

Models of Regional Skew Based on Bayesian GLS Regression

Andrea M. Gruber[1], Dirceu S. Reis Jr.[2], and Jery R. Stedinger[3]

[1] Graduate Research Assistant, School of Civil and Environ. Engineering, Cornell Univ., 220 Hollister Hall, Ithaca, NY 14853-3501; Email: amg66@cornell.edu
[2] Water Resources Dept., Fundaçã Cearense de Meteorologia e Recursos Hídricos, Avenida Rui Barbosa, 1246 Fortaleza, Ce, Brazil. email: dirceu@funceme.br
[3] Professor, School of Civil & Environ. Engineering, Cornell Univ., 220 Hollister Hall, Ithaca, NY 14853-3501 USA; PH (607) 255-2351; Email: jrs5@cornell.edu.

Abstract:
The skew map developed by Hardison in 1974 was based on records of at least 25 years in lengths. It is still used today, over 30 years later. The first edition of Bulletin 17 states: "It is expected that Plate 1 [the skew map] will be revised as more data become available and more extensive studies are completed." Today, tremendous advances in computing power and spatial statistical methods allow for a much better analysis of the larger data set now available. This paper describes a Bayesian Generalized Least Squares (B-GLS) framework together with diagnostic statistics introduced by Reis *et al*. (2005) that can be used to develop regional skew relationships. An example using data from the Illinois River Basin illustrates useful diagnostic statistics including pseudo R^2, Bayesian plausibility, leverage, influence and σ-influence. The B-GLS framework and diagnostic statistics developed in this analysis are being applied to an ongoing study in the southeast United States which will produce a regional skew estimator.

Introduction

This paper further develops the Bayesian Generalized Least Squares (B-GLS) regression framework first presented in Reis *et al.* (2005). This operational regression methodology is used in the estimation of regional shape parameters, as well as flood quantiles. The focus of this paper is implementing the B-GLS framework developed by Reis *et al.* (2005) in conjunction with diagnostic statistics presented by Reis *et al.* (2005), Reis (2005), and Griffis and Stedinger (2006). New diagnostic statistics for use with regression analyses presented in this paper include pseudo adjusted R-squared (R^2_{GLS}), Bayesian plausibility value (ψ) and σ-Influence. These new statistics in conjunction with the average variance of prediction for a new site (AVP_{new}), error variance ratio (EVR), misrepresentation of the beta variance (MBV), leverage and influence allow for a comprehensive examination of the regression models.

Hydrologic Regression Analysis

The B-GLS regression framework developed by Reis *et al.* (2005) can be used with streamflow data in order to derive empirical relationships between hydrologic characteristics at a site, such as the log-space skewness coefficient used to fit a log-Pearson Type III distribution, and physical watershed characteristics. The GLS model used in the B-GLS analysis was developed by Stedinger and Tasker (1985) and Tasker and Stedinger (1986). In the analysis it is assumed that the actual value of the quantity of interest y_i for a given site i can be described by a function of physiographic characteristics with an additive error

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ij} + \delta_i \qquad i = 1,2,\ldots n \text{ sites} \tag{1}$$

where X_{ij} (j=1…k) are the elements of a matrix of k explanatory variables based upon the physical characteristics for each site used in the regression model, and δ_i are the independently distributed model errors with mean zero and variance σ_δ^2. However, in most analyses, only an estimate of y_i is available, and thus a time-sampling error η_i should be introduced into the model. As formulated in Reis *et al.* (2005, eqn. 6), the GLS model becomes

$$\hat{y} = \mathbf{X\beta} + \boldsymbol{\eta} + \boldsymbol{\delta} = \mathbf{X\beta} + \boldsymbol{\varepsilon} \qquad \text{where } \hat{y}_i = y_i + \eta_i \qquad i = 1,2,\ldots n \text{ sites} \tag{2}$$

Thus the observed regression model errors ε_i are the sum of the model errors δ_i and the sampling errors η_i. The total error vector ε has mean zero and a covariance matrix

$$E\left[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T\right] = \boldsymbol{\Lambda}(\sigma_\delta^2) = \sigma_\delta^2 \mathbf{I} + \boldsymbol{\Sigma}(\hat{\mathbf{y}}) \tag{3}$$

where $\boldsymbol{\Sigma}(\hat{\mathbf{y}})$ is the covariance matrix of the sampling errors in the sample estimators.

The B-GLS regression framework constructed by Reis *et al.* (2005) for the regionalization of hydrologic data requires specification of prior distributions for both the β parameters and the model error variance σ_δ^2. An almost non-informative multivariate normal distribution with a mean of zero and a large variance is used here as the β prior. An exponential distribution with parameter λ was used as the model error variance σ_δ^2 prior. The parameter λ is the reciprocal of the prior mean for σ_δ^2. Following Reis *et al.* (2005), we set λ to 6 with the understanding that as experience accumulates a smaller value could be justified. After determining the prior distributions, Reis *et al.* (2005) calculated the posterior moments of the β parameters and the full posterior distribution of the model error variance σ_δ^2. In doing so, they showed that B-GLS provides a more realistic description of possible values of the model error variance, especially in cases where the sampling error variances are larger than the model error variance.

Model Selection

In order to determine which covariates, if any, should be included in a regression model, descriptive statistics have been developed to evaluate how well the model describes the data. The goal of model selection is to resolve which set of possible explanatory variables best fit the data affording the most accurate skew prediction, while also allowing for the simplest model possible. Traditional diagnostic statistics available for model selection include R^2, likelihood ratios, Mallows C_p statistic, Akaike Information Criterion (AIC), and the Bayesian Information Criterion

(BIC) [Linhart and Zucchini, 1986; Gelman *et al.*, 2004]. Many of these statistics penalize additional complexity: thus, a sufficient improvement in the model's prediction ability must result so as to support the inclusion of an additional independent variable. Below, we develop descriptive statistics to evaluate the B-GLS regression parameters.

Average Variance of Prediction

The Average Variance of Prediction (AVP) is a natural metric to use in evaluating models because the main motivation of creating the model is to be able to make accurate skew predictions at both gauged and ungauged sites. However, because the AVP accounts for the sampling variance of the parameters, the greater the number of parameters, the larger penalty potentially assessed.

As noted in Reis *et al.* (2005) [following Tasker and Stedinger, 1986], our AVP_{new} assumes that the sites used in the regression are representative of sites where skew predictions will be made because values of their independent variables are used to compute the average variance of prediction for new site AVP_{new}, as follows:

$$AVP_{new} = E[\sigma_\delta^2] + \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i Var[\boldsymbol{\beta} \mid \hat{\mathbf{y}}] \mathbf{x}_i^T \quad (4)$$

If the prediction is being computed for an old site, i.e. site *i* used in the regression analysis, one needs the average variance of prediction for an old site AVP_{old},:

$$AVP_{old} = E[\sigma_\delta^2] + \frac{1}{n}\sum_{i=1}^{n} \left\{ \mathbf{x}_i Var[\boldsymbol{\beta} \mid \hat{\mathbf{y}}] \mathbf{x}_i^T - 2E\left[\sigma_\delta^2 \mathbf{x}_i (\mathbf{X}^T \boldsymbol{\Lambda}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{\Lambda}^{-1} \mathbf{e}_i\right] \right\} \quad (5)$$

Here e_i, is a unit column vector with 1 at the *i*-th row and zero otherwise.

Bayesian Plausibility Value

The Bayesian Plausibility Value, ψ, developed by Reis (2005) describes whether zero is a plausible value for each β-parameter in a regression model given the prior and the data. In the Bayesian framework, the posterior pdf of the β-parameters is obtained. As discussed by Lindley (1965) and Zellner (1971), given the posterior pdf of β and the data available, one can construct a credible region for the regression parameters. This credible region is a summary of the posterior beliefs about the parameter and can then be the basis of a hypothesis test that concludes that a parameter is zero if zero is included in a 90% or a 95% credible region. This allows one to perform the equivalent of classical hypothesis tests within a Bayesian framework using the posterior distribution of each parameter. Here we define the plausibility level for zero to be the smallest probability ψ such that zero is in a 100(1-ψ)% credible region for a parameter. This is analogous to the P-value computed in classical statistics to describe the statistical significance of an estimate and whether the parameter should be included in a model, rather than being assigned a value of zero, which is the default in this case. As such, a plausibility value of 5% or less (or 10% if a 90% credible region is used), suggests that the model would be improved by setting the value of the tested parameter to zero. The plausibility value is computed as

$$\psi = 2 \underset{\sigma_\delta^2}{E}\left\{ \Phi\left[-\upsilon \frac{b(\sigma_\delta^2)}{\sigma_b(\sigma_\delta^2)} \right] \right\} \quad (6)$$

wherein Φ is the standard normal cdf, and the conditional mean $b(\sigma_\delta^2)$ and standard error $\sigma_b(\sigma_\delta^2)$ for β_ι are both dependent on σ_δ^2; υ = sign[μ_β] = 1 for $\mu_\beta \geq 0$ and -1 for $\mu_\beta < 0$;.

The Bayesian P-value discussed by Bayarri and Berger (2000) and Robins et al. (2000) corresponds to the probability that another random sample X would generate a more extreme value of a test statistic than that which was observed, and thus is a statistic more consistent with the classical P-value. These authors and others have tried to develop a Bayesian P-value that strictly reflects the data and not the prior. However, the Bayesian Plausibility Value reflects the Bayesian point of view that the prior is also information about the parameters, and thus, it is appropriate to use such information when deciding when to include a parameter in a model.

Pseudo-R^2_{GLS} and Pseudo ANOVA table

The traditional Ordinary Least Squares (OLS) measures, R^2 and $\overline{R}^2$, explain the degree to which a model explains the variability in the data. These measures use the partitioning of the sum of squared deviations and associated degrees of freedom to analyze the variance of the signal versus the model error. However, for Weighted Least Squares (WLS) and GLS these measures are inappropriate because they group together both the sampling $\mathbf{\Sigma}(\hat{\mathbf{y}})$ variance and the model error $\sigma_\delta^2\mathbf{I}$ variance.

In the B-GLS framework, the error of most concern is the model error variance because the sampling error is unexplainable and represents noise that complicates the analysis. Thus, a new measure is needed in which the sampling error variance is separated from the total error variance, leaving behind the fraction of the variance accounted for by the model and by the model error. Such a statistic for the B-GLS regression model, Pseudo-R^2_{GLS}, was developed by Reis (2005). We proposed that it be calculated, as:

$$\text{Pseudo - } R^2_{GLS} = \frac{n\left[\hat{\sigma}_\delta^2(0) - \hat{\sigma}_\delta^2(k)\right]}{n\hat{\sigma}_\delta^2(0)} = 1 - \frac{\hat{\sigma}_\delta^2(k)}{\hat{\sigma}_\delta^2(0)} \qquad (7)$$

where $\hat{\sigma}_\delta^2(k)$ is the estimated model error variance with k explanatory variables and $\hat{\sigma}_\delta^2(0)$ is the estimated model error variance when no explanatory variables are present. Pseudo-R^2_{GLS} is a direct extension of the traditional adjusted R^2 in that it uses the ratio of unbiased estimators of the variance of the error δ and the variance of y. If $\hat{\sigma}_\delta^2(k) = 0$, then $\bar{R}^2_{GLS} = 1$ as it should, even though the model is not perfect because Var[$\eta_i + \delta_i$] is still not zero because Var[η_i] > 0.

Table 1 presents a pseudo Analysis of Variance (ANOVA) table for WLS or GLS. This table describes how much of the variation in the observations can be attributed to the regional model, and how much of the residual variation can be attributed to model error and sampling error, respectively. The problem is that we cannot actually resolve what the model errors are because we do not know the values of the sampling errors η_ι for each i. But we can describe the total sampling error sum of squares by its mean value, which is $tr[\Sigma(\hat{\mathbf{y}})]$, where tr[A] is the trace of matrix A. And because there are n equations, the total variation due to the model error δ for a

model with k parameters has a mean equal to $n\sigma_\delta^2(\mathrm{k})$. That provides descriptions of two of the three sources of variation.

Now, for a model with no parameters other than the mean, the estimated model error $\sigma_\delta^2(0)$ describes all of the variation in $\hat{y}_i = y_i + \eta_i$ not explained by the sampling errors η_i. Thus, it should on average equal the actual variation in y due to regression and the variation due to the model errors δ. We describe the TOTAL expected sum of squares variation due to model, model error, and sampling error as $n\sigma_\delta^2(0) + tr[\Sigma(\hat{\mathbf{y}})]$. Therefore we would attribute to the model an expected sum of squares equal to $n[\sigma_\delta^2(0) - \sigma_\delta^2(k)]$. This is called a pseudo ANOVA because the contributions of the three sources of error are estimated or constructed, rather than being determined from the computed residual errors and the observed model predictions, and the impact of correlation among the sampling errors is ignored.

Table 1: Pseudo ANOVA table

Source	Degrees-of-Freedom	Sum of squares
Model	k	$n[\sigma_\delta^2(0) - \sigma_\delta^2(k)]$
Model Error	n-k-1	$n\sigma_\delta^2(k)$
Sampling Error	n	$tr[\sum(\hat{y})]$
Total	2n-1	$n\sigma_\delta^2(0) + tr[\sum(\hat{y})]$
EVR =	$\frac{1}{n} tr[n(\hat{y})]/\sigma_\delta^2(k)$	
MBV =	$\frac{1}{n} w^T \Lambda(\sigma_\delta^2) w$ where w is the vector $(1/\sqrt{\Lambda_{ii}})$	

Error Variance Ratio and the Misrepresentation of the Beta Variance

The Error Variance Ratio (EVR) is a modeling diagnostic used to determine if a simple OLS regression is sufficient or a more sophisticated WLS or GLS analysis is appropriate. EVR is the ratio of the average sampling error variance to the model error variance. An EVR greater than 20%, indicating that the sampling variance is not negligible when compared to the model error variance, suggests the need for a WLS or GLS regression analysis. The EVR is calculated as follows,

$$EVR = \frac{\mathrm{SS(sampling\,error)}}{\mathrm{SS(model\,error)}} = \frac{\mathrm{tr}[\boldsymbol{\Sigma}(\hat{\mathbf{y}})]}{n\sigma_\delta^2(k)} \tag{8}$$

Although the EVR distinguishes between OLS and WLS/GLS analyses, it does not determine whether a WLS or GLS regression is best suited for the data. Thus, the Misrepresentation of the Beta Variance (MBV) statistic was developed to determine whether a WLS regression is sufficient or if a GLS regression is appropriate (Griffis and Stedinger, 2006; Griffis, 2006). The MBV describes the error produced by a WLS regression analysis in its evaluation of the precision of b_0^{WLS}, which is the estimator of the constant β_0^{WLS}, as the covariance among the estimated y_i's generally has its greatest impact on the precision of the constant term (Stedinger and Tasker, 1985). If the MBV is substantially greater than 1, then a GLS analysis should be employed. The MBV is calculated as follows,

$$MBV = \frac{Var\left[b_0^{WLS} \mid GLS\ analysis\right]}{Var\left[b_0^{WLS} \mid WLS\ analysis\right]} = \frac{w^T \Lambda w}{n} \text{ where } w_i = \frac{1}{\sqrt{\Lambda_{ii}}} \quad (9)$$

Leverage and Influence

Leverage and influence are two descriptive statistics used to evaluate the fit of the regression model to the data, model adequacy and data quality. Leverage, as adopted by Tasker and Stedinger (1989, eqn. 23), considers whether a specific observation, or x-value, is unusual, and thus likely to have a large impact on the estimated regression coefficients. When this leverage statistic is applied to WLS and GLS regressions, both the x-value of the observation is taken into account as well as the statistical weight placed on the point. Thus, leverage measures the marginal/unit impact of the residuals ε_i on the estimated y_i-values. The Tasker and Stedinger (1989) leverage for WLS and GLS regression analysis is,

$$leverage(\hat{y}_i, \mathbf{x}_i) = \frac{\partial(\mathbf{x}_i \mathbf{b})}{\partial \varepsilon_i} = \underset{\sigma_\delta^2}{E}\left[\mathbf{x}_i (\mathbf{X}^T \mathbf{\Lambda}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{\Lambda}^{-1}\right]_i \quad (10)$$

where $x_i b$ is the estimator of y_i associated with x_i. The average values of the leverage statistic is (k+1)/n, where k is the dimension of β and n is the number of sites in the regression, thus a leverage value greater than 2(k+1)/n can be considered to be large (Tasker and Stedinger, 1989).

Unlike leverage which highlights points which are likely to affect the fit of the regression, influence describes those points which did have an unusual impact on the regression analysis. In many cases, those observations with high influence also have high leverage. High influence requires a combination of leverage and a large residual error. The influence measure below, proposed by Tasker and Stedinger (1989), is based on Cook's D (Cook and Weisberg, 1982; Clarke, 1994),

$$D_i = \frac{k_{ii} \hat{\varepsilon}_i^2}{(k+1)(\lambda_{ii} - k_{ii})^2} \quad (11)$$

where k_{ii} and λ_{ii} are the diagonal elements of $\mathbf{K} = \mathbf{X}(\mathbf{X}^T \mathbf{\Lambda}^{-1} \mathbf{X})^{-1} \mathbf{X}^T$ and Λ. Influence values greater than 4/n are considered to be large.

σ-Influence

In using regional skew models, the model error variance is very important because it determines the weight placed on the regional value relative to the at-site estimator. Thus, we are interested in knowing which, if any, observations had an unusual impact on the estimated model error variance. The σ-influence statistic describes the influence of each observation on the estimated model error variance, thus identifying points that are individually responsible for potentially inflating the estimated model error variance. The influence statistic D_i described above identifies those points with significant influence on the model predictions, thus revealing the instability of the fitted model's predictions in those x-value regions. However, D_i only describes the influence a point has on the model's prediction of skew at each point's x-value. It does not necessarily describe whether the point has a significant influence on the estimated model error variance. The σ-influence considers if a

residual error ε_i actually had a large impact on the estimated value of the model error σ_δ^2. The σ-influence is calculated,

$$\sigma - Influence_i = \frac{\sum_{j=1}^{n} \hat{\varepsilon}_i (\Lambda^{-1})_{ij} \hat{\varepsilon}_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} \hat{\varepsilon}_i (\Lambda^{-1})_{ij} \hat{\varepsilon}_j} = \frac{\hat{\varepsilon}_i (\Lambda^{-1}\hat{\varepsilon})_i}{\hat{\varepsilon}^T \Lambda^{-1} \hat{\varepsilon}} \quad (12)$$

Here the standardize sum-of-squres $\hat{\varepsilon}^T \Lambda^{-1} \hat{\varepsilon}$ used to compute the likelihood function for the data, and the generalized method of moments model error variance in Stedinger and Tasker (1985), is divided among the different sites. By construction, the average value of σ-influence is 1/n, where n is the number of sites in the regression. σ-influence values greater than 2/n are considered to be large.

Application: Regional Skew Estimation

Bulletin 17B (IACWD, 1981) recommends the use of the Log-Pearson Type 3 distribution for flood frequency analysis. The data available at a given site is usually too short to provide a good estimate of the skewness coefficient. In order to improve the precision of the skewness estimator, Bulletin 17B recommends combining a regional skew with the at-site skew [Hardison, 1975; McCuen, 1979 and 2001; IACWD, 1981; Stedinger et al., 1993; Griffis and Stedinger, 2007].

Reis *et al.* (2005) developed regional skew models using 17 sites in the Tibagi River basin and using 44 sites in the Muskingum River basin. In this study, a regional skew model was developed using a larger data set: the Illinois River basin with 62 sites whose record lengths vary from 14 to 90 years. Current efforts are considering data from ten states in the southeast U.S. to use B-GLS to create a new regional skewness model. This southeast undertaking is an ongoing project for which the results to date are presented, as well our expectations for future work.

The Illinois River basin study considered seven explanatory variables, plus a constant. Two binary variables (Z_1,Z_2) were employed to explore variability in the at-site skews that could be explained by hydrologic region. These two binary variables represented the three Illinois River basin regions: Little Wabash (1,0), Rock (0,1) and Sangamon (0,0), as described in Tasker and Stedinger (1986). The five other explanatory variables were: 1) drainage area expressed in sq mi; 2) main channel slope expressed in ft/mi, 3) lake area expressed percentage of drainage area plus one; 4) forest cover expressed as percentage of drainage area plus one; 5) soil permeability index which varies from 1 (low infiltration) to 6 (high infiltration). The logarithms of the above five explanatory variables were taken and then centered by subtracting their means. This enables the scale of five non-binary variables to match the scale of the constant and binary variables, and thus, allows for easier computation of the regional mean of each hydrologic region.

A sampling covariance matrix was developed for the Illinois Data using the estimation procedure described in Reis *et al.* (2005). The inter-site correlation coefficient between concurrent flows $\rho(d_{ij})$ for the Illinois Basin was modeled as a function of the distances between two sites

$$\rho(d_{ij})^{\kappa} = \theta^{\left(\frac{\kappa d_{ij}}{\alpha . d_{ij}+1}\right)} \quad (13)$$

where d_{ij} *i*s the distance between sites in kilometers, θ = 0.988, α = 0.002, and κ = 3.

Tables 2 and 3 along with Figure 1 present the results of the estimation of the regional skew for the Illinois River basin based on B-GLS regression analyses. All combinations of the seven explanatory variables were used to generate 128 possible skew regression models. The B-GLS1 model with 1 explanatory variable, the ln(channel slope), was chosen as the best model as measured by its minimum variance of prediction for a new site and small model error variance. For comparison, the regional constant model, without any covariates, B-GLS0, is also presented in Table 2.

As shown in Table 2, the model error variance for the B-GLS1 model is 0.133 compared to the constant model's, B-GLS0, model error variance of 0.151. The B-GLS1 average variance of prediction (AVP) at a new site is 0.158, which corresponds to an effective record of 49 years. This indicates that the regional skew would be equivalent to an at-site skewness estimator based on 49 years of data. Still, R^2_{GLS} is equal to only 0.12 which implies that ln(channel slope) explains only 12% of the variability of the true station skews. In comparison, an OLS regression on the same two parameter model yields an adjusted R^2 of only 0.9%, which underestimates the true power of the regional model.

Table 2: Skew Regression for the Illinois River Basin (62 sites).
Bayesian Plausibility (%) and Standard Errors are presented in parenthesis.

	Constant	ln(slope)	Model Error Variance	Average Sampling Variance	AVP (new site)	R^2_{GLS}	Effective Record Length
B-GLS 0	-0.419	-	0.151 (0.056)	0.015	0.166	0	49
B-GLS 1	-0.588	0.127 (3.3%)	0.133 (0.052)	0.025	0.158	0.118	49
B-GLS 1 w/o site 28	-0.533	0.105 (7.8%)	0.120 (0.051)	0.024	0.144	0.071	53
B-GLS 1 w/o site 48	-0.594	0.125 (2.5%)	0.122 (0.049)	0.023	0.145	0.143	53

Table 3 displays the pseudo ANOVA table, where for the B-GLS1 model the sampling error is more than twice as large as the model error. The EVR equal to 2.3 clearly indicates that either a WLS or GLS analysis should be employed as opposed to an OLS analysis. Moreover, because the MBV = 3 exceeds one, a GLS analysis is clearly appropriate; a WLS analysis would overestimate the precision of b_0.

Table 3: Pseudo ANOVA table for the Illinois River Basin (B-GLS 1)

Source	Degrees-of-Freedom		Sum of squares		
	Case 1	Cases 2 & 3	Case 1 (all sites)	Case 2 (w/o site 28)	Case 3 (w/o site 48)
Model	k = 1	k = 1	1.10	0.56	1.24
Model Error	n-k-1 = 60	n-k-1 = 59	8.24	7.30	7.41
Sampling Error	n = 62	n = 61	19.04	18.81	18.29
Total	2n-1 = 123	2n-1 = 121	27.28	26.11	25.70
EVR =			2.31	2.58	2.47
MBV =			3.00	3.03	3.01
R^2_{GLS} =			0.12	0.07	0.14

Figure 1 depicts the influence, leverage, and σ-influence statistics of the most influential sites in the B-GLS1 regression analysis. The sites were ordered by decreasing influence. It is clear that only site 28 with the highest influence also has high leverage (i.e. leverage which surpasses the high leverage threshold).

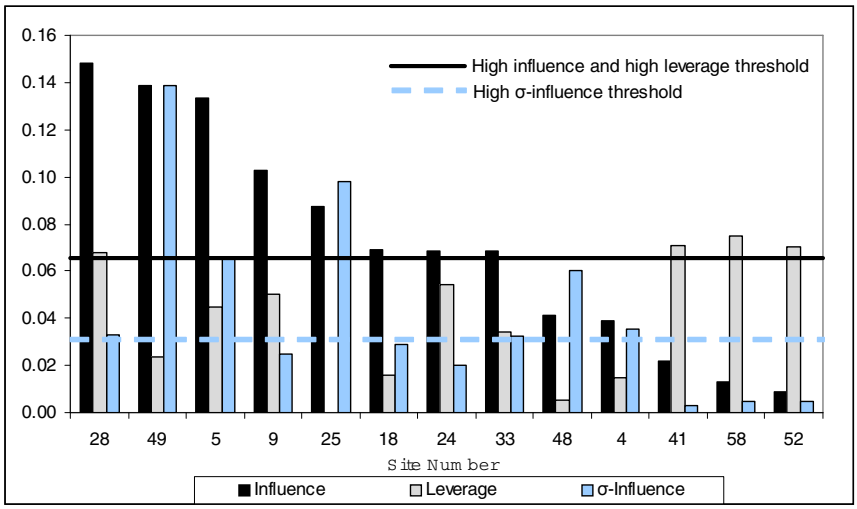

Figure 1: Regression Diagnostics: leverage, influence, and σ-influence for Illinois River basin

As well as having the smallest channel slope in the study, 0.84, site 28 also has a long record, 60 years, and large residual, -0.88. , This makes site 28 an outlier in the regression resulting in large influence, leverage and σ-influence values. As a test, site 28 was removed from the data set and the regression analysis for the B-GLS1 model was repeated. As shown in Table 2, the model error variance estimate decreases from the B-GLS1 value of 0.133, when the entire dataset is included, to a value of 0.120, when site 28 is removed, Table 3 provides the pseudo ANOVA table for the B-GLS1 model w/o site 28.

As shown in Figure 1, site 48 has a high σ-influence value while also having low influence and leverage values. The large residual of site 48, -1.53, is the third largest residual in the Illinois Basin dataset. Site 48 has a short record length of 16 years and a large skew of -1.82. The B-GLS1 model predicts the skew at site 48 as -0.282. As a test, site 48 was removed from the data set and the regression analysis for the B-GLS1 model was repeated. As shown in Table 2, the model error variance estimate is lowered from 0.133 when the entire dataset is included to a value of 0.122 when site 48 is removed. This decrease was expected because the data point with the third highest σ-influence was removed. Even though site 48 has a low influence value, it has a large impact on the model error variance due to its high σ-influence value. This site was missed by the influence statistic but recognized by the σ-influence statistic, thus illustrating the usefulness of the new σ-influence statistic. Table 3 provides the pseudo ANOVA table for the B-GLS1 model w/o site 48.

Conclusions and Ongoing Work

This paper further develops a quasi-analytic Bayesian analysis of a GLS regression model described by Reis *et al.* (2005) into an operational GLS regional hydrologic regression methodology. Regression diagnostic statistics for B-GLS models include a pseudo adjusted R^2, pseudo ANOVA tables, Bayesian Plausibility

value, EVR, MBV, leverage, influence and σ-influence. The regional regression procedure was illustrated with an example of regionalization of the skew parameter for the Log-Pearson Type III distribution.

Currently, this B-GLS procedure is being applied to a dataset with over 800 sites from ten states from the southeast United States. This regionalization study is applying the B-GLS procedure outlined in Reis *et al.* (2005) combined with the diagnostic statistics outlined in this paper. Another important issue that the southeast study has uncovered is the existence of several low-outliers resulting in anomalous skew values, and thus, very large variance for the regional skew estimators. Therefore, in order to censor the data set it is anticipated that the Bulletin 17B (IACWD 1982) low-outlier detection threshold with a conditional probability adjustment along with the expected moments algorithm (EMA) developed by Cohn (1997) will be implemented to determine an appropriate regional skew estimator.

References

Bayarri, M.J., and J.O. Berger, P (2000), Values for Composite Null Models, J. of the Am. Statistical Assoc. 95 (452), 1127-1142, Dec.

Clarke, R. T. (1994), *Statistical Modeling in Hydrology,* John Wiley & Sons Inc.

Cohn, T.A., W.L. Lane, and W.G. Baier (1997), An algorithm, for computing moments-based flood quantile estimates when historical flood information is available, *Water Resour. Res.,* 33(9), 2089-2096.

Cook, R.D. and Weisbeg, S. (1982), Residuals and Influence in Regression, Chapman and Hall, New York, NY, 230 pp.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004), *Bayesian Data Analysis*, Chapman & Hall/CRC, Boca Raton, FL.

Griffis, V. W., and J. R. Stedinger (2006), The Use of GLS Regression in Regional Hydrologic Analyses, manuscript, Cornell University, July.

Griffis, V. W. (2006), Flood Frequency Analysis, Bulletin 17B and Regional Analysis, Ph.D. Thesis, Cornell University, August.

Griffis, V. W., and J. R. Stedinger (2007), The LP3 distribution and its application in flood frequency analysis, 3. Sample Skew and Weighted Skew Estimators, submitted *J. of Hydrol. Engineering.*

Hardison, C. H. (1975), Generalized skew coefficients of annual floods in the United States and their application, *Water Resour Res., 11*(6), 851-854.

Interagency Advisory Committee on Water Data (1982), Guidelines for Determining Flood Flow Frequency, Bulletin #17B, U.S. Department of the Interior, U.S. Geological Survey, Office of Water Data Coordination, Reston Virginia.

Linart, H. and W. Zucchini (1986), *Model Selection*, John Wiley and Sons, Inc., New York.

Lindley, D.V. (1965), *Introduction to Probability and Statistics from a Bayesian Viewpoint, Part2. Inference*. Cambridge: University Press.

Reis, D. S., Jr., J. R. Stedinger, and E. S. Martins (2005), Bayesian generalized least squares regression with application to log Pearson type 3 regional skew estimation, *Water Resour. Res.*, *41*, W10419, doi:10.1029/2004WR003445.

Reis, D.S.,Jr. (2005), Flood Frequency Analysis Employing Bayesian Regional Regression and Imperfect historical Information, Ph.D. Thesis, Cornell University, January.

Robins, J. M. (2000), A. van der Vaart, and V. Ventura, Asymptotic Distribution of P Values in Composite Null Models, , J. of the Am. Statistical Assoc. 95 (452), 1143-1156, Dec.

Stedinger, J.R., and G.D. Tasker (1985), Regional Hydrologic Analysis, 1. Ordinary, Weighted and Generalized Least Squares Compared, *Water Resources Research, 21*(9), 1421-1432.

Tasker, G and J.R. Stedinger (1986a), Correction to "Regional hydrologic analysis, 1, Ordinary, weighted and generalized least squares compared", *Water Res. Research*, 22(5), 844.

Tasker, G and J.R. Stedinger (1986b), Regional hydrologic analysis, 2: Model-error estimators, estimation of sigma and log-Pearson Type 3 distributions, *Water Resour. Res.*, 22(10), 1487-1499.

Tasker, G.D., and J.R. Stedinger (1989), An Operational GLS Model for Hydrologic Regression, *Journal of Hydrology, 111*(1-4), 361-375.

Zellner, A. (1971), *An Introduction to Bayesian Inference in Econometrics*, John Wiley and Sons, Inc., New York.