Mixed-Station Extension of Monthly Streamflow Records

By William M. Alley[1] and Alan W. Burns[2]

Abstract: Monthly streamflow records at a site are sometimes extended by exploiting interstation correlation of streamflow, often through simple linear regression with a base station having a long-term record. An approach is presented which selects a base station from among several in a region for filling in missing data. It differs from traditional approaches in that different stations can be selected as the base station to fill in different values. The approach also provides a decision rule for using only flow values from the same month or all flow values in developing the extension equation used for estimating a particular missing value. A previously presented extension equation to minimize variance bias of the flows estimated at the short-record station is shown to be applicable to this technique. The possibility of using this extension equations to adjust the results of multiple regression in order to account for variance reduction is also suggested.

Introduction

Monthly streamflow records are used in a wide variety of water resources studies. Unfortunately, records available for many streams are either too short to contain a sufficient range of hydrologic conditions or have periods of missing data. One solution to this problem is to rely on transfer of information from nearby stream gages. This can be done using the historic record of streamflows and extending it in time by exploiting the correlation between flows at the site of interest and concurrent flows at a nearby (base) gage (3). This is commonly done using simple linear regression. In the event that several nearby stations with concurrent long-term records are available, multiple regression or other multivariate approaches can be used (3).

Several deficiencies exist in record extension as commonly practiced using simple linear regression. The first of these deficiencies is that a single station is often used for extending the entire record. That station is usually one of a small set of nearby stations that have a complete long-term record of streamflow data. This method may ignore gaged flows at many other potentially important stations which could be used for filling in some of the missing record but have records covering different periods of time. Streamflow at these stations may come from drainage basins of similar topography and geology as the short-record station with resultant high correlations.

As an example, an extended streamflow record was desired for stations 09093500 (Parachute Creek at Parachute) and 09095000 (Roan Creek near DeBeque) which gage the flow of two streams in the Piceance basin

[1] Hydro., U.S. Geological Survey, Reston, Va.

[2] Hydro., U.S. Geological Survey, Lakewood, Colo.

Note.—Discussion open until March 1, 1984. To extend the closing date one month, a written request must be filed with the ASCE Manager of Technical and Professional Publications. The manuscript for this paper was submitted for review and possible publication on November 9, 1982. This paper is part of the *Journal of Hydraulic Engineering*, Vol. 109, No. 10, October, 1983. ©ASCE, ISSN 0733-9429/83/0010-1272/$01.00. Paper No. 18306.

of Colorado. The only long-term record available (for a nearby station without storage impoundment or transmountain diversions) was for station number 09304500 (White River near Meeker). However, the correlations between the two short-term sites and this base station were low. The period of record of the two Piceance stations and long-term station are shown in Fig. 1, as well as the period of record of several other nearby stations. These additional stations, which had no storage impoundments or transmountain diversions, were located on streams draining similar terrain, and had at least five years of concurrent data with one of the Piceance stations. It was thought that a method of including these stations in the information transfer would be useful. One approach would be multiple regression. However, the fact that each station was gaged during a different period of time hampers this approach. Two of the potential predictor stations (09303000 and 09304500) had fairly long concurrent records. However, the coefficent of the second variable in a multiple regression equation using both of these stations as predictor variables was not statistically significant at the $\alpha = 0.05$ level for either of the two dependent stations. An alternative would be to use streamflow at different stations as the independent variable to fill in different missing values. The station selected would be that expected to generate the least error. This is the approach that will be discussed in this paper.

Another common limitation of record extension techniques arises from the cyclic or seasonal nature of streamflow. A single regression model may be formulated for a station using the data from all concurrent flow periods. This approach (herein referred to as the noncyclic approach) assumes that the variability in flows is random rather than partly cyclic and partly random. An alternative is to develop separate equations for flows in each month or season. That approach (herein referred to as the cyclic approach) requires the estimation of more parameters (by a factor of 12, if monthly equations are developed) and is more tedious to apply. Often, either the cyclic or noncyclic approach is selected without a strong justification for the approach selected.

A third common deficiency of record extension techniques arises from the general limitations of regression analysis in preserving statistical moments. If the square of the product-moment correlation coefficient is less than 1.0, the expected value of the sample variance will be less than the population variance (8). This underestimation of variance is a disturbing

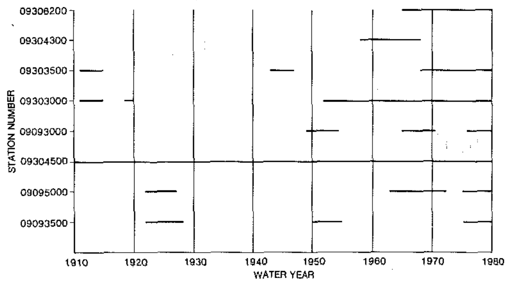

FIG. 1.—Years of Gaged Streamflow Data

feature of using regression analysis for record extension, since the extended record is often used for evaluating the severity and duration of hydrologic extremes.

Problem Definition

The nomenclature of Matalas and Jacobs (8) and Hirsch (5) is used here. All extensions are performed using log-transformed data. For the base station, the log-transformed flow values are denoted $x(i)$, where i is an index of time. For the short-record station the log-transformed flow values are denoted $y(i)$. The measured events for the two sequences are represented as $x(1), \ldots, x(N_1), x(N_1 + 1), \ldots, x(N_1 + N_2)$ and $y(1), \ldots, y(N_1)$ in which N_1 = length of short sequence and $(N_1 + N_2)$ = length of long sequence. Term N_1 is also the length of concurrent record. It is not necessary for the two sequences to begin simultaneously, nor need the measurements be consecutive. However, there is no loss of generality if the two sequences are represented as aforementioned. Additionally, x may denote variables other than streamflow.

The estimates of y based on an extension equation are denoted $\hat{y}(i)$, $i = 1, \ldots, N_1 + N_2$ and the complete extended record is denoted $\tilde{y}(i)$, $i = 1, \ldots, N_1 + N_2$, in which

$$\begin{aligned} \tilde{y}(i) &= y(i) \quad i = 1, \ldots, N_1 \\ &= \hat{y}(i) \quad i = N_1 + 1, \ldots, N_1 + N_2 \end{aligned}$$

The concurrent measurements of x and y are assumed to have a bivariate normal probability distribution with parameters μ_x, μ_y, σ_x^2, σ_y^2, and ρ, in which μ_x and σ_x^2 denote the population mean and variance for x and μ_y and σ_y^2 the population mean and variance for y. The parameter ρ is the product-moment correlation coefficient.

The sample mean and variance of a complete extended record are denoted $m(\tilde{y})$ and $s^2(\tilde{y})$. Other estimates of mean and variance are denoted in a similar manner. The subscripts 1 and 2 are used when the estimates are based on the measurement periods N_1 and N_2, respectively. No numerical subscript is used with those estimates based on the entire period $(N_1 + N_2)$. For example, the sample mean and variance of the first N_1 values of x are denoted $m(x_1)$ and $s^2(x_1)$ while the sample mean and variance for the entire base station record ($N_1 + N_2$ values) are denoted $m(x)$ and $s^2(x)$.

For simple linear regression, missing values of y are filled in by

$$\hat{y}(i) = \hat{\beta}_1 + \hat{\beta}_2 (x(i) - m(x_1)) \quad \ldots\ldots \quad (1)$$

in which $\hat{\beta}_1 = m(y_1)$

$$\hat{\beta}_2 = r \frac{s(y_1)}{s(x_1)}$$

in which r = product-moment correlation coefficient between the N_1 concurrent measurements of x and y.

An extended streamflow record may be desired to obtain: (1) Estimates of parameters such as the mean and variance of the dependent variable; (2) prediction of dependent station flows given values of x; or (3) a typical realization of dependent station flows given the values of x. The parameters for purpose 1 can be estimated without resorting to record extension (8). The usual intent of record extension is to produce a time series that possesses statistical properties like those of an actual record for the station. Therefore, purpose 3 is probably more often of concern than purpose 2 and is emphasized in this paper.

METHODOLOGY

Our method selects the base stations from among several in a region for filling in missing data. It differs from traditional approaches since a different station can be selected as the base station for different missing values of the same short-term station. Flow values from a station having missing values may be used to fill in missing values for another station. However, estimated values are never used to fill in other estimated values. The method can consider both cyclic and noncyclic regression equations for each particular prediction. Monthly streamflow data may display large serial correlation. In that event, one may want to consider only cyclic regression equations.

A skeleton flow chart of the method is shown in Fig. 2. It consists of 11 basic steps.

Step 1.—Initialize variables and setup matrix **Z**. Each row of matrix **Z** contains the log-transformed monthly streamflow data for either a dependent or potential independent station. Missing values are designated by large negative values (−100). Thus, matrix **Z** is of order N_s by N where N_s is the total number of stations and $N = N_1 + N_2$.

Step 2.—For a station at which missing flow values are to be estimated, say row k of matrix Z, create a $1 \times N$ matrix Y from row k (i.e., $y(i) = z(k, i)$ in which y and z are elements of the **Y** and **Z** matrices). Matrix **Y** is used to store the extended record for the dependent station. Also, create a $1 \times N$ matrix V consisting of the error criteria for the flow values. Initially,

$$v(i) = 0 \quad \text{if } z(k, i) > -100 \text{ (i.e., not a missing value)}$$

$$= 10^6 \quad \text{otherwise}$$

Step 3.—Select a row 1 of matrix Z $(1 \neq k)$ as a potential predictor variable.

Step 4.—Estimate noncyclic regression coefficients, $\hat{\beta}_1$ and $\hat{\beta}_2$ using row 1 of matrix **Z** as independent variable and row k as dependent variable. If $\hat{\beta}_2$ is significantly different from zero at the $\alpha = 0.05$ level, proceed to step 5. Otherwise, go to step 7.

Step 5.—Identify a col i of matrix **Z** such that $z(k, i) \leq -100$ and $z(1, i) > -100$.

Step 6.—Compute error criterion, $e(i)$. If $e(i) < v(i)$, then let $x(i) = z(1, i)$, and solve equation 1 for $\hat{y}(i)$. Store this value in matrix Y and replace the current value of $v(i)$ in matrix V by its new estimate of $e(i)$. Repeat steps 5 and 6 until the noncyclic extension equation between variables k and 1 has been evaluated for all missing values of the dependent variable.

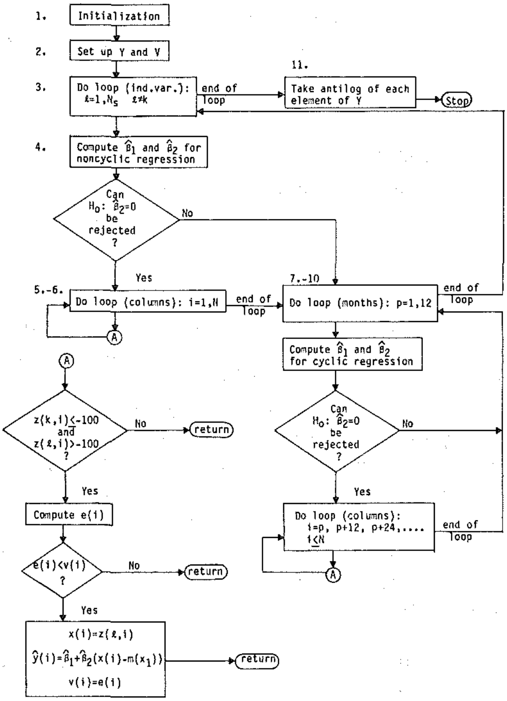
1.
Initialization
2.
Set up Y and V
3.
Do loop (ind.var.): ℓ=1,N_s ℓ≠k
end of loop
11.
Take antilog of each element of Y
Stop
4.
Compute β̂1 and β̂2 for noncyclic regression
Can H0: β̂2=0 be rejected ?
No
Yes
5.-6.
Do loop (columns): i=1,N
end of loop
A
7.-10
Do loop (months): p=1,12
end of loop
Compute β̂1 and β̂2 for cyclic regression
Can H0: β̂2=0 be rejected ?
No
Yes
Do loop (columns): i=p, p+12, p+24,.... i≤N
end of loop
A
A
z(k,i)≤-100 and z(ℓ,i)>-100 ?
No
return
Yes
Compute e(i)
e(i)<v(i) ?
No
return
Yes
x(i)=z(ℓ,i)
ŷ(i)=β̂1+β̂2(x(i)-m(x1))
v(i)=e(i)
return

FIG. 2.—Schematic of Mixed-Station Extention Methodology for a Single Dependent Station (Dependent Station Index = k, Independent Station Index = 1)

Step 7.—Estimate $\hat{\beta}_1$ and $\hat{\beta}_2$ for a cyclic regression of variable k on variable 1 using only cols of matrix **Z** corresponding to January flows. If $\hat{\beta}_2$ is significantly different from zero at the $\alpha = 0.05$ level, proceed to step 8. Otherwise, repeat step 7 for the next month.

Step 8.—Identify a column i of matrix **Z** corresponding to January (or month of interest) such that $z(k,i) \leq -100$ and $z(1,i) > -100$.

Step 9.—Compute error criterion, $e(i)$. If $e(i) < v(i)$, then let $x(i) = z(1,i)$, and solve Eq. 1 for $\hat{y}(i)$. Store this value in matrix Y and replace the current value of $v(i)$ in matrix V by its new estimate of $e(i)$. Repeat steps 8 and 9 until the cyclic extension equation between January (or the month of interest) flows at stations k and 1 has been evaluted for all the missing January (or the month of interest) flow values at station k.

Step 10.—Repeat steps 7–9 for February, March, . . ., and December flows. Then return to step 3 and try another potential independent variable. Continue in this manner until all potential independent variables have been investigated.

Step 11.—Take the antilog of all elements of Y to determine the extended streamflow record.

The above procedure outlines the methdology for extending the record of a single dependent station. To extend records at an additional site steps 2–11 are repeated for a new dependent variable, k, and so on.

The method estimates each missing value using the independent station that results in the "least error." Thus, the definition of the error criterion is critical to the analysis. The criterion selected was the standard error of prediction of an individual value, SEP(i)

$$\mathrm{SEP}(i) = \mathrm{SEE}\sqrt{1 + \frac{1}{N_1} + \frac{[x(i) - m(x_1)]^2}{(N_1 - 1)\, s^2(x_1)}} \quad \cdots\cdots (2)$$

in which SEE = standard error of estimate

$$(\mathrm{SEE})^2 = \frac{1}{N_1 - 2} \sum_{i=1}^{N_1} [\hat{y}(i) - y(i)]^2 \quad \cdots\cdots (3)$$

Assuming that the errors are lognormally distributed (1), the standard deviation of the untransformed variable q is

$$\sigma_q = \mu_q[\exp(\sigma_y^2) - 1]^{1/2} \quad \cdots\cdots (4)$$

in which σ_y^2 = population variance of the transformed flow value and μ_q and σ_q = population mean and standard deviation of the flow value q. Thus, $e(i)$, the standard error of prediction in percent, can be estimated by dividing Eq. 4 by μ_q, replacing σ_y by SEP (i) and multiplying by 100

$$e(i) = 100\, [\exp(\mathrm{SEP}(i)^2) - 1]^{1/2} \quad \cdots\cdots (5)$$

Application to Piceance Streams

Returning to the set of stations shown in Fig. 1, the missing data for the two Piceance stations (09093500 and 09095000) were filled in for four different cases. For cases 1–3, station 09304500 (White River near Meeker) was selected as the only potential independent variable. Only noncyclic regressions were performed for case 1 and only cyclic regressions were performed for case 2. For case 3 either the cyclic regression or noncyclic regression was selected for estimating a particular missing value based on equation 5. Case 4 selected either cyclic or noncyclic regressions and

TABLE 1.—Average Standard Error of Prediction for Stations in Piceance Basin

Case (1)	Independent station(s) (2)	Consideration of seasonality (3)	Average standard error of prediction for flows at station 09093500, as a percentage (4)	Average standard error of predictions for flows at station 09095000, as a percentage (5)	Cyclic regression predictions, as a percentage (6)	Noncyclic regression predictions, as a percentage (7)
1	09304500 only	Noncyclic regressions only	157	67	0	100
2	09304500 only	Cyclic regressions only	95 (17)[a]	63 (8)[a]	100	0
3	09304500 only	Cyclic or noncyclic	92	49	70	30
4	All streamflow gaging stations	Cyclic or noncyclic	83	46	74	26

[a]Numbers in parentheses are the percent of missing values for which there were no significant regressions at the $\alpha = 0.05$ level.

the independent variable was selected using the flows at the remaining 7 stations and the mixed-station methodology.

Brief descriptions of the four cases and average standard errors of prediction in percent for each of the two Piceance stations are shown in Table 1. Considerable reduction in estimated prediction errors is observed as one examines case 1–case 4. In regressions with station 09304500 as the only independent variable the results for cyclic regression (case 2) were in general better than those for noncyclic regressions (case 1), although some of the missing values could not be filled in for case 2.

For cases 3 and 4 both cyclic and noncyclic regressions were common. For case 4 all stations were used to estimate some of the missing values. Station 09304500 was, of course, used as an independent variable the most times largely due to the dearth of data at the other stations between 1910 and 1950. However, for case 4 it contributed to less than 50 percent of the predictions.

Maintenance of Variance

The procedure discussed thus far is an efficient method for reducing estimated errors in predicted values of streamflow. Unfortunately, the use of regression analysis often results in underestimates of the variance in the extended record. This variance reduction should be lessened by the mixed-station methodology, since better overall regressions between measured and simulated data should be achieved. However, the importance of accurately estimating hydrologic extremes suggests that the procedure should incorporate some method of preserving variance.

Matalas and Jacobs (8) demonstrated that unbiased estimates of mean and variance are achieved if noise is added to the regressed values, where the noise is a random variable with zero mean and variance proportional to the variance of the observations for the short sequence about the regression line. One problem with modifying their approach for data fill in is that studies of the same sequence of x and y by different investigators will almost surely lead to different values of $\hat{y}(i)$, $i = N_1 + 1, \ldots, N_1 + N_2$. However, this approach has found use in practice (2), and it may be particularly useful when preservation of interstation correlations between the site of interest and the base station is important.

Hirsch (5) suggests two other approaches which he refers to as MOVE.1 and MOVE.2 (Maintenance of Variance Extension, Types 1 and 2). The MOVE.1 equation is

$$\hat{y}(i) = m(y_1) + \frac{s(y_1)}{s(x_1)}(x(i) - m(x_1)) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (6)$$

Eq. 6 is the same as Eq. 1 except the r in $\hat{\beta}_2$ is omitted. This equation has often been referred to as the line of organic correlation, and is discussed in a hydrologic context by Kritskiy and Menkel (7). Using Eq. 6, the estimates of $m(\tilde{y})$ and $s^2(\tilde{y})$ are both asymptotically unbiased as N_1 approaches infinity (5).

Hirsch (5) presents Eq. 6 as an alternative to least-squares regression for filling in missing values. However, it can be demonstrated that the same results would be obtained (for a single independent variable) if regression was applied first, followed by application of MOVE.1 as a post-regression adjustment. For example, assume that values of $\hat{y}(i)$, $i = 1, \ldots, N_1 + N_2$ have been determined using regression analysis. Considering $\hat{y}(i)$ as the independent variable and $y(i)$ again as the dependent variable and applying MOVE.1 gives

$$\hat{\hat{y}}(i) = m(y_1) + \frac{s(y_1)}{s(\hat{y}_1)}\,[\hat{y}(i) - m(\hat{y}_1)] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (7)$$

in which $m(\hat{y}_1)$ and $s(\hat{y}_1)$ are the mean and standard deviation of $\hat{y}(i)$, $i = 1, \ldots, N_1$. Substituting Eq. 1 into Eq. 7 gives

$$\hat{\hat{y}}(i) = m(y_1) + \frac{s(y_1)}{s(\hat{y}_1)}\left[m(y_1) + r\,\frac{s(y_1)}{s(x_1)}\,(x(i) - m(x_1)) - m(\hat{y}_1)\right] \ldots\ldots\ldots\ldots \quad (8)$$

Since $m(y_1) = m(\hat{y}_1)$ and $r = s(\hat{y}_1)/s(y_1)$, Eq. 8 reduces to $\hat{\hat{y}}(i) = m(y_1) + [s(y_1)/s(x_1)][x(i) - m(x_1)]$ which is equivalent to Eq. 6. Thus, for a single independent station MOVE.1 can be viewed as an adjustment to regression in order to maintain variance, as well as an alternative to regression. MOVE.1 may have utility for applications other than those addressed by Hirsch (5) or this paper. For example, if streamflow records are extended using multiple regression, MOVE.1 might be used to adjust the regression results to adjust for variance reduction.

If MOVE.1 is considered to be a post-regression adjustment, then Eq. 5 is a reasonable error criterion for comparing extension equations. However, it should only be interpreted as an error criterion for comparing extension equations and not as standard error of prediction for MOVE.1.

Use of MOVE.1 results in preservation of the sample estimates of variance (and mean) from the historical record. For short records the estimate of variance may be unreliable. However, Monte Carlo and empirical experiments by Hirsch (5) suggests that, even for relatively small values of N_1, MOVE.1 tends to produce less biased estimates of the variance and extreme order statistics of an extended record, than does simple linear regression.

In MOVE.1 the only four parameters are the sample means and variances of x and y estimated from the first N_1 observations. In MOVE.2 these same parameters are used but the sample estimates of mean and variance for x are based on $N_1 + N_2$ observations and the sample mean and variance estimates for y are based on the historical values of y and on information transfer from the x sequence. The MOVE.2 equation is

$$\hat{y}(i) = \hat{m}(y) + \frac{\hat{s}(y)}{s(x)}\,[x(i) - m(x)] \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (9)$$

in which $\hat{m}(y)$ and $\hat{s}^2(y)$ are unbiased estimators of μ_y and σ_y^2 developed by Matalas and Jacobs (8)

$$\hat{m}(y) = m(y_1) + \frac{N_2}{(N_1 + N_2)}\,r\,\frac{s(y_1)}{s(x_1)}\,(m(x_2) - m(x_1)) \ldots\ldots\ldots\ldots\ldots\ldots \quad (10)$$

$$\hat{s}^2(y) = \frac{1}{N_1 + N_2 - 1}\left\{(N_1 - 1)s^2(y_1) + (N_2 - 1)\,r^2\,\frac{s^2(y_1)}{s^2(x_1)}\,s^2(x_2)\right.$$

$$+ \frac{N_2\,(N_1 - 4)(N_1 - 1)}{(N_1 - 3)(N_1 - 2)}\,(1 - r^2)\,s^2(y_1)$$

$$\left. + \frac{N_1 N_2}{(N_1 + N_2)}\,r^2\,\frac{s^2(y_1)}{s^2(x_1)}\,(m(x_2) - m(x_1))^2 \right\} \quad \text{(11)}$$

Matalas and Jacobs (8) also present criteria for judging whether or not Eqs. 10 and 11 are better estimators of the population values of the mean and variance than the estimates based on the short sequence. Typically, the correlation coefficient should exceed about 0.65 to obtain improvement in the estimate of variance using Eq. 11. Eq. 5 was also used in this study with MOVE.2 as the error criterion for comparing extension equations.

Empirical Examination of Methods.—An empirical example was designed to examine the utility of MOVE.1 and MOVE.2, in conjunction with mixed-station record extension, for preserving the variance and low-flow duration characteristics of a streamflow record. The example used concurrent, 50-year monthly streamflow records from nine streamflow-gaging stations in west central Virginia. The stations that were used are listed in Table 2.

The experimental design is as follows: Each of the first eight stations listed in Table 2 was considered to be the short-record station with only 10 years of data. This record was considered, in turn, to be years 1–10, 11–20, 21–30, 31–40, and 41–50, respectively. Thus, 40 different realizations of an extended streamflow record (8 possible dependent stations times 5 nonoverlapping 10-year base periods) were generated for a given approach. This analysis was performed for each of four approaches:

1. REG-LITTLE—The full 50-year record at the Little River at Graysonton was assumed to be available as the sole base station. Record extension was by regression.
2. REG-SEP—In addition to the Little River at Graysonton the full 50-

TABLE 2.—Information on the Nine Streamflow-Gaging Stations

Station number and name (1)	Drainage area, in square kilometers (2)	Mean discharge, in cubic meters per second (3)	Historic correlation coefficient with flows of station 03170000 (4)
02030500 Slate R. nr. Arvonia, Va.	585	6.5	0.68
02040000 Appomattox R. at Mattoax, Va.	1,880	20	0.74
02013000 Dunlap Cr. nr. Covington, Va.	425	4.7	0.74
02016000 Cowpasture R. nr. Clifton Forge, Va.	1,194	15	0.75
02017500 Johns Cr. at New Castle, Va.	269	3.6	0.82
02018000 Craig Cr. at Parr, Va.	852	11	0.83
02055000 Roanoke R. at Roanoke, Va.	1,023	11	0.92
03167000 Reed Cr. at Grahams Forge, Va.	640	7.6	0.80
03170000 Little R. at Graysonton, Va.	777	10	1.0

year streamflow records at the other stations were assumed to be available for record extension. Each missing value of the dependent station was filled in using the mixed-station methodology. Record extension was by regression.

3. MOVE.1-SEP—This approach was the same as REG-SEP, except MOVE.1 was applied in place of regression.

4. MOVE.2-SEP—This approach was also the same as REG-SEP, except MOVE.2 was applied in place of regression.

A decision to select a cyclic or noncyclic regression was also determined using equation 5 for all four approaches.

For each of the 40 extended records for a particular approach, various statistics were recorded. These included the first- and fifth-order statistics (lowest and fifth lowest) in the annual series of minimum 1-month volumes for the extended portion of the record and the standard deviation of monthly flows. The ratio, *U*, of a statistic for the extended record to that for the historic record was computed for each extended record.

The results of the analysis are summarized in Figs. 3 and 4. In these figures the box plots represent the distribution of all 40 values of *U* for a given statistic and approach. The accuracy of each approach can be evaluated by the degree of dispersion in the box plots for that approach, by the closeness of the median to a value of 1.0, and by the symmetry of the box about a value of 1.0. Although, one would expect some skewness in the box plots since the value of *U* is bonded below by zero.

The reduction of variance of the two unadjusted regression approaches is demonstrated in Fig. 3 as the standard deviations of the flow data tended to be underestimated. REG-SEP resulted in a smaller dis-

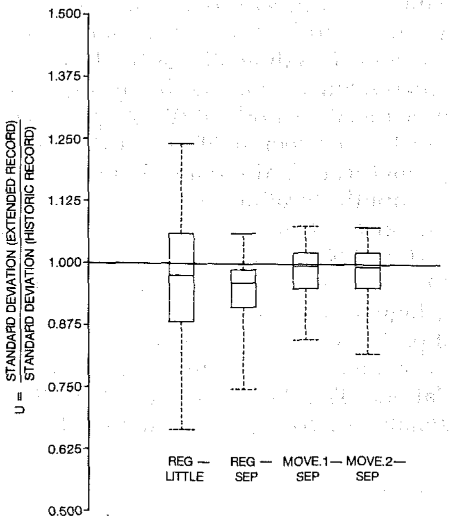

FIG. 3.—Box Plots of *U* Values for Standard Deviation of Monthly Flows Box Plots Show Median, Upper and Lower Quartiles, and Maximum and Minimum Values. The Sample Size is 40

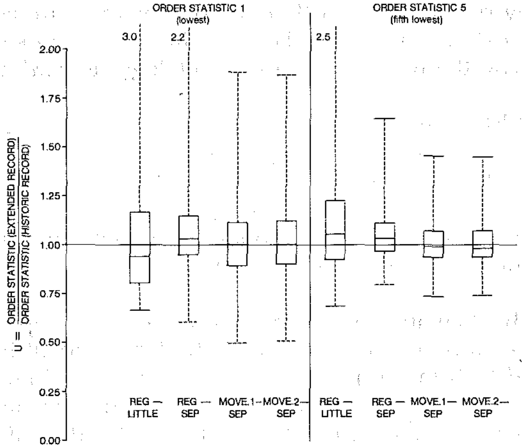

FIG. 4.—Box Plots of *U* Values for 1-month Duration Low Flows. Box Plots Show Median, Upper and Lower Quartiles, and Maximum and Minimum Values. The Sample Size is 40

persion of the values of *U* about 1.0. However, over 75 percent of the computed standard deviations were less than the historical values. Use of MOVE.1 and MOVE.2 results in median *U* values close to 1.0 and somewhat symmetrical distributions about 1.0.

The effect of variance reduction on low-flow statistics is illustrated in Fig. 4. With one exception, the median values of *U* for all box plots of REG-LITTLE and REG-SEP are greater than 1.0 and the box plots are skewed toward values of *U* greater than 1.0. Thus, as expected, the severity of low flows tends to be underestimated by both regression approaches. This problem is less dramatic for the REG-SEP regressions than the REG-LITTLE regressions, probably due to the better fit of the REG-SEP models. Use of MOVE.1 and MOVE.2 again results in median *U* values very close to 1.0 (between 0.997 and 1.008 for the four box plots) and in general symmetrical distributions about 1.0. Similar results were found for 2- and 3-month duration low flows.

Root-mean-square-errors (RMSE) of predicted streamflow were 5.66 m^3/s for REG-LITTLE, 3.58 m^3/s for REG-SEP, and 3.65 m^3/s for both MOVE.1 and MOVE.2. Thus, though MOVE.1-SEP and MOVE.2-SEP resulted in slightly higher RMSE's than REG-SEP, these increases in RMSE were a very small part of the difference in RMSE between REG-SEP and REG-LITTLE. This results to a certain extent from the generally high interstation correlations. The RMSE's of MOVE.1-SEP and MOVE.2-SEP were indistinguishable at two significant figures for each station.

Summary and Conclusions

An approach for streamflow record extension using simple linear regression has been presented. This approach selects a base station from among several in a region for filling in missing data. It differs from traditional approaches in that different stations can be selected as the base station at different times. The approach also provides a decision rule for using only flow values from the same month or all flow values in developing the extension equation used for estimating a particular missing value. An alternative, and perhaps preferred, application of the methodology is to simply use the mixed-station approach to identify separate periods during which a particular station will be used for data fill in and to determine whether to use cyclic or noncyclic extension equations.

A procedure is presented to adjust the regression equations to better maintain the variance and hydrologic extremes of the short-record station. If MOVE.1 and MOVE.2 are viewed as post-regression adjustments, their range of application is extended. For example, if streamflow records are extended using multiple regression, MOVE.1 and MOVE.2 might be used to adjust the regression results.

Appendix I.—References

1. Aitchison, J., and Brown, J. A. C., *The Lognormal Distribution*, Cambridge University Press, London, England, 1957.
2. Beard, L. R., Fredrich, A. J., and Hawkins, E. F., "Estimating Monthly Streamflows Within a Region," *Tech. Paper 18*, The Hydrologic Engineering Center, U.S. Army Corps of Engineers, 1970, 14 pp.
3. Fiering, M. B., "On the Use of Correlation to Augment Data," *Journal of the American Statistical Association*, Vol. 57, Mar., 1962, pp. 20–32.
4. Hirsch, R. M., "An Evaluation of Some Record Reconstruction Techniques," *Water Resources Research*, Vol. 15, No. 6, 1979, pp. 1781–1790.
5. Hirsch, R. M., "A Comparison of Four Streamflow Record Extension Techniques," *Water Resources Research*, Vol. 18, No. 4, 1982, pp. 1081–1088.
6. Kottegoda, N. T., and Elgy, J., "Infilling Missing Flow Data," *Proceedings*, The Fort Collins Third International Hydrology Symposium on Applied and Theoretical Hydrology, Water Resources Publications, 1979, pp. 60–73.
7. Kritskiy, S. N., and Menkel, J. F., "Some Statistical Methods in the Analysis of Hydrologic Series," *Soviet Hydrology: Selected Papers*, Issue No. 1, 1968, pp. 80–98.
8. Matalas, N. C., and Jacobs, B. A., "A Correlation Procedure for Augmenting Hydrologic Data," *U.S. Geol. Surv. Prof. Paper 434-E*, 1964.

Appendix II.—Notation

The following symbols are used in this chapter:

e = error criterion of current extension equation;
i = index of time;
k = index of dependent station;
l = index of independent station;
m = sample mean;
N = number of variables;
q = untransformed flow value;
r = sample estimate of product-moment correlation coefficient;
SEE = standard error of estimate;
SEP = standard error of prediction;
U = ratio of statistic for the extended record to that for historic record;
v = minimum value of error criterion;
x = logarithm (base e) of flow at independent station;

y = logarithm (base *e*) of flow at dependent station;

$\hat{y}$ = logarithm (base *e*) of estimated flow at dependent station;

$\tilde{y}$ = logarithm (base *e*) of flow that is part of complete extended record;

z = element of matrix of log-transformed flow values;

$\hat{\beta}_1$ = estimate of regression intercept;

$\hat{\beta}_2$ = estimate of regression slope;

ρ = population value of product-moment correlation coefficient;

σ = population value of standard deviation; and

μ = population value of mean.

Subscripts

1 = member of sequence containing concurrent record; and

2 = member of extended portion of record.