Universiteit Utrecht

Stochastic Hydrology

Hydrological statistics and extremes

Marc F.P. Bierkens

Professor of Hydrology
Faculty of Geosciences

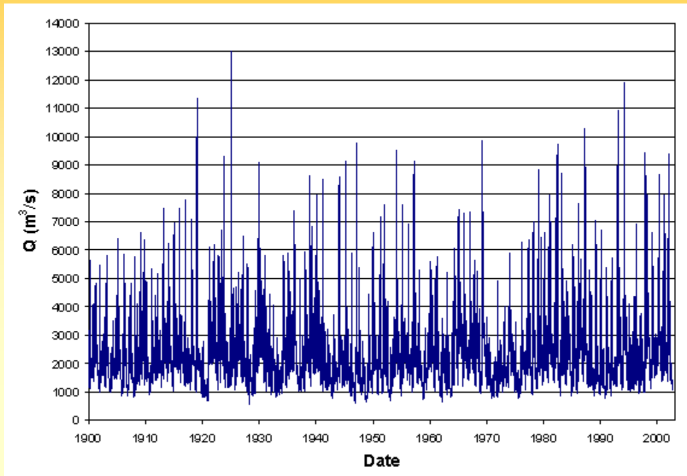
Universiteit Utrecht

Hydrological statistics

Mostly concernes with the statistical analysis of hydrological time series in relation to *extremes*, i.e. floods and droughts.

Universiteit Utrecht

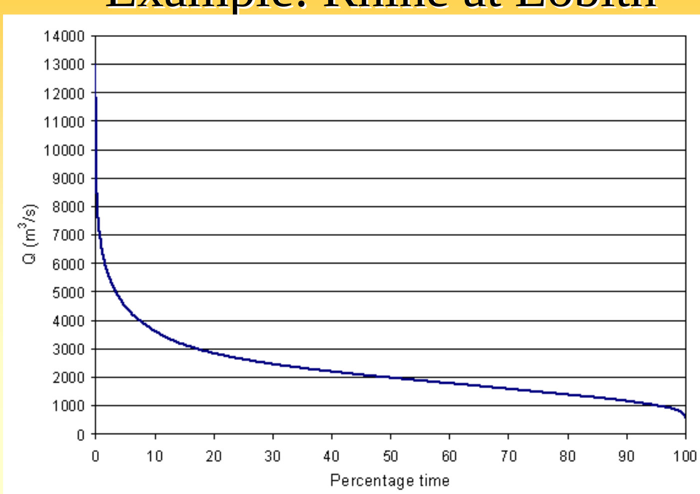
Example: Rhine at Lobith



Daily averaged discharge (m^3/s) 1902-2002

Universiteit Utrecht

Example: Rhine at Lobith



Flow duration curve of Rhine discharge

Universiteit Utrecht

Extreme events

Want to know:

What is the probability distribution that a flood of given size occurs?

What is the size of a flood that belongs to a given design frequency?

First question that must be answered is:

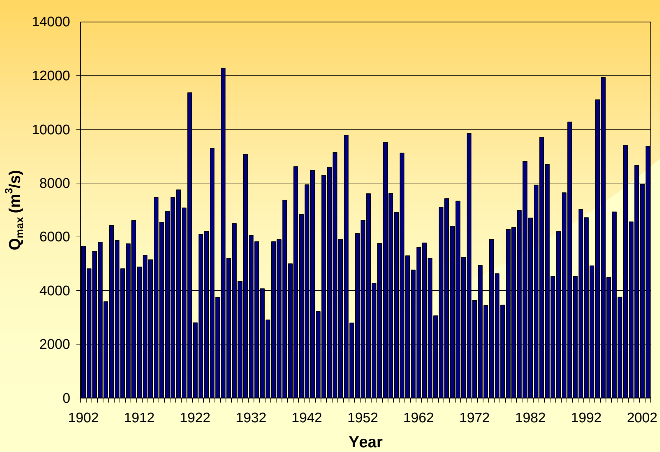
What constitutes a flood?

Two methods:

1. The largest discharge per year (Maximum values)
2. All discharge values above a certain threshold (Peak over Threshold (POT) data or partial duration data)

Universiteit Utrecht

Maximum values of Rhine at Lobith



Maximum daily averaged discharge (m^3/s) for each year in 1902-2002

Assumptions about maximum values

1. The maximum values are realisations of independent random variables.
2. There is no trend in time.
3. The maximum values are identically distributed.

In this case a single probability distribution can be assumed for the maximum values.

Probability and recurrence times

Cumulative probability distribution: $F(y) = \Pr(Y \leq y)$

Exceedence probability: $P(y) = \Pr(Y \geq y)$

Return period or recurrence time: $T(y) = \dfrac{1}{P(y)} = \dfrac{1}{1 - F(y)}$

Recurrence time: the *average* number years between two consecutive flood events of a given size

Note: the actual number of years between flood events of a given size is itself random.

Universiteit Utrecht

Recurrence times from data

Analysis of maximum values of Rhine discharge at Lobith for recurrence time

Y	Rank	*F*(*y*)	*P*(*y*)	*T*(*y*)
2790	1	0.00971	0.99029	1.0098
2800	2	0.01942	0.98058	1.0198
2905	3	0.02913	0.97087	1.0300
3061	4	0.03883	0.96117	1.0404
3220	5	0.04854	0.95146	1.0510
3444	6	0.05825	0.94175	1.0619
3459	7	0.06796	0.93204	1.0729
.	.	.	.	.
.	.	.	.	.
9140	90	0.87379	0.12621	7.9231
9300	91	0.88350	0.11650	8.5833
9372	92	0.89320	0.10680	9.3636
9413	93	0.90291	0.09709	10.3000
9510	94	0.91262	0.08738	11.4444
9707	95	0.92233	0.07767	12.8750
9785	96	0.93204	0.06796	14.7143
9850	97	0.94175	0.05825	17.1667
10274	98	0.95146	0.04854	20.6000
11100	99	0.96117	0.03883	25.7500
11365	100	0.97087	0.02913	34.3333
11931	101	0.98058	0.01942	51.5000
12280	102	0.99029	0.00971	103.0000

$$\hat{F}(y) = \frac{i}{n+1}$$

$$\hat{P}(y) = 1 - \hat{F}(y)$$

$$\hat{T}(y) = \frac{1}{1-\hat{F}(y)}$$

Universiteit Utrecht

Large recurrence times

From a series of n maximum values: largest recurrence time to be assessed: n+1 years!

For design purposes: y_T for large T is necessary!

To this end:

1. Fit a probability distribution (which one?)
2. Use it to extrapolate to large values of y_T (maximum values y with $F(y)$ close to 1).

Universiteit Utrecht

The Gumbel distribution

It can be proven (see section 4.2.2. Syllabus) that maximum values of the following distributions follow a Gumbel distribution:

- Exponential
- Gaussian
- logGaussian
- Gamma
- Logistic
- Gumbel

pdf: $f_Y(y) = be^{-b(y-a)}\exp(-e^{-b(y-a)})$

cpdf: $F_Y(z) = \exp(-e^{-b(y-a)})$

Universiteit Utrecht

Fitting the Gumbel distribution

Types of methods:

- graphical using Gumbel paper and linear regression
- method of moments
- maximum likelihood estimation

Universiteit Utrecht

Gumbel paper

Taking the double logarithm of the cpdf:

$$\ln[F_Y(y)] = \ln[(\exp(-e^{-b(y-a)})] = -e^{-b(y-a)}$$

$$\ln\{-\ln[F_Y(y)]\} = -b(y-a)$$

$$y_T = a - \frac{1}{b}\ln\{-\ln[F_Y(y)]\} \quad \frac{1}{\ }$$

$$y_T = a - \frac{1}{b}\ln\{-\ln[\frac{T(y)-1}{T(y)}]\}$$

Plot maximum y_i vs $-\ln\{-\ln[\frac{i}{n+1}]\}$

Plot maximum y_i vs $T(y_i)$ on special Gumbel (double logarithmic) paper

Universiteit Utrecht

Gumbel paper

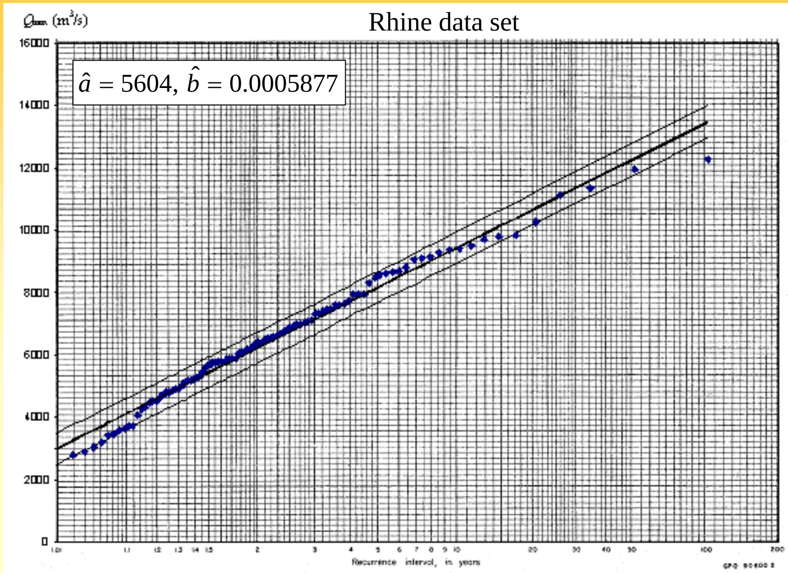
1. Plot maximum y_i vs $-\ln\{-\ln[\frac{i}{n+1}]\}$

   or, plot maximum y_i vs $T(y_i)$ on special Gumbel (double logarithmic) paper

2. Fit a straight line by eye or by regression to determina *a* and *b*.

Universiteit Utrecht

Gumbel paper



Universiteit Utrecht

Method of moments

Mean and variance of a Gumbel variate *Y*:

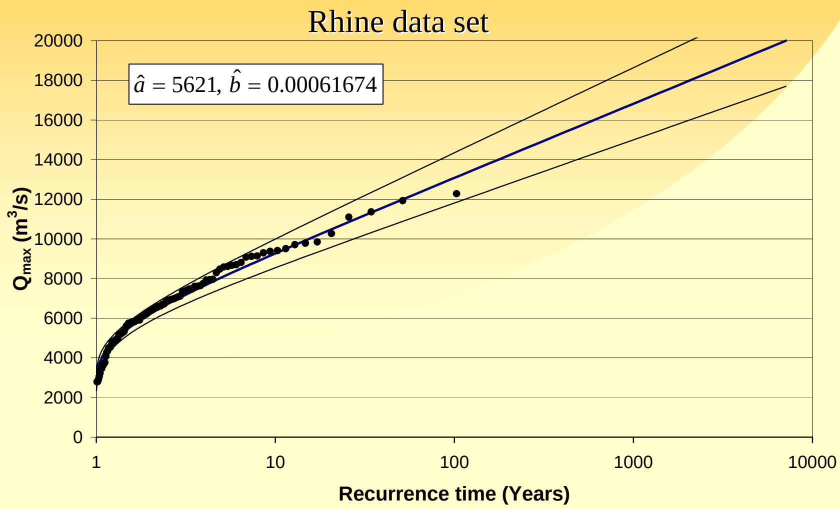
$$\mu_Y = a + \frac{0.5772}{b} \qquad \sigma_Y^2 = \frac{\pi^2}{6b^2}$$

1) Estimate mean m_Y and variance s_Y^2
2) Equate these to the above expressions
3) Solve for *a* and *b*:

$$\frac{1}{\hat{b}} = \sqrt{\frac{6 s_Y^2}{\pi^2}} \qquad \hat{a} = m_Y - \frac{0.5772}{\hat{b}}$$

Universiteit Utrecht

Method of moments



Universiteit Utrecht

Estimation of the *T*-year event

$$\hat{y}_T = \hat{a} - \frac{1}{\hat{b}} \ln(-\ln(\frac{T-1}{T}))$$

In case of the Rhine data set:

MoM parameters:

$$y_{1250} = 5621 - 1621 \cdot \ln(-\ln(0.9992)) = 17182 \text{ m}^3/s.$$

with parameters from regression: 17736 m^3/s

Universiteit Utrecht

Estimation of confidence limits

In case of regression:

$$\hat{y}_T - t_{95} s_{\hat{Y}}(T) \le y_T \le \hat{y}_T + t_{95} s_{\hat{Y}}(T)$$

t_{95} : the 95-point of the student's *t*-distribution

and the standard error of prediction $s_{\hat{Y}}(T)$ estimated as:

$$s^2_{\hat{Y}}(T) = \left(1 + \frac{1}{N} + \frac{(x(T)-\bar{x})^2}{\sum (x(T_i)-\bar{x})^2}\right) \frac{1}{N-1} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

with $x(T_i) = -\ln[-\ln((T_i - 1)/T_i)]$

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x(T_i)$$

Universiteit Utrecht

Estimation of confidence limits

In case of method of moments:

$$\hat{\mathrm{Var}}(\hat{y}_T) \approx \frac{(1.11 + 0.52x + 0.61x^2)}{\hat{b}^2 N}$$

with $x(T) = -\ln[-\ln((T-1)/T)]$

Assuming a Gaussian estimation error:

$$\hat{y}_T - 1.96\sqrt{\hat{\mathrm{Var}}(y_T)} \le y_T \le \hat{y}_T + 1.96\sqrt{\hat{\mathrm{Var}}(y_T)}$$

Universiteit Utrecht

The number T-year events in a given period

The *expected* number of T-year events in N years:

$$Np = N/T$$

The *actual* number n of T-year events in N years is a random variable obeying a binomial distribution:

$$\Pr(n \text{ events } y_T \text{ in } N \text{ years}) = \binom{N}{n} p^n (1-p)^{N-n}$$

Examples: $\Pr(1 \text{ events } y_{100} \text{ in } 10 \text{ years}) = \binom{10}{1} 0.01 \cdot (1-0.01)^9 = 0.0914$

Pr(*one or more* flood events occur in 10 years) = 1-Pr(no events) = $1-0.99^{10} = 0.0956$.

Universiteit Utrecht

The time until the next *T*-year event

The *expected* number of years n until the next T-year event: T

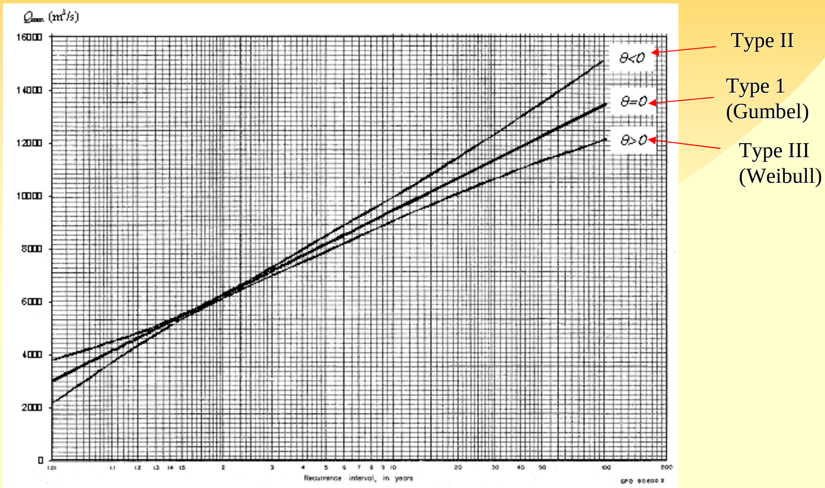
The *actual* number of years n until the next T-year event is a random variable obeying a geometric distribution (with p=1/T):

$$\Pr(m \text{ years until event } y_T) = (1-p)^{m-1} p$$

Universiteit Utrecht

Other extreme value distributions

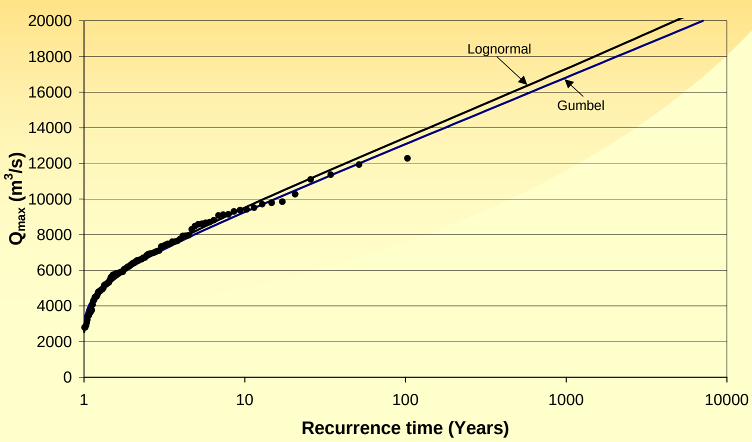
Generalized extreme value distributions



Universiteit Utrecht

Other extreme value distributions

Other maximum value distributions used for maximum values: log-normal, log-Pearson-type III.



Universiteit Utrecht

Minimum values (e.g. low flows)

- Take -Z or $1/Z$
- or use Weibull distribution on Z_{min}

Universiteit Utrecht

Testing the assumptions

Independence: Von Neuman's Q

$$Q = \frac{\sum_{i=1}^{n-1}(Y_{i+1} - Y_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

Lower critical area.

If larger than critical value: no evidence that data are dependent!

Rhine data set: Q=2.071 ; upper critical value at α=0.05: 1.618 -> no evidence for dependence

Universiteit Utrecht

Testing the assumptions

Trends: Mann-Kendall test

$$T = \sum_{i=2}^{n} \sum_{j=1}^{i-1} \mathrm{sgn}(Y_i - Y_j)$$

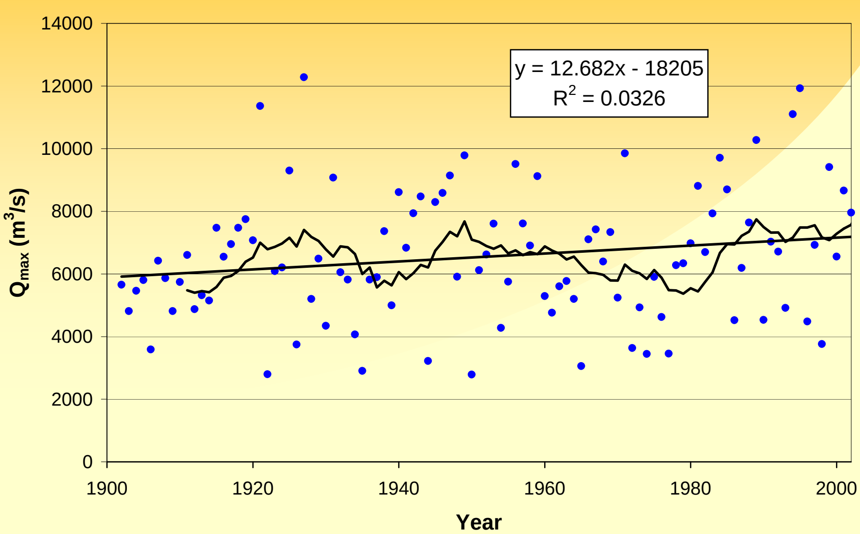
$$T' = 18T / [n(n-1)(2n+5)]$$

For n>40 T' has standard Gaussian distribution with two sided critical area (i.e. significant at 95% (α=0.05) accuracy trend is $T' < -1.96$ or $T' > 1.96$); otherwise no evidence of a trend.

Rhine data set: T' = 1.992 -> significant trend at 95% accuracy.

Universiteit Utrecht

Yearly maxima of average dayly runoff of the Rhine at Lobith



Universiteit Utrecht

Testing the assumptions

Testing for some distribution: χ^2 test

1. Define a *m* classes (as in a histogram) and assign the *n* data values
2. Count the number of data falling in each each class *i*: n_i
3. Fit the proposed distribution function $F(Y)$ to the data
4. Calculate the expected number of data falling into each class *i* as:

$$n_i^e = n\left(F_Y(y_{\text{up}}) - F_Y(y_{\text{low}})\right)$$

5. The following test statistic is calculated:

$$\mathrm{X}^2 = \sum_{i=1}^{m} \frac{(n_i - n_i^e)^2}{n_i^e}$$

Universiteit Utrecht

Testing the assumptions

Testing for some distribution: χ^2 test

X^2 follows a chi-squared distribution with *m*-1 degress of freedom: χ^2_{m-1}

There is an upper critical area for the 0-hypothesis that the data follow the proposed distribution.

Rhine data and proposed Gumbel distribution and 20 classes: $\mathrm{X}^2 = 23.70$

Rhine data and proposed lognormal distribution and 20 classes: $\mathrm{X}^2 = 14.36$

Lower boundary critical area for *m*-1 = 19 degrees of freedom and α=0.05: 30.144 -> both distributions cannot be discarded!