Climate-Biased Decisions Via Partial Historical Sampling

Thomas E. Croley II*

*Research Hydrologist, Great Lakes Environmental Research Laboratory, National Oceanic and Atmospheric Administration, US Department of Commerce, 2205 Commonwealth Blvd., Ann Arbor, Michigan, 48105-2945; PH (734) 741-2238; FAX (734) 741-2055; email: croley@glerl.noaa.gov

Abstract

A heuristic approach for incorporating probabilistic meteorology outlooks, via operational hydrology, into derivative hydrology probability forecasts or storm frequency distributions is described. It constructs a weighted set of future possibilities that agree with selected meteorology outlooks. Many times, calculated weights are zero-valued and a question arises on how to properly consider them in the biased sample. After exploring the effects of directly using zero-valued weights, an alternative is presented that omits historical observations from the biased sample corresponding to zero-valued weights (partial historical sampling). This requires adjustment of the non-zero weights and redefinition of hydrology forecast statistics that are based on the biased sample. Examples of simple storm frequency estimation, using El Niño conditional probabilities, illustrate the problem with zero-valued weights for some estimators and their negligible effect with other estimators.

Introduction

Hydrology probabilities can be forecast *indirectly* from past historical records of meteorology with watershed (and other) models via operational hydrology approaches (Croley 1996, 1997, 2000a). Future storm frequencies can be estimated *directly* from past historical records of sufficient length (Croley 2001). Both approaches build a set of “possibilities” for the future, to be treated as a “sample” from which to estimate various statistics: hydrology outlook probabilities, storm frequencies, or other parameters. These estimates ignore changing climate. Now, numerous probabilistic meteorology forecasts of the changing climate are available to water resource engineers and hydrologists for use in the estimation of derivative hydrology probability forecasts or storm frequencies. The historical sample may be biased to match forecast meteorological probability statements by “weighting” it appropriately (Croley 2000a). Basically, those groups of meteorology segments (from the historical record) matching probabilistic meteorology forecasts are given more weight than those not matching. Boundary condition equations for the weights are constructed corresponding to probabilistic meteorology forecasts and solved for physically relevant weights. The solution may involve an optimization when there is more than one set of weights possible. Croley (2000a) describes the formulation of objectives to be used in such optimizations and the appropriate solution methodology.

Origin of the Weighted Sample

Consider building an arbitrarily large biased sample (size N) with repeated observations from the original sample (size n) to force certain events to constitute a larger or smaller portion of the sample (Croley 2000a). Figure 1 illustrates this for an example that forces the relative frequency of September

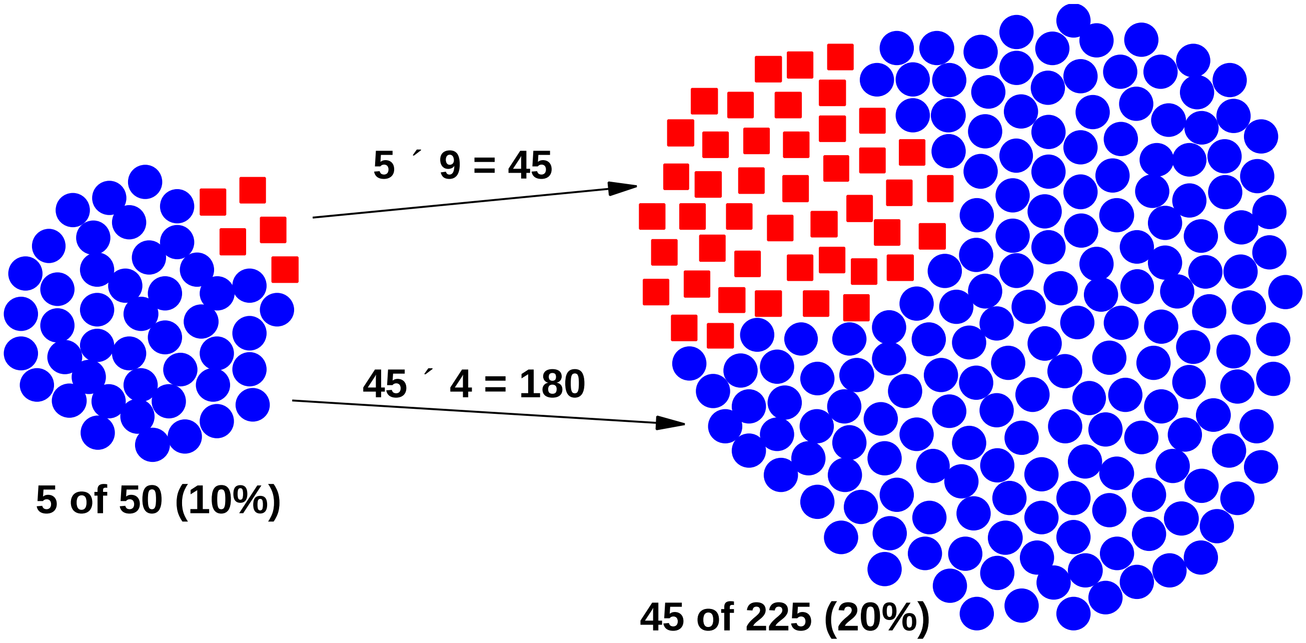

Figure 1. Building a biased sample. For example, each square could represent a scenario in which September Air Temperature > 7°C and each circle September Air Temperature $\leq$ 7°C (Croley 2000).

air temperatures greater than 7°C to be 20%. Each scenario, ω_i, ($i = 1, 2, \ldots, n$) is duplicated J_i times. By judiciously choosing these duplication numbers ($J_1, J_2, \ldots, J_n$), we can force the relative frequency of any group of scenarios in the structured sample to any desired value. Note also that

$$\sum_{i=1}^{n} J_i = N \tag{1}$$

One can estimate the probability of event A, $P[\mathrm{A}]$, by its relative frequency, $\hat{P}[\mathrm{A}]$, defined as the number of observations in the sample for which A occurs (i.e., for which event A is true), N_{A} divided by the total number of observations, N:

$$\begin{aligned}\hat{P}[\mathrm{A}] &= \frac{N_{\mathrm{A}}}{N} = \frac{1}{N}\sum_{k\,|\,\mathrm{A}} 1 = \frac{1}{N}\sum_{k=1}^{N} I\left(w'_k \in \mathrm{A}\right) \\ &= \frac{1}{N}\sum_{i=1}^{n} J_i\, I\left(w_i \in \mathrm{A}\right) = \frac{1}{N}\sum_{i\,|\,\mathrm{A}} J_i = \frac{1}{n}\sum_{i\,|\,\mathrm{A}} w_i\end{aligned} \tag{2}$$

where the sum is taken over all k (members of the very large sample) for which A occurs, denoted as $k|\mathrm{A}$, and where ω_i is scenario i from the original sample, ω'_k is scenario k from the large sample, $I(\omega \in \mathrm{A})$ (the indicator function) is unity if ω is included in event A and zero if not, and

$$w_i = \frac{n}{N} J_i \tag{3}$$

Note that (1) and (3) guarantee that all weights sum to the original sample size, n. Likewise, other statistics can be estimated from the very large sample or, equivalently, from the original sample by applying weights. Let z_k, $k = 1, \ldots, p$ denote a set of statistics with the following form:

$$z_k = f_k(n) \sum_{i=1}^{n} g_k\left(x_i, z_{k-1}, \ldots, z_1\right), \qquad k = 1, \ldots, p \tag{4}$$

where f_k is an arbitrary function of n, g_k is an arbitrary function of several sample values and lower-order statistics, and x_i are sample values of random variable X in the original sample of size n. Define an analogous set of statistics, $z_k^{\bullet}$, $k = 1, \ldots, p$ over the very large sample of size N and derive a weighted set of statistics similar to (4):

$$\begin{aligned} z_k^{\bullet} &= \frac{f_k(n)\, n}{N} \sum_{i=1}^{N} g_k\left(x_i', z_{k-1}^{\bullet}, \ldots, z_1^{\bullet}\right) \\ &= \frac{f_k(n)\, n}{N} \sum_{i=1}^{n} \vartheta_i \; g_k\left(x_i, z_{k-1}^{\bullet}, \ldots, z_1^{\bullet}\right) \\ &= f_k(n) \sum_{i=1}^{n} w_i \; g_k\left(x_i, z_{k-1}^{\bullet}, \ldots, z_1^{\bullet}\right), \qquad k = 1, \ldots, p \end{aligned} \tag{5}$$

where x_i' are sample values of random variable X in the large sample of size N. Generally, almost any practical statistic can be written in the form of (4), and therefore has a weighted counterpart in the form of (5) which is derivable by using the “large-sample” reasoning employed in (2) and (5) and illustrated in Figure 1. For example, the following set of statistics is described by (4):

$$\begin{aligned} \overline{x} &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ s_X^2 &= \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^2 \\ \hat{\psi}_X &= \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(x_i - \overline{x}\right)^3 \Big/ \left(\sqrt{s_X^2}\right)^3 \\ \hat{P}\left[X > x_{i(\ell)}\right] &= \frac{\ell}{n+1} = \frac{1}{n+1} \sum_{k=1}^{\ell} 1, \qquad \ell = 1, \ldots, n \end{aligned} \tag{6}$$

where $\overline{x}$ = sample mean of X, s_X^2 = sample variance of X, $\hat{\psi}_X$ = sample skew coefficient of X, and and $i(\ell)$ is the number of the value in the unordered sample corresponding to the $\ell^{\underline{\mathrm{th}}}$ order (largest value is order 1). The last statistic in (6) is the Weibull estimator of exceedance probabilities. The corresponding set of statistics defined by (5) is:

$$\begin{aligned}
\overline{x}^{\bullet} &= \frac{1}{n}\sum_{i=1}^{n} w_i \; x_i \\
\left(s_X^2\right)^{\bullet} &= \frac{1}{n-1}\sum_{i=1}^{n} w_i \left(x_i - \overline{x}^{\bullet}\right)^2 \\
\hat{\psi}_X^{\bullet} &= \frac{n}{(n-1)(n-2)}\sum_{i=1}^{n} w_i \left(x_i - \overline{x}^{\bullet}\right)^3 \Big/ \left(\sqrt{\left(s_X^2\right)^{\bullet}}\right)^3 \\
\hat{P}^{\bullet}\left[X > x_{i(\ell)}\right] &= \frac{1}{n+1}\sum_{k=1}^{\ell} w_{i(k)}, \qquad \ell = 1, \ldots, n
\end{aligned} \tag{7}$$

Note that the weighted estimators of (5) are equivalent to (4) if all weights are unity. This is intuitively appealing. It guarantees that weighted statistics degenerate to unweighted statistics when there are no weights (i.e., when weights are unity). However other properties of (4) and (5) may not be similar. For example, consider the sample variance in (6) and (7). Statisticians recognize the former as an unbiased estimator of the population variance from a sample of size n while the latter results from the application of the following to the very large sample of size N, as in (5):

$$\left(s_X^2\right)^{\bullet} = \frac{1}{n-1}\frac{n}{N}\sum_{i=1}^{N}\left(x_i' - \overline{x}^{\bullet}\right)^2 \tag{8}$$

which would be a biased estimate of the population variance from a sample of size N if the sample was a true random sample (had not been specifically constructed in the manner described here). However, whenever $f_k(n) = 1/n$ then statistic properties in (4) and (5) will be similar. This is because (4) and (5) become identical (except for sample size):

$$\begin{aligned}
z_k &= \frac{1}{n}\sum_{i=1}^{n} g_k\left(x_i, z_{k-1}, \ldots, z_1\right), \qquad k = 1, \ldots, p \\
z_k^{\bullet} &= \frac{1}{N}\sum_{i=1}^{N} g_k\left(x_i', z_{k-1}^{\bullet}, \ldots, z_1^{\bullet}\right), \qquad k = 1, \ldots, p
\end{aligned} \tag{9}$$

Zero Weights

It is observed that weighting solutions often involve zero-valued weights. Thus, the corresponding historical observations, weighted by these zero values, are not represented in the “sample” from which derivative forecasts are made. See, for example, Figure 2. There are multiple manners in which to consider zero-valued weights. One way is to simply use the zero weights directly in the weighted statistics as if the corresponding observations (scenarios) are still in the sample. Unfortunately, this can result in strange behavior in some of the statistics. For example, the Weibull estimator of exceedance probabilities [last line in (7)] will assign the same exceedance probability estimate to more than one sample value of *X*. Successive exceedance probability estimates differ by the weight assigned to the smaller *X* value divided by (n + 1) and, if the weight is zero-valued, the successive estimates are identical. This is undesirable if, for example, these successive estimates are later used to linearly interpo-

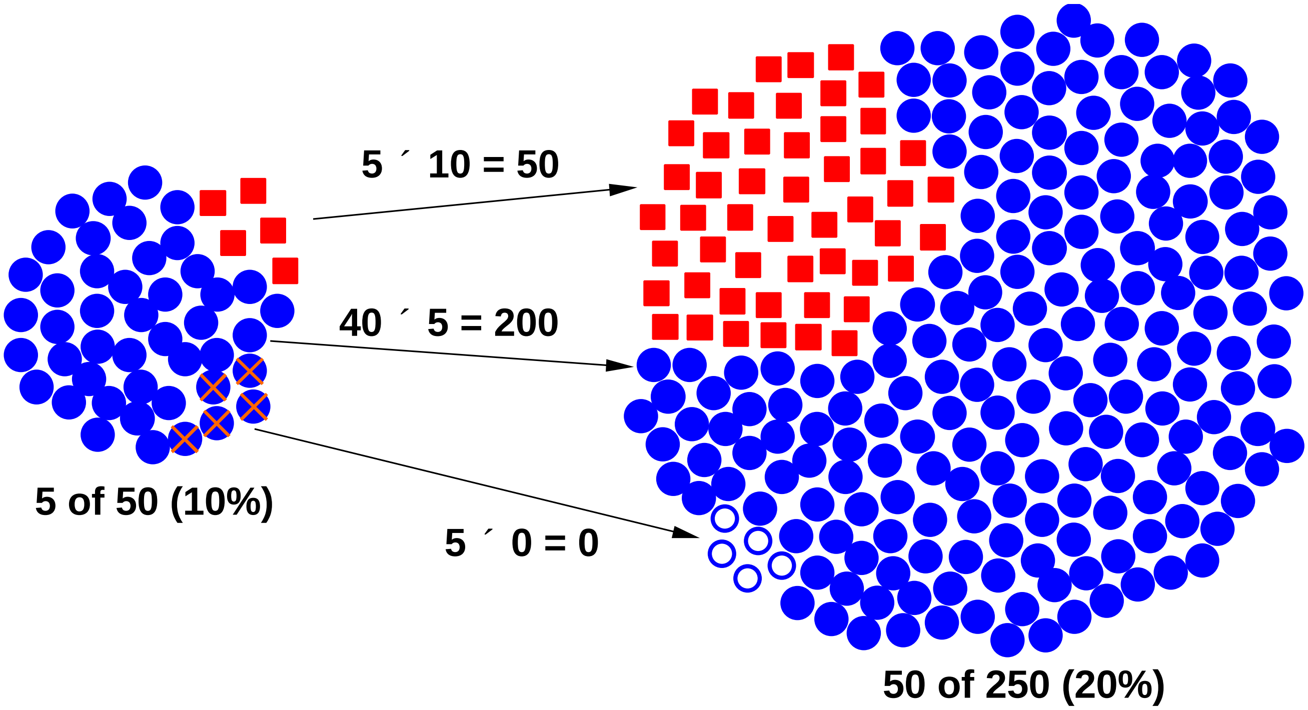

Figure 2. Building a biased sample with some zero-valued "weights."

late for a value of X from an exceedance probability; the many (infinite) possible values of X make it indeterminate. This will be illustrated shortly in the Examples section.

An alternate method of considering zero-valued weights involves a reformulation of the sample. Consider a weighted sample of n observations, where only d weights are non-zero ($d < n$), as a weighted sample of d observations where all weights are non-zero (i.e., eliminate the zero-weighted observations). The discussion of the preceding section concerned non-zero-valued weights, but a slight adjustment can be used to include zero-valued weights. Suppose that for the set of statistics defined in (4), some of the weights are zeroes. For the sake of notation, suppose that the last $n-d$ weights are zeroes. The unweighted statistics over the smaller sample are:

$$z_k \quad = \quad f_k(d)\sum_{i=1}^{d} g_k\left(x_i, z_{k-1}, \ldots, z_1\right), \qquad k \quad = \quad 1, \ldots, p \tag{10}$$

The corresponding weighted statistics then become:

$$z_k^{\bullet} \quad = \quad f_k(d)\sum_{i=1}^{d} w_i \; g_k\left(x_i, z_{k-1}^{\bullet}, \ldots, z_1^{\bullet}\right), \qquad k \quad = \quad 1, \ldots, p \tag{11}$$

That is, the sample size is effectively reduced by the presence of zero-valued weights. This set of statistics still has the desired properties established above; the set of statistics in (11) approaches the unweighted version in (10) as the weights go to unity.

Recall that even though there may be zero-valued weights in a set of *n* weights, their sum is equal to *n*. For all the weights to sum to the (new) sample size, the *d* non-zero-valued weights must be recomputed relative to the new sample size (*d*).

$$w_i' = \frac{d}{n} w_i, \qquad i = 1, \ldots, d \tag{12}$$

where w_i' = recomputed (or adjusted) weight. The adjusted weights now sum to the new sample size,

$$\sum_{i=1}^{d} w_i' = \sum_{i=1}^{d} \frac{d}{n} w_i = \frac{d}{n} \sum_{i=1}^{n} w_i = d \tag{13}$$

Likewise, the adjusted weights still satisfy (2), now for the smaller sample size:

$$\hat{P}[\mathrm{A}] = \frac{1}{n} \sum_{i|\mathrm{A}} w_i = \frac{1}{n} \sum_{i|\mathrm{A}} \frac{n}{d} w_i' = \frac{1}{d} \sum_{i|\mathrm{A}} w_i' \tag{14}$$

and still satisfy an equation like (5) with some substitutions for the arbitrary leading constant term, now for the smaller sample size, to guarantee that the statistics revert to their unweighted form in (10):

$$\begin{aligned} z_k^\bullet &= \frac{f_k(d)\,d}{N} \sum_{i=1}^{N} g_k\left(x_i', y_i', \ldots, z_{k-1}^\bullet, \ldots, z_1^\bullet\right) \\ &= \frac{f_k(d)\,d}{N} \sum_{i=1}^{d} J_i \; g_k\left(x_i, y_i, \ldots, z_{k-1}^\bullet, \ldots, z_1^\bullet\right) \\ &= \frac{f_k(d)\,d}{n} \sum_{i=1}^{d} w_i \; g_k\left(x_i, y_i, \ldots, z_{k-1}^\bullet, \ldots, z_1^\bullet\right) \\ &= f_k(d) \sum_{i=1}^{d} w_i' \; g_k\left(x_i, y_i, \ldots, z_{k-1}^\bullet, \ldots, z_1^\bullet\right), \qquad k = 1, \ldots, p \end{aligned} \tag{15}$$

Examples

Historical data for the Maumee River basin were used to construct a sample of annual maximum daily river flows from the Maumee River into Lake Erie (Croley 2000b, 2001). Conventional storm frequency estimates are calculated for the Maumee River annual maximum daily flow with the Weibull estimator in (6) as the blue line in Figure 3. Croley (2000b, 2001) then used El Niño conditional probabilities to calculate weights to bias storm frequency estimates for a forecast made in September 1999. The weights are given in Table 1. Climate-*biased* storm frequencies for the annual maximum daily flow are estimated by applying the weights in Table 1 to the historical sample of annual Maumee River basin flow extremes with the Weibull estimator in (7), to estimate the storm frequencies as the red line in Figure 3. As mentioned in the last section, the presence of zero-valued weights leads to strange behavior in the estimate when they are used directly in the weighted statistics. The Weibull estimator of exceedance probabilities [last line in (7)] assigns the same exceedance probability estimate to more than one sample value of *X* when there are zero weights. This gives rise to horizontal line segments in

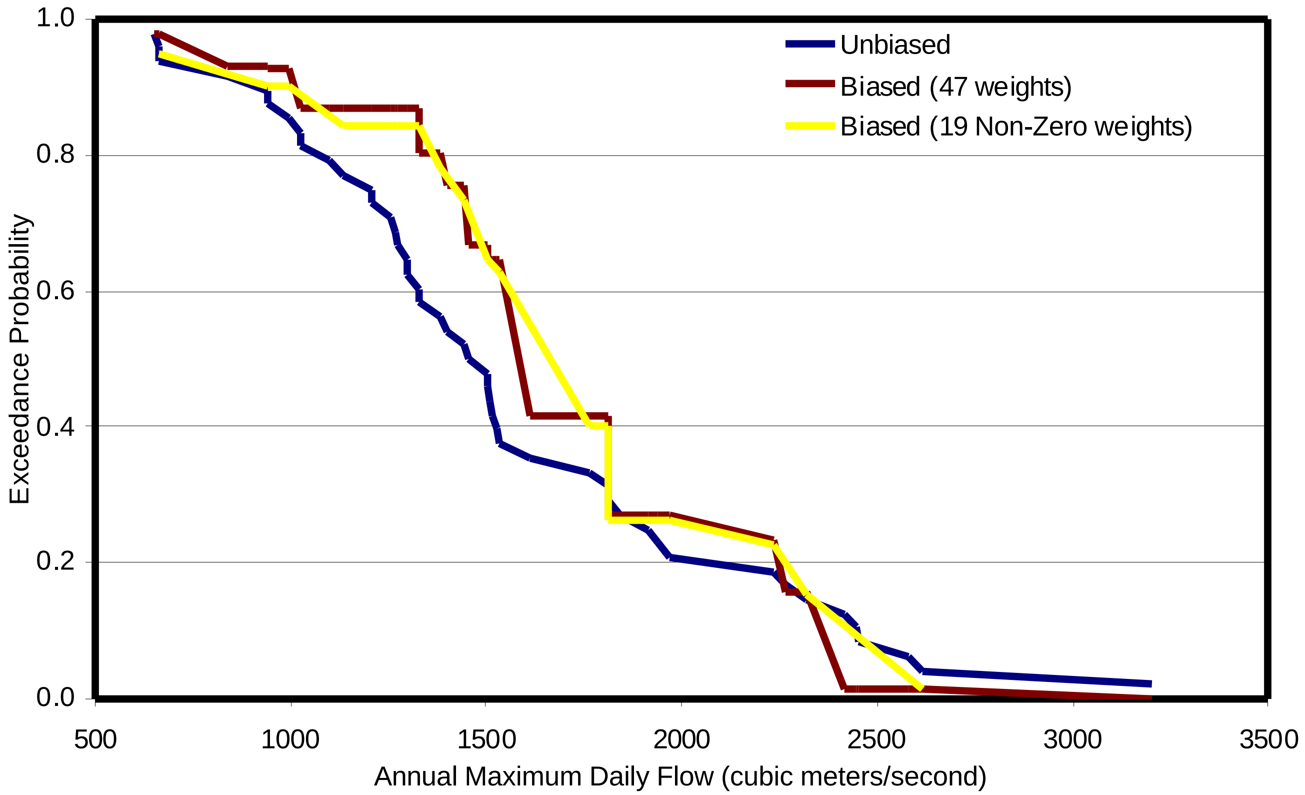

Figure 3. Non-parametric estimates of exceedance probability with the Weibull statistic.

Table 1. Maumee River Weights for Biasing Annual Maximum Daily Flow Probability Exceedance Estimates.

Year	Weight	Year	Weight	Year	Weight
(1)	(2)	(3)	(4)	(5)	(6)
1949	0.003264	1965	0	1981	0
1950	0.716331	1966	3.560669	1982	0
1951	0	1967	2.376111	1983	11.123379
1952	0.979820	1968	0	1984	0
1953	0.007180	1969	0	1985	0
1954	0.018277	1970	0	1986	0
1955	0	1971	0	1987	2.377790
1956	0	1972	0	1988	0
1957	0.005875	1973	0.000839	1989	0
1958	0	1974	1.900889	1990	6.854586
1959	0	1975	0	1991	0
1960	0	1976	0	1992	0
1961	0	1977	0.004943	1993	0
1962	0.000839	1978	0	1994	6.891700
1963	2.836320	1979	0	1995	4.243991
1964	3.097198	1980	0		

Table 2. Adjusted Non-Zero Weights for Biased Maumee River Exceedance Probability Estimates.

Year (1)	Weight (2)	Year (3)	Weight (4)	Year (5)	Weight (6)
1949	0.001319	1963	1.146597	1983	4.496685
1950	0.289581	1964	1.252059	1987	0.961234
1952	0.396097	1966	1.439419	1990	2.771003
1953	0.002903	1967	0.960556	1994	2.786006
1954	0.007389	1973	0.000339	1995	1.715656
1957	0.002375	1974	0.768444		
1962	0.000339	1977	0.001998		

the red line in Figure 3. This is undesirable since these successive estimates may be later used to linearly interpolate for a value of *X* (annual maximum daily flow) given an exceedance probability; the many (infinite) possible values of *X* make it indeterminate. Also note another problem with the red line in Figure 3; the exceedance probability goes to zero as the flow gets large. This results because the weight associated with the largest value in the sample is zero-valued. However, the Weibull estimator does not have this property in the unweighted case and was not intended to.

Nineteen of the weights in Table 1 are non-zero; their values are adjusted with (12) and appear in Table 2. Climate-*biased* storm frequencies for the annual maximum daily flow are estimated by applying the weights in Table 2 to the historical sample of annual Maumee River basin flow extremes with the Weibull estimator in (7), as modified by (15) for using only non-zero valued weights, to estimate the storm frequencies as the yellow line in Figure 3. Now there are no horizontal portions to the estimate (the two segments that appear horizontal are actually slightly sloped). Thus, the yellow line in Figure 3 may be inverted to linearly interpolate a flow value given an exceedance probability. Also, the yellow line in Figure 3 shows that the exceedance probability does not go to zero as the flow gets large. This is a desirable property for the Weibull estimator.

Finally in Figure 3, note that the weighted (biased) Weibull estimators shift the distribution from the unweighted (unbiased) case. Both of the weighted estimates are steeper than the unweighted estimate and increase the exceedance probability for low flows and decrease it for high flows. These observations can be sharpened and refined by using a parametric estimator; that is, a family of distributions is assumed to apply to the sample data. Croley (2000b, 2001) used the first 3 equations in (6) and (7) to fit Log-Pearson Type III distributions to unweighted and weighted samples (respectively) of Maumee River data. Figure 4 shows the resulting Log-Pearson Type III distributions. The blue line in Figure 4 corresponds to the unweighted (unbiased) estimate; the red line corresponds to the biased estimate made with all the weights in Table 1; the yellow line corresponds to the biased estimate made only with the 19 non-zero-valued weights in Table 2. Two interesting observations can be made in Figure 4. First, we can refine our estimates of the effect of biasing the estimation to reflect the El Niño conditions used to calculate the weights. Both weighted estimates are steeper than the unweighted estimate and increase the exceedance probability for flows below about 2200 m^3s^{-1} and decrease it above. Second, we observe that both weighted (biased) estimates are very similar (almost indistinguishable) at this scale. This is a point of information on how important (or not) it is to account for zero-valued weights. In this example, the use of zero-valued weights and sample moments unadjusted for zero-valued weights has little effect on the biasing. In general, using all weights (including zero values) often is sufficient; using only non-zero values can have practical advantages but is not necessarily theoretically better.

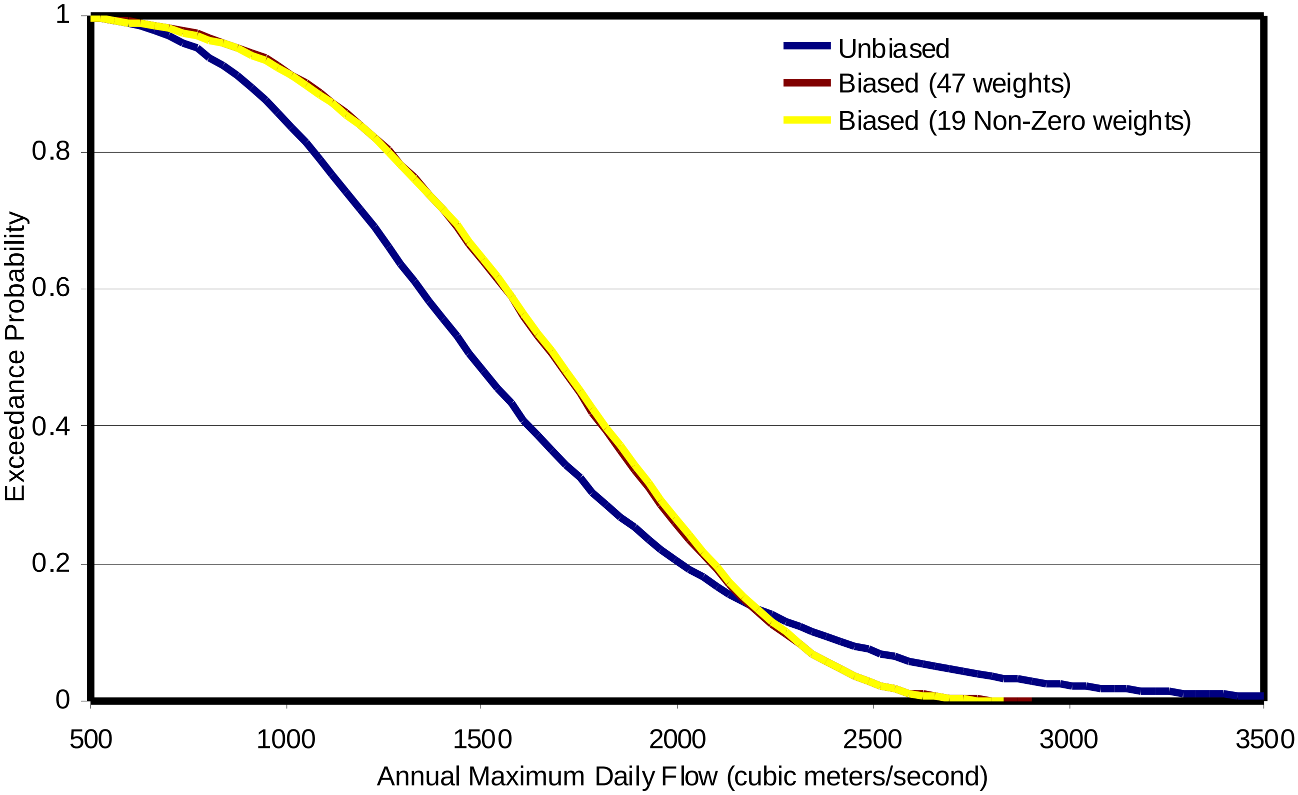

Figure 4. Parametric estimates of exceedance probability using Log-Pearson Type III distribution.

Summary

The origin of a weighting technology for biasing a historical sample of observations lies in the construction of a very large biased sample from the original sample. Weighted sample statistics can be defined equivalently over the original sample, preserving the form of the unweighted statistics when the weights are all unity and then extended for the case where some weights are zero-valued via partial sampling. Corrections must be made to the weights in this case to preserve desirable statistical properties. Historical data for the Maumee River basin and biasing weights (from others' studies of El Niño) were used to construct non-parametric storm frequency estimates with the Weibull estimator and parametric estimates using the Log-Pearson Type III distribution. The presence of zero-valued weights leads to strange behavior in the Weibull estimator, allowing the same exceedance probability for more than one sample value of *X*. Inverting this distribution estimate is therefore not possible since there are many values of *X* for a given value of exceedance probability. Another problem is that the exceedance probability can be zero at the maximum sample flow, which is not a desired property of the Weibull estimator. Using only non-zero-valued weights in the biasing eliminates these problems. On the other hand, the use of zero-valued weights may make only small differences in parametric estimates. The example Log-Pearson Type III distributions, fitted both with all weights and with only non-zero-valued weights, are very similar. The use of zero-valued weights and sample moments unadjusted for zero-valued weights had little effect on the biasing. Complete software, in the form of an easy-to-use interactive *WindowsÔ* graphical user interface, and worked examples are available free of charge over the World Wide Web. The software, examples, and tutorial materials may be acquired in a self-installing file by visiting from the web site entitled http://www.glerl.noaa.gov/wr/OutlookWeights.html and downloading.

Acknowledgments

This is GLERL contribution no. 1186.

References

Croley, T.E., II (1996). Using NOAA's new climate outlooks in operational hydrology. *Journal of Hydrologic Engineering*, ASCE, **1**(3):93—102.

Croley, T.E., II (1997). Mixing probabilistic meteorology outlooks in operational hydrology. *Journal of Hydrologic Engineering*, ASCE, **2**(4):161—168.

Croley, T. E., II (2000a). *Using Meteorology Probability Forecasts In Operational Hydrology*. ASCE Press, American Society of Civil Engineers, Reston, Virginia, 214 pp.

Croley, T. E., II (2000b). Climate-Corrected Storm-Frequency Examples. *NOAA Technical Memorandum ERL GLERL-118*, Great Lakes Environmental Research Laboratory, Ann Arbor, Michigan, 30 pp.

Croley, T.E., II (2001). Climate-biased storm-frequency estimation. *Journal of Hydrologic Engineering*, ASCE, (in press).

Croley, T. E., II. Climate-Biased Decisions Via Partial Historical Sampling. Proceedings, World Water & Environmental Resources Congress, Bridging the Gap: Meeting the World's Water and Environmental Resources Challenges May 20-24, 2001, Orlando, Florida. D. Phelps and G. Sehlke (eds.). Environmental Water Resources Institute, American Society of Civil Engineers, Washington, DC, Compact Disc (2001).