Drought Frequency Analysis for California from Observed, Synthetic, and Proxy Data

Frederick Wade Freeman and John A. Dracup

Drought frequency analysis can be performed with statistical techniques developed for determining recurrence intervals for extreme precipitation and flood events (Linsley *et al* 1992). The drought analysis method discussed in this paper uses the log-Pearson Type III distribution, which has been widely used in flood frequency research.

Some of the difficulties encountered when using this distribution for drought analysis are investigated. These difficulties arise because this distribution has been developed from statistics derived from large-sample-size datasets, while streamflow time series data generally contain only a limited number of drought events and, therefore, most drought datasets are of small sample size. For example, the 86-year record (starting in water year 1906) of unimpaired flow data for the 4-River Index, which represents streamflow in Northern California, contains only 19 drought events, 10 of which are minor events with durations of only 1 year. Large-sample-size statistics are typically based on datasets of 30 events or more. The problem the hydrologist faces is that the analysis of recurrence intervals for drought events of severity critical to the management and operation of water delivery systems is often based on datasets of very small sample size. If an adequate amount of data were to become available, then the methods developed for floodflows could be applied to droughts with only moderate adjustments (Lee *et al* 1986).

Two possible responses to the limiting condition of small-sample-size datasets are:

- Development and application of small-sample-size statistics specifically for use in drought frequency analysis, or
- Extension of the flow data record to increase the number of drought events, thus allowing analysis by standard methods such as the log-Pearson Type III distribution.

The first response has been investigated by a number of researchers, including Sen (1980), Lee *et al* (1986), and Nathan and McMahon (1990).

The second possibility, which is explored in this paper, involves expansion of the number of drought events in the time series record for a particular drainage basin. Two options exist for expanding the sample size of drought events for a given drainage basin. They are:

In: K.T. Redmond and V.L. Tharp, Editors. 1993. Proceedings of the Ninth Annual Pacific Climate (PACLIM) Workshop, April 21-24, 1992. California Department of Water Resources, Interagency Ecological Studies Program, Technical Report 34.

- Use of statistical modeling to generate additional years of synthetic flow data with statistical characteristics similar to the original dataset. These statistical models can be operated until a sufficient number of synthetic drought events have been created to increase sample size for use with large-sample-size statistics. A review of statistical streamflow models capable of generating synthetic droughts has been presented by Lawrance and Kottegoda (1977).
- Augmentation of the observed flow record with proxy flow data reconstructed from regression analysis of paleoclimate records and, in the specific case of this research, from tree-ring chronologies. An example of reconstructing streamflows from tree-rings has been presented by Stockton (1975).

Methodology and Results

Discussed here is the application of the log-Pearson Type III distribution to various interpretations of the flow data record for the 4-River Index. This distribution is used to relate drought severity to recurrence interval. A limitation of this analysis is that it cannot determine drought duration, although Lee *et al* (1986) have presented an approach for frequency analysis of drought duration that can be employed in conjunction with the method presented here.

Figure 1. The rivers of the 4-River Index and locations of their streamflow gauge stations.

The hydrologic system under investigation is the 4-River Index of the Sacramento River Basin, for which annual unimpaired flow data exist for 1906 through 1991 (data provided by California Department of Water Resources). Figure 1 identifies the rivers constituting the 4-River Index. Tree-ring sampling locations used to reconstruct annual 4-River Index flows for 1560 to 1980 are shown in Earle and Fritts (1986).

Flow data are shown in Figure 2 for the observed, synthetic, and proxy records. Mean flow values are included for each of these three interpretations of the flow record.

Figure 3 shows two interpretations of the observed 86-year record. These interpretations are the theoretical and empirical values for (1) streamflow frequency for each water year and (2) drought frequency for each drought event (generally multi-

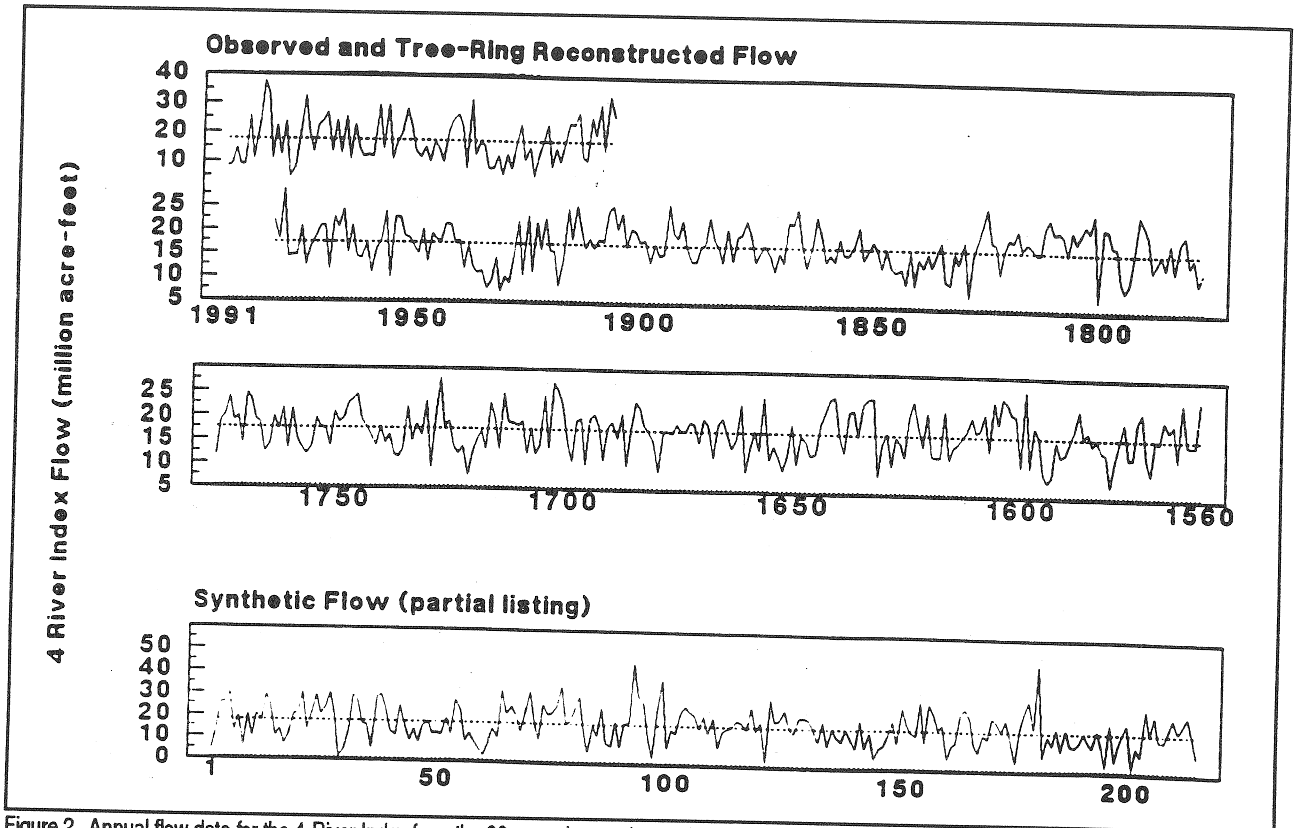

Figure 2. Annual flow data for the 4-River Index from the 86-year observed record (top graph), the 420-year tree-ring reconstructed record (top and middle graphs), and a partial listing of the 10,000-year synthetically generated record (bottom graph). Mean annual flows are 17.81 million acre-feet for the observed and synthetic records and 17.46 million acre-feet for the reconstructed record.

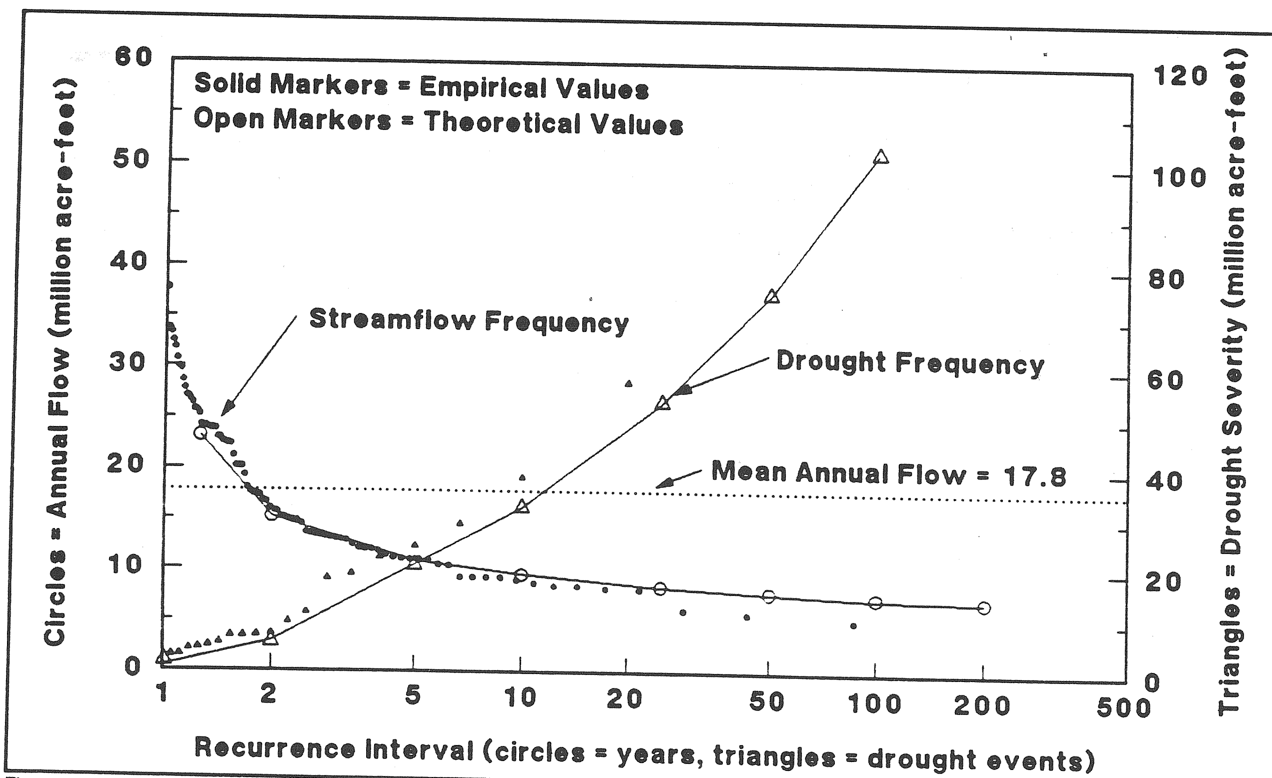

Figure 3. Streamflow and drought frequency analysis for the 4-River Index from observed data. Solid markers represent empirical values determined with equation (1). Open markers indicate theoretical values determined by the log-Pearson Type III distribution. The axes are labeled differently for each curve. Mean annual flow is shown by the dotted line.

year in duration). Theoretical frequencies were computed by the log-Pearson Type III distribution, and the empirical recurrence interval, T_r , is determined by:

$$T_r = N+1/m \tag{1}$$

(Linsley *et al* 1992), where N = sample size and m = rank of each drought event. The curve labeled "streamflow frequency" shows that for a water year similar to 1991 (8.4 million acre-feet, which is 47 percent of the annual mean flow) the theoretical recurrence interval is about 18 years. This is plotted by the open-circle markers. Theoretical values compare well with those computed empirically, which are plotted by the solid-circle markers, with the exception of the three lowest flows, further discussed below.

Two limitations of this application of the log-Pearson Type III distribution are that it applies only to single-year flows and not to multi-year droughts and that it does not account for the non-independence of annual flow data.

The curve in Figure 3 labeled "drought frequency" details an alternative procedure for applying the log-Pearson Type III distribution to the same 86-year record of observed flows to overcome the above-stated limitation. This procedure calculates theoretical frequency curves from the series of drought events, each with a measurable severity and duration (after Dracup *et al* 1980), contained within the observed flow record. The threshold flow defining a drought year is set at 5 percent below the mean annual flow. This curve relates drought severity to recurrence interval. However, a consequence of extracting the series of drought events from the annual flow data is that the series loses its temporal character; that is, N in equation (1) is no longer a measure of the number of years in the observed flow record but, instead, is a measure of the number of drought events occurring over this observed record. This results in a recurrence interval measured by the number of interceding drought events (all having less severity) and not interceding years. This unit of measure is difficult to use in drought frequency analyses.

This difficulty is overcome by empirically determining a recurrence interval measured in years. This is achieved by determining the average interval between drought event onsets calculated from the total number of droughts occurring over the total record length. For the 4-River Index, 19 drought events have occurred during the 86-year observed record, which is equivalent to 4.53 years between events.

Table 1 summarizes this recurrence interval calculation for a frequency analysis of a drought with a severity of 38 million acre-feet. Column A of Table 1 shows that a recurrence interval of 14 drought events is calculated from the observed frequency curve for a drought of the given severity. The recurrence interval is then converted into units of years,

Table 1
SUMMARY OF RECURRENCE INTERVALS
(For a Drought of Severity = 38 Million Acre-Feet)

Data Source	Recurrence Interval from Graphs (in Drought Events) A	Length of Record (in Years) B	Number of Drought Events C	Years Between Onsets (per Drought Event) D (D=B/C)	Recurrence Interval (in Years) E (E=AxD)
Observed	14	86	19	4.53	63
Synthetic	24	10,000	2,233	4.48	108
Proxy	70	420	85	4.94	346

column E, by the use of the empirical return frequency determined in column D. For the observed data, the recurrence interval is 63 years. Synthetic and proxy data have longer recurrence intervals, 108 and 346 years respectively.

Noticeable on both curves in Figure 3 is how the theoretical curve underestimates severity when compared to the empirical data. For the streamflow frequency curve, this is limited to the extreme events, which are water years 1931, 1924, and the lowest year on record, 1977. This behavior may suggest that extreme low flow years follow a different distribution or are caused by a physical mechanism separate from that causing the other low-flow years. For the drought frequency curve, all events are underestimated. This may be due to the small sample size of drought events contained in the observed record.

Figure 4 attempts to improve the accuracy of the theoretical drought frequency curve of Figure 3 when compared to empirical data by displaying curves for expanded versions of the 4-River Index flow obtained from (1) synthetically generated flow values and (2) proxy flow values reconstructed from tree-ring chronologies. The theoretical drought frequency curve from synthetic data accurately reproduces the trend of the empirical data as determined from 2,233 drought events contained within 10,000 years of annual flow data generated from an ARMA(1,1) statistical model. The ARMA(1,1) model was selected over other autoregressive models because it had the best AIC value (Akaike 1976). This type of model has been described by Box and Jenkins (1970). For clarity, not all 2,233 drought events used to determine the theoretical curve are shown by the solid-triangle markers plotted in Figure 4. The theoretical drought frequency curve for the proxy data from tree-ring reconstructions also accurately reproduces the trend of the empirical data from the 85 drought events within the 420 years of reconstructed flow, with the exception of the single most severe reconstructed drought event of 1928-1939. This 12-year period recorded a total flow deficit of 72.8 million acre-feet and, when plotted, lies far above the theoretical curve. These two theoretical curves better represent the severity of moderate and extreme drought events (events with deficits of about 20 to 60 MAF) than the theoretical drought curve of Figure 3 when compared to the empirical values.

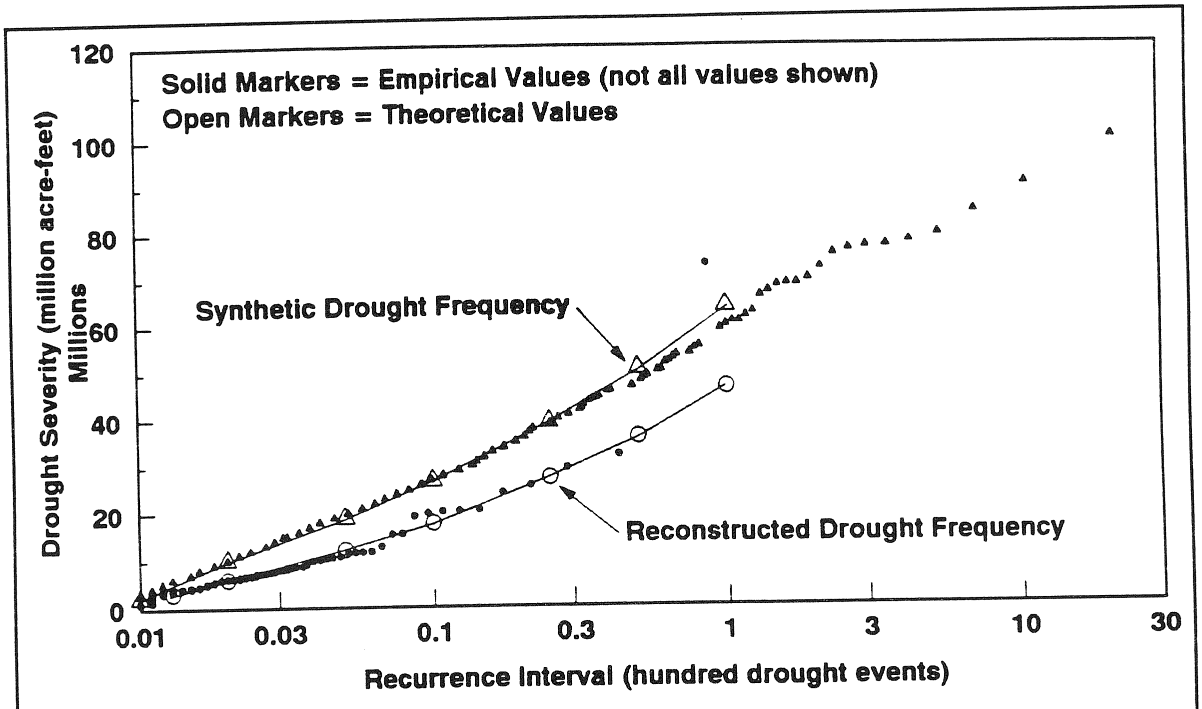

Figure 4. Drought frequency analysis for the 4-River Index from ARMA(1,1) synthetically generated data and tree-ring reconstructed proxy data. Solid markers represent empirical values determined with equation (1). Open markers indicate theoretical values determined by the log-Pearson Type III distribution. The horizontal axis is scaled in drought events.

Figure 4 suggests that analyzing synthetic data from an appropriate statistical model or proxy data from tree-ring reconstructions may provide better frequency analysis from distributions developed for large sample sizes, such as the log-Pearson Type III distribution. However, there are deficiencies in both the synthetic and proxy datasets. The ARMA(1.1) model computes an occasional impossible negative annual flow value. For the analysis presented here, all negative flows were set to zero, although zero or near-zero flows have never been recorded by the 4-River Index, suggesting severity of a drought event containing a zero flow year would be unduly exaggerated. Reconstructed flow from tree-ring data, on the other hand, display less variance than observed flow in that it makes estimates for high or low flows that are less accurate than estimates for flows close to the mean (Earle and Fritts 1986).

The log-Pearson Type III method assumes independence of the input data. To check for independence, lag-one autocorrelation coefficients were determined for each of the four frequency curves presented in Figures 3 and 4. The autocorrelation results are well within the confidence interval in each case and are summarized in Table 2.

Figure 5 compares the two log-Pearson Type III theoretical frequency curves of Figure 4 and the same curve for the observed data from Figure 3. The curve from the observed data provides the shortest recurrence intervals; the proxy data curve provides the longest. The recurrence interval of drought events with severity of 38 million acre-feet is considered as an example (refer to Table 1). This severity is about equal to the first five years (1987-1991) of the current 6-year California drought. This

Table 2
SUMMARY OF LAG-ONE AUTOCORRELATION COEFFICIENTS

Data Source	Type of Frequency Curve	Curve Location	N (from Eq. 1)	Autocorrelation Coefficient	95% Significance Level	
Observed	Annual Flows	Figure 3	86 Years	0.095	0.167	-0.190
Observed	Drought Events	Figure 3	19 Events	0.048	0.328	-0.446
Synthetic	Drought Events	Figure 4	2,233 Events	0.009	0.034	-0.035
Proxy	Drought Events	Figure 4	85 Events	-0.081	0.167	-0.192

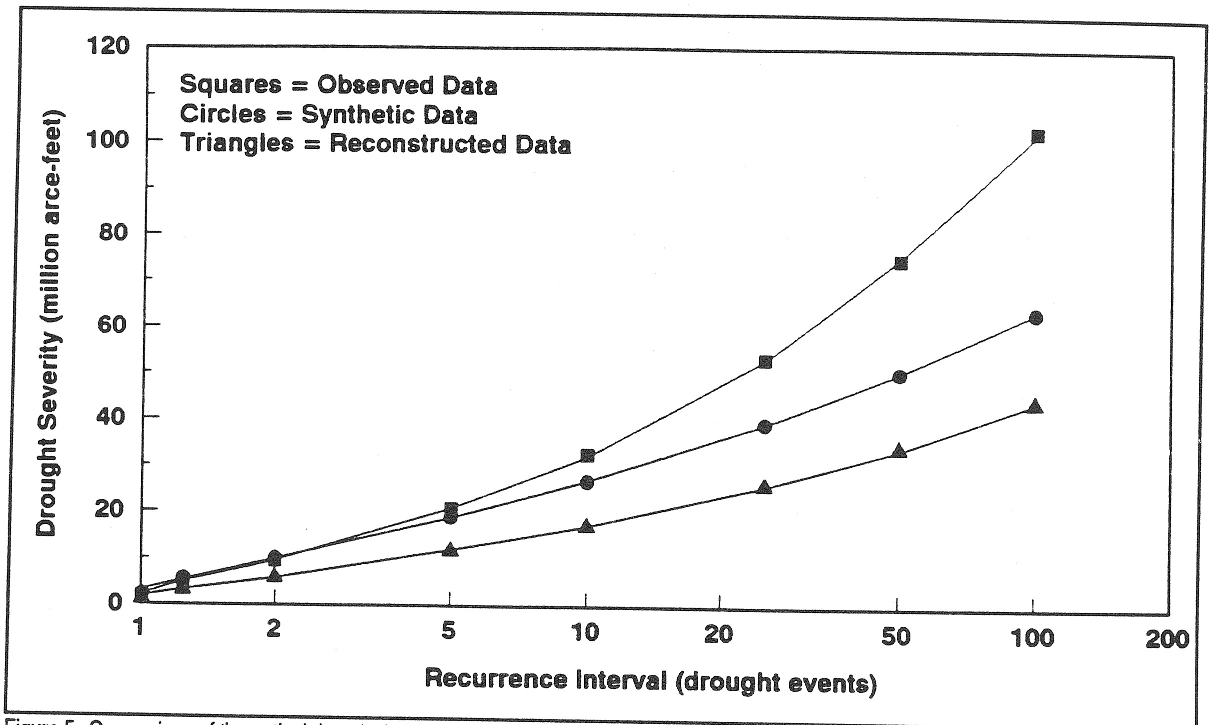

Figure 5. Comparison of theoretical drought frequency curves from Figures 3 and 4 for observed, synthetic, and tree-ring reconstructed (proxy) data for the 4-River Index. The horizontal axis is scaled in drought events.

is the second-worst drought, behind the drought of the 1930s, for the observed and reconstructed records.

The disparity between the observed curve and the synthetic and proxy curves in Figure 5 becomes even greater for larger severities. This suggests the observed dataset is being adversely affected by its small sample size during log-Pearson Type III analysis.

Conclusions

Frequency curves from both synthetic and proxy data reproduce the trend of the empirical values better than the frequency curve from the observed data. The divergence of the observed data frequency curve from the synthetic and proxy curves may represent the influence of the inappropriate application of the log-Pearson Type III distribution to datasets of limited sample size.

Both synthetic and proxy data-based analyses have advantages and drawbacks. The synthetic data sample size can be expanded as desired; however, it is based on the sample statistics of the observed dataset, which is often of limited temporal duration (86 years for the 4-River Index) and, therefore, may not represent the best available sampling of the population. While tree-ring reconstructed data, on the other hand, may be adversely affected due to a loss of observed variance, the data do have the advantage of being based on the much longer sample size, in this case 420 years of annual flow values. However, analysis of the results presented here suggests the loss of variance in the tree-ring reconstructions maybe sufficient to cause significant overestimation of recurrence intervals, as is apparent from column E of Table 1 for the proxy data, although no quantitative measure of this has been made.

Acknowledgment

Support for this research was provided by the National Science Foundation through the Small Grants for Exploratory Research program contract BCS 9114753.

References

Akaike, H. 1976. Canonical correlation analysis of time series and the use of an information criterion. In R Mehra and DG Lainiotis (Editors), *Advances and Case Studies in System Identification*, pp. 27-96. Academic Press, New York.

Box, GEP, and GM Jenkins. 1970. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Dracup, JA, KS Lee, and EG Paulson. 1980. On the definition of droughts. *Water Resources Research* 16(2):297-302.

Earle, CJ, and HC Fritts. 1986. *Reconstructing Riverflow in the Sacramento Basin Since 1560*. Prepared by the Laboratory of Tree-Ring Research, University of Arizona, Tucson for the California Department of Water Resources, Sacramento.

Lawrance, AJ, and NT Kottegoda. 1977. Stochastic modelling of riverflow time series. *Journal of the Royal Statistical Society*, A.140(1):1-47.

Lee, KS, J Sadeghipour, and JA Dracup. 1986. An approach for frequency analysis of multiyear drought duration. *Water Resources Research* 22(5):655-662.

Linsley, RK, JB Franzini, DL Freyberg, and G Tchobanoglous. 1992. Chapter 5 in *Water-Resources Engineering*, Fourth Edition. McGraw-Hill, New York.

Nathan, RJ, and TA McMahon. 1990. Practical aspects of low-flow frequency analysis. *Water Resources Research* 26(9):2135-2141.

Sen, Z. 1980. Statistical analysis of hydrologic critical droughts, *Journal of the Hydraulics Division, Proceedings of the American Society of Civil Engineers* 106:99-115.

Stockton, CW. 1975. *Long Term Streamflow Reconstructed From Tree Rings*. University of Arizona Press, Tucson.