Summary of U.S. Geological Survey Open-File Report 85-95

“Low-flow Frequency Estimation Using Base-flow Measurements”

by J. R. Stedinger and W. 0. Thomas, Jr.

The motivation for writing U.S. Geological Survey Open-File Report 85-95 was to (1) show that the linear regression approach of transferring low-flow statistics (e.g., 7-day, 10-year low flow) from an index station to a partialrecord site was biased and (2) propose an alternative unbiased estimator. Most analysts who use a mathematical technique, use linear regression analysis to relate base-flow measurements to concurrent daily flows at an index station. Therefore it is important to document that this technique is biased. The purpose of this summary is to clarify the important conclusions presented in Open-File Report 85-95.

Riggs (1972) describes how base-flow measurements at partial-record sites can be used to estimate low-flow statistics such as the 7-day, 10-year low flow. Riggs suggested the following estimator

$$\hat{Y}_T^{(R)} = a + b\hat{X}_T \qquad (1)$$

where $\hat{Y}_T^{(R)} =$ D-day, T-year low-flow value in log units at the partial-record site,

$\hat{X}_T =$ corresponding D-day, T-year low-flow value in log units at the index station, and

a, b = regression constant and coefficient.

Unless otherwise noted, all estimates of low flows, means and variances are expressed in log units (base 10). Estimates of a and b are made by relating base-flow measurements at the partial-record site to concurrent daily mean flows at the index station. Riggs (1972) used graphical techniques to estimate a and b. As the use of computers became more prevalent, many analysts began to use linear regression analysis to define a and b. This latter technique is shown to be biased in Open-File Report 85-95 and an unbiased technique is proposed.

The proposed unbiased technique (Stedinger and Thomas, 1985) involves first estimating the mean and variance of the annual D-day low flows at the partial-record site. These two statistics are then used to estimate the D-day, T-year low flows at the partial-record site. This is in contrast to equation 1 where the D-day, T-year low flow is estimated directly from the corresponding value at the index station. The following estimators for the mean and variance were proposed.

$$\hat{\mu}y = a + b\,m_x \quad \text{and} \qquad (2)$$

$$\hat{\sigma}_y^{\,2} = b^2 s_x^{\,2} + s_e^{\,2}\left[1 - \frac{s_x^{\,2}}{(L-1)s_x^{\,2}}\right] \qquad (3)$$

where $\hat{\mu}_y$ = mean of annual D-day low flows at the partial-record site,

$\hat{\sigma}_y{}^2$ = variance of annual D-day low flows at the partial-record site,

m_x, $s_x{}^2$ = mean and variance, respectively, of the annual D-day low flows at the index station,

$s_e{}^2$ = squared standard error of estimate of the relationship between the base-flow measurements and concurrent daily mean flows,

$s_x{}^2$ = variance of concurrent daily mean flows at the index station,
L = number of base-flow measurements at the partial-record site, and

a, b = regression constant and coefficient of base-flow measurements/daily mean flow relationship.

The D-day, T-year low flow $(\hat{Y}_T{}^{(M)})$ is then estimated by

$$\hat{Y}_T{}^{(M)} = \hat{\mu}_y + K_y\hat{\sigma}_y \tag{4}$$

where K_y is the Pearson Type III standard deviate for recurrence interval T and is estimated by K_x, the corresponding value at the index station. This substitution assumes that the logarithms of the D-day low flows at the partialrecord site and index station are both distributed as Pearson Type III random variables with equal skew coefficients. This assumption implies that one should choose index stations that have watershed characteristics (such as drainage area and soil type) that are similar to the watershed characteristics of the partial-record site.

A third estimator was also evaluated. This estimator is similar to equation 4 except that $\hat{\sigma}_y$ is estimated differently. Gilroy (1972) suggested that $\sigma_y{}^2$ could be estimated by

$$\hat{\sigma}_y{}^{2(G)} = b^2\, s_x{}^2 + (1 - r^2)\left(\frac{L - 4}{L - 2}\right)\frac{s_{\tilde{y}}{}^2\, s_x{}^2}{s_{\tilde{x}}{}^2} \tag{5}$$

where $\hat{\sigma}_y{}^{2(G)}$ = the variance of annual D-day low flows at the partial-record site,

r_2 = squared correlation coefficient of the base-flow measurements/daily mean flow relationship, and

$s_{\tilde{y}}{}^2$ = variance of base-flow measurements at the partial-record site.

The other terms are defined earlier. By substituting a $\hat{\sigma}_y{}^{(G)}$ into equation 4, a different estimator $\left(\hat{Y}_T{}^{(G)}\right)$ is obtained. This estimator of σ_y requires the assumption that the variance of the annual peaks at the index station (s_x^2) is equal to the variance of the concurrent daily mean flows $\left(s_{\tilde{x}}^2\right)$ at the index station (Stedinger and Thomas, 1985). On the average $s_{\tilde{x}}^2$ will be less than s_x^2 (see tables 2 and 4 in Open-File Report 85-95) because $s_{\tilde{x}}^2$ is a restrictive sample of daily mean flows corresponding to times when base-flow measurements were made at the partial-record site. Therefore estimation of σ_y^2 by equation 3 is theoretically more appealing and less restrictive than equation 5.

The final estimator evaluated was the maintenance of variance extension (MOVE.1) proposed by Hirsch (1982). This estimator is similar to equation 1 except that the regression coefficient (b) is estimated differently (i.e., r is assumed to equal 1.0). The MOVE.1 estimator $\hat{Y}_T^{(H)}$ of the D-day, T-year low flow is as follows.

$$\hat{Y}_T^{(H)} = m_{\tilde{y}} + \frac{s_{\tilde{y}}}{s_{\tilde{x}}}\left(\hat{X}_T - m_{\tilde{x}}\right) \qquad (6)$$

where $m_{\tilde{y}}$ = mean of base-flow measurements of partial-record site,

$m_{\tilde{x}}$ = mean of concurrent daily mean flows at index station, and other terms previously defined.

In order to compare this estimator to the linear regression approach, equation 1 can be rewritten as follows

$$\hat{Y}_T^{(R)} = m_{\tilde{y}} + r\frac{s_{\tilde{y}}}{s_{\tilde{x}}}\left(\hat{X}_T - m_{\tilde{x}}\right) \qquad (7)$$

where all the terms are previously defined. Note the only difference between equations 6 and 7 is the way the slope of the line is computed. The MOVE.1 estimator was originally suggested (Hirsch, 1982) for record extension or augmentation because it perserves or maintains the variance of the observed record. Hirsch (1982) did not recommend the use of MOVE.1 for estimating low flows at partial-record sites but Stedinger and Thomas (1985) elected to evaluate the technique because of its applicability in record augmentation.

In summary four estimators are evaluated. Two estimators directly relate the D-day, T-year low flow at the partial-record site to the corresponding value at the index station (linear regression, MOVE.1). The other two estimators (Stedinger and Thomas, 1985 and Gilroy, 1972) require estimating the mean and variance of the D-day annual low flows at the partial-record site. These statistics are then used to estimate the D-day, T-year low flow at the partial-record site.

Stedinger and Thomas (1985) show that the linear regression estimator ($\hat{Y}_T^{(R)}$ from equation 1) will be unbiased only if

$$K_y = r\,K_x \qquad (8)$$

where K_y and K_x are the Pearson Type III standard deviates introduced earlier and r is the sample correlation coefficient of the base-flow measurement/daily mean flow relationship. If the skew coefficients of annual D-day low flows at the partial-record site and index station are approximately equal, then K_y and K_x will be approximately equal. Under these conditions, $\hat{Y}_T^{(R)}$ is unbiased only if r = 1.0. Since r is always less than or equal to1.0, the tendency is for $\hat{Y}_T^{(R)}$ to underestimate the D-day, T-year low flow. This is illustrated later using an actual data set. The assumption of approximately equal skew coefficients at the partial-record site and index station (i.e., $K_y = K_x$) is more reasonable than assuming that the K_y value for the partial-record site is always less than K_x for the index station (see tables 4 and 5 in Open-File Report 85-95). This latter assumption is necessary for $\hat{Y}_T^{(R)}$ to be unbiased for $r \neq 1.0$.

The difference between the MOVE.1 estimator $\hat{Y}_T^{(H)}$ and the linear regression estimator $\hat{Y}_T^{(R)}$ can be determined by substracting equation 7 from equation 6 obtaining

$$\hat{Y}_T^{(H)} - \hat{Y}_T^{(R)} = (1 - r)\frac{s_{\tilde{y}}}{s_{\tilde{x}}}\left(\hat{X}_T - m_{\tilde{x}}\right) \qquad (9)$$

For recurrence intervals T greater than about 5 years, $\hat{X}_T$ will generally be less than $m_{\tilde{x}}$, the mean of the concurrent daily mean flows at the index station. Therefore the difference $\hat{Y}_T^{(H)} - \hat{Y}_T^{(R)}$ will be negative, implying that the linear regression estimate will generally be larger than the MOVE.1 estimate. Regardless of the recurrence interval T, equation 9 will give the difference between the two estimates.

The four estimators described above were applied to an actual data set of 20 pairs of continuous-record stations as described by Stedinger and Thomas (1985). For each pair of stations, one station was designated as a partial-record site and the estimates of the D-day, T-year low flows from the four techniques were compared to corresponding estimates based on the actual record. The four estimators were compared by computing the bias (BIAS) and root-mean-square error (RMSE) in both log units and cubic feet per second by the following equations

$$\mathrm{BIAS}_j = \frac{1}{20}\sum_{i-1}^{20}\left(\hat{Y}_{T_{(i)}}^{(j)} - Y_{T_{(i)}}\right) \qquad (10)$$

$$\mathrm{RMSE}_j = \left[\frac{1}{20}\sum_{i=1}^{20}\left(\hat{Y}_{T_{(i)}}^{(j)} - Y_{T_{(i)}}\right)^2\right]^{0.5} \qquad (11)$$

where $\hat{Y}_{T_{(i)}}^{(j)}$ is the j estimator for station i and $Y_{T_{(i)}}$ is the D-day, T-year low flow based on actual record at the ith partial-record site. The BIAS and RMSE values in cubic feet per second were obtained by estimating the D-day, T-year low flow in log units, converting the estimate to ft^3/s and comparing that to $Y_{T_{(i)}}$ in ft^3/s. The results of this analysis are shown in tables 1-3.

Table I.—Summary of bias and root-mean-square error for the 7-day, 2-year low flow.

Estimator	BIAS $\log_{10}$ units	RMSE $\log_{10}$ units	BIAS ft^3/s	RMSE ft^3/s
Regression $\hat{Y}_T^{(R)}$	0.048	0.237	3.33	10.94
Stedinger $\hat{Y}_T^{(M)}$	0.049	0.242	3.53	11.53
Gilroy $\hat{Y}_T^{(G)}$	0.053	0.241	3.58	11.23
MOVE.1 $\hat{Y}_T^{(H)}$	0.003	0.278	1.94	8.60

Table 2.—Summary of bias and root-mean-square error for the 7-day, 10-year low flow.

Estimator	BIAS $\log_{10}$ units	RMSE $\log_{10}$ units	BIAS ft^3/s	RMSE ft^3/s
Regression $\hat{Y}_T^{(R)}$	0.125	0.248	3.15	6.97
Stedinger $\hat{Y}_T^{(M)}$	0.042	0.189	1.46	4.12
Gilroy $\hat{Y}_T^{(G)}$	0.021	0.228	1.68	5.28
MOVE.1 $\hat{Y}_T^{(H)}$	-0.059	0.294	0.75	3.76

Table 3.—Summary of bias and root-mean-square error for the 7-day, 20-year low flow.

Estimator	BIAS $\log_{10}$ units	RMSE $\log_{10}$ units	BIAS ft^3/s	RMSE ft^3/s
Regression $\hat{Y}_T^{(R)}$	0.189	0.409	3.04	6.10
Stedinger $\hat{Y}_T^{(M)}$	0.083	0.325	1.19	3.10
Gilroy $\hat{Y}_T^{(G)}$	0.052	0.373	1.46	4.18

MOVE.1	$\hat{Y}_T^{(H)}$	-0.037	0.393	0.66	3.00

The conclusions to be drawn from tables 1-3 are as follows:

1. There is little difference among the methods when estimating the 7-day, 2-year low flow. This results because K_y and K_x in equation 8 given earlier are close to zero and, therefore, equation 8 is approximately satisfied for any value of r.

2. For estimating the 7-day, 10-year and 20-year low flows, the linear regression estimator $\hat{Y}_T^{(R)}$ is the most biased and generally has the highest RMSE of all estimators evaluated.

3. Furthermore, the linear regression estimator is the only biased estimator as the other three methods are about equal relative to bias.

4. The moment estimator $\hat{Y}_T^{(M)}$ suggested by Stedinger and Thomas (1985) is theoretically the most correct approach but really does not perform significantly better than Gilroy's or the MOVE.1 methods.

The overall conclusion to be drawn from tables 1-3 is that the linear regression estimator is biased and should not be used in actual studies. Although the data set is very limited, the empirical results support the theoretical argument given earlier. Either MOVE.1 $\left(\hat{Y}_T^{(H)}\right)$ or Stedinger $\left(\hat{Y}_T^{(M)}\right)$ will give better results than the linear regression estimator. The MOVE.1 method apparently performs well because the assumption that r always equals 1.0 satisfies equation 8.

Stedinger and Thomas (1985) provide an estimate of the variance of their estimator $\left(\hat{Y}_T^{(M)}\right)$. This variance can be computed from the following equation

$$\text{Var } Y_T^{(M)} \cong \frac{\sigma_e^{\,2}}{(L-1)}\left\{1+\frac{\left(m_x-m_{\tilde{x}}\right)^2}{s_{\tilde{x}}^{\,2}}+\frac{K_y^{\,2}}{2\hat{\sigma}_y^{\,2}}\left[\sigma_e^{\,2}+\frac{2b^2 s_x^{\,4}}{s_{\tilde{x}}^{\,2}}\right]+\frac{2bK_y\left(m_x-m_{\tilde{x}}\right)s_x^{\,2}}{\hat{\sigma}_y s_{\tilde{x}}^{\,2}}\right\}$$

$$+\frac{b^2 s_x^{\,2}}{(n-1)}\left\{1+\frac{b^2 K_y^{\,2} s_x^{\,4}}{2\hat{\sigma}_y^{\,2}}\right\} \qquad (12)$$

where m_x = mean of annual D-day low flows at the index station,

K_y is estimated by K_x at the index station,

$\hat{\sigma}_y^{\,2}$ is estimated by equation 3,

n = number of years of record at index station, and

all other terms are previously defined.

The variance of the D-day, T-year low flow at the partial-record site is a function of (1) the accuracy of the base-flow measurement/daily flow relationship represented by $\frac{\sigma_e{}^2}{(L-1)}$ { • • • } in equation 12 and (2) the accuracy of the D-day, T-year low flow at the index station represented by $\frac{b^2 s_x{}^2}{(n-1)}$ { • • • } in equation 12. Using "representative" values of the terms in equation 12, Stedinger and Thomas (1985) demonstrate that little improvement in accuracy results after about 20 base-flow measurements are obtained. An added incentive for using the Stedinger-Thomas estimator is that its variance can be evaluated as a function of the number of base-flow measurements (L) and the length of record at the index station (n).

W. 0. Thomas, Jr.
December 2, 1985

REFERENCES

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

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

Riggs, H. C., 1972, Low flow investigations: U.S. Geological Survey Techniques of Water-Resources Investigations, Book 4, Chapter Bl.

Stedinger, J. R., and Thomas, W. O., Jr., 1985, Low-flow frequency estimation using base-flow measurements: U.S. Geological Survey Open-File Report 85-95.