# User Manual for Program *PeakfqSA*
# Flood-Frequency Analysis with the Expected Moments Algorithm

John England, USACE
Tim Cohn, USGS

12 June 2019

# Contents

## Abstract

Bulletin 17B specifies the government-mandated method for conducting flood frequency studies in the United States. Although B17B has remained unchanged since 1981, researchers have recently proposed changing B17B to employ a more general parameter-estimation method, specifically the Expected Moments Algorithm (EMA), which can employ interval data in the context of the moments-based method currently in use. *PeakfqSA* is a computer software package that implements EMA. It is intended to be used by researchers and hydrologists who wish to experiment with EMA. This draft manual documents both the usage and the technical details of *PeakfqSA*. The manual is intended to be used in conjunction with Flynn et al. 2006, which documents practical aspects of flood frequency analysis, such as how flood data are collected and interpreted, as well as the conceptual basis for *PEAKFQ*, an approved software package which implements B17B.

# 1 Introduction

Many federal projects and activities in the United States are vulnerable to flooding, and for this reason the federal government has developed procedures for estimating flood quantiles. The most important application involves estimating the 100-year flood for the National Flood Insurance Program, where use of a consistent method helps ensure fairness and economic efficiency.

Since 1981, Bulletin 17B (IACWD, 1982, “B17B”) has codified the standard methodology for conducting flood frequency studies in the United States. However, B17B is not flawless. Its authors included a section identifying specific components of B17B that might benefit from additional studies. Based on studies performed in response to this recommendation, a group of researchers and practicing hydrologists have recently called for modest changes to the B17B Guidelines Griffis and Stedinger (2007).

Adoption of the Expected Moments Algorithm (EMA) is among the changes that has been suggested. EMA directly addresses several concerns identified in B17B (IACWD, 1982, p. 27, “Future Studies”), while maintaining the essential structure of existing B17B procedures. In particular, EMA permits efficient use of interval data, which arise in the context of historical flood information Cohn et al. (1997), low outliers Griffis et al. (2004b); Cohn et al. (2013), uncertain data points Cohn et al. (2001), and in special other situations. EMA also provides a method for computing reasonably accurate confidence intervals Cohn et al. (2001). EMA, though computationally demanding, retains the structure and moments-based approach of Bulletin 17B and is therefore easily accommodates the other B17B procedures Cohn et al. (1997, 2001); England et al. (2003b,a); Griffis et al. (2004a,b).

# 2 Estimating Flood Frequency at a Streamgage

Flood frequency at a streamgage site is usually estimated by statistical analysis of the historical experience at that site. The gage record is examined and a flood frequency distribution – the Log-Pearson Type 3 (LP3) – is fit to the annual peak flow data. The fitted distribution is then used to provide flood-quantile estimates.

## 2.1 Data Sources

The U.S. Geological Survey maintains a peak-flow file in the National Water Information System (NWIS), which is available on-line at nwis.usgs.gov and described in detail in Flynn et al. (2006). Because EMA can derive useful information by incorporating non-standard data – data that are not typically collected as part of a flood study and are not usually available through NWIS – meeting EMA's data requirements may require some changes in standard data collection practices.

## 2.2 How EMA Characterizes Peak Flow Data

In a traditional B17B analysis, peak-flow data are associated with a regular streamgaging program. These are called systematic data. However, at many sites one finds evidence of "historical" floods as well – flows that were recorded because they were sufficiently large to lead to "historic" records (Flynn et al., 2006, see for additional discussion). In short, B17B recognizes two categories of data:

**Systematic** Annual peaks observed in the course of "systematic" streamgaging at the site;

**Historic** Records of floods that occurred outside the period of regular streamgaging. Such data may be available from anthropogenic records such as flood lines on bridge piers or media accounts, from botanical evidence such as flood scars on trees, or physical evidence such as slack-water deposits Jarrett and Malde (1987); Stedinger and Baker (1987); Hupp (1987); Wohl (1995); Jarrett and Tomlinson (2000).

EMA employs a more general description of peak-flow data and, in doing so, eliminates the need to distinguish between historical and systematic data (floods never notice this distinction anyway). For every year $Y$ during the "historical period" – which can extend as far back in time as one can find useful information – we assume there was a peak discharge, $Q_Y$. We describe our knowledge of $Q_Y$ by the interval $(Q_{Y,lower}, Q_{Y,upper})$. For traditional systematic data, this generalized description reduces to $(Q_{Y,lower}, Q_{Y,upper}) = (Q_Y, Q_Y)$.

For sites with historical information, the generalized description is more complicated. For example, on the Naugatuck River (CT), it is known that the great flood of 19 August 1955, whose magnitude was determined to be $Q_{1955} = 106000[cfs]$ at Beacon Falls, CT (01208500), exceeded all other flows from 1642 to the present. Thus, for every year $Y$ from 1642 on, we can at least say that $(0 < Q_Y \leq Q_{1972} = 106000[cfs])$, or, using the notation above, that $(Q_{Y,lower}, Q_{Y,upper}) = (0, 106000)$. Of course, more precise information is available for many of the years, and thus tighter intervals can be constructed. Nonetheless, it has been found that even such broad non-exceedance information can provide the equivalent of decades to hundreds of years worth of systematic streamgage data Stedinger and Cohn (1987, 1986); Cohn and Stedinger (1987); England et al. (2003b).

The discharge interval corresponding to a year in the very distant past may approach $(Q_{Y,lower}, Q_{Y,upper}) = (0, \infty)$. Such vague information will have no effect on the fitted frequency distribution, but it does no harm to include it, either.

## 2.3 Thresholds: A Critical Concept in EMA

While B17B employed an *artificial* distinction between *historical* and *systematic* data, the distinction served an important purpose: To differentiate observations taken when the entire range of flows would have been recorded (i.e. systematic data) from peak flows that were recorded *because* their magnitudes were sufficient to lead to creation of a permanent record Stedinger and Cohn (1986). The authors of B17B recognized that these cases have to be treated differently, and, generally, one needs to understand the sampling properties of each peak flow in order to determine what it tells us about the peak-flow distribution. The B17B "historic moments adjustment" reflects this.

EMA makes this distinction explicit. It employs a record of "observational thresholds," denoted $(T_{Y,lower}, T_{Y,upper})$, which reflect the range of flows whose magnitude would have been recorded had they occurred. The lower bound, $T_{Y,lower}$, represents the smallest annual peak flow that would result in a permanent record. For periods of systematic record, this is typically represented by the "Gage-base discharge," which is typically 0.

For most sites, $T_{upper}$ is assumed to be infinite; bigger floods typically get recorded. However, one can imagine cases where a really big flood might do so much damage that no one can estimate exactly how big it was except to say that it was bigger than some previous flood. In that case, $T_{upper}$ might be set equal to the largest flood that could be expected to result in a reliable discharge estimate.

EMA employs the flow intervals, $\{(Q_{Y,lower}, Q_{Y,upper})\}$, to estimate the fitted distribution. EMA requires the corresponding thresholds, $\{(T_{Y,lower}, T_{Y,upper})\}$, to estimate the confidence intervals and other measures of uncertainty in frequency estimates. It is therefore important to estimate the thresholds accurately, which has not been necessary in the past when using *PEAKFQ*.

For further discussion of available data, see Cohn et al. (2001); Flynn et al. (2006).

# 3 Regional Information

*PeakfqSA* accepts regional information that can be used to improve estimates of the at-site skew ($G$, which is prescribed in Bulletin 17B) and the at-site mean ($M$) and standard deviation ($S$), which is not described in Bulletin 17B, but is needed in special cases where there is inadequate at-site data to fit the LP3 distribution.

For $M$, $S$, and $G$ one can specify use of at-site data only, regional data only, or a weighted average of the two. Weights are computed inversely proportional to the MSE of each estimate:

$$W = \frac{MSE_{X_{Reg}} X_{AtSite} + MSE_{X_{AtSite}} X_{Reg}}{MSE_{X_{Reg}} + MSE_{X_{AtSite}}} \tag{1}$$

where $X$ refers to one of the estimators $\{M, S^2, G\}$.

## 3.1 Weighting Regional Estimators

Regional information is usually specified in terms of an estimated value of the parameter and a mean square error (MSE). It is assumed that the regional estimators for each of the parameters are independent. One can specify that EMA employ the regional information alone (”Generalized”), the at-site information along (”Station”), or a weighted combination where the weights are calculated to minimize the overall MSE (”Weighted”). This works reasonably well for the mean, $M$, and the skew, $G$, but is problematic when it comes to the variance, $S^2$.

## 3.2 Weighting Regional Variance

The MSE of the at-site variance, $MSE[S^2_{AtSite}]$, is proportional to the unknown value of $\sigma^4$. Because the at-site and regional estimates of $\sigma^4$, $S^2_{Reg}$ and $S^2_{AtSite}$, respectively, typically differ, use of the estimated MSE will typically result in placing too much weight on the smaller of the estimates for $S^2$. The result would be a biased estimator.

However, one way to avoid this problem is to express both estimates in terms of ”degrees of freedom” – the number of independent observations that each estimator employs. $DF$ for an ordinary $\chi^2$ random variable can be expressed as the ratio of twice the expected value squared divided by the variance: $DF \equiv 2E^2[X]/Var[X]$. The weighted estimator for $S^2$ is then computed by dividing the estimated MSE of the at-site and regional estimators by the corresponding parameter estimates for $S^2$, and then defining the weight, $W_{DF}$, corresponding to the at-site estimate as:

$$W_{DF} \equiv \frac{MSE[S^2_{Reg}]/S^4_{Reg}}{MSE[S^2_{Reg}]/S^4_{Reg} + MSE[S^2_{AtSite}]/S^4_{AS}} \tag{2}$$

where $MSE[S^2_{Reg}]$ is the mean square error of the regional variance estimator. An algebraic equivalent is to define an adjusted regional MSE

$$MSE_{Radj} \equiv MSE[S^2_{Reg}](S^2_{AtSite}/S^2_{Reg})^2 \tag{3}$$

which results in the same weightings as achieved by equation 2. Employing $MSE_{Radj}$, the weighted estimator is then given by equation 1.

Because the variance of the estimators for $M$ and $G$, and their respective estimated $MSEs$, are nearly independent of the true values, $\mu$ and $\gamma$, when estimated weighted values for $M$ and $G$ it is not necessary to provide any comparable correction to the one for $S^2$ set forth in equation 3.

## 4 Running *PeakfqSA*: Practical Issues

While it is possible to conduct an EMA or B17B analysis ”by hand,” the computational requirements are onerous and tedious. As a result, B17B Guidelines have been implemented independently as computer software by various various federal agencies, researchers, and private-sector individuals and corporations. The USGS has had several software packages, beginning with Bill Kirby's *J407*, and currently *PEAKFQ*. *PeakfqSA* provides similar relief for those wishing to use EMA.

### 4.1 Step 1: Downloading the *PeakfqSA* Software

*PeakfqSA* software is available from the author as executable or source code. To obtain a copy of *PeakfqSA*, one can:

1. Download a binary executable from `dfdddfadadf`
2. Download source code from `dfdddfadadf` and compile it using a *Fortran-77*compiler.

The source code is typically stored in a “zip” file. This will need to be “un-zipped” before going to the next step.

### 4.2 Installing *PeakfqSA*

*PeakfqSA* can be installed by simply dragging the executable and accompanying data and specification files to a convenient diretory. When properly installed, you should see the following files:

```
BigSandy.dat
BigSandy.psf
BigSandy_hist.psf
PeakfqSA.out (or PeakfqSA.exe)
```

## 4.3 Running the Test Case

*PeakfqSA* is compiled as command-line executable, which must be run through a "terminal" interface. In many cases, double-clicking on the *PeakfqSA* icon will start *PeakfqSA* inside a client Terminal Session (MacOS or Windows). In MacOS the default application will be *Terminal.* The Terminal application should immediately display something like:

```
********************************************************************************
                                       EMA ANALYSIS
                                Computed by PeakfqSA 0.981
                                     Tim Cohn, USGS
                                     tacohn@usgs.gov
                                      703/648-5711
********************************************************************************
```

This will be followed by a request to enter a SPECIFICATION filename.

```
 Enter SPECIFICATION filename (a80)
BigSandy_hist.psf
 Specification lines read:           31
 Data lines read:                    51
 Finished Processing: Normal Termination
```

The numbers printed in the right-hand column indicate the number of lines read from each file.

## 4.4 The Output File

*PeakfqSA* will create one or more output files. For the case given above, the main output file will be named "BigSandy_hist.out". The name is usually the same as that of the specification file, with the only change being that the extension ".out" replaces the ".psf" extension on the specification file. If there is a conflict with filenames, a small integer may be appended to the first component of the filename in order not to overwrite an existing file.

The output file consists of a set of tables designed to describe:

**calibration specifications** The options and parameter values used for the particular flood frequency analysis

**fitted frequency curve** The computed quantiles of the fitted LP3 distribution. Corresponding 95-percent confidence intervals are also printed.

**input data** The data, and corresponding plotting positions, used to calibrate the frequency model.

The output file is intended to replicate the standard output file provided by *PEAKFQ* Flynn et al. (2006). However, because EMA is more general than B17B, it also requires that some additional information be printed.

# 5 The Specification File

The specification file is an ASCII file modeled after the specification file currently used by *PEAKFQ*. It contains a number of lines each containing a keyword and possibly a parameter. The keywords that are currently enabled are

**# or !** Comment line (ignored)

**I** * Filename for data file (see instructions below for format)

**O** Output filename; currently not used

**BEGYEAR** * First year of period for which flood information is available

**ENDYEAR** * Last year of period for which flood information is available

**GAGEBASE** Gage base discharge

**A_S_SKEW_OPT** Method for computing the at-site skew. Options are: *B17B*, *EMA*, and *ADJE*. These refer respectively to the MSE equation in Bulletin 17B (p. 13), the asymptotic method given in Cohn et al. (2001), and an adjusted method that is calibrated to match with Bulletin 17B if only systematic data is present but recognizes information content of historical/interval data info.

**HISTPERIOD** Historical period length (unused)

**HISYS** Threshold for identifying high systematic outliers

**HITHRESH** High threshold (unused)

**LATITUDE** Latitude (unused)

**LOHIST** Observation threshold for historical data (unused; see PCPT_THRESH below)

**LONGITUDE** Longitude (unused)

**PP_ALPHA** Value of $\alpha$ used in computing plotting positions for the data: $P_{i:N} = (i - \alpha)/(N + 1 - 2\alpha)$

**CSV** Option to create ".csv" file for export to Excel

**CLT** Option to censor a percentage of the left-hand tail. The coefficient is expressed as a percentage. When this option is used, the percentage would normally be 50.

**LOMETHOD** Low outlier identification method. Choices are (

- **NONE** No low-outlier test is performed
- **GBT** Standard (B17B) Grubbs-Beck test
- **MGBT** Multiple Grubbs-Beck test
  - **Alphaout** $\alpha$-level used to test all observations below median (outward sweep)
  - **Alphain** $\alpha$-level used to test all observations starting at result from inward sweep (inward sweep)
  - **Alphazeroin** $\alpha$-level used to test all observations starting from smallest observation (inward sweep)
- **FIXED** A fixed critical value, given in LOTHRESH, is employed

**LOTHRESH** Low outlier threshold; if LOMETHOD is "FIXED", this value is used as the threshold, $GB_{thresh}$ below which values are treated as low outliers.

**GENMEAN** Generalized (regional) MEAN

**MEANMSE** Mean square error of generalized mean.

**MEANOPT** Four skew options are available:

- **WEIGHTED** Computes the weighted mean based on the MSE of the at-site and regional mean estimates
- **STATION** Employs the at-site mean, with MSE computed
- **GENERALIZED** Employs the regional mean, with MSE assumed to be zero

**MEANSD** The standard deviation of the skew. If this and MEANMSE are both specified, the latest line in the file determines which is used

**MSE_SD_RATIO** Mean Square Error of the Standard Deviation divided by the true value of $\sigma$. For complete samples, this is asymptotically $\approx 1/(2 * N)$

**SDGEN** The "generalized" (regional) standard deviation applicable to the site. This is not a regular component of Bulletin 17B, but it is included to handle situations that arise, for example, in the Mojave Desert. The default value of SDGEN is 1.0.

**SDMSE** Mean square error of regional standard deviation. There is no default value for this.

**SDOPT** Three standard deviation options are available:

- **WEIGHTED** Computes the weighted standard deviation based on the MSE of the at-site and regional standard deviation estimates
- **STATION** Employs the at-site standard deviation, with MSE computed internally.
- **GENERALIZED** Employs the regional standard deviation, with MSE assumed to be zero.

**SDSD** The standard deviation of the standard deviation ($\sqrt{\text{SDMSE}}$). If this and SDMSE are both specified, the latest line in the specification file determines which is used.

**GENSKEW** Generalized (regional) skew

**SKEWMSE** Mean square error of generalized skew. The default is 0.3025.

**SKEWOPT** Four skew options are available:

- **WEIGHTED** Computes the weighted skew based on the MSE of the at-site and regional skew estimates
- **STATION** Employs the at-site skew, with MSE computed
- **GENERLALIZED** Employs the regional skew, with MSE assumed to be zero
- **GEN/VAR** Employs the regional skew, with MSE given by SKEWMSE

**SKEWSD** The standard deviation of the skew. Default value is 0.55. If this and SKEWMSE are both specified, the latest line in the file determines which is used

**STATION** A character string of not more than 80 characters, used to identify and describe the particular site or case being considered. This character string is printed on the output file.

**PCPT_THRESH** * Threshold information. For every year in the historical period, an interval is specifed for which flood magnitudes would have been measured. Each PCPT_THRESH card contains four parameters following the keyword "PCPT_THRESH":

`PCPT_THRESH start_year end_year` $T_{Y,lower}$ $T_{Y,upper}$

where

**start_year** ($I4$) the first year for which threshold applies

**end_year** ($I4$) the last year for which threshold applies

$T_{Y,lower}$ lower bound on threshold (usually 0 for systematic data)

$T_{Y,upper}$ upper bound on threshold (by convention, $1.e50$ for systematic data)

Note that more than one PCPT_THRESH card can refer to the same year. It is the *last* card whose time interval includes the year that determines which threshold applies for that year.

Periods of Broken Record: For years in which no information is available, the perception thresholds should be set at $T_{Y,lower} = 1.d99$ and $T_{Y,upper} = 1.d99$. The value $1.d99$ is treated as infinity, indicating that we would never know the magnitude of a flood during that period regardless of how large it was.

Table 5 shows an example with some the different keywords and parameters one might find in a specification file.

## 5.1 Creating Comma-Separated-Value (CSV) Ouput

If the CSV keyword is present, with "Yes" specified, an additional data file will be created: *BigSandy_hist.csv* This file contains some of the specification information, the fitted frequency analysis ("ffa") and the original data ("data") combined with plotting position information to facilitate plotting in Excel. The file can be opened directly into an Excel spreadsheet. The file can also be imported into $R$ if one is careful about picking out the necesary rows by using commands like:

```
data <- read.csv(file="BigSandy_hist.csv",skip=26)
```

```
ffa <- read.csv(file="BigSandy_hist.csv",skip=8)[1:15,]
```

Table 1: Specification file for *PeakfqSA*

| Keyword | Parameter<br>Example Value | Definition |
|---|---|---|
| I | BigSandy.dat | Input datafile for discharge data (note format) |
| O | whatever.out | Not used by PeakfqSA |
| STATION | BIG SANDY (03606500) | Station header info |
| LOTHRESH | 0.0 | Low-outlier threshold (if 0, GB crit. value used) |
| SKEWOPT | WEIGHTED | Employ weighted skew option |
| GENSKEW | -0.2 | Regional skew |
| SKEWSD | 0.55 | Std Dev of regional skew |
| BEGYEAR | 1890 | Beginning year of period of record (everything): |
| ENDYEAR | 1973 | Ending year of entire period of record |
| GAGEBASE | 0 | Employ a gagebase of 0.0 [cfs] |
| CSV | YES | Create exportable (Excel) file |
| PCPT_THRESH | 1890 1929 18000 1.d99 | Threshold 1890-1929 is 18000 [cfs] |
| PCPT_THRESH | 1930 1973 0.000 1.d99 | Threshold 1930-1973 is 0 [cfs] |

## 5.2 Deliberate Censoring of Left-Hand Tail (CLT)

At some sites there may be a profound lack-of-fit between the observed annual peak flow data and the LP3 distribution. Usually, this can be attributed to unusual physical circumstances. For example, in regulated systems small annual peaks may be completely controlled; in arid regions, there may be years where no rain occurs; and in certain geologies, such as the granitic environment of the Sierra in California, the groundwater system may be effectively absent thus leading to an irregular left-hand tail.

In such circumstances, it may be desirable to censor the left-hand-tail of the observed data. This can be justified because the smallest annual peak flows often tell us little about the extreme quantiles we are most interested in. Rather, they are mostly a nuisance. Thus, censoring the data at the median value of the systematic record may in some cases make a great deal of sense. This is achieved by entering a line in the specification file:

```
CLT 50.0
```

One should be careful about using this option in the context of a regional skew study, however. Censoring the left-hand tail (CLT 50) will increase the MSE of the at-site skew by a factor of from 2 to 10 – thereby greatly reducing the information available to estimate a regional skew coefficient.

# 6 The Data File

As with *PEAKFQ*, *PeakfqSA* requires annual peak flow data. However, unlike *PEAKFQ*, the flow data must be formatted as a keyword-based file (N.B. Standard NWIS formats are not employed, in part because NWIS does not support interval data). Because the format will likely be unfamiliar to most users, it may be useful to look closely at the example in the Appendix. However, most users will find that it is relatively easy to convert an NWIS file into a file for *PeakfqSA*. The only keywords employed for the datafile are:

**# or !** Comment line (ignored)

**Q** A water-year followed by a single discharge value

```
Q 1897 25000
```

**QINT** A water-year followed by two discharge values which together define an interval bracketing the true discharge

```
QINT 1898 0 18000
```

```
QINT 1899 18500 1.d99
```

Note that the specification file and data file use non-overlapping keywords, and thus one can choose to store the data and specifications in the same file.

# 7 Multiple Batch Processing

*PeakfqSA* has the capability of running multiple sites from a single command. If you create a file containing a list of specification filenames and give it a ".cmd" extension (e.g. "ListOfSpc.cmd"), you can then enter this filename instead of the ".spc" file and *PeakfqSA* will run each of the specification files in sequence.

If you first copy all of the specification files into a directory, it is easy to create the ".cmd" file (in a unix environment, anyway) with the command:

```
ls -l *.spc > ListOfSpc.cmd
```

# 8 Conclusions and Cautions

This manual describes a first-draft implementation of a possible future "Bulletin 17C" procedure. The motivation for producing *PeakfqSA* was to provide a practical tool to enable further studies and evaluation of EMA. However, *PeakfqSA* and this user manual should not be construed as an explicit endorsement of EMA as a generic substitute for the procedures specified in B17B. Where EMA is employed, at least for now, it represents a deviation from the B17B guidelines and will therefore require justification.

Good luck!

# 9 Appendices

## 9.1 Low-Outlier Identification

*PeakfqSA* employs an iterative scheme for identifying low outliers. The algorithm involves iterating 5 steps

1. The LP3 distribution is fit to all the data, using both systematic and historical information. Regional skew information is employed.
2. As in B17B, the Grubbs-Beck test is applied to identify a *critical value*.
3. The smallest observation exceeding the *critical value* becomes a new *low-outlier threshold*. The *low-outlier threshold* applies to all of the data.
4. Observations smaller than the *low-outlier threshold* are classified as *low outliers*. The magnitude of each such observations is recoded as interval $(Q_{Y,lower}, Q_{Y,upper}) = (0, T_L,$ where $T_L$ is the magnitude of the *low-outlier threshold*.
5. The observational thresholds for all the data are adjusted to reflect $T_L$ as a lower bound on what would have been observed.

The sequence is repeated until no additional low outliers are identified.

## 9.2 Plotting Positions for Censored Data

Hirsch and Stedinger 1987 provide an algorithm for assigning plotting positions to measured values for the case of left-censored data. Their formula handles multiple thresholds of the form $(T_{Y,lower}, T_{Y,upper}) = (\tau_i, \infty)$. The algorithm involves the following steps:

1. Arrange the thresholds, $\{\tau_1 = 0 < \tau_2, ..., < \tau_m < \tau_{m+1} = \infty\}$, in increasing order;
2. Define $A_j$ as the number of measured floods $Q$ for which it is known that $\tau_j \leq Q < \tau_{j+1}$;
3. Define $B_j$ as the number of floods $Q$ for which it is known that $Q < \tau_j$
4. Define $p_{e,j} \equiv P[Q \geq \tau_j]$ as the exceedance probability corresponding to $tau_j$.

We can now define a downward recursion formula to estimate the threshold exceedance probabilities:

$$\hat{p}_{e,j} = \hat{p}_{e,j+1} + (\frac{A_j}{A_j + B_j})(1 - \hat{p}_{e,j+1}) \tag{4}$$

Noting that $\hat{p}_{e,m+1} = 0$, we can recursively evaluate $\{\hat{p}_{e,m}, \hat{p}_{e,m-1}, ..., \hat{p}_{e,1}\}$. The $A_j$ measured floods, $\{Q_1^j, Q_2^j, ..., Q_{A_j}^j\}$ in the range $[\tau_j, \tau_{j+1})$ can now be assigned plotting positions in $[\hat{p}_{e,j}, \hat{p}_{e,j+1})$ using a standard plotting position formula:

$$\hat{p}_{e,i}^j = (1 - \hat{p}_{e,j}) + (\hat{p}_{e,j} - \hat{p}_{e,j+1})\frac{i - a}{A_j + 1 - 2a} \tag{5}$$

where $a \equiv 0$ is the standard Weibull formula employed by B17B, $a \equiv 0.44$ corresponds to Gringorten, and $a \equiv 0.5$ corresponds to Hazen plotting positions.

## 9.3 Estimating the Mean Square Error of At-Site Skew

Efficient weighting of the regional and at-site skew requires an estimate of the mean square errors of both quantities. The MSE of the regional skew is provided by the skew map or skew estimation procedure; details are provided in Reis Jr et al. (2003). The MSE of the at-site skew can be estimated in several ways depending on the data available. Bulletin 17B (IACWD, 1982, p. 13) provides the following equations:

$$\begin{aligned} MSE_{\hat{\gamma}} &= 10^{A - B*log(n/10)/log(10)} && (6) \\ A &\equiv \begin{cases} -0.33 + 0.08 * abs(\hat{\gamma}) & abs(\hat{\gamma}) \le 0.9 \\ -0.52 + 0.3 * abs(\hat{\gamma}) & o.w. \end{cases} && (7) \\ B &\equiv \begin{cases} 0.94 - 0.26 * abs(\hat{\gamma}) & abs(\hat{\gamma}) \le 1.5 \\ -0.55 & o.w. \end{cases} && (8) \end{aligned}$$

where $n$ is the sample size. Griffis Griffis et al. (2004b) employs an equation originally found in Bobee (1973)

$$\begin{aligned} MSE_{\hat{\gamma}} &= (6/n + a) * (1 + (9/6 + b) * \hat{\gamma}^2 + (15/48 + c) * \hat{\gamma}^4) && (9) \\ a &\equiv -17.75/n^2 + 50.06/n^3 && (10) \\ b &\equiv 3.93/n^{0.3} - 30.97/n^{0.6} + 37.1/n^{0.9} && (11) \\ c &\equiv -6.16/n^{0.56} + 36.83/n^{1.12} - 66.9/n^{1.68} && (12) \end{aligned}$$

Although B17B recommends using $H$ – the length of the historical period – when historical information is present, this does not lead to satisfactory results when dealing with interval (censored) data. In the presence of censored data, both equations 6 and 9 are in general highly inaccurate, underestimating the MSE by a factor of 5 or more when the entire left-hand tail is censored.

The approach adopted here, which seems to work reasonably well, is to rely on first-order estimates of the MSE (Cohn et al., 2001, equation 55). These are never more than slightly biased (on the order of 15 percent high) for cases of interest, they are asyptotically unbiased, and they perform very well when interval data are present. Given that the only function of the MSE is to determine the weighting between the at-site and regional skew, both of which are assumed to be unbiased, employing a slightly biased estimator will not bias the frequency estimates.

## 9.4 Non-standard Features

Where *PeakfqSA* is run from a command-line, *PeakfqSA* checks to see if an additional argument have been entered on the command line immediately following *PeakfqSA*. If it exists, it is interpreted to be the name of the ".psf" file. This feature can be useful when running multiple sites.

## 9.5 Development History of *PeakfqSA*

All of the source code for *PeakfqSA* was developed and written by Tim Cohn of the U.S. Geological Survey. Over the years, Greg Baier, John England, and many others, have provided help testing and debugging, and have offered many suggestions about how to improve and better document the code. However, Tim Cohn remains responsible for any errors.

The source code for the *PeakfqSA* front-end routine amounts to approximately 1300 lines of code, and was written during the period 24-26 August 2007. The underlying computational software was written over approximately a dozen years starting in 1995 and ending in the present year. The subroutines and functions supporing EMA amount to approximately 45000 additional lines of code.

*PeakfqSA* is written in *Fortran-95*, though care has been taken to ensure backward compatibility with *Fortran-77*.

## 9.6 Example Specification File for *PeakfqSA*

```
#  This is the specification file for the PeakfqSA program written by
#  Tim Cohn, USGS, 23 Aug 2007
#
#  Format follows loosely upon the format of the PEAKFQ program specification
#  file.

I  BigSandy.dat           !   Input datafile for discharge data (note format)
STATION  BIG SANDY RIVER AT BRUCETON, TN, 1890-1973
LOTHRESH         0.0      !   Low-outlier threshold (if 0, GB crit. value used)
SKEWOPT          WEIGHTED !   Choices:  STATION WEIGHTED GENERALIZED GENERAL/VAR
GENSKEW          -0.5     !   Regional skew
SKEWSD           0.55     !   Std Dev of regional skew
BEGYEAR          1890     !   Beginning year of period of record (everything):
ENDYEAR          1973     !   Ending year of entire period of record
GAGEBASE            0     !
CSV              YES
#  Thresholds:  Note that it must include year in which threshold starts to
#  apply, year it ceases to apply, lower value on what would have been observed
#  during the interval, and upper value on what would have been observed
#  N.B.  Later PCPT\_THRESH statements supersede earlier ones where there is an
#        overlap
PCPT\_THRESH        1890 1929 18000 1.d99
PCPT\_THRESH        1930 1973 0.000 1.d99
```

## 9.7 Example Data File for ***PeakfqSA***

```
!   3602120881344004747017SW06040005205.00        380.58
!   BIG SANDY RIVER AT BRUCETON, TN
QINT   1895      0       18000 ! N.B. Not needed; PCPT\_THRESH
QINT   1896      0       18000 ! card already specifies this
Q      1897  25000
Q      1919  21000
Q      1927  18500
Q      1930   9100
Q      1931   2060
Q      1932   7820
Q      1933   3220
Q      1934   5580
Q      1935  17000
Q      1936   6740
Q      1937  13800
Q      1938   4270
Q      1939   5940
Q      1940   1680
Q      1941   1200
Q      1942  10100
Q      1943   3780
Q      1944   5340
Q      1945   5630
Q      1946  12000
Q      1947   3980
Q      1948   6130
Q      1949   4740
Q      1950   9880
Q      1951   5230
Q      1952   4260
Q      1953   5000
Q      1954   3320
Q      1955   5480
Q      1956  11800
Q      1957   5150
Q      1958   3350
Q      1959   2400
Q      1960   1460
Q      1961   3770
Q      1962   7480
Q      1963   2740
Q      1964   3100
```

```
Q       1965   7180
Q       1966   1920
Q       1967   9060
Q       1968   3080
Q       1969   2800
Q       1970   4330
Q       1971   5080
Q       1972  12000
Q       1973   7640
```

## 9.8 Example Output File for ***PeakfqSA***

```
********************************************************************************


                                EMA ANALYSIS
                         Computed by PeakfqSA 0.910
                               Tim Cohn, USGS
                              tacohn@usgs.gov
                                703/648-5711

********************************************************************************


                    BIG SANDY RIVER AT BRUCETON, TN, 1890-1973

                          Spec. file         BigSandy_hist.psf
                          Data file          BigSandy.dat

                          Historical Period Begins         1890
                          Historical Period Ends           1973
                          Length of Historical Period        84
                          Total No. of Observations          84
                          Years Without Information           0
                          Number of Peaks                    47
                          Peaks Not Used                      0
                          Systematic Peaks                   44
                          Historic Peaks                      3
                          Years of Historic Record           84
                          Skew Option                     WEIGHTED
                          Generalized Skew                -0.50
                          Standard Error                 0.5500
                          (Pseudo)-MSE Employed       0.303
                          Gage Base Discharge               0.0
                          Low-Outlier Bound           Iterated GB
                          Iteration                           1
                          - No. LOs in Iter. 1                0
                          - Crit. Val. Iter. 1            838.6
                          - Thresh. Val. Iter. 1         1200.0
                          Number of Low Outliers              0


                 ANNUAL FREQUENCY CURVE PARAMETERS  -- LOG-PEARSON TYPE III
```

```
         Log10 Moments                           M         S         G
         -----------------------------------------------------------------
         EMA, At-Site Data, w/o Reg. Skew     3.715691  0.287326  0.001893
         EMA, At-Site Data, w/Reg. Skew       3.717272  0.289200 -0.118702


                                  EMA FREQUENCY ESTIMATES

-----------------------------------------------------------------------------
Ann. Exc. Prob.  EMA Est.                               Ci-Low      CI-High
-----------------------------------------------------------------------------
EMA-Q 0.9950       871.25                                 268.16       1336.76
EMA-Q 0.9900      1045.59                                 390.27       1524.62
EMA-Q 0.9500      1706.18                                 980.17       2249.92
EMA-Q 0.9000      2203.77                                1483.72       2814.62
EMA-Q 0.8000      2990.15                                2256.78       3727.25
EMA-Q 0.6667      3957.50                                3131.46       4872.08
EMA-Q 0.5000      5284.36                                4264.02       6472.14
EMA-Q 0.4292      5948.01                                4819.28       7278.88
EMA-Q 0.2000      9166.15                                7447.82      11237.90
EMA-Q 0.1000     12134.65                                9766.00      15218.32
EMA-Q 0.0400     16276.60                               12842.97      22107.11
EMA-Q 0.0200     19617.73                               15154.99      29124.18
EMA-Q 0.0100     23158.65                               17388.03      37986.08
EMA-Q 0.0050     26912.12                               19506.95      48956.47
EMA-Q 0.0020     32217.14                               22116.98      67352.73


********************************************************************************

1


                                PERCEPTION THRESHOLDS FOR EMA
                                -----------------------------

          -------------------------------------------------------------
           T   Beg. Year End Year              T_lower        T_upper
          -------------------------------------------------------------
           1        1890     1929             18000.        1.00000E+50
           2        1930     1973             0.0000        1.00000E+50
```

```
                      EMA REPRESENTATION OF DATA
                      --------------------------

            BIG SANDY RIVER AT BRUCETON, TN, 1890-1973

----------------------------------------------------------------
Year        Q_lower        Q_upper      T_lower        T_upper
----------------------------------------------------------------
1890            0.0        18000.0    18000.         0.10000E+11
                         . . .
1896            0.0        18000.0    18000.         0.10000E+11
1897        25000.0        25000.0    18000.         0.10000E+11
1898            0.0        18000.0    18000.         0.10000E+11
                         . . .
1918            0.0        18000.0    18000.         0.10000E+11
1919        21000.0        21000.0    18000.         0.10000E+11
1920            0.0        18000.0    18000.         0.10000E+11
                         . . .
1926            0.0        18000.0    18000.         0.10000E+11
1927        18500.0        18500.0    18000.         0.10000E+11
1928            0.0        18000.0    18000.         0.10000E+11
1929            0.0        18000.0    18000.         0.10000E+11
1930         9100.0         9100.0    838.57         0.10000E+11
1931         2060.0         2060.0    838.57         0.10000E+11
1932         7820.0         7820.0    838.57         0.10000E+11
1933         3220.0         3220.0    838.57         0.10000E+11
1934         5580.0         5580.0    838.57         0.10000E+11
1935        17000.0        17000.0    838.57         0.10000E+11
1936         6740.0         6740.0    838.57         0.10000E+11
1937        13800.0        13800.0    838.57         0.10000E+11
1938         4270.0         4270.0    838.57         0.10000E+11
1939         5940.0         5940.0    838.57         0.10000E+11
1940         1680.0         1680.0    838.57         0.10000E+11
1941         1200.0         1200.0    838.57         0.10000E+11
1942        10100.0        10100.0    838.57         0.10000E+11
1943         3780.0         3780.0    838.57         0.10000E+11
1944         5340.0         5340.0    838.57         0.10000E+11
1945         5630.0         5630.0    838.57         0.10000E+11
1946        12000.0        12000.0    838.57         0.10000E+11
1947         3980.0         3980.0    838.57         0.10000E+11
1948         6130.0         6130.0    838.57         0.10000E+11
1949         4740.0         4740.0    838.57         0.10000E+11
1950         9880.0         9880.0    838.57         0.10000E+11
```

```
1951         5230.0         5230.0   838.57         0.10000E+11
1952         4260.0         4260.0   838.57         0.10000E+11
1953         5000.0         5000.0   838.57         0.10000E+11
1954         3320.0         3320.0   838.57         0.10000E+11
1955         5480.0         5480.0   838.57         0.10000E+11
1956        11800.0        11800.0   838.57         0.10000E+11
1957         5150.0         5150.0   838.57         0.10000E+11
1958         3350.0         3350.0   838.57         0.10000E+11
1959         2400.0         2400.0   838.57         0.10000E+11
1960         1460.0         1460.0   838.57         0.10000E+11
1961         3770.0         3770.0   838.57         0.10000E+11
1962         7480.0         7480.0   838.57         0.10000E+11
1963         2740.0         2740.0   838.57         0.10000E+11
1964         3100.0         3100.0   838.57         0.10000E+11
1965         7180.0         7180.0   838.57         0.10000E+11
1966         1920.0         1920.0   838.57         0.10000E+11
1967         9060.0         9060.0   838.57         0.10000E+11
1968         3080.0         3080.0   838.57         0.10000E+11
1969         2800.0         2800.0   838.57         0.10000E+11
1970         4330.0         4330.0   838.57         0.10000E+11
1971         5080.0         5080.0   838.57         0.10000E+11
1972        12000.0        12000.0   838.57         0.10000E+11
1973         7640.0         7640.0   838.57         0.10000E+11


               SORTED OBSERVATIONS
       (Hirsch/Stedinger Plotting Positions)
       -------------------------------------


-----------------------------------------------------
Year Plot Pos     Obs. Q      Fit Value Q      % Diff
-----------------------------------------------------
1897   0.0089    25000.         23758.          -5.23
1919   0.0179    21000.         20183.          -4.05
1927   0.0268    18500.         18186.          -1.73
1935   0.0571    17000.         14630.         -16.20
1937   0.0786    13800.         13197.          -4.57
1972   0.1000    12000.         12135.           1.11
1946   0.1214    12000.         11292.          -6.27
1956   0.1429    11800.         10593.         -11.39
1942   0.1643    10100.         9998.0          -1.02
1950   0.1857    9880.0         9478.8          -4.23
1930   0.2071    9100.0         9018.4          -0.90
```

```
1967   0.2286   9060.0         8604.6         -5.29
1932   0.2500   7820.0         8228.6          4.97
1973   0.2714   7640.0         7883.8          3.09
1962   0.2929   7480.0         7565.2          1.13
1965   0.3143   7180.0         7268.7          1.22
1936   0.3357   6740.0         6991.3          3.59
1948   0.3571   6130.0         6730.3          8.92
1939   0.3786   5940.0         6483.6          8.38
1945   0.4000   5630.0         6249.5          9.91
1934   0.4214   5580.0         6026.4          7.41
1955   0.4429   5480.0         5813.2          5.73
1944   0.4643   5340.0         5608.6          4.79
1951   0.4857   5230.0         5411.7          3.36
1957   0.5071   5150.0         5221.7          1.37
1971   0.5286   5080.0         5037.9         -0.84
1953   0.5500   5000.0         4859.4         -2.89
1949   0.5714   4740.0         4685.8         -1.16
1970   0.5929   4330.0         4516.4          4.13
1938   0.6143   4270.0         4350.6          1.85
1952   0.6357   4260.0         4188.0         -1.72
1947   0.6571   3980.0         4028.1          1.19
1943   0.6786   3780.0         3870.3          2.33
1961   0.7000   3770.0         3714.2         -1.50
1958   0.7214   3350.0         3559.1          5.88
1954   0.7429   3320.0         3404.6          2.49
1933   0.7643   3220.0         3250.0          0.92
1964   0.7857   3100.0         3094.6         -0.17
1968   0.8071   3080.0         2937.5         -4.85
1969   0.8286   2800.0         2777.7         -0.80
1963   0.8500   2740.0         2613.9         -4.83
1959   0.8714   2400.0         2444.1          1.80
1931   0.8929   2060.0         2265.8          9.08
1966   0.9143   1920.0         2074.8          7.46
1940   0.9357   1680.0         1864.1          9.88
1960   0.9571   1460.0         1619.3          9.84
1941   0.9786   1200.0         1300.3          7.71
```

# References

Bobee, B. (1973). Sample error of T-year events computed by fitting a Pearson Type 3 Distribution. *Water Resources Research*, 9(5):1264–1270.

Cohn, T. A., England, J. F., Berenbrock, C. E., Mason, R. R., Stedinger, J. R., and Lamontagne, J. R. (2013). A generalized Grubbs-Beck test statistic for detecting multiple potentially influential low outliers in flood series. *Water Resources Research*, 49(8):5047–5058.

Cohn, T. A., Lane, W. L., and Baier, W. G. (1997). An algorithm for computing moments-based flood quantile estimates when historical flood information is available. *Water Resources Research*, 33(9):2089–2096.

Cohn, T. A., Lane, W. L., and Stedinger, J. R. (2001). Confidence intervals for Expected Moments Algorithm flood quantile estimates. *Water Resources Research*, 37(6):1695–1706.

Cohn, T. A. and Stedinger, J. R. (1987). Use of historical flood information in a maximum likelihood framework. *Journal of Hydrology*, 96:215–223.

England, J. F., Jarrett, R. D., and Salas, J. D. (2003a). Data-based comparisons of moments estimators using historical and paleoflood data. *Journal of Hydrology*, 278(1):172–196.

England, J. F., Salas, J. D., and Jarrett, R. D. (2003b). Comparisons of two moments-based estimators that utilize historical and paleoflood data for the log Pearson type III distribution. *Water Resources Research*, 39(9):1243.

Flynn, K. M., Kirby, W. H., and Hummel, P. (2006). Users Manual for PeakFQ, Annual Flood Frequency Analysis Using Bulletin 17B Guidelines: U.S. Geological Survey Techniques and Methods Chapter 4 of Book 4, Section B.

Griffis, V. and Stedinger, J. (2007). Editorial on Why We Need to Update Bulletin 17b. *Journal of Hydrologic Engineering*, xx:yyy.

Griffis, V. W., Stedinger, J. R., and Cohn, T. A. (2004a). Extension of EMA to Address Regional Skew and Low Outliers. *World Water & Environmental Resources Congress 2003 and Related Symposia*.

Griffis, V. W., Stedinger, J. R., and Cohn, T. A. (2004b). Log Pearson type 3 quantile estimators with regional skew information and low outlier adjustments. *Water Resources Research*, 40(7).

Hirsch, R. M. and Stedinger, J. R. (1987). Plotting positions for historical floods and their precision. *Water Resources Research*, 23(4):715–727.

Hupp, C. R. (1987). Botanical evidence of floods and paleoflood history. In Singh, V. P., editor, *Regional flood frequency analysis*, pages 355–369. D. Reidel Publishing Company.

IACWD (1982). Guidelines for determining flood flow frequency, Bulletin 17B. Technical report, Interagency Committee on Water Data, Hydrology Subcommittee.

Jarrett, R. D. and Malde, H. E. (1987). Paleodischarge of the late Pleistocene Bonneville flood, Snake River, Idaho, computed from new evidence. *Geological Society of America Bulletin*, 99:127–134.

Jarrett, R. D. and Tomlinson, E. M. (2000). Regional interdisciplinary paleoflood approach to assess extreme flood potential. *Water Resources Research*, 36(10):2957–2984.

Reis Jr, D., Stedinger, J., and Martins, E. (2003). Bayesian GLS Regression with Application to LP3 Regional Skew Estimation. *Proceedings, World Water and Environmental Resources Congress 2003*, pages 23–26.

Stedinger, J. R. and Baker, V. R. (1987). Surface water hydrology: Historical paleoflood information. *Reviews of Geophysics*, 25:119–124.

Stedinger, J. R. and Cohn, T. A. (1986). Flood frequency analysis with historical and paleoflood information. *Water Resources Research*, 22(5):785–793.

Stedinger, J. R. and Cohn, T. A. (1987). Historical flood-frequency data: Its value and use. In Singh, V. P., editor, *Regional flood frequency analysis*, pages 273–286. D. Reidel Publishing Company.

Wohl, E. (1995). Estimating Flood Magnitude in Ungauged Mountain Channels, Nepal. *Mountain Research and Development*, 15(1):69–76.