A Comparison of Four Streamflow Record Extension Techniques

ROBERT M. HIRSCH

U.S. Geological Survey, Reston, Virginia 22092

One approach to developing time series of streamflow, which may be used for simulation and optimization studies of water resources development activities, is to extend an existing gage record in time by exploiting the interstation correlation between the station of interest and some nearby (long-term) base station. Four methods of extension are described, and their properties are explored. The methods are regression (REG), regression plus noise (RPN), and two new methods, maintenance of variance extension types 1 and 2 (MOVE.1, MOVE.2). MOVE.1 is equivalent to a method which is widely used in psychology, biometrics, and geomorphology and which has been called by various names, e.g., 'line of organic correlation,' 'reduced major axis,' 'unique solution,' and 'equivalence line.' The methods are examined for bias and standard error of estimate of moments and order statistics, and an empirical examination is made of the preservation of historic low-flow characteristics using 50-year-long monthly records from seven streams. The REG and RPN methods are shown to have serious deficiencies as record extension techniques. MOVE.2 is shown to be marginally better than MOVE.1, according to the various comparisons of bias and accuracy.

INTRODUCTION

Current practice in many aspects of water resources planning and management involves the use of hydrologic time series to simulate the outcome of decisions. The decisions may include waste treatment plant designs, establishment of operating policies for water supply systems, hydropower production scheduling, entry into river compacts and intergovernmental agreements, and construction of water storage and withdrawal facilities. The kinds of outcomes that may be of interest include frequency and duration of unacceptable water quality conditions, frequency and duration of supply shortfalls for municipal, industrial or agricultural water users, dependable rate of hydropower production during peak demand periods, frequency and severity of river compact violations, or, more abstractly, the expectation and variance of project benefits or costs. There are several different methods for developing time series for use in simulation. The following is a list of the general categories of methods for developing such time series for streamflows at a single site.

1. Use the historic record of streamflows [*Rippl*, 1883; *Hazen*, 1914].

2. Use the historic record of streamflows and extend it in time by exploiting the correlation between flows at the site and concurrent flows at some nearby long-term gage (a base station) [*Riggs*, 1972; *Matalas and Jacobs*, 1964; *Hirsch*, 1979].

3. Reconstruct historic flow records by transposing records from a base station to the site of interest by using some function in which the coefficients are derived from a regional streamflow basin characteristics regression equation [*Hirsch*, 1979].

4. Generate multiple synthetic streamflow records [*Fiering*, 1967] where the parameters are based on historic flow values at the site or on regional streamflow characteristics [*Benson and Matalas*, 1967].

5. Develop and calibrate a conceptual model of the basin and use it to generate a streamflow record by using historical meteorological records as inputs [*Crawford and Linsley*, 1966].

6. Develop and calibrate a conceptual model of the basin and use it to generate multiple streamflow records by using synthetic meteorological records as inputs [*Leclerc and Schaake*, 1973].

Each of these categories, and the specific methods in each category, have certain advantages and disadvantages for various applications. The selection of an appropriate method would depend on the relevant time step of analysis (hours, days, weeks, seasons, years, or decades) and the benefits of increased accuracy in estimation of outcomes in comparison to the cost of applying a more complex method. The more historically based methods (categories 1, 2, 3, and 5) have their greatest applicability when the time scales are fine (for example, small storages, run of river withdrawals, or water quality analysis). The synthetic methods (categories 4 and 6) have their greatest applicability where the time scales are coarse (for example, large storages for control of multiyear droughts). An additional consideration in the selection of methods is the potential for analysis of errors. In general, this becomes more difficult as the number of parameters and model assumptions grow and this grows rapidly as the time step of analysis becomes finer. Another consideration in many cases is the adaptability of the method to changes in watershed characteristics. In general, only those methods that involve conceptual models have this capability.

This paper will focus on methods within the second category and comparison of these with category 1. It is the intent of this paper to evaluate the characteristics of some methods within category 2. This is not done out of a belief that this category is generally superior to the others, but because it may be superior for some uses and is easy to apply and well accepted by many practicing water resource engineers. The evaluation of these techniques will rely on measures of the accuracy of low-flow duration and severity estimates as indicators of the suitability of the techniques for use in water resource system simulations.

In the next section of the paper, four methods of record extension are defined and some of their properties are discussed. A set of Monte Carlo trials are then carried out to evaluate the bias and error in estimating means, variances,

This paper is not subject to U.S. copyright. Published in 1982 by the American Geophysical Union.

Paper number 2W0479.

and order statistics based on extended records where marginal distributions of flows are normal. Finally, the methods are applied to historical data sets and the results are examined for accuracy in reproducing certain historical order statistics.

Problem Definition

For the base station the flow values are denoted $x(i)$ where i is an index of time. For the short-record station the flow values are denoted $y(i)$. The observed events for the two sequences are represented as

$$x(1), \cdots, x(N_1), x(N_1 + 1), \cdots, x(N_1 + N_2)$$

$$y(1), \cdots, y(N_1)$$

where N_1 is the length of the shorter sequence and $(N_1 + N_2)$ is the length of the longer sequence. N_1 is also the length of concurrent record. It is not necessary for the two sequences to begin or end simultaneously, nor need the observations be consecutive, but there is no loss of generality if the two sequences are represented as above.

The estimates of the missing values are denoted $\hat{y}(i)$, $i = N_1 + 1, \cdots, N_1 + N_2$ and the complete extended record is denoted $\tilde{y}(i)$, $i = 1, \cdots, N_1, N_1 + 1, \cdots, N_1 + N_2$; where

$$\tilde{y}(i) = y(i) \qquad i = 1, \cdots, N_1$$

$$\tilde{y}(i) = \hat{y}(i) \qquad i = N_1 + 1, \cdots, N_1 + N_2$$

Table 1 gives the naming conventions for the various sample statistics referred to in the report. Much of the notation and some of the derivations used in this paper were developed by *Matalas and Jacobs* [1964]. There is, however, a fundamental difference between the intent of that paper and the intent of the present one. Matalas and Jacobs were concerned with the quality (bias and variance) of the estimates $m(\tilde{y})$ and $S^2(\tilde{y})$, and it was not their intent to consider methods of producing an extended record $\tilde{y}(i)$. In this paper the goal is the development of this extended record. The properties of statistics of this record, such as $m(\tilde{y})$ and $S^2(\tilde{y})$ are used as measures (but not the only measures) of the quality of the extended record. In the next section, four methods will be presented, and in each case the expectation of $m(\tilde{y})$ and of $S^2(\tilde{y})$ will be evaluated under the assumption that concurrent observations of x and y are stationary, serially independent, and have a bivariate normal probability distribution with parameters μ_x, μ_y, σ_x^2, σ_y^2, and ρ, where 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, respectively, for y. The parameter ρ is the population product moment correlation coefficient, and $\rho\sigma_y/\sigma_x$ is the population value of the slope of the linear regression of y on x.

TABLE 1. Definitions of Sample Statistics

Statistic	Definition
	Sample Mean of
$m(x_1)$	$x(1), \cdots, x(N_1)$
$m(x_2)$	$x(N_1 + 1), \cdots, x(N_1 + N_2)$
$m(x)$	$x(1), \cdots, x(N_1), x(N_1 + 1), \cdots, x(N_1 + N_2)$
$m(y_1)$	$y(1), \cdots, y(N_1)$
$m(\tilde{y})$	$y(1), \cdots, y(N_1), \hat{y}(N_1 + 1), \cdots, \hat{y}(N_1 + N_2)$
	*Sample Variance of**
$S^2(x_1)$	$x(1), \cdots, x(N_1)$
$S^2(x_2)$	$x(N_1 + 1), \cdots, x(N_1 + N_2)$
$S^2(x)$	$x(1), \cdots, x(N_1), x(N_1 + 1), \cdots, x(N_1 + N_2)$
$S^2(y_1)$	$y(1), \cdots, y(N_1)$
$S^2(\tilde{y})$	$y(1), \cdots, y(N_1), \hat{y}(N_1 + 1), \cdots, \hat{y}(N_1 + N_2)$
	Product Moment Correlation Coefficient of
r	$x(1), \cdots, x(N_1)$ and $y(1), \cdots, y(N_1)$

*Variance computed with sample size minus 1 as the divisor.

The Four Methods of Extension

Regression (REG)

The first method of extension is linear regression. The missing values are filled in by the equation

$$\hat{y}(i) = a + bx(i) \qquad (1)$$

where the parameters a and b are those values which minimize

$$Z = \sum_{i=1}^{N_1} (\hat{y}(i) - y(i))^2$$

The solution for a and b is found by solving the normal equations [*Draper and Smith,* 1966, p. 59]. The optimal solution to (1), rearranged for convenience, is

$$\hat{y}(i) = m(y_1) + r\frac{S(y_1)}{S(x_1)}(x(i) - m(x_1)) \qquad (2)$$

Matalas and Jacobs [1964] show that $m(\tilde{y})$ is an unbiased estimate of μ_y but $S^2(\tilde{y})$ is a biased estimate of σ_y^2 for $\rho^2 < 1.0$. Specifically,

$$E[m(\tilde{y})] = \mu_y$$

$$E[S^2(\tilde{y})] = \sigma_y^2\left\{1 - \frac{(1 - \rho^2)N_2(N_1 - 4)}{(N_1 + N_2 - 1)(N_1 - 3)}\right\}$$

Table 2 gives some example values of $E[S^2(\tilde{y})]/\sigma_y^2$ for various combinations of ρ, N_1, and N_2. Given that a common purpose of record extension is the evaluation of the severity and duration of hydrologic extremes, this consistent underestimation of variance is an alarming feature of REG. It is in fact the intent of each of the following three methods to eliminate (or at least minimize) this bias in the variance.

Regression Plus Independent Noise (RPN)

Matalas and Jacobs [1964] demonstrated that unbiased estimates of mean and variance are achieved if the following equation is used to calculate $\hat{y}(i)$:

$$\hat{y}(i) = m(y_1) + r\frac{S(y_1)}{S(x_1)}(x(i) - m(x_1)) + \alpha(1 - r^2)^{1/2}S(y_1)e(i) \qquad (3)$$

where $e(i)$ is a normal independent random variable with zero mean and unit variance and

$$\alpha^2 = \frac{N_2(N_1 - 4)(N_1 - 1)}{(N_2 - 1)(N_1 - 3)(N_1 - 2)}$$

This procedure of adding independent noise to regression estimates [*Matalas and Jacobs,* 1964, p. E4]

> . . . is not too appealing. Independent studies of the same sequence of x and y by several investigators lead to different

TABLE 2. Values of $E[S_y^2(\hat{y})]/\sigma_y^2$ Using the Regression (REG) Method of Record Extension

		ρ		
N_1	N_2	0.5	0.7	0.9
10	10	0.66	0.77	0.91
20	20	0.64	0.75	0.91
30	30	0.63	0.75	0.91
50	10	0.88	0.92	0.97
10	50	0.46	0.63	0.86

values of $[m(\hat{y}), S^2(\hat{y})$ and $\hat{y}(i)$, $i = N_1 + 1, \cdots, N_1 + N_2]$, because the same sequence of pseudo-random numbers is unlikely to be used by the investigators.

The estimates $\hat{y}(i)$ are hybrids: a weighted sum of the historically observed random variable $x(i)$ and a unrelated, computer generated random variable, $e(i)$. One may feel uncomfortable making a management decision when only one realization of these series of random numbers is used, and yet the interpretation of results from multiple realizations would be, at best, ambiguous because the realizations would not be independent of each other. Nevertheless, the RPN method does have the desired properties of an unbiased mean and variance, and it has found use in practice [*Beard et al.*, 1970].

Maintenance of Variance Extension, Type 1 (MOVE.1)

An alternative to the RPN approach is to specify that the extension equation must be of the form given in (1) but that a and b are to be set not to minimize squared errors, but rather to maintain the sample mean and variance. The idea which led to the development of MOVE.1 was to find some values of a and b in (1) which satisfy the following two equalities:

$$\sum_{i=1}^{N_1} \hat{y}(i) = \sum_{i=1}^{N_1} y(i)$$

$$\sum_{i=1}^{N_1} (\hat{y}(i) - m(y_1))^2 = \sum_{i=1}^{N_1} (y(i) - m(y_1))^2$$

One such solution is

$$\hat{y}(i) = m(y_1) + \frac{S(y_1)}{S(x_1)} (x(i) - m(x_1)) \qquad (4)$$

In spite of the obvious similarity between (2) and (4), it should be recognized that they arise from completely different motivations. For the former it is the minimization of squared errors of $\hat{y}(i)$; for the latter it is the desire for the sample mean and variance of the $\hat{y}(i)$ to equal the sample mean and variance of the $y(i)$ for $i = 1, \cdots, N_1$. When this equation is used for record extension it can be shown that

$$E(m(\hat{y})) = \mu_y \qquad (5a)$$

$$E[S^2(\hat{y})] = \frac{\sigma_y^2}{N_1 + N_2 - 1} \left\{ N_1 - 1 + N_2 \left(\frac{N_1 - 1 - 2\rho^2}{N_1 - 3} \right) \right\} \qquad (5b)$$

Table 3 gives values of the ratio $E\ S^2(y)/\sigma_y^2$ for various combinations of ρ, N_1, and N_2. It is clear from Tables 2 and 3 that the magnitude of the bias is substantially lower for MOVE.1 than for REG, and that REG underestimates variance, while MOVE.1 overestimates it. Note that $S^2(y)$ is an asymptotically unbiased estimator of σ_y^2 as $N_1 \rightarrow \infty$.

It should be noted that the estimates $y(i)$ in (4) lie between the estimates of $y(i)$ from a regression of y on x and the estimate of $y(i)$ from the inverse of a regression of x on y. This property was noted and discussed in a hydrologic context by *Kritskiy and Menkel* [1968], who called (4) 'the unique solution.' *Till* [1973] refers to this line as the reduced major axis and gives the standard error of the slope and intercept estimates; he also gives some applications in geomorphology. Additional information on the mathematics is found in the work by *Kruskal* [1953]. The first known reference to this line is by *Pearson* [1901] and discussions of its applications can be found in works by *Imbrie* [1956] (biometrics) and *Greenall* [1949] (psychology).

Kirby [1974] has proposed least normal squares, which also results in an equation for a line falling between the regression of x on y and y on x. His equation coincides with (4) only where $r = 1.0$ and/or $S(x_1) = S(y_1)$. Neither the line represented by (4) nor Kirby's line coincide with the line which bisects the angle between these two regression lines except where $r = 1.0$ and/or $S(x_1) = S(y_1)$.

Maintenance of Variance Extension, Type 2 (MOVE.2)

In MOVE.1 the only four parameters are the means and variances of x and y estimated from the first N_1 observations. In MOVE.2 these same parameters are used, but their estimation is based on more information. The extension equation is

$$\hat{y}(i) = \hat{m}(y) + \frac{\hat{S}(y)}{S(x)} [x(i) - m(x)] \qquad (6)$$

The mean and variance estimates for x are based on all $N_1 + N_2$ observations, and the mean and variance estimates for y are based on the historical values of y and on information transfer from the x sequence. The parameters $\hat{m}(y)$ and $\hat{S}^2(y)$ were developed by *Matalas and Jacobs* [1964] and are themselves unbiased estimates of μ_y and σ_y^2.

$$\hat{m}(y) = m(y_1) + \frac{N_2}{(N_1 + N_2)} r \frac{S(y_1)}{S(x_2)} (m(x_2) - m(x_1)) \qquad (7)$$

$$\hat{S}^2(y) = \frac{1}{N_1 + N_2 - 1} \left\{ (N_1 - 1)S^2(y_1) \right.$$
$$+ (N_2 - 1)\, r^2 \frac{S^2(y_1)}{S^2(x_1)} S^2(x_2) + (N_2 - 1)\, \alpha^2(1 - r^2)S^2(y_1)$$
$$\left. + \frac{N_1 N_2}{(N_1 + N_2)} r^2 \frac{S^2(y_1)}{S^2(x_1)} (m(x_2) - m(x_1))^2 \right\} \qquad (8)$$

The complexity of these expressions prevented the discovery of an analytical solution for the bias of the mean and variance. In Monte Carlo experiments of 2000 trials with each of 15 different combinations of ρ, N_1, and N_2, the hypothesis that $E[m(\hat{y})] = \mu_y$ could not be rejected at the

TABLE 3. Values of $E[S_y^2(\hat{y})]/\sigma_y^2$ Using the Maintenance of Variance Type 1 (MOVE.1) Method of Record Extension

N_1	N_2	ρ		
		0.5	0.7	0.9
10	10	1.19	1.08	1.03
20	20	1.05	1.03	1.03
30	30	1.03	1.02	1.01
50	10	1.01	1.00	1.00
10	50	1.18	1.12	1.05

0.05 significance level. The results for $S^2(\hat{y})$ are shown in Table 4. These results suggest that for all practical purposes, MOVE.2 satisfies the need for an extension method which produces unbiased values of $m(\hat{y})$ and $S^2(\hat{y})$.

MONTE CARLO EVALUATION OF ERRORS

In estimating frequency distributions of streamflows, the hydrologist will not only consider the statistical moments of the sample but also some of the extreme order statistics. If some method of record extension introduces a bias into the value of the more extreme order statistics, this will lead to bias in the estimates of the probability of exceedance of selected extreme values or, conversely, bias in the estimation of distribution quantiles.

Let w represent a statistic of a (simulated) historic time series (such as a moment or order statistic) and $\tilde{w}$ represent the same statistic of a (simulated) extended record. The error in any realization of w is $(w - E[w])$, and the error in any realization of $\tilde{w}$ is $(\tilde{w} - E[w])$, where $E[w]$ is the expectation of the statistic for the sample size and population being considered. The estimated bias of $\tilde{w}$, $B(\tilde{w})$, is the average error $(\tilde{w} - E[w])$ over a large number of Monte Carlo trials (2000 trials in this case). The root mean square error for statistic w is denoted $R(w)$, and for the statistic $\tilde{w}$ the root mean square error is $R(\tilde{w})$.

The statistics w considered here are $m(y)$, $S^2(y)$, y_1, y_2, and y_5 where y_k is the kth order statistic of the series $y(i)$ $i = 1, \cdots, N_1 + N_2$. The statistics $\tilde{w}$ are $m(\hat{y})$, $S^2(\hat{y})$, $\hat{y}_1$, $\hat{y}_2$, and $\hat{y}_5$ where $\hat{y}_k$ is the kth order statistic of the series $\hat{y}(i)$, $i = 1, \cdots, N_1 + N_2$. The Monte Carlo trials were carried out for (N_1, N_2) values of (10, 50) and (20, 40) and for ρ values of 0.5, 0.7, and 0.9. In all cases the flows are assumed to be bivariate normal $\mu_x = \mu_y = 0$, $\sigma_x^2 = \sigma_y^2 = 1$ with no autocorrelation.

The results of these trials are given in Table 5. For each statistic w, the null hypothesis that $B(w)$ or $B(\tilde{w})$ equals zero (unbiased) was tested and those cases found significant at the 5% level are marked by the asterisk. The $B(w)$ are unbiased by construction.

For REG all of the statistics except $m(\hat{y})$ appear to be biased for all of the cases considered. The absolute value of the bias decreases with an increase in ρ or an increase in N_1. The biases for REG all show the regression towards the mean, the estimated variances are too low, and the low-order statistics are too high. For example, with $N_1 = 10$, $N_2 = 50$, and $\rho = 0.5$, the second order statistic is, on the average, -1.261 rather than -1.935, which is its expectation; thus $B(\hat{y}_2) = (-1.261) - (-1.935) = 0.674$.

For the other three methods the absolute value of the biases for all w except $m(\hat{y})$ are in all cases substantially less than the bias with REG. For RPN there were no cases of significant bias for $S^2(\hat{y})$, and for the order statistics the bias was either significant and positive or not significant. For MOVE.1 the bias for $S^2(\hat{y})$ was in all cases positive and significant, and the bias for the order statistics was either negative and significant or not significant. In no case did the results show the bias to be significantly (at the 5% level) different from the value determined by (5*b*). For MOVE.2, in only one case was there significant bias in $S^2(\hat{y})$ (negative bias for $N_1 = 20$, $N_2 = 40$, $\rho = 0.5$) and it was not significant at the 2.5% level. The bias in the order statistics were either positive and significant or not significant.

All biases observed in $m(\hat{y})$ were very small. Two of them were significant but even if all methods were in fact unbiased in $m(\hat{y})$, it would not be unreasonable to find two significantly biased values (at the 5% level) out of 24 cases. Neither of these two cases is significant at the 2.5% level.

Thus, in summary, one can conclude that all four methods produce extended records that are unbiased in the mean, that REG substantially reduces variability, MOVE.1 slightly increases variability, and RPN and MOVE.2 preserve the variance but slightly decreases variability in the more extreme events. The method with the lowest value of the root mean square error (RMSE) for all statistics except $m(\hat{y})$ is MOVE.2. For $m(\hat{y})$ there is little difference between the RMSE of REG and of MOVE.2. For all methods and all statistics the RMSE decreases with an increase in N_1 or an increase in ρ.

EMPIRICAL CHECK OF THE METHODS

In the previous sections of this paper, the random variable being considered had very simple and hydrologically unrealistic properties (normal, independent, and without a cyclic component). In this section some real data are used to explore the techniques under conditions of nonnormal distribution, serial dependence, and seasonal cycles. The data used are monthly volume data. Some decisions were necessary on the particulars of how to apply the four techniques to such data.

The first decision was whether or not to use only one extension equation for all months, 12 different ones for the 12 months, or to make some compromise such as two or four seasonal equations. The choice involves a trade-off between greater sample size for estimating the parameters versus the ability to preserve real month to month differences that may exist in the base station to short-record station relationship. The choice made here, based on some experiments with the data, was to use a single extension equation for all of the months in each of the four techniques. This problem has

TABLE 4. Values of Sample Mean of $S^2(\hat{y})/\sigma_y^2$ Using the Maintenance of Variance, Type 2 (MOVE.2) Method of Record Extension

N_1	N_2	ρ		
		0.5	0.7	0.9
10	10	0.99	0.99	1.02*
20	20	0.99	1.00	1.00
30	30	0.99	0.99	1.00
50	10	1.00	0.99	1.01
10	50	0.99	1.01	1.00

Based on 2000 Monte Carlo trials.

*The hypothesis H_0: $E[S^2(\hat{y})]/\sigma_y^2 = 1.00$ is rejected at the 5% level.

TABLE 5. Biases $B(\ )$ and Root Mean Square Error $R(\ )$ From 2000 Monte Carlo Trials

				$B(\omega)$ or $B(\hat{\omega})$					$R(\omega)$ or $R(\hat{\omega})$				
n1	n2	rho	method	$m(y)$	$S^2(y)$	y_5	y_2	y_1	$m(y)$	$S^2(y)$	y_5	y_2	y_1
10	50	0.5	HIST	0.001	0.004	−0.002	−0.002	−0.006	0.131	0.183	0.239	0.332	0.459
10	50	0.5	REG	−0.003	−0.557*	0.603*	0.674*	0.656*	0.294	0.658	0.784	0.871	0.910
10	50	0.5	RPN	−0.005	−0.001	0.030*	0.067*	0.097*	0.306	0.517	0.487	0.616	0.734
10	50	0.5	MOVE.1	−0.009	0.164*	−0.059*	−0.069*	−0.090*	0.334	0.783	0.593	0.758	0.930
10	50	0.5	MOVE.2	−0.004	−0.017	0.044*	0.083*	0.102*	0.292	0.487	0.467	0.590	0.726
10	50	0.7	HIST	−0.001	0.003	−0.000	−0.005	−0.010	0.128	0.184	0.233	0.323	0.442
10	50	0.7	REG	0.004	−0.362*	0.358*	0.447*	0.476*	0.251	0.533	0.578	0.707	0.804
10	50	0.7	RPN	0.003	0.007	0.021*	0.041*	0.057*	0.262	0.458	0.439	0.563	0.705
10	50	0.7	MOVE.1	0.002	0.137*	−0.045*	−0.069*	−0.090*	0.271	0.670	0.501	0.641	0.803
10	50	0.7	MOVE.2	0.004	0.004	0.031*	0.046*	0.058*	0.250	0.445	0.427	0.533	0.670
10	50	0.9	HIST	0.004	−0.002	0.001	0.010	0.013	0.129	0.186	0.235	0.333	0.448
10	50	0.9	REG	0.003	−0.148*	0.129*	0.176*	0.217*	0.186	0.352	0.367	0.491	0.610
10	50	0.9	RPN	0.004	−0.013	0.026*	0.037*	0.050*	0.191	0.346	0.342	0.472	0.589
10	50	0.9	MOVE.1	0.002	0.040*	−0.012	−0.018*	−0.017	0.188	0.382	0.348	0.473	0.603
10	50	0.9	MOVE.2	0.003	−0.011	0.022*	0.033*	0.047*	0.186	0.333	0.340	0.456	0.577
20	40	0.5	HIST	0.002	−0.000	−0.001	0.004	0.001	0.130	0.187	0.245	0.334	0.452
20	40	0.5	REG	0.006	−0.484*	0.476*	0.470*	0.446*	0.208	0.533	0.582	0.624	0.687
20	40	0.5	RPN	0.006	−0.006	0.023*	0.044*	0.067*	0.228	0.347	0.375	0.472	0.586
20	40	0.5	MOVE.1	0.005	0.054*	−0.009	−0.016	−0.014	0.227	0.418	0.386	0.510	0.641
20	40	0.5	MOVE.2	0.005	−0.013*	0.027*	0.049*	0.077*	0.208	0.325	0.341	0.443	0.556
20	40	0.7	HIST	−0.004	0.003	−0.008	−0.003	−0.011	0.128	0.184	0.237	0.334	0.458
20	40	0.7	REG	−0.007*	−0.314*	0.267*	0.327*	0.344*	0.185	0.412	0.423	0.531	0.624
20	40	0.7	RPN	−0.007	0.010	−0.005	0.010	0.026*	0.199	0.330	0.343	0.447	0.560
20	40	0.7	MOVE.1	−0.005	0.052*	−0.027*	−0.022*	−0.031*	0.194	0.375	0.355	0.470	0.604
20	40	0.7	MOVE.2	−0.006	0.003	−0.002	0.022*	0.037*	0.186	0.313	0.328	0.427	0.544
20	40	0.9	HIST	0.002	−0.003	0.006	0.015*	0.005	0.129	0.180	0.240	0.331	0.459
20	40	0.9	REG	0.004	−0.124*	0.099*	0.138*	0.156*	0.151	0.268	0.296	0.395	0.521
20	40	0.9	RPN	0.006*	−0.004	0.010	0.019*	0.019	0.158	0.259	0.285	0.389	0.535
20	40	0.9	MOVE.1	0.004	0.014*	−0.007	−0.001	−0.012	0.152	0.257	0.283	0.383	0.524
20	40	0.9	MOVE.2	0.004	−0.005	0.005	0.019*	0.017	0.151	0.245	0.280	0.375	0.510

The rows marked HIST represent calculations based on $(N_1 + N_2)$ observations and involving no record extension. $E(\omega) = 0.000, 1.000, -1.430, -1.935, -2.319$, for the five statistics, respectively.

*Bias significant at the 5% level.

been explored by *Alley and Burns* [1981], who provide a means for selecting the best method for any given case.

The second decision was whether or not to do the extension with the real data or with their logarithms (that is, take logarithms of all of the data, do the extension, and then take antilogs of these extended records). This question has been dealt with by *Hirsch* [1979] and, in a somewhat different context, by *Stedinger* [1980]. In both cases the conclusion was to work with the logarithms, and this was the approach taken here. The consequences of this choice are that for any of the techniques, the sample mean of the extended record of the logarithms is an unbiased estimate of the mean of the logarithms, but the sample mean of the extended record of flows is not an unbiased estimate of the mean of the flows. However, the objective of the technique is to produce records with sample cumulative distribution functions (CDF's) which are close approximations of the CDF's actual records (particularly in the tails). Where the station to station relationship more nearly approximates a bivariate lognormal distribution rather than a bivariate normal, transforming to logarithms will better achieve the desired result.

The evaluation described here will forcus on the ability of the techniques to produce low-flow duration intensity characteristics like those of the actual records they are intended to represent. The example used concurrent, 50-year-long (commencing in 1928) U.S. Geological Survey monthly volume records from seven stream gaging stations in west central Virginia. Table 6 provides some information on the stations, their drainage basins and the observed flow characteristics. The experimental design is the following.

1. Each station is, in turn, considered to be the short-record station with only 10 years of data available. The historic values of the first, second, and fifth-order statistics in the annual series of minimum 1-month, 3-month, and 6-month volumes from the entire 50-year record are recorded. They are denoted $Q(r, d, is)$ where $r = 1, 2, 5$ (index of order or rank), $d = 1, 3, 6$ (duration in months), and $is = 1, 2, \cdots, 7$ (index of short-record station). The year is considered to be the climatic year commencing in April.

2. For each given short-record station, each of the other six stations is considered, in turn, to be the base station with the full 50-year record available.

3. For each short-record station–base station pair, the period of record overlap is considered, in turn, to be years 1–10, 11–20, 21–30, 31–40, and 41–50, respectively.

4. For each short-record station–base station–overlap period, (there are 7 × 6 × 5 = 210 such 3-tuples) each of the four extension methods is applied to the logarithms of the streamflow volume data. From each of these extended flow records the same flow statistics are computed. They are

TABLE 6. Information on the Seven Gages

Station Number	Location	Drainage area mi^2	Mean basin elevation in feet	Forest cover in percent	Monthly Volumes: Mean 10^6 m^3	Monthly Volumes: Coefficient of Variation	Monthly Volumes: Coefficient of Skewness	Logarithms of Monthly Volumes: Standard deviation	Logarithms of Monthly Volumes: Coefficient of skewness	Logarithms of Monthly Volumes: Coefficient of kurtosis
02018000	Craig Creek at Parr, Va.	329	2150	89	28.3	0.92	1.26	1.04	−0.12	1.81
02017500	Johns Creek at New Castle, Va.	104	2210	91	9.3	0.97	1.33	1.11	−0.12	1.80
02055000	Roanoke River at Roanoke, Va.	395	1680	75	26.3	0.83	1.65	0.83	−0.02	2.12
02013000	Dunlap Creek near Covington, Va.	164	2230	88	12.3	1.02	1.48	1.12	−0.00	1.76
02016000	Cowpasture River near Clifton Forge, Va.	461	2030	82	38.6	0.88	1.33	0.93	−0.03	1.87
03167000	Reed Creek at Grahams Forge, Va.	247	2500	44	19.7	0.78	1.59	0.74	0.21	2.01
03170000	Little River at Graysonton, Va.	300	2470	47	27.0	0.57	1.54	0.54	−0.03	2.59

denoted $\hat{Q}(r, d, is, ib, sq, m)$, $ib = 1, 2, \cdots, 7$ (index of base station), $ib \neq is$, $sq = 1, 2, \cdots, 5$ (index of sequence), and m = {REG, RPN, MOVE.1, MOVE.2}.

5. The measure of accuracy of each estimate is $U(r, d, is, ib, sq, m)$

$$U(r, d, is, ib, sq, m) = \frac{\hat{Q}(r, d, is, ib, sq, m)}{Q(r, d, is)}$$

The correlation coefficient (of the log values) was also recorded for each of the 210 3-tuples. The median correlation coefficient is 0.90, the lower and upper quartiles are 0.85 and 0.93, and the minimum and maximum are 0.71 and 0.99.

The accuracy of the methods is summarized in Figures 1, 2, and 3. In these figures the box plots represent the distribution of all 210 U values for a given rank, duration, and method. The accuracy of the method may be judged by the degree of dispersion in the box plot for that method, by the degree that the median approaches a value of 1.0, and by the symmetry of the box about a value of 1.0. The following are some observations about the results.

1. Both MOVE.1 and MOVE.2 have median values for U very close to 1.0 (in fact, between 0.980 and 1.007) for all flow statistics.

2. At all durations REG had median U values > 1.00 (the most extreme being 1.11). For durations of 1 and 3 months and for all three order statistics, approximately 75% of the computed values were higher than the historical values. This result was expected, of course, given the tendency for regression to reduce variance resulting in low extremes which are 'too high.'

3. For RPN at a 1-month duration, the median U values as well as the upper quartile U values are <1.00. This arises because of the kurtosis of the log volume values. For all seven streams, the kurtosis is significantly (α = 0.01) less than 3.0. For the three other methods, REG, MOVE.1, and MOVE.2, the kurtosis of the extended record will closely approximate the kurtosis from the base station. But in RPN the kurtosis of the extended record will be some weighted average (depending on r) of the base station kurtosis and the kurtosis of the noise term (a value of 3). Thus the kurtosis of

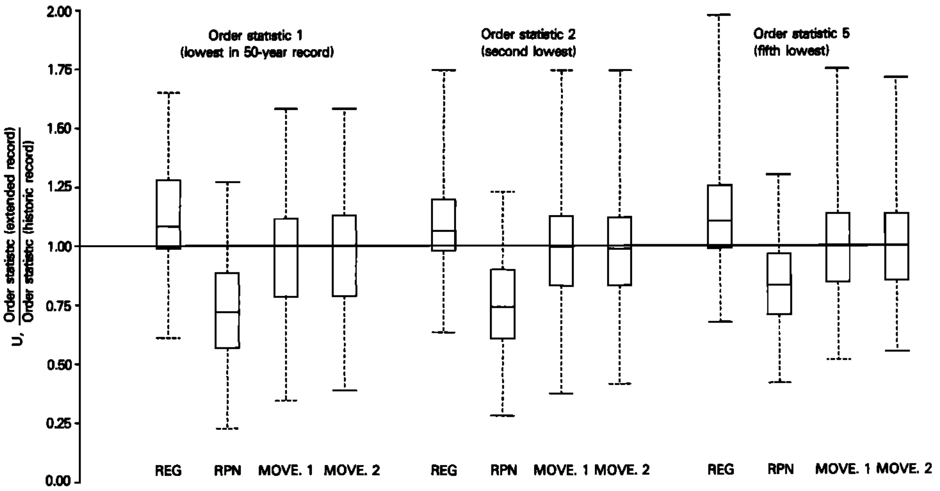

Fig. 1. Box plots of U values for 1-month-duration low flows. Box plots show median, upper and lower quartiles, and maximum and minimum values. The sample size is 210.

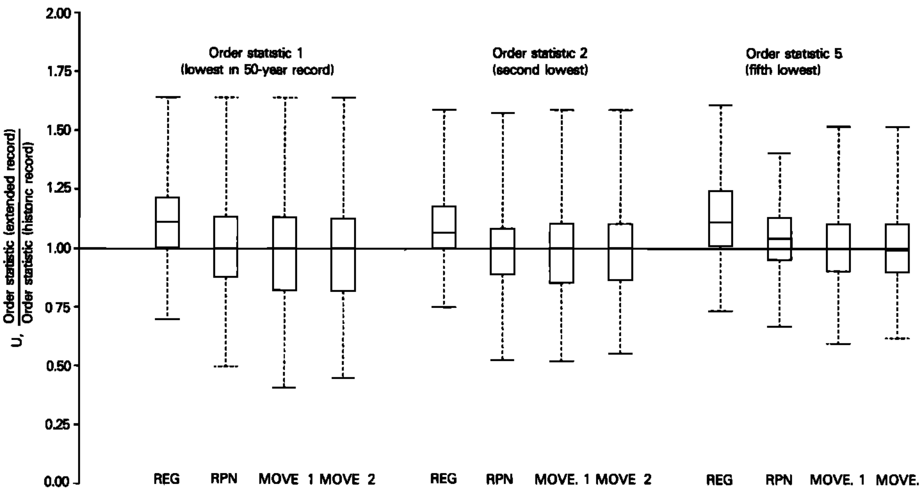

Fig. 2. Box plots of U values for 3-month-duration low flows. Box plots show median, upper and lower quartiles, and maximum and minimum values. The sample size is 210.

the extended records using RPN will be higher than those of the historic records. Given the approximate preservation of variance and inflation of kurtosis, the extremes of the extended records will be more extreme than the historic. It is not known how common it is for kurtosis values for such time series to fall in the range 1.5–2.5 (as these seven records do). It is not, however, unexpected. If one postulates that log monthly volumes are the sum of a normal random component and a sine curve of a period of 1 year (with the random component standard deviation equal to about one third of the semiamplitude of the sine curve), the kurtosis of the sum is approximately 2.

4. For RPN at a duration of 6 months the median U values for RPN are >1.00, and in fact, nearly 75% of all U values are >1.00. Thus RPN extended records have too low low-flow values at a 1-month duration and too high low-flow values at a 6-month duration. The reason for these too high values at 6-month duration is that the noise in (3) is independent. Thus the estimated portion ($i = N_1 + 1, \cdots, N_1 + N_2$) of the extended record will have a serial correlation coefficient lower (in absolute value) than the base station, but for the other three methods it will match the base station exactly.

5. At a 3-month duration the U values for RPN are reasonably symmetrical about a value of 1.0. Presumably, the two contrary effects described above (3 and 4) are approximately in balance at this duration.

6. The box plots of U values for MOVE.1 and MOVE.2 are very similar in all cases. In general, the interquartile range and overall range for MOVE.2 is slightly smaller than for MOVE.1.

An additional question considered here was the matter of selecting a base station from among a variety of possible long-record stations located nearby. The hypothesis considered was that using the base station with the highest correlation coefficient (of the log values) for the 10-year overlap period would result in a smaller dispersion of U values than was found using all possible base stations. This hypothesis was born out by the results: The interquartile ranges of U values using only the best correlated base station was

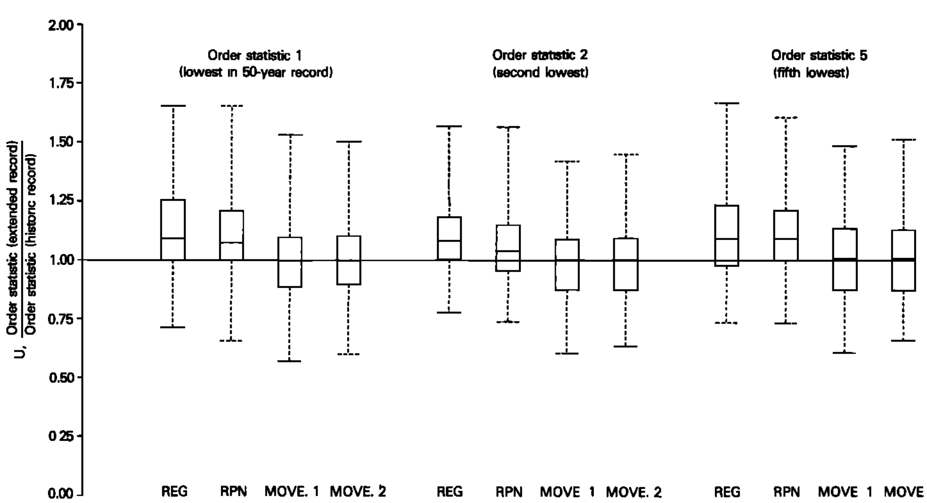

Fig. 3. Box plots of U values for 6-month-duration low flows. Box plots show median, upper and lower quartiles, and maximum and minimum values. The sample size is 210.

smaller for all nine order statistics considered. The average ratio of the interquartile range using these selected base stations to the interquartile range for all possible base stations was 0.63. Thus maximum sample correlation appears to be a reasonable selection criterion; obviously, others might exist which are better.

SUMMARY AND CONCLUSIONS

In using record extension, as applied in the preceding section, one is transferring the characteristics of distribution shape, serial correlation, and seasonality from the base station to the short-record station with adjustments of location and scale appropriate to the short-record station. The analytical derivations, the Monte Carlo study of samples from bivariate normal populations, and the empirical analysis all show that REG and RPN fall substantially short of achieving the desired result of creating a realistic extended record. Specifically, REG cannot be expected to provide records with the appropriate variability and RPN cannot be expected to provide records with the appropriate distribution shape or serial correlaton.

MOVE.1 and MOVE.2 differ in the ways that they utilize the available data for parameter estimation (the latter uses more of the base station data than does the former). The differences in their performance, as seen in the evaluation of biases of moments and order statistics and in the empirical study, all show MOVE.2 to have slightly more desirable properties than MOVE.1 and rather substantially better properties than REG or RPN. Of course, if the base station and short record station have substantial differences in terms of distribution shapes, serial correlation, or seasonality, then these methods cannot be expected to perform very well, because they will transfer these characteristics from the base station to the short record station.

The intent of record extension is to produce a time series which is relatively long and is also plausible; that is, one that possesses statistical characteristics believed to be like those of an actual record for the station. The reason for producing such a record is for use in simulation and optimizations related to potential water management decisions. The use of regression for these purposes is inappropriate. The aim of regression is to provide the 'best estimate' (minimum squared error) for each individual streamflow value and not to preserve any particular statistical characteristics of the record. The three other methods are attempts to produce series which do preserve certain desired characteristics. The results of this study suggest that MOVE.2 is the most effective of all four of the methods in terms of producing time series with properties (such as variance and extreme order statistics) most like the properties of the records they are intended to represent.

Acknowledgments. Ed Gilroy of the U.S. Geological Survey helped with the derivation of the bias of the variance for MOVE.1. Bill Kirby of the U.S. Geological Survey suggested the idea of MOVE.2 as a variant of MOVE.1.

REFERENCES

Aitchison, J., and J. A. C. Brown, *The Log-normal Distribution,* Cambridge University Press, London, 1957.

Alley, W. A., and A. W. Burns, Mixed-station extension of monthly streamflow records, report, U.S. Geol. Surv., Reston, Virginia, 1981.

Beard, L. R., A. J. Fredrich, and E. F. Hawkins, Estimating monthly streamflows within a region, *Tech. Pap. 18,* 14 pp., The Hydrol. Eng. Center, U.S. Army Corps of Eng., Davis, Calif., 1970.

Benson, M. A., and N. C. Matalas, Synthetic hydrology based on regional statistical parameters, *Water Resour. Res., 3,* 931–936, 1967.

Crawford, N. H., and R. K. Linsley, Jr., Digital simulation in hydrology: Stanford watershed model iv, *Tech. Rep. 39,* Dep. of Civ. Eng., Stanford Univ., Stanford, Calif., July 1966.

Draper, N. R., and H. Smith, *Applied Regression Analysis,* John Wiley, New York, 1966.

Fiering, M. B, *Streamflow Synthesis,* Harvard University Press, Cambridge, Massachusetts, 1967.

Greenall, P. D., The concept of equivalent scores in similar tests, *Br. J. Stat. Psychol., 2,* 30–40, 1949.

Hazen, A., Storage required for the regulation of streamflow, *Trans. Am. Soc. Civ. Eng., 77,* 1914.

Hirsch, R. M., An evaluation of some record reconstruction techniques, *Water Resour. Res., 15,* 1781-1790, 1979.

Imbrie, J., Biometrical methods in the study of invertebrate fossils, *Bull. Am. Mus. Nat. Hist., 108,* 215–52, 1956.

Kirby, W., Straight line fitting of an observation path by least normal squares, *U.S. Geol. Surv. Open-File Rep., 74-187,* 1974.

Kritskiy, S. N., and M. F. Menkel, Some statistical methods in the analysis of hydrologic series, *Sov. Hydrol. Select. Pap., 7,* 80–98, 1968.

Kruskal, W. H., On the uniqueness of the line of organic correlation, *Biometrics, 9,* 47–58, 1953.

Leclerc, G., and J. C. Schaake, Jr., Methodology for assessing the potential impact of urban development on urban runoff and the relative efficiency of runoff control alternatives, *Rep. 167,* Dep. of Civ. Eng., Mass. Inst. of Technol., Cambridge, Mass., 1973.

Matalas, N. C., Mathematical assessment of synthetic hydrology, *Water Resour. Res., 3,* 937-945, 1967.

Matalas, N. C., and B. A. Jacobs, A correlation procedure for augmenting hydrologic data, *U.S. Geol. Surv. Prof. Pap., 434-E,* 1964.

Pearson, K., On lines and planes of closest fit to systems of points in space, *Philos. Mag., 2,* pp. 559-572, 1901.

Riggs, H. C., Low-flow investigations, *U.S. Geol. Surv. Tech. Water Resour. Invest., 4,* 1972.

Rippl, W., The capacity of storage reservoirs for water supply, *Proc. Inst. Civ. Eng., 71,* 1883.

Stedinger, J. R., Fitting log normal distributions to hydrologic data, *Water Resour. Res., 16,* 481–490, 1980.

Till, R., The use of linear regression in geomorphology, *Area, 5,* 303-308, 1973.

(Received October 29, 1981;
revised March 9, 1982;
accepted March 22, 1982.)