Likelihood methods for qualitative and quantitative historical flood data

Peter A. Henrys

BHS meeting

Historical records for flood estimation

Introduction

- Aims
  - Use any historical data available to attempt to improve fit of model.
  - Use Bayliss and Reed graphical plotting approach to plot gauged data and historic data and assess fits of distributions
  - The aim cannot be to determine the absolute most correct model and method, rather the aim is to improve previously fitted distributions in light of additional historic information
  - As always results and interpretations should be carefully assessed and care taken to avoid incorrect conclusions.
- The project this work was born out of was to assess to most severe events in terms of return periods in light of additional historic information.

Fit a Distribution

- L-moments
  - Use probability weighted moments and L-moments of specified distribution to estimate parameters.
- Maximum Likelihood
  - Find the parameters that maximise the likelihood
- Bayesian methods and MCMC approaches
  - Repeatedly sample values for the parameters and retain or update these values according to some probability distribution function.

- Following work by Stedinger and Cohn (1986), we follow a likelihood based approach.

Likelihood Techniques

Statistical likelihood - the likeliness of distribution given observed data

- Likelihood of getting observed data
- Random Variable $X = \{x_1, x_2, \ldots, x_n\}$
- Prob(x_1) and Prob(x_2) and Prob(x_3) and ...

$$L(X;\Theta) = \prod_{i=1}^{n} f(x_i)$$

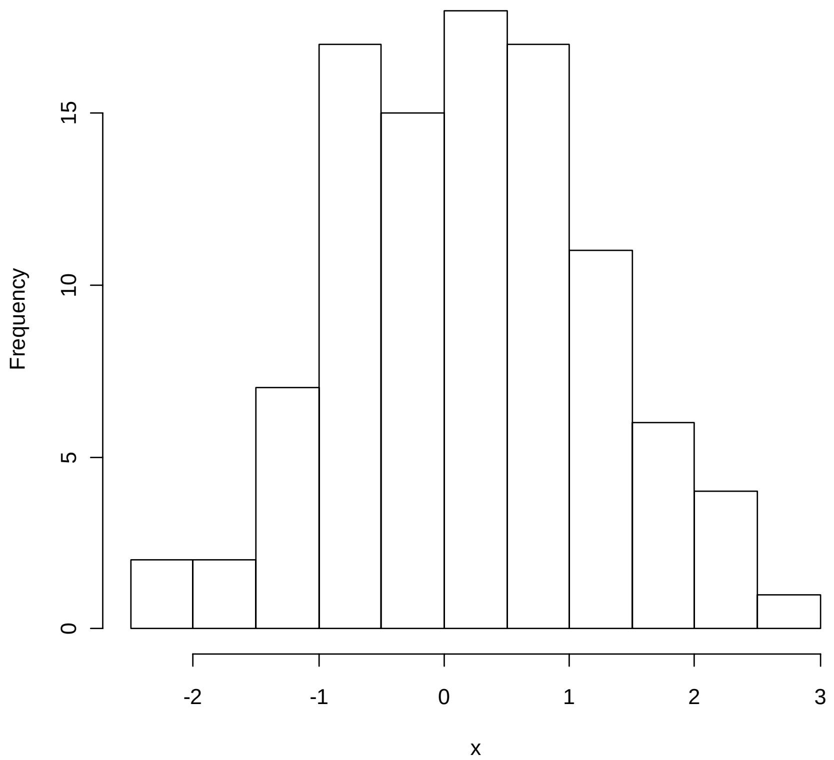
- Suppose that the data can be split into two sections. We could write the data set as

$$\begin{aligned} X &= \{x_1 , \ldots , x_h , x_{h+1} , \ldots , x_{h+s}\} \\ &= \{y_1 , \ldots , y_h , z_1, \ldots , z_s\} \end{aligned}$$

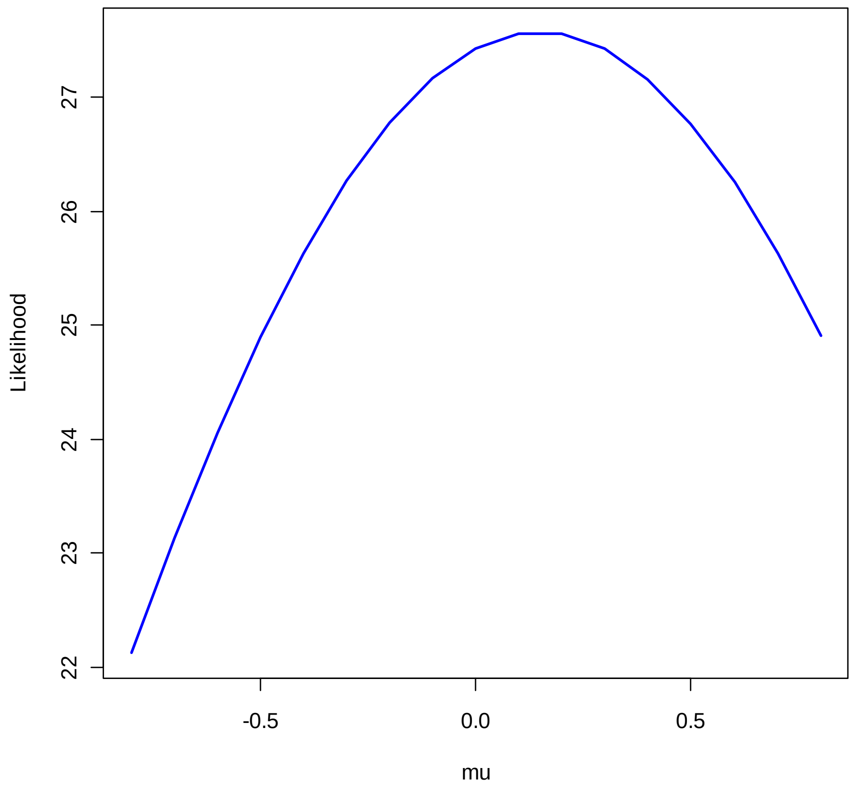
- And the likelihood could be written as

$$L(X;\Theta) = \prod_{i=1}^{s} f(y_i) \cdot \prod_{j=1}^{h} f(z_i)$$

Histogram of the observed data



Plot of likelihood as a function of parameter



Example

- Probability density function of the GEV distribution

$$f(x) = \frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi-1} \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

- So the likelihood in this case is a function of three parameters

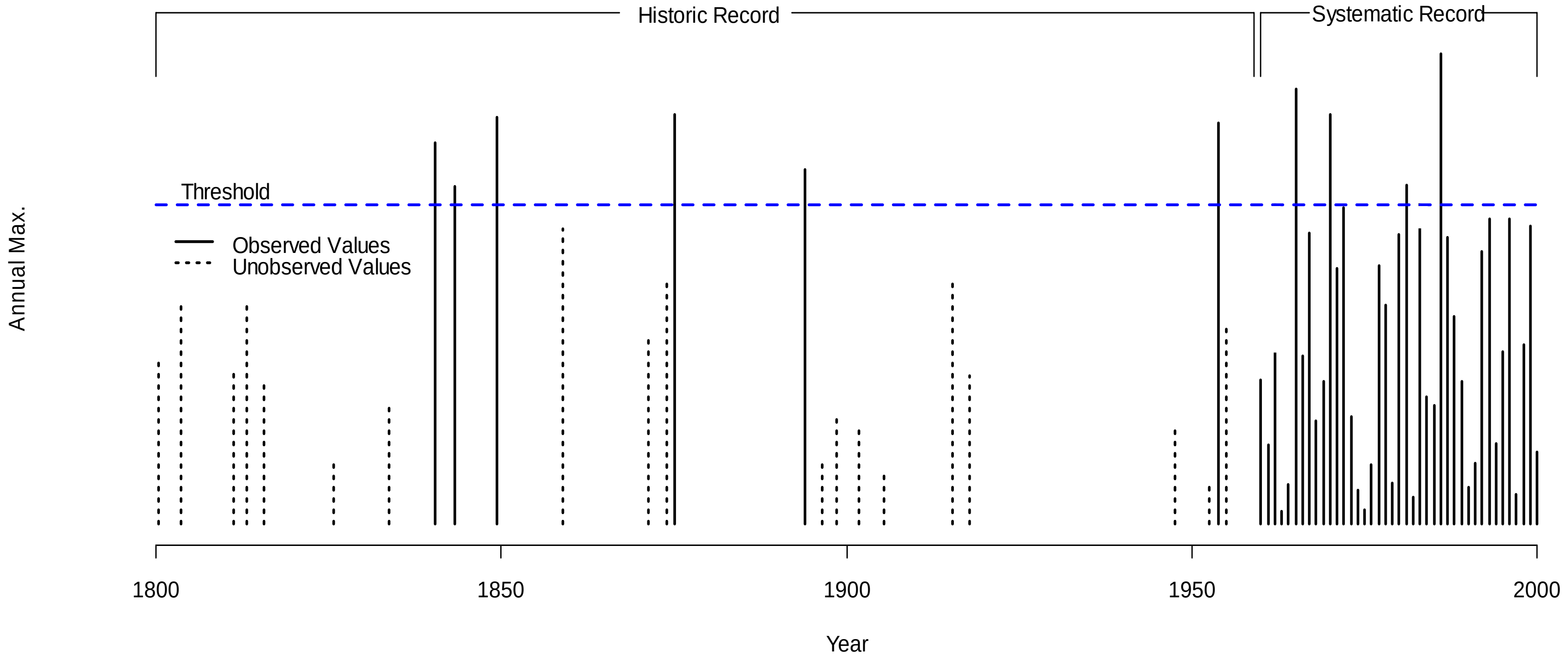
Stedinger and Cohn (1986),

- Present likelihood based methods for inclusion of historical data
- They realise that historical observations can be used directly in a likelihood approach
- The key, however, is to note that the historical data is censored
- Likelihood methods can accommodate any additional components that contribute to the overall probability of obtaining the observed data
- Adopt a binomial distribution into likelihood to account for censoring, i.e. the probability that we “saw” an historical event.

Binomial Data

- Binomial data – success or failure
- Example – Roll of a die
  - I roll a die ten times, what is the probability I get three sixes?
  - A success in this scenario is rolling a six
  - So the probability of success is 1/6
  - The probability of three sixes is 1/6 x 1/6 x 1/6
  - However, there have also been 7 instances we haven't got a 6
  - The probability of this is 5/6 x 5/6 x 5/6 x 5/6 x 5/6 x 5/6 x 5/6
  - These successes and failure can happen in any order

$$\Pr(x=6)=\binom{10}{3}(1/6)^3(5/6)^7$$



"Success" defined as observation above threshold (T)

Probability of success from distribution function

$$\Pr(X > T) = 1 - \Pr(X \leq T) = 1 - F_X(T)$$

- So if we have k threshold exceedances from the h years of historic data, that is k successes
- Then,

$$\Pr(k \text{ successes}) = \binom{h}{k} \cdot (1 - F_X(T))^k \cdot (F_X(T))^{h-k}$$

- h and k are known, so including this in a likelihood with the distribution function introduces no more parameters

Full Likelihood

- The likelihood consists of three components. They are:
  - The probability distribution of the systematic observations
  - The binomial distribution corresponding to the number of exceedances
  - The probability distribution of the historical data

$$L(x, y; \Theta) = \prod_{i=1}^{s} f_x(x_i) \left\{ \binom{h}{k} [1 - F_x(T)]^k [F_X(T)]^{h-k} \right\} \prod_{j=1}^{k} f_y(y_j)$$

- Because the distribution of the historical data can be written as a scaled version of the distribution of the gauged data, the likelihood becomes

$$f_y(y) = \frac{f_x(y)}{(1 - F_x(T)))}$$

$$L(x, y; \Theta) = \prod_{i=1}^{s} f_x(x_i) \left\{ \binom{h}{k} [F_x(T)]^{h-k} \right\} \prod_{j=1}^{k} f_x(y_j)$$

Qualitative Analysis

- Sometimes with historical data, the occurrence of a large flood event is known, perhaps via newspapers or personal accounts, but the extent of the flood is unknown
- Stedinger and Cohn also look into such situations and show that even when the presence of an exceedance is all that is known, improvements can be made on the model fit.
- So we keep the binomial component in the likelihood for censoring and known exceedances, but we drop the data on the quantitative observation

$$L(x, y; \Theta) = \prod_{i=1}^{s} f_x(x_i) \left\{ \binom{h}{k} [1 - F_x(T)]^k [F_x(T)]^{h-k} \right\}$$

Graphical Checks

- Use same approach as outlined in Bayliss and Reed, to plot the gauged and historic event on the same plot.

- Standard plotting position formula -

$$P_i = \frac{i - \alpha}{n + 1 - 2\alpha}$$

- Plotting position formula for data including historical events –

  - Above threshold

$$P_i = \frac{i - \alpha}{k + 1 - 2\alpha} \frac{k}{n}$$

  - Below threshold

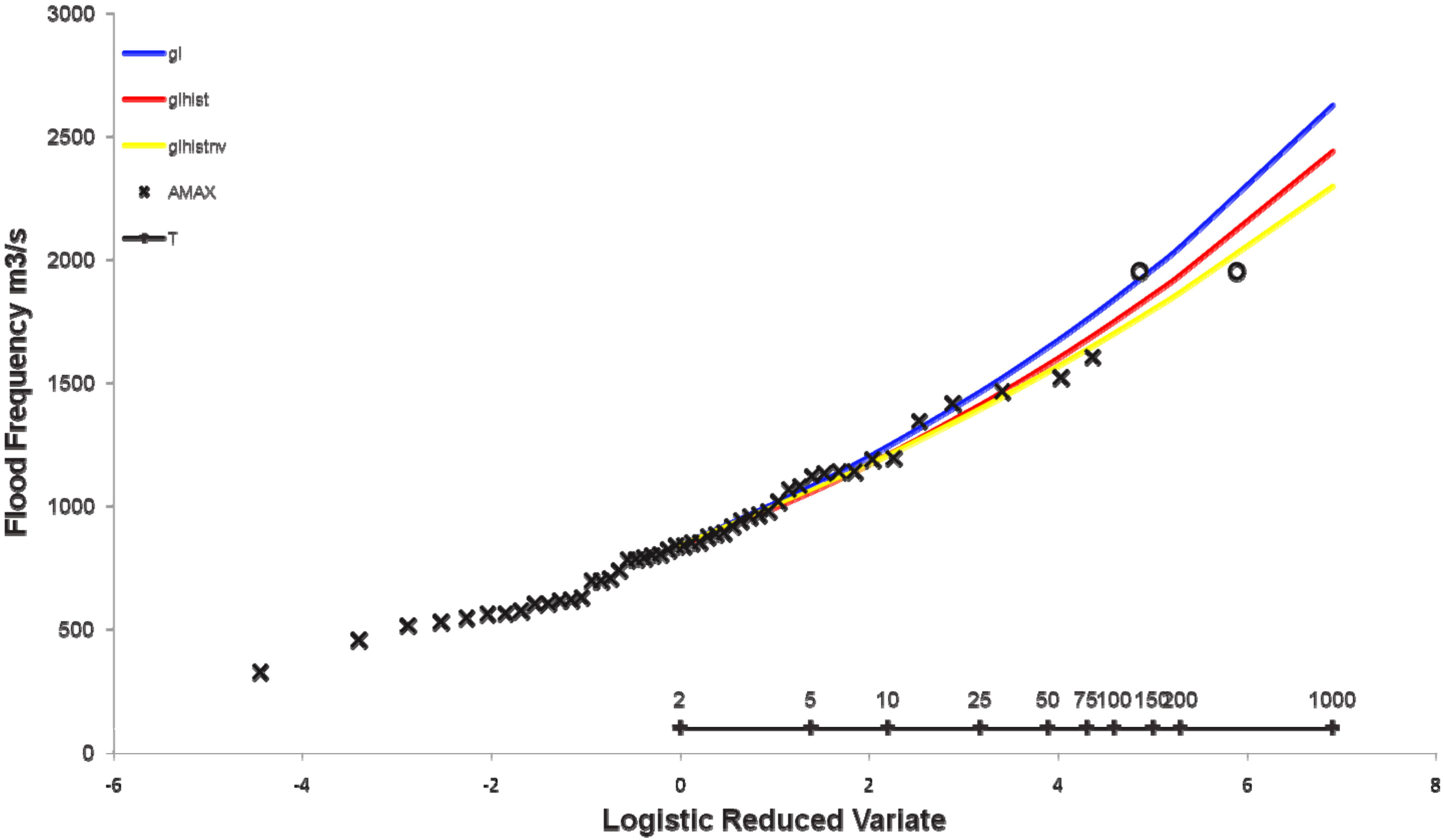
$$P_i = \frac{k}{n} + \frac{n - k}{n} \frac{i - k - \alpha}{s - e + 1 - 2\alpha}$$

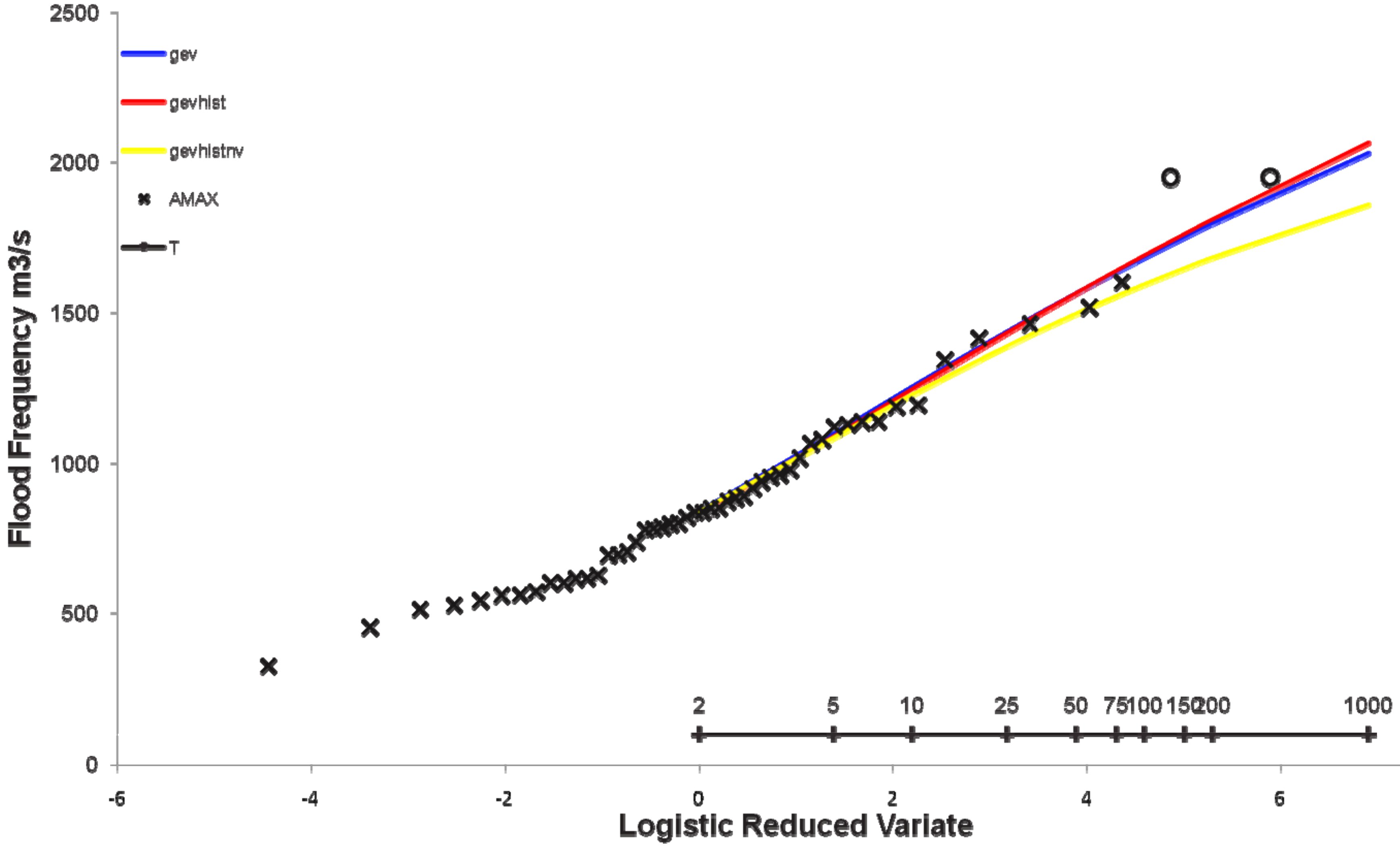
- Adding the estimated distributions to the observed data plotted using the plotting position formulae enables inspection of the fit of the distributions.

Example – River Tweed

- Three historical events were sourced.
- One is a qualitative event as the flow is unknown
- Summary of flood ranking

1948 = Rank 1	Q = 1950 m^3/sec	Historical
1831 = Rank 1	Q = 1950 m^3/sec	Historical
2002 = Rank 3	Q = 1602 m^3/sec	Gauged
1956 = Rank 3	Q = >1518 m^3/sec	Historical
1982 = Rank 5	Q = 1518 m^3/sec	Gauged


Flood Frequency m3/s
3000
2500
2000
1500
1000
500
0
gl
glhist
glhistnv
AMAX
T
2
5
10
25
50
75
100
150
200
1000
-6
-4
-2
0
2
4
6
8
Logistic Reduced Variate


2500
2000
1500
1000
500
0
gev
gevhist
gevhistnv
AMAX
T
Flood Frequency m3/s
2
5
10
25
50
75
100
150
200
1000
-6
-4
-2
0
2
4
6
8
Logistic Reduced Variate

Conclusions and Further Work

- If we have collected historical data, we should use it to its full extent
- Use the approaches presented, we can improve the fit of the distribution by including historical data
- Graphical checking is crucial. Although we may fit numerous distributions, we should always compare to the fits to the data.
- When we fit a distribution there is mathematical theory underpinning the extreme values that allows for randomness in the sample.
- A full Bayesian approach would perhaps do a better job at estimating the parameters given the uncertainty and censoring of the historical events.