WATER RESOURCES RESEARCH, VOL. 21, NO. 9, PAGES 1421–1432, SEPTEMBER 1985

Regional Hydrologic Analysis
1. Ordinary, Weighted, and Generalized Least Squares Compared

JERY R. STEDINGER[1] AND GARY D. TASKER

U.S. Geological Survey, Reston, Virginia

Streamflow gaging networks provide hydrologic information which is often used to derive relationships between physiographic variables and streamflow statistics. This paper compares the performance of ordinary, weighted, and generalized least squares estimators of the parameters of such regional hydrologic relationships in situations where the available streamflow records at gaged sites can be of different and widely varying lengths and concurrent flows at different sites are cross-correlated. A Monte Carlo study illustrates the performance of an ordinary least squares (OLS) procedure and an operational generalized least squares (GLS) procedure which accounts for and directly estimates the precision of the predictive model being fit. The GLS procedure provided (1) more accurate parameter estimates, (2) better estimates of the accuracy with which the regression model's parameters were being estimated, and (3) almost unbiased estimates of the model error. The OLS approach can provide very distorted estimates of the model's predictive precision (model error) and the precision with which the regression model's parameters are being estimated. A weighted least squares procedure which neglects the cross correlations among concurrent flows does as well as the GLS procedure when the cross correlation among concurrent flows is relatively modest. The Monte Carlo examples also explore the value of streamflow records of different lengths in regionalization studies.

INTRODUCTION

Hydrologic data collection networks provide regional-hydrologic information which is often used to derive relationships between relevant physiographic variables, such as drainage area and channel slope, and streamflow characteristics (see, for example, *Thomas and Benson* [1970]). These regression relationships can be used to estimate the 50- or 100-year flood flow at ungaged sites when such statistics are needed for water resources and flood plain planning. They also serve to provide prior probability distributions based on regional information for use in Bayesian analyses aimed at deriving the posterior flood risk distribution at a site (see *Stedinger* [1983*b*], *Kuczera* [1983], and references therein). Traditionally, the parameters of these regression models have been estimated using ordinary least squares (OLS). However, these regionalization problems with hydrologic data violate the commonly made assumptions that the residual errors associated with the individual observations are homoscedastic and independently distributed; in the case of regional hydrologic data, variations in the length of the available streamflow records and cross correlations among concurrent flows result in estimates of the 50-year or 100-year flood and other flow statistics which are cross-correlated and of varying precision. *Matalas and Benson* [1961], *Matalas and Gilroy* [1968], *Hardison* [1971], *Moss and Karlinger* [1974], *Tasker and Moss* [1979], and *Moss* [1976, 1979] have all examined the statistical precision and properties of OLS procedures with hydrologic data sets.

What has received relatively little attention is how best to estimate the parameters of regional-hydrologic relationships given that OLS procedures will not identify the most efficient estimates of a regression model's parameters when the residual errors are not homoscedastic and independently distributed. Moreover, as is shown by the studies cited above, an OLS procedure's estimates of the standard error of prediction and the estimated precision of the estimated parameters can be highly biased. Good estimates of the precision of general hydrologic relationships are an essential component of procedures which attempt to optimally combine regional and at-site information to obtain the best possible estimators of flood quantiles at gaged sites (see, for example, *Kuczera* [1982*a*, *b*] and references in the work by *Stedinger* [1983*b*]). In fact, if OLS procedures with regional hydrologic data sets correctly described the standard error of prediction and the precision of the estimated parameters, there would be no need for the complex Network Analysis for Regional Information procedure developed by the U.S. Geological Survey [*Moss and Karlinger*, 1974; *Tasker and Moss*, 1979].

Weighted and generalized least squares techniques were developed to deal with situations like those encountered in hydrology where a regression model's residuals are heteroscedastic and perhaps cross-correlated [*Draper and Smith*, 1981; *Johnston*, 1972]. *Tasker* [1980] has, in fact, used a weighted least squares procedure to account for unequal record lengths. *Marin* [1983] and *Kuczera* [1982*a*, *b*, 1983] developed Bayesian and empirical Bayesian methodologies which deal with some of these issues. In particular, *Kuczera* [1983] presented a Bayesian analysis of a generalized least squares model as part of a general scheme for deriving the posterior flood frequency distribution at sites with short streamflow records.

An obstacle to the use of weighted least squares (WLS) and generalized least squares (GLS) procedures with hydrologic data is the need to provide an estimate of the covariance matrix of the residual errors; that covariance matrix is a function of the precision with which the true model can predict the values of the streamflow statistic of concern as well as the sampling error in the available estimates of that statistic. The discussions and examples in the works by *Tasker* [1980] and *Kuczera* [1983] illustrate difficulties associated with estimation of this matrix. These problems are addressed here and procedures are developed for estimating the precision of the underlying regression model as well as the sampling variability of streamflow statistic estimators and their cross corre-

[1]Now at the Department of Environmental Engineering, Cornell University, Ithaca, New York.

Paper number 5W0476.

lation for use in WLS and GLS algorithms. A Monte Carlo analysis with synthetically generated flow sequences allows a comparison of the performance of the OLS procedure with that of a GLS procedure. In situations where the available streamflow records at gaged sites are of different and widely varying length and concurrent flows at different sites are cross-correlated, the GLS procedure provided more accurate parameter estimates, better estimates of the accuracy with which the regression model's parameters were being estimated, and almost unbiased estimates of the variance of the underlying regression model's residual errors. A simpler WLS procedure neglects the cross correlations among concurrent flows. The WLS algorithm is shown to do as well as the GLS procedure when the cross correlation among concurrent flows are relatively modest. Finally, examples illustrate the relative value of flow records of various lengths in regional regression analyses.

BASIC PROBLEM

The regional hydrologic regression problem can be described as follows. At each site i ($i = 1, \cdots, N$) one has available an n_i-year flow record $\{x_{i1}, \cdots, x_{in}\}$. The observed flows at each site are assumed to be a sample of independent random variables corresponding to annual maximum floods, annual streamflow volumes, seasonal flow volumes, or some transformation thereof. Flows are assumed to be temporally independent resulting in independent identically distributed observations at each site; however, concurrent observations at different sites can be cross-correlated. This corresponds to the spatial correlation among concurrent events so often observed in hydrology.

With each flow record one can calculate the sample mean $\bar{x}_i$ and the unbiased sample variance s_i^2; these are estimators of the mean and variance of the flows at that site, respectively. The term "flow" is used here to describe the x variable even if it is a nonlinear normalizing transformation of the actual flow.

Let $\hat{\theta}_i$ denote either an estimator of the mean flow $\bar{x}_i$, of the standard deviation s_i, or of the 100p percentile of the x distribution $\bar{x} + K_p s_i$, where K_p is the frequency factor associated with the 100p percentile. Then $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \cdots, \hat{\theta}_N)^T$ is the vector of $\boldsymbol{\theta}$ estimators. Our task is to use $\hat{\boldsymbol{\theta}}$ to derive a relationship that specifies the expected value of $\hat{\boldsymbol{\theta}}$ as a function of relevant physiographic parameters and other basin characteristics at the sites of interest. Common explanatory variables include drainage area, channel slope, area in lakes, forest cover, and precipitation amounts.

Let $\boldsymbol{\theta}$ be the vector of true values of the flow characteristics of interest: either the at-site means $(\mu_1, \cdots, \mu_N)^T$, standard deviations $(\sigma_1, \cdots, \sigma_N)^T$, or 100p percentiles $(\mu_1 + K_p\sigma_1, \cdots, \mu_N + K_p\sigma_N)^T$. We assume that $\hat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}$, so that

$$E[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta} \tag{1}$$

with

$$E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] = \Sigma(\hat{\boldsymbol{\theta}})$$

Here $\Sigma(\hat{\boldsymbol{\theta}})$ is the sampling covariance matrix of the vector estimator $\hat{\boldsymbol{\theta}}$.

When $\bar{x}$ and s^2 are uncorrelated, as occurs if the distribution of x is symmetric, then the elements of the sampling covariance matrix for $\hat{\theta}_i = \bar{x} + K_p s_i$ can be approximated by

$$\Sigma(\hat{\boldsymbol{\theta}})_{ij} = \sigma_i^2\left[1 + K_p^2\left(\frac{\kappa - 1}{4}\right)\right]\Big/ n_i \qquad i = j \tag{3}$$

$$\Sigma(\hat{\boldsymbol{\theta}})_{ij} = \frac{\rho_{ij} m_{ij} \sigma_i \sigma_j}{n_i n_j}\left[1 + \rho_{ij} K_p^2\left(\frac{\kappa - 1}{4}\right)\right] \qquad i \neq j \tag{4}$$

where

$$\rho_{ij} = E[(x_{it} - \mu_i)(x_{jt} - \mu_j)]/\sigma_i\sigma_j$$

κ is the kurtosis of the x distribution, and m_{ij} is the number of concurrent observations at sites $_i$ and j. For normally distributed x, (3) and (4) become

$$\Sigma(\hat{\boldsymbol{\theta}})_{ij} = \sigma_i^2[1 + z_p^2/2]/n_i \qquad i = j \tag{5}$$

$$\Sigma(\hat{\boldsymbol{\theta}})_{ij} = \frac{\rho_{ij} m_{ij} \sigma_i \sigma_i}{n_i n_j}[1 + \rho_{ij} z_p^2/2] \qquad i \neq j \tag{6}$$

where z_p is the 100p percentile of a standard normal distribution [*Moss*, 1973; *Stedinger*, 1983*a*].

Given $\hat{\boldsymbol{\theta}}$ and $\Sigma(\hat{\boldsymbol{\theta}})$, or some estimate thereof, the goal is to estimate the parameters of a regional regression model. Assume that the individual components of i are in expectation linearly related to a set of basin characteristics. For ease of notation and without loss of generality, let the basin characteristics which might be employed be represented by just the logarithm of the drainage area, $\ln A_i$. Our model is then

$$\theta_i = \alpha + \beta \ln A_i + \varepsilon_i \tag{7}$$

where the ε_i are normal and independently distributed with mean zero and variance γ^2. Thus γ^2 is the model error variance or residual unexplained variance. In matrix notation, (7) can be written

$$\boldsymbol{\theta} = \Xi\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{8}$$

with

$$\Xi = \begin{bmatrix} 1 & \ln A_1 \\ \vdots & \vdots \\ 1 & \ln A_N \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \alpha \\ \beta \end{bmatrix} \qquad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

Combining (1) and (8) yields

$$E[\hat{\boldsymbol{\theta}}] = \Xi\boldsymbol{\beta} \tag{9}$$

where the covariance of $\hat{\boldsymbol{\theta}}$ about $\Xi\boldsymbol{\beta}$ is now

$$E[(\hat{\boldsymbol{\theta}} - \Xi\boldsymbol{\beta})(\hat{\boldsymbol{\theta}} - \Xi\mathbf{B})^T] = \Lambda(\gamma^2) = \gamma^2 I_N + \Sigma(\hat{\boldsymbol{\theta}}) \tag{10}$$

given the sampling covariance $\Sigma(\hat{\boldsymbol{\theta}})$ of $\hat{\boldsymbol{\theta}}$ about the unknown parameter vector $\boldsymbol{\theta}$ and the covariance of $\boldsymbol{\theta}$ about its mean $\Xi\boldsymbol{\beta}$, which is $\gamma^2 I_N$, where I_N is the $N \times N$ identity matrix.

Were $\Lambda(\gamma^2)$ in (10) known, the best linear unbiased estimator of $\boldsymbol{\beta}$ would be the generalized least squares estimator

$$\hat{\boldsymbol{\beta}}_{GLS} = [\Xi^T\Lambda(\gamma^2)^{-1}\Xi]^{-1}\Xi^T\Lambda(\gamma^2)^{-1}\hat{\boldsymbol{\theta}} \tag{11}$$

[*Johnston*, 1972]. $\hat{\boldsymbol{\beta}}_{GLS}$ has covariance

$$\Sigma[\hat{\boldsymbol{\beta}}_{GLS}] = [\Xi^T\Lambda(\gamma^2)^{-1}\Xi]^{-1} \tag{12}$$

if $\Lambda(\gamma^2)$ is indeed $E[(\hat{\boldsymbol{\theta}} - \Xi\boldsymbol{\beta})(\hat{\boldsymbol{\theta}} - \Xi\boldsymbol{\beta})^T]$.

Unfortunately, γ^2 will not in general be known so that an estimator must be employed. If $\Lambda(\gamma^2)$ were known, then

$$E[(\hat{\boldsymbol{\theta}} - \Xi\hat{\boldsymbol{\beta}})^T\Lambda(\gamma^2)^{-1}(\hat{\boldsymbol{\theta}} - \Xi\hat{\boldsymbol{\beta}})] = N - k \tag{13}$$

where N and k are the dimension of $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\beta}$, and where $\hat{\boldsymbol{\beta}}$ is given by (11) [*Johnston*, 1972, p. 210]. Our method-of-moments or generalized mean-square error estimator of γ^2 is obtained by solution of

$$(\hat{\boldsymbol{\theta}} - \Xi\hat{\boldsymbol{\beta}})^T[\hat{\gamma}_{GLS}^2 I_N + \hat{\Sigma}(\hat{\boldsymbol{\theta}})]^{-1}(\hat{\boldsymbol{\theta}} - \Xi\hat{\boldsymbol{\beta}}) = N - k \tag{14}$$

where $\hat{\Sigma}(\hat{\theta})$ is a reasonable estimate of $\Sigma(\hat{\theta})$. In general, the left-hand side of (14) is a decreasing function of $\hat{\gamma}_{GLS}^2$, making it relatively easy to find a positive solution to that equation if one exists. However, occasionally one finds that

$$(\hat{\theta} - \Xi\beta)^T[\hat{\Sigma}(\hat{\theta})]^{-1}(\hat{\theta} - \Xi\beta) < N - k \quad (15)$$

In these instances the sampling covariance matrix $\hat{\Sigma}(\hat{\theta})$ more than accounts for the observed differences between $\hat{\theta}$ and $\Xi\beta$ and no positive value of $\hat{\gamma}_{GLS}$ will satisfy (14). Such circumstances often arose in the Monte Carlo experiments described below when the true value of γ^2 was zero or at least small compared to $\Sigma(\hat{\theta})$. In these instances, $\hat{\gamma}_{GLS}$ is taken to be zero.

The Monte Carlo experiments will also evaluate the performance of a WLS regression procedure. In that algorithm, $\hat{\gamma}_{WLS}^2$ is obtained in the same fashion as $\hat{\gamma}_{GLS}^2$ with the additional assumption or constraint that $\rho_{ij} = 0$ so that $\hat{\Lambda}(\hat{\gamma}_{WLS}^2)$ in (14) is diagonal. Use of (14) to estimate γ^2 is completely analogous to the residual mean square error calculated by ordinary linear regression procedures and would reduce to that value were $\hat{\Sigma}(\hat{\theta}) = 0$.

Finally, the ordinary least squares β estimator can be obtained by replacing $\Lambda(\gamma^2)$ by $\gamma^2 I_N$. Then the best linear unbiased estimator of β becomes

$$\beta_{OLS} = (\Xi^T\Xi)^{-1}\Xi^T\hat{\theta} \quad (16)$$

with covariance

$$\Sigma[\hat{\beta}_{OLS}] = \gamma^2(\Xi^T\Xi)^{-1} \quad (17)$$

if $\Lambda(\gamma^2)$ is indeed $\gamma^2 I_N$.

Monte Carlo Experiments and Results

Here we report the results of coupling implementations of the OLS, WLS, and GLS β estimators with a multivariate stochastic streamflow generator to evaluate the relative performance of the three parameter estimation procedures. The four experiments reported illustrate the potential benefits from use of the GLS procedure in regional hydrologic network analysis.

For these experiments it is necessary to make some assumptions about the true underlying regression model. The generated x values are normally distributed with a mean and standard deviation which were themselves generated randomly using the regional models

$$\mu_i = \alpha_\mu + \beta_\mu \ln A_i + \varepsilon_i \quad (18)$$

$$\sigma_i = [\alpha_\sigma + \beta_\sigma \ln A_i] \exp(\delta_i) \quad (19)$$

with $\varepsilon \sim \text{NID}(0, \sigma_\varepsilon^2)$ and $\delta \sim \text{NID}(-0.5\sigma_\delta^2, \sigma_\delta^2)$, where NID means normal and independently distributed with the specified mean and variance.

In these experiments $\alpha_\mu = 0$, $\beta_\mu = 0.75$, $\alpha_\sigma = 1.5$, and $\beta_\sigma = -0.14$. The logarithms of the drainage area, $\ln A_i$, were drawn randomly from a uniform distribution ranging from ln 10 to ln 20,000. Synthetic streamflows were generated with a constant cross correlation ρ between all concurrent flows. Additional information on how and why these underlying models were selected is provided in Appendix A.

Reasonable values of σ_ε for the μ model in (18) and σ_δ for the σ model in (19) will also be needed. The appropriate values of these variances depends on whether one is considering annual flow volumes, monthly flow volumes, or maximum annual floods. The Network Analysis for Regional Information (NARI) model [*Moss and Karlinger*, 1973; *Tasker and Moss*, 1979; *Moss*, 1976, 1979, 1982] could provide rough estimates of the values of these statistics. *Thomas and Benson* [1970] and others also report the standard error of prediction from their ordinary least squares analyses; however, their values can be inflated because they include both the standard error of prediction and the time sampling error Var $[\hat{\theta}_i]$ of the at-site estimators [*Hardison*, 1971]. From these studies one can infer that for the mean annual flow, the standard error of prediction is likely to be in the range 10–25%. For maximum annual flood quantiles, the standard error of prediction may be 30–90%; in any region, these values seem to be fairly independent of p. Based on these observations, σ_ε was varied over the range [0, 1], with $\sigma_\delta = \sigma_\varepsilon/4$. This implies, as generally seems to be the case, that regional models can estimate σ more accurately than μ.

For the WLS and GLS estimators, an estimate of $\Sigma(\hat{\theta})$ is needed that is independent, or nearly so, of $\hat{\theta}$. Such an estimate was constructed by replacing each σ_i in (3)–(6) by

$$\hat{\sigma}(A_i) = \hat{\alpha}_\sigma + \hat{\beta}_\sigma \ln A_i \quad (20)$$

where $\hat{\alpha}_\sigma$ and $\hat{\beta}_\sigma$ are sample regression coefficients determined by regressing sample standard deviations, s_i, against $\ln A_i$ using a three-step ordinary-weighted least squares procedure (details are given in Appendix B).

The generalized least squares algorithm also requires estimates of the cross correlations ρ_{ij} as a function of the distance between the drainage basins, their size, geology, soils, and other characteristics of the streamflow network, both basins and the region's hydrology, and storm patterns. In some cases, the concurrent record lengths may be sufficient to justify use of sample cross-correlation estimators, perhaps in combination with predictions based on physiographic and other characteristics of the basins. While in practice the best estimators of ρ_{ij} for a particular basin and a given data set should be developed by experienced hydrologists, it seems unnecessary to encumber this initial investigation with such refinements.

In the Monte Carlo experiments to follow, the cross correlation ρ_{ij} between any two flows generated in the same year will have a common value ρ. Thus the natural estimator of the common correlation ρ for use in this study is

$$\hat{\rho} = f \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left[\frac{\min(n_i, n_j)}{\sum_{t=1} (x_{it} - \bar{x}_i)(x_{jt} - \bar{x}_j)/(s_i s_j)} \right] \Big/ M$$

$$M = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \min(n_i, n_j) \quad (21)$$

$$f = [n_i/(n_i - 1)][n_j/(n_j - 1)]$$

This is the average of the sample cross-correlation coefficients weighted by the concurrent sample sizes.

Note that (21) is not a bad estimator of ρ_{ij} if the true cross correlations do not exhibit marked variation with i and j. A generalized least squares procedure based upon use of $\hat{\rho}$, the average cross-correlation of the observed flows, should, in general, be more reasonable than a weighted least squares procedure which corresponds to use of $\hat{\rho} = 0$ for all i and j.

In the Monte Carlo experiments, negative ρ estimators are set to zero reflecting an absolute prior belief that $\rho \geq 0$. Also $\hat{\rho}$ is constrained to be less than 0.99 to eliminate numerical difficulties experienced with the inversion of $\Lambda(\gamma^2)$.

Because of the selected σ model (see equation (19)), the accuracy with which the true model can estimate either σ_i or $\mu_i + z_p\sigma_i$ will actually depend upon $\ln A_i$. In particular,

$$\gamma^2(A_i) = \text{Var}[\mu_i + z_p\sigma_i] = \sigma_\varepsilon^2 + z_p^2\sigma^2(A_i^2)[\exp(\sigma_\delta^2) - 1] \quad (22)$$

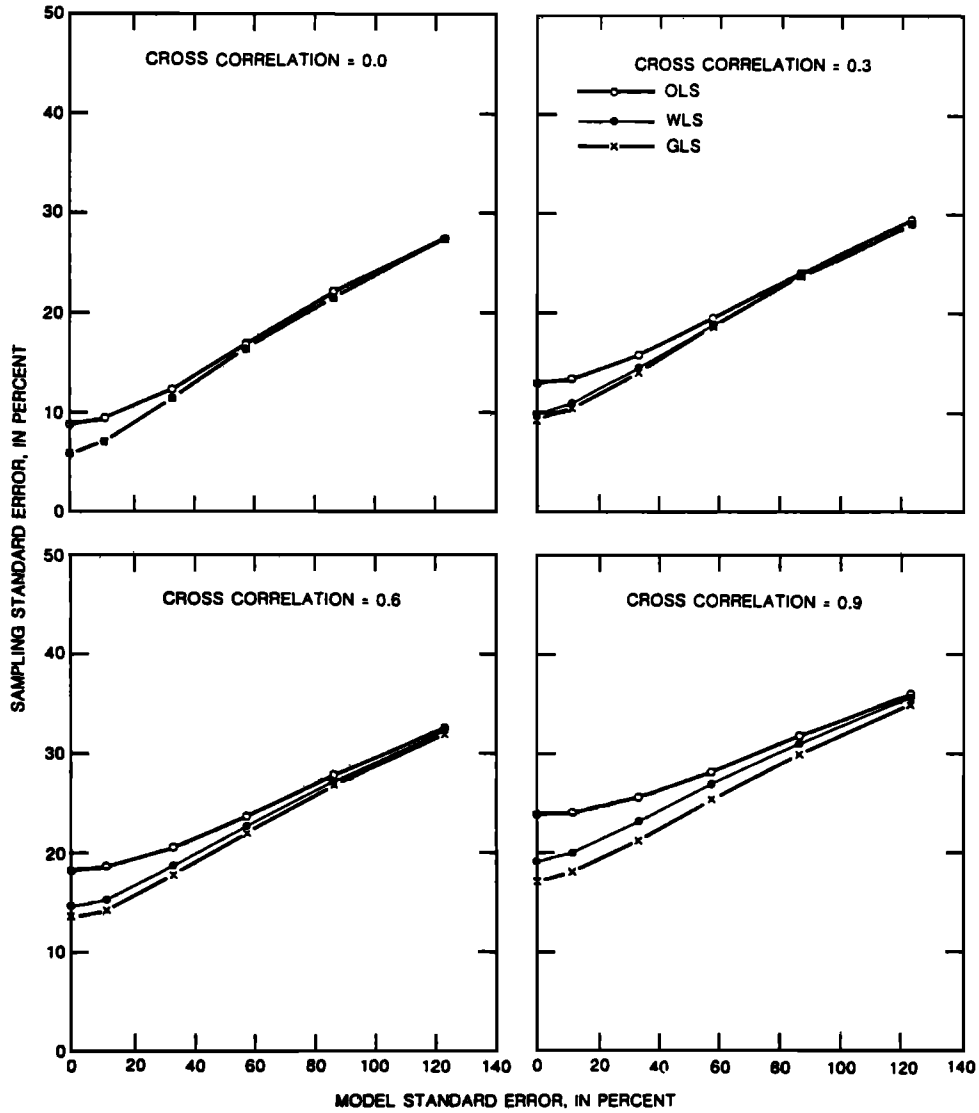

Fig. 1. Monte Carlo results from experiment 1 show prediction accuracy obtained with GLS, WLS, and OLS routines. The regression estimators of the 50-year peak were obtained with flow data from 37 sites: 10 sites with 50 years of data, 10 sites with 20 years of data, and 10 sites with 10 years of data.

Thus the residuals of the true regression for the 100p percentile of the x distribution as a function of ln A_i are really heteroscedastic. However, for $\sigma_\delta = 0.25\sigma_\varepsilon$, the departures from homoscedasticity were modest for $0.01 \leq p \leq 0.99$. The WLS and GLS algorithms provide estimates of the average model error variance

$$\bar{\gamma}^2 = E[\gamma^2(A)] = \sigma_\varepsilon{}^2 + z_p{}^2 E[\sigma^2(A)][\exp(\sigma_\delta{}^2) - 1] \quad (23)$$

where

$$E[\sigma^2(A)] = (\alpha_\sigma + \beta_\sigma E[\ln A])^2 + \beta_\sigma{}^2 \operatorname{Var}[\ln A] \quad (24)$$

The OLS procedure's residual mean square error corresponds to $\bar{\gamma}^2$ plus an average of the sampling variance of the estimators $\hat{\theta}_i$ of each θ_i.

A reasonable criterion for comparing the performance of OLS, WLS, and GLS estimators is their average "sampling" mean square error. Let $\tilde{\theta} = \mu + z_p\sigma$ be the 100p percentile of the x distribution at a randomly selected site; the tilde is added to emphasize that $\tilde{\theta}$ is a random variable. In addition, let the conditional mean of $\tilde{\theta}$, given that a site's drainage area is A, be

$$\theta(A) = E[\tilde{\theta} \mid A]$$

The sampling mean square error is defined as the average over A of the mean square error with which the conditional mean of $\tilde{\theta}$, $\theta(A)$, would be approximated by the estimated regression relationship. In particular, let $\hat{\theta}(A)$ be the estimator of $\theta(A)$ obtained with either OLS, WLS, or GLS. Then the average sampling mean square error is

$$\begin{aligned} \text{mse}_S &= E_{A,\hat{\alpha},\hat{\beta}}\{[\theta(A) - \hat{\theta}(A)]^2\} \\ &= \operatorname{Var}(\hat{\alpha}) + 2\,E[\ln A]\operatorname{Cov}[\hat{\alpha}, \hat{\beta}] + E[(\ln A)^2]\operatorname{Var}(\hat{\beta}) \end{aligned} \quad (25)$$

Note that the expectation is taken over both the sampling distribution of $\hat{\alpha}$ and $\hat{\beta}$ as well as over the distribution of A.

With these definitions, the average mean square with which $\tilde{\theta}$ can be estimated can be written as

$$\begin{aligned} (\text{mse of prediction}) &= E\{_{A,\hat{\alpha},\hat{\beta}}\{E_{\tilde{\theta}|A}[[\tilde{\theta} - \hat{\theta}(A)]^2]\}\} \\ &= E_A\{E_{\tilde{\theta}|A}[[\tilde{\theta} - \theta(A)]^2]\} \\ &\quad + E_{A,\hat{\alpha},\hat{\beta}}\{[\theta(A) - \hat{\theta}(A)]^2\} \\ &= \bar{\gamma}^2 + \text{mse}_S \end{aligned} \quad (26)$$

TABLE 1. Partial Results for Experiment 1

ρ	Model Error Variance	Method	Mean of Estimated Model Error Variance	Standard Deviation of Estimated Model Error Variance	Sampling Mean Square Error, mse_s	Variance of $\hat{\alpha}$	Mean of Predicted Variance of $\hat{\alpha}$	Variance of $\hat{\beta}$	Mean of Predicted Variance of $\hat{\beta}$
0	0	OLS	0.094	0.040	0.008	0.050	0.028	0.00084	0.00069
		WLS	0.005	0.010	0.004	0.020	0.021	0.00028	0.00032
		GLS	0.005	0.010	0.004	0.020	0.022	0.00028	0.00033
	0.011	OLS	0.104	0.042	0.009	0.055	0.032	0.00093	0.00077
		WLS	0.015	0.015	0.005	0.028	0.027	0.00044	0.00045
		GLS	0.015	0.015	0.005	0.028	0.027	0.00044	0.00045
	0.102	OLS	0.194	0.063	0.015	0.088	0.060	0.00165	0.00143
		WLS	0.104	0.045	0.013	0.069	0.063	0.00130	0.00127
		GLS	0.104	0.045	0.013	0.069	0.063	0.00130	0.00127
	0.284	OLS	0.374	0.110	0.028	0.152	0.115	0.00307	0.00276
		WLS	0.284	0.098	0.027	0.139	0.123	0.00282	0.00269
		GLS	0.284	0.098	0.027	0.139	0.124	0.00282	0.00269
	0.557	OLS	0.645	0.184	0.048	0.248	0.200	0.00520	0.00476
		WLS	0.554	0.175	0.046	0.239	0.210	0.00501	0.00474
		GLS	0.554	0.175	0.046	0.239	0.210	0.00501	0.00474
	0.922	OLS	1.008	0.284	0.073	0.375	0.312	0.00804	0.00745
		WLS	0.917	0.276	0.073	0.368	0.324	0.00789	0.00744
		GLS	0.917	0.276	0.073	0.368	0.324	0.00789	0.00745
0.3	0	OLS	0.083	0.035	0.017	0.088	0.025	0.00112	0.00061
		WLS	0.003	0.007	0.010	0.048	0.020	0.00051	0.00029
		GLS	0.005	0.008	0.009	0.044	0.043	0.00048	0.00049
	0.011	OLS	0.094	0.037	0.018	0.093	0.029	0.00122	0.00069
		WLS	0.011	0.012	0.012	0.057	0.025	0.00067	0.00041
		GLS	0.015	0.013	0.011	0.053	0.049	0.00064	0.00062
	0.102	OLS	0.183	0.057	0.025	0.125	0.056	0.00194	0.00135
		WLS	0.097	0.042	0.021	0.100	0.060	0.00154	0.00121
		GLS	0.103	0.043	0.020	0.097	0.089	0.00151	0.00146
	0.284	OLS	0.363	0.105	0.038	0.189	0.112	0.00337	0.00268
		WLS	0.274	0.094	0.035	0.173	0.120	0.00308	0.00262
		GLS	0.282	0.095	0.035	0.170	0.152	0.00306	0.00290
	0.557	OLS	0.635	0.179	0.057	0.285	0.196	0.00551	0.00468
		WLS	0.544	0.171	0.056	0.273	0.207	0.00530	0.00466
		GLS	0.553	0.171	0.056	0.271	0.241	0.00527	0.00497
	0.922	OLS	0.998	0.280	0.084	0.412	0.309	0.00837	0.00737
		WLS	0.906	0.272	0.082	0.403	0.321	0.00819	0.00737
		GLS	0.915	0.273	0.082	0.402	0.357	0.00817	0.00769
0.6	0	OLS	0.065	0.030	0.033	0.158	0.020	0.00167	0.00048
		WLS	0.001	0.005	0.021	0.097	0.019	0.00092	0.00026
		GLS	0.004	0.006	0.018	0.083	0.076	0.00079	0.00074
	0.011	OLS	0.076	0.031	0.034	0.161	0.023	0.00174	0.00055
		WLS	0.006	0.009	0.023	0.107	0.022	0.00110	0.00034
		GLS	0.014	0.011	0.020	0.092	0.083	0.00095	0.00088
	0.102	OLS	0.165	0.050	0.041	0.192	0.051	0.00242	0.00121
		WLS	0.085	0.038	0.034	0.155	0.056	0.00196	0.00111
		GLS	0.102	0.039	0.031	0.140	0.127	0.00182	0.00173
	0.284	OLS	0.345	0.098	0.054	0.253	0.107	0.00381	0.00255
		WLS	0.260	0.089	0.050	0.230	0.116	0.00349	0.00250
		GLS	0.281	0.090	0.047	0.217	0.197	0.00337	0.00322
	0.557	OLS	0.616	0.172	0.074	0.347	0.191	0.00591	0.00455
		WLS	0.528	0.164	0.071	0.331	0.202	0.00568	0.00454
		GLS	0.552	0.165	0.069	0.321	0.292	0.00558	0.00533
	0.922	OLS	0.979	0.272	0.100	0.472	0.303	0.00873	0.00723
		WLS	0.889	0.265	0.099	0.461	0.316	0.00854	0.00724
		GLS	0.914	0.266	0.097	0.453	0.411	0.00846	0.00809
0.9	0	OLS	0.040	0.033	0.055	0.253	0.012	0.00240	0.00029
		WLS	0.002	0.011	0.036	0.164	0.019	0.00147	0.00027
		GLS	0.002	0.004	0.029	0.132	0.119	0.00118	0.00107
	0.011	OLS	0.051	0.033	0.056	0.254	0.016	0.00244	0.00038
		WLS	0.004	0.013	0.039	0.177	0.021	0.00169	0.00030
		GLS	0.012	0.008	0.032	0.142	0.126	0.00133	0.00119
	0.102	OLS	0.141	0.048	0.063	0.281	0.044	0.00303	0.00104
		WLS	0.072	0.037	0.052	0.229	0.051	0.00251	0.00100
		GLS	0.101	0.033	0.044	0.193	0.175	0.00217	0.00206
	0.284	OLS	0.322	0.092	0.076	0.338	0.100	0.00433	0.00237
		WLS	0.242	0.084	0.070	0.306	0.110	0.00398	0.00237
		GLS	0.281	0.082	0.062	0.275	0.253	0.00371	0.00361
	0.557	OLS	0.593	0.165	0.096	0.428	0.184	0.00634	0.00438
		WLS	0.508	0.158	0.092	0.407	0.196	0.00610	0.00440
		GLS	0.553	0.157	0.086	0.382	0.355	0.00588	0.00579
	0.922	OLS	0.957	0.264	0.122	0.549	0.297	0.00907	0.00707
		WLS	0.869	0.257	0.120	0.534	0.310	0.00889	0.00709
		GLS	0.916	0.258	0.115	0.516	0.481	0.00872	0.00860

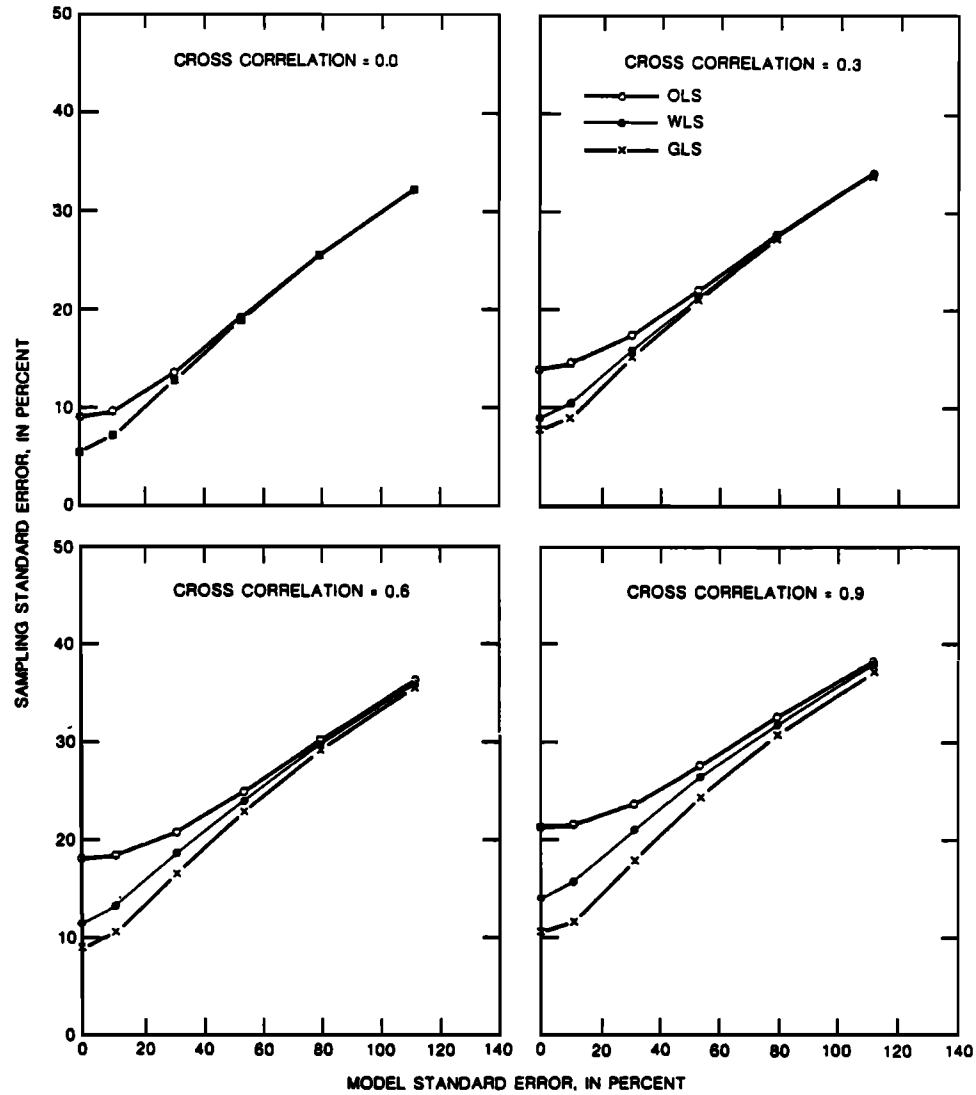

Fig. 2. Monte Carlo results from experiment 2 show prediction accuracy obtained with GLS, WLS, and OLS routines. The regression estimators of the mean annual flow were obtained with flow data from 20 sites: 5 with 50 years of data, 5 with 10 years of data, and 10 with 5 years of data.

Here mse_S is the average mean square error of $\hat{\theta}(A)$ about $\theta(A)$, and $\bar{\gamma}^2$ is the average model error variance of $\hat{\theta}$ about its conditional mean $\theta(A)$.

Experiment 1

In the first Monte Carlo experiment the dependent variables was the 50-year peak flood estimated at 30 sites: 10 sites with 50 years of data, 10 sites with 20 years of data, and 10 sites with 10 years of data. Results were obtained for $\rho = 0.0$, 0.3, 0.6, and 0.9, and $\sigma_\varepsilon = 0.0$, 0.10, 0.30, 0.50, and 0.90. These values of σ_ε (with $\sigma_\delta = 0.25\ \sigma_\varepsilon$) correspond to average model error variances $\bar{\gamma}^2$ of 0.00, 0.011, 0.102, 0.284, 0.557, and 0.922.

The resultant sampling mean square errors mse_S are illustrated in Figure 1; the sampling standard error in percent for $q_p = \exp(x_p)$ where $x_p = \mu + z_p\sigma$ was obtained from the mse_S by the transformation

$$\text{PSE} = 100[\exp(\text{mse}_S) - 1]^{0.5} \tag{27}$$

whereas the model standard error for q_p, in percent, is given by

$$\text{PME} = 100[\exp(\bar{\gamma}^2) - 1]^{0.5} \tag{28}$$

The GLS estimator was best for all model errors tested when $\rho \geq 0.6$ and was as good or better than WLS and OLS for $\rho < 0.6$. The WLS estimator was as good or better than the OLS estimator for all combinations of $\bar{\gamma}^2$ and ρ tested. The difference between the performance of the estimators is greatest for small model errors.

Table 1 lists the mean and standard derivation of the estimators of the average model error variance $\bar{\gamma}^2$. The OLS estimator of $\bar{\gamma}^2$ was taken to be the residual mean square error; WLS and GLS estimators were obtained by solution of (14) as discussed in the text. From Table 1 and results from the other experiments, one can conclude that OLS provides a highly biased estimate of the model error variance, particularly when the cross correlation is high or the true model error is relatively small. This deficiency with OLS was one factor that led *Moss and Karlinger* [1973, p. 427] to develop the NARI procedure for hydrologic data network design.

Table 1 also lists the mean and variance of the $\hat{\alpha}$ and $\hat{\beta}$ values obtained in the 1000 Monte Carlo replications of each procedure. Also given are the average variances of $\hat{\alpha}$ and $\hat{\beta}$ predicted by the various procedures were its basic assumptions satisfied. (See equations (12) and (17).)

TABLE 2. Partial Results for Experiment 2

ρ	Model Error Variance	Method	Mean of Estimated Model Error Variance	Standard Deviation of Estimated Model Error Variance	Sampling Mean Square Error, mse_S	Variance of $\hat{\alpha}$	Mean of Predicted Variance of $\hat{\alpha}$	Variance of $\hat{\beta}$	Mean of Predicted Variance of $\hat{\beta}$
0	0	OLS	0.065	0.034	0.008	0.049	0.031	0.00087	0.00076
		WLS	0.005	0.010	0.003	0.017	0.019	0.00027	0.00031
		GLS	0.005	0.010	0.003	0.017	0.019	0.00027	0.00032
	0.01	OLS	0.076	0.037	0.009	0.054	0.036	0.00097	0.00088
		WLS	0.014	0.015	0.005	0.029	0.028	0.00051	0.00050
		GLS	0.014	0.016	0.005	0.029	0.029	0.00051	0.00051
	0.09	OLS	0.156	0.060	0.018	0.095	0.075	0.00190	0.00181
		WLS	0.094	0.047	0.016	0.082	0.079	0.00167	0.00165
		GLS	0.094	0.047	0.016	0.082	0.080	0.00167	0.00167
	0.25	OLS	0.317	0.110	0.036	0.178	0.153	0.00381	0.00369
		WLS	0.255	0.103	0.035	0.172	0.164	0.00371	0.00364
		GLS	0.255	0.103	0.035	0.172	0.165	0.00371	0.00365
	0.49	OLS	0.559	0.189	0.063	0.302	0.270	0.00668	0.00649
		WLS	0.496	0.184	0.063	0.299	0.284	0.00664	0.00650
		GLS	0.497	0.185	0.063	0.300	0.286	0.00665	0.00651
	0.81	OLS	0.881	0.295	0.099	0.467	0.427	0.01053	0.01023
		WLS	0.818	0.292	0.099	0.467	0.442	0.01052	0.01027
		GLS	0.819	0.292	0.099	0.467	0.443	0.01053	0.01028
0.3	0	OLS	0.052	0.027	0.019	0.099	0.025	0.00122	0.00060
		WLS	0.003	0.007	0.008	0.036	0.017	0.00039	0.00027
		GLS	0.004	0.008	0.006	0.028	0.029	0.00033	0.00036
	0.01	OLS	0.062	0.030	0.021	0.104	0.030	0.00132	0.00072
		WLS	0.010	0.012	0.011	0.050	0.024	0.00066	0.00042
		GLS	0.013	0.013	0.008	0.041	0.040	0.00057	0.00056
	0.09	OLS	0.143	0.053	0.030	0.145	0.069	0.00225	0.00166
		WLS	0.086	0.043	0.025	0.115	0.075	0.00187	0.00154
		GLS	0.093	0.044	0.023	0.108	0.101	0.00183	0.00178
	0.25	OLS	0.304	0.105	0.048	0.228	0.147	0.00415	0.00353
		WLS	0.245	0.099	0.045	0.211	0.158	0.00394	0.00351
		GLS	0.254	0.100	0.044	0.209	0.194	0.00395	0.00383
	0.49	OLS	0.546	0.184	0.075	0.353	0.264	0.00702	0.00634
		WLS	0.485	0.181	0.074	0.343	0.278	0.00691	0.00636
		GLS	0.495	0.181	0.073	0.343	0.319	0.00694	0.00672
	0.81	OLS	0.868	0.291	0.111	0.519	0.421	0.01087	0.01008
		WLS	0.807	0.288	0.110	0.513	0.436	0.01081	0.01013
		GLS	0.817	0.289	0.110	0.515	0.480	0.01086	0.01051
0.6	0	OLS	0.039	0.023	0.032	0.151	0.018	0.00160	0.00045
		WLS	0.002	0.009	0.013	0.057	0.016	0.00055	0.00026
		GLS	0.002	0.005	0.008	0.037	0.035	0.00037	0.00038
	0.01	OLS	0.049	0.026	0.033	0.156	0.023	0.00170	0.00057
		WLS	0.007	0.012	0.017	0.075	0.022	0.00085	0.00037
		GLS	0.012	0.010	0.011	0.051	0.048	0.00061	0.00059
	0.09	OLS	0.130	0.049	0.042	0.197	0.063	0.00263	0.00151
		WLS	0.078	0.041	0.034	0.151	0.070	0.001214	0.00144
		GLS	0.092	0.040	0.027	0.127	0.117	0.00193	0.00185
	0.25	OLS	0.291	0.100	0.060	0.280	0.141	0.00453	0.00338
		WLS	0.235	0.096	0.056	0.254	0.153	0.00425	0.00339
		GLS	0.253	0.095	0.051	0.239	0.217	0.00415	0.00395
	0.49	OLS	0.533	0.180	0.087	0.404	0.258	0.00741	0.00619
		WLS	0.474	0.177	0.085	0.389	0.273	0.00724	0.00623
		GLS	0.495	0.177	0.082	0.381	0.348	0.00721	0.00689
	0.81	OLS	0.855	0.287	0.124	0.570	0.414	0.01126	0.00992
		WLS	0.795	0.285	0.122	0.560	0.430	0.01116	0.00998
		GLS	0.816	0.285	0.120	0.558	0.513	0.01118	0.01072
0.9	0	OLS	0.026	0.025	0.044	0.202	0.012	0.00196	0.00030
		WLS	0.003	0.010	0.019	0.081	0.017	0.00073	0.00026
		GLS	0.001	0.002	0.011	0.048	0.041	0.00043	0.00038
	0.01	OLS	0.036	0.026	0.045	0.207	0.017	0.00206	0.00042
		WLS	0.007	0.013	0.024	0.103	0.021	0.00106	0.00035
		GLS	0.010	0.006	0.013	0.061	0.053	0.00065	0.00058
	0.09	OLS	0.117	0.046	0.054	0.248	0.056	0.00300	0.00135
		WLS	0.072	0.040	0.043	0.189	0.066	0.00243	0.00134
		GLS	0.091	0.035	0.031	0.143	0.127	0.00201	0.00186
	0.25	OLS	0.278	0.096	0.073	0.330	0.135	0.00491	0.00323
		WLS	0.225	0.094	0.067	0.297	0.148	0.00459	0.00326
		GLS	0.252	0.090	0.057	0.264	0.235	0.00430	0.00402
	0.49	OLS	0.520	0.176	0.100	0.455	0.252	0.00781	0.00603
		WLS	0.463	0.174	0.096	0.434	0.267	0.00761	0.00609
		GLS	0.493	0.172	0.090	0.414	0.372	0.00745	0.00701
	0.81	OLS	0.842	0.282	0.136	0.622	0.408	0.01168	0.00977
		WLS	0.783	0.281	0.134	0.608	0.424	0.01155	0.00984
		GLS	0.818	0.280	0.129	0.596	0.542	0.01148	0.01088

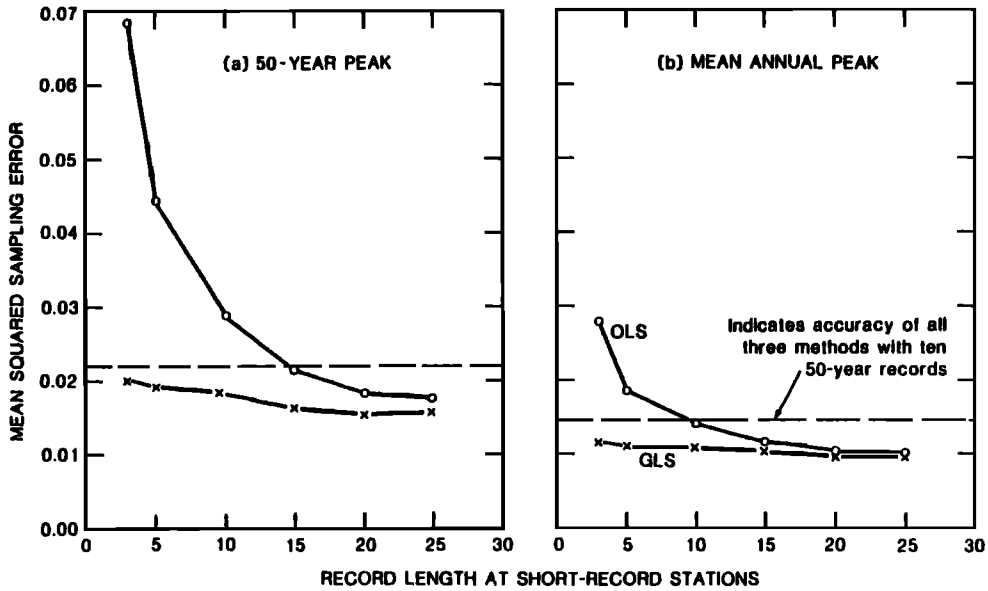

Fig. 3. Monte Carlo results from experiment 3 show the impact of short record sites on the precision with which the OLS and GLS procedures estimate the mean annual and 50-year peak flood. For this example, $\rho = 0.3$ and $\bar{\gamma}^2 = 0.0400$ (mean) and 0.0454 (50-year peak). The dashed line indicates the accuracy of OLS, WLS, and GLS estimators when only flow data for ten sites each with 50-year records is available. The other points indicate estimator precision using flow data from ten 50-year sites and an additional ten short-record sites with the record lengths indicated. The WLS points are not shown but fell slightly above the GLS values.

From Table 1 and results from the other experiments one can conclude that only the GLS procedure accurately estimates the variances of $\hat{\alpha}$ and $\hat{b}$ when $\rho > 0$. Because of the distortion in these values with the OLS procedure, tests of significance and variable selection procedures may lead to erroneous conclusions.

Experiment 2

In the second experiment the dependent variable was the mean annual peak estimated at 20 sites: 5 with 50 years of data, 5 with 10 years of data, and 10 with 5 years of data. Results were obtained for the same combinations of ρ and σ_ε yielding $\bar{\gamma}^2 = 0.0$, 0.01, 0.09, 0.25, 0.49, or 0.81 for $z_p = 0$. Results in terms of sampling mean square error (Figure 2) are similar to those for experiment 1. The GLS procedure out performed the WLS and OLS procedures for $\rho \geq 0.6$. The GLS and WLS both performed as well as or better than the OLS procedure in all cases. Results in terms of estimates of $\bar{\gamma}^2$ and variances of $\hat{\alpha}$ and $\hat{\beta}$ (Table 2) were similar to those of experiment 1 and lead to the same conclusions.

Experiment 3

The third experiment was designed to show the effect of adding short-record sites to a regional regression analysis. The dependent variable was the 50-year peak in each of seven cases. The first case had 10 sites each with 50 years of record. The second case had 20 sites: 10 with 50 years of record and an additional 10 stations with 3 years of record. Cases 3–7 had 20 stations: 10 stations with 50 years of record and an additional 10 short-record stations with 5 (case 3), 10 (case 4), 15 (case 5), 20 (case 6), or 25 (case 7) years of record. The results in Figure 3*a* demonstrate that stations with short records may be counter productive when OLS is used. This observation has led some analysts to recommend dropping certain short-record stations from a regional regression (for example, *Tasker and Moss* [1979]). On the other hand, the GLS and WLS procedures allow one to effectively use short-record stations to improve the regression model's parameter estimates.

The experiment was repeated with the mean annual peak discharge as the the dependent variable. Results in Figure 3*b* are similar to the results for the 50-year peak, although the differences between OLS and GLS were not as striking. The difference in results for the mean and 50-year peaks is probably due to the large uncertainties in calculated s_i values based on very short records. Still, including records as short as 3 years improves the GLS estimator, regardless of whether the regression is for the mean or 50-year peak. Moreover, with the GLS procedure, a site with 5–10 years of record can be seen in these cases to provide almost as much regional information as it would were 25 years of record available. One should not conclude from the results in experiment 3 that long records are not valuable. To show the effect of having some long record stations in a network, a fourth experiment was performed.

Experiment 4

The final experiment was designed to show the value of having longer records at a few sites in a network. Five cases were run; in the first case the 20 stations each had 10 years of data. Cases 2 through 5 also had 20 stations, where in each case, 15 stations had 10 years of data. The remaining 5 stations had either 20 years (case 2), 30 years (case 3), 40 years (case 4), or 50 years (case 5) of data.

Separate runs were made for the mean annual peak and 50-year peak with cross correlations of 0.3 and 0.6. The results are summarized in Figure 4. They show that having a few long-record stations decreases the mean-square sampling error for both OLS and GLS estimators. However, it is clear by the increasing distance between the two curves that the GLS estimator makes more efficient use of the longer records than does the OLS estimator.

CONCLUSIONS

This paper has examined the statistical performance of three-parameter estimation and analysis procedures (OLS,

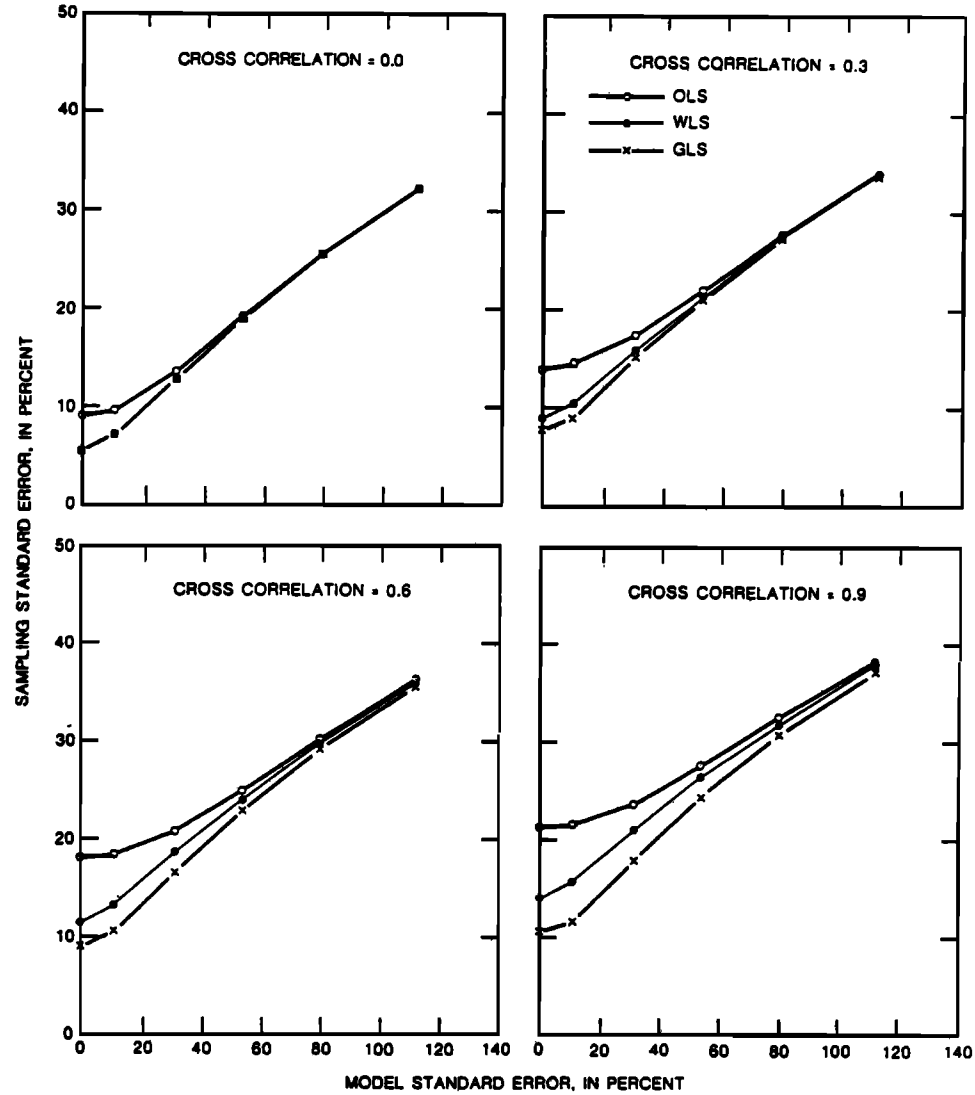

Fig. 4. Monte Carlo results from experiment 4 shown the impact of having a network of 15 short-record (10-year) sites and 5 long-record sites on the precision with which the OLS and GLS procedures estimate the mean annual and 50-year peak floods. For the mean annual flow, $\bar{\gamma}^2 = 0.0400$; for the 50-year peak, $\bar{\gamma} = 0.0454$.

WLS, GLS). The basic theory and motivation of the procedure were presented along with four sets of Monte Carlo results. Several important lessons relating to the estimation of the parameters of such models and hydrologic network design can be learned from this study.

1. The analyses clearly demonstrate that use of the proposed WLS and the GLS estimators instead of the popular OLS can result in remarkable improvements in the precision with which the parameters of regional hydrologic regression models can be estimated. This is particularly true when the length of record varies widely from site to site. In some cases there was almost no difference in the precision of the WLS and GLS estimators. In other cases, particularly when the cross correlation of the flows was 0.6 or greater, and the model error variance was relatively modest, the GLS estimators had substantially greater precision than the WLS estimators.

The numerical results presented here are for the case where $x = \ln(q)$ is normally distributed. However, the general procedures are equally applicable to other distributions for which reasonably accurate estimators of Var $(\hat{\theta}_i)$ and Cov $(\hat{\theta}_i, \hat{\theta}_j)$ can be developed; in this regard, an extremely important constraint on any estimator of $\Sigma(\hat{\boldsymbol{\theta}})$ is that the variance and covariance estimators associated with each $\hat{\theta}_i$ should be nearly independent of the differences $\{\hat{\theta}_i - E[\theta_i | A]\}$ if unbiased and minimum variance estimators of α and β are to be obtained (see Appendix B).

2. The Monte Carlo results demonstrate that the accuracy of regional regression models can be improved by incorporating into the analysis sites with as little as 3 years of record. Certainly, it would be unwise to depend on only 3 years of at-site data to estimate the 50-year flood peak at a site for flood plain management. However, such data and such estimates can be used to improve regional estimators of various flood flow statistics. These opportunities were not available when model parameters were estimated with OLS because sites with short flow records often degraded the estimators' precision and thus had a negative impact on model accuracy [*Tasker and Moss*, 1979].

3. The results in Tables 1 and 2 demonstrate the tremendous differences between the estimates of Var $(\hat{\alpha})$ and Var $(\hat{\beta})$ generated by OLS and WLS routines and the actual accuracy

of their estimators. This was even true in instances when the WLS estimators were essentially as efficient as those generated with the GLS algorithm. This has major implications for the use of these estimators in step-wise regression procedures and other automated and nonautomated algorithms which rely on the estimated statistical significance of the regression coefficients to select an appropriate regional hydrologic model. Because the estimated variances of the parameter estimators with OLS and WLS are often in error by a factor of two, model selection based on the apparent statistical significance of the parameters could be a questionable procedure.

4. Finally, parameter estimation algorithms determine the relationship between model estimation accuracy and available and potentially available data. Optimal network design decisions are critically related to the ability of parameter estimation procedures to make use of streamflow records which are or could be available at each gage. Moreover, to make optimal network design decisions one needs to estimate what is or that might be the impact of various data on model prediction accuracy. The inability of OLS procedures to adequately describe model accuracy led *Moss and Karlinger* [1974] and others [*Moss*, 1976, 1979; *Tasker and Moss*, 1979] to develop and employ the complex NARI package. However, as has been seen, generalized least squares procedures can change the environment and constraints within which regional hydrologic network analyses are performed. The GLS procedure's ability to reasonably describe the precision of the resultant regression model and of a model's prediction error should mean that indirect procedures such as NARI can be avoided.

APPENDIX A: MODELS FOR POPULATION PARAMETERS

Here μ_i and σ_i will denote the mean and variance of the population from which the n_i values $\{x_{it}\}$ available at site i were drawn; $\mu_i + z_p\sigma_i$ will be the 100p percentile of that distribution. Suppose that the actual flows q in a river basin are lognormally distributed so that

$$x = \ln (q) \quad \text{with}$$

$$\mu_q = E[q] = \exp [\mu + 0.5\sigma^2] \tag{A1}$$

$$\sigma_q{}^2 = \text{Var } [q] = E[q]^2[\exp (\sigma^2) - 1] \tag{A2}$$

where

$$P_r[q \leq \exp (\mu + z_p\sigma)] = p \tag{A3}$$

Hence q_p equals $\exp (x_p)$, where q_p and x_p are the 100p percentiles of the corresponding distributions.

Typically, regional regression equations are derived for such statistics as q's mean μ_q, q's standard deviation σ_q, or q_p by regressing the logarithms of such quantities on the logarithms of various basin characteristics (see, for example, *Thomas and Benson* [1970]). Here drainage area A is used as a surrogate for all the basin characteristics that might be employed. The resultant models of q_p are

$$\ln q_p = \alpha_p + \beta_p \ln A + \varepsilon \tag{A4}$$

where ε is the residual error. In fact, this very model was employed by *Benson* [1962] and *Tasker and Moss* [1979].

Here we need to assume a true model for the mean μ_i and standard deviation σ_i of the x_{it} as a function of A_i for use in the Monte Carlo experiments. For the mean we use

$$\mu_i = \alpha_\mu + \beta_\mu \ln A_i + \varepsilon_i \tag{A5}$$

where

$$\varepsilon_i \sim \text{NID}(0, \sigma_\varepsilon{}^2)$$

This is entirely consistent with (A4), which reduces to (A5) for $p = 0.50$ corresponding to the 2-year flood. Following *Tasker* [1980], we took $\alpha_\mu = 0$ and $\beta_\mu = 0.75$. Here α_μ is only a location parameter and its value will have no effect on most of the statistics reported.

The appropriate form of the σ model is less clear. *Moss and Karlinger* [1974], *Tasker and Moss* [1979], and *Tasker* [1980] all assumed that σ was constant, independent of ln A and other drainage basin characteristics. While this was perhaps a reasonable assumption for those investigations, studies relating flow statistics with basin characteristics have shown that σ varies inversely with drainage area [*Benson*, 1962, equations A-21 to A-23; *Cruff and Rantz*, 1965; *Thomas and Benson*, 1970]. A reasonable model for $E[\sigma]$ consistent with (A4) is

$$E[\sigma] = \alpha_\sigma + \beta_\sigma \ln A \tag{A6}$$

In fact, the values of α_σ and β_σ can be obtained by differencing (A4) for ln $[q_{0.50}]$ and ln $[q_p]$ with $p \neq 0.50$. By using the equations given in the work *Tasker and Moss* [1979] we obtained

$$\alpha_\sigma = 1.5 \qquad \beta_\sigma = -0.14$$

Benson's [1962] analysis suggested that $\beta_\sigma = -0.17$ was appropriate for parts of New England.

A remaining issue is the choice of distribution for the σ-model's errors. If one writes the σ model as

$$\sigma_i = \alpha_\sigma + \beta_\sigma \ln A_i + \eta_i \tag{A7}$$

where σ_i is to be nonnegative, it is clear that η_i must satisfy

$$\eta_i \geq -(\alpha_\sigma + \beta_\sigma \ln A_i) \tag{A8}$$

Here the lower bound on η_i depends on ln A_i. Furthermore, if $E[\sigma_i]$ is on the order of 1 or larger, then values of Var $[\eta_i]$ on the order of $(0.1)^2$ or $(0.2)^2$ seem plausible, given the variability in σ_i one would expect to see among basins with similar drainage basin characteristics. However, if $E[\sigma_i] \simeq 0.2$, then values of Var $[\eta_i]$ on the order of $(0.1)^2$ or $(0.2)^2$ seem much too large, for they imply that the coefficient of variation of the σ_i values would be 0.5 and 1.0, respectively. A reasonable model for σ_i consistent with (A4) is

$$\sigma_i = [\alpha_\sigma + \beta_\sigma \ln A_i] \exp (\delta_i) \tag{A9}$$

where

$$\delta_i \sim \text{NID}(-\sigma_\delta{}^2/2, \sigma_\delta{}^2)$$

With this model, for given A_i, σ_i is a lognormally distribution with mean $\sigma(A)$, given by (A6), and variance $\sigma^2(A)[\exp (\sigma_\delta{}^2) - 1]$. Thus σ_i is strictly positive and unbounded above.

Equation (A9) also yields a reasonable model of the likely variation in the coefficient of variation of the q distribution across sites. Given that q has a lognormal distribution, its coefficient of variation is

$$CV_q = \frac{\sigma_q}{\mu_q} = [\exp (\sigma^2) - 1]^{0.5} \tag{A10}$$

When the variance $\sigma_i{}^2$ of the x_{it} is small,

$$CV_q = [\exp (\sigma_i{}^2) - 1]^{0.5} \cong \sigma_i \tag{A11}$$

so that the coefficient of variation of the q_{it} is essentially σ_i. Equation (A9) states that for given A_i, the coefficient of variation of σ_i across sites is $[\exp(\sigma_\delta^2) - 1]^{0.5} \simeq \sigma_\delta$; thus σ_δ will very nearly be the coefficient of variation of CV_q across sites with similar drainage characteristics.

Often the logarithms of the flow's mean μ_q and either σ_q or σ_q^2 are modeled as if they were linear functions of ln A and additive error terms. (for examples, see *Thomas and Benson* [1970]). When the models of μ_q are σ_q are

$$\begin{aligned} \ln \mu_q &= \alpha_1 + \beta_1 \ln A + \varepsilon_1 \\ \ln \sigma_q &= \alpha_2 + \beta_2 \ln A + \varepsilon_2 \end{aligned} \tag{A12}$$

This would imply that

$$\begin{aligned} \ln CV_q &= \ln (\sigma_q/\mu_q) \\ &= \ln (\sigma_q) - \ln (\mu_q) \\ &= \alpha_{cv} + \beta_{cv} \ln A + \varepsilon_{cv} \end{aligned} \tag{A13}$$

where

$$\begin{aligned} \alpha_{cv} &= \alpha_2 - \alpha_1 \\ \beta_{cv} &= \beta_2 - \beta_1 \\ \varepsilon_{cv} &= \varepsilon_2 - \varepsilon_1 \end{aligned} \tag{A14}$$

Here ε_{cv} is a random error term. Recall from (A11) that at site i, $CV_q \cong \sigma_i$. Thus (A13) implies that

$$\ln (\sigma_i) \cong \alpha_{cv} + \beta_{cv} \ln A_i + \varepsilon_{cv,i} \tag{A15}$$

or

$$\sigma_i = c[A_i]\beta_{cv} \exp (\varepsilon_{cv,i}) \tag{A16}$$

where $c = \exp(\alpha_{cv})$. Thus the error model for σ_i in (A16) which results from linear models in ln (A) of ln (σ_q) and of ln (μ_q) is the same as the model employed in (A9); the difference between (A9) and (A16) is how $E[\sigma]$ depends on A or equivalently on ln (A). While (A16) provides a reasonable model of $\sigma(A)$ yielding only positive values, the model (A9) will be employed here. Use of (A5) and (A9) together as models of μ_i and σ_i implies that $x_p = \ln q_p$ is of the form in (A4), regardless of the value of p. This would not be the case if (A16) were used as a model of σ. However, for large enough A_i and negative β_i, (A9) could yield negative values of σ_i, whereas (A16) could not.

APPENDIX B: ESTIMATION OF $\Sigma(\hat{\boldsymbol{\theta}})$

Equations (1)–(14) present the basic theory and the motivation behind the weighted and generalized least squares $\boldsymbol{\beta}$ estimators. Both of these estimators require estimates of all or part of $\Sigma(\hat{\boldsymbol{\theta}})$, the sampling covariance matrix for the at-site estimators $\hat{\theta}_i$. This section discusses one approach to estimating $\Sigma(\hat{\boldsymbol{\theta}})$ and the reasons for its selection in this study.

Problem

In general, both the weighted and generalized least squares estimators can be written

$$\hat{\boldsymbol{\beta}}_G = (\Xi^T M \Xi)^{-i}\, \Xi^T M \hat{\boldsymbol{\theta}} \tag{B1}$$

where M is some weighting matrix. Regardless of the choice of the constant matrix M, $\hat{\boldsymbol{\beta}}_G$ will be unbiased

$$E[\hat{\boldsymbol{\beta}}_G] = E[(\Xi^T M \Xi)^{-1}\, \Xi M \hat{\boldsymbol{\theta}}] = \boldsymbol{\beta} \tag{B2}$$

if $E[\hat{\boldsymbol{\theta}}] = \Xi\boldsymbol{\beta}$; $\hat{\boldsymbol{\beta}}_G$ is still unbiased even if M is a random variable provided its distribution is independent of $(\hat{\boldsymbol{\theta}} - \Xi\boldsymbol{\beta})$.

While any fixed value of M yields an unbiased estimator, some values of M yield some values of $\hat{\boldsymbol{\beta}}_G$ with smaller variances. In particular, M^{-1} equal to

$$\begin{aligned} &E[(\hat{\boldsymbol{\theta}} - \Xi\boldsymbol{\beta})(\hat{\boldsymbol{\theta}} - \Xi\boldsymbol{\beta})^T] \\ &\quad = E[(\boldsymbol{\theta} - \Xi\boldsymbol{\beta})(\boldsymbol{\theta} - \Xi\boldsymbol{\beta})^T] + E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T] \\ &\quad = \gamma^2 I_N + \Sigma(\hat{\boldsymbol{\beta}}) \end{aligned} \tag{B3}$$

where γ^2 is the assumed model error variance, yields the minimum variance $\boldsymbol{\beta}$ estimator. This suggests that the closer our estimator $\hat{\Sigma}(\hat{\boldsymbol{\theta}})$ is to $\Sigma(\hat{\boldsymbol{\theta}})$, the more efficient will be the resultant estimator of $\boldsymbol{\beta}$. Unfortunately, that goal can conflict with the requirement that M is fixed or least independent of $(\hat{\boldsymbol{\theta}} - \Xi\boldsymbol{\beta})$.

Consider the derivation of regional relationship for the x distribution's 100p percentile $\mu_i + z_p\sigma_i$, $z_p \neq 0$. The natural estimator of this quantity is $\bar{x}_i + z_p s_i$ with sampling variance, to first order, equal to $(1 + z_p^2/2)(\sigma_i^2/n_i)$. Clearly, $\hat{\theta}_i = \bar{x}_i + z_p s_i$ and the natural sample estimator of its variance $(1 + z_p^2/2)$ (s_i^2/n_i) both depend on s_i and hence are likely to be highly correlated. We initially tried such combinations of quantile and quantile-variance estimators with disastrous results. Even use of σ_i, were its value known, to estimate $\hat{\theta}_i$ variance as $(1 + z_p^2/2)(\sigma_i^2/n_i)$ is likely to cause troubles; in this second case, there would still be correlation between σ_i and the model residuals $(\boldsymbol{\theta} - \Xi\boldsymbol{\beta})$ which depend on the difference between σ_i and its conditional expectation $E[\sigma_i | A_i]$. These same problems arise when deriving a regional regression estimator of σ_i based upon (A9).

Solution

A reasonable estimator of $\Sigma(\hat{\boldsymbol{\theta}})$ is needed that is or is nearly independent of $\hat{\boldsymbol{\theta}} - E[\hat{\boldsymbol{\theta}} | A]$. Such an estimator was constructed by replacing each σ_i in (3)–(6) by

$$\hat{\sigma}(A_i) = \hat{\alpha}_\sigma + \hat{\beta}_\sigma \ln A_i \tag{B4}$$

where $\hat{\alpha}_\sigma$ and $\hat{\beta}_\sigma$ are estimators of the α_σ and β_σ which appear in (A6). Thus each σ_i is replaced by an estimate of its expected value, given the drainage area and other basin characteristics of that site. These estimates will be nearly independent of the σ-model's residuals $[\sigma_i - \sigma(A_i)]$ and the time sampling errors $[s_i - \sigma_i]$. The only dependence which remains is that which rises from use of estimators of α_σ and β_σ. Let $\mathbf{s} = (s_1, \cdots, s_N)^T$; then the OLS estimates of α_σ and β_σ are

$$\begin{pmatrix} \hat{\alpha}_{\sigma-\text{OLS}} \\ \hat{\beta}_{\sigma-\text{OLS}} \end{pmatrix} = (\Xi^T\Xi)^{-1}\Xi^T\mathbf{s} \tag{B5}$$

In many cases, (B5) provides reasonable values of the $\hat{\alpha}_\sigma$ and $\hat{\beta}_\sigma$ needed in (B4). However, the estimators in (B5) ignore the possible differences in record lengths among the stations and the variations in the variance of the residuals $[\sigma_i - \sigma(A_i)]$ with A_i implied by the σ error model in (A9).

To obtain even better estimators of α_σ and β_σ for use in (B4), the OLS estimators in (B5) were used to construct improved weighted least squares estimators:

$$\begin{pmatrix} \hat{\alpha}_{\sigma-\text{WLS}} \\ \hat{\beta}_{\sigma-\text{WLS}} \end{pmatrix} = (\Xi^T W_\sigma \Xi)^{-1}\Xi^T\, W_\sigma\mathbf{s} \tag{B6}$$

where W_σ is a diagonal weighting matrix with

$$(W_\sigma)_{ii}^{-1} = \left\{\frac{1}{2n_i} + [\exp(\sigma_\delta^2) - 1]\right\}\hat{\sigma}^2(A_i) \tag{B7}$$

Here $\hat{\sigma}^2(A_i)/(2n_i)$ describes the variance of the time sampling error $(s_i - \sigma_i)$, while $\hat{\sigma}^2(A_i)[\exp(\sigma_\delta^2) - 1]$ describes the as-

sumed variance of the σ-model's residuals. The value of σ_δ^2 needed for solution of (B6) was obtained by solution of (14) with W_σ replacing $\Lambda(\delta^2)^{-1}$.

Equations (B6) and (B7), in practice, could be used iteratively by employing the $\hat{\sigma}_{\sigma-\mathrm{WLS}}$ and $\hat{\beta}_{\sigma-\mathrm{WLS}}$ values from one iteration to calculate the $\hat{\sigma}(A_i)$ used in the next iteration. While use of (B6) following (B5) sometimes resulted in substantial improvements in $\hat{\sigma}(A_i)$, repeated use of (B6) and (B7) yielded only marginal accuracy increases. In our simulations we generally followed an initial iteration of the OLS algorithm with two iterations of the WLS algorithm. In practice, one could use generalized least squares estimators of α_σ and β_σ.

Acknowledgments. We want to express our thanks to M. Moss, E. Gilroy, M. Karlinger, and W. Thomas for their encouragement and useful comments during the course of this study.

REFERENCES

Benson, M. A., Evolution of methods for evaluating the occurence of floods, *U.S. Geol. Surv. Water Supply Pap., 1580-A*, 30 pp., 1962.

Cruff, R. W., and S. E. Rantz, A comparison of methods used in flood frequency studies for coastal basins in California, *U.S. Geol. Surv. Water Supply Pap., 1580-E*, 56 pp., 1965.

Draper, N. R., and H. Smith, *Applied Regression Analysis*, 2nd ed., John Wiley, New York, 1981.

Hardison, C. H., Prediction error of regression estimates of streamflow characteristics at ungaged sites, *U.S. Geol. Surv. Prof. Pap., 750-C*, C228–C236, 1971.

Johnston, J., *Econometric Methods*, McGraw-Hill, New York, 1972.

Kuczera, G., Combining site-specific and regional information: An empirical Bayes approach, *Water Resour. Res., 18*(2), 306–314, 1982*a*.

Kuczera, G., Robust flood-frequency models, *Water Resour. Res., 18*(2), 315–324, 1982*b*.

Kuczera, G., Effect of sampling uncertainty and spatial correlation on an empirical Bayes procedure for combining site and regional information, *J. Hydrol., 65*(4), 373–398, 1983.

Marin, C., Parameter uncertainty in water resources planning, Ph.D. thesis, Harvard Univ., Cambridge, Mass., 1983.

Matalas, N. C., and M. A. Benson, Effects of interstation correlation on regression analysis, *J. Geophys. Res., 66*(10), 3285–3293, 1961.

Matalas, N. C., and E. J. Gilroy, Some comments on regionalization in hydrologic studies, *Water Resour. Res., 4*(6), 1361–1369, 1968.

Moss, M. E., Cross correlation of the logarithms of estimates of mean streamflows, *Water Resour. Res., 9*(6), 1681–1633, 1973.

Moss, M. E., Design of surface water data networks for regional information, *Hydrol. Sci. Bull., 21*(1), 113–127, 1976.

Moss, M. E., Space, time, and the third dimension (model error), *Water Resour. Res., 15*(6), 1797–1800, 1979.

Moss, M. E., Concepts and techniques in hydrological network design, *Oper. Hydrol. Rep. 19*, 30 pp., World Meteorological Organization, Geneva, 1982.

Moss, M. E., and M. R. Karlinger, Surface water network design by regression analysis simulation, *Water Resour. Res., 10*(3), 427–433, 1974.

Stedinger, J. R., Estimating a regional flood frequency distribution, *Water Resour. Res., 19*(2), 503–510, 1983*a*.

Stedinger, J. R., Design events with specified flood risk, *Water Resour. Res., 19*(2), 511–522, 1983*b*.

Tasker, G. D., Hydrologic regression and weighted least squares, *Water Resour. Res., 16*(6), 1107–1113, 1980.

Tasker, G. D., and M. E. Moss, Analysis of Arizona flood data network for regional information, *Water Resour. Res., 15*(6), 1791–1796, 1979.

Thomas, D. M., and M. A. Benson, Generalization of streamflow characteristics, from drainage-basin characteristics, *U.S. Geol. Surv. Water Supply Pap., 1975*, 55 pp., 1970.

J. R. Stedinger, Department of Environmental Engineering, Cornell University, Ithaca, NY 14853.

G. D. Tasker, U.S. Geological Survey, 430 National Center, Reston, VA 22092.

(Received May 10, 1984;
revised May 2, 1985;
accepted June 3, 1985.)