Short Course

Statistische Methoden in der Meteorologie — Extremwertverteilungen

Dr. Manfred Mudelsee
Prof. Dr. Werner Metz
LIM—Institut für Meteorologie
Universität Leipzig

Donnerstag, 13. Februar 2003, 10:00 bis 12:00 und 13:00 bis ≈ 15:00, LIM
Sommersemester 2002

Contents

MORNING PART: SOME THEORY

Chapter 1

Probability Theory

1.1 Random variable

What is the current temperature in 2 km height above Mount Everest?
What was the summer temperature at Grünberg (Silesia) in A.D. 1690?

Measurement and systematic errors (e. g., tree-ring width as temperature indicator):

$\Rightarrow$ uncertainty, unknown variable value,
$\Rightarrow$ probability, **random variable**, here denoted as X.

1.2 Probability density function of a random variable

$f(x) =$ probability per interval dx ($\int f = 1$).

Excursion: probability density function.

1.3 Probability distribution function of a random variable

$F(x) = \int_{-\infty}^{x} f(y)dy.$

Excursion: probability distribution function.

1.4 Expectation of a random variable

$E[X] = \int x \cdot f(x)dx.$

Note: $\int$ normally goes from $-\infty$ to $+\infty$.
Expectation of a function g of a random variable X:
$E[g(X)] = \int g(x) \cdot f(x)dx$.

1.5 Moments of a random variable: mean, variance

Moment of order 1: **mean** $= E[X] = \int_{-\infty}^{+\infty} x \cdot f(x)dx$.
Central moment of order 2: **variance** $= E[(X - E[X])^2]$.
Standard deviation $=$ std $= \sqrt{\text{variance}}$
Some formulae: $E(aX + b) = a \cdot E[X] + b$.

$$\begin{aligned} \text{Variance} &= VAR = E[(X - E[X])^2] \\ &= E[X^2 - 2XE[X] + E[X]^2] = E[X^2] - 2E[X] \cdot E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2. \end{aligned}$$

1.6 Normal or Gaussian density function

$f_N(x) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$,

parameters: μ, σ. Notation: $X \sim \mathrm{N}(\mu, \sigma^2)$. $\mathrm{N}(0,1)$: standard normal density.

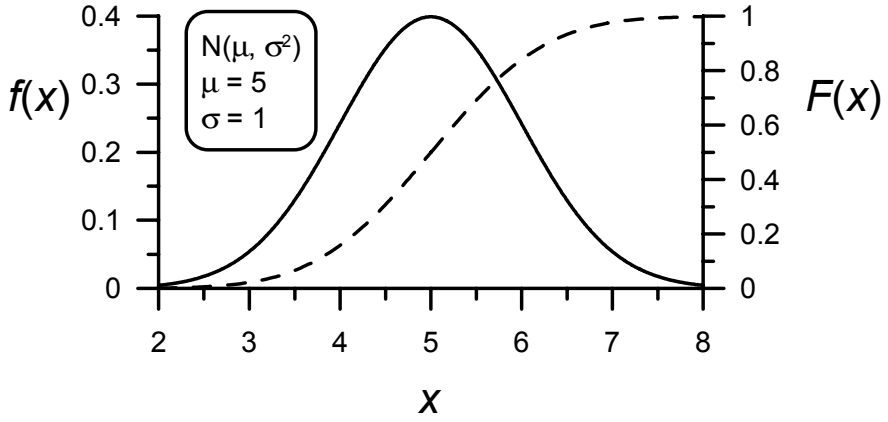

Figure 1.1: Normal density (solid), normal distribution (dashed) functions. Notes: (1) inventor = de Moivre $\neq$ Gauss. (2) $F(x)$: no closed form $\Rightarrow$ approximations.

1.6.1 Moments of the normal density

We recall: $\int f_{N(0,1)} = 1$ (note normalizing constant).

Mean $= E[X] = \int_{-\infty}^{+\infty} x \cdot f(x)dx$.

Excursion: (via substitution) mean $= \mu$.

Variance $= E[(X - E[X])^2] = E[(X - \mu)^2] =$ (homework) $= \sigma^2$.

(Standardized) **skewness**: $\gamma_1 = E[((X - \mu)/\sigma)^3] = 0$ (symmetric).

1.6.2 Parameters of the normal density

f_N is a two-parameter density: μ and σ. These parameters determine (here actually: they equal) mean and standard deviation. In applications it is therefore important to know their true values.

Chapter 2

Parametric Estimation

We have a data sample and a candidate density function. We wish to know:

- What are the true values of the parameters of the density?

- Is the candidate density a good model for the data?

2.1 Sample

Observations: $x(1), x(2), \ldots, x(n) = \{x(i), i = 1, n\}$,
n: number of observations, data size or sample size.

Note: x (sample) vs. X (theory, model, population).

We use the sample to **estimate** model parameters. This is one part of statistical inference (the other is hypothesis testing).

2.2 Maximum likelihood estimation of f_N parameters μ, σ^2

We say:

Well, the sample $\{x(i), i = 1, n\}$ could stem from $f_N(x; \mu_1, \sigma_1^2)$. The probability for that is proportional to

$$f_N(x(1); \mu_1, \sigma_1^2) \cdot f_N(x(2); \mu_1, \sigma_1^2) \cdot \ldots \cdot f_N(x(n); \mu_1, \sigma_1^2) = \Pi_{i=1}^n f_N(x(i); \mu_1, \sigma_1^2).$$

(We assume that the data are Excursion: independent and probabilities can, hence, be multiplied.)

Eventually, $\Pi_{i=1}^{n} f_N(x(i); \mu_2, \sigma_2^2)$ is higher. Then we would prefer (μ_2, σ_2^2) as estimate.

Define the **likelihood function**,

$$L_{\{x(i),i=1,n\}}(\mu, \sigma^2) = \Pi_{i=1}^{n} f_N(x(i); \mu, \sigma^2),$$

a function of the parameters, and maximize it. What amounts to the same, but is normally easier to calculate: maximize the **log-likelihood function**,

$$l_{\{x(i),i=1,n\}}(\mu, \sigma^2) = \log\left[L_{\{x(i),i=1,n\}}(\mu, \sigma^2)\right] = \sum_{i=1}^{n} \log\left[f_N(x(i); \mu, \sigma^2)\right].$$

For f_N, therefore, maximize

$$l_{\{x(i),i=1,n\}}(\mu, \sigma^2) = \sum_{i=1}^{n} \left\{-(1/2)\log(\sigma^2) - (1/2)\log(2\pi) - (1/2)\left[x(i) - \mu\right]^2/\sigma^2\right\}.$$

(We take log = natural logarithm because this is easier here.) Now, take 1st derivative, set zero, look whether 2nd derivative is negative (homework!?), to obtain:

$$\widehat{\mu}_{\mathrm{ML}} = (1/n) \sum_{i=1}^{n} x(i)$$

$$\widehat{\sigma^2}_{\mathrm{ML}} = (1/n) \sum_{i=1}^{n} \left[x(i) - \mu_{\mathrm{ML}}\right]^2 .$$

Plug in $\widehat{\mu}$ in last Eq. Notes: hats indicate estimators, index ML for "Maximum Likelihood".

2.3 Bias and variance of estimators

Other principles than maximum likelihood may yield other estimators.

Bias $= E[\widehat{\theta}] - \theta$ under assumed density, f.

Variance $= VAR[\widehat{\theta}]$.

Excursion: 4 cases: bias and variance large and small.

MSE = mean-squared error, to evaluate quality of an estimator:

$$\begin{aligned}
\text{MSE} &= E\left[\left(\widehat{\theta}-\theta\right)^2\right] \\
&= E\left[\left\{\left(\widehat{\theta}-E[\widehat{\theta}]\right)-\left(\theta-E[\widehat{\theta}]\right)\right\}^2\right] \\
&= E\left[\left(\widehat{\theta}-E[\widehat{\theta}]\right)^2\right]+\left(\theta-E[\widehat{\theta}]\right)^2-2\left(\theta-E[\widehat{\theta}]\right)\cdot\underbrace{E\left[\widehat{\theta}-E[\widehat{\theta}]\right]}_{0} \\
&= VAR\left[\widehat{\theta}\right]+(\text{bias})^2 .
\end{aligned}$$

Chapter 3

Nonparametric Estimation

Recall: we wish to estimate a density function. Until now, we assumed a certain density model, which has some parameters. We estimated those parameters and obtained the estimated density.

Here we try to estimate the density without specifying a model and parameters: nonparametric density estimation.

3.1 Histogram

Excursion: idea of histogram.

Problems with histogram: unsmooth estimate, where to place class bounds, what to do if $x(i)$ sits on class bound?

Excursion: histogram Oder floods.

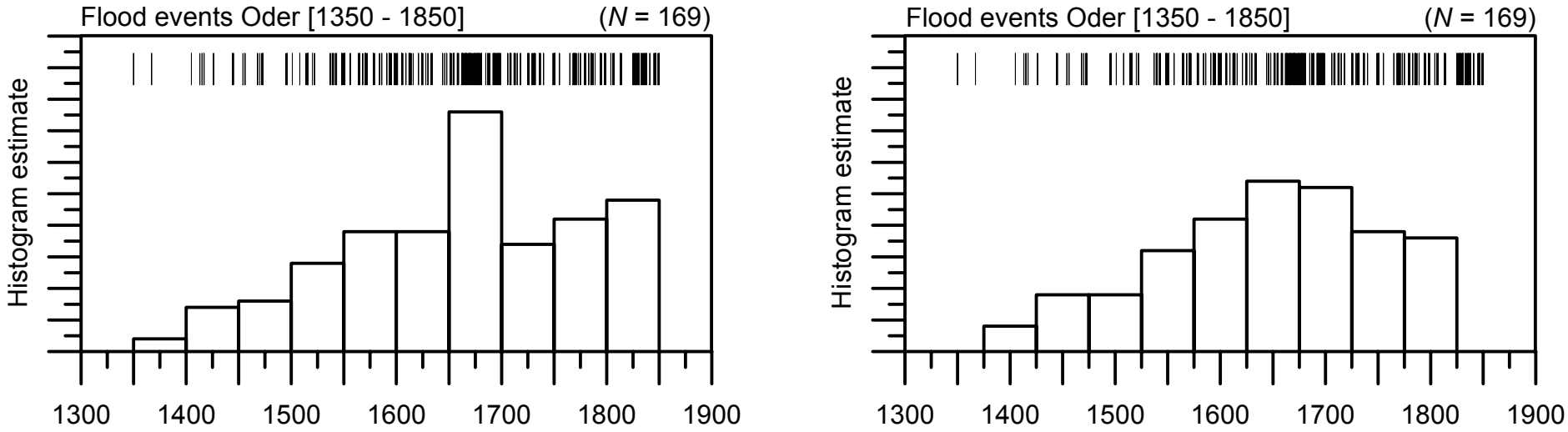

Figure 3.1: Floods river Oder, histogram estimation. Original dates as bars. Note effect of choice of class bounds. Data from Weikinn sources of documentary data (Weikinn 1958–2002, partly edited at LIM).

3.2 Kernel density estimation

Excursion: explicit description of idea of kernel density estimation.

$$\widehat{f}_h(x) = 1/(nh) \sum_{i=1}^{n} K_h(x - x(i)), \qquad K_h(\cdot) = h^{-1} K(\cdot/h)$$

- continually shifted $\Rightarrow$ no class bound problems
- kernel function, K:
  - $\int K = 1$
  - (in most applications:) positive and symmetric
  - smooth (e. g., Gaussian) $\Rightarrow$ smooth estimate
  - histogram K = uniform kernel ("naive estimator" (Silverman 1986))

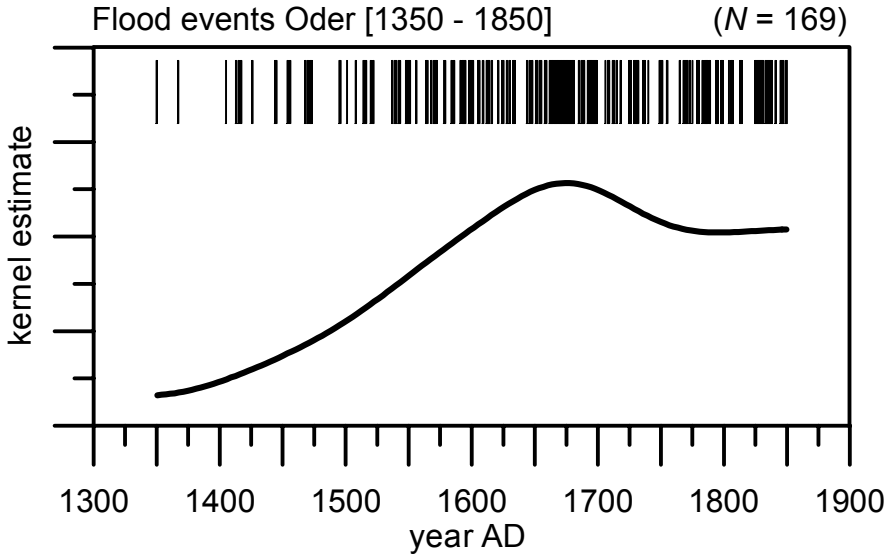

Figure 3.2: Floods river Oder, kernel estimation (K = Gaussian with std $\sigma = h =$ 50 years).

3.2.1 Smoothing problem

Choice of bandwidth, h, is crucial.

Excursion: kernel density estimation (h large/small)

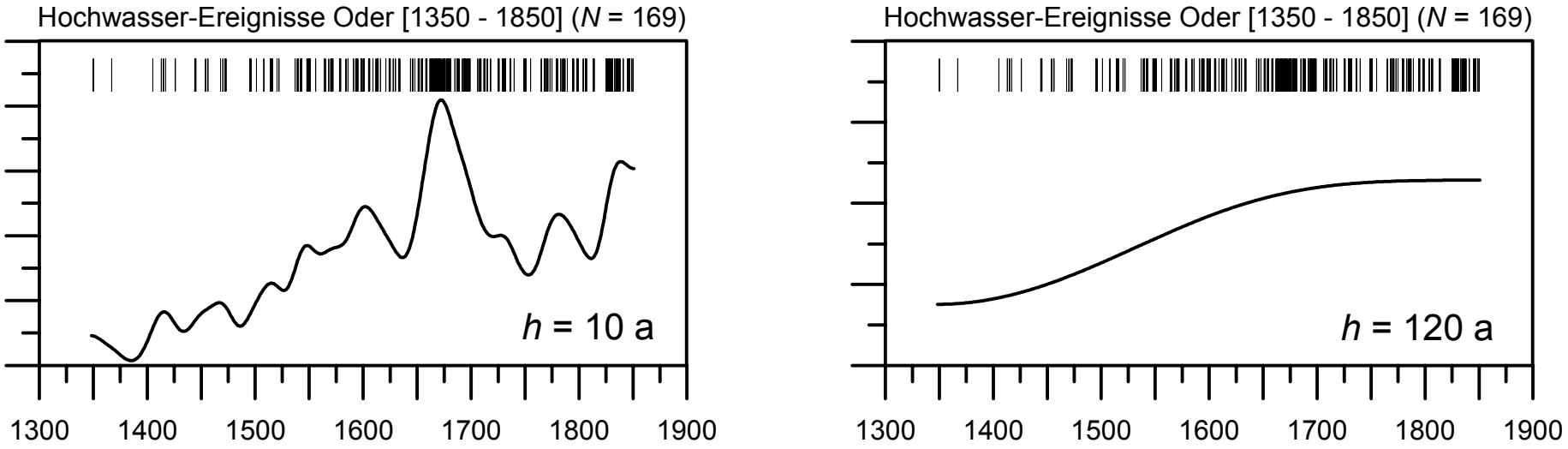

Figure 3.3: Floods river Oder, kernel estimation with large/small h.

- Too small h produces many spurious "wiggles."
- Too large h produces too flat estimate
- Also histogram exhibits smoothing problem.

How to choose h for a normal density?

Here we can only sketch how to obtain optimal h.

$$\begin{aligned} \text{Bias}[\widehat{f}(x)] &= E[\widehat{f}(x)] - f(x), \\ VAR[\widehat{f}(x)] &= f(x)/(nh) \ \ \text{(OK, we believe that)}, \end{aligned}$$

which gives mean-squared error, $\text{MSE}[\widehat{f}(x)] = \text{bias}^2 + VAR$.

Integrating MSE over x gives MISE (mean integrated squared error). MISE depends on σ of f_N and n, and is minimized by

$$h_{\text{opt, MISE}} = 1.059\sigma n^{-1/5}. \tag{3.1}$$

Plugging in an estimate for σ gives a data-based bandwidth choice.

3.3 Model suitability

We will not use kernel density estimates for estimating tail probabilities and related quantities, for example, return periods. For that purpose, kernel estimates are not accurate enough. However, they are useful for us to judge whether the model (i. e., the density function type) is suited for the data. For example to answer questions as: "Should we use a normal density model or a Weibull model?"

Another tool for assessing model suitability are Q–Q plots ("quantile–quantile"), also denoted as (normal) probability plots.

Excursion: f_N, F_N, empirical distribution function, transforming y-scale (note: uses numerical approximation).

KURZE PAUSE

Chapter 4

Extreme Value Distributions

We look at Johnson *et al.* (1995) Ch. 22.1–2, who describe that subject concisely.

4.1 Generalized Extreme Value or Extreme Value or von Mises type extreme value or von Mises–Jenkinson-type distribution

Excursion: Eq. 22.4 in Johnson *et al.* (1995).

4.1.1 Notation

Well, as often, there is no general agreement on that subject.

Johnson *et al.* (1995)	EXTREMES
ξ	μ
$1/\alpha$	γ
θ	σ/γ

4.2 Type I or doubly exponential, or Gumbel

Excursion: EXTREMES.

4.3 Type II or Fréchet

Excursion: EXTREMES.

4.4 Type III or Weibull

Excursion: EXTREMES.

4.5 Remarks

- We will estimate using maximum likelihood. Since the density functions are rather complex, minimization of the likelihood function is done numerically. Typically: guess start value, move in $-$gradient direction, stop when parameter-values changes are within a pre-defined precision. More details cannot be brought here

on estimation. :-(

- In my assessment, it would be nice if a meteorology student could remember the density formulae. But that is not mandatory. It is sufficient if he or she would know where to look up them.

Guten Appetit!

AFTERNOON PART: PC PROGRAM *EXTREMES*

Chapter 5

Extreme Time Series

5.1 Time series

A time series is a set of observations, $x(i)$, of a random variable, $X(i)$, made at time $t(i)$.

$\{x(i), t(i), i = 1, \ldots, n\}$.

Geosciences: discrete time, continuous X.

Equidistance: $t(i) - t(i-1) = d = \text{const.}$ (d, spacing).

5.2 Elbe runoff time series

Elbe:

- Catchment area at station Dresden (Augustusbrücke): 53096 km^2
- Low-mountainous climate (Sudeten Mountains, Erzgebirge, Bohemia)
  - Floods: stationary depression over catchment area required
  - Winter floods (November to April): enhanced by breaking river ice

- Runoff Q [m^3/s]
  - $\Rightarrow$ Precipitation
  - $\Rightarrow$ Land-use
    - $*$ Interception
    - $*$ Evapotranspiration
    - $*$ Infiltration
  - $\Rightarrow$ River engineering: dikes, reservoirs
  - $\Rightarrow$ Runoff efficiency: Q/precipitation
- Runoff less influenced by river geometry than water level

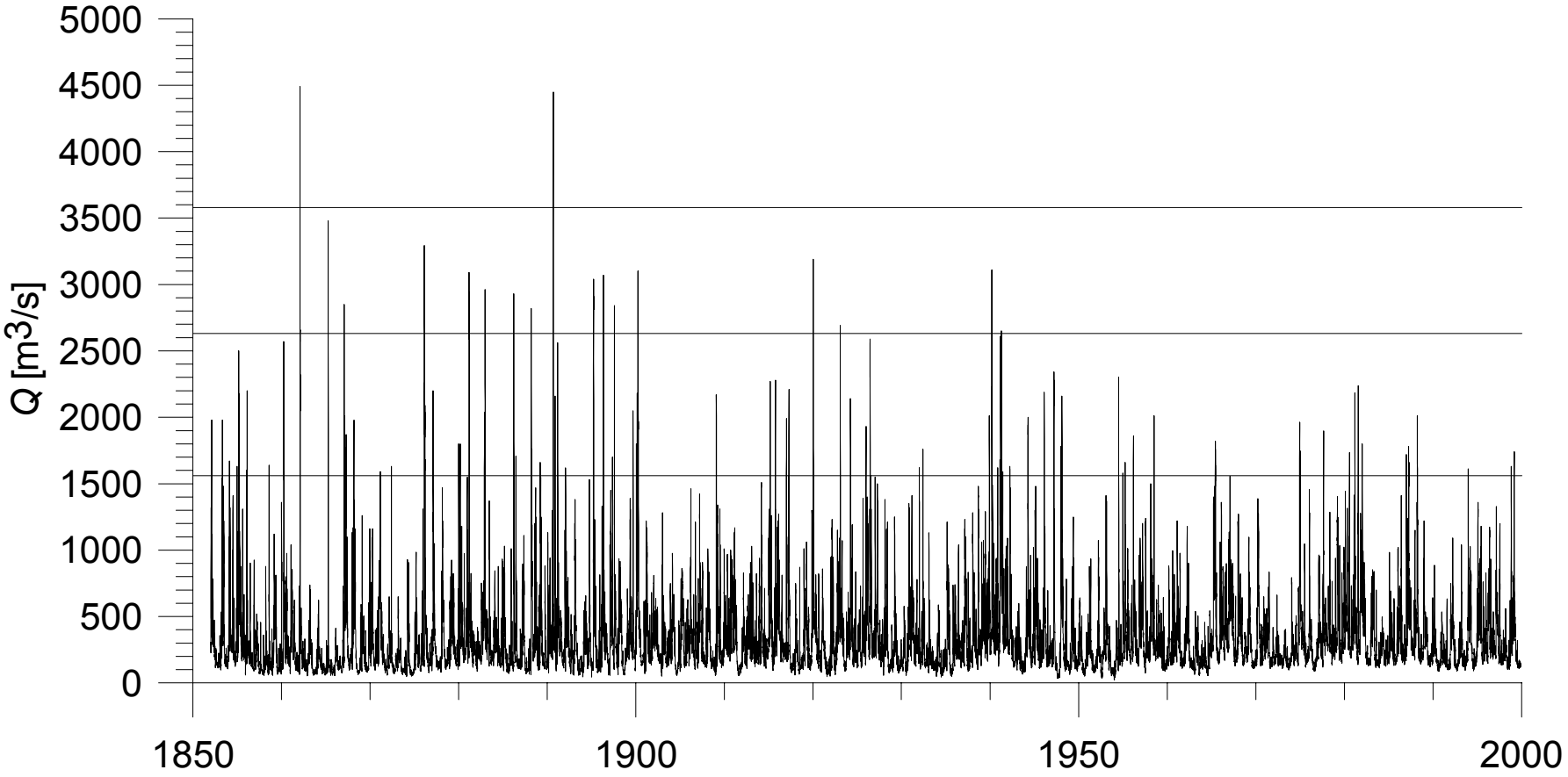

Figure 5.1: Daily runoff Elbe (Dresden), whole interval (1852–2000) ($n = 54020$). Data from Global Runoff Data Centre (Koblenz, Germany). Shown also are thresholds (1560, 2630 and 3580 m^3/s) used to define floods (3 magnitude classes).

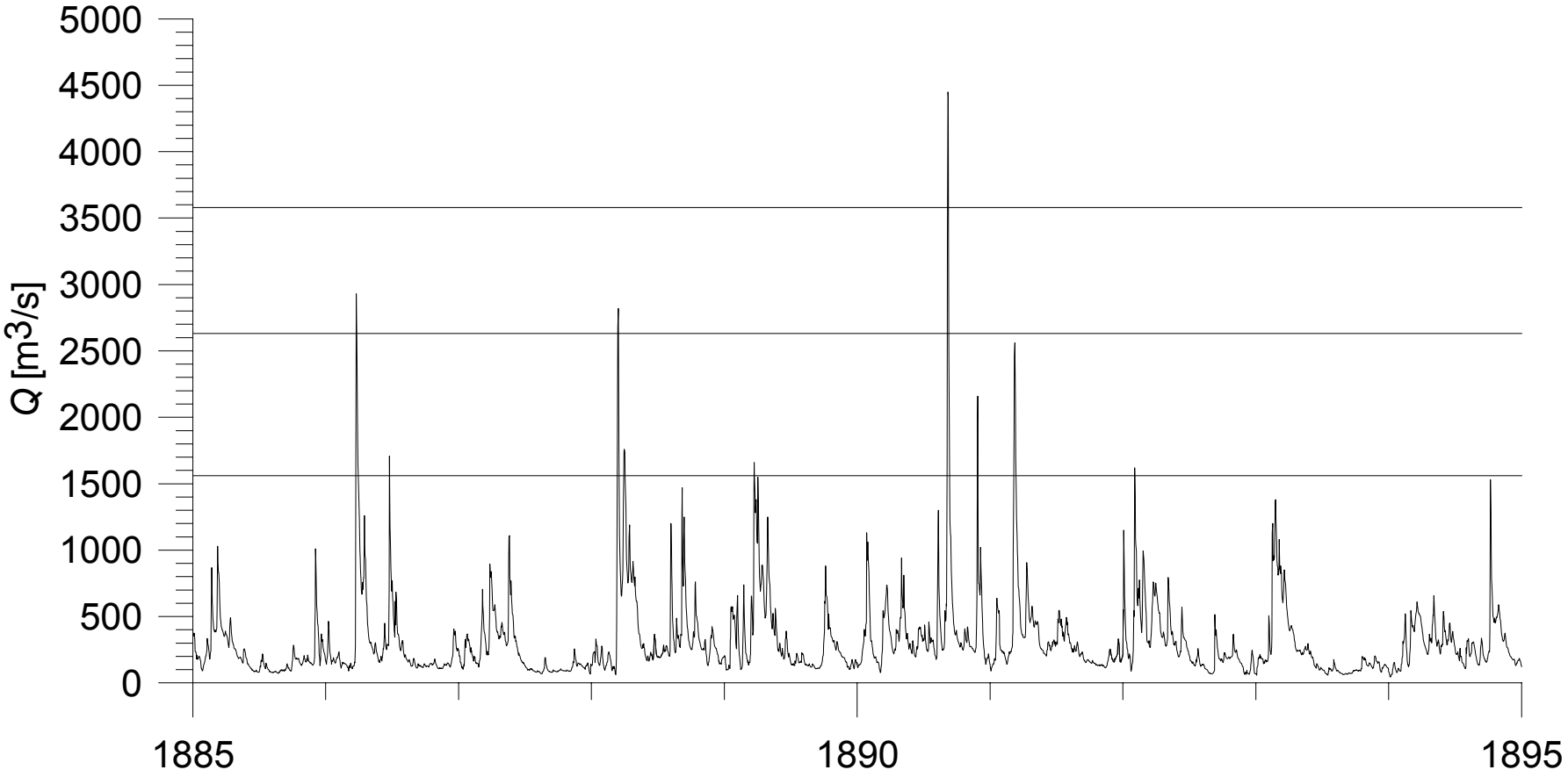

Figure 5.2: Daily runoff Elbe (Dresden), 1885–1895.

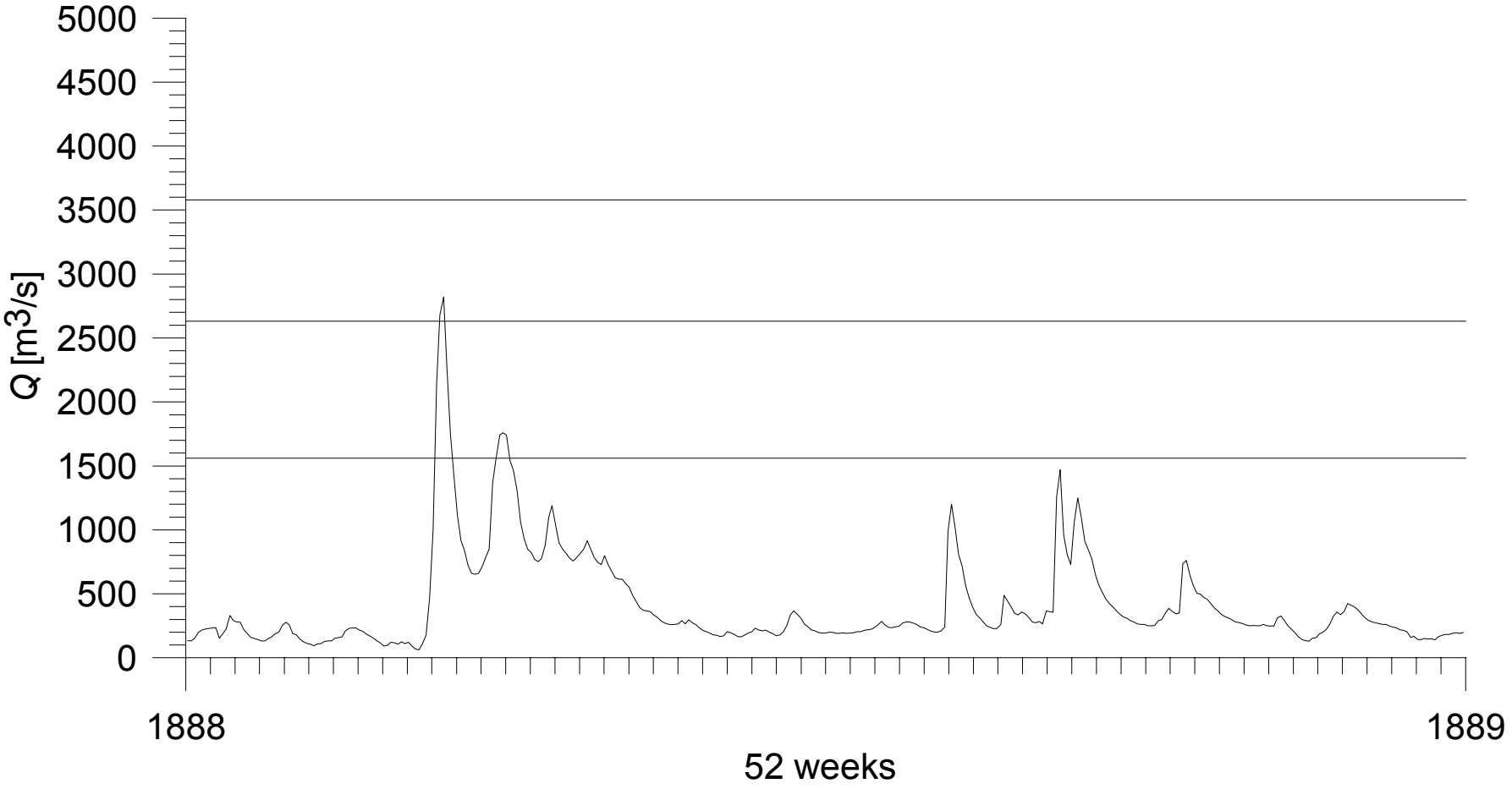

Figure 5.3: Daily runoff Elbe (Dresden), 1888.

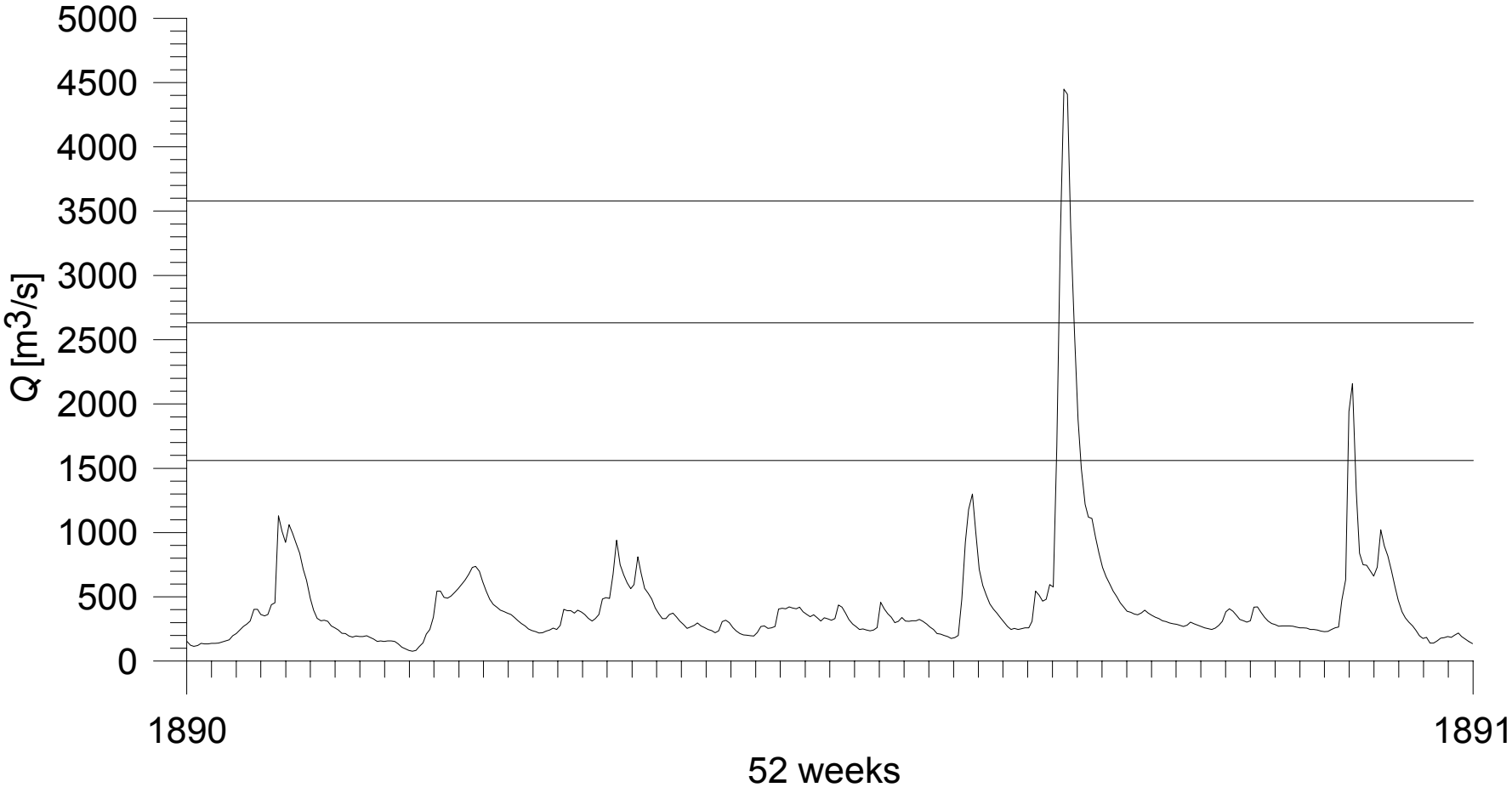

Figure 5.4: Daily runoff Elbe (Dresden), 1890.

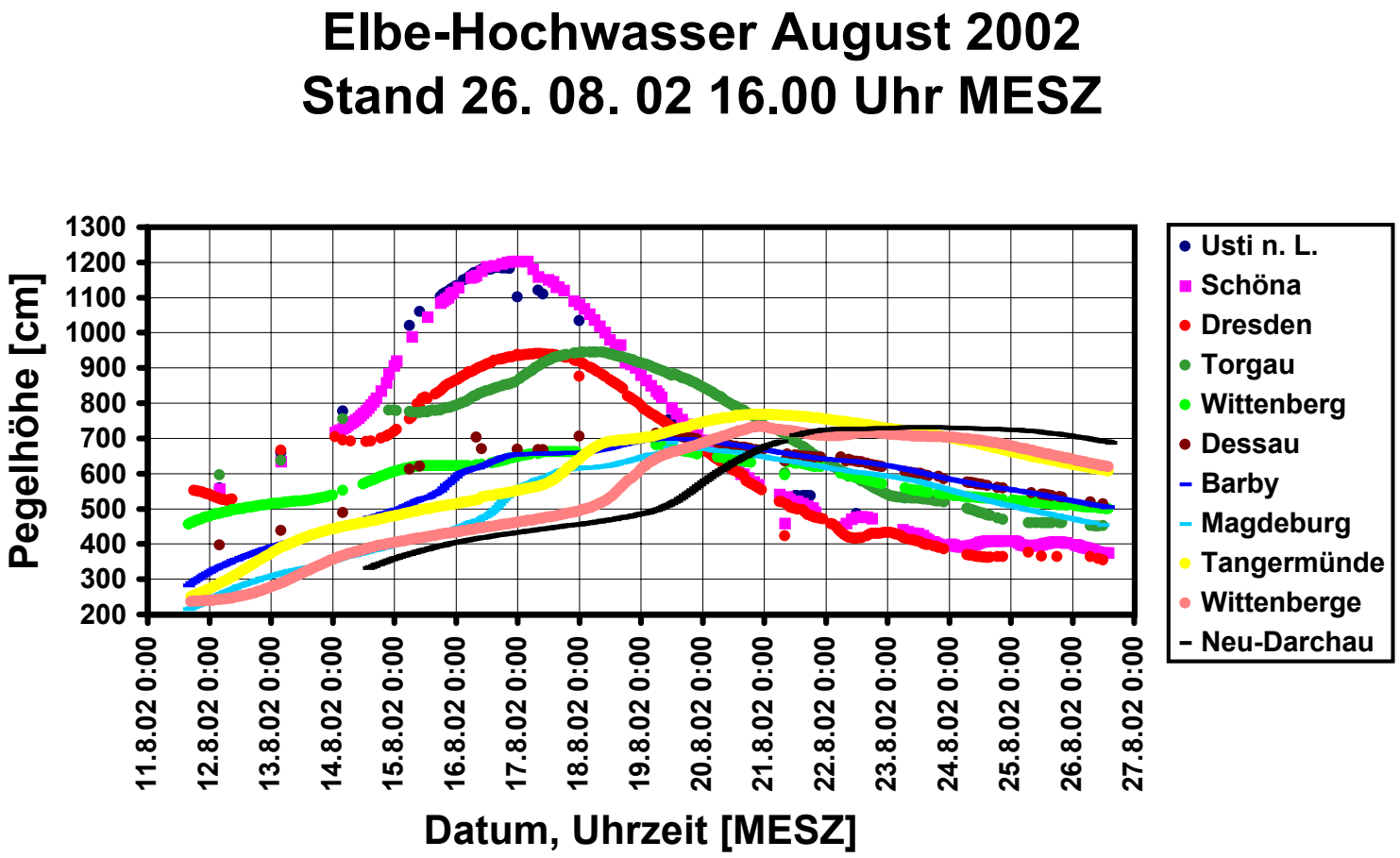

Figure 5.5: Elbe flood August 2002. Note: the maximum water level at Dresden (940 cm) corresponds (measured) to $Q \approx 5000$ m^3/s. Data from `http://www.wetteronline.de`.

5.3 Return period and T-year event

> Return period: "Expected number of observations [x; made at times with spacing d] of a variable [X] needed to obtain one observation in excess of a specified quantity, θ."

Kotz and Johnson (1988), notes by M.M. in brackets

Excursion: calculation of return period for IID random variable, X.

Note: "IID" means "independent and identically distributed."

Note: No agreement exists on how to denote return period. We use $T_{\text{return},\,\theta}$.

Result:

$$T_{\text{return},\,\theta} = d \cdot 1/\left(1 - F(\theta)\right).$$

Invert relation: T-year event.

$$T_{\text{return},\,\theta} = T' \Rightarrow \theta = \theta_{T'}.$$

Hydrology: θ_T for $T' = 100$ years denoted as $HQ100$, etc.

5.4 IID assumption

First, winter and summer floods should be treated separately because they are influenced/caused by different processes.

Second, to calculate the return period for Elbe floods using above formula, the IID assumption has to be satisfied. We have to check two points: (1) independence and (2) identical distribution.

5.4.1 Independence

See above: floods may last a few weeks. Therefore, we take **bimonthly maxima** from the daily Q time series. Eight weeks are likely long enough $\Rightarrow$ bimonthly maxima are likely independent.

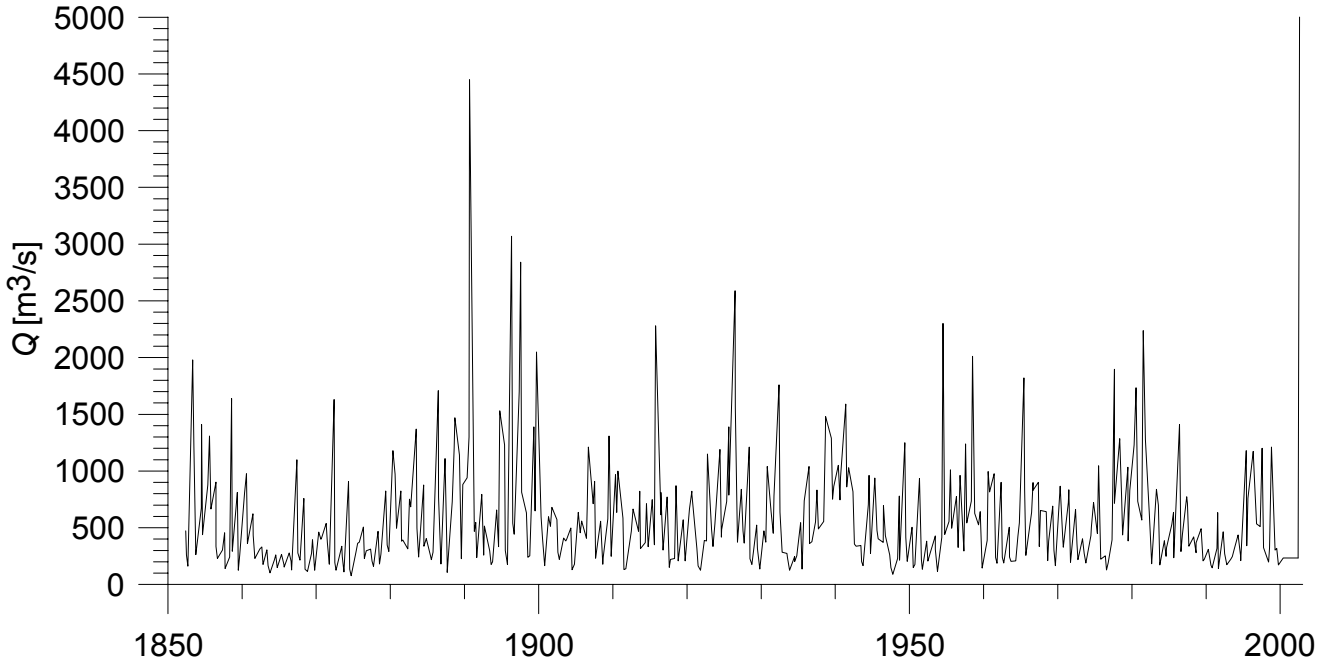

Figure 5.6: Bimonthly maximum runoff Elbe (Dresden), summer, 1852–2002.

5.4.2 Identical distribution: stationarity

Winter floods of the Elbe at Dresden, 1852–2002, can be shown to exhibit **nonstationarity**, that means, the distribution of X changes with time. To detect nonstationarity, the statistical test after Cox and Lewis (described in Mudelsee (2002)) can be used. Since the winter flood time series likely reflects nonstationarity, we do not calculate return periods from it. More details cannot be brought here.

To investigate stationarity of summer floods, we compare the distributions for two time intervals: 1852–1919 and 1920–2002.

Excursion: EXTREMES: comparison of time intervals.

Finally, we estimate the 100-year runoff and the return period for summer floods of the Elbe at Dresden, using bimonthly maxima, ML estimation of the parameters of a Generalized Extreme Value distribution.

Excursion: EXTREMES: calculation of 100-year runoff.

Result:

$$HQ100 = 3880 \text{ m}^3/\text{s}.$$

Excursion: EXTREMES: calculation of return period for August 2002 flood.

Result:

$$T_{\text{return, } 5000 \text{ m}^3/\text{s}} = 168 \text{ years}.$$

BEFORE YOU LEAVE

Chapter 6

Further Topics

- Estimation uncertainty of parameters: confidence intervals, bootstrap resampling
- Hypothesis tests for model suitability
- Nonstationarity—time-dependent density function: estimate occurrence rate of extreme events over time (done here at LIM in case of Elbe and Oder floods; M Mudelsee)

Chapter 7

Literature

Bronstein IN, Semendjajew KA (1980) Taschenbuch der Mathematik. 19. Edn., Harri Deutsch, Thun, 860 pp. [where you look up formulae, cheap!?]

Johnson NL, Kotz S, Balakrishnan N (1994) Continuous Univariate Distributions. Vol. 2, 2nd Edn., Wiley, New York, 719 pp. [series contains everything on distributions, this Vol. contains Extreme Value distributions; expensive]

Kotz S, Johnson NL (Eds) (1988) Encyclopaedia of statistical sciences. Vol. 8, Wiley, New York, 870 pp. [contains everything about statistics (excluding recent

research), readable, very expensive]

Mudelsee M (2002) XTREND: A computer program for estimating trends in the occurrence rate of extreme weather and climate events. In: Raabe A, Arnold K (Eds.) Wissenschaftliche Mitteilungen aus dem Institut fr Meteorologie der Universitt Leipzig. Vol. 26, Institute for Meteorology, University Leipzig, 149–195. [`http://www.uni-leipzig.de/~meteo/MUDELSEE/publ/pdf/xtrend.pdf`, contains description of statistical test of stationarity hypothesis]

Mudelsee M (2003) Statistical Methods in Meteorology—Time Series Analysis. University of Leipzig, Lecture Notes. [`http://www.uni-leipzig.de/~meteo/MUDELSEE/teach/stats.pdf`, contains material on time series analysis and also on statistics in general]

Reiss R-D, Thomas M (1997) Statistical Analysis of Extreme Values. Birkhäuser, Basel, 316 pp. [not optimal from a didactical point, contains EXTREMES program]

Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman

and Hall, London, 175 pp. [extremely readable]

Simonoff JS (1996) Smoothing Methods in Statistics. Springer, New York, 338 pp. [contains other types of smoothing besides density estimation, readable]

von Storch H, Zwiers FW (1999) Statistical Analysis in Climate Research. Cambridge University Press, Cambridge, 484 S. [authoritative as regards the title, many examples, paperback version exists!?]

Weikinn, C (1958–2002) Quellentexte zur Witterungsgeschichte Europas von der Zeitwende bis zum Jahre 1850: Hydrographie, Parts 1–4 (Akademie-Verlag, Berlin, 1958–1963), Parts 5–6 (Eds. Börngen M and Tetzlaff G) (Gebrüder Borntraeger, Berlin, 2000–2002). [contains about 23160 entries, primary and secondary sources, high coverage for Germany and neighbours; internet version in preparation; if you are interested: ask Dr Börngen :-)]