Estimation of 7-day, 10-year low-streamflow statistics using baseflow correlation

Christine F. Reilly[1] and Charles N. Kroll

Department of Environmental Resource Engineering, State University of New York College of Environmental Science and Forestry, Syracuse, New York, USA

[1] Estimators of low-streamflow statistics are often required for water quality and quantity management at gauged and ungauged sites. Here the technique of baseflow correlation is investigated as an estimator of the 7-day, 10-year low-streamflow statistic at ungauged sites. This method uses an information transfer technique to estimate streamflow statistics at an ungauged site by correlating a nominal number of measured streamflow discharges during baseflow conditions at the ungauged site with those at a nearby gauged site. A regional assessment of baseflow correlation estimators is made by employing daily streamflow values at more than 1300 USGS HCDN stream gauge sites. A jackknife simulation is performed in each of the 18 USGS water resource regions located within the conterminous United States. Potential gauged sites arebselected using a variety of watershed, geographic, topographic, and geologic classification systems. The results of this study indicate that baseflow correlation performs well in the United States when baseflow measurements are nearly independent and potential gauged sites are located within 200 km. The method is improved as the number of baseflow measurements is increased, although some leveling off of performance was observed with more than 15 baseflow measurements. A comparison of baseflow correlation with regional regression shows that baseflow correlation is the preferred method for estimating low flows in much of the United States. *INDEX TERMS:* 1860 Hydrology: Runoff and streamflow; 1854 Hydrology: Precipitation (3354); 1894 Hydrology: Instruments and techniques; *KEYWORDS:* baseflow correlation, low-flow estimation, streamflow

1. Introduction

[2] Knowledge of the magnitude and frequency of low-flow events is required in order to plan water-supply systems, wastewater discharges, reservoirs, and irrigation systems, and to maintain the quality of water for wildlife and recreation [*Smakhtin*, 2001]. In the United States, the 7-day, 10-year low flow ($Q_{7,10}$) is a commonly used low-flow statistic [*Riggs*, 1980]. The $Q_{7,10}$ is the average annual 7-day minimum flow that is expected to be exceeded on average in 9 out of every 10 years, which is equivalent to the tenth percentile of the distribution of 7-day annual minimum streamflows.

[3] When a historic record of streamflow measurements is available, low-streamflow statistics can be estimated using a frequency analysis [*Riggs*, 1972; *Stedinger et al.*, 1993]. In the absence of a long record, reliable methods for estimating streamflow statistics are still needed. A common method used in the United States for estimating low-flow statistics at ungauged river sites is regional regression, a technique that assumes catchments with similar climatic, topographic, and geologic characteristics will have similar streamflow responses [*Smakhtin*, 2001]. Low-flow regional regression models have generally performed poorly in practice, producing estimators with unacceptably large errors [*Thomas and Benson*, 1970; *Barnes*, 1985; *Hammett*, 1985; *Arihood and Glatfelter*, 1986; *Vogel and Kroll*, 1992; *Ries*, 1994; *Kroll et al.*, 2003]. The current study examines baseflow correlation, an alternative method for estimating low-streamflow statistics that uses the cross-correlation of a nominal number of baseflow measurements from an ungauged and a nearby gauged river site to estimate the streamflow statistics at the ungauged site [*Stedinger and Thomas*, 1985].

[4] *Stedinger and Thomas* [1985] examined 20 pairs of river sites that were considered representative of sites in the humid eastern region of the United States. The stations were paired based on similar watershed size and baseflow recession characteristics. The authors used these stations to compare five different methods of estimating the $Q_{7,10}$, including regional regression and baseflow correlation, and found baseflow correlation to be the best estimator. The average root-mean square error of the baseflow correlation method examined by Stedinger and Thomas was 0.19, which corresponds to a standard error of 46%. Because of

this high error, the authors outlined ideas for improving the accuracy of the baseflow correlation estimator. They suggested selecting a gauged site with baseflows that are highly correlated ($\rho > 0.70$) with the baseflows at the ungauged site. Another suggestion was increasing the number of baseflow observations; however, Stedinger and Thomas noted that there is a point where more observations will not significantly reduce error.

[5] *Potter* [2001] used a similar baseflow correlation method as *Stedinger and Thomas* [1985] to estimate low-streamflow statistics. In addition to the assumptions made by Stedinger and Thomas, Potter assumed the baseflow discharges at the gauged and ungauged sites have similar log-variances. The variance in log-space is used because of Potter's assumption of lognormality. Two pairs of adjacent gauged sites in southwestern Wisconsin were selected for Potter's experiment, with only one of these pairs having similar log-variances. The results indicated that Potter's method provides estimators with little bias and low standard error when the sites have similar log-variances, but a large drop in performance occurs when sites do not have similar log-variances. One weakness of Potter's study was the small number of sites examined.

[6] The current study expands upon these previous baseflow correlation experiments. A computer simulation is used to examine the baseflow correlation technique with various experimental parameters at more than 1300 river sites located throughout the United States. The performance of the baseflow correlation technique in different USGS water resource regions is compared in order to examine regional variations in method performance. The computer simulation compares the baseflow correlation technique when (1) the number of measured baseflow days are 5, 10, 15, and 20, (2) various methods for selecting potential gauged sites based on geographic and watershed characteristics are implemented, and (3) different techniques for selecting baseflow measurements from the ungauged site are employed.

2.   Methods

[7] This section outlines the methods used for this experiment and contains four subsections. First, the mathematics of the baseflow correlation technique are presented. The second subsection describes the streamflow database. The computer simulation is outlined in the third subsection. Finally, the methods for selecting potential gauged sites and baseflow segments are presented.

2.1.   Baseflow Correlation

[8] The baseflow correlation method described by *Stedinger and Thomas* [1985] is outlined below. This method is based on an assumed linear relationship between y_i, the logarithm of the annual minimum d-day flows at the ungauged site, and those at the gauged site, x_i:

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

where α and β are model parameters, and ε_i are independent normal error terms with a mean of zero and a constant variance, σ_ε^2. This experiment uses 7-day annual minimum flows for y_i and x_i. Because annual minimum flows are not available for the ungauged site, the linear relationship from equation 1 is adapted to correlate daily (instantaneous)

baseflow measurements at the ungauged site, $\tilde{y}_i$, with corresponding daily baseflows at the gauged site, $\tilde{x}_i$:

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

The assumption that the relationship between annual minimum d-day flows is similar to the relationship between daily baseflows appears reasonable for 7-day annual minimum flows [*Stedinger and Thomas*, 1985]. However, *Stedinger and Thomas* [1985] suggested that this assumption be tested before using d-day annual minimum flows where d is significantly greater than seven.

[9] Here it is assumed that the annual minimum streamflows are described by a log Pearson type 3 (LP3) distribution. The LP3 distribution has been repeatedly employed by the United States Geological Survey (USGS) for describing annual minimum streamflow series in the United States [*Rumenik and Grubbs*, 1996; *Wandle and Randall*, 1993; *Barnes*, 1985]. Using the LP3 distribution, the logarithm of the $Q_{7,10}$ at the ungauged site can be estimated by:

$$\ln\big(\hat{Q}_{7,10}\big) = \hat{\mu}_y + K_y \hat{\sigma}_y \qquad (3)$$

where $\hat{\mu}_y$ is an estimator of the log-space mean, $\hat{\sigma}_y$ is an estimator of the log-space variance, and K_y is the associated frequency factor for the LP3 distribution [*Stedinger et al.*, 1993]. The frequency factor is a function of the log-space skew of the 7-day minimum flows and the percentile of interest, which is the tenth percentile in this case. For the baseflow correlation technique, the frequency factor for the ungauged site, K_y, is assumed equal to the frequency factor for the gauged site, K_x; thus only estimators of $\hat{\mu}_y$ and $\hat{\sigma}_y$ are required.

[10] *Stedinger and Thomas* [1985] suggested the log-space mean and variance of the annual d-day minimum flows at the ungauged site be estimated by:

$$\hat{\mu}_y = a + b m_x \qquad (4)$$

$$\hat{\sigma}_y^2 = b^2 s_x^2 + s_e^2 \left(1 - \frac{s_x^2}{(L-1)s_{\tilde{x}}^2}\right) \qquad (5)$$

where m_x and s_x^2 are the log-space mean and variance of the annual 7-day minimum flows at the gauged site, $s_{\tilde{x}}^2$ is the sample variance of the logarithms of the daily flows at the gauged site, and L is the number of concurrent daily baseflow measurements. The variables a, b, and s_e^2 are the ordinary least squares estimators of the parameters α, β, and σ_ε^2 from equation 2 [*Draper and Smith*, 1966]:

$$a = m_{\tilde{y}} - b m_{\tilde{x}} \qquad (6)$$

$$b = \sum_{i=1}^{L} \frac{\big(\tilde{y}_i - m_{\tilde{y}}\big)\big(\tilde{x}_i - m_{\tilde{x}}\big)}{s_{\tilde{x}}^2(L-1)} \qquad (7)$$

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

In these equations $m_{\tilde{y}}$ and $m_{\tilde{x}}$ are the sample means of the logarithms of the baseflow measurements at the ungauged

site and corresponding flows at the gauged site, respectively. *Stedinger and Thomas* [1985] showed that $\hat{\mu}_y$ and $\hat{\sigma}_y^2$ are unbiased estimators of the log-space mean and variance. The variance of the logarithm of the $Q_{7,10}$ estimator [*Stedinger and Thomas*, 1985] can be estimated by:

$$\text{Var}\left[\ln\left(\hat{Q}_{7,10}\right)\right] = \text{Var}\left(\hat{\mu}_y\right) + \frac{K_y^2}{4\hat{\sigma}_y^2}\text{Var}\left(\hat{\sigma}_y^2\right) + \frac{K_y}{\hat{\sigma}_y}\text{Cov}\left(\hat{\mu}_y, \hat{\sigma}_y^2\right)$$

(9)

Stedinger and Thomas [1985] detail the derivation of this equation to:

$$\text{Var}\left[\ln\left(\hat{Q}_{7,10}\right)\right] \cong \frac{s_e^2}{L} + \frac{(m_x - m_{\bar{x}})^2 s_e^2}{(L-1)s_{\bar{x}}^2} + \frac{b^2 s_x^2}{n} + \frac{K_y^2}{4\hat{\sigma}_y^2}$$
$$\cdot \left(\frac{4b^2 s_x^4 s_e^2}{L s_{\bar{x}}^2} + \frac{2b^4 s_x^4}{n} + \frac{2s_e^4}{L}\right) + \frac{2bs_x^2(m_x - m_{\bar{x}})K_y s_e^2}{L\hat{\sigma}_y s_{\bar{x}}^2}$$

(10)

where the first three terms on the right hand side of equation 10 correspond with $\text{Var}\left(\hat{\mu}_y\right)$, and the last two terms in equation 10 correspond with the last two terms in equation 9, respectively.

2.2. Streamflow Database Development

[11] This study uses data from river sites contained within the Hydro-Climatic Data Network (HCDN), a streamflow data set provided by the USGS [*Slack and Landwehr*, 1992]. The HCDN data set is useful for studying surface water and was specifically developed for examining the effects of climate change on hydrologic conditions. The stream gauge stations selected for inclusion in the HCDN are from locations that are not affected by "artificial diversions, storage, or other human-made works in or on the natural stream channels or watersheds" [*Slack and Landwehr*, 1992]. In addition to streamflow records, the HCDN includes some data regarding the topography, climate, and geology of each watershed. The HDCN has been employed in many streamflow studies [see, e.g., *Vogel et al.*, 1997; *Douglas et al.*, 2000; *Kroll et al.*, 2003].

[12] The HCDN consists of over 1600 stream gauge stations located throughout the United States and its Territories. Most of the stations have at least 20 years, and as much as 114 years of daily average streamflow measurements, for a total of more than 26 million days of streamflow data. On average, each HCDN station has approximately 44 years of recorded data [*Slack and Landwehr*, 1992]. This experiment considers the approximately 1300 HCDN stream gauge stations located within the continental United States that are designated as having data suitable at a daily time step. In addition, HCDN sites with at-site $Q_{7,10}$ estimates of zero are also excluded from this study.

2.3. Computer Simulation

[13] A jackknife simulation was performed to evaluate the use of the baseflow correlation method for estimating low-streamflow statistics. The simulation begins by designating one HCDN site as the ungauged site and marking the days that site has baseflow conditions. The remainder of the
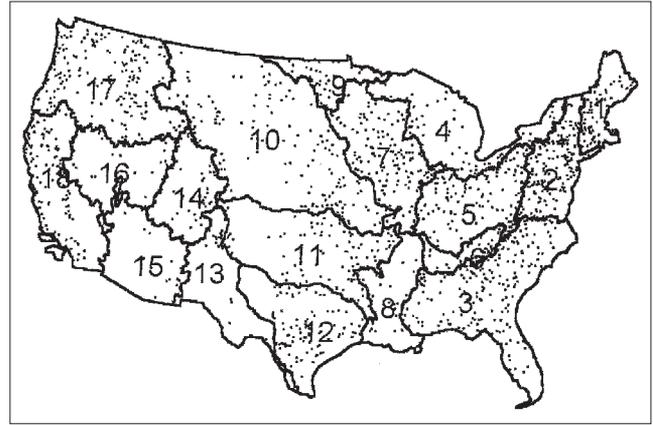
Figure 1. USGS water resource regions with HCDN stations indicated by dots. From *Douglas et al.* [2002], reproduced by permission of the American Society of Civil Engineers (www.pubs.asce.org).

sites are designated as potential gauged sites. To designate days where streamflow is only comprised of baseflow, an empirical formula based on a watershed's drainage area has been suggested:

$$N = A^{0.2}$$

(11)

where N is the time in days from the peak of the hydrograph to the end of surface runoff and A is the drainage area, in square miles, of the watershed (the work of Linsley et al., 1949, as discussed by *Bras* [1990]). In this experiment the number of days of decreasing streamflow (flow less than or equal to the previous day's flow) until baseflow is present at the ungauged site was 6 days, and at a gauged site 4 days. From equation 11, a 6 day designation corresponds to a drainage area of approximately 7800 mi^2, and a 4 day designation approximately 1000 mi^2. In the HCDN database, more than 98% of the sites had a drainage area less than 7800 mi^2, and more than 79% had a drainage area less than 1000 mi^2; therefore these designations appear to be relatively conservative. This experiment uses a longer period of decreasing flow to designate baseflow conditions at the ungauged site than at the gauged site, with the motivation of increasing the likelihood of finding baseflow conditions simultaneously at the gauged and ungauged sites. It should be noted that streamflow values of zero are not included in this analysis, since the logarithm of the flow is required to calculate the sample statistics in equations 6, 7 and 8.

[14] Only flows during the months of July through October are examined because this is the time period when annual low flows occur in most areas of the continental United States, and thus baseflow conditions are more likely to be present during these months. It should be noted that in some regions of the United States (mostly the northern portions of hydrologic regions 4, 7, 9, and 10 in Figure 1), annual d-day minimum flows sometimes occur during the winter months. This is due to prolonged below freezing temperatures which cause precipitation to be stored as snow and ice on the watershed. Histograms of the occurrence of 7-day annual minimums were constructed for the entire United States using the HCDN database. Even in the northern regions noted above, 7-day annual minimums
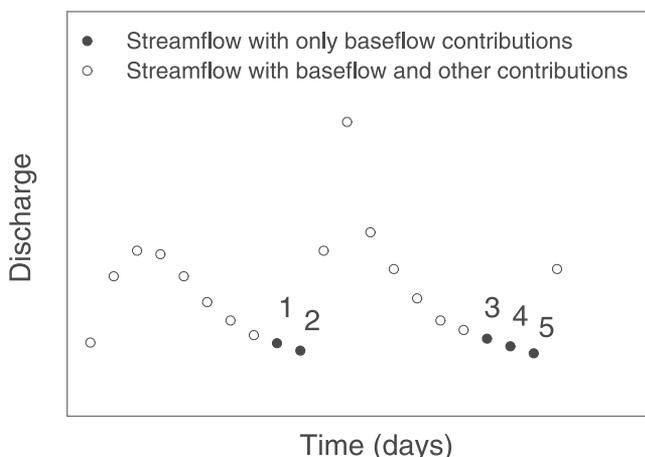
Figure 2. Streamflow record indicating baseflow conditions present on the fourth day of decreasing streamflow.

frequently occurred during the months of July through October.

[15] A specified number of daily streamflow values under baseflow conditions are chosen from the ungauged site using one of the selection methods described in the Experimental Parameters section below. This sequence of flows is called a baseflow segment. Potential gauged sites are searched to find candidate gauged sites that have baseflow on the same days as those in the baseflow segment. Criteria for determining potential gauged sites are also discussed in the Experimental Parameters section below. If a potential gauged site has baseflow conditions for all of the days in a specific baseflow segment, then the site is designated as a candidate gauged site. The baseflow correlation method can be used to estimate the $Q_{7,10}$ at the ungauged site using each candidate gauged site. If a baseflow segment has fewer than three candidate gauged sites the segment was discarded. This was done to avoid the chance of having only a few candidate gauged sites which might produce $Q_{7,10}$ estimators with large errors. The baseflow correlation procedure is performed using each candidate gauged site to estimate the $Q_{7,10}$ (equation 3) and the variance of the log $Q_{7,10}$ estimator at the ungauged site (equation 10). The $Q_{7,10}$ estimator that provides the minimum variance is selected to provide the best estimate of the $Q_{7,10}$ at the ungauged site for a particular baseflow segment. This process is repeated for all baseflow segments at an ungauged site.

2.4. Experimental Parameters

[16] Two groups of experimental parameters are examined: the selection of potential gauged sites and the selection of baseflow segments at ungauged sites. Potential gauged sites are selected by comparing geographic, topographic, geologic, and watershed characteristics with those at the ungauged site. Eight methods of selecting potential gauged sites based on the characteristics of the ungauged site were examined: (1) sites within the same USGS water resource region (see Figure 1); (2) sites within the same region having a drainage area within 25% of the drainage area of the ungauged site; (3) sites within 200 miles; (4) sites within 200 miles with a drainage area

within 25%; (5) sites within 200 miles with a channel slope within 25%; (6) sites within 200 miles with a soil index within 25%; (7) sites within 200 miles with a stream length within 25%; and (8) sites within 100 miles. The watershed characteristics used in methods 2, 4, 5, 6, and 7 above were obtained from the HCDN data set.

[17] Three methods for obtaining the baseflow segments at ungauged sites are examined. With all three methods, baseflow segment lengths of 5, 10, 15, and 20 days are considered. In this simulation, as the baseflow segment length increases the number of candidate gauged sites decreases due to the decreasing chance of having baseflow conditions simultaneously at both the gauged and ungauged sites for all days in the baseflow segment. The total number of baseflow segments examined using each of these methods was equal to the total number of baseflow days at the ungauged site divided by the segment length.

[18] The first method selects consecutive baseflow days at the ungauged site to build a baseflow segment. This method will be referred to as "consecutive." For example, one five-day consecutive baseflow segment exists in the streamflow record shown in Figure 2. The 5-day baseflow segment is comprised of baseflow days labeled 1, 2, 3, 4, and 5. Because the derivation of the baseflow correlation method assumes independent baseflow measurements, one would expect it to be best not to employ flows from the same baseflow recession [*Stedinger and Thomas*, 1985]. Therefore the second and third methods obtain baseflow segments in a more random manner.

[19] The second baseflow segment selection method is referred to as "random." In this method a baseflow segment is built by choosing one random baseflow day from a random starting year, and one random baseflow day from each year thereafter until the baseflow segment reaches the specified length. To illustrate the random baseflow segment selection process, assume that the streamflow record shown in Figure 2 starts on July 1 and ends on October 31 of the randomly selected starting year. The first day of the random baseflow segment would be one randomly selected baseflow from this record (either baseflow day 1, 2, 3, 4, or 5). While the random selection method employs nearly uncorrelated baseflow measurements, the length of time required (with one flow measured a year) would hinder its application in practice.

[20] The third baseflow segment selection method randomly chooses one baseflow day from consecutive baseflow recessions. This method will be referred to as "recession." In order to build a baseflow segment this method randomly picks a starting year from the ungauged data set and then picks a starting baseflow recession within that year. The baseflow segment is built by randomly choosing one day from the starting baseflow recession, and one day from each following baseflow recession until the baseflow segment reaches the specified length. Assuming the first baseflow recession shown on Figure 2 is the randomly chosen starting recession, the first day of the recession baseflow segment would be randomly chosen from baseflow day 1 or 2. The second day of the recession baseflow segment would be randomly chosen from baseflow days 3, 4, and 5. The recession selection method represents a more realistic sampling scenario, as one could

potentially gather a sufficient number of streamflow estimates over 1 to 2 low-flow seasons. This method also produces baseflow measurements that are more independent than the consecutive method.

3. Results

[21] The first portion of this section outlines the performance metrics used to evaluate the baseflow correlation method. The second sub-section examines the performance of the three baseflow segment selection methods with results aggregated over the conterminous United States. Also included in this section is an examination of a minimum acceptable correlation coefficient between the baseflows at the ungauged and gauged sites. In section 3.3, the performance of the baseflow correlation technique using the recession method of selecting a baseflow segment is examined in greater detail for USGS water resource regions 5, 10, and 17. Finally, the $Q_{7,10}$ estimators from baseflow correlation is compared with those from regional regression.

3.1. Performance Metrics

[22] After finding the gauged site that corresponds to the smallest variance of the $\ln(Q_{7,10})$ estimator (equation 10) for each of the baseflow segments, the following three performance metrics were calculated for a specific ungauged site:

Average relative absolute difference

$$\text{ARAD} = \frac{\sum\limits_{i=1}^{M}\left(\frac{|\hat{Q}_{7,10i}-Q_{7,10}|}{Q_{7,10}}\right)}{M} \quad (12)$$

Relative bias

$$R-BIAS = \frac{\sum\limits_{i=1}^{M}\left(\frac{\hat{Q}_{7,10i}-Q_{7,10}}{Q_{7,10}}\right)}{M} \quad (13)$$

Relative mean square error

$$R-MSE = \frac{\sum\limits_{i=1}^{M}\left(\frac{\hat{Q}_{7,10i}-Q_{7,10}}{Q_{7,10}}\right)^2}{M} \quad (14)$$

where $\hat{Q}_{7,10i}$ is the ith estimate of the $Q_{7,10}$ at the ungauged site, $Q_{7,10}$ is the "true value" of the $Q_{7,10}$ at the ungauged site, and M is the number of baseflow segments with an associated $\hat{Q}_{7,10i}$. $Q_{7,10}$ is the at-site estimator obtained using the entire historic record at the site, fitting a log-Pearson type III distribution by method of moments [*Stedinger et al.*, 1993], and estimating the tenth percentile of the distribution. The ARAD performance metric is primarily used to discuss the results of this experiment. The ARAD measures of the average percent deviation of the $Q_{7,10}$ estimator, and thus is easily interpretable. For example, an ARAD of 0.05 indicates a 5% error on average, and an ARAD of 1.00 indicates a 100% error on average. R-BIAS and R-MSE are also included to further assess method performance.

[23] Each HCDN river site is sequentially considered as the ungauged river site in this experiment. The ARAD, R-BIAS, and R-MSE are calculated for each ungauged site. The results for the ungauged sites in a group of G sites, such

as those within the same USGS water resource region or all sites in the conterminous United States, are summarized by finding the average of each performance measure, weighted by the record length (in years) of the site:

$$\text{Average ARAD} = \frac{\sum\limits_{i=1}^{G}\text{ARAD}_i * \text{Record Length}_i}{\sum\limits_{i=1}^{G}\text{Record Length}_i} \quad (15)$$

Average R-BIAS and Average R-MSE were calculated similarly. Thus sites with longer records, which typically have at-site $Q_{7,10}$ estimators with a smaller variance, will receive a greater weight. Average ARAD, Average R-BIAS, and Average R-MSE are used to compare the performance of the baseflow correlation method using different experimental parameters, though in general results are discussed in terms of ARAD.

3.2. Average Results for the Continental United States

[24] Initial results examined the previously described eight methods of selecting potential gauged sites based on physical characteristics and location of the ungauged site. All of the methods where potential gauged sites are within 200 miles of the ungauged site (methods 3 through 7) produced similar results. The two methods where potential gauged sites are within the same USGS water resource region as the ungauged site (methods 1 and 2) also had similar performance. One would expect gauged sites with similar watershed characteristics to provide better estimators. The observed results may be due to data limitation. For instance, restrictions were placed on the minimum number of candidate sites to accept a baseflow segment. When fewer than three candidate sites are available the baseflow segment was discarded, thus reducing the sample size. In addition there was also a restriction on the distance between the gauged and ungauged sites; in some portions of the United States the HCDN sites are sparse with few gauged sites within 100 or 200 km. Since the results were similar, potential gauged site selection methods 1 (sites within the same USGS water resource region), 3 (sites within 200 km), and 8 (sites within 100 km), are the only methods of selecting potential gauged sites for which results are presented. For each of these gauged site selection methods, the three methods of compiling a baseflow segment were examined.

[25] *Stedinger and Thomas* [1985] suggested a minimum correlation coefficient of 0.7 between the baseflow measurements at the gauged and ungauged river sites. This means for each baseflow segment at an ungauged site, the correlation between the flows at the gauged and ungauged site must be greater than 0.7 or that estimate using the gauged site is discarded from the results. The initial results presented here employ this restriction. Later in this sub-section the impact of varying this restriction is examined. Figures 3a, 3b, and 3c present the average ARAD, R-BIAS, and root R-MSE (RR-MSE) for all sites in the conterminous United States when the baseflow segments are comprised of consecutive days, randomly selected days, and days from consecutive recessions, respectively. The RR-MSE is presented instead of the R-MSE so that the units of all performance metrics are the same. The x axis displays the criteria for selecting potential gauged sites: those within the
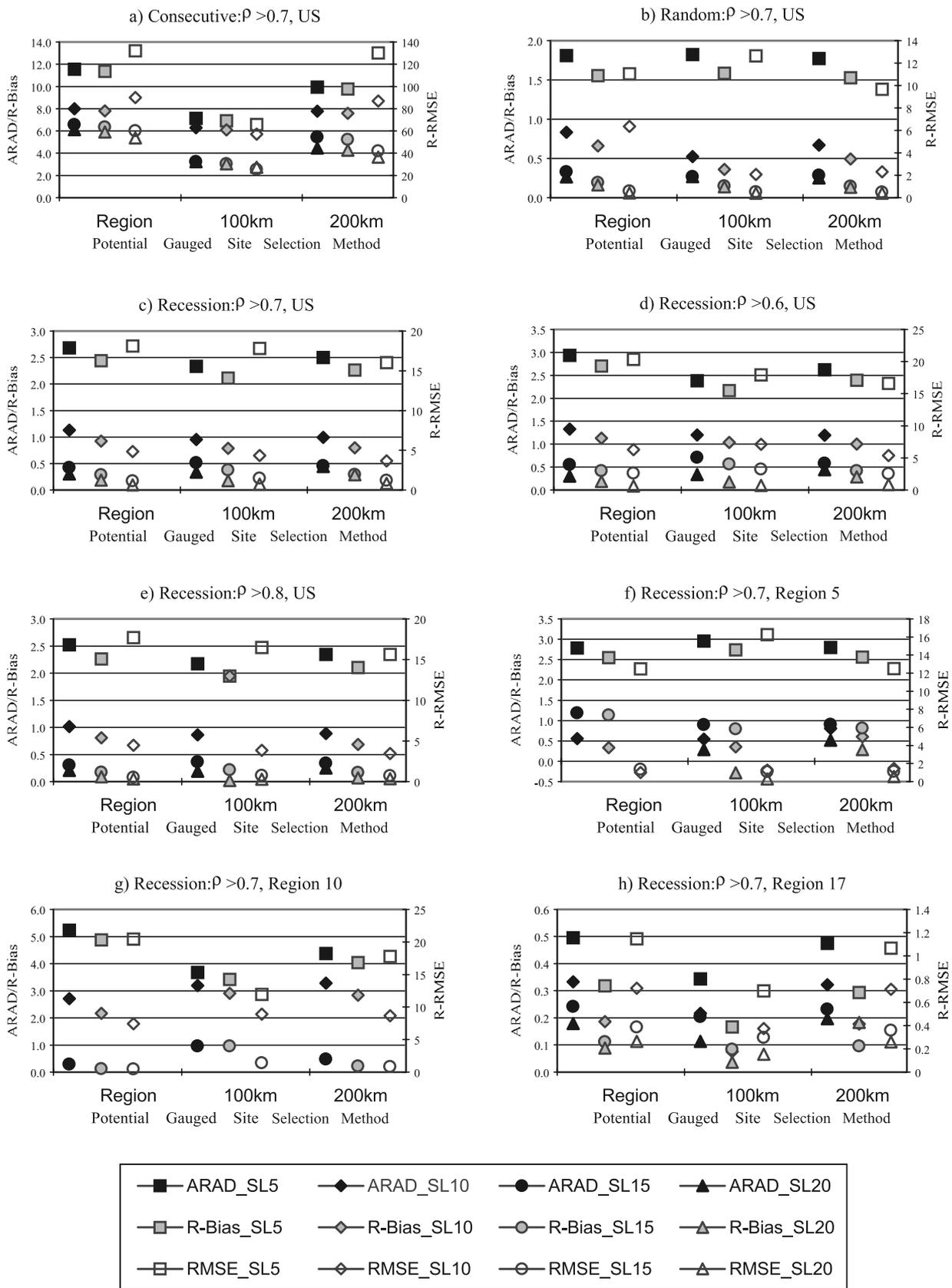
Figure 3. Average relative absolute deviation (ARAD), relative bias (R-BIAS), and relative root mean square error (R-RMSE) for (a) consecutive baseflows with $\rho > 0.7$, (b) random baseflows with $\rho > 0.7$, (c) recession baseflows with $\rho > 0.7$, (d) recession baseflows with $\rho > 0.6$, (e) recession baseflows with $\rho > 0.8$, (f) recession baseflows with $\rho > 0.7$ in region 5, (g) recession baseflows with $\rho > 0.7$ in region 10, and (h) recession baseflows with $\rho > 0.7$ in region 17.

same region as the ungauged site (region), those within 100 km of the ungauged site (100 km), and those within 200 km of the ungauged site (200 km). There are twelve points plotted within each potential gauged site selection method. Of these points, the square represents a baseflow segment length of 5, the diamond a segment length of 10, the circle a segment length of 15, and the triangle a segment length of 20 days. In addition, the ARAD results are represented by solid-filled symbols, the R-BIAS with gray-filled symbols, and the RR-MSE with unfilled symbols. The left y axis presents the scale for ARAD and R-BIAS while the right y axis presents the scale for RR-MSE. Note that the scale of the y axis is different for each of these plots due to the different range of performance metric for each of the baseflow segment selection methods.

[26] From these plots it is evident that over the conterminous United States the random and recession methods for compiling a baseflow segment perform much better than the consecutive method, thus verifying the need for nearly independent baseflow measurements. Longer baseflow segments improve performance; however, this improvement typically levels off at segment lengths of 15 to 20 days. For the random and recession methods of compiling a baseflow segment, the criteria for selecting potential gauged sites do not have a large impact on performance. Selecting potential gauged sites within 100 km of the ungauged sites performs slightly better than the other methods for baseflow segments comprised of random days.

[27] The best results presented in this section are for the random method with baseflow segments of 15 days or 20 days and gauged sites within 100 km, with an Average ARAD of 0.27 corresponding to an average error of 27%. In practice one would most likely employ a baseflow sampling method more similar to the recession method, since the random method would require an inordinate amount of time to obtain baseflow measurements. The best performance for the recession method was with a 20 day baseflow segment length and sites within the same water resource region. This result had an Average ARAD of 0.30. For a 15 day baseflow segment length the performance is reduced, with an Average ARAD of 0.43.

[28] Results for R-BIAS and RR-MSE followed a similar pattern to those for ARAD. One important note is that the R-BIAS is almost always positive and nearly as large as the ARAD, suggesting that there is a systematic upward bias to the baseflow correlation estimators (i.e., very few sites have a negative bias). This bias appears most pronounced at sites with smaller at-site $Q_{7,10}$ estimates. At these sites the relative performance measures employed in this study tend to be inflated due to dividing by a small $Q_{7,10}$.

[29] Since in practice one would employ a sampling method similar to the recession method, only results for the recession method are presented in the rest of this paper. In Figures 3a, 3b, and 3c the minimum correlation coefficient between the baseflows at the gauged and ungauged sites was 0.7. Figures 3d and 3e present the results for the recession method when the correlation coefficient restriction is 0.6, and 0.8, respectively, and can be compared to Figure 3c (ρ > 0.7). When the correlation coefficient restriction is 0.8, the performance improves to a minimum Average ARAD of 0.19 for a segment length of

20 days and gauged sites within 100 km. For a correlation coefficient restriction of 0.6, the results for 20 day segment lengths are similar to those for 0.7, but for a segment length of 15 days the correlation coefficient restriction of 0.7 produces noticeably improved results (ARAD of 0.42 versus ARAD of 0.55). Again the results for R-BIAS and RR-MSE follow a similar trend as those for ARAD.

3.3. USGS Water Resource Regions 5, 10, and 17

[30] The results presented above were for the entire conterminous United States. Because some of the USGS water resource regions performed well and others performed poorly, this section examines three USGS water resource regions (see Figure 1) in greater detail to compare regions that perform well with those that have poor performance. The water resource regions are: region 5, which had moderate performance as compared with the other regions; region 10, a poor performer; and region 17, which had relatively good performance. This section only considers baseflow segments comprised of days from the recession method with a correlation coefficient between concurrent baseflows greater than 0.7.

[31] Region 5 is located between the mid-Atlantic and mid-west regions of the United States and includes western Pennsylvania, West Virginia, Kentucky, Indiana, the eastern edge of Illinois, and most of Ohio. Region 10 is located in the northern mid-west region of the United States and includes most of Montana, the western half of North Dakota, most of Wyoming and South Dakota, the northeast corner of Colorado, the north half of Nebraska, the western edge of Iowa, Kansas, and most of Missouri. Region 17 is located in the northwestern United States and includes Washington, most of Idaho, the western edge of Montana, and most of Oregon.

[32] Figures 3f, 3g, and 3g show the Average ARAD, R-BIAS, and RR-MSE for regions 5, 10, and 17, respectively. These plots have a similar format as Figures 3a through 3e described above, including a different scale for the y axis of each plot. If a point is missing on these plots, then there were no baseflow segments at any sites in the region where there were three or more candidate gauged sites, and thus the baseflow correlation technique was never applied. For regions 10 and 17 longer baseflow segments improve the performance of the baseflow correlation method. For region 5 the performance improves between segment lengths of 5 and 10 days and 15 and 20 days, but declines between segment lengths of 10 and 15 days. The best performance for regions 5 and 17 is when potential gauged sites are located within 100 km of the ungauged site. In region 10 the best performance is when potential gauged sites are located within the same region as the ungauged site. The best Average ARAD over these three regions, with a value of 0.11, is in region 17 when potential gauged sites are within 100 km of the ungauged site and the baseflow segment length is 20 days. As with the previous results, trends based on R-Bias and RR-MSE were similar to those based on ARAD. Again R-Bias results indicated a systematic upward bias of the baseflow correlation method.

[33] One potential reason for the poor performance of region 10 may be due to the large area of this region, thus potentially increasing the heterogeneity across watersheds. In addition, Region 10 has a relatively small density of sites
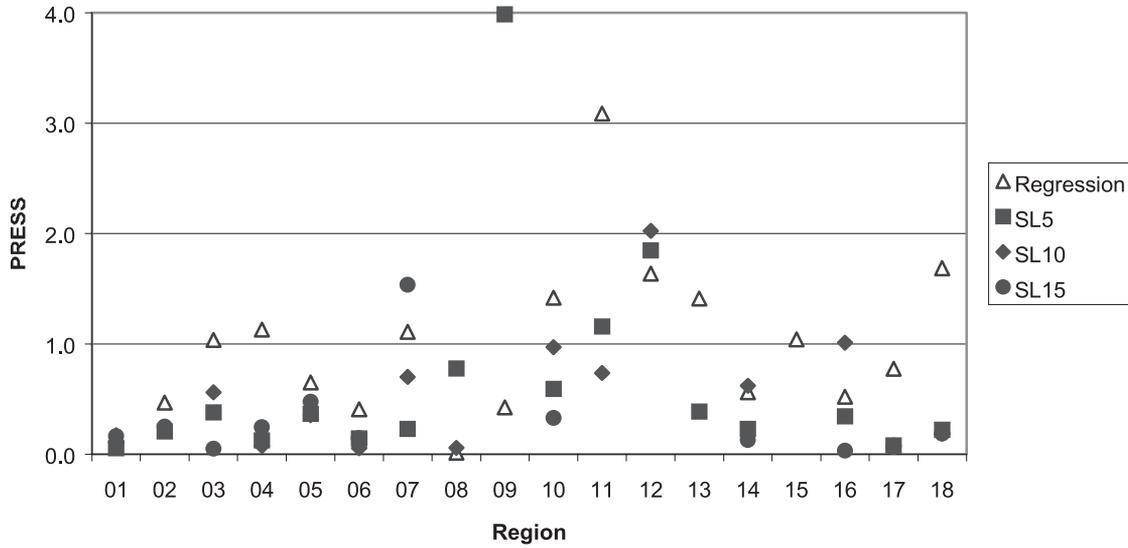
Figure 4. PRESS statistic for regional regression and baseflow correlation with baseflow segments of length 5, 10, and 15, $\rho > 0.7$, and gauged sites within 200 km.

(see Figure 1) compared with other regions. Thus there are fewer potential gauged sites available for each ungauged site. With fewer potential gauged sites and less homogeneity among the sites, one would expect a decrease in performance. Another reason for the poor performance may be the large number of sites in region 10 with small at-site $Q_{7,10}$ estimates, which may impact the "relative" performance metrics of equation 12–14.

3.4. Comparison With Regional Regression

[34] In order to compare the baseflow correlation technique with regional regression, a prediction error sum of squares (PRESS) statistic was calculated. The PRESS statistic is a validation-type estimator of error commonly employed in regression analyses [*Helsel and Hirsch*, 1992], and is applied here as:

$$\text{PRESS} = \frac{\sum_{i=1}^{G} \left[\left(\ln(\hat{Q}_i) - \ln(Q_i) \right)^2 \right]}{G - 1} \qquad (16)$$

where $\hat{Q}_i$ is the estimate of the $Q_{7,10}$ at the ith ungauged site, Q_i is the "true value" of the $Q_{7,10}$ at the ith ungauged site obtained using the entire historic record at the site, and G is the number of stream gauge sites in a region. In regression analyses, the PRESS statistic is obtained by sequentially removing one site at a time from a region, fitting the regression model using the other sites in the region, and then predicting the $Q_{7,10}$ at the site that has been removed. The regression results presented here are based on the work of *Kroll et al.* [2003]. In the Kroll et al. study low-flow regional regression models were fit for various regions across the United States by employing a new database of digitally derived watershed characteristics. This database was employed here with regression models fit for each USGS water resource region using OLS regression techniques. For the baseflow correlation method, $\ln(\hat{Q}_i)$ was calculated by averaging the logarithms of the estimated $Q_{7,10}$ across all baseflow segments at an ungauged site.

[35] Figure 4 presents the PRESS statistic for regional regression (Regression) and for baseflow correlation with segment lengths of 5 (SL5), 10 (SL10), and 15 (SL15). The baseflow correlation results shown are for gauged sites within 200 km, with a correlation coefficient between baseflows of at least 0.7, and baseflows obtained using the recession method (consecutive recessions). In 15 of the 18 regions the baseflow correlation technique performs better than regional regression. In regions 8, 9, and 12 regional regression outperforms baseflow correlation. These regions tend to be where baseflow correlation performs the worst, in the midwestern and southern United States.

[36] It should be noted the comparison presented in this section might not be particularly fair to the regression method. While the baseflow correlation method employs only sites within 200 km of the ungauged site, the regressions were developed using sites within an entire USGS water resource region. Often these regions were quite large and did not contain homogeneity amongst the sites. A nearest neighbor regression analysis [*Tasker*, 1996] might provide more comparative results with the baseflow correlation method, but by limiting the analysis to HCDN sites, relatively few sites would be available for the nearest neighbor regression analysis and thus it was not performed. In addition, some sites were not employed in the baseflow correlation method because there were never three or more gauged sites with baseflows on the same days as the ungauged site. These sites, though, were included in the regression analysis, and may be responsible for the poor performance of this technique. Regardless, the baseflow correlation method does appear to provide good results in many regions of the United States.

4. Conclusions

[37] The goal of this experiment was to further examine the baseflow correlation method for estimating low-streamflow statistics at ungauged river sites. The baseflow correlation method requires a nominal number of measured streamflows during baseflow conditions at the ungauged

site of interest, and a nearby gauged site with baseflow conditions on the same days. Three potential techniques were examined for measuring streamflow during baseflow conditions: (1) consecutive baseflow days (consecutive), (2) a random baseflow day from consecutive years (random), and (3) a random baseflow day from consecutive baseflow recessions (recession). Numerous methods were also used to choose potential gauged sites based on geographic location, water resource region, and the physical characteristics of the ungauged site. In addition, the impact of varying the number of streamflow measurements from the ungauged site was examined, as well as the requirement for the correlation coefficient between baseflow measurements at the gauged and ungauged river site to be above a specific level. On the basis of this experiment, the following conclusions were reached.

[38] 1. The method of selecting days for streamflow measurements greatly impacts the performance of the baseflow correlation method. Having nearly independent baseflow measurements (random) provided the best results, while choosing consecutive baseflow measurements (consecutive) provide much poorer results than the other 2 methods.

[39] 2. Choosing random baseflows from consecutive recessions (recession) provided a slight reduction in performance, but is a much more realistic sampling procedure than the "random" method (which would require many years of measurements), and thus is recommended in practice.

[40] 3. In general, performance of the baseflow correlation method improved as the number of baseflow measurements increased, but some leveling off of performance was observed between 15 and 20 measurements.

[41] 4. As expected, the higher the correlation coefficient between the baseflow measurements at the gauged and ungauged sites, the better the performance of the baseflow correlation method. Restricting the correlation coefficient to be above 0.8 produced a large increase in method performance, but such a restriction greatly reduces the number of potential gauged sites. In general, 0.7 appears to be a more reasonable selection.

[42] 5. The use of watershed characteristics at the ungauged site to select potential gauged river sites produced no increase in performance. This may be due to the limited number of sites examined in this analysis.

[43] 6. While choosing gauged sites within 100 km of the ungauged site generally improves performance, only a slight decrease in performance was observed when gauged sites within 200 km were employed.

[44] 7. Results were generally the same when any one of three performance metrics were examined: average relative absolute difference (ARAD), relative bias (R-BIAS), and root relative mean square error (RR-MSE). The R-BIAS metric indicated a systematic upward bias of the baseflow correlation method, though this result appears to be driven by sites with smaller at-site $Q_{7,10}$ estimates where the relative performance measures employed in this study can be inflated.

[45] 8. A brief comparison of baseflow correlation with regional regression models indicated the baseflow correlation method is preferred in nearly all of the United States, especially in the northeastern and northwestern United States. In general, baseflow correlation performed worse in the midwestern and southern United States.

[46] On the basis of this analysis, baseflow correlation appears to be an excellent method for estimating low-streamflow statistics at ungauged river sites. Results indicate this method should involve (1) selecting one baseflow measurement from consecutive streamflow recessions, where baseflows are indicated by at least a 6 day drop in discharge; (2) measuring at least 10 and preferably 15 discharges at the ungauged river site; (3) choosing gauged sites from within 200 km of the ungauged site; and (4) having a correlation coefficient between the measured streamflows at the gauged and ungauged sites of at least 0.7. It should be noted that in practice one would expect the baseflow correlation method to perform better than the results presented here. This is because ungauged sites will typically be correlated to gauged sites that are closer to each other than the average distance between the HCDN sites employed in this study. In addition, similarity of watershed characteristics (climate, geology, etc.) will generally be considered in selecting gauged sites. Further investigation of the baseflow correlation method is warranted, especially in arid regions where low-streamflow analyses are further confounded by intermittent streamflow conditions.

References

Arihood, L. D., and D. R. Glatfelter, Method for estimating low-flow characteristics of ungaged streams in Indiana, *U.S. Geol. Surv. Open File*, 86-323, 1986.

Barnes, C. R., Method for estimating low-flow statistics for ungaged streams in the lower Hudson River basin, *U.S. Geol. Surv. Water Resour. Invest. Rep.*, 85-4070, 1985.

Bras, R. L., *Hydrology: An Introduction to Hydrologic Science*, Addison-Wesley-Longman, Reading, Mass., 1990.

Douglas, E. M., R. M. Vogel, and C. N. Kroll, Trends in floods and low flows in the United States: Impact of spatial correlation, *J. Hydrol.*, 240(1–2), 90–105, 2000.

Douglas, E. M., R. M. Vogel, and C. N. Kroll, Impact of streamflow persistence on hydrologic design, *J. Hydrol. Eng.*, 7(3), 220–227, 2002.

Draper, N. R., and H. Smith, *Applied Regression Analysis*, John Wiley, Hoboken, N.J., 1966.

Hammett, K. M., Low–flow frequency analysis for streams in west central Florida, *U.S. Geol. Surv. Water Resour. Invest. Rep.*, 84-4299, 1985.

Helsel, D. R., and R. M. Hirsch, *Statistical Methods in Water Resources*, Elsevier Sci., New York, 1992.

Kroll, C., J. Luz, B. Allen, and R. M. Vogel, Developing a watershed characteristics database to improve low streamflow prediction, *J. Hydrol. Eng*, in press, 2003.

Potter, K. W., A simple method for estimating baseflow at ungauged stations, *J. Am. Water Resour. Assoc.*, 37(1), 177–184, 2001.

Ries, K. G., Development and application of generalized-least-squares regression models to estimate low-flow duration discharges in Massachusetts, *U.S. Geol. Surv. Water Resour. Invest. Rep.*, 94-4155, 1994.

Riggs, H. C., Low flow investigations, *U.S. Geol. Surv. Tech. Water Resour. Invest.*, 4(B1), 1972.

Riggs, H. C., Characteristics of low flows, *J. Hydraul. Eng.*, 106(5), 717–731, 1980.

Rumenick, R. P., and J. W. Grubbs, Methods for estimating low-flow characteristics for ungaged streams in selected areas, northern Florida, *U.S. Geol. Surv. Water Resour. Invest. Rep.*, 96-4124, 28 pp., 1996.

Slack, J. R., and J. M. Landwehr, Hydro-Climatic Data Network: A U.S. Geological Survey streamflow data set for the United States for the study of climate variations, 1874–1988, *U.S. Geol. Surv. Open File Rep., 92-129*, 1992. (Available at http://water.usgs.gov/pubs/wri/wri934076/1st_page.html)

Smakhtin, V. U., Low flow hydrology: A review, *J. Hydrol.*, *240*, 147–186, 2001.

Stedinger, J. R., and W. O. Thomas Jr., Low-flow frequency estimation using base-flow measurements, *U.S. Geol. Surv. Open File Rep.*, *85–95*, 1985.

Stedinger, J. R., R. M. Vogel, and E. Foufoula-Georgiou, Frequency analysis of extreme events, in *Handbook of Hydrology*, edited by D. R. Maidment, McGraw-Hill, New York, 1993.

Tasker, G. D., Region of influence regression for estimating the 50-year flood at ungauged sites, *J. Am. Water Resour. Assoc.*, *32*(1), 163–170, 1996.

Thomas, D. M., and M. A. Benson, Generalization of streamflow characteristics from drainage-basin characteristics, *U.S. Geol. Surv. Water Supply Pap.*, *1975*, 1970.

Vogel, R. M., and C. N. Kroll, Regional geohydrologic-geomorphic relationships for the estimation of low-flow statistics, *Water Resour. Res.*, *28*(9), 2451–2458, 1992.

Vogel, R. M., C. J. Bell, and N. M. Fennessey, Climate, streamflow and water supply in the northeastern United States, *J. Hydrol.*, *198*(1–4), 42–68, 1997.

Wandle, S. W., Jr., and A. D. Randall, Effects of surficial geology, lakes and swamps, and annual water availability on low flows of streams in central New England, and their use in low-flow estimation, *U.S. Geol. Surv. Water Resour. Invest. Rep.*, *93-4092*, 57 pp., 1993.

―――――――――――――

C. N. Kroll, Department of Environmental Resource Engineering, State University of New York College of Environmental Science and Forestry, 312 Bray Hall, 1 Forestry Drive, Syracuse, NY 13210, USA. (cnkroll@mailbox.syr.edu)

C. F. Reilly, Department of Computer Sciences, University of Wisconsin-Madison, Madison, WI 53706, USA.