Journal of Hydrology (2007) 344, 82–95

available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/jhydrol

The use of GLS regression in regional hydrologic analyses

V.W. Griffis [a,*], J.R. Stedinger [b,1]

[a] Department of Civil and Environmental Engineering, Michigan Technological University, 1400 Townsend Drive, Houghton, MI 49931-1295, USA
[b] School of Civil and Environmental Engineering, Cornell University, Hollister Hall, Ithaca, NY 14853-3501, USA

Received 27 October 2006; received in revised form 2 May 2007; accepted 30 June 2007

KEYWORDS
Generalized least squares regression;
Flood frequency analysis;
Regional skew;
Log-Pearson type 3 distribution

Summary To estimate flood quantiles and other statistics at ungauged sites, many organizations employ an iterative generalized least squares (GLS) regression procedure to estimate the parameters of a model of the statistic of interest as a function of basin characteristics. The GLS regression procedure accounts for differences in available record lengths and spatial correlation in concurrent events by using an estimator of the sampling covariance matrix of available flood quantiles. Previous studies by the US Geological Survey using the LP3 distribution have neglected the impact of uncertainty in the weighted skew on quantile precision. The needed relationship is developed here and its use is illustrated in a regional flood study with 162 sites from South Carolina. The performance of a pooled regression model is compared to separate models for each hydrologic region: statistical tests recommend an interesting hybrid of the two which is both surprising and hydrologically reasonable. The statistical analysis is augmented with new diagnostic metrics including a condition number to check for multicollinearity, a new pseudo-$\overline{R}^2$ appropriate for use with GLS regression, and two error variance ratios. GLS regression for the standard deviation demonstrates that again a hybrid model is attractive, and that GLS rather than an OLS or WLS analysis is appropriate for the development of regional standard deviation models.

Introduction

An important problem in hydrology is estimation of flood quantiles for ungauged locations, or sites with very short records. Regional generalized least squares (GLS) analyses are commonly used to estimate such statistics using physiographic characteristics of a catchment such as drainage

* Corresponding author. Tel.: +1 906 487 1079; fax: +1 906 487 2943.
E-mail addresses: vgriffis@mtu.edu (V.W. Griffis), jrs5@cornell.edu (J.R. Stedinger).
[1] Tel.: +1 607 255 2351; fax: +1 607 255 9004.

0022-1694/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhydrol.2007.06.023

area, main-channel slope, and land use and land cover indices (Tasker and Stedinger, 1989). Stedinger and Tasker (1985, 1986a,b) and Kroll and Stedinger (1998) show that GLS estimators are much more appropriate and efficient for use with hydrologic data than ordinary least squares (OLS) estimators. Unlike OLS estimators, the GLS estimators account for differences in the variance of streamflows from site-to-site due to different record lengths, and cross-correlation among the estimators due to cross-correlation among concurrent streamflows (Tasker, 1980; Kuczera, 1983). In this paper, the traditional GLS estimators are extended to correctly reflect uncertainty in a weighted skewness estimator. In addition, metrics are developed that correctly describe the value of GLS regression models, and an example demonstrates the benefit of pooled regression models.

The GLS procedure has been used extensively throughout the US and the world. Applications in hydrologic studies include the regionalization of flood quantiles, water quality parameters, and extreme rainfall (Tasker, 1978; Curtis, 1987; Tasker and Driver, 1988; Tasker and Stedinger, 1989; Landers and Wilson, 1991; Moss and Tasker, 1991; Ludwig and Tasker, 1993; Madsen et al., 1995; Madsen and Rosbjerg, 1997; Kroll and Stedinger, 1999; Pope et al., 2001; Feaster and Tasker, 2002; Kjeldsen and Rosbjerg, 2002; Reis et al., 2005). GLS has also been used as a regression method to regionalize flood quantiles using region-of-influence techniques (Tasker et al., 1996; Law and Tasker, 2003; Eng et al., 2005). And, GLS has been used as the basis of hydrologic network design (Medina, 1987; Tasker and Stedinger, 1989; Moss and Tasker, 1991; Soenksen et al., 1999). Reis et al. (2005) and Gruber et al. (2007) suggest a Bayesian analysis of the GLS model for hydrologic analyses. The GLS procedure proposed by Tasker and Stedinger (1989) and commonly used in the US for regionalization of flood quantiles is considered here.

The weighted least squares (WLS) procedure, which considers only differences in record length, has been used for regionalization of the standard deviation in US Geological Survey (USGS) studies as suggested by Stedinger and Tasker (1985, 1986b) and Tasker and Stedinger (1989). In addition to regionalization of flood quantiles, GLS is employed in this paper to develop a regional model of the standard deviation.

In the United States, flood quantiles at gauged sites are estimated as a function of the mean, standard deviation, and skew coefficient of the logarithms of the sample data using the guidelines contained in *Bulletin 17B* (IACWD, 1982). To improve the accuracy of the at-site sample skewness estimator, it is often combined with a regional skewness estimator to obtain a weighted skewness estimator (IACWD, 1982; Stedinger et al., 1993). Thus, the variance of the estimated quantile at a given site is a function of the sampling error in the sample mean, the standard deviation, and the weighted skewness estimator.

One problem with using the GLS procedure with hydrologic data is that the covariance matrix of the residual errors must be estimated from the data (Stedinger and Tasker, 1985). To do so, the variance of the residuals is separated into two components: (1) the model error variance, which is a measure of the precision with which the true model can predict flood quantiles, and (2) the sampling error in the flood quantile estimates. The GLS procedure proposed by Tasker and Stedinger (1989) first estimates the sampling error matrix using available streamflow data. An iterative procedure is then employed to estimate the regression coefficients and the model error variance. Equations are developed here to correctly include in the analysis the sampling error in a weighted skewness estimator for the LP3 distribution estimated with log space moments. Tasker and Stedinger (1989) employ an estimator of the sampling covariance matrix which only includes the sampling error in the mean and the standard deviation, and ignores the error in the estimated skew; this is the estimator currently employed by the US Geological Survey (see for example Feaster and Tasker, 2002; and Law and Tasker, 2003). This paper compares the precision of flood quantile estimators obtained with Tasker and Stedinger's (1989) estimator of the sampling error matrix to that of flood quantile estimators obtained using the new estimator of the sampling covariance matrix which correctly includes the sampling error in the weighted skewness estimator.

Section 2 of this paper develops the GLS regression framework and presents the new estimator of the sampling covariance matrix that includes the error in the weighted skew and its interaction with other moment estimators. Section 3 presents an application of GLS regression for South Carolina. This includes the development of a regional model of the standard deviation using GLS regression, and models for the 100-year event obtained using both estimators of the sampling covariance matrix. Model selection is based on a condition number which identifies problems with multicollinearity, and three metrics for model error: the model error variance, the average variance of prediction, and a new pseudo-$\bar{R}^2$ appropriate for use with GLS regression. Error variance ratios document the need for a full GLS regression analysis. In addition, this paper includes an evaluation of the benefit of pooling data from different regions to increase the number of sites used to develop regression models of the 100-year event, as opposed to simply developing a separate model for each region.

Model description and assumptions

Estimation of quantiles at gauged sites

Consider a region containing N gauging stations. At each gauged site i ($i = 1, \ldots, N$), a record of length n_i-years is available consisting of maximum annual flood peaks denoted $\{Q_{i1}, \ldots, Q_{in_i}\}$. Flows at a given site i are assumed to be temporally independent and identically distributed; however, concurrent observations at different sites ($i \neq j$) may be cross-correlated, resulting in correlation among estimators of different statistics (Stedinger and Tasker, 1985).

Flood quantiles at each site are estimated following the guidelines published in *Bulletin 17B* (B17) which suggests fitting a log-Pearson Type 3 (LP3) distribution to annual flood series. In particular, one uses the method of moments to fit a Pearson Type 3 (P3) distribution to the base-10 logarithms of the flood peaks, denoted $\{z_{i1}, \ldots, z_{in_i}\}$. Estimates of the mean $\bar{z}_i$, standard deviation s_i, and skew coefficient g_i of the logarithms of the sample data are computed using traditional moment estimators (Stedinger et al., 1993).

The record available at a site is often limited to 100 years, and is typically less than 30 years. These short records yield sample skews which are sensitive to extreme events. To improve the accuracy of the skewness estimator, B17 recommends combining the at-site sample skew g_i with a regional skew G_i to obtain a weighted skew $\widetilde{G}_i$ where:

$$\widetilde{G}_i = W_i g_i + (1 - W_i) G_i \quad \text{with } W_i = \mathrm{MSE}_G/(\mathrm{MSE}[g_i] + \mathrm{MSE}_G) \tag{1}$$

Here $\mathrm{MSE}[g_i]$ is the estimated mean square error (equal to the variance plus bias-squared) of the sample skew, and MSE_G is the mean square error of the regional skew. Estimates of the regional skew and its variance are obtained from a separate regional analysis such as that described by McCuen (1979), IACWD (1982), or Reis et al. (2005).

B17 recommends approximating $\mathrm{MSE}[g_i]$ as a function of the sample skew g_i and the record length n_i using the equation provided therein. That equation is based on the Monte Carlo study reported in Wallis et al. (1974), and yields relative errors as large as 10% within the hydrologic region of interest (Griffis, 2003). Griffis et al. (2004) provide a relatively more precise approximation which is also consistent with the asymptotic variance for g_i provided by Bobee (1973).

The pth quantile of the fitted P3 distribution is computed as $\hat{y}_i = \bar{z}_i + K_i s_i$, where K_i is the standard P3 frequency factor for the pth percentile given the weighted skew $\widetilde{G}_i$. Thus, $\hat{y}_i$ is an estimate of the log of the desired flood quantile, i.e. the 100-year peak flow (corresponding to $p = 0.99$):

$$\hat{y}_i = y_i + \eta_i \tag{2}$$

where y_i is the true value of the log of the 100-year event and η_i is a random error, referred to as the time-sampling error. Assuming that $\hat{y}_i$ is an unbiased estimate of y_i, then η_i has mean zero. Its variance is a function of the error in the estimated sample moments; Chowdury and Stedinger (1991) provide the following first-order approximation of the sampling error:

$$\begin{aligned}\mathrm{Var}[\hat{y}_i] = {} & \left[1 + K_i\gamma_i + K_i^2\left(\frac{1}{2} + \frac{3}{8}\gamma_i^2\right) + W_i K_i \frac{\partial K_i}{\partial g_i}\left(3\gamma_i + \frac{3}{4}\gamma_i^3\right)\right.\\ & \left. + W_i^2\left(\frac{\partial K_i}{\partial \gamma_i}\right)^2\left(6 + 9\gamma_i^2 + \frac{15}{8}\gamma_i^4\right)\right]\frac{s_i^2}{n_i}\\ & + (1 - W_i)^2 s_i^2 \mathrm{MSE}_{G_i}\left(\frac{\partial K_i}{\partial \gamma_i}\right)^2\end{aligned} \tag{3}$$

Chowdury and Stedinger (1991) show that this first-order approximation works fairly well for a number of cases despite the nonlinear relationship between $\hat{y}$ and γ; however, if the skewness estimator is poorly resolved by the sample and regional information, then these first-order estimates of the variance of LP3 quantiles can be inaccurate.

For a regional analysis, the estimates $\hat{y}_i$ of the log of the 100-year event at each site are combined into a $(N \times 1)$ vector $\widehat{\mathbf{Y}}$. It is assumed that $\widehat{\mathbf{Y}}$ is an unbiased estimator of $\mathbf{Y} = \{Y_1, \ldots, Y_N\}^T$. The sampling covariance matrix of $\widehat{\mathbf{Y}}$ is then given by $\boldsymbol{\Sigma} = E[(\widehat{\mathbf{Y}} - \mathbf{Y})(\widehat{\mathbf{Y}} - \mathbf{Y})^T]$. Due to the cross-correlation among concurrent flows, the quantile estimates $\hat{y}_i$ and $\hat{y}_j (i \neq j)$ are correlated, and thus the off-diagonal elements of $\boldsymbol{\Sigma}$ are nonzero.

Tasker and Stedinger (1989) provide the following estimator of the components of Σ which neglects the possible error in the estimated skew:

$$\begin{aligned}\Sigma_{ii} &= \left[1 + K_i\gamma_i + \frac{1}{2}K_i^2(1 + 0.75\gamma_i^2)\right]\frac{\sigma_i^2}{n_i} \quad \text{for } i = j\\ \Sigma_{ij} &= \left[1 + \frac{1}{2}K_i\gamma_i + \frac{1}{2}K_j\gamma_j + \frac{1}{2}K_iK_j(\rho_{ij} + 0.75\gamma_i\gamma_j)\right]\rho_{ij}\frac{m_{ij}\sigma_i\sigma_j}{n_i n_j}\\ &\quad \text{for } i \neq j\end{aligned} \tag{4}$$

where m_{ij} is the concurrent record length between sites i and j, ρ_{ij} is the lag zero cross-correlation of flows between sites i and j, and σ_i and σ_j are the population standard deviations at sites i and j, respectively. To avoid correlation between the residuals and the fitted quantiles, Tasker and Stedinger recommend that (i) ρ_{ij} is estimated as a function of the distance between sites i and j, (ii) the standard deviations σ_i and σ_j are estimated using a separate regional regression on drainage area, and (iii) the regional skew G_i is used in place of the population skew γ_i. Kroll and Stedinger (1998) demonstrate for a number of cases that this smoothing results in negligible loss of efficiency.

Because the sample skew is an estimate with an associated estimation error, the approximation in Eq. (4) should be extended to reflect all of the sampling error in the quantile estimates. The needed estimator of the sample covariance matrix is:

$$\begin{aligned}\Sigma_{ii} &= \mathrm{Var}[\hat{y}_i] \quad \text{for } i = j\\ \Sigma_{ij} &= \left[1 + \frac{1}{2}K_i\gamma_i + \frac{1}{2}K_j\gamma_j + \frac{1}{2}K_iK_j(\rho_{ij} + 0.75\gamma_i\gamma_j)\right.\\ &\quad + \frac{1}{2}W_iK_j\gamma_i\frac{\partial K}{\partial \gamma_i}(3\rho_{ij} + 0.75\gamma_i\gamma_j) + \frac{1}{2}W_jK_i\gamma_j\frac{\partial K}{\partial \gamma_j}(3\rho_{ij} + 0.75\gamma_i\gamma_j)\\ &\quad \left. + W_iW_j\sigma_i\sigma_j\frac{\partial K}{\partial \gamma_i}\frac{\partial K}{\partial \gamma_j}\mathrm{Cov}[g_i, g_j]\right]\rho_{ij}\frac{m_{ij}\sigma_i\sigma_j}{n_i n_j} \quad \text{for } i \neq j\end{aligned} \tag{5}$$

where $\mathrm{Var}[\hat{y}_i]$ is as specified in Eq. (3). The parameters ρ_{ij}, σ_i, and γ_i are estimated as specified in the preceding paragraph to avoid correlation between the residuals and the fitted quantiles. The partial derivative $\partial K/\partial \gamma_i$ is computed using the approximation provided by Chowdury and Stedinger (1991), and the weight W_i is computed using $\mathrm{MSE}[g_i]$ estimated using the approximation provided by Griffis et al. (2004) wherein the regional skew is used to estimate the population skew γ. Griffis (2006) considers different estimators of γ when computing weights for use in Eq. (1), and ultimately recommends use of the regional skew.

To include skew uncertainty in Eq. (5), the needed covariance terms such as $\mathrm{Cov}[\bar{x}_i, g_j]$ and $\mathrm{Cov}[s_i, g_j]$ were estimated assuming the relationship between concurrent observations y_i and y_j at sites i and j could be modeled by a multivariate gamma distribution for which the two correlated variables are generated by the sums

$$\begin{aligned}y_i &= z_1 + z_2\\ y_j &= z_1 + z_3\end{aligned} \tag{6}$$

wherein z_1, z_2, and z_3 are independent standard gamma random variables. Here z_1 is common to both y_i and y_j, thereby introducing a common signal, whereas z_2 and z_3 are

independent of one another causing y_i and y_j to be less than perfectly correlated.

The off-diagonal elements of the sampling covariance matrix estimator in Eq. (5) also include the term Cov[g_i, g_j] which is the covariance between the two at-site skew estimators g_i and g_j. This term is obtained from

$$\mathrm{Cov}[g_i, g_j] = \rho_{g_i g_j}\sqrt{\mathrm{Var}[g_i]\mathrm{Var}[g_j]} \tag{7}$$

where the cross-correlation $\rho_{g_i g_j}$ is estimated using the approximation developed by Martins and Stedinger (2002):

$$\hat{\rho}_{g_i g_j} = \mathrm{Sign}(\hat{\rho}_{ij})cf_{ij}|\hat{\rho}_{ij}|^{v} \tag{8}$$

wherein $cf_{ij} = m_{ij}/\sqrt{(m_{ij}+n_i)(m_{ij}+n_j)}$, and values of v are tabulated by Martins and Stedinger (2002) for $|\gamma| \leqslant 1.0$. In addition, Var[g_i] and Var[g_j] are evaluated using the following approximation derived by Griffis (2003):

$$\mathrm{Var}[g_i] = \left[\frac{6}{n_i} + a(n_i)\right]\left[1 + \left(\frac{9}{6} + b(n_i)\right)\gamma_i^2 + \left(\frac{15}{48} + c(n_i)\right)\gamma_i^4\right] \tag{9}$$

wherein $a(n_i)$, $b(n_i)$, and $c(n_i)$ are corrections for small samples:

$$a(n_i) = -\frac{17.75}{n_i^2} + \frac{50.06}{n_i^3};$$
$$b(n_i) = \frac{3.92}{n_i^{0.3}} - \frac{31.1}{n_i^{0.6}} + \frac{34.86}{n_i^{0.9}};$$
$$c(n_i) = -\frac{7.31}{n_i^{0.59}} + \frac{45.9}{n_i^{1.18}} - \frac{86.5}{n_i^{1.77}}$$

The regional skew G_i is used in Eq. (9) in place of the population skew γ_i to avoid correlation between the residuals and the fitted quantiles. The approximation for MSE[g] in Griffis et al. (2004) has the same form as the approximation for Var[g] above. The approximations are asymptotically equivalent, and are identical for $\gamma = 0$ because for that case the skewness estimator is unbiased.

Griffis (2006) shows that for reasonable parameter values the variance of $\hat{y}_i$ in Eq. (4) underestimates the value of the diagonal elements of Σ. This error increases with an increase in either n or MSE_G, and can be substantial when the error in the regional skew is on the order of 0.3 as suggested by *Bulletin 17B*. The relative difference between Eqs. (4) and (5) is independent of G. Evaluation of the error in the off-diagonal elements of Eq. (4) is more complicated because of the terms involving the cross-correlation ρ_{ij} and the concurrent record length m_{ij}. However, results in Griffis (2006) suggest that Eq. (4) overestimates the off-diagonal elements of Σ for negative values of regional skew, whereas the elements are overestimated for positive regional skews, and the magnitude of the error increases with the magnitude of m_{ij}, ρ_{ij}, G, and MSE_G. Section 3 of this paper includes an evaluation of the impact of using Eq. (4) instead of Eq. (5) on a GLS regression analysis and model selection using data for South Carolina.

The analysis presented in Section 3 employs LP3 quantile estimates obtained for each site following B17 guidelines. While other distributions and parameter estimation methods could be employed, Griffis and Stedinger (2007a,b) demonstrate that the LP3 distribution provides a reasonable model of annual maximum flood series in the United States, and that log space method of moments with regional skew information is an efficient LP3 parameter estimation technique relative to alternative methods such as maximum likelihood estimation and real space method of moments. Overall, the results of Griffis and Stedinger (2007a,b) indicate that the quantile estimators employed herein are reasonable. Nonetheless, the procedures and statistics developed in the following sections can easily be applied to other distributions; one would only need an appropriate estimator of Σ to describe the variance and covariance of the quantile estimates obtained using the chosen distribution.

GLS regression model

The goal of a generalized least squares (GLS) regression analysis is to identify the best model one can for estimating flood quantiles, such as the 100-year peak flow, at an ungauged site given a set of k basin characteristics. These basin characteristics are assumed to be measured with negligible error. If y_i can be expressed as a linear function of the logs of the basin characteristics (x's) and the model error δ_i, then one has the model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \delta_i \tag{10}$$

The errors δ_i are assumed to be normal and independently distributed with mean zero and a variance of σ_δ^2. Here σ_δ^2 is the model error variance, or the residual variance not explained by sampling error.

Combining Eqs. (2) and (10) yields

$$\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \delta_i + \eta_i \tag{11}$$

Thus, when $y_i = \log(Q_i)$, the log–log model in Eq. (11) corresponds to the real space model:

$$\widehat{Q}_i = b_0 R_{i1}^{b_1}, R_{i2}^{b_2}, \ldots, R_{ik}^{b_k} 10^{\delta_i + \eta_i} \tag{12}$$

wherein b_j is an estimate of the jth beta coefficient ($j = 0, \ldots, k$) and R_{ij} are the observed values of the k basin characteristics ($R_{ij} = 10^{x_{ij}}$ for $j = 1, \ldots, k$).

Eq. (11) in matrix notation can be written as

$$\widehat{\mathbf{Y}} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{13}$$

where $\mathbf{X}$ is a $[N \times (k+1)]$ matrix of k basin characteristics augmented by a column of one's, $\boldsymbol{\beta}$ is a $[(k+1) \times 1]$ vector of regression parameters, and $\boldsymbol{\varepsilon} = \boldsymbol{\delta} + \boldsymbol{\eta}$ is a $(N \times 1)$ vector of random errors, for which $E[\varepsilon] = 0$ and $E[\varepsilon\varepsilon^{\mathrm{T}}] = \boldsymbol{\Lambda}$.

Due to the correlation between the residuals, the traditional OLS analysis is not appropriate, and a GLS analysis should be used to relate the fitted quantiles to the specified basin characteristics and to describe the errors. The GLS estimator of $\boldsymbol{\beta}$ is

$$\mathbf{b} = (\mathbf{X}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\widehat{\mathbf{Y}} \tag{14}$$

This estimator would be the best linear unbiased estimator of $\boldsymbol{\beta}$ if $\boldsymbol{\Lambda}$ were known (Johnston, 1984). However, $\boldsymbol{\Lambda}$ is not known, but can be estimated by

$$\widehat{\boldsymbol{\Lambda}}(\sigma_\delta^2) = \sigma_\delta^2 \mathbf{I}_N + \widehat{\boldsymbol{\Sigma}} \tag{15}$$

wherein $\mathbf{I}_N$ is an $(N \times N)$ identity matrix, and $\widehat{\boldsymbol{\Sigma}}$ is computed using either Eq. (4) or (5).

The model error variance σ_δ^2 and the vector of regression coefficients **b** are estimated jointly by iteratively searching for a nonnegative solution to the equation (Stedinger and Tasker, 1985):

$$(\widehat{\mathbf{Y}} - \mathbf{Xb})^{\mathrm{T}}[\hat{\sigma}_\delta^2 \mathbf{I}_N + \widehat{\boldsymbol{\Sigma}}]^{-1}(\widehat{\mathbf{Y}} - \mathbf{Xb}) = N - (k+1) \tag{16}$$

where $\hat{\sigma}_\delta^2$ is the estimator of the unknown model error variance and **b** is given by Eq. (14).

Measures of model and prediction error

The purpose of the regression model is to estimate flood quantiles at ungauged sites. Therefore, given a site with basin characteristics $\mathbf{x}_0$, a concern is how well the GLS regression model predicts the true quantile, y_0 (Tasker et al., 1986). Under the assumption that the observed data were collected at representative sites at which predictions will be made, the average variance of prediction (AVP) over the available dataset is a measure of how well the GLS regression model predicts the true quantile on average (Tasker and Stedinger, 1986), where:

$$\mathrm{AVP}_{\mathrm{GLS}} = \hat{\sigma}_\delta^2 + \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i(\mathbf{X}^{\mathrm{T}}\widehat{\boldsymbol{\Lambda}}^{-1}\mathbf{X})^{-1}\mathbf{x}_i^{\mathrm{T}} \tag{17}$$

Here the x_i's are row vectors containing the physiographic characteristics of each site. When comparing hydrologic regression models a smaller AVP is preferred. (See studies cited in Reis et al., 2005.) Additionally, if the residuals have nearly a normal distribution, then the standard error of prediction in percent (SEP%) for the true flood quantile estimator (rather than its common logarithm) is described by

$$\mathrm{SEP}\% = 100\sqrt{10^{\ln(10)\mathrm{AVP}_{\mathrm{GLS}}} - 1} \tag{18}$$

In order to assess the precision of a model, the AVP and the model error variance are preferred over more common metrics such as the traditional R^2 and adjusted-R^2 ($\bar{R}^2$), which misrepresent the true power of the model. These R^2 metrics measure the proportion of variance in the observed $\hat{y}_i$ values explained by the fitted model. Unfortunately, that proportion considers the total error ε, which includes the sampling error η. However, our interest is actually to quantify the proportion of the variance among the unobserved y_i explained by the model. Let $\hat{\sigma}_\delta^2(k)$ be the estimated model error variance for the regression model with k explanatory variables, and $\hat{\sigma}_\delta^2(0)$ be the estimated model error variance when no explanatory variables are employed. Then a pseudo-$\bar{R}^2$ appropriate for use with GLS regression is

$$R^2_{\mathrm{GLS}} = 1 - \frac{\hat{\sigma}_\delta^2(k)}{\hat{\sigma}_\delta^2(0)} \tag{19}$$

Both our pseudo-$\bar{R}^2$ and the traditional adjusted-R^2 correct for the degrees-of-freedom lost when k parameters are estimated.

An important question that should be addressed is whether a full GLS regression is needed, or if WLS, or even OLS would be sufficient. Unlike OLS regression which includes a single error term, WLS and GLS regression analyses divide the total error into two parts: sampling error and model error. Therefore, OLS regression should be sufficient if the sampling variance is negligible when compared to the model error variance. The relative magnitude of the average sampling variance to the model error variance, and thus the necessity of a WLS or GLS regression analysis, can be described using the error variance ratio (EVR) computed as

$$\mathrm{EVR} = \frac{\mathrm{tr}(\widehat{\boldsymbol{\Sigma}})}{N\hat{\sigma}_\delta^2(k)} \tag{20}$$

If the EVR is greater than 20%, then a WLS or GLS regression analysis is recommended instead of OLS.

Similarly, the Misrepresentation of the Beta Variance (MBV) statistic can be used to determine whether a full GLS regression is necessary, or if WLS regression is sufficient. This is accomplished by measuring the impact the correlation among the y_i values has on the regression analysis. A statistic that is particularly sensitive to cross-correlation is the variance of the estimator of the constant term in the model, as illustrated by the results in Table 1 of Stedinger and Tasker (1985). For example, for a population correlation of +0.3 and the true model error variance of 0.011, their results indicate that WLS regression misrepresents the variance of the constant by a factor of 2.3; the variance of the slope coefficient is misrepresented to a lesser degree by a factor of 1.6. Thus to illustrate the impact of cross-correlation for a particular problem we consider the variance a GLS and a WLS analysis would ascribe to the WLS estimator b_0 of the constant in the model β_0. Let w be a $(N \times 1)$ vector with components

$$w_i = \frac{1}{\sqrt{\Lambda_{ii}}} \tag{21}$$

corresponding to the inverse of the square root of the diagonal elements of Λ. Then the WLS estimator of the average of the y_i 's, computed as $w^{\mathrm{T}}y/w^{\mathrm{T}}v$ where v is a $(1 \times N)$ vector of ones, has sampling variance $w^{\mathrm{T}}[\mathrm{diag}(\Lambda_{ii})]\ w/(w^{\mathrm{T}}v)^2 = N/(w^{\mathrm{T}}v)^2$; the GLS analysis would estimate the variance of the WLS estimator as $w^{\mathrm{T}}\Lambda w/(w^{\mathrm{T}}v)^2$. Thus the MBV can be computed as

$$\mathrm{MBV} = \frac{\mathrm{Var}[b_0^{\mathrm{WLS}}|\text{GLS analysis}]}{\mathrm{Var}[b_0^{\mathrm{WLS}}|\text{WLS analysis}]} = \frac{w^{\mathrm{T}}\Lambda w}{N} \tag{22}$$

This ratio provides a direct measure of the distortion in the estimated variance of the constant that results when a WLS analysis is employed rather than a GLS analysis. If the MBV statistic is notably greater than 1, then WLS regression is insufficient and GLS regression should be employed. Here the appropriate threshold depends upon the issues of concern. If an analysis addresses the precision of the estimated constant, and in particular if one is considering indicator variables to represent differences among regions, then a relatively precise description of the variance of such constants is needed. To ensure less than a 10% error in the estimated standard errors, GLS should be employed when the MBV is greater than 1.2. Errors in the computed standard errors for other beta estimators are also a concern if different explanatory variables are considered.

Alternative definitions of the MBV statistic could be employed. Griffis (2006) considers four possible definitions of the MBV ratio (including the definition above) reflecting different ways the impact of the correlation among the y_i values on the regression analysis might be measured. Three of the MBV ratios (including Eq. (22)) measure the distortion of

Table 1 Pseudo ANOVA table for GLS regression analyses

Source	Degrees-of-freedom	Sum of squares
Regression model	k	$N[\sigma_\delta^2(0) - \sigma_\delta^2(k)]$
Model error δ	$N-k-1$	$N\sigma_\delta^2(k)$
Sampling error η	N	$\mathrm{tr}(\widehat{\boldsymbol{\Sigma}})$
Total error	$2N-1$	$N\sigma_\delta^2(0) + \mathrm{tr}(\widehat{\boldsymbol{\Sigma}})$

the computed variance of the OLS, WLS and GLS estimators b_0 of the constant β_0 when the residuals are cross-correlated. Of these three definitions, the MBV in Eq. (22) is preferred because it directly addresses the distortion in the variance of the constant estimator from a WLS analysis if the residuals were really cross-correlated, and thus a GLS analysis were appropriate.

The fourth definition considered measures the loss of efficiency from using the WLS estimator of the constant rather than the GLS estimator by comparing the variances of the two estimators. This loss of efficiency could be substantial, however, the more sensitive statistic appears to be the errors made in the estimated precision of the β_0 parameter represented by the MBV statistic. Such an error can result in a hydrologist adding variables to a model that are not statistically justified because their inclusion actually increases the average variance of prediction.

Following Reis (2005), Table 1 presents a proposed pseudo Analysis of Variance (ANOVA) table for a GLS analysis. It describes how the total variation among the $\hat{y}_i$ values can be partitioned among that explained by the model and the expected residual variation due to model error and sampling error. This is an extension of the ANOVA table for OLS regression to correctly separate the expected sampling and model error variances. In a traditional OLS regression, the observed total sum of squares (SST) is simply divided into the regression sum of squares (SSR) and the error sum of squares (SSE), where SST = SSR + SSE. Table 1 partitions the expected variation due to sampling error and model error, rather than the actual error because the actual errors are not observed.

Application of GLS regression in South Carolina

Data for N = 162 sites in South Carolina and adjacent states, compiled by Feaster and Tasker (2002), was used in a GLS regression analysis to develop models of the 100-year event. Models were obtained using the sampling error matrix in Eq. (4) and the more appropriate estimator in Eq. (5) which accounts for error in the skew. Model development, comparison, and selection were based on several statistical innovations. The USGS data include moments of the fitted P3 distribution at each site and the resulting estimate of the 99th quantile, the latitude and longitude of each site, estimates of the regional skew and its precision, estimates of the lag-zero correlation of concurrent flows between two sites, and the at-site and concurrent record lengths. For each site, Feaster and Tasker (2002) also provide data on eight basin characteristics: (i) drainage area measured in square-miles; (ii) main-channel slope measured in feet per mile; (iii) main-channel length measured in miles; (iv) elevation reported as the mean basin elevation above sea-level; (v) forest cover measured as a percent of drainage area; (vi) storage measured as the percent of drainage area consisting of lakes, ponds, and swamps; (vii) precipitation measured in inches at the centroid of the basin; and, (viii) runoff measured in inches at the centroid of the basin. A concern addressed in the following subsections is that data for all eight characteristics was not available for 28 of the 162 sites.

The study area was divided into three major physiographic regions: the Blue Ridge, the Piedmont, and the Coastal Plain. The Coastal Plain is further divided into the upper and lower Coastal Plain. The Blue Ridge (region 1) is a mountainous region with steep terrain. The Piedmont (region 2) has rolling hills and valleys with its highest elevations located in the Blue Ridge foothills. Both the Blue Ridge and the Piedmont are rocky regions with poorly permeable soil. The upper Coastal Plain (region 3) has more gradual slopes with extensive swamps and wide floodplains. The lower Coastal Plain (region 4) is also swampy, with highly permeable terrain which absorbs rainfall and reduces runoff. Table 2 reports the number of sites contained in each region, as well as the minimum, maximum, and average record length (Feaster and Tasker, 2002). There is one region containing several sites and three regions with a modest number of sites.

Because of the similarity of the two coastal regions, quantile models for regions 3 and 4 are likely to be very similar. Likewise, quantile models should be similar in the more mountainous terrain of the Blue Ridge and the Piedmont. Conversely, estimates in the highly permeable Coastal Plain might differ from estimates for the poorly permeable terrain of the Blue Ridge and the Piedmont. To test these hypotheses, three indicator variables were introduced:

d_1 = 1, if station is in the Blue Ridge; 0, otherwise; d_2 = 1, if station is in the Piedmont; 0, otherwise; d_4 = 1, if station is in the lower Coastal Plain; 0, otherwise.

Table 2 Number of sites per region and record length statistics

Region	Number of sites (N)	Minimum record length	Maximum record length	Average record length
Blue Ridge	30	11	82	40
Piedmont	84	12	100	34
Upper Coastal Plain	22	11	77	34
Lower Coastal Plain	26	12	72	34
Total Study Area	162	11	100	35

Indicator variables were not employed by Feaster and Tasker (2002) who simply considered each of the four physiographic regions separately.

A critical issue here is the precision of the regional skewness estimator. Feaster and Tasker (2002) conducted a regional analysis for the skew using the arithmetic mean procedure recommended by B17. The skews in the Piedmont were determined to be significantly different from those in the Blue Ridge and Coastal Plain regions, but the differences in the latter three regions were insignificant at the 5% level. Therefore, Feaster and Tasker conducted a separate regional analysis for the Piedmont and obtained a mean regional skew of −0.190 with a variance of 0.090; a mean regional skew of 0.082 with a variance of 0.105 was obtained using the combined data for the Blue Ridge and Coastal Plain regions. These variances correspond to effective record lengths on the order of 60 years (see Griffis et al., 2004), but they were simply computed as the variance of the station skew estimates about the regional average, and thus may overestimate the actual error of the regional skew. Using GLS regression with the same data, Reis et al. (2004) obtained a model error variance of 0.025, which has an effective record length over 200 years. The following section describes a separate regional GLS analysis for the standard deviation.

Regression model for standard deviation

To avoid correlation between the residuals and the fitted quantiles, an estimator of the standard deviation σ_i other than the at-site value s_i is required to compute Σ. Based on their simulation studies, Stedinger and Tasker (1985) and Kroll and Stedinger (1998) observe that the GLS estimators are not very sensitive to the estimator of σ_i, and thus $\hat{\sigma}_i$ need not be highly accurate, though unbiased estimates are desired. The important requirement is that the estimator be independent of the model's residuals and the time-sampling error. Stedinger and Tasker (1985) were concerned that it should also be consistent with the model of y_i in Eq. (10). Such an estimator is given by the model:

$$\sigma_i = \alpha + \beta \log_{10}(A_i) + \varepsilon_i \tag{23}$$

wherein A_i is the drainage area of the basin containing site i, and the error ε_i is lognormally distributed with zero mean and variance $\sigma_\varepsilon^2 = \sigma^2(A)[10^{\ln(10)\sigma_\delta^2} - 1]$. Additional basin characteristics could be used; however, several studies (for example, Benson, 1962; Cruff and Rantz, 1965; Thomas and Benson, 1970) indicate that the drainage area is the most significant explanatory variable and is sufficient for the estimator's purpose here.

The pooled regression model in Eq. (23) assumes a common intercept and slope for all four regions. To test for differences between the intercepts of each region, a fixed effects model using indicator variables was considered:

$$\sigma_i = \alpha_0 + \alpha_1 d_1 + \alpha_2 d_2 + \alpha_4 d_4 + \beta \log_{10}(A_i) + \varepsilon_i \tag{24}$$

With this model, values of α_1, α_2, and α_4 reflect the difference between the intercepts in regions 1, 2, and 4, respectively, relative to the common intercept α_0 that then corresponds to region 3.

To estimate the coefficients of the standard deviation model, Stedinger and Tasker (1985, 1986b) suggest a WLS regression. However, if there is substantial cross-correlation between sites, then a GLS model of the standard deviation would be more appropriate. To estimate the coefficients of the standard deviation model using GLS regression, one needs the sampling covariance matrix $\Sigma(s)$ for the standard deviations. Stedinger and Tasker (1986b) provide an expression for $\Sigma(s)$ for normal variates. From Eq. (4), a formula for P3 variates is

$$\Sigma(s)_{ii} = \mathrm{Var}(s) = \frac{1}{2}(1 + 0.75\gamma_i^2)\frac{\sigma_i^2}{n_i} \quad \text{for } i = j$$
$$\Sigma(s)_{ij} = \mathrm{Cov}[s_i, s_j] = \frac{1}{2}(\rho_{ij} + 0.75\gamma_i\gamma_j)\rho_{ij}\frac{m_{ij}\sigma_i\sigma_j}{n_i n_j} \quad \text{for } i \neq j \tag{25}$$

To avoid correlation between the residuals and the estimated standard deviations, (i) ρ_{ij} is estimated as a function of the distance between sites i and j, (ii) the regional skew G_i is used in place of the population skew γ_i, and (iii) the standard deviations σ_i and σ_j are estimated using a separate OLS regression on drainage area. Estimates of α and β are obtained by iteratively solving Eqs. (14) and (16) wherein $\boldsymbol{\Lambda}$ is estimated using

$$\hat{\Lambda}_{ii} = E[\sigma_i]^2\left[10^{\ln(10)\sigma_\delta^2} - 1\right] + \hat{\Sigma}(s)_{ii} \quad \text{and} \quad \hat{\Lambda}_{ij} = \hat{\Sigma}(s)_{ij} \tag{26}$$

Coefficients for the pooled regression model in Eq. (23) and the fixed effects model in Eq. (24) were obtained using GLS regression. Griffis (2006) reports the estimated coefficients, and the t-values and their p-values for a two-sided hypothesis test of whether or not each coefficient is significantly different from 0. The p-values indicate that α_1 and α_2 in the fixed effects model are not statistically significant at the 5% level relative to $\alpha_3 = 0$, but α_4 is highly significant with a p-value of 0.01%. These results suggest that regions 1, 2, and 3 are similar to one another, but are different from region 4 corresponding to the lower Coastal Plain. Thus, the recommended model for the standard deviation is

$$\sigma_i = \alpha_0 + \alpha_4 d_4 + \beta \log_{10}(A_i) + \varepsilon_i \tag{27}$$

Table 3 reports summary statistics for the models including the average (over A) of the estimated model error variances $\hat{\sigma}_\varepsilon^2$, the AVP, the traditional adjusted-R^2 ($\overline{R}^2$), the pseudo-$\overline{R}^2$

Table 3 Summary statistics for GLS regression models for standard deviation

Model	$\hat{\sigma}_\varepsilon^2$	AVP	$\overline{R}^2$	R^2_{GLS}	EVR	MBV
Pooled model (Eq. (23))	0.0020	0.00211	0.006	0.154	0.666	4.630
Fixed effects model with $\alpha_3 = 0$ (Eq. (24))	0.0017	0.00185	0.111	0.300	0.806	5.081
Fixed effects model with only α_4 (Eq. (27))	0.0016	0.00178	0.109	0.306	0.815	5.101
Interaction terms (Eq. (29))	0.0017	0.00187	0.111	0.291	0.801	5.052

(denoted R^2_{GLS}), and the EVR and MBV. The EVR is about 70 to 80% indicating that the average sample error variance almost equals the model error variance. Thus, a WLS or GLS analysis is called for. The values of the MBV exceed 4, which is appreciably greater than 1 indicating that a WLS analysis for the standard deviation could be very misleading.

Of the three models, the fixed effects model in Eq. (27) has the smallest model error variance and AVP, and the largest pseudo-$\overline{R}^2$. It is interesting that the model in Eq. (24) actually has a larger traditional $\overline{R}^2$ value than the model in Eq. (27), even though the model in Eq. (27) yields a slightly smaller AVP. This illustrates how the traditional $\overline{R}^2$ can misrepresent the true power of the GLS model.

Fig. 1 plots the residuals of the pooled regression model by region wherein each site was assigned an index number between 1 and $N = 162$. The pooled regression model yields residuals with a mean greater than zero in region 4, whereas the mean of the residuals is negative in the other three regions. Use of the additional intercept term in the fixed effects model appropriately decreases the mean of the residuals in region 4, and increases the mean for the other three regions.

To further test if the restrictions imposed by this model are justified, the following test statistic may be used:

$$\tilde{\chi}_{[J]} = [\hat{b}_{\text{unres}} - \hat{b}_{\text{res}}]^{\text{T}} \{\text{Var}[\hat{b}_{\text{unres}}|X]\}^{-1} [\hat{b}_{\text{unres}} - \hat{b}_{\text{res}}] \quad (28)$$

where J is the number of coefficients estimated, $\hat{b}_{\text{unres}}$ is the vector of coefficients estimated for the model in Eq. (24), and $\hat{b}_{\text{res}}$ is the vector of coefficients for Eq. (27). This test statistic is approximately chi-squared distributed with J degrees-of-freedom (Greene, 2003, p. 347). The test statistic $\tilde{\chi}_{[5]}$ has a value of 1.11, which is much less than the critical value of 11.07 for a one-sided 5% test. This test, Fig. 1, and the p-values for a two-sided hypothesis test for individual coefficients indicate that use of a common intercept term is appropriate for regions 1, 2, and 3, but an adjustment is statistically justified for region 4.

In addition to the differences in the intercepts for the four regions, there may also be differences in the slopes. Interactions between region and drainage area were employed to test for possible differences in the slopes using the model

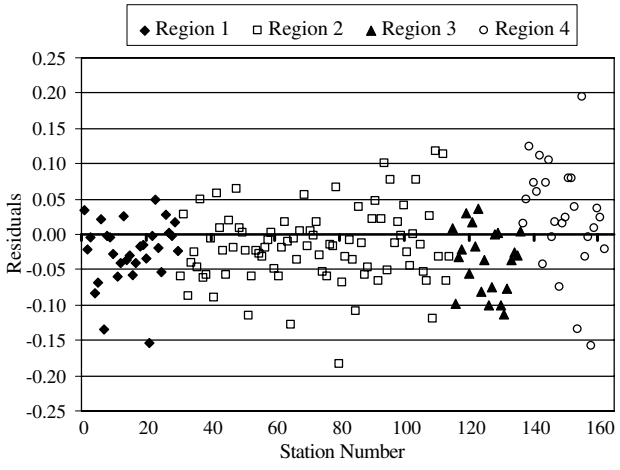

Figure 1 Residuals plotted by region for the pooled regression model (Eq. (23)).

$$\sigma_i = \alpha_0 + \alpha_4 d_4 + (\beta_0 + \beta_1 d_1 + \beta_2 d_2 + \beta_4 d_4) \times \left[\log_{10}(A_i) - \frac{1}{N}\sum_{j=1}^{N} \log_{10}(A_j)\right] + \varepsilon_i \quad (29)$$

Here the data are centered to allow for an adjustment of the intercept in region 4 while also changing the slope. The p-values for a two-sided hypothesis test indicate that only the additional intercept term for region 4 (α_4) is significant at the 5% level (p-value = 0.05%); none of the interaction terms are significant at the 5% level relative to $\beta_3 = 0$. Further, the common slope term (β_0) is just barely not significant at the 5% level when all of the interaction terms are included because the standard error of β_0 doubled. This clearly illustrates the value of pooling the data when estimating the relationship between the standard deviation and drainage area. Overall, these results indicate that the fixed effects model in Eq. (27) is an appropriate model of the sample standard deviation in rural South Carolina.

GLS regression model for the 100-year peak flow

A goal of this analysis is to use GLS regression to identify and estimate the parameters of a model to estimate the 100-year peak flow at ungauged basins in South Carolina. A preliminary analysis indicated that the logarithms of the 99th quantile estimates were linearly related to the logs of several explanatory variables: drainage area (A), main-channel slope (S), main-channel length (L), basin storage (St), and precipitation (P). Distinct relationships with the logs of the elevation (E), forest coverage (F), and runoff (RO) could not be found; thus, these variables are unlikely to contribute to the estimation of the 99th percentile. The following analyses consider the model:

$$y_i = b_o + b_1 \log(A_i) + b_2 \log(S_i) + b_3 \log(L_i) + b_4 \log(E_i) + b_5 \log(F_i) + b_6 \log(St_i) + b_7 \log(P_i) + b_8 \log(RO_i) + e_i \quad (30)$$

where y_i is the base-10 logarithm of the 99th percentile at site i and e_i is the estimation error. Cross-products of the explanatory variables were investigated by Feaster and Tasker (2002) in an attempt to try to improve the estimator $\hat{y}_i$. However, it is not clear that this is the correct approach as the general understanding of hydrology is not advanced enough to know whether or not this model specification is reasonable. Furthermore, by using a log–log transformation, such a multiplicative relationship between the variables should be captured in the resulting real space quantile estimator. Therefore, cross-products of explanatory variables were not considered in this study.

The coefficients of the model in Eq. (30) were estimated using a GLS analysis. The regional skew information and other data provided by Feaster and Tasker (2002) were used in conjunction with regional estimators of the sample standard deviation obtained using Eq. (27). The sampling error matrix Σ can be computed using Eq. (4), or more appropriately using the new relationship in Eq. (5) that accounts for the error in the skew. The impact on the GLS analyses due to correctly including the skew error in Σ is investigated below.

GLS regression model using new estimator of Σ with $MSE_G \approx 0.100$

Using the new estimator of Σ in Eq. (5), various forms of the model in Eq. (30) were fit using GLS regression. In addition to the constant term shown in Eq. (30), indicator variables were again employed to test for differences in the four regions. For the purpose of model comparison, models were fit using only the 134 sites out of the total 162 sites for which information was available for all eight explanatory variables. Griffis (2006) reports the estimated coefficients and *p*-values for a two-sided hypothesis test. Table 4 reports summary statistics for each model. The EVR and MBV demonstrate that the full GLS model is needed. The condition number (CN) is reported to provide a check for problems with multicollinearity, where

$$CN = \sqrt{\frac{\text{Max(eigenvalue of } (X^T X))}{\text{Min(eigenvalue of } (X^T X))}} \quad (31)$$

Generally a CN greater than 20 indicates a problem with multicollinearity (Greene, 2003, p. 58).

Model 1a. Model 1a employed all eight explanatory variables and a single constant term. This is the pooled regression model which assumes the intercept is the same for all four regions. Four of the variables [log(S), log(L), log(F), and log(St)], are not statistically significant at the 5% level, although the coefficient for log(St) is significant at the 6% level. Also, CN = 297 $\gg$ 20 which suggests serious multicollinearity; the residuals are plotted by region in Fig. 2 to investigate possible causes.

Fig. 2 indicates that over the 134 sites there are problems with the model specification because the residuals in regions 1 and 3 tend to be more negative, while the residuals in region 2 tend to be more positive; the residuals in region 4 appear reasonably centered around a mean of zero. This suggests there are differences between the four regions that are not captured in the pooled model. Therefore, in addition to the eight explanatory variables and a common intercept term, indicator variables were added for regions 1, 2, and 4 resulting in a fixed effects model denoted as Model 1.

Model 1. Model 1 employed all eight explanatory variables and three indicator variables in addition to a common intercept term. Only the variable log(A) is statistically significant at the 5% level. The variables log(S), log(L), log(E), log(F), and log(St) have *p*-values ranging from 16% to 60%, whereas log(P) and log(RO) have *p*-values greater than 90%. Because the *p*-values of log(P) and log(RO) are so large relative to those of the other variables, only these two variables were removed from the model to see if the significance of the other variables would be improved. The resulting model is denoted Model 2.

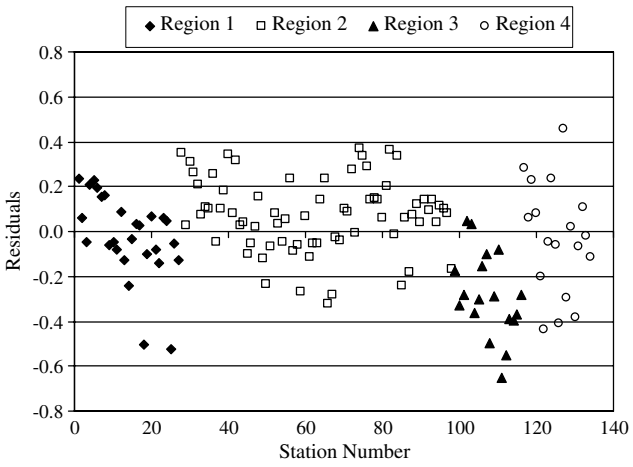

Figure 2 Plot of residuals by location for GLS regression model 1a using the new estimator of Σ.

Model 2. Model 2 employed three indicator variables in addition to a common intercept, and six of the eight explanatory variables, excluding log(P) and log(RO). Again, log(A) is the only significant explanatory variable at the 5% level (other *p*-values ranged from 15% to 57%). Therefore, the variables log(S), log(L), log(E), log(F), and log(St) were excluded resulting in Model 3.

Model 3. Model 3 employed the explanatory variable log(A) and three indicator variables in addition to a common intercept. All of the variables are highly significant with *p*-values less than 0.01%. The results in Table 4 show that Model 3 yields a slightly larger model error variance and a slightly smaller pseudo-$\overline{R}^2$ than Model 2, but the AVP for Model 3 is noticeably smaller than those of Models 1 and 2. Furthermore, Models 1 and 2 have large CN values and the coefficients are not all statistically significant at the 5% level as they are for Model 3. Plots of the residuals by region did not reveal any heteroscedasticity.

Table 4 also reports the traditional $\overline{R}^2$ for each model considered. The traditional $\overline{R}^2$ is consistently smaller than the pseudo-$\overline{R}^2$, illustrating how the traditional $\overline{R}^2$ underestimates the performance of the model.

Table 4 Summary statistics for GLS regression models of the logarithm of the 100-year peak flow obtained using the new estimator of Σ with $MSE_G \approx 0.100$ and $N = 134$ Sites

	CN	$\hat{\sigma}^2_\delta$	AVP	$\overline{R}^2$	R^2_{GLS}	EVR	MBV
Model 1a	297	0.0332	0.0373	0.850	0.890	0.363	3.32
Model 1	330	0.0159	0.0196	0.913	0.947	0.760	4.76
Model 2	102	0.0155	0.0186	0.914	0.949	0.780	4.82
Model 3	12.8	0.0156	0.0177	0.913	0.948	0.775	4.80
Model 4	9.5	0.0146	0.0177	0.919	0.951	0.825	4.94

Model 4. Interaction terms were introduced to test for possible differences in the slopes resulting in Model 4:

$$\hat{y}_i = a_0 + a_1 d_1 + a_2 d_2 + a_4 d_4 + (b_0 + b_1 d_1 + b_2 d_2 + b_4 d_4) \times \left[\log(A_i) - \frac{1}{N} \sum_{j=1}^{N} \log(A_j) \right] \quad (32)$$

None of the interaction terms are significant at the 5% level relative to $b_3 = 0$. This suggests the restrictions imposed by Model 3 are appropriate. However, Eq. (28) can be used to test if the interaction terms in Model 4 are jointly significant; the resulting test statistic $\tilde{\chi}_{[8]}$ has a value of 7.98, which is much less than the critical value 15.51 for a one-sided 5% test. This indicates that use of a single slope coefficient for all four regions is appropriate, and thus Model 3 is recommended. Table 5 reports the estimated coefficients using all 162 sites, their standard errors, and the corresponding *t*-values and *p*-values for a two-sided hypothesis. A pseudo ANOVA table is provided in Table 6. Table 7 reports additional summary statistics.

The recommended model has a different constant for each region, but a common slope. Although there may be physiographic differences among the four regions, these differences are insignificant in terms of the estimated slope coefficients. This makes sense if one believes that the basin time of concentration should scale with drainage area in a consistent fashion, such as the power law. However, the differences in the intercepts allow for differences in runoff volume due to differences in soil characteristics, land cover, storage area, and slope, even though individual slope coefficients for these variables were found to be insignificant.

Use of a pooled regression model versus separate regional models. Many regional regression studies, such as Feaster and Tasker (2002), define physiographic regions and then develop models of a given flood quantile for each region individually. The analysis here demonstrates this may be unwise. It may be beneficial to pool data across regions in which the physiographic differences are not significant, particularly when individual regions contain relatively few sites. Consider separate models of the form $\hat{y}_i = a + b\log(A_i)$ for each region. The coefficients and select summary statistics obtained for each region using GLS regression with the new estimator of Σ are reported in Table 8.

Using the data for all four regions, an AVP of 0.0201 was obtained for the pooled model (Model 3 above). This corresponds to AVPs of 0.0206, 0.0194, 0.0206, and 0.0211 for regions 1, 2, 3 and 4, respectively, using the physiographic characteristics of each region and the pooled model estimate of the variance. When separate models are used for each region, the change in the AVP for regions 2 and 4 is relatively modest, however, the AVP for region 1 is reduced by roughly 43%, and the AVP for region 3 is increased by about 28%. For region 1, the difference in the AVP is predominantly due to the observed decrease in the model error variance for that separate region, though the set of *x* values upon which the AVP is computed has also changed. The model error variance for the pooled model (Model 3) is 0.0182, but when using a separate model for region 1, the estimated model error variance decreases by roughly 50%, thereby yielding a large reduction in the AVP; but the model error variance estimate is now based on only 30 sites. It is also important to note that the variance of the slope coefficients in the models for regions 1, 3 and 4 are now two to three times greater than the variance of the common slope coefficient in the pooled model. Therefore, the separate model for region 1 may not be suitable for application at ungauged sites, particularly if they have unusual drainage areas.

Table 5 Estimated coefficients for recommended GLS regression models for the logarithm of the 100-year peak flow obtained using various estimators of Σ with $N = 162$ (with base-10 logarithms)

	b	SE	*t*-Value	*p*-Value
Using Eq. (4) to Compute Σ with $MSE_G = 0$				
Constant	2.1861	0.0631	—	—
d_1	0.6982	0.0631	11.07	0.0000
d_2	0.5667	0.0496	11.42	0.0000
d_4	0.2948	0.0593	4.968	0.0000
log(A)	0.6443	0.0176	36.61	0.0000
Using Eq. (5) to Compute Σ with $MSE_G \approx 0.100$				
Constant	2.1912	0.0632	—	—
d_1	0.6960	0.0634	10.98	0.0000
d_2	0.5641	0.0492	11.46	0.0000
d_4	0.2924	0.0598	4.892	0.0000
log(A)	0.6433	0.0175	36.74	0.0000
Using Eq. (5) to Compute Σ with $MSE_G = 0.302$				
Constant	2.1963	0.0629	—	—
d_1	0.6932	0.0631	10.99	0.0000
d_2	0.5611	0.0486	11.55	0.0000
d_4	0.2890	0.0594	4.864	0.0000
log(A)	0.6426	0.0173	37.11	0.0000

Table 6 Pseudo ANOVA table for recommended GLS regression models of the logarithm of the 100-year peak flow obtained using various estimators of Σ with $N = 162$

Source	Degrees-of-freedom	Sum of squares
Using Eq. (4) to Compute Σ with $MSE_G = 0$		
Regression model	4	49.0
Model error δ	156	3.32
Sampling error η	162	1.64
Total error	323	54.0
Using Eq. (5) to compute Σ with $MSE_G \approx 0.100$		
Regression model	4	49.0
Model error δ	156	2.94
Sampling error η	162	2.01
Total error	323	54.0
Using Eq. (5) to compute Σ with $MSE_G = 0.302$		
Regression model	4	49.1
Model error δ	156	2.59
Sampling error η	162	2.36
Total error	323	54.1

Table 7 Summary statistics for recommended GLS regression models of the logarithm of the 100-year peak flow obtained using various estimators of Σ with $N = 162$

	$\hat{\sigma}^2_\delta$	AVP	$\overline{R}^2$	R^2_{GLS}	EVR	MBV
Eq. (4) ($MSE_G = 0$)	0.0205	0.0224	0.909	0.937	0.495	4.76
Eq. (5) ($MSE_G \approx 0.100$)	0.0182	0.0201	0.909	0.943	0.684	4.68
Eq. (5) ($MSE_G = 0.302$)	0.0160	0.0178	0.909	0.950	0.912	4.61

Table 8 Estimated coefficients for GLS regression models for the log of the 100-year peak flow for each region using the new estimator of Σ with $MSE_G \approx 0.100$ (with base-10 logarithms)

	Region			
	1	2	3	4
a	2.7685	2.7965	2.0853	2.4527
b	0.7016	0.6159	0.6869	0.6493
SE(a)	0.0932	0.0521	0.1275	0.0955
SE(b)	0.0416	0.0232	0.0520	0.0411
$\hat{\sigma}^2_\delta$	0.0090	0.0207	0.0218	0.0166
AVP	0.0118	0.0221	0.0264	0.0207
$\overline{R}^2$	0.917	0.904	0.902	0.915
R^2_{GLS}	0.958	0.934	0.938	0.953
N	30	84	22	26

For region 3, the model developed using only the 22 sites in the region has a larger model error variance, and thus a larger estimated AVP than the joint model. Consider a particular site (#02102910) in region 3 for which the log of the drainage area is 0.3424: this site has the smallest drainage area in region 3, and the largest residual error using the region 3 model in Table 8. Now consider a new site in region 3 with the same drainage area. To demonstrate the benefit of pooling data across regions, the variance of prediction at this new site using the region 3 specific model in Table 8 will be compared to the variance of prediction using the pooled model (Model 3). For a given model, the variance of prediction at a site *i* is computed as

$$VP_{GLS} = \hat{\sigma}^2_\delta + \mathbf{x}_i(\mathbf{X}^T\widehat{\mathbf{\Lambda}}^{-1}\mathbf{X})^{-1}\mathbf{x}_i^T \quad (33)$$

where for a fair comparison, the estimated model error variance of 0.0182 from the pooled model is used in both computations. The variance of prediction for this new site is 0.0306 using the region-specific model for region 3 in Table 8, whereas the variance is only 0.0217 using the pooled model (Model 3), corresponding to roughly a 40% reduction. This indicates how the prediction error can be reduced by pooling the data from all four regions.

GLS regression model using alternative estimators of Σ

One goal of this study is to compare the precision of models estimated using the new and more appropriate estimator of Σ in Eq. (5) that accounts for the error in the skew versus the estimator in Eq. (4). In order to assess the potential impact of appropriately estimating Σ as in the previous section, the analysis was repeated using Eq. (4). Overall, the results and conclusions using Eq. (4) are similar to those reported above for the new estimator of Σ in Eq. (5). The same form of the fitted model of the log of the 100-year peak flow was chosen and the estimated coefficients are very similar, as were the AVP values. This is likely due to the precision of the regional skew model, which was assumed to be on the order of 0.100. Table 5 reports estimated coefficients using Eq. (4) with all 162 sites, their standard errors, and the *t*-values and *p*-values for a two-sided hypothesis test. A pseudo ANOVA table is provided in Table 6, and additional summary statistics are presented in Table 7.

What if the regional skew was not so precise? When using the regional skew map contained in *Bulletin 17B* an estimation error of 0.550, corresponding to $MSE_G = 0.302$, is to be employed. This is three times the variance of the regional skew model developed by Feaster and Tasker (2002). $MSE_G = 0.302$ has been used in many USGS studies; see for example Giese and Franklin (1996) and Walker and Krug (2003).

In order to assess the impact of a less precise regional skew on the estimated model error variance and AVP of the model of the 100-year peak flow, the analysis was repeated using the new estimator of Σ wherein $MSE_G = 0.302$ was employed in all four regions. Again, Model 3 was selected. The results are included in Tables 5–7. The estimated coefficients and their variances are very similar to those obtained using the more informative regional skew model with $MSE_G = 0.100$. However, when $MSE_G = 0.302$, the model error variance and the average variance of prediction are smaller than those obtained with $MSE_G = 0.100$. In this case we reduce the estimated model error variance by correctly accounting for the error attributed to the skew. However, this example suggests that gross errors would not result from use of Eq. (4) to compute Σ with $MSE_G = 0.302$.

Table 6 reveals the most telling differences among the three cases: when the error attributed to the skew increases, the total sampling error increases, and the estimated model error variance decreases accordingly, while the variation attributed to the model and the total variability in the data remain essentially fixed. Indeed the total variation in the data is fixed, and Eq. (16) seeks to partition the variability not explained by the model between the sampling error and the model error variances.

The results in Table 7 indicate that for models of the log of the 100-year peak flow both the model error variance and the AVP are significantly overestimated when the regional skew is relatively imprecise and it is incorrectly assumed that there is no error in the skew. This is because the error in Eq. (4) increases as the true precision of the regional model decreases. (See Griffis, 2006). And, the error in the estimator of Σ is dominated by the diagonal elements which are underestimated when the error in the skew is ignored. Thus, there is a trade-off between the magnitude of the elements of Σ and the estimated model error variance. By ignoring the error in the computed skew, the elements of Σ are underestimated, thereby resulting in the inflation of the estimate of the model error variance. This inflation is exacerbated if the regional skew is relatively imprecise as suggested by *Bulletin 17B* with an effective record length of only 17 years. However, if the regional skew is relatively precise as suggested by more recent studies, such as Tasker and Stedinger (1986) and Reis et al. (2003, 2004, 2005), with an effective record length in excess of 60 years, then the misrepresentation of the model error variance when using Eq. (4) is likely to be inconsequential. Here the average record length is 35 years across the 162 sites with a standard deviation of 21 years. The misrepresentation of the model error variance due to using Eq. (4) when $MSE_G \approx 0.100$ is likely to be even less important with shorter records because the Σ estimator would be dominated by at-site sampling error in the mean and standard deviation, rather than the imprecision in the regional skew. This is discussed in more detail below.

Application with fewer sites

The analysis presented above is based on a large number of sites, with 162 sites total across the four regions. Would the same trade-off between the sampling and model error be observed if fewer sites were available? Would the misrepresentation of the model error variance when using Eq. (4) still be modest when the regional skew is relatively precise? Consider region 3 which contains only 22 sites, with an average record length of 34 years and a standard deviation of 21 years. For the individual regions in this study area, Feaster and Tasker (2002) found that the most appropriate model of the log of the 100-year event is $\hat{y}_i = a + b\log(A_i)$. Table 8 contains the estimated coefficients and some summary statistics for this model obtained using the data for region 3 and GLS regression with the new estimator of Σ wherein $MSE_G \approx 0.100$. Similar coefficients are obtained using Eq. (5) with $MSE_G = 0.302$ and Eq. (4) to compute Σ. Summary statistics for these models obtained using the three estimators of Σ are summarized in Table 9.

With fewer sites available to develop the regression models, the values of $\hat{\sigma}^2_\delta$ and AVP reported in Table 9 for the

Table 9 Summary of model error variance and AVP for models of the log of the 100-year peak flow in region 3 obtained using various estimators of Σ with $N = 22$

	$\hat{\sigma}^2_\delta$	AVP	$\overline{R}^2$	R^2_{GLS}	EVR	MBV
Eq. (4) ($MSE_G = 0$)	0.0244	0.0290	0.903	0.931	0.422	1.75
Eq. (5) ($MSE_G \approx 0.100$)	0.0218	0.0264	0.902	0.938	0.586	1.80
Eq. (5) ($MSE_G = 0.302$)	0.0195	0.0240	0.902	0.944	0.765	1.81

models obtained using data for region 3 alone are larger than those in Table 7 for the models obtained using data pooled across all four regions. However, again the estimated model error and the AVP increase as the precision of the regional skew increases. Thus, the same trade-off between the sampling covariance and the model error variance is observed, as Eq. (16) suggests would occur. Furthermore, when the regional skew is relatively precise, the overestimation error due to incorrectly ignoring the error in the skew is again only on the order of 10% in terms of both the estimated model error variance and the AVP. For regions with a greater number of sites, the AVP will essentially equal the model error variance $\hat{\sigma}_\delta^2$, so any distortion in $\hat{\sigma}_\delta^2$ will be reflected in the computed AVP.

Conclusions

The appropriate estimator of the sampling error matrix for the log of LP3 quantiles obtained with an estimated skew is developed for use in GLS regression. This new estimator correctly accounts for the error in quantile estimates due to the error in a weighted skewness estimator. The results of this study indicate that the estimated model precision decreases significantly if it is incorrectly assumed there is no error in the skew when the regional skew is actually relatively imprecise. However, if the regional skew is relatively precise as recent studies have shown, then use of the estimator that ignores skew uncertainty results in little distortion with typical record lengths.

While these results were obtained using data for the State of South Carolina, they should generalize to other basins. For a number of record lengths and reasonable parameter values, Griffis (2006) revealed that the error in the estimator of the sampling covariance matrix which ignores skew uncertainty is dominated by the error in the diagonal elements which are consistently underestimated when the error in the computed skew is ignored. This underestimation of the sample variance results in an inflated estimate of the model error variance, as suggested by Eq. (16) which implicitly equates the sum of the sampling error and model error variances to the total variation in the residual errors. Nonetheless, this trade-off between the two variances should not impact model selection as the variance of the estimated coefficients depends on the total error.

This paper also presented innovative approaches to GLS regression in hydrologic applications. New metrics were employed to evaluate model performance. A condition number was used to check for multicollinearity, and a pseudo-$\overline{R}^2$ appropriate for use with GLS regression was used in place of the traditional $\overline{R}^2$ which misrepresents the fraction of the true variation in the statistic of interest that is explained by the model. In addition, variance ratios were used to verify that the full GLS regression analysis was necessary.

A full GLS model was developed to support modeling the standard deviation as a function of basin characteristics. Previous studies had used either ordinary or weighted least squares. The misrepresentation of beta variance (MBV) statistic had a value of 5, indicating that a WLS analysis would underestimate the actual variance of the estimator of the constant in the model by a factor of 5. Given that much of the model identification effort focuses on whether or not different constants were appropriate for different regions, the need for an honest and accurate estimate of the precision of estimators of such constants is important. Thus, future studies should seriously consider a GLS analysis if alternative models of the standard deviation are to be considered.

Pooled regression models were also employed to combine data across physiographic regions, thereby increasing the number of sites and information available for model estimation. When a region contains relatively few sites, pooling the data with neighboring regions allows for the development of a more accurate model, both in terms of the average variance of prediction and the precision of the estimated coefficients and the model error variance. The analysis found that a common slope parameter was appropriate, whereas the individual regions had statistically different constants. This makes sense if one believes that the basin time of concentration should scale with area to a power, while the different constants allow for differences in runoff volume due to differences in soil characteristics, land cover, storage area and slope. Similarly, for the standard deviation, a hybrid model with a separate constant for the lower coastal plain (region 4) was recommended based on the GLS regression analysis.

Acknowledgements

We greatly acknowledge support provided by a Water Resources Institute Internship Award #02HQGR0128 by the US Geological Survey, US Department of the Interior. Comments by Ken Eng and Dirceu Reis are gratefully acknowledged.

References

Benson, M.A., 1962. Evolution of Methods for Evaluating the Occurrence of Floods. US Geological Survey Water Supply Paper, 1580-A, pp. 30.

Bobee, B., 1973. Sample error of T-year events computed by fitting a Pearson type 3 distribution. Water Resour. Res. 9 (5), 1264–1270.

Chowdury, J.U., Stedinger, J.R., 1991. Confidence interval for design floods with estimated skew coefficient. J. Hydraul. Eng. 117 (7), 811–831.

Cruff, R.W., Rantz, S.E., 1965. A Comparison of Methods used in Flood Frequency Studies for Coastal Basins in California. US Geological Survey Water Supply Paper, 1580-E, pp. 56.

Curtis, G.W., 1987. Technique for Estimating Flood-peak Discharges and Frequencies on Rural Streams in Illinois. Water Resources Investigations Report 87-4207, US Geological Survey, Columbia, South Carolina, Urbana, Illinois.

Eng, K., Tasker, G.D., Milly, P.C.D., 2005. An Analysis of Region-of-Influence Methods for Flood Regionalization in the Gulf-Atlantic Rolling Plains. J. Am. Water Resour. Assoc. 41 (1), 135–143.

Feaster, T.D., Tasker, G.D., 2002. Techniques for Estimating the Magnitude and Frequency of Floods in Rural Basins of South Carolina. 1999, Water-Resources Investigations Report 02-4140, US Geological Survey, Columbia, South Carolina.

Giese, G.L., Franklin, M.A., 1996. Magnitude and Frequency of Floods in the Suwannee River Water Management District, Florida. Water-Resources Investigations Report 96-4176, US Geological Survey, Reston, VA, pp. 14.

Greene, W.H., 2003. Econometric Analysis. Prentice Hall, NJ.
Griffis, V.W., 2003. Evaluation of Log-Pearson type 3 Flood Frequency Analysis Methods Addressing Regional Skew and Low Outliers. M.S. Thesis, Cornell University.
Griffis, V.W., 2006. Flood Frequency Analysis: Bulletin 17, Regional Information, and Climate Change. Ph.D. Dissertation, Cornell University.
Griffis, V.W., Stedinger, J.R., Cohn, T.A., 2004. Log Pearson type 3 quantile estimators with regional skew information and low outlier adjustments. Water Resour. Res. 40, W07503. doi:10.1029/2003WR00269.
Griffis, V.W., Stedinger, J.R., 2007a. The log-Pearson type 3 distribution and its application in flood frequency analysis, 1. Distribution characteristics. J. Hydrol. Eng. 12(4), in press.
Griffis, V.W., Stedinger, J.R., 2007b. The log-Pearson type 3 distribution and its application in flood frequency analysis, 2. Parameter estimation methods. J. Hydrol. Eng. 12(4), in press.
Gruber, A.M., Reis Jr., D.S., Stedinger, J.R., 2007. Models of regional skew based on Bayesian GLS regression. In: World Environmental & Water Resources Conference, Tampa, Florida, May 15–18, 2007.
Interagency Committee on Water Data (IACWD), 1982. Guidelines for Determining Flood Flow Frequency: Bulletin 17-B (revised and corrected). Hydrol. Subcomm., Washington, DC, March 1982, pp. 28.
Johnston, J., 1984. Econometric Methods. McGraw-Hill, New York.
Kjeldsen, T.R., Rosbjerg, D., 2002. Comparison of regional index flood estimation procedures based on the extreme value type I distribution. Stoch. Environ. Res. Risk Assess. 16, 358–373.
Kroll, C.N., Stedinger, J.R., 1998. Regional hydrologic analysis: ordinary and generalized least squares revisited. Water Resour. Res. 34 (1), 121–128.
Kroll, C.N., Stedinger, J.R., 1999. Development of regional regression relationships with censored data. Water Resour. Res. 35 (3), 775–784.
Kuczera, G., 1983. Effect of sampling uncertainty and spatial correlation on an empirical Bayes procedure for combining site and regional information. J. Hydrol. 65, 373–398.
Landers, M.N., Wilson Jr., K.V., 1991. Flood Characteristics of Mississippi Streams. Water Resources Investigations Report 91-4037, US Geological Survey in cooperation with Mississippi State Highway Department, Jackson, Mississippi.
Law, G.S., Tasker, G.D., 2003. Flood-frequency Prediction Methods for Unregulated Stream of Tennessee, 2000. Water-Resources Investigations Report 03-4176, US Geological Survey, Nashville, Tennessee.
Ludwig, A.H., Tasker, G.D., 1993. Regionalization of Low-flow Characteristics of Arkansas Streams. US Geological Survey Water-Resources Investigations Report 93-4013, pp. 16.
Madsen, H., Rosbjerg, D., 1997. Generalized least squares and empirical Bayes estimation in regional partial duration series index-flood modeling. Water Resour. Res. 33 (4), 771–782.
Madsen, H., Rosbjerg, D., Harremoes, P., 1995. Application of the Bayesian approach in regional analysis of extreme rainfalls. Stoch. Hydrol. Hydraul. 9, 77–88.
Martins, E.S., Stedinger, J.R., 2002. Cross correlations among estimators of shape. Water Resour. Res. 38 (11), 1252. doi:10.1029/2002WR00158.
McCuen, R.H., 1979. Map skew. J. Water Res. Plan. Manag. Div., ASCE 105 (WR2), 269–277.
Medina, K.D., 1987. Analysis of Surface-water Data Network in Kansas for Effectiveness in Providing Regional Streamflow Information. US Geological Survey Water-Supply Paper 2303. US Geological Survey, Reston, VA, pp. 28.
Moss, M.E., Tasker, G.D., 1991. An intercomparison of hydrological network-design technologies. Hydrol. Sci. J. 36 (3), 209.
Pope, B.F., Tasker, G.D., Robbins, J.C., 2001. Estimating the Magnitude and Frequency of Floods in Rural Basins of North Carolina – Revised. Water-Resources Investigations Report 01-4207, US Geological Survey, Reston, VA, pp. 44.
Reis Jr., D.S., 2005. Flood Frequency Analysis Employing Bayesian Regional Regression and Imperfect Historical Information. Ph.D. Dissertation, Cornell University.
Reis Jr., D.S., Stedinger, J.R., Martins, E.S., 2003. Bayesian GLS regression with application to LP3 regional skew estimation. In: Bizier, P., DeBarry, P. (Eds.), Proceedings of the World Water and Environmental Resources Congress, Philadelphia, PA, June 23–26, 2003. American Society of Civil Engineers, Reston, VA.
Reis Jr., D.S., Stedinger, J.R., Martins, E.S., 2004. Operational Bayesian GLS regression for regional hydrologic analyses. In: Sehlke, G., Hayes, D.F., Stevens, D.K. (Eds.), Critical Transitions in Water and Environmental Resources Management, Proceedings of the World Water & Environmental Resources Congress, Salt Lake City, Utah, June 27–July 1, 2004. American Society of Civil Engineers, Reston, VA.
Reis Jr., D.S., Stedinger, J.R., Martins, E.S., 2005. Bayesian GLS regression with application to LP3 regional skew estimation. Water Resour. Res. 41, W10419. doi:10.1029/2004WR00344.
Soenksen, P.J., Miller, L.D., Sharpe, J.B., Watton, J.R., 1999. Peak-flow Frequency Relations and Evaluation of the Peak-flow Gaging Network in Nebraska. Water-Resources Investigation Report 99-4032.
Stedinger, J.R., Tasker, G.D., 1985. Regional hydrologic analysis 1. Ordinary, weighted, and generalized least squares compared. Water Resour. Res. 21 (9), 1421–1432.
Stedinger, J.R., Tasker, G.D., 1986a. Correction to Regional hydrologic analysis 1. Ordinary, weighted, and generalized least squares compared. Water Resour. Res. 22 (5), 844.
Stedinger, J.R., Tasker, G.D., 1986b. Regional hydrologic analysis 2. Model-error estimators, estimation of sigma and log-Pearson type 3 distributions. Water Resour. Res. 22 (10), 1487–1499.
Stedinger, J.R., Vogel, R.M., Foufoula-Georgiou, E., 1993. Frequency analysis of extreme eventsHandbook of Hydrology. McGraw-Hill Book Co., NY, pp. 18.1–18.66 (Chapter 18).
Tasker, G.D., 1978. Flood frequency analysis with a generalized least squares. Water Resour. Res. 14 (2), 373–376.
Tasker, G.D., 1980. Hydrologic regression with weighted least squares. Water Resour. Res. 16 (6), 1107–1113.
Tasker, G.D., Driver, N.E., 1988. Nationwide regression model for predicting urban runoff water quality at unmonitored sites. Water Resour. Bull. 24 (5), 1091–1101.
Tasker, G.D., Stedinger, J.R., 1986. Estimating generalized skew with weighted least squares regression. J. Water Res. Plan. Manage. 112 (2), 225–237.
Tasker, G.D., Stedinger, J.R., 1989. An operational GLS model for hydrologic regression. J. Hydrol. 111, 361–375.
Tasker, G.D., Eychaner, J.H., Stedinger, J.R., 1986. Application of Generalized Least Squares in Regional Hydrologic Regression Analysis. US Geological Survey Water-Supply Paper 2310, pp. 107–115.
Tasker, G.D., Hodge, S.A., Barks, C.S., 1996. Region of influence regression for estimating the 50-year flood at ungauges sites. Water Resour. Bull. 32 (1), 163–170.
Thomas, D.M, Benson, M.A., 1970. Generalization of Streamflow Characteristics from Drainage-basin Characteristics. US Geological Survey Water Supply Paper 1975, pp. 55.
Walker, J.F., Krug, W.R., 2003. Flood-frequency Characteristics of Wisconsin Streams. Water-Resources Investigations Report 03-4250, US Geological Survey, Reston, VA, pp. 42.
Wallis, J.R., Matalas, N.C., Slack, J.R., 1974. Just a moment! Water Resour.Res. 10 (2), 211–219.