that generated streamgage pairs with at least 80 years of concurrent systematic peaks, which resulted in 546 streamgage pairs. A logit model, termed the Fisher Z transformation (Z), provided a convenient transformation of the sample correlations r_{ij} from the (–1, +1) range to the (–∞, +∞) range. It was developed as follows:

$$Z=\log\left(\frac{1+r}{1-r}\right), \tag{19}$$

where r is a sample correlation.

Various models relating the $\hat{\rho}_{ij}$ to various basin characteristics were considered. The model that was adopted uses only one explanatory variable for estimating the $\hat{\rho}_{ij}$ and is based on the distance, in miles, between basin centroids (D_{ij}), calculated as follows:

$$\hat{\rho}_{ij} = \frac{exp\left(2Z_{ij}\right)-1}{exp\left(2Z_{ij}\right)+1}, \tag{20}$$

where

$$Z_{ij} = \exp\left(0.83 - 0.24\left(\frac{D_{ij}^{0.14}-1}{0.14}\right)\right). \tag{21}$$

An OLS regression analysis based on 546 streamgage pairs with at least 80 years of concurrent record indicated that the cross-correlation model is as accurate as having 270 years of concurrent annual peaks from which to calculate the cross correlation. The fitted OLS regression relation between Z_{ij} and D_{ij} from the 546 streamgage pairs (fig. 2*A*) shows an exponential decline in the cross correlation for streamgages within 50 miles of each other. A similar decline is found in the cross correlation and distance between basin centroids for the untransformed streamgage pairs (fig. 2*B*). The cross-correlation model was used to estimate streamgage-to-streamgage cross correlations for concurrent annual peak discharges at all streamgage pairs used in the regional skew study.

Regression Analyses

The B–WLS/B–GLS method (as it appears in Veilleux and others, 2012) is described briefly, followed by the results of applying these methods in the New England regional skew study; details on the method can be found in Veilleux (2011) and Veilleux and others (2011). The method uses an OLS analysis to fit an initial regional skew model that is then used to generate a stable regional skew estimate for each streamgage. That stable regional estimate is the basis for computing the variance of each streamgage skew employed in the WLS analysis. Then, B–WLS is used to generate estimators of the regional skew model parameters (discussed in the "Weighted Least Squares Analysis" section). Finally, B–GLS is used to estimate the precision of the WLS parameter values, to estimate the model error variance and its precision, and to compute various diagnostic statistics.

Ordinary Least Squares Analysis

The first step in the B–WLS/B–GLS regional skew analysis is the estimation of a regional skew model using OLS. The OLS regional regression yields the parameters and a model that can be used to generate unbiased and relatively stable regional estimates of the skew for all streamgages, as follows:

$$\tilde{y}_{OLS} = X\hat{\beta}_{OLS}, \tag{22}$$

where

$\tilde{y}_{OLS}$	is the estimated regional skew value;
X	is an $[n\times k]$ matrix of basin characteristics for the streamgage;
n	is the number of streamgages;
k	is the number of basin parameters, including a column of ones to estimate the constant; and
$\hat{\beta}_{OLS}$	is the $(k\times 1)$ vector of OLS regression parameters.

These $\tilde{y}_{OLS}$ values are then used to calculate unbiased regional skew variances for all streamgages in New England using the equations reported in Griffis and Stedinger (2009b). These regional skew variances are based on the regional OLS estimator of the skew instead of the at-site skew estimator, thus making the weights in the subsequent steps relatively independent of the at-site skew estimates.

Weighted Least Squares Analysis

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

The first step in the WLS analysis is to estimate the model error variance ($\sigma^2_{\delta,B-WLS}$) using an iterative approach, as described by Reis and others (2005). Using a B–WLS approach to estimate the model error variance avoids the possible pitfall of estimating the model error variance as zero, which can occur when using method-of-moments WLS. It is important to note that the Bayesian analysis produces an estimate of the distribution of the model error variance; however, only the mean $\sigma^2_{\delta,B-WLS}$ estimator is considered in this analysis.

Given the $\sigma^2_{\delta,B-WLS}$ estimator, a diagonal covariance matrix [$\Lambda_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)$] is created as follows:

$$\Lambda_{WLS}\left(\sigma^2_{\delta,B-WLS}\right) = \sigma^2_{\delta,B-WLS} I + diag\left(Var\left[\hat{\gamma}\right]\right), \tag{23}$$

where

I	is an $n\times n$ identity matrix,
$diag\left(Var\left[\hat{\gamma}\right]\right)$	is the $n\times n$ matrix containing the variance of the unbiased at-site skew ($Var\left[\hat{\gamma}_i\right]$) on the diagonal and zeros on the off-diagonal directions, and
$\sigma^2_{\delta,B-WLS}$	is the model error variance for the B–WLS regression.

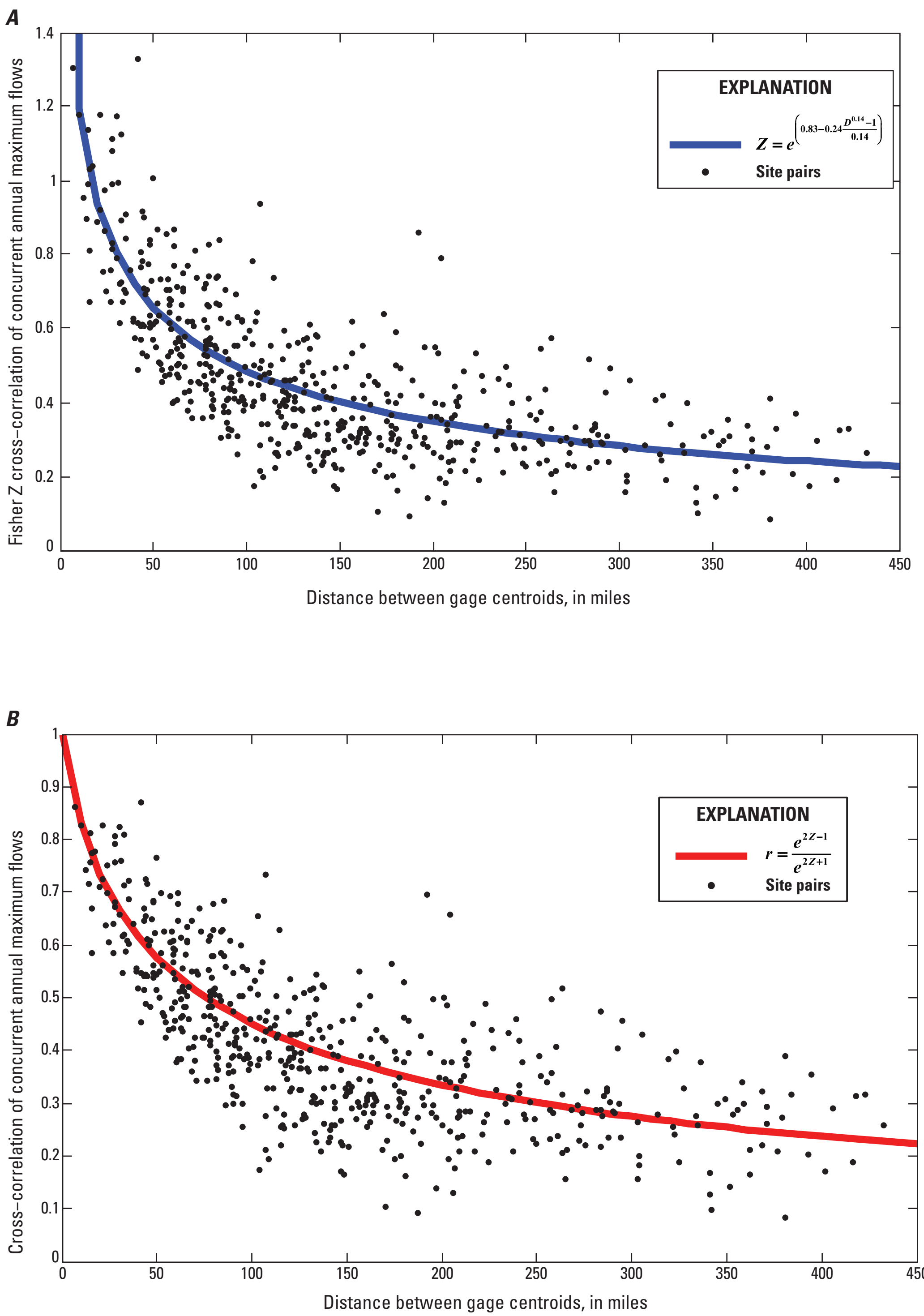


Figure 2. Cross correlation of annual peak flow in the New England skew study. *A*, Relation between Fisher Z-transformed cross-correlation of logarithms of annual peak discharge and distance between basin centroids based on 546 streamgage pairs with concurrent record lengths greater than or equal to 80 years from 34 streamgages in New England. *B*, Relation between untransformed cross-correlation of logs of annual peak discharge and distance between basin centroids based on 546 streamgage pairs with concurrent record lengths greater than or equal to 80 years from 34 streamgages in New England.

The diagonal elements of the covariance matrix are the sum of the estimated $\sigma^2_{\delta,B-WLS}$ and the $Var\left[\hat{\gamma}_i\right]$, which depends upon on the at-site record length and the estimate of the previously calculated $\tilde{y}_{OLS}$ regional skew. The off-diagonal elements of $\Lambda_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)$ are zero because cross correlations between streamgages are not considered in the WLS analysis. Using that covariance matrix, a B–WLS analysis is used to generate the $k\times n$ weight matrix (W) of the WLS weights as follows:

$$W = \left[X^T \Lambda_{WLS}\left(\sigma^2_{\delta,B\text{-}WLS}\right)^{-1} X\right]^{-1} X^T \Lambda_{WLS}\left(\sigma^2_{\delta,B-WLS}\right)^{-1}. \quad (24)$$

These weights (W) are used to compute the final estimates of the $k\times 1$-matrix-vector regression parameters $\hat{\beta}_{WLS}$ as follows:

$$\hat{\beta}_{WLS} = W\hat{\gamma}\cdot \quad (25)$$

Generalized Least Squares Analysis

After the regression model $\hat{\beta}_{WLS}$ parameters and W weights are determined with a WLS analysis, the precision of the fitted model and of the regression skews are estimated using a B–GLS regression framework for regional skew developed by Reis and others (2005) with the posterior probability density function for $\sigma^2_{\delta,B-GLS}$ being as follows:

$$f\left(\sigma^2_{\delta,B-GLS} \mid \hat{\gamma}, \hat{\beta}_{WLS}\right) \propto \frac{\xi\left(\sigma^2_{\delta,B-GLS}\right)}{\sqrt{\left|\Lambda_{GLS}\left(\sigma^2_{\delta,B-GLS}\right)\right|}} \times \exp\left[\frac{-0.5\left(\hat{\gamma} - X\hat{\beta}_{WLS}\right)^T}{\Lambda_{GLS}\left(\sigma^2_{\delta,B-GLS}\right)}\left(\hat{\gamma} - X\hat{\beta}_{WLS}\right)\right], \quad (26)$$

where

$\hat{\gamma}$ represents the skew data, and

$\xi\left(\sigma^2_{\delta,B-GLS}\right)$ is the exponential prior for the model error variance described by the following equation:

$$\xi\left(\sigma^2_{\delta,B-GLS}\right) = \lambda e^{-\lambda\left(\sigma^2_{\delta,B-GLS}\right)} \quad (27)$$

where $\sigma^2_{\delta,B-GLS} > 0$. A value of 10 was adopted for lambda (λ), corresponding to a mean $\sigma^2_{\delta,B-WLS}$ of 0.1. That exponential prior assigns a 63 percent probability to the interval [0, 0.1], 86 percent probability to the interval [0, 0.2], and 95 percent probability to the interval [0, 0.3].

The mean B–GLS model error variance ($\sigma^2_{\delta,B-GLS}$) can then be used to compute the precision of the regression parameters ($\hat{\beta}_{WLS}$) that were calculated with the WLS weights W. The GLS covariance matrix for the WLS estimator $\hat{\beta}_{WLS}$ is as follows:

$$\sum \hat{\beta}_{WLS} = W\Lambda_{GLS}\left(\sigma^2_{\delta,B-GLS}\right)W^T \quad (28)$$

where $\Lambda_{GLS}\left(\sigma^2_{\delta,B-GLS}\right)$ is an $n\times n$ GLS covariance matrix calculated as follows:

$$\Lambda_{GLS}\left(\sigma^2_{\delta,B-GLS}\right) = \sigma^2_{\delta,B-GLS} I + \sum \hat{\gamma} \quad (29)$$

where

I is an $n\times n$ identity matrix,

$\sum \hat{\gamma}$ is a full $n\times n$ matrix that contains the sampling variances of the $Var[\hat{\gamma}]$ and the covariances of the skew $\hat{\gamma}$.

The off-diagonal values of $\sum \hat{\gamma}$ are determined by the cross correlation of concurrent systematic annual peak flows and the cf_{ij} factor (eq. 17; Martins and Stedinger, 2002, eq. 3). When calculating the cf_{ij} factor using the ratio between the number of concurrent peak flows at streamgage pairs and the total number of peak flows at both streamgages, only the systematic records and historic peaks are considered. Thus, any additional information provided by perception thresholds and censored peaks included in the EMA analysis would have been neglected in the calculation of the cross correlation of peak flows and the cf_{ij} factor. Precision metrics include the standard error of the regression parameters [$SE\left(\hat{\beta}_{WLS}\right)$], the model error variance ($\sigma^2_{\delta,B-GLS}$), the pseudocoefficient of determination (pseudo-R^2_δ), as well as the average variance of prediction at a streamgage not used in the regional model (AVP_{new}). The pseudo-R^2_δ, describes the estimated fraction of the variability in the true skew from streamgage to streamgage explained by each model (Gruber and others, 2007; Parrett and others, 2011).

New England Regional Skew Study Results

All available basin characteristics compiled for the New England regional skew study using the B–WLS/B–GLS regression methodology were initially considered as possible explanatory variables and evaluated but none produced a pseudo-R^2_δ greater than 10 percent, indicating that the basin characteristics did not explain the variation in at-site skews within New England. Hence, a constant model, one that does not vary with basin characteristics, was developed to predict the regional skew for New England from the sample mean of 153 streamgages with at least 30 years of P_{RL} (table 2); this model produced a skew of 0.37, the only statistically significant skew for the region.

The constant model does not explain any variability in the skew, so the pseudo-R^2_δ equals 0. However, the addition of any of the available basin characteristics or combination thereof did not produce a pseudo-R^2_δ greater than 13 percent. This indicates that the inclusion of basin characteristics as explanatory variables in the regression help explain only 13 percent of the total variability in the true skew for New England. Thus, the addition of basin characteristics is not warranted because the increased model complexity does not result

Table 2. Regional skew model and model fit for New England regional skew constant model.

[Pseudocoefficient of variation (pseudo- R_δ^2) describes the percentage of the variability in the true skews explained by each model (Gruber and others, 2007). σ_δ^2 , model error variance; *ASEV*, average sampling error variance; AVP_{new}, average variance of prediction for a new streamgage; XX, not applicable]

Model	Regression constant	σ_δ^2	*ASEV*	AVP_{new}	Pseudo- R_δ^2 (percent)
Skewness coefficient	0.37	0.13	0.013	0.14	0
Standard deviation	0.11	0.03	XX	XX	XX

in a gain in model precision. The constant model provides the best estimate of regional skew for New England.

The posterior mean of the constant model error variance (σ_δ^2) is 0.13 (table 2). The average sampling error variance of the constant model was 0.013, which represents the average error in the regional skew as measured in the at-site skew at streamgages used in the analysis. The average variance of prediction at a new streamgage (AVP_{new}) corresponds to the *MSE* (Interagency Advisory Committee on Water Data, 1982) used to describe the precision of the generalized skew. The constant model has an AVP_{new} equal to 0.14, which corresponds to an effective record length of 56 years. An AVP_{new} of 0.14 is a marked improvement compared with the average variance of prediction shown in the national skew map of B17B, where the reported MSE of 0.302 has a corresponding effective record length of only 17 years. Thus, the new regional model has more than three times the information content (as measured by effective record length) of that calculated for the map of B17B.

Bayesian Regression Diagnostics

To determine if a model is a good representation of the data and which regression parameters, if any, should be included in a regression model, diagnostic statistics have been developed to evaluate how well a model fits a regional hydrologic dataset (Griffis, 2006; Gruber and Stedinger, 2008). In a regional skew study, possible explanatory variables are evaluated to provide a region with the most accurate skew prediction while keeping the model as simple as possible. This section details the diagnostic statistics for a B–WLS/B–GLS analysis for the New England regional skew study.

A pseudoanalysis of variance (pseudo-ANOVA) table contains regression diagnostics and goodness of fit statistics (table 3). In particular, these statistics describe how much of the variation in the observations can be attributed to the regional model and how much of the residual variation can be attributed to model error and sampling error. Difficulties arise in determining these quantities. The model errors cannot be resolved because the values of the sampling error (η_i) for each streamgage *i* are not known. However, the sum of the squares of the sampling error η_i can be described by the mean value of the sampling errors, $\sum_{i=1}^{n} Var[\hat{\gamma}_i]$ where there are *n* equations

Table 3. Pseudoanalysis of variance for the New England regional skew constant model.

[*n*, number of streamgages; *k*, number of basin parameters]

Source	Degrees of freedom: Formula	Degrees of freedom: Value	Equations	Sum of squares
Model	*k*	0	$n\left[\sigma_\delta^2(0)-\sigma_\delta^2(k)\right]$	0
Model error	*n*–*k*–1	107	$n\left[\sigma_\delta^2(k)\right]$	20
Sampling error	*n*	108	$\sum_{i=1}^{n} Var(\hat{\gamma}_i)$	21
Total	2*n*–1	215	$n\left[\sigma_\delta^2(k)\right]+\sum_{i=1}^{n} Var(\hat{\gamma}_i)$	41
$EVR=\dfrac{\sum_{i=1}^{n} Var(\hat{\gamma}_i)}{n\left[\sigma_\delta^2(k)\right]}$				1.0
$MBV^*=\dfrac{Var\left[b_0^{WLS}\mid GLS\ analysis\right]}{Var\left[b_0^{WLS}\mid WLS\ analysis\right]}=\dfrac{w^T\Lambda w}{w^T v}$			where $w_i=\dfrac{1}{\Lambda_{ii}}$ and $v=\lvert n\times 1\rvert$	7.8
$R_\delta^2=1-\dfrac{\sigma_\delta^2(k)}{\sigma_\delta^2(0)}$				0 percent

(equal to the number of streamgages in the study) and the total variation due to the model error (δ) for a model with k parameters has a mean equal to $n\sigma_\delta^2(k)$.

For a model with no explanatory parameters (γ_i) other than the mean (as was found by the constant model in this study), the estimated model error variance ($\sigma_\delta^2(0)$) describes all the variation in $\gamma_i = \mu + \delta_i$, where μ is the mean of the estimated at-site skews and δ_i is the model error for each streamgage i. Thus, the total expected variation in the sum of squares due to δ_i and η_i is as follows:

$$\eta_i = \hat{\gamma}_i - \gamma_i = \sigma_\delta^2(0) + \sum_{i=1}^{n} Var(\hat{\gamma}_i). \tag{30}$$

For a nonconstant regional model, the expected sum of squares for n streamgages attributed with k parameters equals $n\left[\sigma_\delta^2(0) - \sigma_\delta^2(k)\right]$ because the sum of the model error variance $n\sigma_\delta^2(k)$ and the variance explained by the model must equal $n\sigma_\delta^2(0)$. This division of the variation in the observations is referred to as a pseudo-ANOVA because the contributions of the three sources of error are estimated or constructed rather than being determined from the computed residual errors and the observed model predictions, while also ignoring the effect of correlation among the sampling errors.

The error variance ratio (*EVR*) is a diagnostic value used to evaluate if a simple OLS regression is sufficient or if a more sophisticated WLS or GLS analysis is appropriate. The *EVR* is the ratio of the average sampling error variance to the model error variance. An *EVR* greater than 0.20 generally indicates that the sampling variance is not negligible when compared to the model error variance, suggesting the need for a WLS or GLS regression analysis. The *EVR* is calculated as follows:

$$EVR = \frac{SS(\text{sampling error})}{SS(\text{model error})} = \frac{\sum_{i=1}^{n} Var(\hat{\gamma}_i)}{n\sigma_\delta^2(k)}, \tag{31}$$

where *SS* is the sum of squares.

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

$$MBV^* = \frac{Var\left[b_0^{WLS} \mid GLS\ analysis\right]}{Var\left[b_0^{WLS} \mid WLS\ analysis\right]} = \frac{w^T \Lambda w}{\sum_{i=1}^{n} w_i}, \tag{32}$$

where w is the vector of weights for the WLS analysis.

The pseudo-ANOVA results for the constant model ($k = 0$), along with other model diagnostics, determined for the New England regional skew model are provided in table 3. The constant model does not have any explanatory variables, thus the variation attributed to the model is 0. The constant model has a sampling error variance greater than its model error variance (table 3); the *EVR* for the constant model has a value of 1.05. The sampling variability in the sample skew was greater than the error in the regional model. Thus, an OLS model that neglects sampling error in the at-site skew may not provide a statistically reliable analysis of the data. Given the variation of record lengths among streamgages, a WLS or GLS analysis is needed to evaluate the final precision of the model rather than a simple OLS analysis.

The *MBV** for the constant model is equal to 7.8. This is a large value, indicating the cross correlation among the skew estimators has had an effect on the precision with which the regional average skew can be estimated; if a WLS precision analysis were used for the estimated constant parameter in the constant model, then the variance would be underestimated by a factor of 7.8. Thus, a WLS analysis would seriously misrepresent the variance of the constant in the New England regional skew model. Moreover, a WLS model would result in underestimation of the variance of prediction given that the sampling error in both the WLS and GLS models was sufficiently large enough to make an appreciable contribution to the average variance of prediction.

Leverage and Influence

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

Influential observations often have high leverage. If p is the number of estimated regression skews ($p = 1$ for a constant model) and n is the sample size (or number of streamgages in the study), then leverage values have a mean of p/n and values greater than $2p/n$ are generally considered large. Influence values greater than $4/n$ are typically considered large (Tasker and Stedinger, 1989; Veilleux, 2011; Veilleux and others, 2011). Veilleux and others (2011) and Veilleux (2011) provide

Table 4. Streamgages with high influence in the New England regional skew constant model.

[High influence is defined as Cook's D values greater than 4/*n* (or 4/153 = 0.026). Each of the streamgages was assigned a rank from 1 to 153 for each category (pseudorecord length [P_{RL}], unbiased at-site skew, unbiased mean squared error [MSE] of at-site skew, residual) where a rank of 1 corresponds to the highest positive value in each category. ME, Maine; NH, New Hampshire; MA, Massachusetts; CT, Connecticut]

Index number	USGS streamgage number	State	Cook's D	Leverage	Pseudo ERL (years)		Unbiased (at-site skew)		Unbiased MSE (at-site skew)		Residual	
					Value	Rank	Value	Rank	Value	Rank	Value	Rank
19	1055000	ME	0.137	0.0074	82	29	–1.5	153	0.28	14	–1.88	153
166	1150900	NH	0.120	0.0075	84	22	2.1	2	0.41	7	1.71	2
115	1176000	MA	0.087	0.0079	99	11	1.7	5	0.29	13	1.31	5
33	1124000	CT	0.053	0.0053	42	133	2.3	1	1.03	1	1.98	1
110	1173500	MA	0.042	0.0055	45	124	2.0	3	0.58	3	1.65	3
62	1209700	CT	0.033	0.0058	49	113	1.7	4	0.42	6	1.35	4
99	1162000	MA	0.031	0.0078	94	14	1.2	10	0.17	43	0.81	10
163	1142500	NH	0.029	0.0092	163	3	1.0	19	0.09	132	0.63	19
120	1185500	MA	0.027	0.0079	98	12	1.1	12	0.15	56	0.74	12

a detailed description of the equations used to determine influence and leverage for a B–WLS/B–GLS analysis.

For the New England regional skew constant model, high influence is defined as influence greater than 0.026 (or $4p/n$ = 4/153) and high leverage is defined as leverage greater than 0.013 (or $2p/n$=(2×1)/153). No sites in the study had high leverage, and thus the differences in the leverage values for the constant model reflect the variation in record lengths among sites. Nine sites in the study have high influence (table 4) and thus have an unusually large effect on the fitted regression relation. The five sites with the greatest unbiased at-site skew are part of the nine sites with high influence. The most influential streamgage (USGS 01055000, Swift River near Roxbury, Maine), as measured by Cook's distance (Cook, 1977) and the unbiased skew and residual ranks, also has the smallest unbiased at-site skew (that is, the most negative skew) of all streamgages used in the study. Each of the 153 streamgages used in the regional skew study was assigned a rank from 1 to 153, where a rank of 1 corresponds to the highest positive value in each category.

Summary

The flood frequency guidelines for the United States recommend fitting a log-Pearson type III distribution to a series of annual flood maximums. This distribution, in the specific case of flood frequency analysis, is described by three moments: the mean, the standard deviation, and the skewness coefficient (or skew) of the logarithms of the flow. The skew is very sensitive to extreme events, such as large floods, because they cause a sample to be highly skewed, or asymmetrical.

The magnitude of annual exceedance probability floods is greatly affected by the skew of the annual peak flows at a streamgage. Thus, the recommended method for estimating streamgage skew is to weight the station skew with a regional skew to better reflect regional and long-term hydrology.

In order to determine the at-site skew for each streamgage in the regional skew analysis, an expected moments algorithm with multiple Grubbs-Beck censoring of potentially influential low floods was used to compute moments of the logarithms for the log-Pearson type III distribution. The Bayesian weighted least squares/Bayesian generalized least squares regression method was used to develop a regional skew model for New England.

The study area for the New England regional skew analysis consists of the entire States of Connecticut, Maine, New Hampshire, Rhode Island, Vermont, most of Massachusetts (excluding Cape Cod and the islands), and the easternmost part of New York. The region spans about 500 miles from southeastern Connecticut to northern Maine. A total of 186 streamgages were evaluated for use in the analysis. Streamgage basin characteristics were obtained from the Geospatial Attributes of Gages for Evaluating Streamflow, version II (GAGES II) database. Various basin characteristics were considered as possible explanatory variables in the Bayesian weighted least squares/Bayesian generalized least squares regression analysis for regional skew. All available basin characteristics were evaluated, but none produced a pseudocoefficient of determination greater than 13 percent, indicating that basin characteristics did not help in explaining the variation in at-site skews in New England. Based on the results of these analyses, a regional skew constant model was chosen. The constant model yielded a regional skew of 0.37, a model error variance of 0.13, and an average variance of prediction at a new site of 0.14. An average variance of prediction of 0.14 is a marked improvement from the skew from the nationally recommended flood frequency guidelines, whose reported mean squared error is 0.302.

References Cited

Beard, L.R., 1974, Flood flow frequency techniques—A report: University of Texas at Austin, [variously paged].

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

Cember, R.P., and Wilks, D.S., 1993, Climatological atlas of snowfall and snow depth for the northeastern United States and southeastern Canada: Northeast Regional Climate Center Research Publication RR 93–1, 216 p.

Cohn, T.A., England, J.F., Berenbrock, C.E., Mason, R.R., Stedinger, J.R., and Lamontagne, J.R., 2013, A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series: Water Resources Research, v. 49, no. 8, p. 5047–5058. [Also available at https://doi.org/10.1002/wrcr.20392.

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

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

Cook, R.D., 1977, Detection of influential observations in linear regression: Technometrics, v. 19, no. 1, p. 15–18. [Also available at https://doi.org/10.2307/1268249.]

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

Denny, C.S., 1982, Geomorphology of New England: U.S. Geological Survey Professional Paper 1208, 18 p. [Also available at https://doi.org/10.3133/pp1208.]

Eash, D.A., Barnes, K.K., and Veilleux, A.G., 2013, Methods for estimating annual exceedance-probability discharges for streams in Iowa, based on data through water year 2010: U.S. Geological Survey Scientific Investigations Report 2013–5086, 63 p. [Also available at https://doi.org/10.3133/sir20135086.]

Falcone, J.A., 2011, GAGES–II—Geospatial attributes of gages for evaluating streamflow: U.S. Geological Survey dataset, accessed October 23, 2016, at https://doi.org/10.3133/70046617.

Feaster, T.D., Gotvald, A.J., and Weaver, J.C., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 3, South Carolina: U.S. Geological Survey Scientific Investigations Report 2009–5156, 226 p. [Also available at https://doi.org/10.3133/sir20095156.]

Gotvald, A.J., Barth, N.A., Veilleux, A.G., and Parrett, C., 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., data, accessed September 12, 2016, at https://doi.org/10.3133/sir20125113.

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. [Also available at https://doi.org/10.3133/sir20095043.]

Griffis, V.W., 2006, Flood frequency analysis—Bulletin 17, regional information, and climate change: Cornell University Ph.D. dissertation, 482 p.

Griffis, V.W., and Stedinger, J.R., 2007a, Log-Pearson type 3 distribution and its application in flood frequency analysis. II—Parameter estimation methods: Journal of Hydrologic Engineering, v. 12, no. 5, p. 492–500. [Also available at https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(492).]

Griffis, V.W., and Stedinger, J.R., 2007b, The use of GLS regression in regional hydrologic analyses: Journal of Hydrology, v. 344, nos. 1–2, p. 82–95. [Also available at https://doi.org/10.1016/j.jhydrol.2007.06.023.]

Griffis, V.W., and Stedinger, J.R., 2009a, Closure to "Log-Pearson type 3 distribution and its application in flood frequency analysis. II—Parameter estimation methods": Journal of Hydrologic Engineering, v. 14, no. 2, p. 209–212. [Also available at https://doi.org/10.1061/(ASCE)1084-0699(2009)14:2(209).]

Griffis, V.W., and Stedinger, J.R., 2009b, Log-Pearson type 3 distribution and its application in flood frequency analysis. III—Sample skew and weighted skew estimators: Journal of Hydrology, v. 14, no. 2, p. 121–130. [Also available at https://doi.org/10.1061/(ASCE)1084-0699(2009)14:2(121).]

Griffis, V.W., Stedinger, J.R., and Cohn, T.A., 2004, Log Pearson type 3 quantile estimators with regional skew information and low outlier adjustments: Water Resources Research, v. 40, no. 7, article W07503, 17 p., accessed June 23, 2016, at https://doi.org/10.1029/2003WR002697.

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

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

Gruber, A.M., and Stedinger, J.R., 2008, Models of LP3 regional skew, data selection and Bayesian GLS regression, *in* Babcock, R.W., Jr., and Walton, Raymond, eds., World Environmental and Water Resources Congress: American Society of Civil Engineers, paper 596, 10 p.

Hardison, C.H., 1975, Generalized skew coefficients of annual floods in the United States and their application: Water Resources Research, v. 11, no. 6, p. 851–854. [Also available at https://doi.org/10.1029/WR010i004p00745.]

Hoaglin, D.C., and Welsch, R.E., 1978, The hat matrix in regression and ANOVA: The American Statistician, v. 32, no. 1, p. 17–22. [Also available at https://doi.org/10.1080/00031305.1978.10479237.]

Interagency Advisory Committee on Water Data, 1982, Guidelines for determining flood flow frequency: U.S. Geological Survey Bulletin 17B, 28 p. [Also available at https://water.usgs.gov/osw/bulletin17b/dl_flow.pdf.]

Lamontagne, J.R., Stedinger, J.R., Berenbrock, C., Veilleux, A.G., Ferris, J.C., and Knifong, D.L., 2012, Development of regional skews for selected flood durations for the Central Valley region, California, based on data through water year 2008: U.S. Geological Survey Scientific Investigations Report 2012–5130, 60 p., appendix, accessed August 24, 2014, at https://doi.org/10.3133/sir20125130.

Martins, E.S., and Stedinger, J.R., 2002, Cross-correlation among estimators of shape: Water Resources Research, v. 38, no. 11, article 1252, p. 34–1—34–7, accessed July 8, 2016, at https://doi.org/10.1029/2002WR001589.

McCuen, R.H., 1979, Map skew: Journal of the Water Resources Planning and Management Division, v. 105, no. WR2, p. 265–277.

McCuen, R.H., 2001, Generalized flood skew—Map versus watershed skew: Journal of Hydrologic Engineering, v. 6, no. 4, p. 293–299. [Also available at http://dx.doi.org/10.1061/(ASCE)1084-0699(2001)6:4(293).]

National Oceanic and Atmospheric Administration, 2013, Regional climate trends and scenarios for the U.S. national climate assessment—Part 1. Climate of the northeast U.S.: National Oceanic and Atmospheric Administration Technical Report NESDIS 142–1, 79 p., accessed June 23, 2016, at https://scenarios.globalchange.gov/sites/default/files/NOAA_NESDIS_Tech_Report_142-1-Climate_of_the_Northeast_U.S_1.pdf.

National Oceanic and Atmospheric Administration, [undated], Historical hurricane tracks: National Oceanic and Atmospheric Administration Web mapping service, accessed July 23, 2016, at http://www.csc.noaa.gov/hurricanes/. [Moved to https://oceanservice.noaa.gov/news/historical-hurricanes/ at time of publication.]

Northeast Regional Climate Center, 2014, Northeast climate normal maps 1981–2010: Northeast Regional Climate Center Web page, accessed April 16, 2014, at http://www.nrcc.cornell.edu/page_northeast.html. [Moved to https://www.ncdc.noaa.gov/data-access/land-based-station-data/land-based-datasets/climate-normals/1981-2010-normals-data at time of publication.]

Paretti, N.V., Kennedy, J.R., Turney, L.A., and Veilleux, A.G., 2014, Methods for estimating magnitude and frequency of floods in Arizona, developed with unregulated and rural peak-flow data through water year 2010: U.S. Geological Survey Scientific Investigations Report 2014–5211, 61 p., accessed June 24, 2015, at https://doi.org/10.3133/sir20145211.

Parrett, C., Veilleux, A., Stedinger, J.R., Barth, N.A., Knifong, D.L., and Ferris, J.C., 2011, Regional skew for California, and flood frequency for selected sites in the Sacramento-San Joaquin River Basin, based on data through water year 2006: U.S. Geological Survey Scientific Investigations Report 2010–5260, 94 p., data, accessed June 24, 2014, at https://doi.org/10.3133/sir20105260.

Reis, D.S., Jr., Stedinger, J.R., and Martins, E.S., 2005, Bayesian generalized least squares regression with application to the log Pearson type III regional skew estimation: Water Resources Research, v. 41, no. 10, paper W10419, 14 p., accessed February 12, 2016, at https://doi.org/10.1029/2004WR003445.

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

Stedinger, J.R., and Tasker, G.D., 1985, Regional hydrologic analysis—1. Ordinary, weighted, and generalized least squares compared: Water Resources Research, v. 21, no. 9, p. 1421–1432. [Also available at https://doi.org/10.1029/WR021i009p01421.]

Stedinger, J.R., and Cohn, T.A., 1986, Flood frequency analysis with historical and paleoflood information: Water Resources Research, v. 22, no. 5, p. 785–793. [Also available at https://doi.org/10.1029/WR022i005p00785.]

Stedinger, J.R., and Griffis, V.W., 2008, Flood frequency analysis in the United States—Time to update: Journal of Hydrologic Engineering, v. 13, no. 4, p. 199–204. [Also available at https://doi.org/10.1061/(ASCE)1084-0699(2008)13:4(199).]

Tasker, G.D., 1978, Flood frequency analysis with a generalized skew coefficient: Water Resources Research, v. 14, no. 2, p. 373–376. [Also available at https://doi.org/10.1029/WR014i002p00373.]

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. [Also available at https://doi.org/10.1061/(ASCE)0733-9496(1986)112:2(225).]

Tasker, G.D., and Stedinger, J.R., 1989, An operational GLS model for hydrologic regression: Journal of Hydrology, v. 111, nos. 1–4, p. 361–375. [Also available at https://doi.org/10.1016/0022-1694(89)90268-0.]

U.S. Environmental Protection Agency, 2012, ORD/USEPA_Ecoregions_Level_III_and_IV (MapServer): U.S. Environmental Protection Agency map server, accessed July 23, 2014, at http://geodata.epa.gov/ArcGIS/rest/services/ORD/USEPA_Ecoregions_Level_III_and_IV/MapServer.

U.S. Environmental Protection Agency, 2013, Level III ecoregions of the continental United States: U.S. Environmental Protection Agency map, scale 1:7,500,000, accessed March 24, 2014, at https://www.epa.gov/eco-research/level-iii-and-iv-ecoregions-continental-united-states.

U.S. Geological Survey, 2014a, Peak streamflow for the nation: U.S. Geological Survey National Water Information System Web database, accessed June 13, 2014, at https://doi.org/10.5066/F7P55KJN. [Specific information directly accessible at https://nwis.waterdata.usgs.gov/nwis/peak.]

U.S. Geological Survey, 2014b, PeakFQ—Flood frequency analysis based on bulletin 17B and recommendations of the Advisory Committee on Water Information (ACWI) Subcommittee on Hydrology (SOH) Hydrologic Frequency Analysis Work Group (HFAWG): U.S. Geological Survey software page, accessed June 13, 2014, at https://water.usgs.gov/software/PeakFQ/.

Veilleux, A.G., 2009, Bayesian GLS regression for regionalization of hydrologic statistics, floods and bulletin 17 skew: Ithaca, N.Y., Cornell University M.S. thesis, 155 p. [Also available at https://hdl.handle.net/1813/13819.]

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

Veilleux, A.G., Cohn, T.A., Flynn, K.M., Mason, R.R., Jr., and Hummel, P.R., 2014, Estimating magnitude and frequency of floods using the PeakFQ 7.0 program: U.S. Geological Survey Fact Sheet 2013–3108, 2 p., accessed October 23, 2016, at https://doi.org/10.3133/fs20133108.

Veilleux, A.G., Stedinger, J.R., and Eash, D.A., 2012, Bayesian WLS/GLS regression for regional skewness analysis for regions with large crest stage gage networks, *in* Loucks, E.D., ed., World Environmental and Water Resources Congress—Crossing boundaries: American Society of Civil Engineers, p. 2253–2263.

Veilleux, A.G., Stedinger, J.R., and Lamontagne, J.R., 2011, Bayesian WLS/GLS regression for regional skewness analysis for regions with large cross-correlations among flood flows, *in* Beighley, R.E., II, and Killgore, M.W., eds., World Environmental and Water Resources Congress—Bearing knowledge for sustainability: American Society of Civil Engineers, p. 3103–3112. [Also available at https://doi.org/10.1061/41173(414)324.]

Wagner, D.M., and Veilleux, A.G., 2019, Peak-flow files and PeakFQ output files for U.S. Geological Survey streamgages in the New England region that were used in regional skew analysis: U.S. Geological Survey data release, https://doi.org/10.5066/P9MC98OM.

Weaver, J.C., Feaster, T.D., and Gotvald, A.J., 2009, Magnitude and frequency of rural floods in the southeastern United States, 2006—Volume 2, North Carolina: U.S. Geological Survey Scientific Investigations Report 2009–5158, 113 p., data, accessed June 23, 2016, at https://doi.org/10.3133/sir20095158.

Table 1. Streamgages in Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, and easternmost New York used in the New England regional skew analysis.

[USGS, U.S. Geological Survey; EMA, expected moments algorithm; MSE, mean squared error; No-R, streamgage not used in regional skew analysis because of redundancy; No-P, streamgage not used in regional skew analysis because the pseudorecord length (P_{RL}) was less than 30 years]

Regional skew index number	USGS streamgage number	State where streamgage is located	P_{RL}	EMA analysis		Used in regional skew model
				Skew, in log units	MSE of skew, in log units	
1	01011000	Maine	80	-0.09	0.07	Yes
2	01013500	Maine	87	-0.10	0.05	Yes
3	01015800	Maine	54	-0.44	0.13	Yes
4	01016500	Maine	32	0.08	0.16	Yes
5	01018000	Maine	50	-0.11	0.80	Yes
6	01022500	Maine	64	-0.01	0.08	Yes
7	01023000	Maine	60	0.18	0.08	Yes
8	01030500	Maine	109	0.13	0.06	Yes
9	01031500	Maine	117	-0.01	0.05	Yes
10	01033500	Maine	65	0.29	0.10	Yes
11	01035000	Maine	64	0.32	0.10	Yes
12	01038000	Maine	73	0.57	0.11	Yes
13	01046000	Maine	41	0.27	0.14	Yes
14	01047000	Maine	86	-0.14	0.07	Yes
15	01048000	Maine	107	-0.28	0.06	Yes
16	01049000	Maine	83	0.12	0.07	Yes
17	01054200	Maine	51	-0.51	0.14	Yes
18	01054300	Maine	30	-0.56	0.15	Yes
19	01055000	Maine	82	-1.41	0.24	Yes
20	01055500	Maine	92	0.21	0.06	Yes
21	01057000	Maine	104	-0.18	0.05	Yes
22	01058500	Maine	48	0.91	0.19	Yes
23	01059800	Maine	21	-0.74	0.21	No-P
24	01060000	Maine	55	0.06	0.10	Yes
25	01066000	Maine	222	-0.03	0.03	Yes
26	01066500	Maine	70	0.27	0.08	Yes
27	01118300	Connecticut	53	0.11	0.10	Yes
28	01119500	Connecticut	80	0.65	0.11	Yes
29	01120000	Connecticut	59	0.22	0.10	Yes
30	01120500	Connecticut	31	0.29	0.19	Yes
31	01121000	Connecticut	71	0.56	0.11	Yes
32	01123000	Connecticut	60	-0.17	0.10	Yes
33	01124000	Connecticut	42	2.05	0.79	Yes
34	01126000	Connecticut	47	0.30	0.13	Yes
35	01126500	Connecticut	50	0.26	0.12	Yes
36	01127500	Connecticut	81	0.24	0.08	Yes
37	01184100	Connecticut	52	0.14	0.11	Yes
38	01184490	Connecticut	45	0.21	0.13	Yes
39	01184500	Connecticut	56	1.08	0.21	No-R
40	01187300	Connecticut	61	0.06	0.09	Yes

Table 1. Streamgages in Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, and easternmost New York used in the New England regional skew analysis.—Continued

[USGS, U.S. Geological Survey; EMA, expected moments algorithm; MSE, mean squared error; No-R, streamgage not used in regional skew analysis because of redundancy; No-P, streamgage not used in regional skew analysis because the pseudorecord length (P_{RL}) was less than 30 years]

Regional skew index number	USGS streamgage number	State where streamgage is located	P_{RL}	EMA analysis: Skew, in log units	EMA analysis: MSE of skew, in log units	Used in regional skew model
41	01187800	Connecticut	30	0.65	0.23	Yes
42	01188000	Connecticut	80	0.02	0.07	Yes
43	01189000	Connecticut	75	0.42	0.09	Yes
44	01190500	Connecticut	46	0.21	0.13	Yes
45	01191000	Connecticut	27	1.43	0.46	No-P
46	01192500	Connecticut	85	0.08	0.07	Yes
47	01192883	Connecticut	50	-0.36	0.13	Yes
48	01193500	Connecticut	83	0.55	0.10	Yes
49	01194000	Connecticut	51	0.49	0.14	Yes
50	01194500	Connecticut	55	1.34	0.30	No-R
51	01195100	Connecticut	30	1.29	0.38	Yes
52	01196500	Connecticut	81	-0.12	0.07	Yes
53	01196620	Connecticut	35	0.33	0.17	Yes
54	01199000	Connecticut	99	0.50	0.08	No-R
55	01199050	Connecticut	50	0.15	0.11	Yes
56	01200000	Connecticut	79	0.37	0.09	Yes
57	01203600	Connecticut	26	-0.32	0.21	No-P
58	01204000	Connecticut	79	0.68	0.11	Yes
59	01208925	Connecticut	39	-0.07	0.13	Yes
60	01208950	Connecticut	52	0.74	0.16	Yes
61	01208990	Connecticut	50	0.05	0.11	Yes
62	01209700	Connecticut	49	1.53	0.34	Yes
63	01094500	Massachusetts	76	0.37	0.09	Yes
64	01096000	Massachusetts	62	-0.14	0.09	Yes
65	01097000	Massachusetts	70	0.12	0.08	Yes
66	01097300	Massachusetts	48	-0.13	0.12	Yes
67	01099500	Massachusetts	74	-0.21	0.08	No-R
68	01100600	Massachusetts	48	0.24	0.12	Yes
69	01101000	Massachusetts	66	0.47	0.11	Yes
70	01102000	Massachusetts	81	0.12	0.07	Yes
71	01102500	Massachusetts	72	0.37	0.09	Yes
72	01104500	Massachusetts	80	0.26	0.08	Yes
73	01105000	Massachusetts	72	0.53	0.11	No-R
74	01105500	Massachusetts	59	0.61	0.13	Yes
75	01105600	Massachusetts	45	-0.14	0.12	Yes
76	01105730	Massachusetts	45	0.12	0.12	Yes
77	01105870	Massachusetts	45	-0.05	0.12	Yes
78	01106000	Rhode Island	38	-0.57	0.19	Yes
79	01108000	Massachusetts	65	0.05	0.08	Yes
80	01109000	Massachusetts	86	0.27	0.08	Yes

Table 1. Streamgages in Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, and easternmost New York used in the New England regional skew analysis.—Continued

[USGS, U.S. Geological Survey; EMA, expected moments algorithm; MSE, mean squared error; No-R, streamgage not used in regional skew analysis because of redundancy; No-P, streamgage not used in regional skew analysis because the pseudorecord length (P_{RL}) was less than 30 years]

Regional skew index number	USGS streamgage number	State where streamgage is located	P_{RL}	EMA analysis		Used in regional skew model
				Skew, in log units	MSE of skew, in log units	
81	01109060	Massachusetts	45	-0.39	0.14	No-R
82	01109070	Massachusetts	45	0.05	0.12	Yes
83	01109500	Massachusetts	55	0.60	0.14	Yes
84	01110000	Massachusetts	72	0.01	0.07	Yes
85	01110500	Massachusetts	47	0.41	0.14	No-R
86	01111300	Rhode Island	45	0.24	0.13	No-R
87	01111500	Rhode Island	69	0.18	0.09	Yes
88	01112500	Rhode Island	83	0.38	0.08	No-R
89	01114500	Rhode Island	71	-0.14	0.08	Yes
90	01116000	Rhode Island	71	0.60	0.11	Yes
91	01117000	Rhode Island	71	0.96	0.15	Yes
92	01117350	Rhode Island	38	0.92	0.22	Yes
93	01117420	Rhode Island	36	0.52	0.18	No-R
94	01117468	Rhode Island	37	0.84	0.21	Yes
95	01117500	Rhode Island	71	0.96	0.15	Yes
96	01118000	Rhode Island	70	1.04	0.17	Yes
97	01118500	Rhode Island	72	0.96	0.15	No-R
98	01161500	Massachusetts	66	1.28	0.23	No-R
99	01162000	Massachusetts	94	1.11	0.15	Yes
100	01162500	Massachusetts	94	0.57	0.09	Yes
101	01163200	Massachusetts	47	-0.02	0.11	Yes
102	01165000	Massachusetts	32	1.17	0.32	Yes
103	01165500	Massachusetts	66	0.32	0.10	Yes
104	01168500	Massachusetts	98	0.38	0.07	No-R
105	01169000	Massachusetts	72	0.84	0.13	Yes
106	01169900	Massachusetts	45	0.85	0.19	Yes
107	01170100	Massachusetts	44	1.21	0.28	Yes
108	01171300	Massachusetts	37	-0.52	0.18	Yes
109	01171500	Massachusetts	73	-0.34	0.09	Yes
110	01173500	Massachusetts	45	1.78	0.45	Yes
111	01174000	Massachusetts	35	-0.47	0.18	Yes
112	01174500	Massachusetts	75	0.80	0.12	Yes
113	01174900	Massachusetts	36	-0.28	0.16	Yes
114	01175670	Massachusetts	51	0.00	0.10	Yes
115	01176000	Massachusetts	99	1.58	0.26	Yes
116	01179500	Massachusetts	32	0.76	0.23	Yes
117	01180000	Massachusetts	29	0.39	0.21	No-P
118	01180500	Massachusetts	54	0.95	0.18	Yes
119	01181000	Massachusetts	76	0.37	0.09	Yes
120	01185500	Massachusetts	98	1.05	0.14	Yes

Table 1. Streamgages in Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, and easternmost New York used in the New England regional skew analysis.—Continued

[USGS, U.S. Geological Survey; EMA, expected moments algorithm; MSE, mean squared error; No-R, streamgage not used in regional skew analysis because of redundancy; No-P, streamgage not used in regional skew analysis because the pseudorecord length (P_{RL}) was less than 30 years]

Regional skew index number	USGS streamgage number	State where streamgage is located	P_{RL}	EMA analysis		Used in regional skew model
				Skew, in log units	MSE of skew, in log units	
121	01197000	Massachusetts	76	0.25	0.08	No-R
122	01197500	Massachusetts	76	0.40	0.09	Yes
123	01198000	Massachusetts	28	0.84	0.15	No-P
124	01331500	Massachusetts	79	0.69	0.11	Yes
125	01332000	Massachusetts	63	0.50	0.12	Yes
126	01333000	Massachusetts	62	-0.05	0.09	Yes
127	01333500	New York	64	0.53	0.12	Yes
128	01334500	New York	100	0.82	0.10	No-R
129	01360640	New York	21	-0.65	0.30	No-P
130	01052500	New Hampshire	70	0.37	0.10	Yes
131	01064300	New Hampshire	41	0.24	0.14	Yes
132	01064400	New Hampshire	28	-0.14	0.19	No-P
133	01064500	New Hampshire	88	-0.11	0.05	No-R
134	01065000	New Hampshire	62	0.23	0.09	No-R
135	01073000	New Hampshire	77	0.18	0.08	Yes
136	01073500	New Hampshire	77	0.41	0.09	Yes
137	01073600	New Hampshire	47	0.29	0.12	Yes
138	01074520	New Hampshire	49	0.10	0.07	Yes
139	01075800	New Hampshire	35	0.49	0.19	Yes
140	01076500	New Hampshire	108	0.39	0.07	Yes
141	01078000	New Hampshire	122	0.42	0.06	Yes
142	01082000	New Hampshire	68	0.38	0.09	Yes
143	01084000	New Hampshire	50	0.74	0.11	Yes
144	01084500	New Hampshire	63	0.48	0.11	Yes
145	01085800	New Hampshire	49	0.06	0.11	Yes
146	01086000	New Hampshire	52	0.33	0.09	Yes
147	01089100	New Hampshire	59	0.12	0.09	Yes
148	01089500	New Hampshire	62	0.59	0.09	Yes
149	01091000	New Hampshire	44	1.11	0.18	Yes
150	01093800	New Hampshire	41	0.26	0.14	Yes
151	01094000	New Hampshire	160	0.45	0.05	Yes
152	01127880	New Hampshire	21	-0.43	0.27	No-P
153	01130000	New Hampshire	71	-0.17	0.08	Yes
154	01133000	New Hampshire	70	0.84	0.12	Yes
155	01134500	New Hampshire	65	0.37	0.10	Yes
156	01135150	New Hampshire	21	1.49	0.56	No-P
157	01135300	New Hampshire	21	1.25	0.45	No-P
158	01135500	New Hampshire	83	0.36	0.08	Yes
159	01137500	New Hampshire	72	0.24	0.09	Yes
160	01139000	New Hampshire	71	0.25	0.09	Yes

Table 1. Streamgages in Connecticut, Maine, Massachusetts, New Hampshire, Rhode Island, Vermont, and easternmost New York used in the New England regional skew analysis.—Continued

[USGS, U.S. Geological Survey; EMA, expected moments algorithm; MSE, mean squared error; No-R, streamgage not used in regional skew analysis because of redundancy; No-P, streamgage not used in regional skew analysis because the pseudorecord length (P_{RL}) was less than 30 years]

Regional skew index number	USGS streamgage number	State where streamgage is located	P_{RL}	EMA analysis		Used in regional skew model
				Skew, in log units	MSE of skew, in log units	
161	01139800	New Hampshire	53	0.19	0.11	Yes
162	01141800	New Hampshire	39	0.04	0.13	Yes
163	01142500	New Hampshire	163	0.97	0.08	Yes
164	01144000	New Hampshire	96	0.99	0.13	Yes
165	01145000	New Hampshire	71	-0.17	0.08	Yes
166	01150900	New Hampshire	84	1.94	0.36	Yes
167	01152500	New Hampshire	83	0.24	0.08	Yes
168	01153550	New Hampshire	69	0.37	0.10	Yes
169	01154000	New Hampshire	58	0.79	0.16	Yes
170	01155000	New Hampshire	46	0.54	0.16	Yes
171	01157000	New Hampshire	64	0.74	0.15	Yes
172	01160000	New Hampshire	60	0.71	0.15	Yes
173	01329000	Vermont	80	0.45	0.09	Yes
174	01334000	Vermont	80	0.04	0.07	Yes
175	04280000	Vermont	83	0.24	0.08	Yes
176	04280350	Vermont	25	-0.01	0.20	No-P
177	04282000	Vermont	83	-0.12	0.07	Yes
178	04282650	Vermont	22	-0.01	0.22	No-P
179	04282780	Vermont	22	0.31	0.25	No-P
180	04282795	Vermont	22	0.03	0.23	No-P
181	04287000	Vermont	78	0.25	0.08	Yes
182	04288000	Vermont	164	0.76	0.06	Yes
183	04292000	Vermont	85	0.11	0.06	Yes
184	04292700	Vermont	23	0.52	0.18	No-P
185	04296000	Vermont	60	-0.02	0.09	Yes
186	010965852	Vermont	25	0.17	0.21	No-P

Appendix 1. Assessment of New England Regional Skew Constant Model Through Monte Carlo Realizations

The Monte Carlo simulations presented in this appendix address whether the apparent observed geospatial structure in the station skewness coefficient (skews; fig. 1–1) is evidence of model misspecification or is an artifact of random sampling variability, possibly confounded by the covariance structure of the errors. The Monte Carlo simulations were generated from a multivariate normal distribution with a mean equal to the constant from the New England regional skew constant model and a covariance matrix identical to the covariance matrix from the New England regional skew model. Both the mean and covariance of the multivariate normal distribution from which samples were drawn are described in detail below.

Observed, unbiased at-site skews developed for the New England Bayesian weighted least squares/Bayesian generalized least squares (B–WLS/B–GLS) regional skew constant model described in this report are depicted in figure 1–1, along with contour lines and shading to provide a sense of geographic patterns in the skews. The contouring algorithm used to generate the map in figure 1–1 shows a substantial amount of geospatial structure in the pattern of the unbiased at-site skews. In particular, the smaller skews (negative skews) in Maine, especially in northern and northwestern Maine (shaded blue and green), might be a cause for concern that a constant model is not representing all the observed variability in skew.

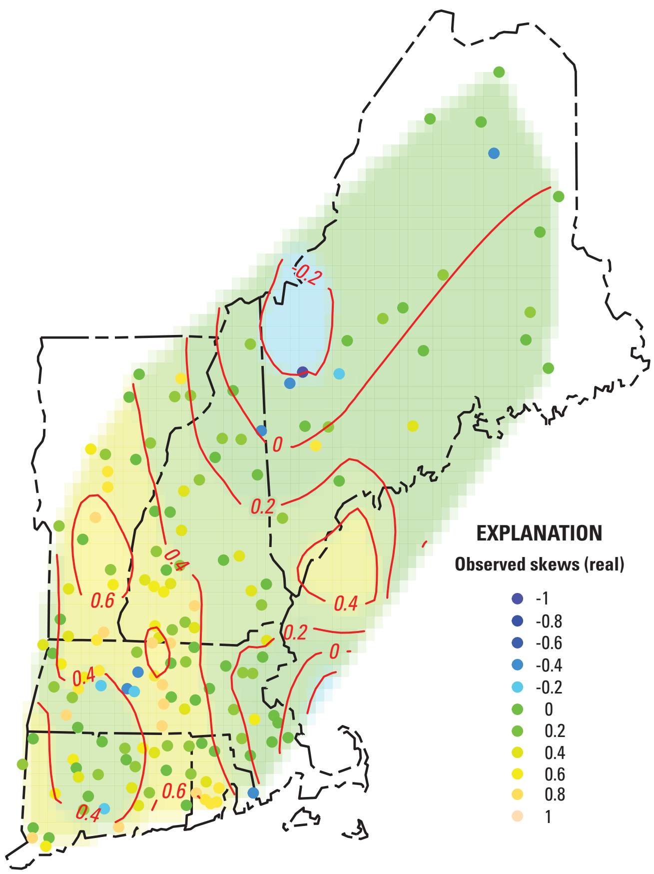


Figure 1–1. Map of at-site unbiased (unweighted) skews computed at 186 streamgages used to develop a regional skew for New England. The contour lines and shading provide a sense of geographic patterns in the skews.

The skew $\hat{\gamma}_{BWLS/BGLS}$ from the constant model for the New England regional skew (table 2) is calculated from the following:

$$\hat{\gamma}_{BWLS/BGLS} = 0.37 + \varepsilon\,, \tag{1–1}$$

where ε represents the total error, calculated as follows:

$$\varepsilon \sim N(0,\ Var(\varepsilon)), \tag{1–2}$$

where N signifies a normal distribution of the total error in constant regional skew model determined in the B–GLS analysis. As described in equation 23, $Var(\varepsilon)$ can be described as follows:

$$\varepsilon\varepsilon^{T} = \Lambda_{GLS}\left(\sigma^2_{\delta,B-GLS}\right) = \sigma^2_{\delta,B-GLS} I + \sum \hat{\gamma} \tag{1–3}$$

where

$\Lambda_{GLS}\left(\sigma^2_{\delta,B-GLS}\right)$ is the $n\times n$ GLS covariance matrix,

$\sigma^2_{\delta,B-GLS}$ is the B–GLS variance of the underlying model error δ,

I is an $n\times n$ identity matrix, and

$\sum \hat{\gamma}$ is the full $n\times n$ covariance matrix of the sampling errors for each streamgage (i).

The covariance matrix of the sampling errors is made up of the sampling variances of the unbiased at-site skew ($Var\left[\hat{\gamma}_i\right]$) and the covariances of skew estimators ($\hat{\gamma}_i$). The off-diagonal values of $\Sigma\left(\hat{\gamma}\right)$ are determined by the cross correlation of concurrent systematic annual peak flows and the *cf* factor that accounts for the sample size difference between streamgages and their concurrent record length (eq. 17; Martins and Stedinger, 2002, eq. 3).

The model error variance σ^2_δ for the New England regional skew constant model is 0.13 (table 2), which was used in the Monte Carlo simulations. The covariance matrix $\Sigma\left(\hat{\gamma}\right)$ used in the Monte Carlo simulations is the same as the one used in the B–WLS/B–GLS (eq. 25).

The results of the Monte Carlo simulations are depicted graphically (fig. 1–2) in twenty realizations of the expected patterns in the station skew if the station skews are normally distributed with a mean equal to 0.37 and with the $Var(\varepsilon)$ given by equation 35. The Monte Carlo simulations reveal no consistent spatial pattern of the station skews, consistent with the observed pattern of the station skews of the constant model (figs. 1–1 and 1–2). Therefore, there is little evidence of a lack of fit based on the geographic patterns observed in the station skews.

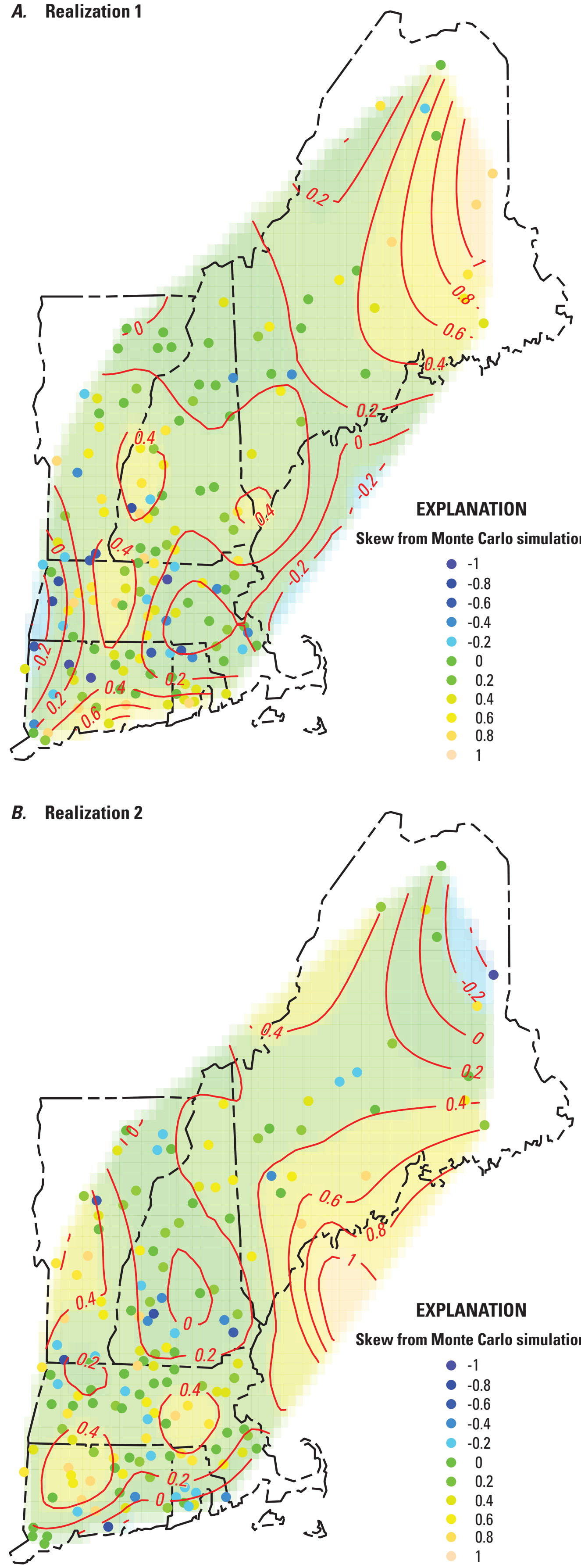


Figure 1–2. Results of *A*–*T*, twenty Monte Carlo simulations of skews at 153 streamgages in New England. Simulations were generated from a multivariate normal distribution with a mean equal to the constant (0.37) from the New England regional skew model and a covariance matrix identical to the covariance matrix from the New England regional skew model. The contour lines and shading provide a sense of geographic patterns in the skews.

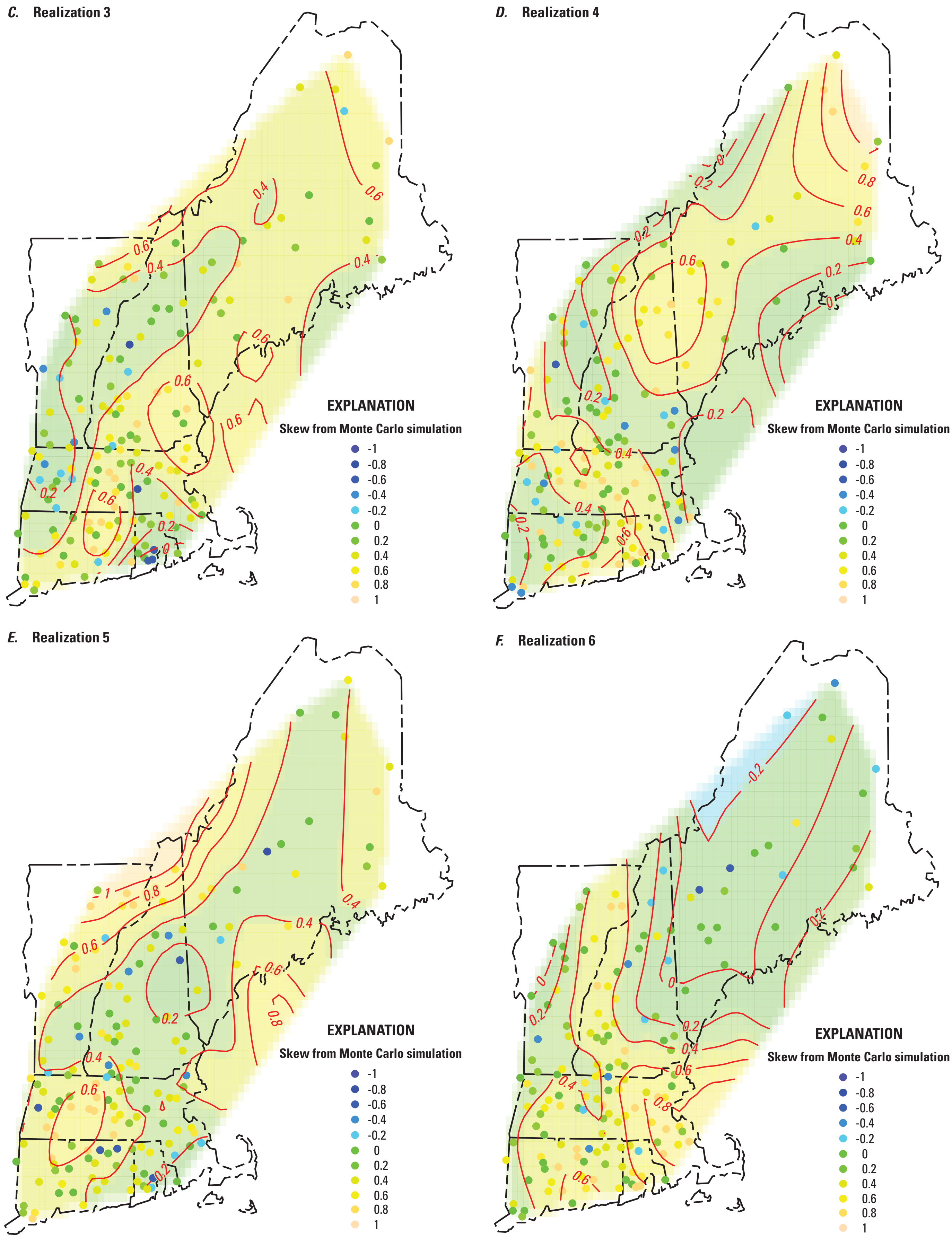


Figure 1–2.—Continued

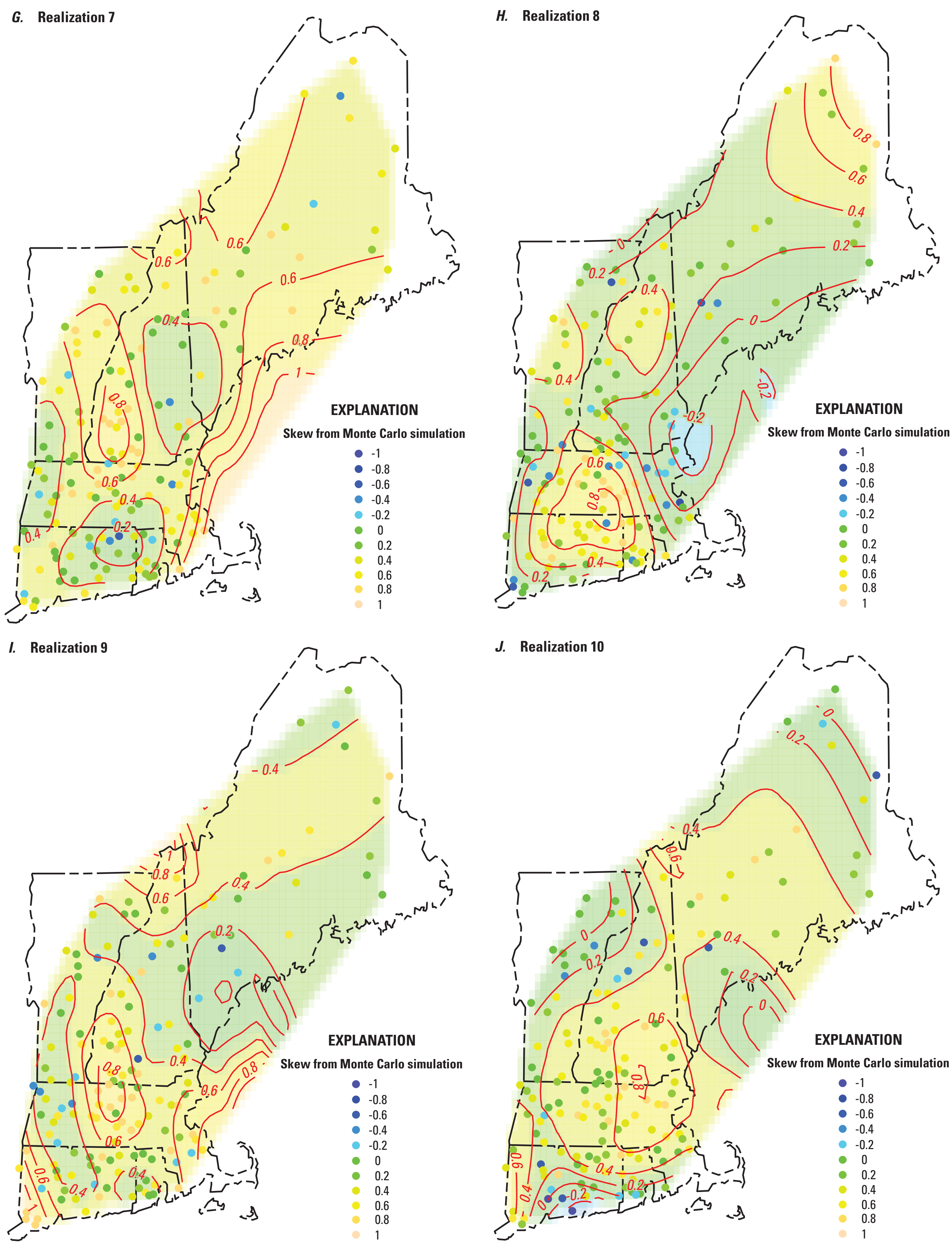


Figure 1–2.—Continued

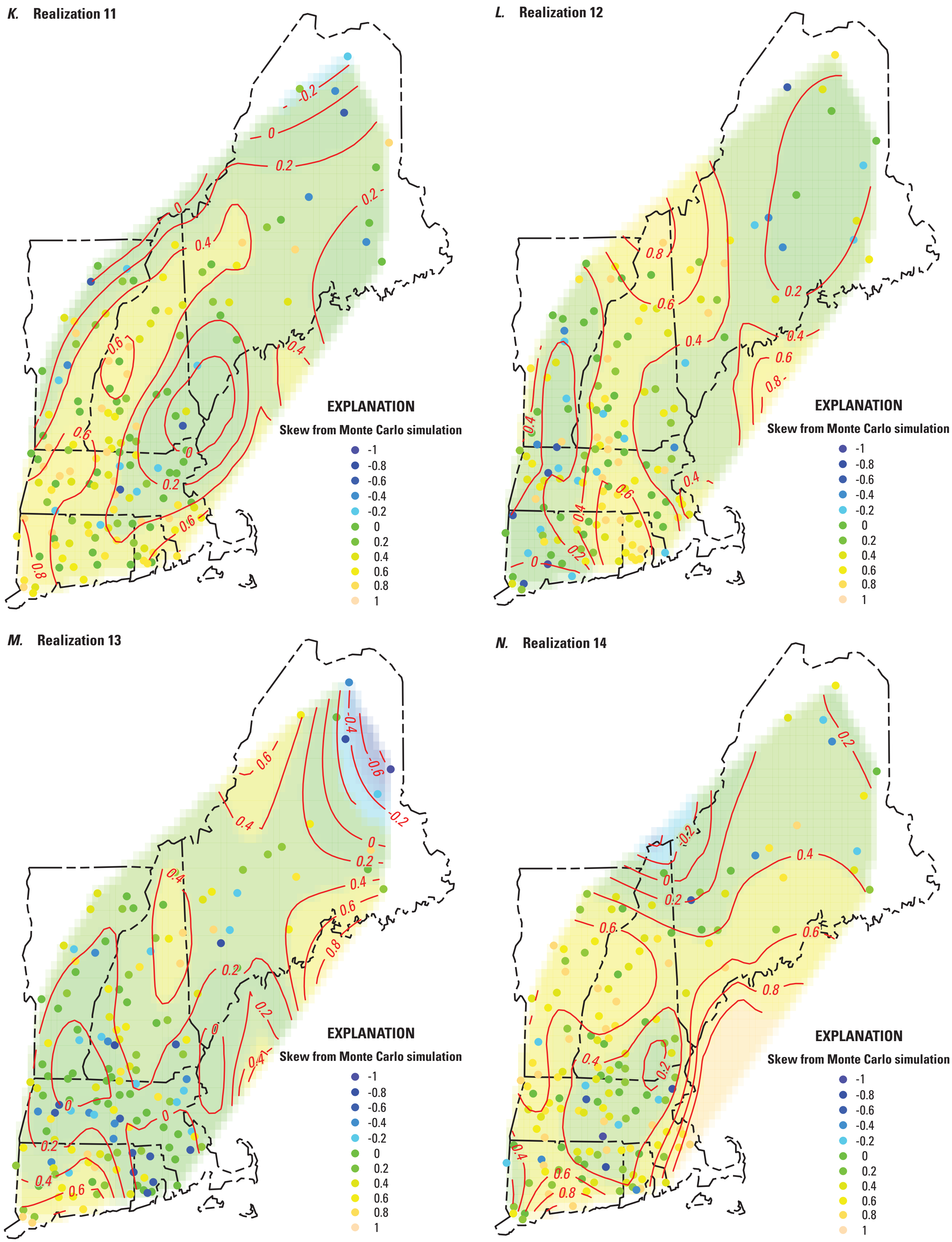
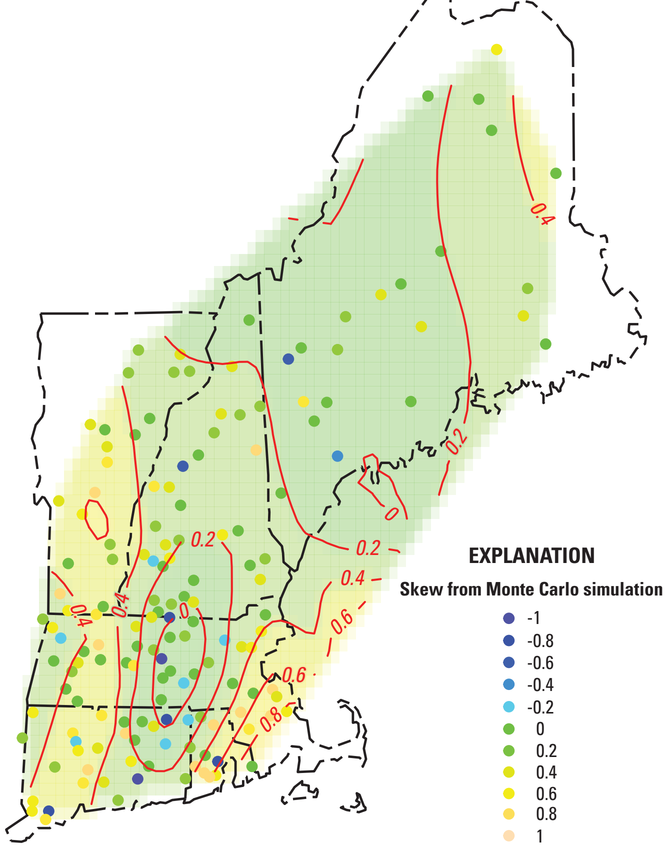
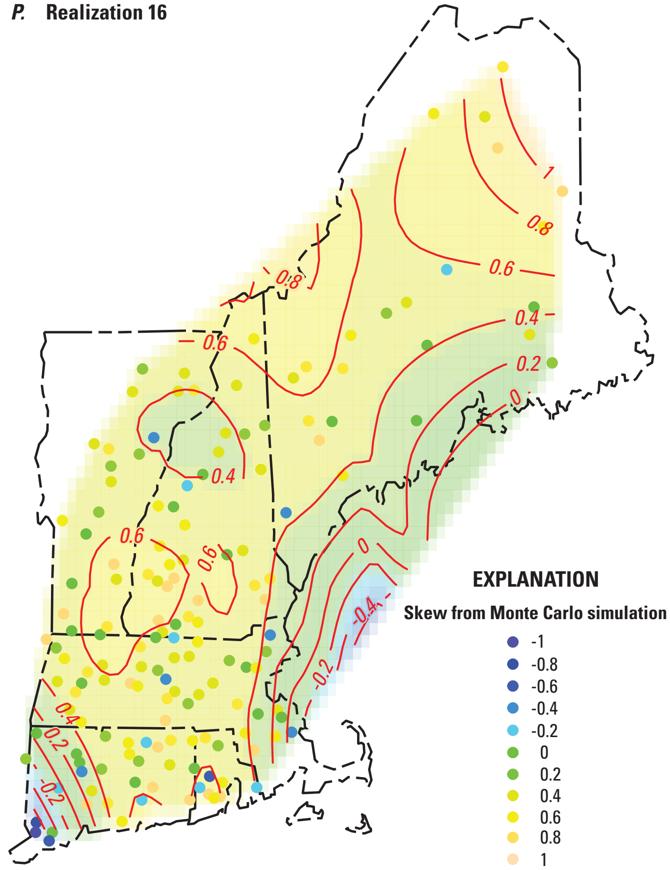


Figure 1–2.—Continued

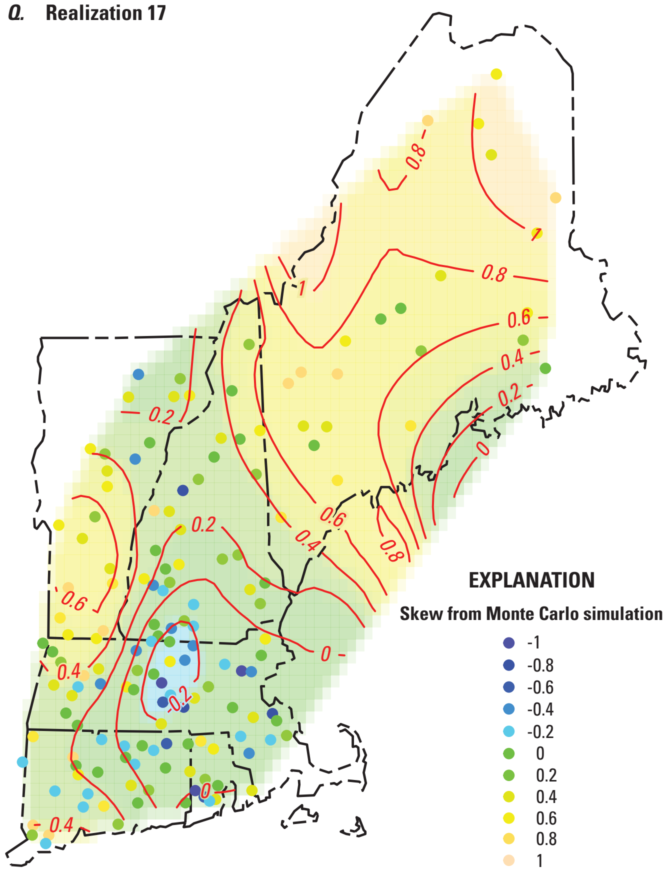
O. **Realization 15**



P. **Realization 16**



Q. **Realization 17**



R. **Realization 18**

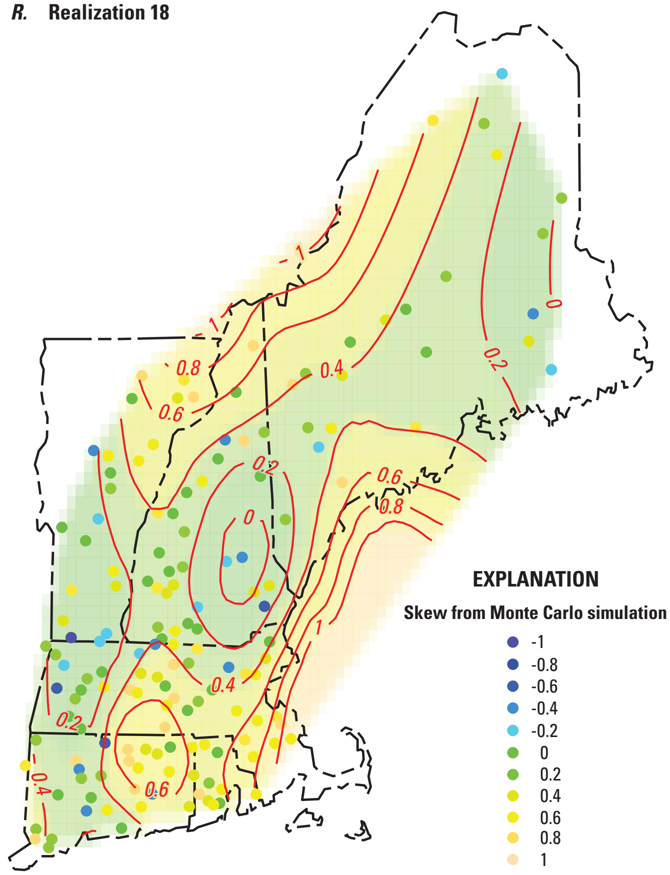
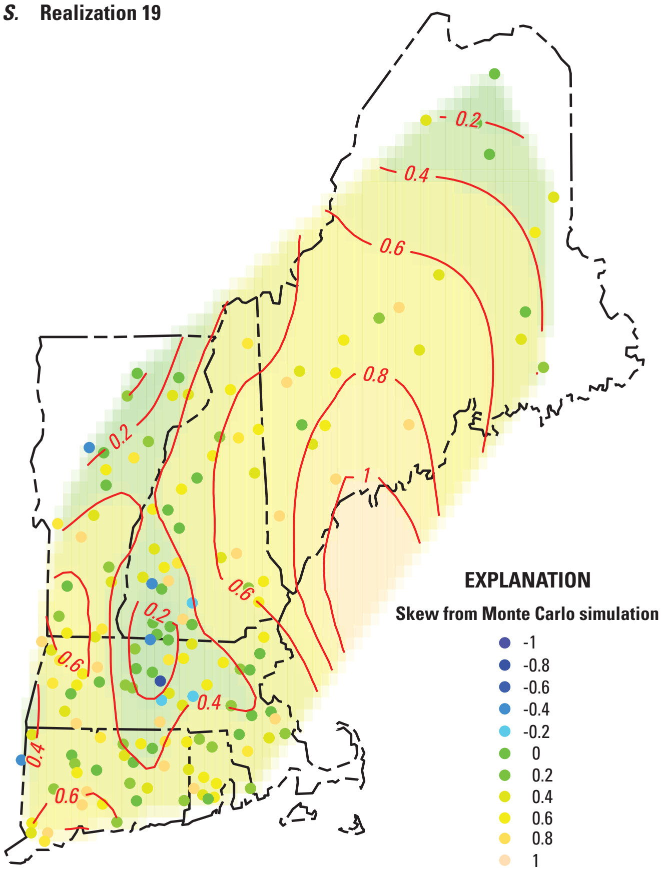


Figure 1–2.—Continued

S. **Realization 19**



T. **Realization 20**

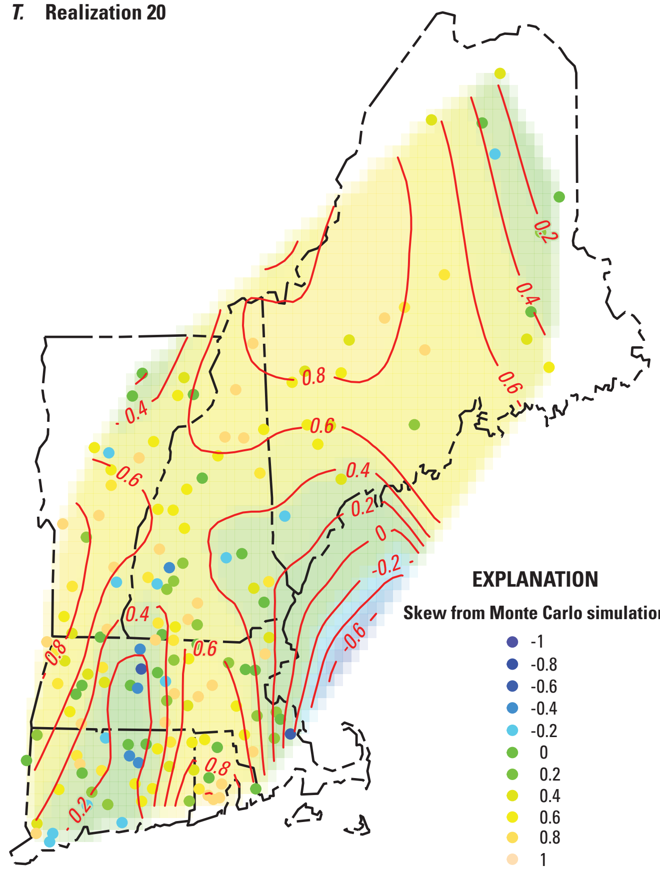


Figure 1–2.—Continued

For more information about this report, contact:
U.S. Geological Survey
Integrated Modeling and Prediction Division
Water Mission Area
12201 Sunrise Valley Drive, MS 415
Reston, VA 20192

Publishing support provided by the
Pembroke Publishing Service Center

Veilleux and others—**Methods for Estimating Regional Coefficient of Skewness for Unregulated Streams in New England**—SIR 2017–5037

Printed on recycled paper

ISSN 2328-0328 (online)
https://doi.org/10.3133/sir20175037