Mixture Methods for the Analysis of Flood Frequency Distributions

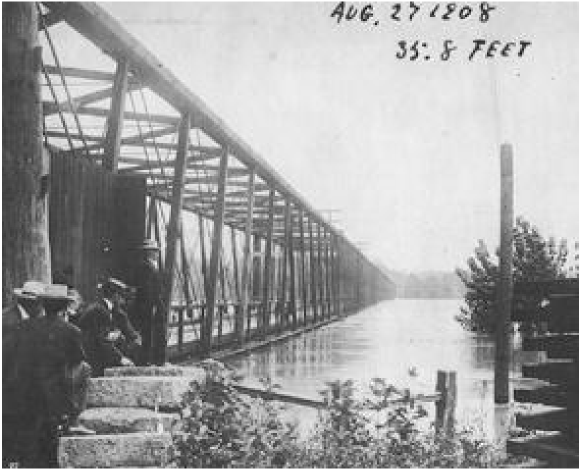
Philip Yates [1] with John Grego [2]

[1]Department of Mathematics & Statistics
California State Polytechnic University, Pomona
payates@csupomona.edu
[2]Department of Statistics
University of South Carolina
grego@stat.sc.edu

June 24, 2008

Congaree River: Columbia, SC

Congaree River: Columbia, SC



AUG, 27 1908
35'. 8 FEET

Flood Frequency Data

Annual Peak Flows

The maximum momentary peak discharge in each year of record

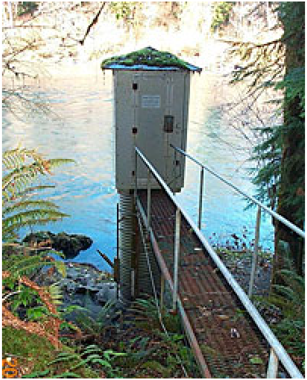
- The year of record begins the previous October and ends in September of the current year
- Peak flows are measured in cubic feet per second (cfs)

Historical Flood Data

Any observation of flood stage or conditions made before actual flood data were collected systematically

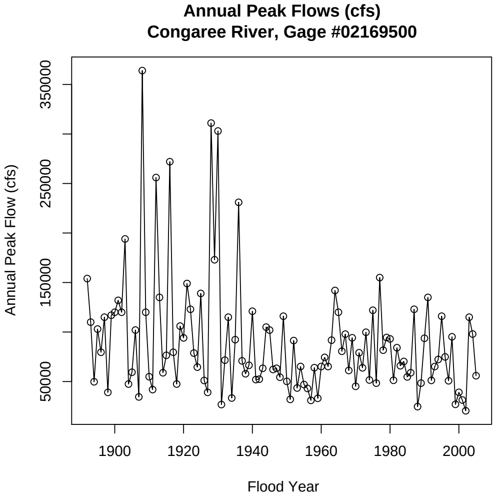
- Data collected from old newspapers, diaries, museums, libraries, etc.

River Gage

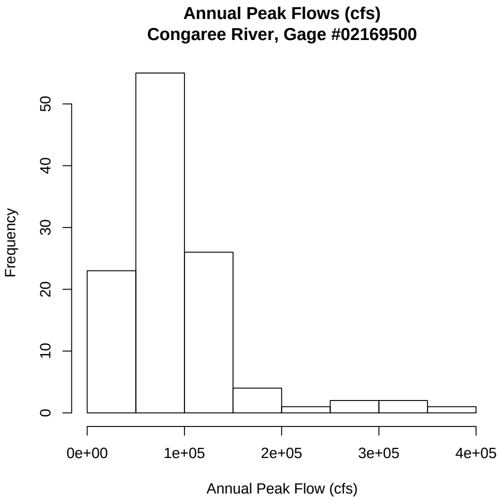


Gage #12041200: Hoh River at U.S. Highway 101 near Forks, WA

Congaree River, Gage #02169500



Congaree River, Gage #02169500



1865: Columbia, SC

- Ruins, as seen from the State House, 1865.
- General Sherman's Union troops were slowed entering Columbia by a major flood on the Congaree River.

Modeling Flood Frequency Data

Annual peak flows are assumed to be independent and identically distributed.

Let $Y = \log_{10} X$, where X is the annual peak flow.

$$f_Y(y) = \frac{(y-\gamma)^{\alpha-1}\exp[-(y-\gamma)/\beta]}{\beta^\alpha \Gamma(\alpha)},$$

$\alpha > 0, \beta > 0, y > \gamma$, where:

α — the shape parameter
β — the scale parameter
γ — the shift or location parameter

Modeling Flood Frequency Data

Bulletin #17B (1981)

- FEMA uses it as the trade standard
- Method of moments used to estimate parameters of log-Pearson type III distribution
- When dealing with mixed populations, splits data into number of groups and fits a separate curve for each group
- When collecting annual peak flow data, estimates the 99th percentile of the log-Pearson type III distribution; this is used as an estimate for 100-year flood — FEMA uses this in regulatory policy for floodplains

Other Works

American River flood frequency analyses (1999)

- Used both systematic annual peak flows and one historical flood observation from 1862
- Used EMA, the Expected Moments Algorithm (Cohn, et al., 1997) to fit the data to a log-Pearson type III distribution
- Recommended the independent identically distributed approach to flood estimation
- Tentative use of mixed models as remedy for certain types of non-stationarity in the flood data

Other Works (Mixtures)

Singh (1987a,b) used a mixture of two normal distributions and two lognormal distributions in order to consider an observed flood that is composed of two component distributions

Hirschboeck (1987) & Diehl and Potter (1987) discussed various reasons to believe that there may be two or more components to the observed flood's distribution

- in the southeastern region of US, tropical storms
- in the western, and even midwestern and northeastern, regions of US, snowmelt

Planned Work

- Investigate methods to find parameter estimates for a mixture of normals
- Develop an estimate of the standard error for the 99th percentile of a finite mixture model with two components
- Use these methods on the Congaree and Broad flood data in order to find the 100-year flood estimate (the 99th percentile of the fitted distribution)
- Compare these results to methods of standard error estimation in Bulletin #17B and Cox et. al. (2002)

Finite Mixture Model

Suppose Y is a random variable or vector that takes values from sample space $\mathcal{Y}$.

$$p(y) = \pi_1 f_1(y) + \ldots + \pi_k f_k(y) \quad (y \in \mathcal{Y}),$$

where

$$\pi_j > 0, \quad j = 1, \ldots, k; \quad \pi_1 + \ldots + \pi_k = 1$$

and

$$f_j(\cdot) \geq 0, \quad \int_{\mathcal{Y}} f_j(y)\,\mathrm{d}x = 1, \quad j = 1, \ldots, k,$$

then Y has a finite mixture distribution

Finite Mixture Model

Likelihood Function for Finite Mixture Model

$$L(\psi) = \prod_{i=1}^{n} p(y_i|\psi) = \prod_{i=1}^{n} \left[\sum_{j=1}^{k} \pi_j f(y_i|\boldsymbol{\theta}_j) \right]$$

$\pi_j f(y_i|\boldsymbol{\theta}_j)$ component keeps track of y_i's contribution to entire mixture density

If it was known which component of the mixture the observation was from, the data would be fully categorized:

Complete Data Likelihood

$$\{x_i, i = 1, \ldots, n\} = \{(y_i, \mathbf{z}_i); i = 1, \ldots, n\},$$

where

$\mathbf{z}_i = (z_{ij} = 1, \ldots, k)$ — indicator vector of length k with 1 in the position corresponding to the appropriate component and zeros elsewhere

$\mathbf{x}$ — complete data (not observed directly)

$\mathbf{y}$ — incomplete data ($\mathbf{x}$ is indirectly observed via $\mathbf{y}$)

Complete Data Likelihood

Complete Data Likelihood

$$g(x_1, \ldots, x_n|\psi) = \prod_{i=1}^{n} \prod_{j=1}^{k} \pi_j^{z_{ij}} f_j(y_i|\theta_j)^{z_{ij}}$$

Complete Data Log-likelihood

$$l(\psi) = \sum_{i=1}^{n} \mathbf{z}_i^{\mathbf{T}} \mathbf{V}(\boldsymbol{\pi}) + \sum_{i=1}^{n} \mathbf{z}_i^{\mathbf{T}} \mathbf{U}_i(\boldsymbol{\theta}),$$

where $\log \pi_j$ is the j^{th} component of $\mathbf{V}(\boldsymbol{\pi})$ and $\log f_j(y_i|\boldsymbol{\theta}_j)$ is the j^{th} component of $\mathbf{U}_i(\boldsymbol{\theta})$

Observable Log-Likelihood

Define for each i and j

$$\mathbf{w}_i = w_{ij} = E(\mathbf{z}_i|y_i, \psi') = \frac{\pi'_j f_j(y_i|\boldsymbol{\theta}'_j)}{p(y_i|\psi')}$$

Observable Log-Likelihood Function

$$Q(\psi|\psi') = E(\log g(\mathbf{x}|\psi)|\mathbf{y}, \psi') =$$
$$\sum_{i=1}^{n} \mathbf{w}_i^{\mathbf{T}} \mathbf{V}(\pi) + \sum_{i=1}^{n} \mathbf{w}_i^{\mathbf{T}} \mathbf{U}_i(\theta)$$

Mixture of Two Normals

$$f(y_i|\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, \pi) = (1-\pi)\phi\left(\frac{y_i - \mu_0}{\sqrt{\sigma_0^2}}\right) + \pi\phi\left(\frac{y_i - \mu_1}{\sqrt{\sigma_1^2}}\right),$$

To implement ECM (Meng & Rubin, 1993) algorithm, let

$$w_i = \frac{\pi\phi\left(\frac{y_i-\mu_1}{\sqrt{\sigma_1^2}}\right)}{(1-\pi)\phi\left(\frac{y_i-\mu_0}{\sqrt{\sigma_0^2}}\right) + \pi\phi\left(\frac{y_i-\mu_1}{\sqrt{\sigma_1^2}}\right)}.$$

Mixture of Two Normals

E-step

$$Q(\psi|\psi^{(p)}) = \sum_{i=1}^{n} w_i^{(p)} \log(\pi) +$$

$$(1 - w_i^{(p)}) \log(1 - \pi) + w_i^{(p)} \log \phi \left(\frac{y_i - \mu_1}{\sqrt{\sigma_1^2}} \right) +$$

$$(1 - w_i^{(p)}) \log \phi \left(\frac{y_i - \mu_0}{\sqrt{\sigma_0^2}} \right).$$

Mixture of Two Normals

CM-steps

$$\begin{aligned}
\pi^{(p+1)} &= \frac{1}{n}\sum_{i=1}^{n} w_i^{(p)} \\
\mu_0^{(p+1)} &= \frac{\sum_{i=1}^{n}(1-w_i^{(p)})y_i}{\sum_{i=1}^{n}(1-w_i^{(p)})} \\
\mu_1^{(p+1)} &= \frac{\sum_{i=1}^{n} w_i^{(p)} y_i}{\sum_{i=1}^{n} w_i^{(p)}}
\end{aligned}$$

Mixture of Two Normals

CM-steps

$$\sigma_0^{2(p+1)} = \frac{\sum_{i=1}^n (1 - w_i^{(p)})(y_i - \mu_0^{(p+1)})^2}{\sum_{i=1}^n (1 - w_i^{(p)})}$$

$$\sigma_1^{2(p+1)} = \frac{\sum_{i=1}^n w_i^{(p)}(y_i - \mu_1^{(p+1)})^2}{\sum_{i=1}^n w_i^{(p)}}$$

100-year Flood Estimation

To find the 100-year flood when the distribution is from a finite mixture model, one needs to find the 99th percentile of the finite mixture distribution. For example, if the distribution was a finite mixture model of normal densities, then the 100-year flood, Q, is

$$\int_{-\infty}^{\log_{10} Q} (1-\pi) f_0(y|\mu_0, \sigma_0^2) + \pi f_1(y|\mu_1, \sigma_1^2) dy = 0.99,$$

where $f_j(y) \sim N(\mu_j, \sigma_j^2)$ and $\log_{10} Q$ is the logged value of the 100-year flood.

To find $\log_{10} \hat{Q}$, first obtain the 99th percentile from each component of the mixing distribution.

100-year Flood Estimation

These percentiles from each mixing component are used as end points to search for the root of

$$\int_{-\infty}^{\log_{10}\hat{Q}} (1-\pi)f_0(y|\mu_0,\sigma_0^2) + \pi f_1(y|\mu_1,\sigma_1^2)dy - 0.99 = 0.$$

In order to obtain a $100(1-\alpha)\%$ confidence interval for $\log_{10} Q$ in this example, a delta method argument can be made to find the standard error of $\log_{10} Q$

100-year Flood Estimation

Define the quantile function for a finite mixture model with two normal densities as:

$$\begin{aligned}\Gamma(\boldsymbol{\theta}, x(\boldsymbol{\theta}, 0.99)) &= \int_{-\infty}^{x} \left\{ \tau \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}(y-\mu_1)^2} + (1-\tau)\frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2\sigma_0^2}(y-\mu_0)^2} \right\} dy \\ &= 0.99,\end{aligned}$$

with the chain rule yielding

$$\nabla_\tau x(\boldsymbol{\theta}, 0.99) = -\frac{\nabla_\tau \Gamma(\boldsymbol{\theta}, x)|_{x=x(\boldsymbol{\theta},0.99)}}{\nabla_x \Gamma(\boldsymbol{\theta}, x)|_{x=x(\boldsymbol{\theta},0.99)}} \quad \text{Uryasev (2000)},$$

where $\boldsymbol{\theta}^{\mathbf{T}} = (\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, \tau)^{\mathbf{T}}$.

100-year Flood Estimation

The numerator of $\nabla_\tau x(\boldsymbol{\theta}, 0.99)$ is,

$$\begin{aligned}
\nabla_\tau \Gamma(\boldsymbol{\theta}, x)|_{x=x(\boldsymbol{\theta},0.99)} &= \int_{-\infty}^{x} \left\{ \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}(y-\mu_1)^2} - \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2\sigma_0^2}(y-\mu_0)^2} \right\} dy \\
&= \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}(y-\mu_1)^2} dy - \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2\sigma_0^2}(y-\mu_0)^2} dy \\
&= \Phi\left(\frac{x(\boldsymbol{\theta}, 0.99) - \mu_1}{\sigma_1}\right) - \Phi\left(\frac{x(\boldsymbol{\theta}, 0.99) - \mu_0}{\sigma_0}\right),
\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function for a standard normal probability distribution.

100-year Flood Estimation

The numerator for the gradient of this quantile function with respect to μ_0 and μ_1 is

$$\begin{aligned}
\nabla_{\mu_0}\Gamma(\boldsymbol{\theta},x)\big|_{x=x(\boldsymbol{\theta},0.99)} &= \int_{-\infty}^{x} \frac{(1-\tau)(y-\mu_0)}{\sqrt{2\pi}\sigma_0^3} e^{-\frac{1}{2\sigma_0^2}(y-\mu_0)^2} dy \\
\nabla_{\mu_1}\Gamma(\boldsymbol{\theta},x)\big|_{x=x(\boldsymbol{\theta},0.99)} &= \int_{-\infty}^{x} \frac{\tau(y-\mu_1)}{\sqrt{2\pi}\sigma_1^3} e^{-\frac{1}{2\sigma_1^2}(y-\mu_1)^2} dy
\end{aligned}$$

100-year Flood Estimation

The numerator for the gradient of this quantile funciton with respect to σ_0^2 and σ_1^2 is

$$\begin{aligned}
\nabla_{\sigma_0^2}\Gamma(\boldsymbol{\theta},x)\big|_{x=x(\boldsymbol{\theta})} &= \int_{-\infty}^{x}\left\{\frac{1-\tau}{2\sqrt{2\pi}\sigma_0^3}e^{-\frac{1}{2\sigma_0^2}(y-\mu_0)^2}\left(\frac{(y-\mu_0)^2}{\sigma_0^2}-1\right)\right\}dy\\
\nabla_{\sigma_1^2}\Gamma(\boldsymbol{\theta},x)\big|_{x=x(\boldsymbol{\theta})} &= \int_{-\infty}^{x}\left\{\frac{\tau}{2\sqrt{2\pi}\sigma_1^3}e^{-\frac{1}{2\sigma_1^2}(y-\mu_1)^2}\left(\frac{(y-\mu_1)^2}{\sigma_1^2}-1\right)\right\}dy.
\end{aligned}$$

100-year Flood Estimation

The denominator for the gradient of this quantile function with respect to any of the parameters in $\boldsymbol{\theta}$ is,

$$\begin{aligned}
\nabla_x \Gamma(\boldsymbol{\theta}, x)\big|_{x=x(\boldsymbol{\theta},0.99)} &= p(x(\tau, 0.99)) \\
&= \tau \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{1}{2\sigma_1^2}(x(\boldsymbol{\theta},0.99)-\mu_1)^2} \\
&\quad +(1-\tau)\frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2\sigma_0^2}(x(\boldsymbol{\theta},0.99)-\mu_0)^2}.
\end{aligned}$$

100-year Flood Estimation

$100(1-\alpha)\%$ Confidence Interval for $\log_{10} Q$

$$\log_{10}\hat{Q} \pm z_{\frac{\alpha}{2}} \left(\nabla \log_{10}\hat{Q}^{\mathsf{T}} I_Y(\hat{\psi})^{-1} \nabla \log_{10}\hat{Q} \right)^{1/2},$$

where I_Y is the observed information matrix.

Observed Information Matrix

Observed Information Matrix

$$
\begin{aligned}
\mathbf{I}_Y(\psi) = {} & E(\mathbf{B}(\mathbf{x}|\psi)|\mathbf{y},\psi) \\
& -E(\mathbf{S}(\mathbf{x}|\psi)\mathbf{S}(\mathbf{x}|\psi)^{\mathbf{T}}|\mathbf{y},\psi) + \mathbf{S}^*(\mathbf{y}|\psi)\mathbf{S}^*(\mathbf{y}|\psi)^{\mathbf{T}}.
\end{aligned}
$$

This is the conditional expected complete data information matrix minus the expected information matrix for the conditional distribution of the complete data, $\mathbf{x}$, given the incomplete data, $\mathbf{y}$, and ψ.

Congaree River

Bulletin #17B

100-year flood is estimated by

$$\log_{10} \hat{Q} = \bar{y} + K_{.01} s$$

where $K_{.01}$ is a Pearson Type III deviate

For the Congaree River, Gage #02169500:

$$\log_{10} Q = \bar{y} + K_{.01} s = 4.884468 + 2.541867 \times 0.241563 = 5.498489$$

This gives a point estimate for the 100-year flood, Q, of 315,129 cfs

Congaree River

Confidence intervals in Bulletin #17B approximate a noncentral *t*-distribution. For ease of computation, the following formulas are suggested as a large sample approximation to the noncentral *t*-distribution:

$$\begin{aligned} K^{U}_{.01,\frac{\alpha}{2}} &= \frac{K_{.01} + \sqrt{K^2_{.01} - ab}}{a} \\ K^{L}_{.01,\frac{\alpha}{2}} &= \frac{K_{.01} - \sqrt{K^2_{.01} - ab}}{a} \end{aligned}$$

where

$$\begin{aligned} a &= 1 - \frac{z^2_{\frac{\alpha}{2}}}{2(n-1)} \\ b &= K^2_{.01} - \frac{z^2_{\frac{\alpha}{2}}}{n} \end{aligned}$$

Congaree River

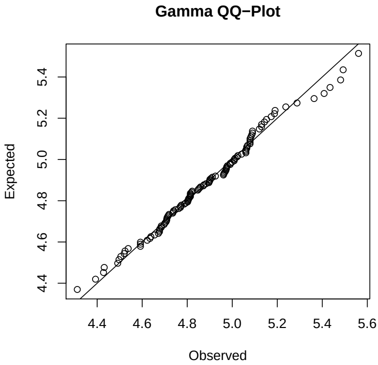
To obtain the 90% confidence interval of the 100-year flood for the Congaree River at gage #02169500, first use $z_{0.05} = 1.645$ to get $K^L_{.01,.05} = 2.252476$ and $K^U_{.01,.05} = 2.892320$.

$$\log_{10} Q^L = \bar{y} + K^L_{.01,.05} s = 5.428582$$

$$\log_{10} Q^U = \bar{y} + K^U_{.01,.05} s = 5.583145.$$

With a confidence level of 90%, the 100-year flood for the Congaree River at gage #02169500 is between 268,276 cfs and 382,953 cfs.

Congaree River



Toward the upper tail of the distribution, the MOMs are not fitting the $\log_{10}$ of the Congaree River's annual peak flows well

Congaree River

Gumbel Probability Density Function

$$f_Y(y|\mu,\beta) = \frac{e^{-(y-\mu)/\beta} e^{-e^{-(y-\mu)/\beta}}}{\beta},$$

$-\infty < \mu < \infty$; $\beta > 0$; $-\infty < y < \infty$

μ is the location parameter; β is the scale parameter.

The Gumbel distribution is used to find the maximum (or minimum) of a number of samples of various distribution.

Congaree River

Using a Gumbel distribution on the non-transformed annual peak flood flows:

Gumbel MLE's $\hat{\mu} = 67050.89$, $\hat{\beta} = 35747.7$

Point Estimate The 99th percentile is 231,496 cfs.

To find the standard error for a quantile from a Gumbel pdf, Cox, Isham, & Northrop (2002) suggest:

$$nV(\hat{Q}) = \beta^2 \left\{ 1 + \frac{(r_p + \psi(2))^2}{1 + \psi'(2)} \right\},$$

where $r_p = -\ln\{-\ln(1-p)\}$ and $\psi(\cdot)$ is the digamma function.

Congaree River

Using a Gumbel distribution on the non-transformed annual peak flood flows:

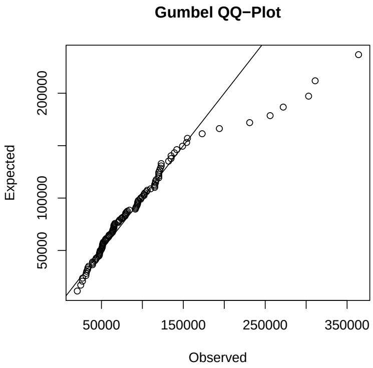
Gumbel MLE's $\hat{\mu} = 67050.89$, $\hat{\beta} = 35747.7$

Point Estimate The 99th percentile is 231,496 cfs.

Standard Error Estimate 13,474 cfs.

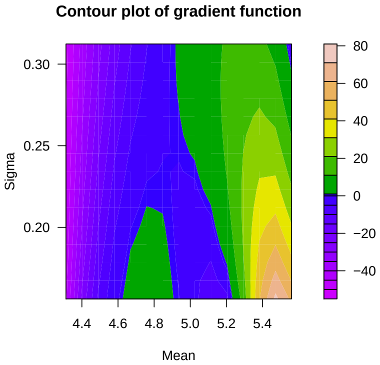
90% CI Q: 209,331 to 253,660 cfs.

Congaree River



The QQ-plot indicates that a Gumbel does not seem to handle the large flood events very well.

Congaree River



The results here suggest two mixing components are needed for the $\log_{10}$ of the Congaree River's annual peak flows

Congaree River

Mixture of Normals $\hat{\mu_0} = 4.859969$, $\hat{\sigma_0^2} = 0.0451328$,
$\hat{\mu_1} = 5.471577$, $\hat{\sigma_1^2} = 0.00341794$, $\hat{\pi} = 0.04005613$

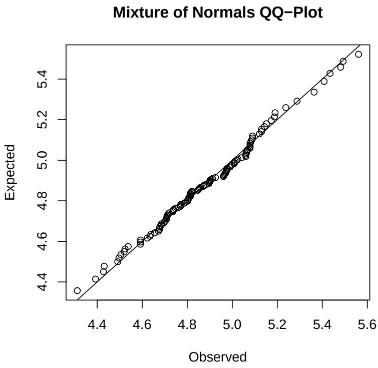
Point Estimate $\log_{10} \hat{Q} = 5.515678$; $\hat{Q} = 327,852$ cfs

Standard Error Estimate 0.03545455

90% CI Q: 286,652 cfs to 374,974 cfs

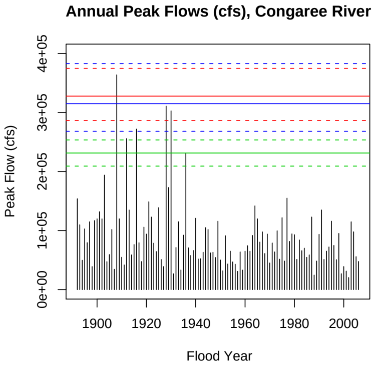
This interval for the 100-year flood is smaller than the one produced by Bulletin #17B.

Congaree River



The QQ-plot indicates that a mixture of normals handles observations of the lower and upper tails of the flood distribution better than Bulletin #17B

Congaree River



The point estimate and 90% CI of Q using the finite mixture model (red), Gumble (green), and Bulletin #17B (blue).

Broad River

For the Broad River, Gage #02161000

Bulletin #17B

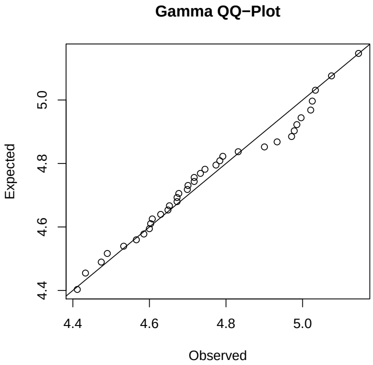
$$\begin{aligned}\log_{10} Q &= \bar{y} + K_{.01} s \\ &= 4.751704 + 2.525838 \times 0.1948259 \\ &= 5.243802\end{aligned}$$

Point Estimate $\log_{10} \hat{Q} = 5.243802$; $\hat{Q} = 175,308$ cfs

90% CI $\log_{10} Q$: (5.150279, 5.37576)

90% CI Q: 141,345 cfs to 237,553 cfs

Broad River



Just before the upper tail of the distribution, the MOMs are not fitting the $\log_{10}$ of the Broad River's annual peak flows well

Broad River

Using a Gumbel distribution on the non-transformed annual peak flood flows:

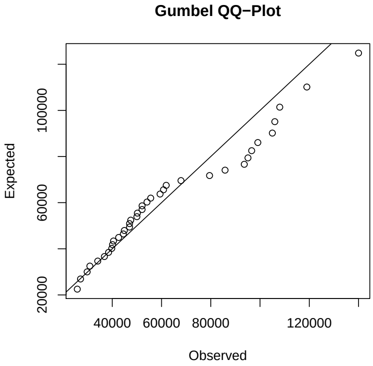
Gumbel MLE's $\hat{\mu} = 49395.73$, $\hat{\beta} = 20818.66$

Point Estimate The 99th percentile is 145,165 cfs.

Standard Error Estimate 15,195 cfs.

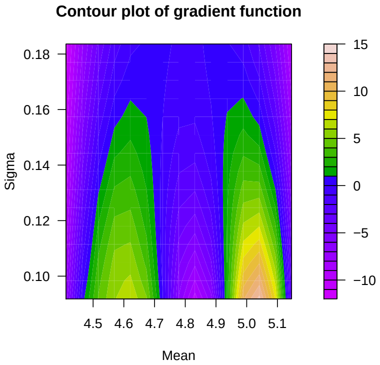
90% CI Q: 120,170 to 170,160 cfs.

Broad River



The QQ-plot indicates that a Gumbel does not seem to handle the large flood events very well.

Broad River



The results here suggest two mixing components are needed for the $\log_{10}$ of the Broad River's annual peak flows

Broad River

Mixture of Normals $\hat{\mu_0} = 4.65069$, $\hat{\sigma_0^2} = 0.01350768$,
$\hat{\mu_1} = 5.010183$, $\hat{\sigma_1^2} = 0.003946671$, $\hat{\pi} = 0.2809903$

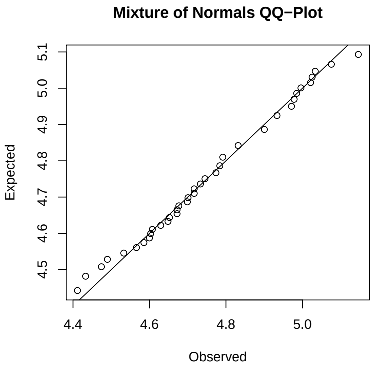
Point Estimate $\log_{10} \hat{Q} = 5.123601$; $\hat{Q} = 132,923$ cfs

Standard Error Estimate 0.03153947

90% CI for Q 117,956 cfs to 149,790 cfs

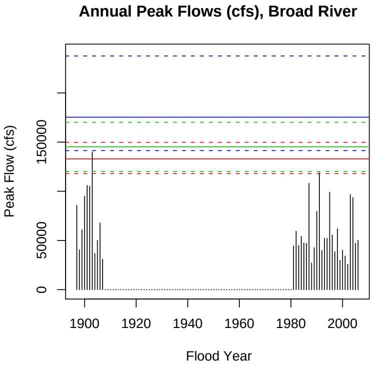
This interval for the 100-year flood is smaller than the one produced by Bulletin #17B.

Broad River



The QQ-plot indicates that a mixture of normals handles observations of the lower and upper tails of the flood distribution better than Bulletin #17B

Broad River



The point estimate and 90% CI of Q using the finite mixture model (red), Gumble (green), and Bulletin #17B (blue).

Recommendations

Use the gradient contour plot developed by Lindsay (1983) to decide whether or not additional components need to be added to a finite mixing distribution

If the plot indicates that additional components are needed, then fit the $\log_{10}$ of the annual peak flood flows to a mixture of two normal distributions

If additional components are not needed, proceed with Bulletin #17B in order to obtain that point estimate and confidence interval for the annual peak flood flows

Future Research

How would short-term dependence affect the estimation of the 100-year flood? The 1929 (March 1, 1929) and 1930 (October 3, 1929) flood years could pose a problem for the Congaree River.

Incorporating historical floods into the flood distribution as censored observations, and use these historical floods in an EM or ECM algorithm to obtain MLEs of the annual peak flood flows. Extend the use of historical floods into the finite mixture model.