Journal of Hydrology, 125 (1991) 259–276
Elsevier Science Publishers B.V., Amsterdam

[1]

THE VALUE OF STREAMFLOW RECORD AUGMENTATION PROCEDURES IN LOW-FLOW AND FLOOD-FLOW FREQUENCY ANALYSIS

RICHARD M. VOGEL[a] and CHARLES N. KROLL[b]

[a] *Department of Civil Engineering, Tufts University, Medford, MA 02155 (USA)*
[b] *Goldberg Zoino & Associates, Inc., 320 Needham Street, Newton Upper Falls, MA 02164 (USA)*

(Received 19 July 1989; accepted after revision 9 October 1990)

ABSTRACT

Vogel, R.M. and Kroll, C.N., 1991. The value of streamflow record augmentation procedures in low-flow and flood-flow frequency analysis. J. Hydrol., 125: 259–276.

Streamflow record augmentation procedures exploit the cross-correlation among streamflows at two or more streamgages to obtain improved estimates of the mean and variance of the flows at a short-record gage. Recent improvements in these procedures provide unbiased estimates of the mean and variance of the flows at the short-record site which always have equal or lower variance than simple at-site sample estimates. Essentially, record augmentation procedures increase the effective record length at a short-record site in proportion to the additional length of the nearby longer record and the cross-correlation of the concurrent streamflows at the two sites. An experiment documents the effective increase in record lengths on a site-by-site basis and on a regional basis using a network of 23 streamgages in or near Massachusetts. The increases in effective record lengths owing to the use of streamflow record augmentation procedures are substantial for the very-short-record sites for both flood-flow and low-flow statistics. However, the serial correlation associated with both flood-flow and low-flow sequences reduces those gains considerably.

INTRODUCTION

A common problem faced by hydrologists is the estimation of low-flow and flood-flow quantiles from short streamflow records. Streamflow record augmentation procedures introduced by Fiering (1963), Matalas and Jacobs (1964) and Vogel and Stedinger (1985) can effectively increase the length of short hydrologic records by exploiting the cross-correlation among nearby longer records. These procedures are documented in the well-known manual of practice referred to as Bulletin 17B (Interagency Advisory Committee on Water Data, 1982, appendix 7). Essentially, the cross-correlation between a long x record and a short y record is used to obtain maximum likelihood estimates of the mean and variance of the flows at the short-record site when the observations are independent across time and arise from a bivariate normal distribution. Vogel and Stedinger (1985) provided a review of the literature pertaining to streamflow record augmentation procedures.

For the short-record site, the streamflow record augmentation procedures introduced by Fiering (1963) and Matalas and Jacobs (1964) do not always generate estimates of the mean and variance of the flows with lower variance than simple at-site sample estimates. More recently, Vogel and Stedinger (1985) introduced improved record augmentation procedures which result in unbiased estimates of the mean and variance of the streamflows at the short-record site: these have equal or lower variance than the simple at-site sample estimates.

The studies by Fiering (1963), Matalas and Jacobs (1964), and Vogel and Stedinger (1985), derived streamflow record augmentation estimators and their corresponding sampling properties. To our knowledge, no studies document the merit of employing such procedures to obtain estimates of design quantiles using actual streamflow records. Here we derive first-order approximations to the standard error of quantile estimators which use the improved streamflow record augmentation procedures introduced by Vogel and Stedinger (1985). To evaluate the merit of these procedures in practice, an experiment is performed that documents the effective increase in the length of the short record which may be obtained by using record augmentation to estimate quantiles at 23 gaged sites in or near Massachusetts.

QUANTILE ESTIMATION USING STREAMFLOW RECORD AUGMENTATION PROCEDURES

This section reviews the procedures for estimating quantiles using both at-site sample estimates and record augmentation estimates of the moments of streamflow sequences. As a two-parameter log-normal model (LN2) often provides an adequate description of the distribution of both annual maximum and annual minimum d day streamflows (see Vogel and Kroll, 1989, and later sections of this study), we employ the LN2 distribution for all streamflow sequences in this study. The events are denoted by

$$x_1, \ldots, x_{n_1}, x_{n_1+1}, \ldots, x_{n_1+n_2-1}, x_{n_1+n_2}$$

$$y_1, \ldots, y_{n_1}$$

where n_1 is the length of the short record, and $n_1 + n_2$ is the length of the long record. The n_1 concurrent observations need not correspond to the first n_1 observations, nor do they need to be consecutive if the observations are serially independent. As the streamflows are hypothesized to arise from an LN2 distribution, x and y denote the natural logarithm of the streamflows. In practice, sequences of annual maximum flood-flows and sequences of annual minimum d day streamflows may exhibit serial dependence. The impact of serial dependence of flood-flow and low-flow series on the effective record length associated with a design quantile is examined below and is discussed in more detail by Tasker (1983a).

If the x and y series arise from a bivariate normal population, then Vogel and Stedinger (1985) advocated the use of the following minimum variance unbiased estimators of the mean and variance of the complete extended y record

$$\hat{\mu}_y^+ = (1 - \theta_1)\,\bar{y}_1 + \theta_1\,\hat{\mu}_y \tag{1}$$

$$\hat{\sigma}_y^{+2} = (1 - \theta_2)\,s_{y_1}^2 + \theta_2\,\hat{\sigma}_y^2 \tag{2}$$

where $\hat{\mu}_y$ and $\hat{\sigma}_y^2$ are the unbiased estimators of the mean and variance of the complete extended y record derived by Matalas and Jacobs (1964):

$$\hat{\mu}_y = \bar{y}_1 + \delta\,\beta\,(\bar{x}_2 - \bar{x}_1)/n_1 \tag{3}$$

$$\hat{\sigma}_y^2 = \lambda\,[(n_1 - 1)\,s_{y_1}^2 + (n_2 - 1)\,\hat{\beta}^2\,s_{x_2}^2 + (n_2 - 1)\,\alpha^2\,(1 - \hat{\rho}^2)\,s_{y_1}^2 + \delta\,\hat{\beta}^2\,(\bar{x}_2 - \bar{x}_1)^2] \tag{4}$$

where

$$\lambda = 1/(n_1 + n_2 - 1)$$

$$\delta = (n_1 n_2)/(n_1 + n_2)$$

$$\alpha^2 = \{n_2(n_1 - 4)\,(n_1 - 1)\}/[(n_2 - 1)\,(n_1 - 3)\,(n_1 - 2)]$$

$$\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$$

$$\bar{x}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} x_i$$

$$\bar{x}_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n_1+n_2} x_i$$

$$s_{y_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} (y_i - \bar{y}_1)^2$$

$$s_{x_1}^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2$$

$$s_{x_2}^2 = \frac{1}{n_2 - 1}\sum_{i=n+1}^{n_1+n_2} (x_i - \bar{x}_1)^2$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)\,(y_i - \bar{y}_1)}{\sum_{i=1}^{n_1} (x_i - \bar{x}_1)^2}$$

$$\hat{\rho} = \hat{\beta}\,\frac{s_{x_1}}{s_{y_1}}$$

$$\theta_1 = \frac{(n_1 - 3)\,\hat{\rho}^2}{(n_1 - 4)\,\hat{\rho}^2 + 1}$$

$$\theta_2 = \frac{(n_1 - 4)\,\hat{\rho}^2}{(n_1 - 8.5)\,\hat{\rho}^2 + 4.5}$$

The estimators $\hat{\mu}_y$ and $\hat{\sigma}_y^2$ derived by Matalas and Jacobs (1964) do not always have lower variance than the alternative at-site sample estimates $\bar{y}_1$ and $s_{y_1}^2$. For example, $\hat{\mu}_y$ only has lower variance than $\bar{y}_1$ when $\hat{\rho}^2 > (n_1 - 2)^{-1}$. The improved estimators $\hat{u}_y^+$ and $\hat{\sigma}_y^{+2}$ in eqns. (1) and (2) contain the weights θ_1 and θ_2 which were derived to assure that $\mathrm{Var}(\hat{\mu}_y^+) \leqslant \mathrm{Var}(\bar{y}_1)$ and $\mathrm{Var}(\hat{\sigma}_y^{+2}) \leqslant \mathrm{Var}(s_{y_1}^2)$.

The pth design quantile using record augmentation is

$$\hat{y}_p^+ = \exp(\hat{\mu}_y^+ + z_p\,\hat{\sigma}_y^+) \tag{5}$$

where z_p is the pth quantile from a standard normal distribution. In this study we compare the precision of estimates of $\hat{y}_p^+$ with the precision associated with the alternative at-site maximum likelihood estimator

$$\hat{y}_p = \exp(\bar{y}_1 + z_p v_{y_1}) \tag{6}$$

where

$$v_{y_1}^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(y_i - \bar{y}_1)^2$$

Stedinger (1980) recommended the use of the maximum likelihood estimator $\hat{y}_p$ for the LN2 distribution. When the cross-correlation, ρ, of the natural logarithms of the streamflows is equal to zero, then $\hat{\mu}_y^+ = \bar{y}_1$ and $\hat{\sigma}_y^{+2} = s_{y_1}^2$, in which case $\hat{y}_p^+$ and $\hat{y}_p$ are almost identical. In general, as ρ and n_2 increase, the variance of $\hat{y}_p^+$ will decrease in comparison to the variance of the at-site estimator $\hat{y}_p$.

AN EXPERIMENT

Twenty-three basins in or near Massachusetts with unregulated streamflows are used to evaluate the merit of using streamflow record augmentation procedures in practice. These basins, described in Table 1, are identical to the basins used in other recent studies by Vogel and Kroll (1989, 1990). Vogel and Kroll (1989) employed probability-plot correlation coefficients (PPCCs) to evaluate various probability distribution functions and parameter estimation procedures for their ability to describe the distribution of annual minimum 7 day low-flows. They showed that the annual minimum 7 day low-flows at these 23 sites are well approximated by a two-parameter log-normal distribution. This study is more general as record augmentation procedures are evaluated for estimating quantiles of the distribution of annual minimum d day low-flows (d = 1, 3, 7, 14, and 30 days) and annual maximum flood-flows. The 7-day

TABLE 1

US Geological Survey gaging stations, record lengths and drainage areas

USGS gage no.	Site No.	Record length (years)	Drainage area (mile^2)
01180500	1	73	52.70
01096000	2	34	63.69
01106000	3	37	8.01
01170100	4	16	41.39
01174000	5	34	3.39
01175670	6	23	8.68
01198000	7	19	51.00
01171800	8	11	5.46
01174900	9	22	2.85
01101000	10	38	21.30
01187400	11	31	7.35
01169000	12	44	89.00
01111300	13	20	16.02
01169900	14	17	24.09
01181000	15	48	94.00
01332000	16	52	40.90
01097300	17	20	12.31
01333000	18	34	42.60
01165500	19	65	12.10
01171500	20	45	54.00
01176000	21	71	150.00
01162500	22	63	19.30
01180000	23	28	1.73

10-year low-flow statistic, $Q_{7,10}$, is the most widely used index of low-flow in the United States (Riggs et al., 1980).

A regional probability distribution of annual minimum d *day low-flows and annual maximum flood-flows*

Filliben (1975), Vogel (1986, 1987) and Vogel and Kroll (1989) described the use of the PPCC for testing alternative at-site distributional hypotheses. Log-normal probability plots were constructed at all 23 sites for the annual minimum d = 1, 3, 7, 14, and 30 day low-flows and the annual maximum flood-flows. In each case a log-normal PPCC test statistic was estimated and the at-site significance level, α, associated with each hypothesis test was computed. In theory, each set of 23 significance levels should be uniformly distributed over the interval [0, 1], if the sites are independent across space. Hence a regional log-normal PPCC test may be performed by constructing uniform probability plots of the significance levels associated with each of the flow series. Vogel and Kroll (1989) provided a detailed discussion of the application of regional uniform PPCC hypothesis tests.

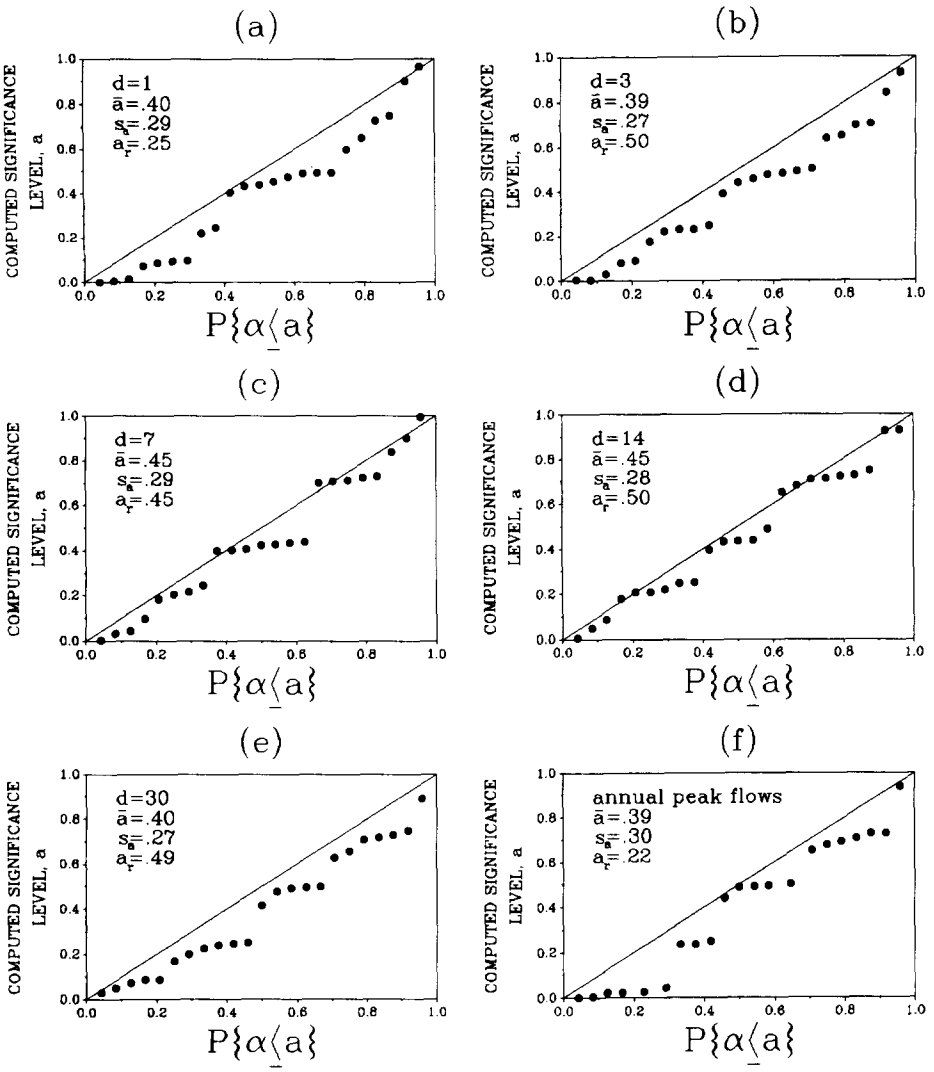

Fig. 1. Regional uniform probability plots for 23 sites in Massachusetts corresponding to fits of a log-normal model to sequences of annual-minimum d-day low-flows and annual maximum flood-flows.

Figure 1 depicts the regional uniform probability plots of the significance levels associated with each of the six flow series considered. In theory, one would expect the average significance level and the standard deviation of the significance levels to be 0.5 and 0.289, respectively, as the significance levels, a, are in theory uniformly distributed over the interval [0,1]. The sample mean significance level $\bar{a}$, and the sample standard deviation of the significance level s_a, associated with each flow series are shown in Fig. 1. In addition, the

regional significance level, a_r, determined from the uniform PPCC hypothesis test (see Vogel and Kroll, 1989), is provided. We would expect a_r to be close to 0.5 if the regional sample evaluated is a 'typical' regional log-normal sample. Figure 1 provides consistent evidence, in terms of the values of $\bar{a}$, s_a, a_r and the uniform probability plots, that the annual minimum $d = 3$, 7, 14, and 30 day low-flow series are well approximated by a two-parameter log-normal distribution on a regional basis in Massachusetts. The annual minimum 1 day low-flows and the annual maximum flood-flow series are not nearly as well approximated by a two-parameter log-normal distribution as the other flow series considered. Nevertheless, we assume that those flow series are two-parameter log-normal for the purposes of evaluating the merit of streamflow record augmentation procedures.

Autocorrelation of flow series

Figure 2 displays estimates of the first-order serial correlation coefficient, $\hat{\rho}_1$, of the logarithms of the annual minimum $d = 1$, 7, and 30 day low-flows and the logarithms of the annual maximum flood-flows for the 23 sites. Here $\hat{\rho}_1$ is the estimator recommended by Jenkins and Watts (1968) which has $E[\hat{\rho}_1] = -1/n$

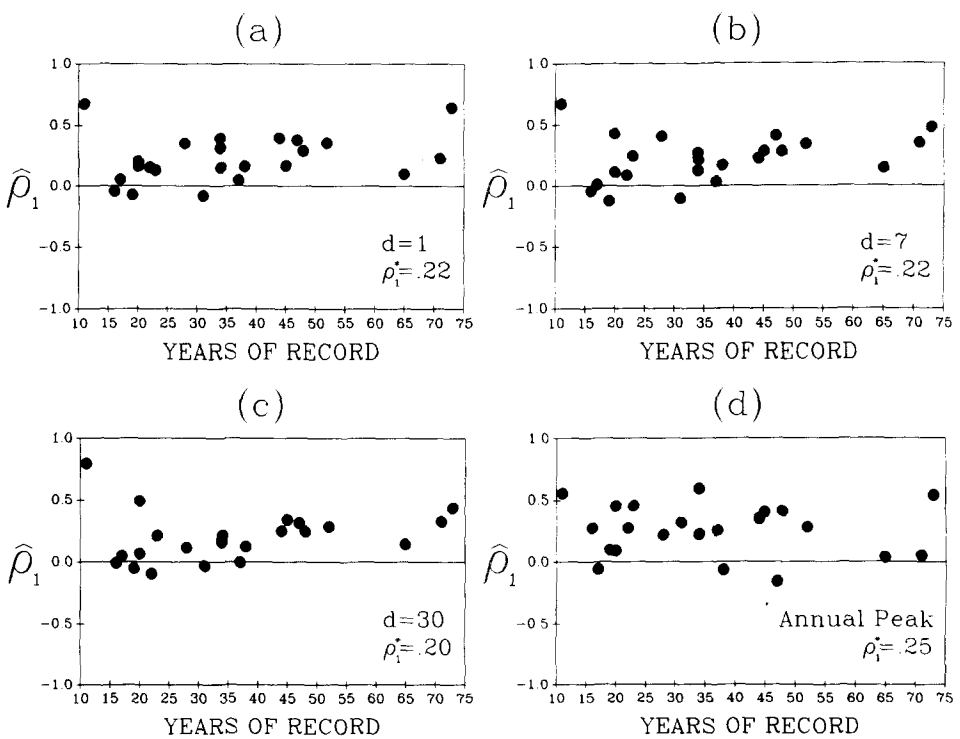

Fig. 2. Estimates of the lag-one serial correlation coefficient $\hat{\rho}_1$ corresponding to sequences of annual-minimum d-day low-flow series and annual maximum flood-flow series (ρ_1^* is the regional mean value of $\hat{\rho}_1$).

and $\text{Var}[\hat{\rho}_1] = 1/n$ if the observations are independent (Loucks et al., 1981, p. 173). Figure 2 depicts $\{E(\hat{\rho}_1) \pm 2[\text{Var}[\hat{\rho}_1]]^{1/2}\}$ using dashed lines. All the flow series depicted in Fig. 2 appear to exhibit serial correlations which are significantly different from zero at a number of sites. In the following experiments we assume that all six flow series arise from a Markov process. Unbiased estimates of ρ_1 for a Markov process were obtained for each flow series using an approach suggested by Tasker (1983b). Individual unbiased at-site sample estimates of ρ_1 tend to contain substantial sampling variability, hence the average of all 23 at-site unbiased estimates, termed ρ_1^*, is assumed to characterize the regional serial correlation of each Markov flow series.

As low-flows are essentially baseflows that originate from groundwater storage, one would expect a fair amount of temporal correlation owing to persistence of carry-over storage effects. The annual maximum flood-flow series also show evidence of autocorrelation, yet most investigators treat such series as independent. Interestingly, of the four flow series in question in Fig. 2, the annual maximum flood-flow series has the largest value of ρ_1^*.

If the flow series are nonstationary then sample estimates of the serial correlation of flow series will be influenced by trends. To confirm that the sample autocorrelations provide evidence of persistence only, and not a trend, an experiment is performed. The logarithms of the annual minimum 7 day flow series were plotted against time, and ordinary least-squares regression was employed to fit relations of the form $\ln(Q_{t,7}) = a + bt$, where $Q_{t,7}$ is the annual minimum 7 day flow in year t, a and b are constants, and t is time. The slope term b, was significantly different from zero (using a 5% significance level) at only four of the 23 sites considered (sites 8, 17, 22, and 23). This experiment provides evidence that the series of annual minimum 7 day low-flows employed in this study do not exhibit a linear trend. Hence, the assumption that these flow-series arise from a stationary Markov process is probably reasonable.

Effective record length of the at-site estimator $\hat{y}_p$

The effective record length of the at-site estimator $\hat{y}_p$ is often considered to be the short-record length, n_1. This will only be the case when the observations are independent in time. If the log-transformed flow series are assumed to originate from a Markov process, then the effective record length of the at-site estimator $\hat{y}_p$ is

$$n_1^- = n_1 \left[\frac{\text{Var}(\hat{y}_p \mid \rho_1 = 0)}{\text{Var}(\hat{y}_p \mid \rho_1 = \rho_1^*)} \right] \tag{7}$$

where $\text{Var}[\hat{y}_p|\rho_1]$ is derived in the Appendix. In general, for a Markov process with $\rho_1 > 0$, $\text{Var}(\hat{y}_p|\rho_1)$ will be larger than for an independent process, hence the series will convey less information about $\hat{y}_p$ than a random (independent) series of the same length. Thus eqn. (7) documents the reduction in the information content or record length associated with $\hat{y}_p$ owing to the autocorrelation

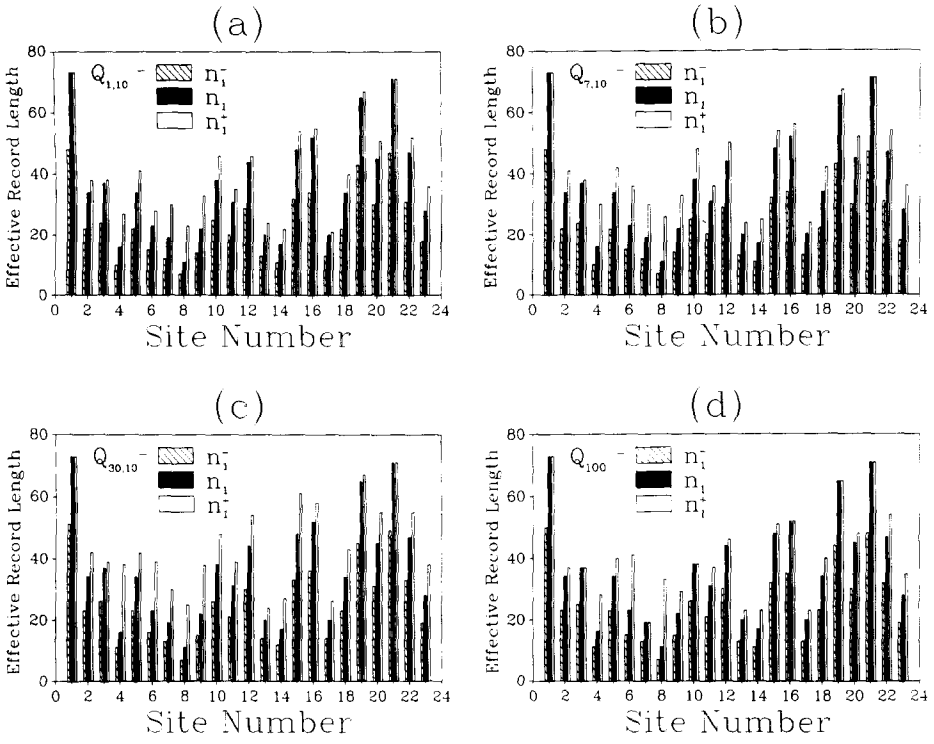

Fig. 3. Comparison of the original short-record length n_1 with the effective record length n_1^- which accounts for serial correlation and n_1^+ which accounts for record augmentation.

of the flow series. When the flow series are independent, ρ_1^* will equal zero in which case $n_1^- = n_1$; otherwise for $\rho_1 > 0$, $n_1^- < n_1$.

Figure 3 compares the original short-record length n_1 and the effective record length n_1^- associated with the at-site estimator y_p at the 23 gaged basins described in Table 1. In Fig. 3 we compare n_1 with the values of n_1^- corresponding to the at-site estimators of the d-day 10-year low-flow statistic, $Q_{d,10}$ (d = 1, 7, and 30 days) and the 100 year flood-flow Q_{100}. There are significant reductions of the effective record length which result from the autocorrelation of the time-series of both annual minimum d day low-flows and annual maximum flood-flows.

Table 2 and Fig. 4 summarize the regional loss of information owing to the autocorrelation of the flow series by summing the values of n_1^- across sites. In Fig. 4(a) and Table 2(a) we consider only the six sites with record lengths, $n_1 \leqslant 20$ years. Similarly, in Figs. 4(b) and 4(c) and Tables 2(b) and 2(c) we consider the nine sites with record lengths $n_1 \leqslant 30$ years and the 15 sites with record lengths $n_1 \leqslant 40$ years, respectively. Finally, in Fig. 4(d) and Table 2(d) we consider all 23 sites. Among the 23 sites in this study, there are $\sum n_1 = 829$ site-years of streamflow records, where $\sum$ denotes the summation over sites.

TABLE 2(a)

Regional gains in effective record length owing to record augmentation compared with regional losses in effective record length owing to serial correlation for the six sites with $n_1 \leqslant 20$

	$Q_{d,10}$; d-day, 10-year low-flow					
	$d = 1$	$d = 3$	$d = 7$	$d = 14$	$d = 30$	Q_{100}
$\sum n_1$	103	103	103	103	103	103
$\sum n_1^-$	66	67	66	68	71	68
$\sum n_1^+$	147	154	159	163	170	149
$\sum n_1^*$	110	118	122	128	138	114

$\sum$ denotes the sum over all 23 gaged sites.

TABLE 2(b)

Regional gains in effective record length owing to record augmentation compared with regional losses in effective record length owing to serial correlation for the nine sites with $n_1 \leqslant 30$

	$Q_{d,10}$; d-day, 10-year low-flow					
	$d = 1$	$d = 3$	$d = 7$	$d = 14$	$d = 30$	Q_{100}
$\sum n_1$	176	176	176	176	176	176
$\sum n_1^-$	113	116	113	117	121	117
$\sum n_1^+$	244	254	264	270	285	254
$\sum n_1^*$	181	194	201	211	230	195

$\sum$ denotes the sum over all 23 gaged sites.

TABLE 2(c)

Regional gains in effective record length owing to record augmentation compared with regional losses in effective record length owing to serial correlation for the 15 sites with $n_1 \leqslant 40$

	$Q_{d,10}$; d-day, 10-year low-flow					
	$d = 1$	$d = 3$	$d = 7$	$d = 14$	$d = 30$	Q_{100}
$\sum n_1$	384	384	384	384	383	384
$\sum n_1^-$	248	256	248	257	263	258
$\sum n_1^+$	482	497	511	520	538	483
$\sum n_1^*$	346	369	375	393	417	357

$\sum$ denotes the sum over all 23 gaged sites.

TABLE 2(d)

Regional gains in effective record length owing to record augmentation compared with regional losses in effective record length owing to serial correlation for all 23 sites

	$Q_{d,10}$; d-day, 10-year low-flow					
	$d = 1$	$d = 3$	$d = 7$	$d = 14$	$d = 30$	Q_{100}
$\sum n_1$	829	829	829	829	829	829
$\sum n_1^-$	542	552	542	555	571	559
$\sum n_1^+$	951	969	988	998	1032	943
$\sum n_1^*$	664	692	701	724	774	673

$\sum$ denotes the sum over all 23 gaged sites.

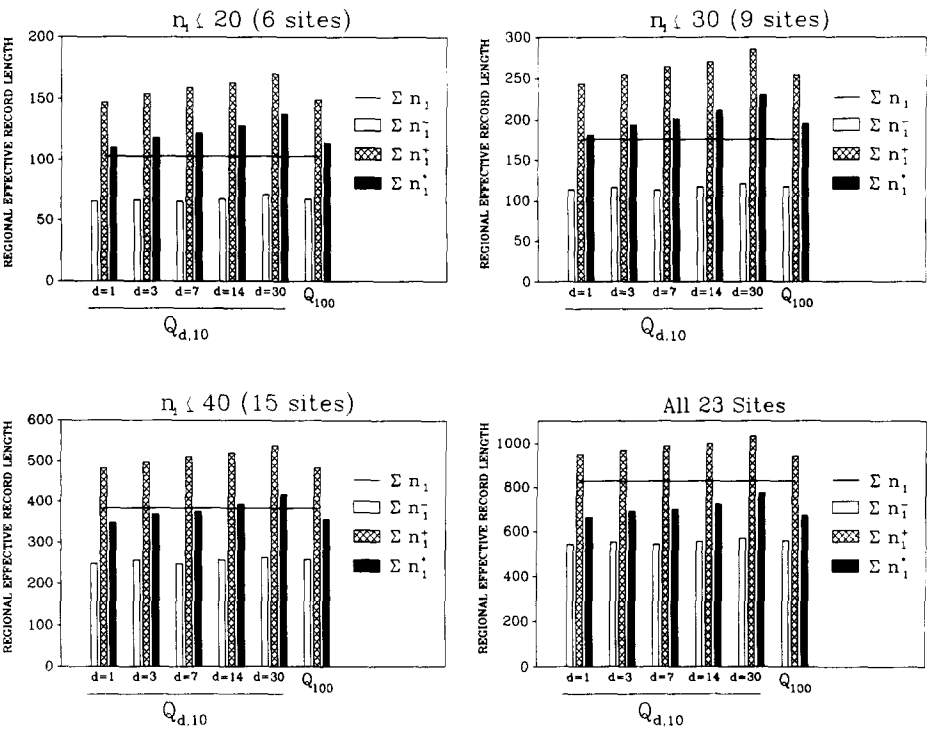

Fig. 4. Comparison of the regional effective record length, which accounts for both serial correlation and record augmentation $\sum n_1^*$, with the regional effective record length accounting for serial correlation only, $\sum n_1^-$, and record augmentation only, $\sum n_1^+$, corresponding to the various low-flow and flood-flow quantile estimates.

The effective record length associated with the at-site estimator $\hat{y}_p$, for the entire region, $\sum n_1^-$, is obtained by applying eqn. (7) to all 23 sites and summing the resulting values of n_1^- across sites. Values of the regional effective record length $\sum n_1^-$ in Table 2(d) and Fig. 4(d) range from 542 to 571 site-years when estimating quantiles of either annual minimum d day low-flows or annual maximum flood-flows. Therefore, the serial correlation of the flow-series has effectively reduced the total amount of regional information by between 258 and 287 site-years or between 31 and 35% of the total information that would be available if these flow series were not serially correlated. The percentage losses of regional information owing to serial correlation among the 6, 9, and 15 sites considered in Tables 2(a)–2(c) and Figs. 4(a)–4(c) are similar to the losses experienced at all 23 sites. Interestingly, the loss of information owing to serial correlation is approximately the same for all the at-site low-flow and flood-flow statistics considered here

Effective record length of the augmented estimator $\hat{y}_p^+$

In the previous section we focused on the decrease in the effective record lengths which results from the serial correlation of both low-flow and flood-flow series. In this section we compare the increase in effective record lengths resulting from the use of record augmentation procedures with the decrease in effective record length owing to serial correlation of the flow series. First, ignoring the impact of the serial correlation of the flow-series, we define the effective record length associated with the streamflow record augmentation estimator $\hat{y}_p^+$ as

$$n_1^+ = n_1 \left[\frac{\mathrm{Var}(\hat{y}_p | \rho_1 = 0)}{\mathrm{Var}(\hat{y}_p^- | \rho_1 = 0)} \right] \tag{8}$$

where $\mathrm{Var}(\hat{y}_p^+)$ and $\mathrm{Var}(\hat{y}_p)$ are derived in the Appendix.

Values of n_1^+ were computed for all possible site interactions and only the largest value of n_1^+ computed for each short-record site is summarized here. Each long-record site x, is chosen as that site which maximizes the transfer of information, using n_1^+ as the index of information transfer. This procedure mimics what is done in practice, because hydrologists normally choose a long-record site as that site which has the highest cross-correlation with the short-record site. Sites with large values of both ρ and n_2 will produce the largest effective record lengths n_1^+.

Again Fig. 3 compares the original short-record length n_1 with the effective record lengths n_1^+ associated with record augmentation estimators of the d-day 10-year low-flow statistics $Q_{d,10}$ (d = 1, 7, and 30 days) and the 100-year flood-flow, Q_{100}. As expected, $n_1^- > n_1 > n_1^-$ at all sites, with the exception of Site 1 which had no neighbors with longer records to exploit. The gains in information as a result of record augmentation may be defined by the difference $n_1^+ - n_1$. Those gains are largest for the short-record sites with record lengths less than about 20 years.

Table 2 and Fig. 4 report the regional effective record lengths $\sum n_1^+$ owing to record augmentation, ignoring the impact of serial correlation, which are simply the summation over sites of the values of n_1^+ computed from (8). Values of $\sum n_1^-$ are always larger than $\sum n_1$. For the low-flow statistic $Q_{d,10}$, values of $\sum n_1^+$ increase as the averaging period, d, increases, though the differences are relatively small. Among the six sites with $n_1 \leqslant 20$ years, Fig. 4(a) and Table 2(a) document that the use of streamflow record augmentation procedures has effectively increased the amount of regional information by 44–67 site-years, or between 43 and 65% of the total information which would be available if record augmentation were not performed, ignoring the impact of serial correlation. Among the 15 sites with $n_1 \leqslant 40$ years, Fig. 4(c) and Table 2(c) document that the use of streamflow record augmentation procedures has effectively increased the amount of regional information by 98–154 site-years, or between 26 and 40% of the total information which would be available if record augmentation were not performed, ignoring the impact of serial correlation. The regional gains in information transfer are lowest when one considers all 23 sites in Table 2(d) and Fig. 4(d) owing to the inclusion of many long-record sites where record augmentation results in little or no transfer of information.

The regional effective record length owing to the use of record augmentation is actually lower than that computed using $\sum n_1^-$ because (8) ignores the impact of serial correlation. A better estimate of the regional effective record length associated with the record augmentation estimator y_p^+ would be

$$\sum n_1^* = \sum n_1^+ + \sum n_1^- - \sum n_1 \tag{9}$$

Table 2 and Fig. 4 contain the values of $\sum n_1^*$ corresponding to the various estimators of low-flow and flood-flow statistics. When one considers only the short-record sites with $n_1 \leqslant 20$ and $n_1 \leqslant 40$ in Figs. 4(a) and 4(b), one observes that $\sum n_1^*$ is larger than $\sum n_1$ for all the flow-series considered. However, when one considers all 23 sites in Fig. 4(d), $\sum n_1^*$ is considerably less than $\sum n_1$. Apparently, the gains in information owing to record augmentation are only greater than the losses owing to serial correlation at sites with $n_1 \leqslant 30$. The regional gains in information owing to record augmentation and the regional losses in information owing to serial correlation are very similar for the flood-flow statistic Q_{100} and the widely used low-flow statistic $Q_{7,10}$. Considering only the sites with record lengths $n_1 \leqslant 30$ in Fig. 4(b) and Table 2(b), the effective regional record length $\sum n_1^*$ ranges from 181 to 230 site-years as compared with the regional record lengths $\sum n_1^-$ which range from 113 to 121 site-years when one accounts for the impact of serial correlation. Thus streamflow record augmentation procedures have increased the regional information content by between 60 and 90% for these flow-series.

Figure 5 compares values of n_1^* and n_1 at each of the 23 sites for the statistics $Q_{d,10}$ (d = 1, 7, and 30 days) where n_1^* is obtained using eqn. (9); $n_1^* = (n_1^+ + n_1^- - n_1)$. Although in most cases $n_1 > n_1^*$, there are a few short-record sites in which $n_1 < n_1^*$. In particular, Sites 4 and 8 depict situations in

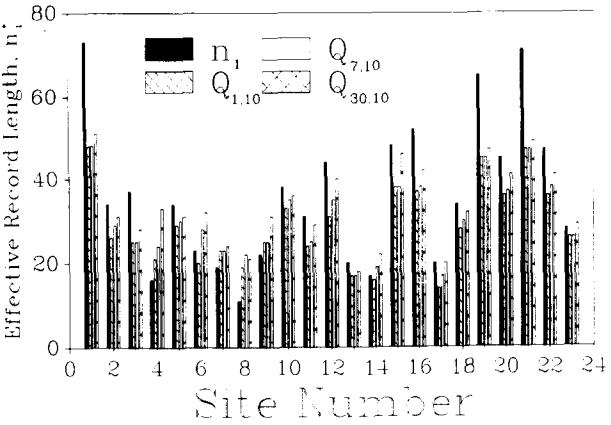

Fig. 5. Comparison of the effective record lengths n_1^* associated with the low-flow statistic $Q_{d,10}$ (d = 1, 7, and 30 days).

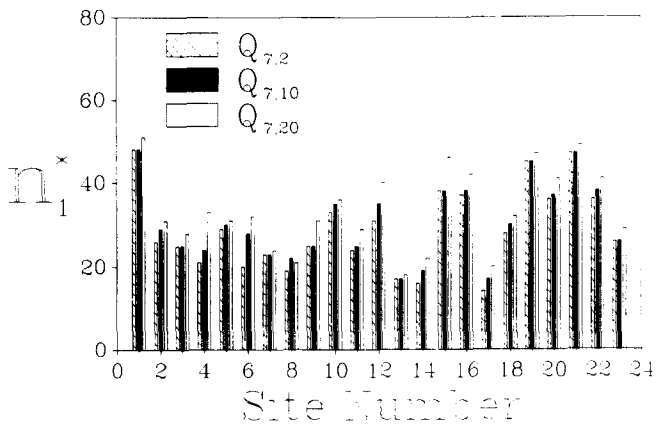

Fig. 6. Comparison of the effective record length n_1^* associated with the low-flow statistic $Q_{7,T}$ for average return periods T = 2, 10, and 20 years.

which record augmentation results in substantial gains in information transfer, even after taking into account the impact of serial correlation.

Figure 6 compares the effective record lengths n_1^*, associated with estimates of $Q_{7,T}$, for average return periods T = 2, 10, and 20 years. Clearly, n_1^* increases with T, although the differences are relatively small.

CONCLUSIONS

The primary objective of this study was to evaluate the merit of using streamflow record augmentation procedures in practice. Using a total of 829

site-years of streamflow records at 23 U.S. Geological Survey gaging stations in Massachusetts, the following conclusions have been reached:

(1) Flow series of annual-minimum d-day low-flows (d = 1, 3, 7, 14, and 30 days) and annual maximum flood-flows exhibit significant serial correlation in Massachusetts. Estimates of the average lag-one serial correlation, ρ_1^*, of these flow series ranged from ρ_1^* = 0.20–0.22 for the low-flow series to ρ_1^* = 0.25 for the flood-flow series (see Fig. 2). The serial correlation of the flow series has effectively reduced the total amount of regional information by between 258 and 287 site-years or between 31 and 35% of the total information which would be available if these flow series were not serially correlated.

(2) Record augmentation procedures, which exploit the cross-correlation of nearby flow series, may be used to increase the information content associated with short streamflow records. For the 23 sites considered in Massachusetts, the use of record augmentation procedures has effectively increased the amount of regional information by between 114 and 203 site-years for these flow series. Ignoring the impact of serial correlation, this increase amounts to between 14 and 24% of the total information which would be available if record augmentation were not performed. The percentage gains in total information are much larger when only short-record sites are considered. For the nine sites with streamflow records of length 30 years or less, the gain resulting from record augmentation procedures represents a 39–62% increase in total information, ignoring the impact of serial correlation. For annual-minimum d-day flow series, the gains in information resulting from record augmentation increase as d increases.

(3) At short-record sites (record lengths of 30 years or less), the increase in regional information resulting from the use of streamflow record augmentation is larger than the decrease in regional information resulting from the serial correlation associated with each flow series. In general, the gains in information transfer were largest at the sites with the shortest record lengths. For example, Fig. 5 documents situations (Sites 4 and 8) in which the effective record length n_1^* is almost double the original record length n_1. Here n_1^* is equal to the effective record length accounting for both the impact of record augmentation and serial correlation.

(4) Using the regional hypothesis tests recommended by Vogel and Kroll (1989), we found that sequences of annual-minimum 3-, 7-, 14- and 30-day low-flows were extremely well approximated by a two-parameter log-normal distribution in Massachusetts. This result is useful for developing generalized regional regression equations for low-flow statistics as only two parameters need to be regionalized (see Vogel and Kroll, 1990).

Tasker (1983a) has already warned us that the serial correlation of flow series can result in dramatic reductions in the information content, or effective record length, associated with those series. Furthermore, the potential gains due to streamflow record augmentation procedures have also been well documented (Fiering, 1963; Matalas and Jacobs, 1964; Vogel and Stedinger, 1985). Perhaps the most important result here is that, in practice, after

employing streamflow record augmentation procedures to counterbalance the loss of information owing to serial correlation, the resulting effective record lengths, n_1^*, are usually larger than the originally assumed record length, n_1, at the very-short-record sites. This result is consistent with the study by Vogel and Stedinger (1985) which suggested that the most significant gains owing to record augmentation should occur at the very-short-record sites.

Given the often dramatic reductions in effective record lengths associated with the impact of serial correlation, the increase in effective record lengths which result from record augmentation become even more significant. To counterbalance the reductions in information owing to serial correlation, more attention should be given to the development of multivariate streamflow record augmentation procedures instead of the two-stream record augmentation procedures used in this study. Multivariate record augmentation procedures were originally recommended by Fiering (1963) when he extended the two-stream model to a three-stream model. Recently, Kuczera (1987) and Grygier et al. (1989) have suggested more general multivariate procedures for extending and augmenting short streamflow series.

ACKNOWLEDGMENTS

This research was supported by a cooperative agreement between Tufts University and the US Geological Survey with matching funds from the Massachusetts Division of Water Pollution Control in the Department of Environmental Protection. The writers are indebted to Jery R. Stedinger for his review of an early draft of this manuscript.

REFERENCES

Fiering, M.B., 1963. Use of correlation to improve estimates of the mean and variance. U.S. Geol. Surv., Prof. Pap., 434-C, pp. 20–32.

Filliben, J.J., 1975. The probability plot correlation coefficient test for normality. *Technometrics*, 17(1): 111–117.

Grygier, J.C., Stedinger, J.R. and Yin, H.B., 1989. A generalized maintenance of variance extension procedure for extending correlated series. *Water Resour. Res.*, 25 (3): 345–349.

Interagency Advisory Committee on Water Data, 1982. Guidelines for Determining Flood Flow Frequency. Washington, DC.

Jenkins, G.M. and Watts, D.G., 1968. *Spectral Analysis and its Applications*. Holden-Day, San Francisco, CA.

Kuczera, G., 1987. On maximum likelihood estimators for the multisite lag-one streamflow model: complete and incomplete data cases. *Water Resour. Res.*, 23(4): 641–645.

Loucks, D.P., Stedinger, J.R. and Haith, D.A., 1981. *Water Resource Systems, Planning and Analysis*. Prentice-Hall, New Jersey, 559 pp.

Matalas, N.C. and Jacobs, B., 1964. A correlation procedure for augmenting hydrologic data. U.S. Geol. Surv., Prof. Pap., 434-E, pp. E1–E7.

Riggs, H.C. et al., 1980. Characteristics of Low Flows. ASCE, *J. Hydraul. Eng.*, 106(5): 717–731.

Stedinger, J.R., 1980. Fitting log normal distributions. *Water Resour. Res.*, 16(3): 481–490.

Tasker, G.D., 1983a, Effective record length for the *T*-year event. *J. Hydrol.*, 64: 39–47.

Tasker, G.D., 1983b. Approximate sampling distribution of the serial correlation coefficient for small samples. *Water Resour. Res.*, 19(2): 579–582.

Vogel, R.M., 1986. The probability plot correlation coefficient test for the normal, lognormal and gumbel distributional hypotheses. *Water Resour. Res.*, 22 (4): 587–590 (see Vogel, 1987).
Vogel, R.M., 1987. Correction to: The probability plot correlation coefficient test for the normal, lognormal and Gumbel distributional hypotheses. *Water Resour. Res.*, 23 (10): 2013.
Vogel, R.M. and Kroll, C.N., 1989. Low-flow frequency analysis using probability plot correlation coefficients. ASCE, *J. Water Resour. Plan. Manage.*, 115(3): 338–357.
Vogel, R.M. and Kroll, C.N., 1990. Generalized low-flow frequency relationships for ungaged sites in Massachusetts. *Water Resour. Bull.*, 26 (2): 241–253.
Vogel, R.M. and Stedinger, J.R., 1985. Minimum variance streamflow record augmentation procedures. *Water Resour. Res.*, 21 (5): 715–723.

APPENDIX

First-order approximation to the variance of a log-normal quantile

Equations (5) and (6) may be rewritten as

$$\hat{y}_p = \exp\left[\hat{\omega}_1 + z_p\,(\hat{\omega}_2)^{1/2}\right] \tag{A1}$$

with $\hat{\omega}_1 = \bar{y}_1$ and $\hat{\omega}_2 = v_{y_1}^2$ yielding the estimator $\hat{y}_p^-$; and with $\hat{\omega}_1 = \hat{\mu}_y$ and $\hat{\omega}_2 = \hat{\sigma}_y^{+2}$ yielding the estimator $\hat{y}_p^+$. Equation (A1) may be approximated using a first-order Taylor series about the true values of ω_1 and ω_2

$$\hat{y}_p \simeq \exp[\omega_1 + z_p\,(\omega_2)^{1/2}] + (\delta\, y_p/\delta\omega_1)\,(\omega_1 - \hat{\omega}_1) + (\delta\, y_p/\delta\omega_2)\,(\omega_2 - \hat{\omega}_2) \tag{A2}$$

where

$$\delta y_p/\delta\omega_1 = \hat{y}_p$$

$$\delta y_p/\delta\omega_2 = \tfrac{1}{2}\,\hat{y}_p z_p \hat{\omega}_2^{-1/2}$$

Now the variance of y_p may be approximated as

$$\mathrm{Var}(\hat{y}_p) \simeq \hat{y}_p^2\,\mathrm{Var}(\hat{\omega}_1) + [(\hat{y}_p\, z_p)^2/(4\,\hat{\omega}_2)]\,\mathrm{Var}(\hat{\omega}_2) + \hat{y}_p^2 z_p\,(\hat{\omega}_2)^{-1/2}\,\mathrm{Cov}(\hat{\omega}_1, \hat{\omega}_2) \tag{A3}$$

An approximation to the variance of the at-site estimator $\hat{y}_p$

The at-site estimator $\hat{y}_p$ is defined by $\hat{\omega}_1 = \bar{y}_1$ and $\hat{\omega}_2 = v_{y_1}^2$ in eqn. (A1). Loucks et al. (1981, pp. 170–173) document that

$$\mathrm{Var}(\bar{y}_1) = \frac{\sigma_y^2}{n_1}\left\{1 + \frac{2\rho_1[n_1(1 - \rho_1) - (1 - \rho_1^{n_1})]}{(1 - \rho_1)^2}\right\} \tag{A4}$$

and

$$\mathrm{Var}(v_{y_1}^2) \simeq \frac{2\sigma_y^4}{n_1}\left(\frac{1 + \rho_1^2}{1 - \rho_1^2}\right) \tag{A5}$$

where ρ_1 is equal to the lag-one serial correlation of the log-transformed

streamflows. Equations (A4) and (A5) correspond to the variance of a sample mean and variance, respectively, if the log-transformed flows are assumed to originate from a Markov process. Equations (A4) and (A5) indicate that the variance of the sample mean and variance are increasing functions of both σ_y^2 and ρ_1. As the log-transformed streamflows are normally distributed

$$\text{Cov}(\bar{y}_1, v_{y_1}^2) = 0 \tag{A6}$$

The variance of the at-site estimator $\hat{y}_p$ is obtained by substitution of eqns. (A4), (A5) and (A6) into eqn. (A3).

An approximation to the variance of augmented estimator $\hat{y}_p^+$

The augmentation estimator $\hat{y}_p^+$ is defined by $\hat{\omega}_1 = \hat{\mu}_y^+$ and $\hat{\omega}_2 = \hat{\sigma}_y^{+2}$ in eqn. (A1). Vogel and Stedinger (1985) showed that

$$\text{Var}(\hat{\mu}_y^+) = \frac{\sigma_y^2}{n_1}\left(1 - \frac{n_2\,\theta_1\,\rho^2}{n_1 + n_2}\right) \tag{A7}$$

Similarly, Vogel and Stedinger (1985) provided expressions for $\text{Var}(\hat{\sigma}_y^{+2})$; however, those formulae are too complex to reproduce here. Monte-Carlo experiments similar to those described in Vogel and Stedinger (1985) were performed for: $n_1 = 6, 10, 25$; $\rho = 0.5, 0.7, 0.9$; and $n_2 = 60$. These experiments revealed that

$$\text{Cov}(\hat{\mu}_y^+, \hat{\sigma}_y^{+2}) = 0 \tag{A8}$$

for all values of n_1, and n_2 and ρ considered. The variance of the augmentation estimator $\hat{y}_p^+$ is obtained by substitution of eqns. (A7), (A8), and $\text{Var}(\hat{\sigma}_y^{+2})$ (from Vogel and Stedinger, 1985) into eqn. (A3).