LOW-FLOW FREQUENCY ESTIMATION USING BASE-FLOW MEASUREMENTS

by J. R. Stedinger and W. O. Thomas, Jr.

Open-File Report 85-95

Reston, Virginia

1985

U.S. DEPARTMENT OF THE INTERIOR

WILLIAM P. CLARK, Secretary

GEOLOGICAL SURVEY

Dallas L. Peck, Director

For additional information write to:

Chief Hydrologist
U.S. Geological Survey, WRD
415 National Center
Reston, Virginia 22092

Copies of this report can be purchased from:

Open-File Services Section
Western Distribution Branch
U.S. Geological Survey
Box 25425, Federal Center
Denver, Colorado 80225

CONTENTS

FIGURES

TABLES

Low-Flow Frequency Estimation Using Base-Flow Measurements

by

Jery R. Stedinger
Wilbert O. Thomas, Jr.

ABSTRACT

Estimates of the d-day T-year low flow, such as the 7-day 10-year, are needed at ungaged sites for water-quality management. Experience has indicated that these low-flow characteristics cannot be accurately estimated by regression on drainage-basin characteristics. An alternative is the use of base-flow measurements at the ungaged site and concurrent daily flows at a nearby gaged site to establish a regression relationship between low flows at the two locations. Traditionally the 7-day 10-year low flow at the ungaged site is estimated by using the computed 7-day 10-year low flow at the nearby gaged site and the established regression relationship. This technique is shown to be biased and an alternative estimator is proposed. The alternative estimator utilizes the same regression relationship to estimate the mean and standard deviation of the annual events at the ungaged site in order to estimate the d-day T-year low flow. The new estimator is shown to be theoretically appropriate. A first-order estimate of its variance is provided. When applied to an actual data set, the new estimator appears to be unbiased and to have the minimum mean square error among the five estimators considered.

INTRODUCTION

Water-quality management often requires estimation of low-flow streamflow characteristics at sites without long or perhaps any daily flow records. In particular, the annual minimum 7-day consecutive low flow which on average will be exceeded 9 of 10 years or in 19 of 20 years is often employed as a design flow. Thomas and Benson (1970) found that such 7-day 10- or 20-year low-flow values cannot be accurately estimated as a function of basin characteristics such as drainage area, stream channel length, or the percentage of the drainage area in forest or lakes. As an alternative, Riggs (1965, 1972) suggested that low- or base-flow measurements be obtained at the site in question and correlated with concurrent daily flows at a nearby gaged site at which a long flow record is available. Ideally the watershed for the nearby gaged site should be of similar drainage area size and geologic characteristics and have similar base-flow recession characteristics.

The base-flow measurements and the concurrent daily flows at the gaged site can be used to establish a relationship between the flows at the two sites. That relationship and the long-term flow record at the gaged site can then be used to estimate the low-flow frequency relationship at the ungaged site. Riggs (1965, 1972) considered primarily graphical procedures. Hardison and Moss (1972) and Gilroy (1972) substituted analytical regression procedures for establishing a linear relationship between the logarithms of the flows and for estimating the accuracy of the d-day T-year low-flow estimate for the ungaged sites. Here deficiencies with their approach are discussed. An improved d-day T-year low-flow estimator is developed and a first-order estimate of its variance provided.

THE BASIC PROBLEM

The analysis here is based on an assumed linear model between the logarithms of the annual minimum d-day low flows y_i at the ungaged site and those x_i at a nearby gaged site:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad \varepsilon_i \sim N(0,\sigma_\varepsilon^2). \tag{1}$$

The ε_i are independent residual errors which are assumed to be uncorrelated with the x_i. Letting μ_x, μ_y, σ_x^2, σ_y^2, and ρ_{xy} denote means, variances, and correlations of y and x, (1) implies that

$$\mu_y = \alpha + \beta\mu_x \tag{2}$$

and

$$\sigma_y^2 = \beta^2\sigma_x^2 + \sigma_\varepsilon^2. \tag{3}$$

Equation (2) can also be written

$$\alpha = \mu_y - \beta\mu_x. \tag{4}$$

Multiplying both sides of (1) by x and taking expectations yields the additional relationship

$$\rho_{xy}\sigma_y\sigma_x = \beta\sigma_x^2$$

or

$$\beta = \rho_{xy}\sigma_y/\sigma_x. \tag{5}$$

In order to use the model in (1) and annual d-day minima at the gaged site to estimate the distribution of d-day low flows at the ungaged site, estimators of α, β, and σ_ε^2 of the model in (1) are required. However, no record of d-day low flows at the ungaged site is available for this purpose. To overcome this difficulty, the logarithms of concurrent base-flow measurements $\tilde{y}_t$ and daily flows $\tilde{x}_t$ are used to estimate those parameters. Such observations should be separated by significant storm events so as to represent reasonably independent observations of the low-flow process. Thus, one would base their analysis on the assumption or approximation that the relationship between $\tilde{y}_i$ and $\tilde{x}_i$ can be described by

$$\tilde{y}_i = \alpha + \beta\,\tilde{x}_i + \varepsilon_i \qquad \varepsilon_i \sim N(0,\ \sigma_\varepsilon^2) \tag{6}$$

where α, β, and σ_ε^2 have the same values as the model in (1). In a subsequent section this assumption is evaluated by comparing values of α and β based on base-flow measurements (equation 6) and annual 7-day minima (equation 1). Although the α and β values vary significantly for given pairs of stations, on the average the assumption of similar values of α and β was reasonable. The assumption that the relationship between instantaneous base flows is the same as the relationship between the minimum 7-day annual low flows at the two sites is a heroic one. While this assumption appears reasonable for 7-day means, it may not be satisfactory for durations significantly longer than 7 days. Therefore, it is not advisable to use α and β values based on base-flow measurements to estimate the d-day T-year low flows (d significantly greater than 7) unless the assumption described above is tested.

The derivations to follow use the following definitions:

$m_x = \frac{1}{n} \sum_{t=1}^{n} x_t$	sample mean of the logarithms of annual d-day low flows at the gaged site
$m_{\tilde{y}} = \frac{1}{L} \sum_{t=1}^{L} \tilde{y}_t$	sample mean of the logarithms of base-flow measurements at the ungaged site
$m_{\tilde{x}} = \frac{1}{L} \sum_{t=1}^{L} \tilde{x}_t$	sample mean of the logarithms of concurrent daily flows at the gaged site
$s_x^2 = \frac{1}{(n-1)} \sum_{t=1}^{n} (x_t - m_x)^2$	sample variance of the logarithms of annual d-day low flows at the gaged site
$s_{\tilde{y}}^2 = \frac{1}{(L-1)} \sum_{t=1}^{L} (\tilde{y}_t - m_{\tilde{y}})^2$	sample variance of the logarithms of the base-flow measurements at the ungaged site
$s_{\tilde{x}}^2 = \frac{1}{(L-1)} \sum_{t=1}^{L} (\tilde{x}_t - m_{\tilde{x}})^2$	sample variance of the logarithms of concurrent daily flows at the gaged site (7)

n = number of years record at the gaged site

L = number of base-flow measurements and concurrent daily flows and also

$$b = \sum_{t=1}^{L} [\tilde{y}_t - m_{\tilde{y}}) (\tilde{x}_t - m_{\tilde{x}})]/(s_{\tilde{x}}^2(L-1))$$

$$a = m_{\tilde{y}} - b m_{\tilde{x}}$$

$$s_e^2 = \frac{1}{L-2} \sum_{t=1}^{L} (\tilde{y}_t - a - b\tilde{x}_t)^2 \qquad (8)$$

Here a, b, and s_e^2 are the ordinary least squares estimators of α, β, and σ_e^2 in (6). Furthermore, assume that the ε_i in (6), corresponding to the base-flow measurements, are independent.

The issue is how to estimate the logarithm of the d-day T-year low flow

$$Y_T = \mu_y + K_y \sigma_y \qquad (9)$$

at the y-site (ungaged site) given the logarithm of the d-day T-year low flow

$$X_T = \mu_x + K_x \sigma_x \qquad (10)$$

at the x-site (gaged site). Here K_y and K_x are the appropriate frequency factors for the two sites for the computed skew values at the T-year recurrence interval. If the logarithms of the d-day low flows at both sites are assumed to have the same standardized distribution, then $K_y = K_x$.

A tempting estimator of Y_T suggested by Riggs (1965, 1972) and Hardison and Moss (1972) is

$$\hat{Y}_T^{(R)} = a + b\,\hat{X}_T \tag{11}$$

where the assumption was made that $\hat{Y}_T^{(R)}$ would be unbiased. Equation 11 is the method I estimator described by Gilroy (1972). However, if $\hat{X}_T = X_T$, and with the assumptions and approximation employed here, then

$$\begin{aligned} \underset{a,b}{E}\,[\hat{Y}_T^{(R)}] &= \underset{a,b}{E}\,[a + b\,X_T] \\ &= \alpha + \beta\,[\mu_x + K_x\,\sigma_x] \\ &= (\mu_y - \beta\mu_x) + \rho_{xy}(\sigma_y/\sigma_x)\,[\mu_x + K_x\sigma_x] \\ &= \mu_y + \rho_{xy}\,K_x\,\sigma_y \end{aligned} \tag{12}$$

$\hat{Y}_T^{(R)}$ will be an unbiased and consistent estimator of Y_T only if

$$K_y = \rho_{xy}\,K_x$$

which is unlikely. If K_y and K_x are approximately equal, then $\hat{Y}_T^{(R)}$ is only unbiased if $\rho_{xy} = 1$, given the other assumptions employed here. In a subsequent section of this paper it is shown that K_y and K_x are approximately equal for watersheds in similar hydrologic environments.

A reasonable, consistent, and simple estimator of Y_T can be obtained by first using the base flows to calculate the estimators a and b of α and β. Then these values can be used with m_x and s_x^2 to estimate μ_y and σ_y^2 via equations (2) and (3). Our moment estimators are

$$\hat{\mu}_y = a + b\,m_x \tag{13a}$$

$$\hat{\sigma}_y^2 = b^2\,s_x^2 + s_e^2\left[1 - \frac{s_x^2}{(L-1)s_{\bar{x}}^2}\right] \tag{13b}$$

The extra factor in brackets in (13b) is employed to obtain an unbiased estimator of σ_y^2 as shown below. Clearly, for independent base-flow observations and annual d-day low-flow measurements at the x-site

$$E[\hat{\mu}_y] = (\mu_y - \beta\mu_x) + \beta(\mu_x) = \mu_y \tag{14a}$$

For fixed $\{\tilde{x}_1,\ldots,\tilde{x}_m\}$

$$E[\hat{\sigma}_y^2] = [\beta^2 + \text{Var}(b)]\ \sigma_x^2 + \sigma_e^2 \left[1 - \frac{\sigma_x^2}{(L-1)s_{\tilde{x}}^2}\right]$$

$$= \beta^2\sigma_x^2 + \sigma_e^2 + \sigma_x^2 \left[\text{Var}(b) - \frac{\sigma_e^2}{(L-1)s_{\tilde{x}}^2}\right] \tag{14b}$$

Thus, $\hat{\sigma}_y^2$ is also unbiased given that for every set $\{\tilde{x}_1,\ldots,\tilde{x}_m\}$,

$$\text{Var}(b) = \frac{\sigma_e^2}{(L-1)s_{\tilde{x}}^2} \tag{15}$$

provided the residuals in (6) are independent. Finally, our moment estimator of Y_T is

$$\hat{Y}_T^{(M)} = \hat{\mu}_Y + K_y\hat{\sigma}_y \tag{16}$$

where K_y is estimated by K_x.

OTHER ALTERNATIVES

Equations (13) and (16) provide one method for estimating low-flow quantiles at the ungaged y-site. This section considers three alternatives. These employ the y-site variance estimator suggested by Gilroy (1972), the maintenance of variance extension (MOVE.1) technique suggested by Hirsch (1982), and a mean scaling estimator, respectively.

Gilroy (1972) suggested that σ_y^2 could be estimated by

$$\hat{\sigma}_y^{2(G)} = b^2 s_x^2 + (1-r^2)\left(\frac{L-4}{L-2}\right)\frac{s_y^2 s_x^2}{s_{\tilde{x}}^2} \tag{17}$$

where r is the sample estimator of $\rho_{\tilde{x}\tilde{y}} = \beta\sigma_{\tilde{x}}/\sigma_{\tilde{y}}$. Equation 17 is the method II estimator of the variance of annual low flows described by Gilroy (1972). With the assumptions made here, this need not be a consistent estimator; consider large L and sample estimators replaced by their population values yielding

$$\hat{\sigma}_y^{2(G)} = \beta^2\sigma_x^2 + (1 - \beta^2\sigma_{\tilde{x}}^2/\sigma_{\tilde{y}}^2)\ \sigma_{\tilde{y}}^2\sigma_x^2/\sigma_{\tilde{x}}^2 \tag{18}$$

From (6) one can show that

$$\sigma_{\tilde{y}}^2 = \beta^2\sigma_{\tilde{x}}^2 + \sigma_e^2 \tag{19}$$

Thus (18) becomes

$$\hat{\sigma}_y^{2(G)} = \beta^2\sigma_x^2 + \sigma_e^2(\sigma_x^2/\sigma_{\tilde{x}}^2) \tag{20}$$

Hence, (17) will only be a consistent estimator of σ_y^2 if $\sigma_x^2 = \sigma_{\tilde{x}}^2$ given that σ_ε^2 is the same in equations (1) and (6). It is shown in a subsequent section that the sample estimators of σ_x^2 and $\sigma_{\tilde{x}}^2$, s_x^2 and $s_{\tilde{x}}^2$ can be quite different. After all, the x_t are the minimum annual d-day low flows in different years whereas the $\tilde{x}_t$ are just a set of low-flow measurements, separated by major storms, taken over a 2- or 3-year period.

Gilroy (1972) proved that (17) provided an unbiased estimator of σ_y^2 but to do so assumed that

$$\sigma_y^2/\sigma_x^2 = \sigma_{\tilde{y}}^2/\sigma_{\tilde{x}}^2. \tag{21}$$

By dividing our equations (3) by σ_x^2 and the corresponding relationship for the model in equation (6) by $\sigma_{\tilde{x}}^2$, one can see that when σ_ε^2 is the same in (1) and (6), (21) is true only if $\sigma_x^2 = \sigma_{\tilde{x}}^2$

Use of (13) corresponds in a sense to record augmentation (or moment transfer) as it has been discussed by Fiering (1962, 1963), Matalas and Jacobs (1964), Moran (1974), Loucks and others (1981), and Vogel and Stedinger (1985). A related procedure is maintenance of variance extension (MOVE.1) as proposed by Hirsch (1982) and Alley and Burns (1983). It is of interest to consider if such a procedure could be used in the situation of interest here.

The basic MOVE.1 procedure would generate synthetic logarithmic annual minimum d-day low flows for the ungaged site via

$$y_t' = m_y + \frac{s_y}{s_x}(x_t - m_x) \tag{22a}$$

This relationship is an approximation of

$$y_t'' = \mu_y + \frac{\sigma_y}{\sigma_x}(x_t - \mu_x) \tag{22b}$$

Equation (22b) could be used to convert a random x_t-sample into a corresponding set of observations y_t'', each of which in expectation would have mean μ_y and variance σ_y^2.

The difficulty posed by our application is that at most three annual minimum d-day low-flow values are available at the ungaged y-site so that there is not even a modest y-sample available with which to calculate the m_y and

s_y needed in equation (22a). In order to use the MOVE.1 technique to estimate d-day T-year low flow at the ungaged site, one can substitute $m_{\tilde{y}}$ and $s_{\tilde{y}}$ for m_y and s_y and $m_{\tilde{x}}$ and $s_{\tilde{x}}$ for m_x and s_x in equation (22a). Hirsch (1982) did not recommend the use of the MOVE.1 technique for this particular problem. However, the authors elected to evaluate the MOVE.1 technique by utilizing the estimates $m_{\tilde{y}}$, $s_{\tilde{y}}$, $m_{\tilde{x}}$, and $s_{\tilde{x}}$ in equation (22a) to translate $\hat{X}_T$ into a $\hat{Y}_T$. This is explained later.

Finally, a mean-scaling estimator of the following form is evaluated

$$\hat{Y}_T^{(S)} = X_T + (m_{\tilde{y}} - m_{\tilde{x}}) \tag{23}$$

This is the simplest estimator of all whereby the d-day T-year low flow at the gaged site is scaled by the difference in the geometric means of the base-flow measurements and daily flows to obtain the d-day T-year low flow at the ungaged site. All low-flow values are in logarithmic units. This estimator is computed without performing a regression analysis. If this estimator were sufficiently accurate, then considerable time would be saved in estimating d-day T-year low flows at ungaged sites. It is also reasonable if the flows at the two sites have the same distribution except for scale.

PRECISION OF $\hat{Y}_T^{(M)}$

Use of (13a,b) to estimate the mean μ_y and variance σ_y^2 of the annual minimum d-day low flows at the y-site to facilitate estimation of $Y_T = \mu_y + K_y\sigma_y$ is theoretically the most attractive alternative considered. Here we derive a first-order estimate of the variance of that estimator assuming that the residuals in (6) are normally distributed. To first order

$$\mathrm{Var}[\hat{Y}_T^{(M)}] = \mathrm{Var}(\mu_y) + \frac{K_y^2}{4\sigma_y^2}\,\mathrm{Var}(\sigma_y^2) + \frac{K_y}{\sigma_y}\;\mathrm{Cov}(\hat{\mu}_y,\ \hat{\sigma}_y^2) \tag{24}$$

Clearly, to first order,

$$\mathrm{Var}[\hat{\mu}_y] = \sigma_e^2 \left[\frac{1}{L} + \frac{(\mu_x - m_{\tilde{x}})^2}{(L-1)s_{\tilde{x}}^2}\right] + \beta^2\left(\frac{\sigma_x^2}{n}\right) \tag{25}$$

where we have neglected the second order term $\mathrm{Var}(b) \cdot \mathrm{Var}(m_x)$. Remember that L is the number of base-flow measurements and n is the number of years of record at the gaged site. To first order in 1/L and 1/n

$$[1 - \sigma_x^2/((L-1)s_{\tilde{x}}^2)]\ \mathrm{Var}(s_e^2) \cong \mathrm{Var}(s_e^2) \tag{26}$$

where terms such as $\mathrm{Var}(b^2)\cdot\mathrm{Var}(s_x^2)$ can be neglected. Thus,

$$\mathrm{Var}[\hat{\sigma}_y^2] \cong \sigma_x^4\ E[(b^2-\beta^2)^2] + \beta^4\ \mathrm{Var}(s_x^2) + \mathrm{VAR}(s_e^2)$$

Also, to first order $(b^2-\beta^2) = (b-\beta)(b+\beta) \cong 2\beta(b-\beta)$ so that

$$E[(b^2-\beta^2)^2] \cong 4\beta^2 \, Var(b) \tag{28}$$

Finally, to first order

$$Cov\ (\hat{\mu}_y, \hat{\sigma}_y^2) \cong 2\beta\sigma_x^2 \, [Cov\ (a,b) + \mu_x \, Var(b)]$$

$$= 2\beta\sigma_x^2 \, Var(b) \, (\mu_x - m_{\tilde{x}}) \tag{29}$$

Combining these results and also assuming that the x_t are themselves independent and normally distributed yields

$$Var[\hat{Y}_T^{(M)}] \cong \frac{\sigma_e^2}{L} + \frac{(\mu_x - m_{\tilde{x}})^2 \, \sigma_e^2}{(L-1)s_{\tilde{x}}^2} + \frac{\beta^2\sigma_x^2}{n}$$

$$+ \frac{K_y^{\,2}}{4\sigma_y^2}\left\{\frac{4\beta^2\sigma_x^4\sigma_e^2}{Ls_{\tilde{x}}^2} + \frac{2\beta^4\sigma_x^4}{n} + \frac{2\sigma_e^4}{L}\right\} + \frac{2\beta\sigma_x^2\left(\mu_x - m_{\tilde{x}}\right)K_y\sigma_e^2}{L\,\sigma_y\, s_{\tilde{x}}^2}$$

$$\cong \frac{\sigma_e^2}{(L-1)}\left\{1 + \frac{(\mu_x - m_{\tilde{x}})^2}{s_{\tilde{x}}^2} + \frac{K_y^{\,2}}{2\sigma_y^2}\left[\sigma_e^2 + \frac{2\beta^2\sigma_x^4}{s_{\tilde{x}}^2}\right] + \frac{2\beta K_y(\mu_x - m_{\tilde{x}})\sigma_x^2}{\sigma_y\, s_{\tilde{x}}^2}\right\}$$

$$+ \frac{\beta^2\sigma_x^2}{(n-1)}\left\{1 + \frac{\beta^2K_y^{\,2}\sigma_x^2}{2\sigma_y^2}\right\} \tag{30}$$

While (30) should be quite adequate for assessing the relative precision or sampling variability of $\hat{Y}_T^{(M)}$, it is only a first-order (in 1/n and 1/L) estimate derived assuming the residuals in (6) as well as the x_t are independent and normally distributed. Moreover, it does not incorporate the error introduced into the analysis by the assumption that the models in (1) and (6) have the same parameter values.

COMPARISON OF VARIOUS ESTIMATORS

Five different estimators of the 7-day 10-year low flow were applied to several sets of base-flow measurements and concurrent daily flows to better understand their relative performance. The data set consisted of 20 pairs of daily flow stations--10 in Pennsylvania, 6 in Indiana, 2 in Arkansas, 1 in Missouri, and 1 in South Carolina. All 40 stations have continuous daily flow records and most are quite long. The stations are listed in table 1 along with the length of record at each station. These stations are considered representative of sites in the humid east where low-flow estimates are most frequently required. For each pair, the first station selected was assumed to be the ungaged site. Then a nearby gaged station was selected with similar watershed size and base-flow recession characteristics. Five daily low flows were selected from each of 2 consecutive years from the published records for

Table 1.- Summary of stations used in the analysis and their lengths of record

Designated as Ungaged site	Gaged site	Length of record (years)		State
		Ungaged	Gaged	
03034500	03038000	43	45	Pennsylvania
03049000	03108000	42	41	Pennsylvania
03072000	03073000	42	51	Pennsylvania
07073500	07074000	41	46	Arkansas
07189000	07187000	43	41	Missouri
02196000	02192500	38	30	South Carolina
03017500	03015500	41	73	Pennsylvania
03324500	03324000	59	39	Indiana
03325500	03325000	34	58	Indiana
03340800	03340000	28	30	Indiana
03349500	03348500	21	55	Indiana
03353800	03353500	25	44	Indiana
03361000	03358000	32	33	Indiana
01518500	01518000	20	44	Pennsylvania
01532000	01531000	69	77	Pennsylvania
01542000	01541000	42	69	Pennsylvania
01567500	01568000	29	52	Pennsylvania
01571500	01574000	37	54	Pennsylvania
01601000	01603500	15	49	Pennsylvania
07049000	07049500	17	10	Arkansas

each pair. The years were selected randomly except that years with floods and high flows during the summer were avoided. The ten daily low flows for the designated ungaged site were assumed to be base-flow measurements and the ten concurrent low flows at the gaged site were assumed to represent the concurrent daily flows (H. C. Riggs, personal communication, 1982). These data were used to estimate a and b, the sample estimates of α and β. The regression coefficients, correlation coefficient, and selected statistics computed from the logarithms of the base-flow measurements and concurrent daily flows are given in table 2.

The regression analysis was repeated for the pairs of stations shown in table 2 by using the annual minimum 7-day low flows rather than the base-flow measurements. The purpose of this analysis was to compare the α and β values for equations (1) and (6). The regression coefficients and correlation coefficients based on the annual minimum 7-day low flows are shown in table 3 along with similar values based on the base-flow measurements (taken from table 2). The sample estimates of α and β, a and b, are quite similar for over half the data set. However, for certain pairs, the corresponding values of a and b differ significantly.

As discussed earlier, the 7-day T-year low flow (Y_T) can be estimated by equation 11 ($\hat{Y}_T^{(R)}$), by equation 16 ($\hat{Y}_T^{(M)}$), or equation 23 ($\hat{Y}_T^{(S)}$). In addition, the variance estimated by equation 17 can be combined with equation 13a to estimate Y_T as (Gilroy, 1972)

$$\hat{Y}_T(G) = \mu_y + K_y\ \sigma_y(G) \tag{31}$$

where K_y is estimated by K_x. The K_x value is computed from the logarithms of the annual d-day low flows at the gaged site using a Pearson Type III distribution and at-site sample skew. In equations 16 and 31, the K_x values shown in table 4 are used to estimate K_y because K_y cannot be computed for ungaged sites. However, the K_y values are given in table 5 for comparative purposes because in our data set observed 7-day low flows are available at the designated ungaged site.

The MOVE.1 technique (Hirsch, 1982) could be employed to transfer X_T to the ungaged site using the following equation

$$\hat{Y}_T^{(H)} = m_{\tilde{y}} + s_{\tilde{y}}/s_{\tilde{x}}\ (\hat{X}_T - m_{\tilde{x}}) \tag{32}$$

This corresponds to Riggs' quantile estimator in equation (11) if one also assumes that $\rho_{\tilde{x}\tilde{y}}$ is unity implying that β should be σ_y/σ_x. Were both $\rho_{\tilde{x}\tilde{y}}$ and ρ_{xy} equal to 1, then the bias problem identified in equation (12) would vanish.

Finally, the mean scaling estimator $\hat{Y}_T^{(s)}$ as shown in equation 23 could be used to estimate the 7-day T-year low flows. The appropriate values of $m_{\tilde{y}}$ and $m_{\tilde{x}}$ are given in table 2 and the X_T values are given in table 4.

The statistics for the logarithms of the minimum annual 7-day low flows and the estimated 7-day 10-year low flow for the gaged sites are given in table 4 ($G_{7,10}$ is the computed 7-day 10-year low flow in ft^3/s). The corresponding data for the ungaged sites are given in table 5 where m_y and s_y are

Table 2.--Summary of selected statistics computed from the base-flow measurements and concurrent daily flows

Identification number	Regression constant a	Regression coefficient b	Correlation coefficient r	Means $m_{\tilde{x}}$	Means $m_{\tilde{y}}$	Standard deviations $s_{\tilde{x}}$	Standard deviations $s_{\tilde{y}}$
	($\log_{10}$)			($\log_{10}$)	($\log_{10}$)	($\log_{10}$)	($\log_{10}$)
03034500-03038000	0.085	0.664	0.473	1.199	0.881	0.170	0.238
03049000-03108000	.537	.332	.649	1.158	.923	.455	.233
03072000-03073000	.502	.694	.792	.152	.608	.370	.325
07073500-07074000	-7.074	4.083	.656	1.189	.641	.109	.682
07189000-07187000	- .263	1.166	.944	2.299	2.418	.225	.278
02196000-02192500	- .814	1.175	.541	1.752	1.245	.401	.345
03017500-03015500	- .120	.970	.520	1.757	1.584	.121	.226
03324500-03324000	1.245	.552	.665	1.285	1.954	.406	.337
03325500-03325000	-1.850	1.355	.839	1.894	.716	.205	.330
03340800-03340000	- .172	.637	.726	1.859	1.012	.175	.153
03349500-03340500	-2.289	1.428	.626	2.163	.801	.201	.458
03353800-03353500	.535	.814	.818	1.171	1.488	.290	.289
03361000-03358000	1.455	.280	.692	1.080	1.757	.381	.154
01518500-01518000	.071	.489	.477	1.203	.659	.200	.205
01532000-01531000	-2.655	1.439	.929	2.436	.851	.329	.510
01542000-01541000	.137	.609	.757	1.723	1.187	.207	.166
01567500-01568000	- .028	.358	.643	1.397	.472	.143	.079
01571500-01574000	1.782	.162	.803	1.406	2.010	.362	.073
01601000-01603500	- .523	3.288	.845	.338	.590	.126	.490
07049000-07049500	-1.044	1.240	.958	1.804	1.195	.317	.410

Table 3.--Comparison of regression coefficients and correlation coefficients based on the base-flow measurements and annual minimum 7-day low flows

Identification number	Based on annual minimum 7-day low flows			Based on base-flow measurements		
	a	b	r	a	b	r
03034500-03038000	-0.462	0.941	0.799	0.085	0.664	0.473
03049000-03108000	- .255	.971	.769	.537	.332	.649
03072000-03073000	.501	.725	.821	.502	.694	.792
07073500-07074000	-6.755	3.994	.748	-7.074	4.083	.656
07189000-07187000	- .553	1.262	.891	- .263	1.166	.944
02196000-02192500	- .336	.809	.639	- .814	1.175	.541
03017500-03015500	- .441	1.108	.829	- .120	.970	.520
03324500=03324000	1.112	.191	.324	1.245	.552	.665
03325500-03325000	-1.515	1.077	.720	-1.850	1.355	.839
03340800-03340000	-1.326	1.241	.908	- .172	.637	.726
03349500-03348500	-3.574	1.941	.866	-2.289	1.428	.626
03353800-03353500	.752	.464	.792	.535	.814	.818
03361000-03358000	1.429	.264	.693	1.455	.280	.692
01518500-01518000	- .450	.964	.903	.071	.489	.477
01532000-01531000	*	*	*	-2.655	1.439	.929
01542000-01541000	.217	.540	.544	.137	.609	.757
01567500-01568000	- .450	.698	.861	- .028	.358	.643
01571500-01574000	1.729	.213	.716	1.782	.162	.803
01601000-01603500	- .258	1.756	.777	- .523	3.288	.845
07049000-07049500	- .900	-1.044	.972	-1.044	1.240	.958

* Not determined - annual 7-day low flows not readily available.

Table 4.--Summary of statistics for the minimum annual 7-day low flows and the computed 7-day 10-year low flow for the gaged site

Station number	Mean m_x	Standard deviation s_x	Frequency factor K_x	Low-flow estimate $X_{7,10}$	Low-flow estimate $G_{7,10}$
	$\log_{10}$	$\log_{10}$		$\log_{10}$	ft^3/s
03038000	1.151	0.312	-1.218	0.771	5.9
03108000	1.166	.230	-1.215	.886	7.7
03073000	.152	.493	-1.369	-.523	0.3
07074000	1.747	.118	-1.146	1.612	40.9
07187000	1.909	.214	-1.335	1.623	42.0
02192500	1.451	.418	-1.324	.898	7.9
03015500	1.728	.173	-1.211	1.519	33.0
03324000	.970	.282	-1.344	.591	3.9
03325000	1.798	.228	-1.290	1.504	31.9
03340000	1.642	.229	-1.317	1.340	21.9
03348500	2.014	.151	-1.301	1.818	65.7
03353500	.566	.630	-1.310	- .260	.55
03358000	.799	.458	-1.360	.176	1.5
01518000	1.277	.243	-1.289	.964	9.2
01531000	2.230	.176	-1.307	2.000	100.0
01541000	1.662	.198	-1.290	1.407	25.5
01568000	1.405	.173	-1.225	1.193	15.6
01574000	1.429	.332	-1.332	.987	9.7
01603500	.494	.188	-1.270	.255	1.8
07049500	1.514	.508	-1.342	.833	6.8

Table 5.--Summary of statistics for the minimum annual 7-day low flows and the computed 7-day 10-year low flow for the designated ungaged sites

Station number	Mean m_y	Standard deviation s_y	Frequency factor K_y	Low-flow estimate $Y_{7,10}$	$U_{7,10}$
	$\log_{10}$	$\log_{10}$	$\log_{10}$	$\log_{10}$	ft³/s
03034500	0.648	0.359	-1.314	0.176	1.5
03049000	.875	.287	-1.242	.519	3.3
03072000	.657	.408	-1.252	.146	1.4
07073500	.274	.618	-1.825	-.854	.14
07189000	1.856	.296	-1.340	1.459	2.8
02196000	.964	.416	-1.543	.322	2.1
03017500	1.496	.232	-1.223	1.212	16.3
03324500	1.294	.176	-1.326	1.061	11.5
03325500	.499	.332	-1.378	.041	1.1
03340800	.845	.283	-1.300	.477	3.0
03349500	.422	.305	-1.248	.041	1.1
03353800	1.081	.343	-1.305	.633	4.3
03361000	1.640	.177	-1.319	1.407	25.5
01518500	.726	.282	-1.292	.362	2.3
01532000	.942	.379	-1.347	.431	2.7
01542000	1.145	.184	-1.285	.908	8.1
01567500	.517	.135	-1.150	.362	2.3
01571500	2.059	.096	-1.292	1.935	86.1
01601000	.421	.403	-1.195	- .060	.87
07049000	.825	.464	-1.338	.204	1.6

the mean and standard deviation in log units for the annual series and $U_{7,10}$ is the computed 7-day 10-year low flow in ft^3/s. Table 6 gives the estimated mean and standard deviation of the annual minimum 7-day low flows as estimated by equations 13a, 13b, and 17. The values in table 6 are estimates of the m_y and s_y in table 5. The data given in tables 2-6 can be used to compute the estimators defined by equations 11, 16, 23, 31, and 32.

The five estimators were compared by computing the bias in log units (BIAS) and the root-mean-square error in log units (RMSE) by the following equations

$$BIAS_j = \frac{1}{20} \sum_{i=1}^{20} \left(Y_{T(i)}^{(j)} - Y_{T(i)} \right) \tag{33}$$

$$RMSE_j = \left[\frac{1}{20} \sum_{i=1}^{20} \left(Y_{T(i)}^{(j)} - Y_{T(i)} \right)^2 \right]^{0.5} \tag{34}$$

where $Y_{T(i)}^{(j)}$ is the j estimator by either equation 11, 16, 23, 31, or 32 for station i and $Y_{T(i)}$ is the 7-day 10-year low flow based on the actual record at the i^{th} ungaged site. The bias, and root-mean-square-errors are summarized in table 7 for the five estimators. As can be seen in table 7, the use of the regression equation to transfer the quantiles directly (equation 11), tends to overestimate the 7-day 10-year low flow. Our quantile MOVE.1 technique (equation 32) tends to underestimate but not seriously. The two estimators which utilize the mean and standard deviation of the annual 7-day low flows (equations 16 and 31) are about equal with Gilroy's estimator exhibiting less bias but having higher root-mean-square-error. Neither of the two methods exhibited a bias which was significantly different from zero at the 10 percent level. Overall, the estimator $\hat{Y}_T^{(M)}$ appears to be slightly better than $Y_T^{(G)}$ when applied to actual data and also requires fewer assumptions in its application. The mean scaling estimator (equation 23) is unbiased but has excessively high root-mean-square error. Because of the high root-mean-square error this estimator should not be used in actual practice.

The bias and root-mean-square error was also computed for the $\hat{Y}_T^{(M)}$ method using statistics based on the annual minimum 7-day low flows rather than the base-flow measurements. The bias was computed to be -0.023 log units and the root-mean-square error was 0.128 log units. These values are similar to the values of bias and root-mean-square error given in table 7 for $\hat{Y}_T^{(M)}$ utilizing the base-flow measurements. This close agreement implies that it is reasonable to use instantaneous base-flow measurements to estimate α and β for equation (1). In other words, the assumption that α and β values in equations (1) and (6) are similar is reasonable.

The point to be made here is that the estimator $\hat{Y}_T^{(R)}$, which has been used in practice, has the undesirable property of consistently overestimating Y_T. The estimator $\hat{Y}_T^{(M)}$, introduced in this paper, is shown to be a better estimator of Y_T and should be used instead.

Table 6.--Means and standard deviations of the annual minimum 7-day low flow at the designated ungaged sites as estimated from the gaged sites

Station number	Mean (equation 13a)	Standard deviation (equation 13b)	Standard deviation (equation 17)
	$\hat{\mu}_y$	$\hat{\sigma}_y$	$\hat{\sigma}_y^{(G)}$
	$\log_{10}$	$\log_{10}$	$\log_{10}$
03034500	0.849	0.268	0.393
03049000	.925	.195	.109
03072000	.608	.388	.412
07073500	.059	.688	.680
07189000	1.963	.265	.261
02196000	.891	.541	.824
03017500	1.556	.242	.292
03324500	1.780	.295	.217
03325500	.586	.353	.354
03340800	.874	.175	.189
03349500	.587	.414	.317
03353800	.996	.526	.601
03361000	1.679	.165	.173
01518500	.700	.206	.224
01532000	.555	.317	.268
01542000	1.150	.160	.151
01567500	.475	.084	.089
01571500	2.014	.069	.064
01601000	1.101	.660	.705
07049000	.834	.638	.651

Table 7.--Summary of bias and root-mean-square error for the five estimators*

Estimator	BIAS, $\log_{10}$ units	RMSE, $\log_{10}$ units
Equation 11, $Y_T^{(R)}$	0.125	0.248
Equation 16, $Y_T^{(M)}$	.042	.189
Equation 23, $Y_T^{(S)}$	.134	.279
Equation 31, $Y_T^{(G)}$	.021	.228
Equation 32, $Y_T^{(H)}$	-.059	.294

* The standard error of the estimated biases are 0.049, 0.042, 0.120, 0.052, 0.066 and of the estimated mean square errors are 0.016, 0.010, 0.036, 0.017, 0.045, for $\hat{Y}_T^{(R)}$, $\hat{Y}_T^{(M)}$, $\hat{Y}_T^{(S)}$, $\hat{Y}_T^{(G)}$, and $\hat{Y}_T^{(H)}$, respectively.

POSSIBLE ACCURACY IMPROVEMENTS

The average root-mean-square error of even the best method was a disappointing 0.19 (log base 10 units). This corresponds to a standard error of 46 percent. Of concern is how this error might be reduced. Two particular variables are sometimes subject to a hydrologist's control; these are ρ_{xy}, the cross correlation of the flows, and L, the number of concurrent measurements upon which the estimates of α, β, and σ_e^2 are based. By selecting a gage site whose low flows are highly correlated with the flows at the ungaged site of interest, the hydrologist can hope to obtain a pair of stations with a high ρ_{xy}. L is clearly an indication of the effort invested to obtain concurrent measurements. Figure 1 illustrates, using the first-order approximation in (30), a likely relationship between the standard error of estimate (in percent)

$$SE = 100 \left[\exp \{(2.3)^2 \, \mathrm{Var}[Y_T^{(M)}]\} - 1\right]^{1/2}$$

and values of ρ_{xy} and L. In this example, n = 25 or 50, Ky = -1.3, $\mu_x \cong m_{\tilde{x}}$ $\sigma_y = \sigma_x = 0.35$, and $\sigma_{\tilde{x}} = 0.25$. Note that $\sigma_e^2 = (1 - \rho_{xy}^2)\sigma_y^2$. The choice of K_y=-1.3 implies that the standard errors shown in figure 1 are comparable to those for the 10-year event.

What figure 1 shows is that for small L, the standard error decreases rapidly as L increases. As L becomes larger, the accuracy of $\hat{Y}_T$ is ultimately determined by the precision of m_x and s_x, the estimators of the moments of the flows at the gaged site. The standard error of the gaged site estimator of Y_T is 22 percent for n = 25 and 16 percent for n = 50. These numbers provide a standard with which to compare the values in figure 1. In this particular example, the precision of $\hat{Y}_T^{(M)}$ increases slowly beyond L = 20.

One can also see in the figure that for small L, the accuracy of $\hat{Y}_T$ is sensitive to ρ_{xy}: higher correlations yield more accurate estimators. This occurs because for fixed σ_y^2, large ρ_{xy} yields relatively small σ_e^2 meaning that a, b, and $\hat{\sigma}_e^2$ are relatively more accurate than they would be if ρ_{xy} were smaller.

In general, $\mathrm{Var}[\hat{Y}_T^{(M)}]$ decreases with increasing ρ_{xy}. However, as can be seen in Figure 1a corresponding to n = 25, the variance of $\hat{Y}_T$ for ρ =0.50 actually becomes slightly less than the variance for ρ = 0.70 or 0.90 when L is large. This makes sense in that our estimator of Y_T for small ρ depends as much or more on the parameters of the regression model and the estimated residual variance as it does on m_x and s_x^2, the sample moments of x_t; see equations (13), (2) or (3). This explains mathematically why with large L, it can happen that

$$\mathrm{Var}[\hat{Y}_T^{(M)} | \rho = 0.5] < \mathrm{Var}[\hat{Y}_T^{(M)} | \rho = 0.7]$$

However, such reversals of precision are probably an illusion because they occur in instances when the basic approximation upon which the analysis is based is probably not satisfactory.

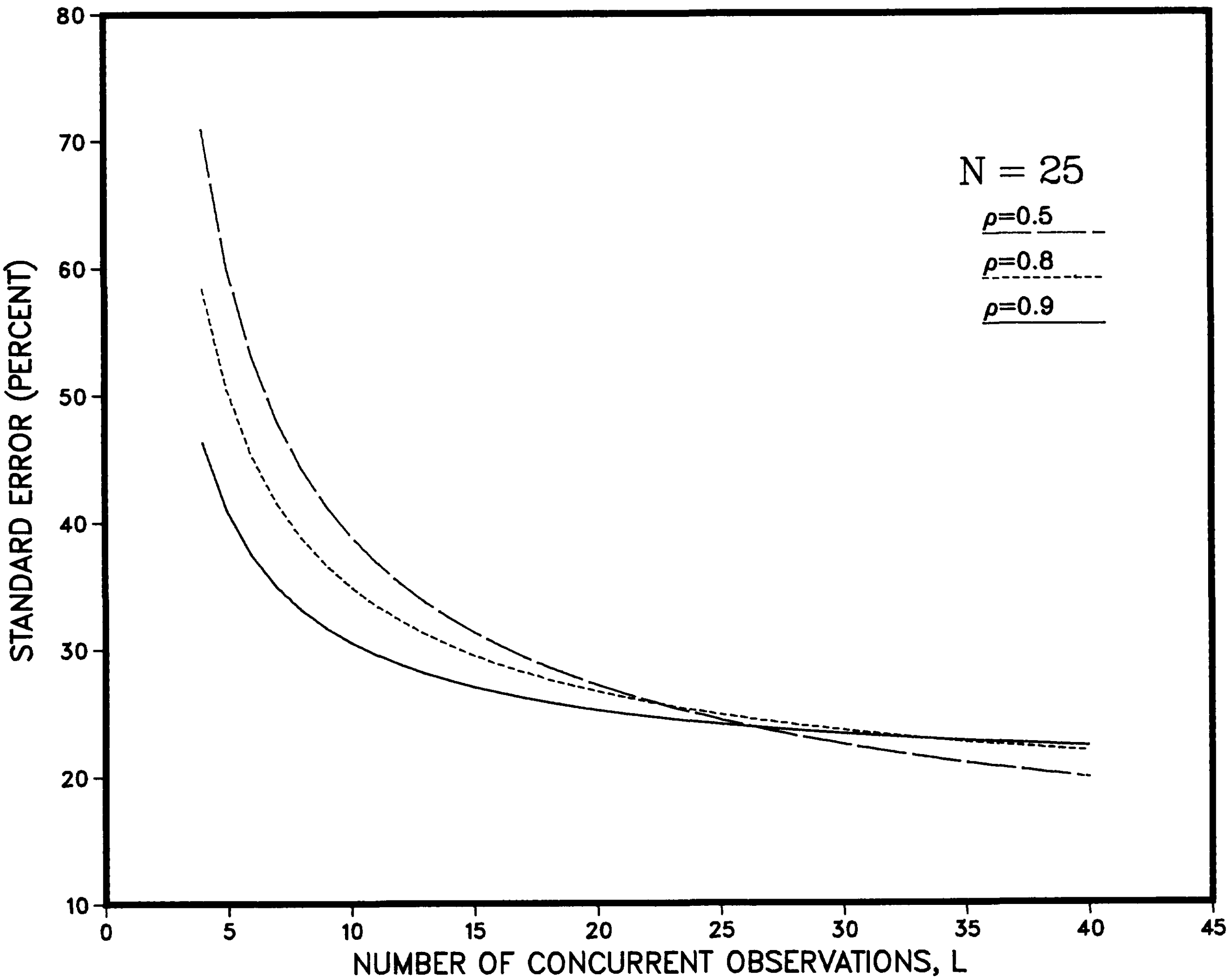

Figure 1a.--Relation of standard error to number of base-flow measurements (L) and correlation coefficient (ρ) for 25 years of record (n) at the gaged site.

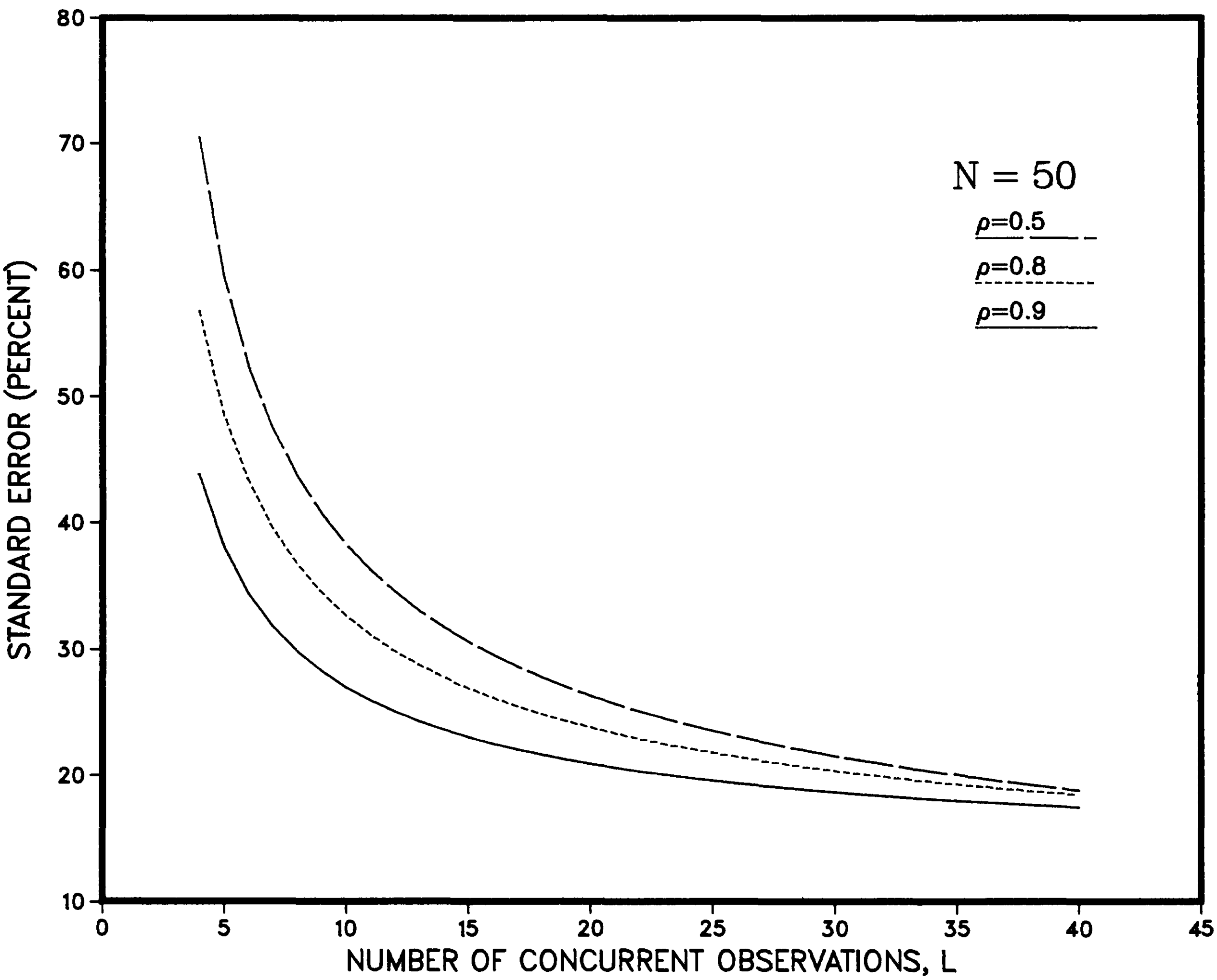

Figure 1b.--Relation of standard error to number of base-flow measurements (L) and correlation coefficients (ρ) for 50 years of record (n) at the gaged site.

The theory leading to our best estimator $Y_T^{(M)}$ was based on the approximation that the parameters α, β, and σ_ε^2 of the models in (1) and (6) were essentially the same. This assumption is probably true when $\rho_{xy} = 1$ but becomes an increasing less precise description of reality as ρ_{xy} decreases. For $\rho_{xy}=\beta=0$, the models in (1) and (6) become

$$y_t = \alpha + \varepsilon_t \qquad E\ \varepsilon_t = E\ \tilde{\varepsilon}_t = 0$$

$$\tilde{y}_t = \alpha + \tilde{\varepsilon}_t \qquad \mathrm{Var}\ \varepsilon_t = \mathrm{Var}\ \tilde{\varepsilon}_t = \sigma_\varepsilon^2$$

implying that both y_t and $\tilde{y}_t$ have the same mean α and variance σ_ε^2. We have argued above that it should be the case that

$$E[y_t] < E[\tilde{y}_t]$$

because y_t are annual minima whereas $\tilde{y}_t$ are only small values, the majority of which will exceed the minima for their year.

Another way of viewing the origin of this problem is by noting that when ρ_{xy} approaches unity, the model in (6) allows low flows $\tilde{x}_t$ at the x-gage to be mapped fairly precisely into the corresponding flows $\tilde{y}_t$ at the y-gage. It is then a reasonable approximation to assume that the annual low flow y_t at the y-gage occurred concurrently with the annual low flow x_t at the x-gage and that

$$y_t = \alpha + \beta\ x_t + \varepsilon_t$$

where α, β, and σ_ε^2 can be estimated using low base-flows and concurrent daily flows. However, when ρ_{xy} assumes small or even modest values, then it will frequently occur that y_t does not occur concurrently with x_t. Then concurrent base-flow and daily flow measurements at the two sites do not provide a reliable means of estimating the relationship between the annual minima. In retrospect, it would be our recommendation that these regression procedures not be employed to estimate the distribution of annual minima at ungaged sites unless ρ_{xy} exceeds about 0.70. Half of our station pairs failed to meet this threshold.

CONCLUSIONS

The problem of estimating the low-flow frequency distribution at an ungaged site based on a limited number of base-flow measurements at the ungaged site and daily flows at a gaged site has been considered. The base-flow measurements and concurrent daily flows can be used to derive a regression relationship between low flows at the two sites. Five different estimators of low-flow statistics at the ungaged site were evaluated. The recommended estimator utilizes what is known of the moments of the logarithms of the minimum d-day low flows at the gaged site to be used to estimate the mean and variance of the logarithms of the minimum d-day low flows at the ungaged sites. These estimates allow construction of a reasonable estimate of the d-day T-year low flow. Unfortunately, if the correlation between flows at the two sites is not relatively high, the relationship between baseflow and concurrent daily flow measurements need not describe the relationship between annual minima at the two sites. A first-order estimate of the variance of the recommended estimator is provided.

ACKNOWLEDGMENTS

We wish to thank E. Gilroy, M. Moss, R. Hirsch, and H. Riggs for their helpful comments and encouragement.

REFERENCES

Alley, W. M., and Burns, A. W., 1983, Mixed-station extension of monthly streamflow records: Journal of Hydraulic Engineering, v. 109, No. 10, American Society of Civil Engineers, p. 1272-1284.

Benson, M. A., and Matalas, N. C., 1967, Synthetic hydrology based on regional statistical parameters: Water Resources Research, v. 3, no. 4, p. 931-935.

Fiering, M. B, 1962, On the use of correlation to augment data: Journal of the American Statistical Association, v. 57, p. 20-32.

_____ 1963, Use of correlation to improve estimates of the mean and variance: U.S. Geological Survey Professional Paper 434-C.

Gilroy, E. J., 1972, Outline of derivations: in U.S. Geological Survey Water-Supply Paper 1542-B, p. 48-55.

Hardison, C. H., and Moss, M. E., 1972, Accuracy of low-flow characteristics estimated by correlation of base-flow measurements: U.S. Geological Survey Water-Supply Paper 1542-B.

Hirsch, R. M., 1982, A comparison of four record extension techniques: Water Resources Research, v. 18, no. 4, p. 1081-1088.

Loucks, D. P., Stedinger, J. R., and Haith, D. A., 1981, Water Resource Systems Planning and Analysis: Prentice-Hall, Inc., New Jersey.

Matalas, N. C., and Jacobs, B., 1964, A correlation procedure for augmenting hydrologic data: U.S. Geological Survey Professional Paper 434-E.

Moran, M. A., 1974, On estimators obtained from a sample augmented by multiple regression: Water Resources Research, v. 10, no. 1, p. 81-85.

Riggs, H. C., 1965, Estimating probability distributions of drought flows: Water and Sewage Works, v. 112, no. 5, p. 153-157.

_____ 1972, Low flow investigations: U.S. Geological Survey Techniques of Water-Resources Investigations, book 4, chapter B1.

Thomas, D. M., and Benson, M. A., 1970, Generalization of streamflow characteristics from drainage-basin characteristics: U.S. Geological Survey Water-Supply Paper 1975, 55 p.

Vogel, R. M., and Stedinger, J. R., 1985, Minimum variance streamflow record augmentation procedures: to appear in Water Resources Research.