MAGNITUDE AND FREQUENCY OF FLOODS IN RURAL BASINS OF GEORGIA, SOUTH CAROLINA, AND NORTH CAROLINA: A MULTI-STATE APPROACH


INTRODUCTION

Reliable estimates of the magnitude and frequency of floods are required for the design of transportation and water-conveyance structures, such as roads, bridges, culverts, dams, and levees. Federal, State, regional, and local officials rely on these estimates to effectively plan and manage land use and water resources, to protect lives and property in flood-prone areas, and to determine flood-insurance rates. Estimates of the magnitude and frequency of floods are needed at locations where gaged stations monitor streamflow continuously as well as at ungaged sites, where streamflow information is not available for use in determining the estimates. A process known as regionalization—where flood-frequency information determined for a group of streamgaging stations within a hydrologic region forms the basis of estimates for ungaged sites within that region—is used to estimate the magnitude and frequency of floods for ungaged sites. Historically, these hydrologic regions were determined independently by each State, which occasionally led to differences in the hydrologic regions at State boundaries. These differences can cause discontinuity and confusion regarding which flood-frequency techniques and results are most appropriate for river basins near or crossing State boundaries. The U.S. Geological Survey (USGS), in cooperation with the Georgia, South Carolina, and North Carolina Departments of Transportation (GDOT, SCDOT, and NCDOT, respectively) and the North Carolina Floodplain Mapping Program, conducted a regional-regression investigation using a multi-state approach by using hydrologic regions that cross State boundaries in order to provide continuity along State boundaries. More in-depth information on this investigation can be found in USGS Scientific Investigation Reports by Feaster and others (2009), Gotvald and others (2009), and Weaver and others (2009).

DESCRIPTION OF STUDY AREA

The study area includes Georgia, South Carolina, and North Carolina, covering an area of about 142,500 square miles ($\text{mi}^2$) within seven U.S. Environmental Protection Agency (USEPA) level III ecoregions—Southwestern Appalachians, Blue Ridge, Ridge and Valley, Piedmont, Southeastern Plains, Southern Coastal Plain, and Middle Atlantic Coastal Plain (fig. 1) (U.S. Environmental Protection Agency, 2007). The ecoregions represent areas of general similarity in ecosystems and in the type, quality, and quantity of environmental resources. The ecoregions provide a spatial framework for the research, assessment, management, and monitoring of
ecosystems and ecosystem components (Griffith and others, 2002). The Fall Line (fig.1) separates the higher elevation Southwestern Appalachians, Blue Ridge, Ridge and Valley, and Piedmont ecoregions from the low lying Southeastern Plains, Southern Coastal Plain, and Middle Atlantic Coastal Plain ecoregions.

Figure 1. Study area and ecoregions in Georgia, South Carolina, North Carolina, and surrounding States.

DATA COMPILATION

For this investigation, a list of rural, USGS streamgaging stations with 10 or more years of annual peak-flow data were compiled. The peak-flow data were reviewed for quality assurance
and quality control (QA/QC) as well as for homogeneity or absence of trends, which implies relatively constant watershed and climatic conditions during the period of record. Once peak-flow records were compiled and reviewed, basin characteristics were determined for each of the streamgaging stations.

**Peak-Flow Data:** Annual peak-flow data at streamgaging stations in Georgia, South Carolina, North Carolina, and adjacent parts of Alabama, Florida, Tennessee, and Virginia were investigated for possible use in this investigation. After the QA/QC review of the peak-flow data, 943 streamgaging stations were considered for use in the study (fig. 2). The 943 streamgaging stations comprise 357 streamgaging stations in Georgia, 82 in South Carolina, 333 in North Carolina, 35 in Alabama, 23 in Florida, 41 in Tennessee, and 72 in Virginia.

**Figure 2.** Locations of rural streamgaging stations in Georgia, South Carolina, North Carolina, and surrounding States considered for use in the regional regression analysis.
Physical and Climatic Basin Characteristics: Peak-flow information can be estimated at ungaged sites through multiple regression analysis, which relates streamflow characteristics such as the 1-percent chance exceedance flow, (also often referred to as the 100-year flood), to selected physical and climatic basin characteristics for gaged drainage basins. Physical and climatic basin characteristics were selected for use as potential explanatory variables in the regression analyses on the basis of the conceptual relation to flood flows and the ability to measure the basin characteristics using digital datasets and Geographic Information System (GIS) technology. Twenty basin characteristics for each of the 943 rural, streamgaging stations were determined, which included drainage area, main channel length, basin perimeter, main channel slope, mean basin slope, basin shape factor, mean basin elevation, maximum basin elevation, minimum basin elevation, percentage of basin that is impervious, percentage of basin that is forested, mean annual precipitation, maximum 24-hour precipitation with recurrence intervals of 2-, 10-, 25-, 50- and 100-years, soil drainage index, hydrologic soil index, and drainage density.

ESTIMATION OF FLOOD MAGNITUDE AND FREQUENCY AT STREAMGAGING STATIONS

A frequency analysis of annual peak-flow data at a streamgaging station provides an estimate of the flood magnitude and frequency at that specific site. Flood-frequency flows in previous USGS reports were expressed as T-year floods based on the recurrence interval for that flood quantile (for example, the “100-year flood”). Some individuals in the water-resources science community now discourage the use of recurrence-interval terminology because it sometimes causes confusion in the general public. The term is sometimes interpreted to imply that there is a set time interval between floods of a particular magnitude, when in fact floods are random processes that are best understood using probabilistic terms. Misunderstandings with the T-year recurrence-interval terminology primarily have to do with the number of times that a peak flow of certain magnitude could occur (and has been equaled or exceeded) during the T-year period. While the T-year recurrence-interval flood is statistically expected to occur, on average, once during the T-year period, it may occur multiple times during the period or not at all.

In an attempt to clarify the issue of flood recurrence, the terminology associated with flood-frequency characterization is undergoing a transition away from referring to flood probabilities in terms of the T-year recurrence-interval flood and instead referring to the P-percent chance exceedance flow (or flood). The use of percent chance exceedance flow is now recommended because it conveys the probability, or odds, of a flood of a given magnitude being equaled or exceeded in any given year. For example, a 1-percent chance exceedance flow (formerly referred to as the "100-year flood") corresponds to the flow magnitude that has a probability of 0.01 of being equaled or exceeded in any given year. That is, a flow with an exceedance probability of 0.01 has a 1 percent chance of being equaled or exceeded in any given year. Recurrence interval and exceedance probability are the mathematical inverses of one another; therefore, a flood with an exceedance probability of 0.01 (1-percent chance exceedance) has a recurrence interval of 1/0.01 or 100 years. Thus, the P-percent chance exceedance is computed as the inverse of the T-year recurrence interval multiplied by 100. For example, a 100-year recurrence interval is
equivalent to a 1-percent chance exceedance (1/100-year recurrence interval x 100 = 1-percent chance exceedance), and a 25-year recurrence interval is equivalent to a 4-percent chance exceedance (1/25-year recurrence interval x 100 = 4-percent chance exceedance).

**Flood Frequency:** Flood-frequency estimates were computed by fitting logarithms (base 10) of the annual peak flows to a Pearson Type III distribution. This follows the guidelines and computational methods described in Bulletin 17B of the Hydrology Subcommittee of the Interagency Advisory Committee on Water Data (IACWD, 1982).

The USGS computer program PEAKFQWin version 5.2.0 was used to compute the flood-frequency estimates for the 943 streamgaging stations considered. PEAKFQWin automates many of the analysis procedures recommended in Bulletin 17B (IACWD, 1982), including identifying and adjusting for outliers and historical periods, weighting of station skews with a generalized skew, and fitting a log-Pearson Type III distribution to the annual peak-flow data. The station skew coefficients were weighted with a new generalized skew coefficient developed for this study.

**Generalized Skew Analysis:** During the development of flood-frequency techniques in the late 1970s and early 1980s, a nationwide generalized skew study was conducted and documented in Bulletin 17B (IACWD, 1982), where the term “generalized” refers to a regional analysis. Station skew coefficients for long-term streamgaging stations throughout the Nation were computed and used to produce a map of isolines of generalized skew; however, the map was prepared at a national scale using data and methods that are now more than 30 years old. In order to generate more accurate generalized skew coefficients, three methods are described in Bulletin 17B (IACWD, 1982) for developing generalized skews using skew coefficients computed from streamgaging stations with long-term peak-flow record (25 or more years of record): (1) plot station skew coefficients on a map and construct skew isolines, (2) use regression techniques to develop a skew prediction equation relating station skew coefficients to some set of basin characteristics, or (3) use the arithmetic mean of station skew coefficients from long-term streamgaging stations in the area.

A generalized skew coefficient analysis using methods 1 and 2 was performed as part of this investigation. The initial dataset for the generalized skew coefficient analysis included 489 streamgaging stations with at least 25 years of annual peak-flow data located in the study area. The station skew coefficients for these 489 streamgaging stations were plotted at the centroid of the watershed for each streamgaging station in order to develop a skew isoline map. The map was then reviewed to determine if any geographic or topographic trends were visually apparent. No clearly definable patterns were found, so regression techniques were then used to develop a skew prediction equation using station skew as the response variable. A Bayesian generalized least squares (GLS) regression model was used for the generalized skew coefficient regression analysis as suggested by Reis and others (2005) and Gruber and others (2007), both of which made use of the methods proposed by Martins and Stedinger (2002). This study also used the methods from Martins and Stedinger (2002) except that the distance between basin centroids was used instead of the distance between gaging stations.

A screening metric was developed to determine redundant station pairs that, in essence, represent the same watershed. After applying the screening metric on the 489 streamgaging stations with at least 25 years of annual peak-flow data, a total of 92 streamgaging stations were removed from
the generalized skew coefficient regression analysis. An additional 55 streamgaging stations were removed because of censored peak-flow data, which are peak flows that are less than the minimum recorded peak flow at a streamgaging station. A total of 342 streamgaging stations were used for the final Bayesian GLS regression analysis of skew.

Based on the Bayesian GLS regression analysis, a constant generalized skew value of –0.019 was determined to be the most reasonable approach to predicting the generalized skew in the study area. More complicated Bayesian GLS models with additional explanatory variables were evaluated but resulted in very modest improvements in accuracy. The modest improvements in the more complicated models were not justified because of the increased complexity associated with the addition of explanatory variables. The mean square error (MSE) associated with the constant generalized skew model is 0.143. This MSE is equivalent to 39 years of record length. This is a substantial improvement over the Bulletin 17B skew map MSE value of 0.302, which is equivalent to 17 years of record. The generalized skew value of –0.019 with an associated MSE of 0.143 was used to compute the flood-frequency estimates for the 943 streamgaging stations according to methods recommended in Bulletin 17B (IACWD, 1982).

ESTIMATION OF FLOOD MAGNITUDE AND FREQUENCY AT UNGAGED SITES

A regional regression analysis was used to develop a set of equations for use in estimating the magnitude and frequency of floods for rural, ungaged sites in Georgia, South Carolina, and North Carolina. These equations relate the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent chance exceedance flows computed from available records for streamgaging stations to measured physical and climatic basin characteristics of the drainage basins for the streamgaging stations. All 943 rural, streamgaging stations used in the flood-frequency analysis were considered for use in the regional regression analysis (fig. 2).

**Regression Analysis:** Selection of the explanatory variables for each hydrologic region was based on all-possible-subsets (APS) regression methods (Neter and others, 1985). The final explanatory variables for each hydrologic region were selected on the basis of several factors, including standard error of the estimate, Mallow’s \( Cp \) statistic, statistical significance of the explanatory variables, coefficient of determination \( r^2 \), and ease of measurement of explanatory variables. Multicollinearity, a situation in which two or more explanatory variables in a multiple regression model are highly correlated, in the candidate exploratory variables also was assessed by the variance inflation factor (VIF) and the correlation between explanatory variables.

Generalized least-squares regression methods, as described by Stedinger and Tasker (1985), were used to determine the final regional P-percent chance exceedance flow regression equations, using the USGS computer program GLSNET (Tasker and Stedinger, 1989; G.D. Tasker, K.M. Flynn, A.M. Lumb, and W.O. Thomas, Jr., U.S. Geological Survey, written commun., 1995). In order to remove the redundancy associated with streamgaging stations that represent the same watershed for the regional GLS regression analysis, two streamgaging stations on the same stream where the percentage change in drainage area from one streamgaging
station to another was within 50 percent were flagged as redundant streamgaging stations. A total of 52 streamgaging stations were omitted from the regional regression analysis because of redundant record, leaving 891 streamgaging stations.

**Regionalization of Flood-Frequency Estimates:** An ordinary least squares (OLS) regression analysis was performed on the 891 remaining rural, streamgaging stations in order to determine the need for separate hydrologic regions within the study area. The residuals from the OLS analysis were plotted for each USEPA level III ecoregion in order to determine the need for dividing the ecoregions into subregions and(or) for the possibility of combining ecoregions to form a hydrologic region. Based on assessments of geographical patterns in the residuals of the regression estimates, five hydrologic regions were established using the USEPA ecoregions (fig. 3): Piedmont-Ridge and Valley (Region 1), Blue Ridge (Region 2), Sand Hills (Region 3), Coastal Plain (Region 4) and Southwest Georgia (Region 5).

![Figure 3](image-url)  
**Figure 3.** Hydrologic regions and locations of rural streamgaging stations in the study area used in the regional regression analysis.
Considerable variability in the predicted P-percent chance exceedance flow for streamgaging stations with drainage areas less than 1 mi² was observed in the OLS regression analyses for each of the five hydrologic regions. This variability can be attributed, in part, to the difficulty in measuring peak flows at streamgaging stations with small drainage basins. In addition to measurement errors, runoff hydrology and hydraulics likely are different for small watersheds, which typically are sensitive to land use and high-intensity rainfalls of short duration (usually not occurring over large basins). Because of the lack of sufficient data in some regions and the variability and uncertainty of data for small watersheds, the 44 streamgaging stations with drainage areas less than 1 mi² were omitted from the final regional regression analysis. Additionally, there was only one streamgaging station with a drainage area greater than 9,000 mi² available for use in the regional regression analysis. This streamgaging station has a drainage area of 13,600 mi² and was omitted from the final regional regression analysis because of the lack of data points needed to develop the regression analysis model for a drainage area greater than 8,930 mi².

**Regional Regression Equations:** Traditionally, flood-frequency equations only apply to a single hydrologic region. As a result, streamgaging stations that have significant drainage from more than one region are not included in the regression analysis. The use of a basin-percentage variable allows for the inclusion of such streamgaging stations and resulted in an additional 83 streamgaging stations used in the analysis for this particular study. The use of basin percentages in the regression coefficients allows for a smooth transition of flood estimates for drainage basins that do not lie exclusively within one hydrologic region. For example, the slopes for the Blue Ridge and Sand Hills hydrologic regions (Regions 2 and 3, respectively) regression lines are visibly different from the slopes for the regression lines for the three remaining hydrologic regions (fig. 4A). As a result, the final equations include a "slope adjustment factor" for hydrologic regions 2 and 3. The transition from a site located exclusively within hydrologic region 1, represented by the "base" slope (for example, 0.594 for the 1-percent chance exceedance flow), to a site located exclusively within hydrologic region 2 is depicted in figure 4B. Note that the slope of the regression line becomes visibly steeper as a basin goes from 100 percent in the Piedmont – Ridge and Valley region (Region 1) to a basin that is 100 percent in the Blue Ridge region (Region 2).
A GLS analysis was run on the final 828 rural, streamgaging stations considered for the regional regression analysis. The slope of the relation between hydrologic regions 2 and 3 was found to be significantly different at the 95-percent probability level for the 50- through 0.2-percent chance exceedance flows ($Q_{50\%}$ and $Q_{0.2\%}$, respectively). A slope adjustment factor was included in the final regression equations for hydrologic regions 2 and 3. The results of the analyses are a single P-percent chance exceedance flow estimate equation for the entire study area, which is composed of five hydrologic regions (fig. 3). The final regional regression equations for the 50-through 0.2-percent chance exceedance flows are as follows:

![Graph showing rural flood-frequency relations by region and basin transition from the Piedmont to the Blue Ridge region.](image-url)
where

\[ Q_{50\%} = 10^{\left(0.0220(PCT_1) + 0.0204(PCT_2) + 0.0141(PCT_3) + 0.0178(PCT_4) + 0.0196(PCT_5)\right)} DA^{0.649 + 0.00130(PCT_2) + 0.00109(PCT_3)} \]
\[ Q_{20\%} = 10^{\left(0.0247(PCT_1) + 0.0253(PCT_2) + 0.0165(PCT_3) + 0.0209(PCT_4) + 0.0230(PCT_5)\right)} DA^{0.627 + 0.00122(PCT_2) + 0.00117(PCT_3)} \]
\[ Q_{10\%} = 10^{\left(0.0260(PCT_1) + 0.0246(PCT_2) + 0.0177(PCT_3) + 0.0224(PCT_4) + 0.0247(PCT_5)\right)} DA^{0.617 + 0.00119(PCT_2) + 0.00123(PCT_3)} \]
\[ Q_{4\%} = 10^{\left(0.0273(PCT_1) + 0.0260(PCT_2) + 0.0189(PCT_3) + 0.0239(PCT_4) + 0.0265(PCT_5)\right)} DA^{0.606 + 0.00118(PCT_2) + 0.00130(PCT_3)} \]
\[ Q_{2\%} = 10^{\left(0.0282(PCT_1) + 0.0268(PCT_2) + 0.0196(PCT_3) + 0.0249(PCT_4) + 0.0276(PCT_5)\right)} DA^{0.600 + 0.00118(PCT_2) + 0.00135(PCT_3)} \]
\[ Q_{1\%} = 10^{\left(0.0289(PCT_1) + 0.0276(PCT_2) + 0.0202(PCT_3) + 0.0258(PCT_4) + 0.0286(PCT_5)\right)} DA^{0.594 + 0.00119(PCT_2) + 0.00139(PCT_3)} \]
\[ Q_{0.5\%} = 10^{\left(0.0295(PCT_1) + 0.0282(PCT_2) + 0.0208(PCT_3) + 0.0265(PCT_4) + 0.0295(PCT_5)\right)} DA^{0.589 + 0.00120(PCT_2) + 0.00144(PCT_3)} \]
\[ Q_{0.2\%} = 10^{\left(0.0303(PCT_1) + 0.0290(PCT_2) + 0.0214(PCT_3) + 0.0274(PCT_4) + 0.0306(PCT_5)\right)} DA^{0.583 + 0.00121(PCT_2) + 0.00149(PCT_3)} \]

\[ Q_{50\%}, Q_{20\%}, \ldots, Q_{0.2\%} \] are the flows for floods with percent chance exceedance of 50 percent, 20 percent, …, and 0.2 percent, in cubic feet per second; and

\( PCT_1, PCT_2, PCT_3, PCT_4, \) and \( PCT_5 \) are the basin percentages in hydrologic regions 1, 2, 3, 4, and 5, in percent; and \( DA \) is the drainage area, in square miles.

The average standard error of prediction for these equations ranged from 34.0 to 47.7 percent. The standard error of prediction is a measure of the accuracy of the regression equations when predicting floods for watersheds not used in the regression analysis. About two-thirds of the regression estimates for ungaged sites have errors less than the standard error of prediction for the regression equations.

**SUMMARY**

This paper presents methods for determining flood magnitude and frequency at rural streamgaging stations and rural, ungaged sites in Georgia, South Carolina, and North Carolina. For the current study, 828 streamgaging stations in or near these three States were used in the regional regression analysis. Streamgaging stations used for the current study included rural basins having at least 10 years of peak-flow data and are not significantly affected by regulation, tidal fluctuations, or urban development. By using a multi-state analysis, continuity in hydrologic regions and regression equations at the State boundaries is maintained so that there is no confusion as to which flood-frequency techniques and results are most appropriate for drainage basins near or crossing the State boundaries.
A regional analysis of station skew coefficients resulted in one generalized skew coefficient that can be used for the entire study area. The generalized skew value of −0.019 was determined by using a Bayesian Generalized Least-Squares (GLS) regression model. The mean square error (MSE) for the new generalized skew value is 0.143, which is substantially less than the 0.302 MSE for the generalized skew map available in Bulletin 17B. A weighted skew coefficient (using the station and generalized skew values) was used with the log-Pearson Type III analysis to compute the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent chance exceedance flows at each streamgaging station.

Regional regression analysis, using GLS regression, was used to develop a set of predictive equations that can be used to estimate the 50-, 20-, 10-, 4-, 2-, 1-, 0.5-, and 0.2-percent chance exceedance flows for rural, ungaged sites in Georgia, South Carolina, and North Carolina. The predictive equations are a function of drainage area and the percentage of drainage basin within each of the five hydrologic regions defined in the study area. Average errors of prediction for these equations ranged from 34.0 to 47.7 percent. Additional data from streamgaging stations draining multiple hydrologic regions were used in the development of the equations by accounting for the proportional drainage area in each hydrologic region in the regional regression analysis. As such, the predictive equations can be used to estimate the P-percent chance exceedance flows for ungaged sites that have a drainage basin in one or more hydrologic regions.

REFERENCES


