Regional hydrologic analysis: Ordinary and generalized least squares revisited

Charles N. Kroll
Environmental Resources and Forest Engineering, College of Environmental Science and Forestry, State University of New York, Syracuse

Jery R. Stedinger
School of Civil and Environmental Engineering, Cornell University, Ithaca, New York

Abstract. Generalized least squares (GLS) regional regression procedures have been developed for estimating river flow quantiles. A widely used GLS procedure employs a simplified model error structure and average covariances when constructing an approximate residual error covariance matrix. This paper compares that GLS estimator ($\hat{\beta}_{GLS}^{MC}$), an idealized GLS estimator ($\hat{\beta}_{GLS}^{E}$) based on the simplifying assumptions of $\hat{\beta}_{GLS}^{MC}$ with true underlying statistics in a region, the best possible GLS estimator ($\hat{\beta}_{GLS}^{T}$) obtained using the true residual error covariance matrix, and the ordinary least squares estimator ($\hat{\beta}_{OLS}$). Useful analytic expressions are developed for the variance of $\hat{\beta}_{GLS}^{T}$, $\hat{\beta}_{GLS}^{E}$, and $\hat{\beta}_{GLS}^{MC}$. For previously examined cases the average sampling mean square error (mse) of the GLS estimators was mostly due to estimating streamflow statistics employed in the construction of the residual error covariance matrix rather than the simplifying assumptions in presently employed GLS estimators. The new analytic expressions were used to compare the performance of the OLS and GLS estimators for new cases representing greater model variability across sites as well as the effect return period has on the estimators’ relative performance. For a more heteroscedastic model error variance and larger return periods, some increase in the mse of both $\hat{\beta}_{GLS}^{E}$ and $\hat{\beta}_{GLS}^{T}$ was observed.

1. Introduction

Regional regression models are often used to estimate flow statistics at ungaged river sites. Relationships between flow statistics and geomorphic, geologic, climatic, and topographic parameters have been formulated in many regions for low flows [Thomas and Benson, 1970; Thomas and Cervione, 1970; Riggs, 1972; Vogel and Kroll, 1992] and for flood flows [Benson, 1962; Matalas and Gilroy, 1968; Thomas and Benson, 1970; Jennings et al., 1994; Tasker et al., 1996]. Traditionally, the parameters of these models were estimated using ordinary least squares (OLS) regression procedures employing data for gaged river sites [Thomas and Benson, 1970]. For OLS parameter estimators to be efficient the model residuals should be independent and homoscedastic.

Tasker [1980] developed a weighted least squares (WLS) regression technique to account for the varying sampling error in the at-site quantile estimators. Stedinger and Tasker [1985, 1986] extended this work by developing generalized least squares (GLS) regression techniques to account for the varying sampling error and the cross correlation among concurrent flows. Tasker and Stedinger [1989] discuss an implementation of GLS estimators which also accounts for varying model error variance among sites and variations in the cross correlation of concurrent observations. Using Monte Carlo simulation, Stedinger and Tasker [1985] demonstrated that GLS procedures provided more accurate parameter estimators, better estimators of parameter sampling variances, and an almost unbiased estimator of the model error variance. In particular, the average sampling mean square error (mse) of the GLS estimators was smaller than the mse of the OLS estimators when the model error variance was small or the cross correlations among the annual flows were large. Moss and Tasker [1991] showed that GLS procedures describe model accuracy in regional analyses better than OLS procedures do.

This paper addresses a number of unresolved issues regarding GLS and OLS regional regression procedures. The paper (1) clarifies assumptions that have been made when implementing GLS and OLS regional regression procedures, (2) examines the loss of efficiency of OLS estimators and GLS estimators that employ a residual error covariance matrix different than the true residual error covariance matrix, (3) determines whether this loss in efficiency in GLS estimators is due to smoothing of the sampling covariance matrix or to implementing an inadequate model of the model error variance, and (4) determines when OLS estimators are adequate and when a GLS estimator which accounts for varying model error variance is needed to achieve efficient parameter estimates. To examine the efficiency of GLS estimators, the practical GLS estimator developed by Stedinger and Tasker [1985] is compared with an idealized GLS estimator that uses the true regional statistics and the simplifying assumptions Stedinger and Tasker used to construct the residual error covariance matrix, and an ideal GLS estimator that uses the true residual error covariance matrix.
This paper is structured as follows. Section 2 presents the regional regression problem. Section 3 describes Stedinger and Tasker’s [1985] Monte Carlo experiments. Section 4 discusses the construction of the residual error covariance matrices. Section 5 compares the results for one of Stedinger and Tasker’s [1985] Monte Carlo experiments with analytic expressions. Section 6 uses analytic expressions to compare the efficiency of OLS and GLS estimators for several new cases not considered by Stedinger and Tasker [1985, 1986]. Finally, section 7 presents our conclusions.

2. Regional Regression Model

Following the notation by Stedinger and Tasker [1985], let $\theta$ be a vector of the true flow statistics for river sites in a region and let $X$ be a matrix of drainage basin characteristics associated with the sites augmented by a column of ones. Assume that the relationship between $\theta$ and $X$ is described by the linear model

$$\theta = X\beta + \epsilon$$  \hspace{1cm} (1)

where $\beta$ contains model parameters and $\epsilon$ contains the residual errors. Here $\text{var}(\epsilon) = \gamma_i^2$ is the model error variance. In practice the true flow statistic, $\theta$, is not known, and an estimator, $\hat{\theta}$, of the statistic of interest is obtained with available streamflow records. Assume that $\hat{\theta}$ is an unbiased estimator of $\theta$ so that

$$E[\hat{\theta}] = \theta$$  \hspace{1cm} (2)

and

$$E[(\hat{\theta} - \theta)(\hat{\theta} - \theta)^T] = \Sigma$$  \hspace{1cm} (3)

where $\Sigma$ is the sampling covariance matrix associated with the estimator $\hat{\theta}$. The covariance of $\hat{\theta}$ about the regression mean $X\beta$ defines the residual error covariance matrix,

$$E[(\hat{\theta} - X\beta)(\hat{\theta} - X\beta)^T] = \Lambda_r = \Gamma + \Sigma$$  \hspace{1cm} (4)

where $\Gamma = \text{diag} \{\gamma_i^2\}$, and $\gamma_i^2$ is the model error variance associated with site $i$.

In practice, OLS regression procedures have been used to estimate the parameters of linear models such as (1). If $\Lambda_r$ is equal to a diagonal matrix with a constant variance $\gamma^2$ along the diagonal, then OLS parameter estimators are efficient since they have minimum variance among all unbiased linear estimators [Johnston, 1984, p. 173]. In practice the variance, and therefore the diagonal elements of $\Lambda_r$, will differ from site to site. However, the assumption of a constant variance (called homoscedasticity) is often adequate for many practical problems [Draper and Smith, 1981].

The OLS estimator of the parameter vector is [Johnston, 1984, equation 5-26]

$$\hat{\beta}_{OLS} = (X^TX)^{-1}X^T\hat{\theta}$$  \hspace{1cm} (5)

Because the residual errors associated with the model in (1) are not independent and identically distributed with equal variances, the standard relationship for the variance of the OLS parameter estimator [Johnston, 1984, equation 5-33]

$$\text{var}(\hat{\beta}_{OLS}) = \gamma^2(X^TX)^{-1}$$  \hspace{1cm} (6)

does not describe the actual sampling variance of the OLS parameter estimator. Instead the correct expression is [Johnston, 1984, equation 8-13]

$$\text{var}(\hat{\beta}_{OLS}) = (X^TX)^{-1}X^T\Lambda_rX(X^TX)^{-1}$$  \hspace{1cm} (7)

Equation (7) involves $\Lambda_r$, the true residual error covariance matrix.

An extension of OLS parameter estimators are GLS parameter estimators that employ some estimate of $\Lambda_r$. These estimators weight each observation to reflect the variance of the residual error associated with that observation and the covariance of the residual error with other residual errors. If $\Lambda_r$ is known, one can employ the optimal GLS estimator [Johnston, 1984, equation 8-20]

$$\hat{\beta}_{GLS} = (X^T\Lambda_r^{-1}X)^{-1}X^T\Lambda_r^{-1}\hat{\theta}$$  \hspace{1cm} (8)

whose sampling variance is [Johnston, 1984, equation 8-12]

$$\text{var}(\hat{\beta}_{GLS}) = (X^T\Lambda_r^{-1}X)^{-1}$$  \hspace{1cm} (9)

Both $\hat{\beta}_{GLS}$ and $\hat{\beta}_{OLS}$ are unbiased estimators of the parameters in the model given by (1); however, $\hat{\beta}_{GLS}$ has a smaller variance than $\hat{\beta}_{OLS}$ because the former correctly weights each of the observations. The parameter estimator $\hat{\beta}_{GLS}$ is the best (minimum variance) linear unbiased estimator (BLUE) of the model parameters [Greene, 1990; Johnston, 1984].

Stedinger and Tasker [1985] were interested in developing an estimator of the parameters of the model in (1) that had an efficiency approaching that of the GLS estimator $\hat{\beta}_{GLS}$. Employing $\hat{\beta}_{GLS}$ is not practical because $\Lambda_r$ is unknown; $\Lambda_r$ requires knowledge of the model error variance associated with each observation, $\gamma_i^2$, and the sampling covariance matrix, $\Sigma$, both of which need to be estimated. Stedinger and Tasker [1985] proposed constructing an approximation of $\Lambda_r$, which we denote as $\Lambda_E$, by approximating $\Sigma$ by $\text{diag} \{\gamma^2\}$ with a constant value of $\gamma^2$ that must be estimated, and by approximating $\Sigma$ using averaged or smoothed estimates of the sampling variances associated with the at-site streamflow statistics of interest. An average value of the cross correlation between concurrent flows in the region was used to calculate the covariance terms which are the off-diagonal elements of $\Sigma$. In later work a relationship between the cross correlation and the distance between gaged river sites was employed, and $\gamma^2$ was allowed to vary across sites [Tasker and Stedinger, 1989].

If smoothed estimates of the variance of individual streamflow observations are available, and the average cross correlation of concurrent streamflows across sites and the average model error variance are known, one can construct $\Lambda_E$ and compute the corresponding idealized GLS estimator

$$\hat{\beta}_{GLS}^e = (X^T\Lambda_E^{-1}X)^{-1}X^T\Lambda_E^{-1}\hat{\theta}$$  \hspace{1cm} (10)

For particular $X$ and $\Lambda_r$ matrices the sampling variance of this GLS parameter estimator is

$$\text{var}(\hat{\beta}_{GLS}^e) = (X^T\Lambda_E^{-1}X)^{-1}X^T\Lambda_E^{-1}\Lambda_r\Lambda_E^{-1}X(X^T\Lambda_E^{-1}X)^{-1}$$  \hspace{1cm} (11)

While Stedinger and Tasker wanted to employ $\Lambda_E$ as an approximation to $\Lambda_r$, as a practical matter they had to employ an estimator of $\Lambda_E$, which we denote as $\Lambda_M$. Stedinger and Tasker [1985] encountered problems when at-site sample variances were used to construct an estimator of the sampling covariance matrix $\Sigma$, because those weights were then correlated with the at-site quantile estimators, the dependent variable in the regression equation. To avoid such problems, the variance of the individual streamflow observations used to compute the elements of $\Sigma$ was estimated using a regression
relationship developed between at-site sample variances and physiographic basin characteristics. \( \Lambda_{MC} \) was constructed using the computed estimate of the variance of individual observations from this regression relationship, a computed average of the sample cross-correlation estimators, and a computed generalized mean square error estimator of the average model error variance. In Stedinger and Tasker’s Monte Carlo experiment, for every replicate of the experiment a different value of \( \Lambda_{MC} \) was constructed to allow computation of a GLS parameter estimator \( \hat{\beta}_{GLS}^{MC} \) for that replicate. They also employed \( \Lambda_{MC} \) to estimate the variance of the GLS parameter estimator as

\[
\text{var} \left( \hat{\beta}_{GLS}^{MC} \right) = (X^T \Lambda_{MC}^{-1} X)^{-1} \tag{12}
\]

Thus Stedinger and Tasker had two estimators of the variance of their GLS parameter estimator: the observed empirical variability of the individual \( \hat{\beta}_{GLS}^{MC} \) estimators across replicates of the Monte Carlo experiment and the average of (12) across replicates. The observed variability of the individual \( \hat{\beta}_{GLS}^{MC} \) estimators across replicates should correspond to an average of analytic expressions were averaged over 100 replicates, with drainage area at that site, \( A_{\rho} \), following the models

\[
\mu_i = \alpha + \beta \ln (A_{\rho}) + \nu_i \tag{14}
\]

\[
\sigma_i = [\alpha + \beta \ln (A_{\rho})] \exp (\delta_i) \tag{15}
\]

where \( \alpha = 0, \beta = 0.75, \alpha = 1.5, \beta = -0.14, \) and \( \nu_i \) and \( \delta_i \) are independent normally distributed random error terms with means 0 and \(-0.03125 \sigma_i^2 \) and variances \( \sigma_i^2 \) and \( \sigma_i^2 = 0.0625 \sigma_i^2 \), respectively. For each site an annual flow record of length \( n_i \) was randomly generated from a lognormal distribution with moments given by (14) and (15), and the cross correlation among the logarithms of concurrent flows in a region equal to a constant value, \( \rho \). Using the generated record for a site, the sample moment estimators, \( \hat{\mu}_i \) and \( \hat{\sigma}_i \), were calculated and used to compute an at-site estimator of a flow quantile

\[
\hat{Q}_{p,i} = \hat{\mu}_i + z_p \hat{\sigma}_i \tag{16}
\]

where \( \hat{Q}_{p,i} \) is an at-site estimate of a flow quantile with a nonexceedance probability of \( p \), and \( z_p \) is the \( p \)th percentile of a standard normal distribution (for the 50-year flood \( p = 0.98 \) and \( z_p = 2.054 \)). Combining (14), (15), and (16), the underlying regression model is

\[
\hat{Q}_{p,i} = \alpha + \beta \ln (A_{\rho}) + \epsilon_i \tag{17}
\]

where \( \alpha = \alpha + \beta \ln (A_{\rho}) \), \( \beta = \beta + \beta \), and \( \epsilon_i \) is the residual error.

In their first experiment, Stedinger and Tasker considered the cases where \( \rho = 0.0, 0.3, 0.6, \) and 0.9, and \( \sigma_i = 0.0, 0.1, 0.3, 0.5, \) and 0.9. For 10 sites the record length was set to \( n_i = 50 \), for 10 sites \( n_i = 20 \), and for the remaining 10 site \( n_i = 10 \).

### 4. Construction of the Residual Error Covariance Matrix

The residual error covariance matrix constructed in Stedinger and Tasker’s Monte Carlo experiment for their GLS estimator, \( \Lambda_{MC} \), is an estimator of the covariance matrix, \( \Lambda_{MC} \). This section examines the assumptions employed to develop the sampling covariance matrix and model error variance for each of the GLS estimators. Table 1 summarizes those assumptions.

#### 4.1. Estimation of Sampling Covariance Matrix

The diagonal elements of the sampling covariance matrix, \( \Sigma \), correspond to the variance of the at-site quantile estimators, and the off-diagonal elements correspond to the covariance among quantile estimators for different sites. When the individual observations are normally distributed, the correct diagonal elements of the sampling covariance matrix, \( \Sigma_{ii} \), are

\[
\Sigma_{ii} = \text{var} \left( \hat{Q}_{p,i} \right) = \text{var} \left( \hat{\mu}_i + z_p \hat{\sigma}_i \right) = \text{var} \left( \hat{\mu}_i \right) + z_p^2 \text{var} \left( \hat{\sigma}_i \right) = \sigma_i^2 + z_p^2 \alpha_i \tag{18}
\]

Equation (18) incorporates an exact expression for the vari-
Table 1. Alternative Assumptions and Equations Employed to Construct Residual Error Covariance Matrices, \( \Lambda \), for the
GLS Estimators \( \hat{\beta}^{\text{GLS}} \), \( \hat{\beta}^{\text{bGLS}} \), and \( \hat{\beta}^{\text{MC}} \).

<table>
<thead>
<tr>
<th>Assumption/Equation</th>
<th>( \hat{\beta}^{\text{GLS}} (\Lambda^T) )</th>
<th>( \hat{\beta}^{\text{bGLS}} (\Lambda_E) )</th>
<th>( \hat{\beta}^{\text{MC}} (\Lambda_{MC}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model error variance for each parameter estimator</td>
<td>( \Gamma_i = \gamma_i ), true model error variance (equation (32))</td>
<td>( \Gamma_i = \bar{\gamma}_i ), expected model error variance (equation (33))</td>
<td>( \Gamma_i = \bar{\gamma}_i ), estimator of average model error variance for each replicate (equation (34))</td>
</tr>
<tr>
<td>Estimator of variance of observations, ( \sigma^2 ), employed to calculate sampling error associated with at-site quantile estimators, ( \Sigma_u = \text{var} (\theta_i) )</td>
<td>( \sigma^2 ), true value of variance (equation (15))</td>
<td>( \text{E}[\sigma^2] ), expected value of variance (equation (28))</td>
<td>( \text{E}(\sigma_o^2) ), square of estimated expected value of standard deviation (equation (29))</td>
</tr>
<tr>
<td>Formula employed to compute the variance of at-site quantile estimator, ( \Sigma_u = \text{var} (\theta_i) )</td>
<td>exact estimator (equation (18))</td>
<td>first-order approximation (equation (20))</td>
<td>first-order approximation (equation (20))</td>
</tr>
<tr>
<td>Correlation coefficient used in first-order approximation (equation (21)) to compute covariances among at-site quantile estimators, ( \Sigma_o = \text{cov} (\theta_i, \theta_j) )</td>
<td>true value</td>
<td>true value</td>
<td>average regional estimate for each replicate</td>
</tr>
</tbody>
</table>

The variance of the at-site estimator of the standard deviation [David, 1981] and is employed in \( \Lambda_T \). Stedinger and Tasker employed the first-order approximation

\[
\text{var} (\hat{\sigma}) = \frac{\sigma^2_i}{2n_i} \tag{19}
\]

for the variance of the at-site estimator of the standard deviation when constructing the sampling covariance matrix [Stedinger, 1983]. Using this approximation, (18) becomes

\[
\text{var} (\hat{Q}_{ij}) = \frac{\sigma^2_i}{n_i} \left[ 1 + \frac{z_i^2}{2} \right] \tag{20}
\]

Stedinger and Tasker used a smoothed estimator of the at-site variance of the observations, \( \sigma^2_i \), in (20) when constructing their residual covariance, \( \Lambda_E \). The idealized residual error covariance matrix, \( \Lambda_E \), implements (20) with the expected value of \( \sigma^2_i \) for each site \( i \).

Stedinger and Tasker also employed a first-order approximation to compute the covariance among at-site quantile estimators which are the off-diagonal elements of the sampling covariance matrix,

\[
\Sigma_o = \frac{\rho_{ij} m_i \sigma_i \sigma_j}{n_{ij}} \left[ 1 + \rho_{ij}^2 / 2 \right] \quad i \neq j \tag{21}
\]

where \( m_{ij} \) is the number of concurrent years of data at sites \( i \) and \( j \), and \( \rho_{ij} \) is the lag zero correlation coefficient between the annual flows at sites \( i \) and \( j \). The exact expression for the covariance among the at-site quantile estimators is quite complex, so (21) will be employed to compute \( \Lambda_T \) and \( \Lambda_E \) in the calculations reported here. \( \Lambda_T \) and \( \Lambda_E \) will employ the true value of the cross-correlation which was constant across the region in Stedinger and Tasker’s Monte Carlo experiment. The matrix associated with the GLS estimator implemented by Stedinger and Tasker, \( \Lambda_{MC} \), used a regional average of at-site estimators of the correlation coefficient.

In (18) and (20), which describe the sampling variance associated with the at-site quantile estimators, an estimate of the variance of the annual flows, \( \sigma^2_i \), is required. The expected variance of the log-space annual flows is

\[
\text{E}[\sigma^2] = \text{var} [\sigma_i] + (\text{E}[\sigma_i])^2 \tag{22}
\]

The formula for the at-site standard deviation given by (15) assumes that the standard deviation is lognormally distributed. In terms of the logarithm of (15)

\[
\ln (\sigma_i) = \ln [\alpha_o + \beta_o \ln (A_i)] + \delta_i \tag{23}
\]

the log space moments of the standard deviation are

\[
\text{E}[\ln (\sigma_i)] = \ln [\alpha_o + \beta_o \ln (A_i)] + \text{E}[\delta_i] \tag{24}
\]

and

\[
\text{var} [\ln (\sigma_i)] = \text{var} [\delta_i] = \sigma^2_o \tag{25}
\]

The real-space moments for the lognormal distribution can be calculated using the log space moments [Loucks et al., 1981], resulting in

\[
\text{E}[\sigma_i] = \alpha_o + \beta_o \ln (A_i) \tag{26}
\]

and

\[
\text{var} [\sigma_i] = [\alpha_o + \beta_o \ln (A_i)]^2 [\exp (\sigma^2_o) - 1] \tag{27}
\]

Substituting (26) and (27) into (22) yields

\[
\text{E}[\sigma^2] = [\alpha_o + \beta_o \ln (A_i)]^2 \exp (\sigma^2_o) \tag{28}
\]

Equation (28) corresponds to the average variance of the annual flows at a site with drainage area \( A_i \), and is used in the residual error covariance matrix, \( \Lambda_E \). Equation (28) is not the same as the variance of the flows in each replicate because \( \delta_i \) in (15) is replaced by an expectation in (28). In this study a Monte Carlo analysis was employed to account for that difference.

In addition, for every replicate of their Monte Carlo experiment, Stedinger and Tasker randomly generated new drainage areas for the region. To account for the variability due to random drainage areas and the difference between \( \sigma_i \) in (15)
and $E[\sigma^2_i]$ in (28), 100 random sets of different drainage areas and $\delta_i$ values were generated. In the construction of the true residual error covariance matrix, $\Lambda_r$, the random values of $\delta_i$ and $A_i$ were used to calculate the true value of $\sigma_i$ at each site using (15), and this value was used in (18) for constructing one realization of $\Sigma_u$. This allows calculation of the average value over 100 replicates of the sampling variance of $\hat{\beta}_G$ and $\hat{\beta}_G^T$ in (9) and (11), respectively, using generated values of $X$ and associated $\Lambda_r$ and $\Lambda_r'$. Stedinger and Tasker used

$$E(\sigma_i^2) = (\hat{\sigma}_i + \hat{\beta}_i \ln (A_i))^2$$  (29)\

instead of (28) for an estimator of the variance of the observations. This estimator systematically underestimates the variance of the observations, but the bias was small in Stedinger and Tasker’s Monte Carlo experiment because $\sigma^2_\alpha$ was much smaller than 1, and thus var $(\sigma_i)$ was much smaller than $(E[\sigma_i^2])^2$.

4.2. Estimation of Model Error Variance

The residual error covariance matrix also depends upon the model error variance, $\gamma^2_i$, for each site $i$. Assuming the annual flows are lognormally distributed, $\gamma^2_i$ is

$$\gamma^2_i = \text{var} [\ln (Q_i)] = \text{var} [\mu_i + z^* \sigma]$$

$$= \text{var} [\mu_i] + z^* \text{var} [\sigma] + 2 \sigma \text{cov} [\mu_i \sigma]$$  (30)

The regional model adopted in the experiment had $\text{cov} (\mu_i, \sigma) = 0$. From (14)

$$\text{var} [\mu_i] = \text{var} [u] = \sigma^2$$  (31)

The var $(\sigma_i)$ is given in (27). Substituting (27) and (31) into (30) yields

$$\gamma^2_i = \sigma^2 + z^* \exp (\sigma^2) - 1$$  (32)

The model error variance is a function of drainage area and thus varies across sites. Stedinger and Tasker [1985] proposed a GLS estimator that uses an average model error variance. This assumption was relatively good in their Monte Carlo experiment because $\sigma^2_\alpha$ was much smaller than 1, and thus var $(\sigma_i)$ was much smaller than $(E[\sigma_i^2])^2$.

The average value of the model error variance is obtained by taking the expectation over the drainage areas of interest on right hand side of (32) to obtain

$$E[\gamma^2_i] = \sigma^2 + z^* [2 \sigma^2 + 2 \alpha \beta \exp (\sigma^2)] [\exp (\sigma^2) - 1]$$  (33)

Stedinger and Tasker’s experiments included six values of $\sigma_i$ which correspond to $E[\gamma^2_i] = 0.0, 0.011, 0.102, 0.284, 0.557,$ and 0.922. These average values can be used to construct a diagonal model error variance matrix $\Lambda_r$ with constant elements to compute the residual error covariance matrix $\Lambda_r$.

When implementing the GLS estimator in their Monte Carlo experiment, Stedinger and Tasker computed a generalized mean square error estimator of the average model error variance by solving

$$(\hat{\theta} - \hat{X}^T \Lambda_{MC}^{-1} (\hat{\theta} - \hat{X}^T \hat{\beta})) = N - k$$  (34)

for $\hat{\gamma}^2$ where $\Lambda_{MC} = \hat{\gamma}^2 I_N + \hat{\Sigma}(\hat{\theta})$, $N$ is the number of sites, $k$ is the number of degrees of freedom in the model, and $\hat{\theta}$ is the parameter estimator for each replicate. This estimator of the average model error variance is dependent upon the at-site data and thus varies from replicate to replicate because of different sets of flows and drainage areas in each replicate.

5. Comparison of Stedinger and Tasker’s Monte Carlo Results With Analytic Expressions

In their Monte Carlo analysis (experiment 1) Stedinger and Tasker [1985] computed estimates of the variance of the parameter estimators and the average sampling mean square error (mse) of the OLS and GLS quantile estimators. The mse of a quantile estimator is the average of variances of a quantile estimator and its true value at such sites. The mse of unbiased quantile estimators is

$$\text{mse} = E_{\hat{\alpha}, \hat{\beta}} [(\hat{\theta} - \theta)^2] = \text{var} (\hat{\alpha}) + 2E [\ln (A)] \text{cov} (\hat{\alpha}, \hat{\beta})$$

$$+ E [(\ln (A))^2] \text{var} (\hat{\beta})$$  (35)

In this study the Monte Carlo results from Stedinger and Tasker [1985] for the mse of $\hat{\beta}_G$ are compared to calculated values of the mse of $\hat{\beta}_{MG}$ and $\hat{\beta}_{MG}^T$ computed using the average values over 100 replicates of the analytic expressions for the variance of the parameter estimators, (9) and (11), respectively. Record lengths of 10, 20, and 50 were randomly assigned to a third of the sites, as in Stedinger and Tasker’s first experiment.

For this comparison the efficiencies of the estimators $\hat{\beta}_{MG}$ and $\hat{\beta}_{MG}^T$ are computed as

$$\text{Efficiency } \hat{\beta}_{MG} = \frac{\text{mse}[\hat{\beta}_{MG}]}{\text{mse}[\hat{\beta}_{MG}]}$$  (36)

$$\text{Efficiency } \hat{\beta}_{MG}^T = \frac{\text{mse}[\hat{\beta}_{MG}^T]}{\text{mse}[\hat{\beta}_{MG}]}$$

The efficiency of an estimator approaches unity when the mses of an estimator is as small as the mses of $\hat{\beta}_{MG}$ which has the smallest possible mse of a linear unbiased estimator. Figure 1a is a plot of these efficiencies over the range of cross correlations and average model error variances examined in Stedinger and Tasker’s experiment 1. The first set of four columns correspond to cross correlations of 0.0, 0.3, 0.6, and 0.9, respectively, when the average model error variance $\gamma^2 = 0.0$. Other sets of four columns correspond to different values of $\gamma^2$. Figure 1a also contains the efficiencies of the OLS estimator $\hat{\beta}_{MG}$ reported by Stedinger and Tasker.

In general, the efficiency of $\hat{\beta}_{MG}$ is nearly 100% and the efficiency of $\hat{\beta}_{MG}^T$ is greater than 90% for the cases examined. The apparent exception is the efficiency for $\hat{\beta}_{MG}$ when $\rho = 0$ and $\gamma^2 = 0.0$, in which case the computed efficiency is only 80%. The computed mse of $\hat{\beta}_{MG}$ for this case (0.004) is small, and Stedinger and Tasker reported only one significant digit; thus this low efficiency is likely due to rounding error. The efficiency of $\hat{\beta}_{MG}$ based on the reported variance of the individual parameter estimators was approximately 90% [Kroll, 1996], which further confirms that this low efficiency is due to rounding error.

In the experiment reported in Figure 1a the high efficiencies of $\hat{\beta}_{MG}$ indicate that the assumptions Stedinger and Tasker made when simplifying the residual error covariance matrix had relatively little effect on the performance of the estimator.
compared to an estimator based on knowing the true residual error covariance matrix. The high efficiencies of \( \hat{\beta}_{\text{GLS}} \) indicate the practical implementation of Stedinger and Tasker’s proposed GLS estimator \( \hat{\beta}_{\text{GLS}} \) in the form of \( \hat{\beta}_{\text{GLS}} \) resulted in little reduction in the performance of the estimator, especially when \( \gamma^2 \) was greater than 0.1.

Figure 1a also includes the efficiency of the OLS estimator \( \hat{\beta}_{\text{OLS}} \) compared to the best GLS estimator \( \hat{\beta}_{\text{GLS}} \). For large values of \( \gamma^2 \) the efficiency of the OLS estimator is close to the efficiency of \( \hat{\beta}_{\text{GLS}} \). For small \( \gamma^2 \) the efficiency of the OLS estimator drops considerably. For moderate model error variances, as the cross correlation increases, the relative efficiency of the OLS estimator decreases. The efficiency of the OLS estimator depends on whether the elements along the diagonal of \( \Lambda \) are close to constant and on the relative magnitude of the off-diagonal elements of \( \Lambda \) compared to the diagonal elements. If the off-diagonal elements are relatively small and the diagonal elements are close to constant, the OLS estimator is almost as efficient as the GLS estimator. For small \( \gamma^2 \) the effect of heteroscedasticity due to the sampling error in the at-site quantile estimators is greater than when \( \gamma^2 \) is large, and thus the OLS estimator performs worse when \( \gamma^2 \) is small. Kroll [1996] showed that the mses, and variance of the OLS estimator reported in Stedinger and Tasker’s [1985] Monte Carlo analysis, \( \hat{\beta}_{\text{OLS}} \), was nearly identical to the average mses and variance of the OLS estimator using (11), \( \hat{\beta}_{\text{OLS}} \), as they should be.

6. Comparison of OLS and GLS Estimators Using Analytic Expressions

Obtaining the mses of the OLS and GLS quantile estimators requires significantly less effort using the average over 100 replicates of the analytic expressions (equations (7), (9), and (11)), instead of Stedinger and Tasker’s complete Monte Carlo simulation using randomly generated streamflows. Using analytic expressions, we examined how the mses of the OLS parameter estimator \( \hat{\beta}_{\text{OLS}} \) compares to the GLS parameter estimator \( \hat{\beta}_{\text{GLS}} \) for a number of different cases. The previous section demonstrated that \( \hat{\beta}_{\text{GLS}} \) performed almost as well as \( \hat{\beta}_{\text{OLS}} \) in Stedinger and Tasker’s Monte Carlo experiment 1.

In experiment 1 the variance of the residual terms in (14) and (15) were related by \( \sigma^2_\hat{e} = 0.0625\sigma^2_\epsilon \). This relationship determines the variability in the model error variance across sites, given by (32). In experiment 1 the maximum variation in the model error variance, \( \gamma^2 \), from the average model error variance, \( \gamma^2 \), in the region was 25%. In this case the model error variance was relatively constant across sites, and thus \( \hat{\beta}_{\text{GLS}} \), which employs a constant model error variance, performed well compared to \( \hat{\beta}_{\text{GLS}} \). Of interest is the relative performance of \( \hat{\beta}_{\text{GLS}} \) and \( \hat{\beta}_{\text{OLS}} \) when the model error variance has greater variability across sites in a region, as well as the effect return period has on the estimators’ performance.

The case where \( \sigma^2_\hat{e} = \sigma^2_\epsilon \) was examined, which yields a maximum variation of \( \gamma^2 \) from \( \gamma^2 \) of 93%. This corresponds to a realistic but perhaps extreme case wherein the relative precision with which the median flood flow and the log space standard deviation of the flood flows can be estimated by their respective models is roughly the same. The same values of \( \gamma^2 \) and \( \rho \) examined in Stedinger and Tasker’s experiment 1 were examined for a quantile estimator with a 50-year return period. Figure 1b contains the efficiency of \( \hat{\beta}_{\text{OLS}} \) and \( \hat{\beta}_{\text{GLS}} \) relative to \( \hat{\beta}_{\text{OLS}} \) for this case. As \( \sigma^2_\hat{e} \) increases relative to \( \sigma^2_\epsilon \), the model error variance varies more across sites and we observe a drop in the relative efficiency of both \( \hat{\beta}_{\text{GLS}} \) and \( \hat{\beta}_{\text{OLS}} \). The decrease in efficiency of \( \hat{\beta}_{\text{GLS}} \) is larger for cases with larger model error variances, because the heteroscedasticity along the diagonal terms of \( \Lambda \) due to variations in the model error variances is greater for these cases. For the cases presented in Figure 1b the efficiency of \( \hat{\beta}_{\text{GLS}} \) is generally greater than 90%. This result indicates that the assumption of a constant model error variance in regions where the model error variance varies as much as 90% from the average model error variance only produces a 10% drop in the efficiency of \( \hat{\beta}_{\text{GLS}} \) for quantile estimators with a 50-year return period. As the cross correlation increases, the efficiency of \( \hat{\beta}_{\text{GLS}} \) increases, since \( \hat{\beta}_{\text{GLS}} \) correctly describes the cross correlation between the quantile estimators. For these cases one would expect that the efficiency of \( \hat{\beta}_{\text{OLS}} \) to track that of \( \hat{\beta}_{\text{GLS}} \) and we are somewhere between the efficiency of \( \hat{\beta}_{\text{GLS}} \) and the OLS estimator \( \hat{\beta}_{\text{OLS}} \).

Also of interest is the effect of return period on the performance of \( \hat{\beta}_{\text{GLS}} \) and \( \hat{\beta}_{\text{OLS}} \). Figures 2a and 2b contain the efficiency of \( \hat{\beta}_{\text{GLS}} \) and \( \hat{\beta}_{\text{OLS}} \) relative to \( \hat{\beta}_{\text{OLS}} \) for a quantile estimator with a 2-year return period when \( \sigma^2_\hat{e} = 0.0625\sigma^2_\epsilon \) and \( \sigma^2_\hat{e} = \sigma^2_\epsilon \), respectively. With a 2-year return period, \( \gamma^2 \) is 0 in (32), and thus the model error variance is constant across sites.
in a region regardless of the relationship between $\sigma_\delta^2$ and $\sigma_y^2$. In Figure 2b, where $\sigma_\delta^2 = \sigma_y^2$, we see a slight loss in efficiency in $\hat{\beta}_{GLS}^E$ compared to Figure 2a, for which $\sigma_\delta^2 = 0.0625\sigma_y^2$. This loss is due to errors incurred when modeling the sampling covariance matrix in $\hat{\beta}_{GLS}^E$. As $\sigma_\delta^2$ increases relative to $\sigma_y^2$, the variance of the at-site standard deviation increases, which increases the error in modeling the variance of the flows as $E[\sigma_i^2]$ by (28), as opposed to the true value $\sigma_i^2$ in (15). The relatively small loss in efficiency observed in Figure 2b indicates that modeling the variance of the flows as $E[\sigma_i^2]$ produces only minor loss of efficiency in the estimator $\hat{\beta}_{GLS}^E$.

In Figures 2a and 2b we also observe that the loss in efficiency of the OLS estimator $\hat{\beta}_{OLS}^T$ for a 2-year return period is much smaller than when the return period was 50-year (Figures 1a and 1b). This is most dramatic for larger model error variances and smaller cross correlations when the efficiency of $\hat{\beta}_{OLS}^T$ is almost as large as the efficiency of $\hat{\beta}_{GLS}^E$. This is because with a 2-year return period the model error variance is constant across sites, so for larger average model error variances the diagonal term of the true residual error covariance matrix is nearly homoscedastic. The effect of cross correlation on the efficiency of $\hat{\beta}_{OLS}^T$ decreases as the average model error variance increases.

7. Conclusions

Stedinger and Tasker [1985] used Monte Carlo simulation to compare the average sampling mean square error (mse) of ordinary least squares (OLS) and generalized least squares (GLS) quantile estimators in regional hydrologic regression...
analyses. Stedinger and Tasker made a number of simplifying assumptions regarding the structure of the model error variance and the sampling error associated with at-site quantile estimators when constructing their residual error covariance matrix. They used smoothed estimators of the at-site variance of the flows, adopted an average regional cross-correlation, and employed a single average generalized mean square model variance estimator in their ideal residual error covariance matrix, \( \Lambda_E \). In practice, all of these parameters and statistics had to be estimated, so the residual error covariance matrix actually employed, \( \hat{\Lambda}_{\text{MC}} \), differs from \( \Lambda_E \). Both of these matrices are different from the true residual error covariance matrix, \( \Lambda_T \), associated with the underlying model in their Monte Carlo experiment.

The variance of the GLS parameter estimator implemented by Stedinger and Tasker in their Monte Carlo experiment, \( \hat{\beta}_{\text{GLS}}^{MC} \), was compared to the variance of GLS parameter estimators based on the residual error covariance matrices \( \Lambda_E \) and \( \hat{\Lambda}_T \), denoted \( \hat{\beta}_{\text{GLS}}^E \) and \( \hat{\beta}_{\text{GLS}}^T \) respectively. Since the parameter estimator \( \hat{\beta}_{\text{GLS}}^{E} \) is based on \( \Lambda_T \), this estimator is unbiased and has minimum variance among all linear unbiased estimators. Using the average over 100 replicates of the new analytic expressions for the mse, \( \text{mse} \), \( \hat{\beta}_{\text{GLS}}^T \) and \( \hat{\beta}_{\text{GLS}}^{E} \), it was shown that the \( \text{mse} \) of \( \hat{\beta}_{\text{GLS}}^T \) is almost indistinguishable from the \( \text{mse} \) of \( \hat{\beta}_{\text{GLS}}^{E} \) for the cases considered by Stedinger and Tasker. For these cases the approximations employed to obtain a smoothed covariance matrix would result in almost no loss of efficiency. In addition, the parameter estimator implemented by Stedinger and Tasker, \( \hat{\beta}_{\text{GLS}}^{MC} \), had an mse almost as small as \( \hat{\beta}_{\text{GLS}}^T \). In most cases the difference was less than 10%. Thus for this case the difference in efficiency between \( \hat{\beta}_{\text{GLS}}^T \) and \( \hat{\beta}_{\text{GLS}}^{MC} \) is relatively small.

The \( \text{mse} \) of the OLS estimator from Stedinger and Tasker’s [1985] Monte Carlo experiment, \( \hat{\beta}_{\text{OLS}}^{MC} \), was also compared to the GLS estimators. For large model error variances, the efficiency of the OLS estimator is close to the efficiency of \( \hat{\beta}_{\text{OLS}}^{MC} \), but for a small model error variance the efficiency of the OLS estimator drops considerably. For moderate model error variances, the relative efficiency of the OLS estimator decreases as the cross correlation increases.

Using analytic expressions for the \( \text{mse} \), the performance of \( \hat{\beta}_{\text{GLS}}^T \) and \( \hat{\beta}_{\text{GLS}}^{E} \) relative to \( \hat{\beta}_{\text{GLS}}^{MC} \) were compared for a number of cases not considered by Stedinger and Tasker [1985]. In particular, a model with a more heteroscedastic model error variance was considered for return periods of 2, 50, and 100 years. For models where the model error variance varied considerably across sites, some loss in efficiency in \( \hat{\beta}_{\text{GLS}}^{E} \) was observed (up to 20%). It was shown that the loss in efficiency of \( \hat{\beta}_{\text{GLS}}^T \) resulted from using an average model error variance and not from smoothing the sampling covariance matrix. The loss of efficiency was particularly apparent for large return periods and moderate to large average model error variances. In this case a GLS estimator which accounts for varying model error variance such as that developed by Tasker and Stedinger [1989] should be implemented. They used a three-parameter error model to account for correlation between the log space means and standard deviations of the flood flows.

Overall, for a small return period and moderate to large model error variance, the OLS estimator \( \hat{\beta}_{\text{OLS}} \) performed nearly as well as \( \hat{\beta}_{\text{GLS}}^{MC} \), especially when the cross correlation of the flows was small. In this case an OLS estimator, which is much easier to implement than a GLS estimator, could be implemented with little or no loss in efficiency. One should note that the efficiency of the \( \beta \) estimators is only one of the advantages of GLS procedures: Stedinger and Tasker [1985] observe that GLS estimators also provide more accurate estimators of model error variances and the precision of estimated parameters than do OLS analyses.

Acknowledgment. This material is based upon work partially supported by the Cooperative State Research Service, USDA, under project 97CRMS06102.

References

C. N. Kroll, Environmental Resources and Forest Engineering, College of Environmental Science and Forestry, State University of New York, Syracuse, NY 13210-2778. (e-mail: cnkroll@mailbox.syr.edu)
J. R. Stedinger, School of Civil and Engineering, Cornell University, Ithaca, NY 14850-3501.

(Received August 20, 1996; accepted September 22, 1997.)