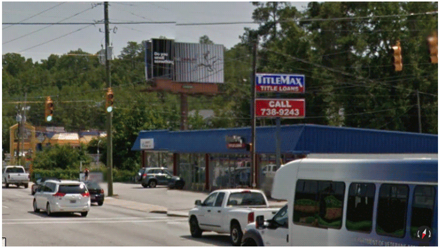
Nonstationary Flood Frequency Analysis: A Mixed and Pooled Approach

Phil Yates

Department of Mathematical Sciences
DePaul University

August 1, 2018

Columbia, SC: October 2015

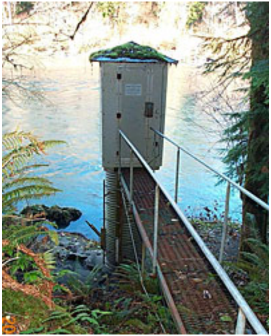


Garners Ferry Road in Columbia, SC before and during the October 2015 flood.

Economists estimated the floods caused about $12 billion damage on the state of South Carolina

Source: Sean Rayford/Getty Images

River Gage



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

Congaree River: Columbia, SC

Gervais Street Bridge, Columbia, SC
First version: 1827 to 1865
Second version: 1870 to 1928
Third version: 1928 to current
Flood gage about 0.2 miles south of current Gervais Street Bridge

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.

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)

Flood Frequency Data

Annual Peak Flows

The maximum momentary peak discharge in each year of record

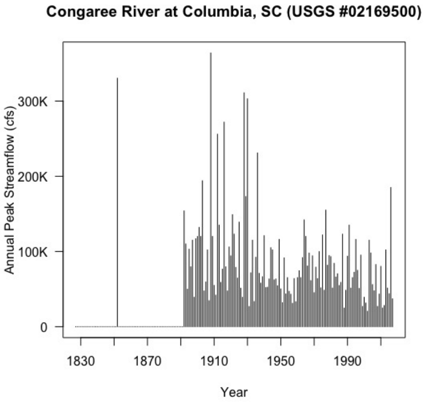
- The year of record begins the previous October and ends in September of the current year
- Peak flows are measured in cubic feet per second (cfs)

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.

Congaree River: Columbia, SC (USGS #02169500)



Historic Floods: 1852, Possible historic floods: 1886, 1888
Flood Record: 1892 to 2017

Modeling Flood Frequency Data

To estimate the 1% chance flood from this flood series:

Modeling Flood Frequency Data

To estimate the 1% chance flood from this flood series:

- Fit a finite mixture model using the $\log_{10}$ transformed annual peak streamflows

Modeling Flood Frequency Data

To estimate the 1% chance flood from this flood series:

- Fit a finite mixture model using the $\log_{10}$ transformed annual peak streamflows
  - Each component in the finite mixture model will be a normal distribution
- Components of the likelihood function:
  - observed annual peak flows: 1892 to 2017

Modeling Flood Frequency Data

To estimate the 1% chance flood from this flood series:

- Fit a finite mixture model using the $\log_{10}$ transformed annual peak streamflows
  - Each component in the finite mixture model will be a normal distribution
- Components of the likelihood function:
  - observed annual peak flows: 1892 to 2017
  - historic period: 1827 to 1891

Modeling Flood Frequency Data

To estimate the 1% chance flood from this flood series:

- Fit a finite mixture model using the $\log_{10}$ transformed annual peak streamflows
    - Each component in the finite mixture model will be a normal distribution
- Components of the likelihood function:
    - observed annual peak flows: 1892 to 2017
    - historic period: 1827 to 1891
    - observed historic flood: 1852

Modeling Flood Frequency Data

To estimate the 1% chance flood from this flood series:

- Fit a finite mixture model using the $\log_{10}$ transformed annual peak streamflows
  - Each component in the finite mixture model will be a normal distribution
- Components of the likelihood function:
  - observed annual peak flows: 1892 to 2017
  - historic period: 1827 to 1891
  - observed historic flood: 1852
  - assume that 1852 is the maximum flood in the $m_1 = 26$ flood year period (1827 to 1852) – this maximum value is the 26th order statistic in this period

Modeling Flood Frequency Data

To estimate the 1% chance flood from this flood series:

- Fit a finite mixture model using the $\log_{10}$ transformed annual peak streamflows
    - Each component in the finite mixture model will be a normal distribution
- Components of the likelihood function:
    - observed annual peak flows: 1892 to 2017
    - historic period: 1827 to 1891
    - observed historic flood: 1852
    - assume that 1852 is the maximum flood in the $m_1 = 26$ flood year period (1827 to 1852) – this maximum value is the 26th order statistic in this period
    - assume that the 1853 to 1891 floods are all smaller than the 1852 flood

Modeling Flood Frequency Data

To estimate the 1% chance flood from this flood series:

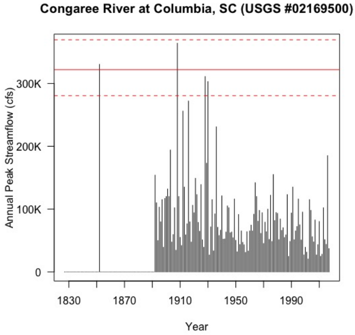
- Fit a finite mixture model using the $\log_{10}$ transformed annual peak streamflows
  - Each component in the finite mixture model will be a normal distribution
- Components of the likelihood function:
  - observed annual peak flows: 1892 to 2017
  - historic period: 1827 to 1891
  - observed historic flood: 1852
  - assume that 1852 is the maximum flood in the $m_1 = 26$ flood year period (1827 to 1852) – this maximum value is the 26th order statistic in this period
  - assume that the 1853 to 1891 floods are all smaller than the 1852 flood
  - 1% chance flood is the 99th percentile of the fitted finite mixture model

Modeling Flood Frequency Data

To estimate the 1% chance flood from this flood series:

- Fit a finite mixture model using the $\log_{10}$ transformed annual peak streamflows
  - Each component in the finite mixture model will be a normal distribution
- Components of the likelihood function:
  - observed annual peak flows: 1892 to 2017
  - historic period: 1827 to 1891
  - observed historic flood: 1852
  - assume that 1852 is the maximum flood in the $m_1 = 26$ flood year period (1827 to 1852) – this maximum value is the 26th order statistic in this period
  - assume that the 1853 to 1891 floods are all smaller than the 1852 flood
  - 1% chance flood is the 99th percentile of the fitted finite mixture model
  - Due to the complexity of the likelihood function, the standard error of the estimate of the 1% chance flood is produced solely from score functions (first partial derivatives)

Initial Results: 1827 to 2017



Parameter Estimates:

$$\hat{\boldsymbol{\mu}}^{\boldsymbol{T}} = (4.854, 5.480)\,, \hat{\boldsymbol{\sigma}}^{\boldsymbol{T}} = (0.228, 0.0498)\,, \hat{\tau} = 0.0278$$

Point Estimate for 1% Chance Flood: 321,972 cfs

95% CI for 1% Chance Flood: 280,568 cfs to 369,486 cfs

Main Assumption

The main assumption to the previous analysis is that the annual peak streamflows are from a stationary process.

The problem?

The main assumption to the previous analysis is that the annual peak streamflows are from a stationary process.

The problem?

POLICYFORUM

CLIMATE CHANGE

Stationarity Is Dead: Whither Water Management?

P. C. D. Milly,[1]* Julio Betancourt,[2] Malin Falkenmark,[3] Robert M. Hirsch,[4] Zbigniew W. Kundzewicz,[5] Dennis P. Lettenmaier,[6] Ronald J. Stouffer[7]

Climate change undermines a basic assumption that historically has facilitated management of water supplies, demands, and risks.

Source: *Science*, Vol. 319, pp. 573-574, 1 February 2008, doi: 10.1126/science.1151915

Possible Solution?

Possible Solution?

- Identify a change point in the flood series.
  - Initial analysis uses the `strucchange` package in R (Zeileis et al. 2002, Zeilieis et al. 2003) to identify the break points.

Possible Solution?

- Identify a change point in the flood series.
    - Initial analysis uses the `strucchange` package in R (Zeileis et al. 2002, Zeilieis et al. 2003) to identify the break points.
- Fit a finite mixture model for each stationary part of the flood series.

Possible Solution?

- Identify a change point in the flood series.
  - Initial analysis uses the `strucchange` package in R (Zeileis et al. 2002, Zeilieis et al. 2003) to identify the break points.
- Fit a finite mixture model for each stationary part of the flood series.
- Pool the results together.

This is the mixed and pooled approach:

- **Pooled:** Annual peak streamflows are identified to belong to a specific component (i.e., before a dam was built vs. after a dam was built) beforehand

Possible Solution?

- Identify a change point in the flood series.
  - Initial analysis uses the `strucchange` package in R (Zeileis et al. 2002, Zeilieis et al. 2003) to identify the break points.
- Fit a finite mixture model for each stationary part of the flood series.
- Pool the results together.

This is the mixed and pooled approach:

- **Pooled:** Annual peak streamflows are identified to belong to a specific component (i.e., before a dam was built vs. after a dam was built) beforehand
  - This is addressed by the change point analysis.

Possible Solution?

- Identify a change point in the flood series.
  - Initial analysis uses the `strucchange` package in R (Zeileis et al. 2002, Zeilieis et al. 2003) to identify the break points.
- Fit a finite mixture model for each stationary part of the flood series.
- Pool the results together.

This is the mixed and pooled approach:

- **Pooled:** Annual peak streamflows are identified to belong to a specific component (i.e., before a dam was built vs. after a dam was built) beforehand
  - This is addressed by the change point analysis.
- **Mixed:** Within each stationary series, it is not known ahead a time if the annual peak streamflow is from a specific component (i.e., was it due to tropical storms? Snow melt?)

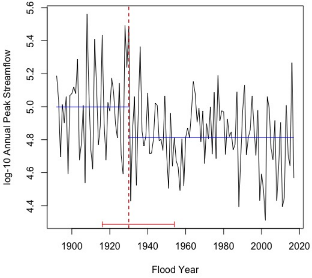
Possible Solution?

- Identify a change point in the flood series.
    - Initial analysis uses the `strucchange` package in R (Zeileis et al. 2002, Zeilieis et al. 2003) to identify the break points.
- Fit a finite mixture model for each stationary part of the flood series.
- Pool the results together.

This is the mixed and pooled approach:

- **Pooled:** Annual peak streamflows are identified to belong to a specific component (i.e., before a dam was built vs. after a dam was built) beforehand
    - This is addressed by the change point analysis.
- **Mixed:** Within each stationary series, it is not known ahead a time if the annual peak streamflow is from a specific component (i.e., was it due to tropical storms? Snow melt?)
    - This is estimated by the finite mixture models.

Congaree River: 1892 to 2017



Results from `strucchange` package in R: two pooled components

- estimated change point is 1930 (95% CI: 1916 to 1954)
- 1930 – Dreher Shoals Dam completed about 10 miles west of flood gage/Columbia, SC
- 1892 to 1930 – pre-dam annual peak streamflows
    - The annual peak streamflow of the 1930 flood year occurred on October 3, 1929.
- 1931 to 2017 – post-dam annual peak streamflows

Proposed Model

Assume that there are $K-1$ change points in the flood series. The pdf of the annual peak streamflows can be written as a linear combination of K finite mixture models.

$$p(y) = \sum_{k=1}^{K} \pi_k \left[\sum_{j=1}^{J_k} \tau_{jk} f\left(y|\boldsymbol{\theta}_{jk}\right) \right],$$

where

$$\sum_{k=1}^{K} \pi_k = 1, \sum_{j=1}^{J_k} \tau_{jk} = 1, \ and \int_{\mathcal{Y}} f\left(y|\boldsymbol{\theta}_{jk}\right) \ dy = 1$$

The likelihood function is

$$L\left(\boldsymbol{\eta}\right) = \prod_{i=1}^{n} p\left(y_i\right) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} \pi_k \left[\sum_{j=1}^{J_k} \tau_{jk} f\left(y_i|\boldsymbol{\theta}_{jk}\right) \right] \right\},$$

where

$$\boldsymbol{\eta} = \left(\boldsymbol{\theta}_j, \tau_j\right), j = 1, \ldots, J_k, k = 1, \ldots, K$$

Complete Data Likelihood & Log-Likelihood

The complete data, $x_i, i = 1, \ldots .n$, are observed indirectly via:

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

where each $\mathbf{z}_i = (z_{i11}, \ldots, z_{ijk})^{\mathbf{T}}$ is an indicator vector of length jk with 1 in the position corresponding to the appropriate **mixing** component and zeroes elsewhere and $\boldsymbol{\zeta}_i = (\zeta_{i1}, \ldots, \zeta_{ik})^{\mathbf{T}}$ is an indicator vector of length k with 1 in the position corresponding to the appropriate **pooling** component and zeroes elsewhere.

Complete Data Likelihood:

$$g(\mathbf{x}|\boldsymbol{\eta}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \prod_{j=1}^{J_k} \pi_k^{\zeta_{ik}} \tau_{jk}^{z_{ijk}\zeta_{ik}} f(y_i|\boldsymbol{\theta}_{jk})^{z_{ijk}\zeta_{ik}}$$

Complete Data Log-Likelihood:

$$\log g(\mathbf{x}|\boldsymbol{\eta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \zeta_{ik} \log \pi_k + \sum_{i=1}^{n} \sum_{k=1}^{K} \sum_{j=1}^{J_k} [z_{ijk}\zeta_{ik} \log \tau_{jk} + z_{ijk}\zeta_{ik} \log f(y_i|\boldsymbol{\theta}_{jk})]$$

A Simple Scenario...

The complete data log-likelihood can be "simplified" when $K = 2$ and $J_1 = J_2 = 2$:

$$
\begin{aligned}
\log g(\mathbf{x}|\boldsymbol{\eta}) &= \sum_{i=1}^{n} \left[\zeta_i \log \pi + (1-\zeta_i)\log(1-\pi)\right] \\
&+ \sum_{i=1}^{n} [z_{i1}\zeta_i \log \tau_1 + (1-z_{i1})\zeta_i \log(1-\tau_1) + z_{i0}(1-\zeta_i)\log \tau_0 + (1-z_{i0})(1-\zeta_i)\log(1-\tau_0 \\
&+ z_{i1}\zeta_i \log f(y_i|\boldsymbol{\theta}_{11}) + (1-z_{i1})\zeta_i \log f(y_i|\boldsymbol{\theta}_{01}) + z_{i0}(1-\zeta_i)\log f(y_i|\boldsymbol{\theta}_{10}) \\
&+ (1-z_{i0})(1-\zeta_i)\log f(y_i|\boldsymbol{\theta}_{00})]
\end{aligned}
$$

The parameters, $\boldsymbol{\eta} = (\tau_1, \tau_0, \boldsymbol{\theta}_{11}, \boldsymbol{\theta}_{10}, \boldsymbol{\theta}_{01}, \boldsymbol{\theta}_{00})$, can be estimated using an ECM algorithm.

In fact, a separate ECM algorithm can be performed on each known pooling component of finite mixture models!

Pooling of Finite Mixture Models: Congaree River Example

Analysis based on:

- observed annual peak flows: 1892 to 2017
- historic period: 1827 to 1891
- observed historic flood: 1852
- two known pooled components: 1827 to 1930, 1931 to 2017

Pooling Weights (based off of 1892 to 2017 flood years):

$$\boldsymbol{\pi}^T = \left(\frac{39}{126} = 0.310, \frac{87}{126} = 0.690 \right)$$

Parameter Estimates (first pooling component):

$$\hat{\boldsymbol{\mu}}^{\boldsymbol{T}} = (4.945, 5.476)\,, \hat{\boldsymbol{\sigma}}^{\boldsymbol{T}} = (0.222, 0.0440)\,, \hat{\tau} = 0.0804$$

Parameter Estimates (second pooling component):

$$\hat{\boldsymbol{\mu}}^{\boldsymbol{T}} = (4.803, 4.970)\,, \hat{\boldsymbol{\sigma}}^{\boldsymbol{T}} = (0.217, 0.00554)\,, \hat{\tau} = 0.0595$$

Point Estimate for 1% Chance Flood: 314,424 cfs

95% CI for 1% Chance Flood: 279,633 cfs to 353,543 cfs