Extreme Value Analysis for Climate Time Series

Whitney Huang

UVic/ CANSSI & SAMSI

STAT 457/554: Time Series Analysis
November 7, 2018

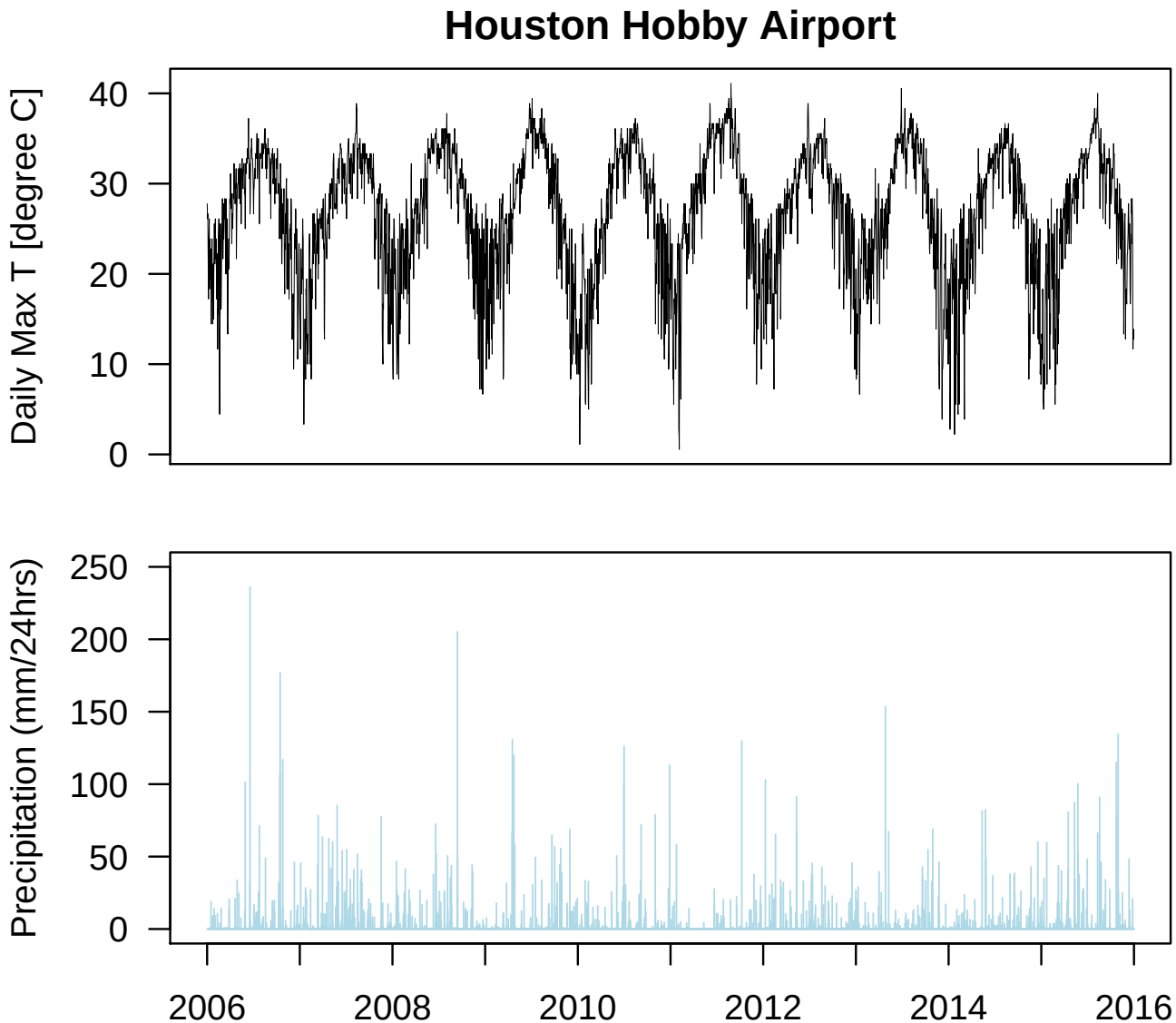
University of Victoria

Outline

- Background: Climate time series
- Extreme Value Analysis
- A new approach to modeling extremes

Part I: Background

Examples of time series of climate variables



Climate time series

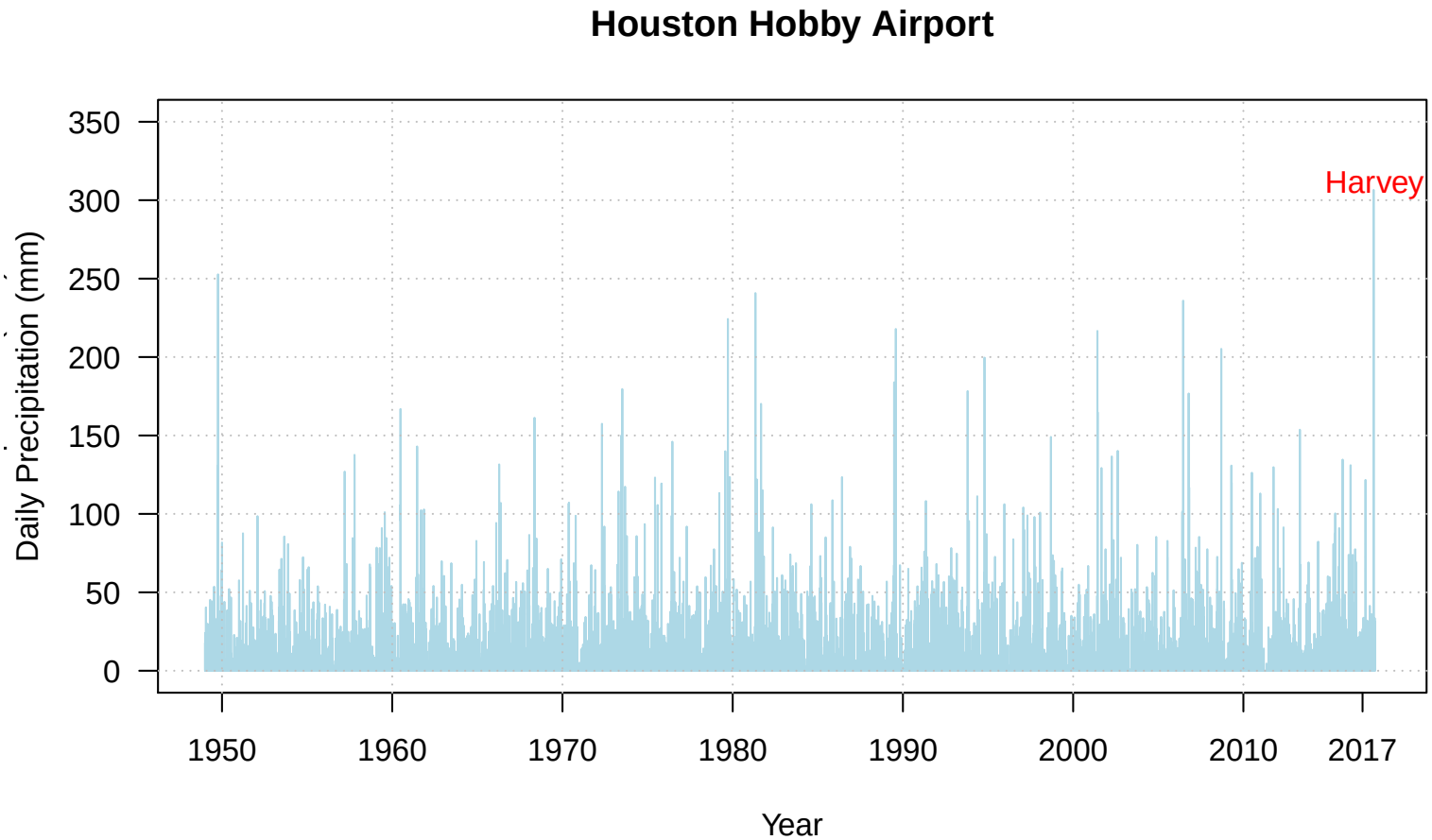
- "Climate is what you expect, but weather is what you get." – attributed to many different authors, including Mark Twain

> Statistical point of view: Climate is the underlying **distribution** that generates weather events

- Common features: seasonality, temporal dependence, and maybe long term trend due to external forcings (e.g., greenhouse gas emissions)

- Today I am going to focus on climate extremes, i.e., the **tail distribution of a climate variable** (e.g., annual maximum daily rainfall)

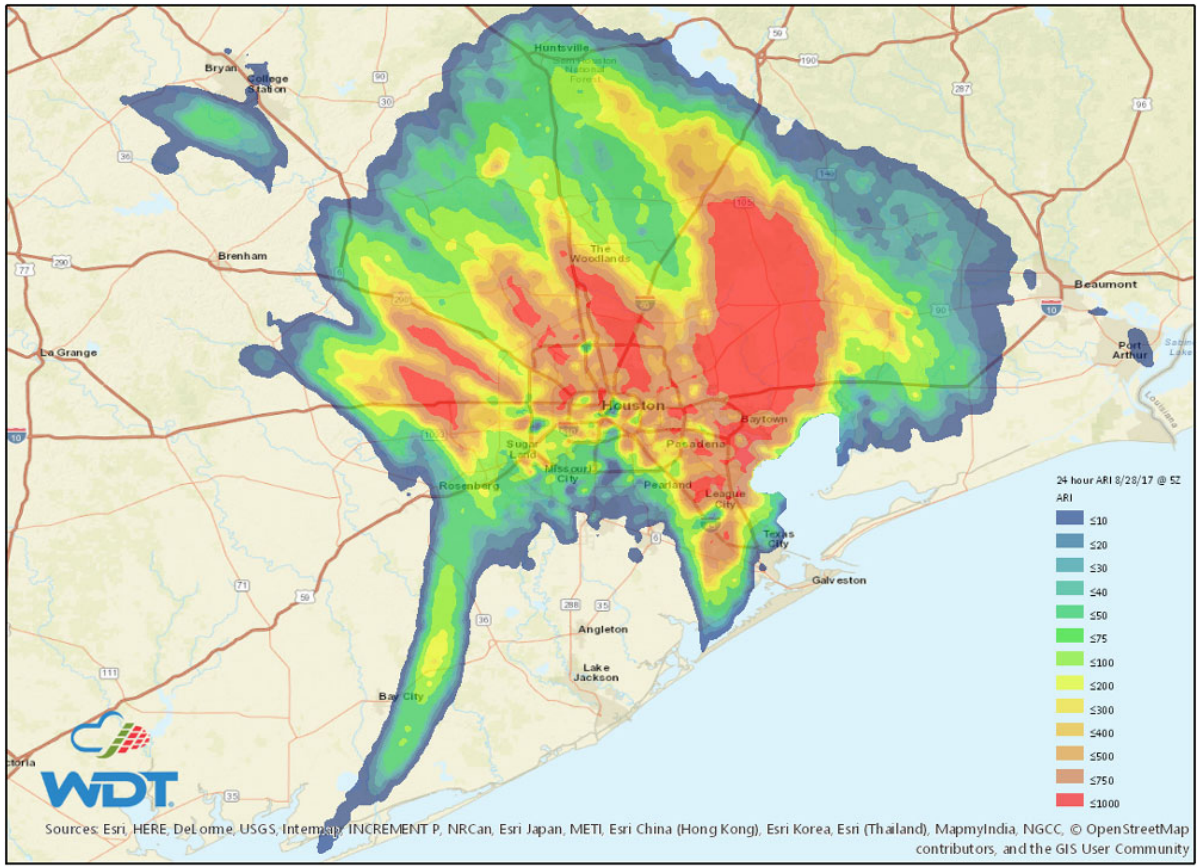
Houston daily precipitation (data source: GHCN)



How to quantify climate extremes?

r-year return level (RL_r): the value whose probability of exceedance is $1/r$ in any given year

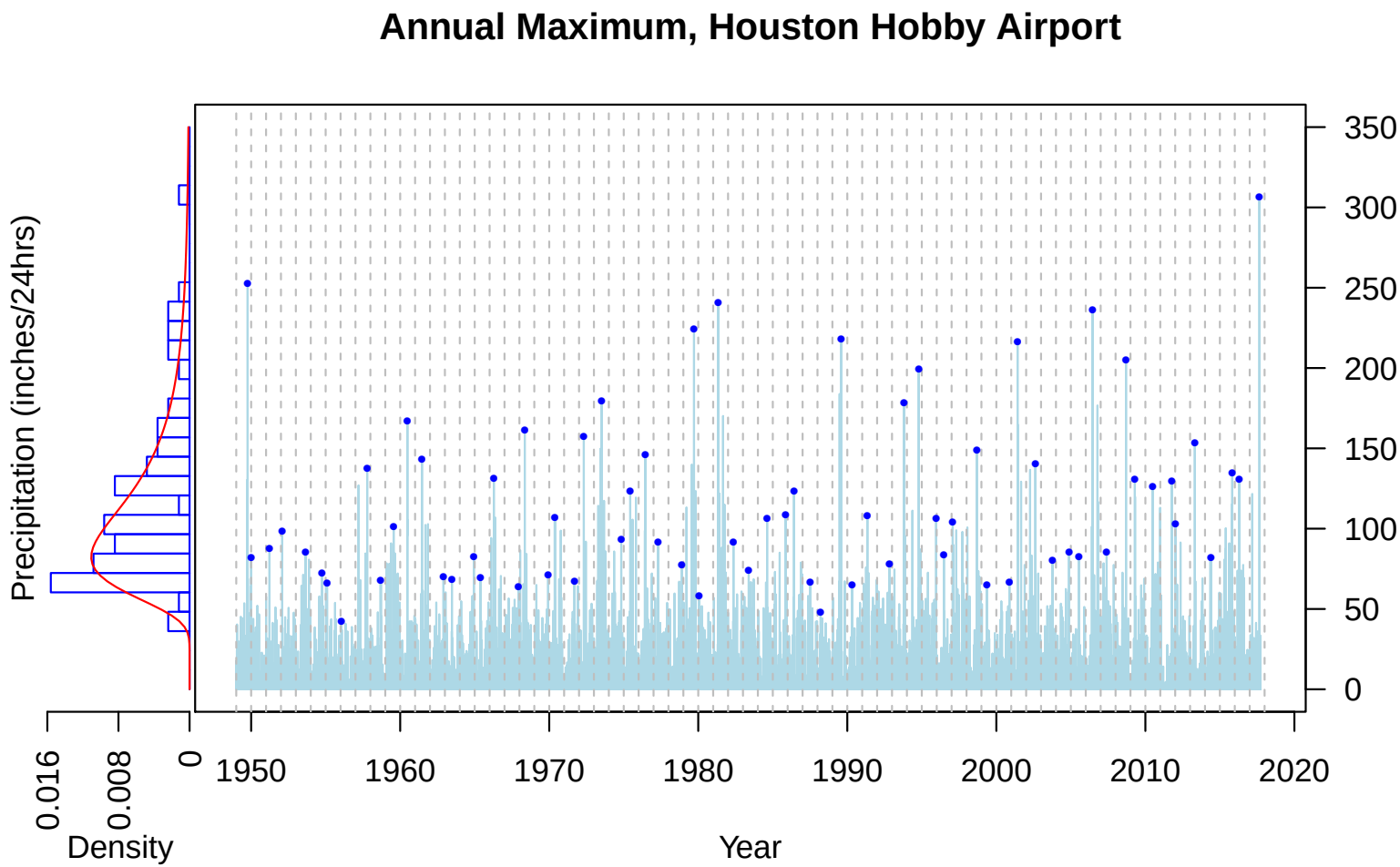
Hurricane Harvey



- "A storm forces Houston, the limitless city, to consider its limits" – The New York Times (8.31.17)
- "Storm Harvey breaks record of rainfall and worsens disaster" – La Nacion (8.30.17)

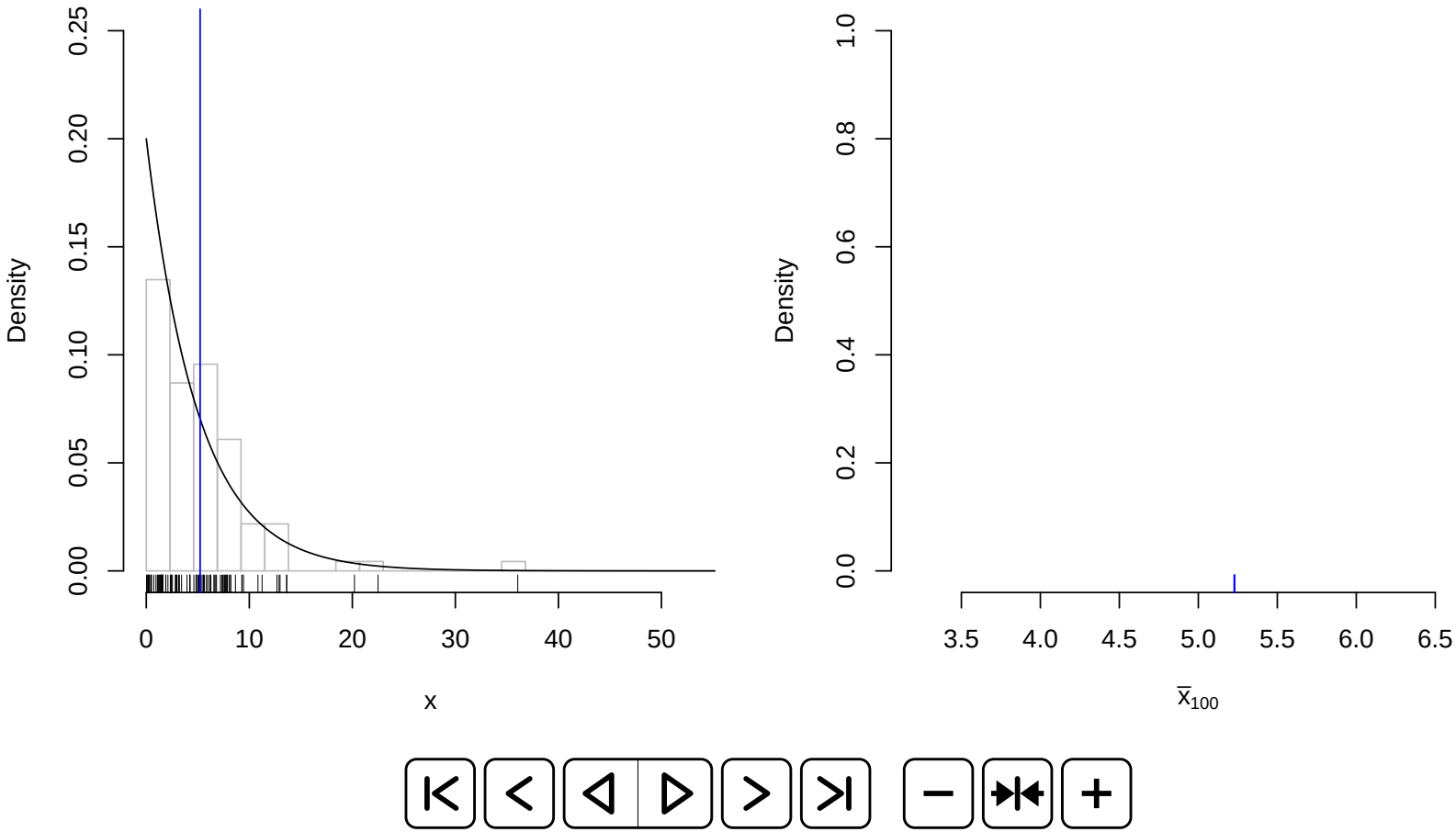
Part II: Introduction to extreme value theory

Estimating extremes using block maxima (Gumbel 1958)

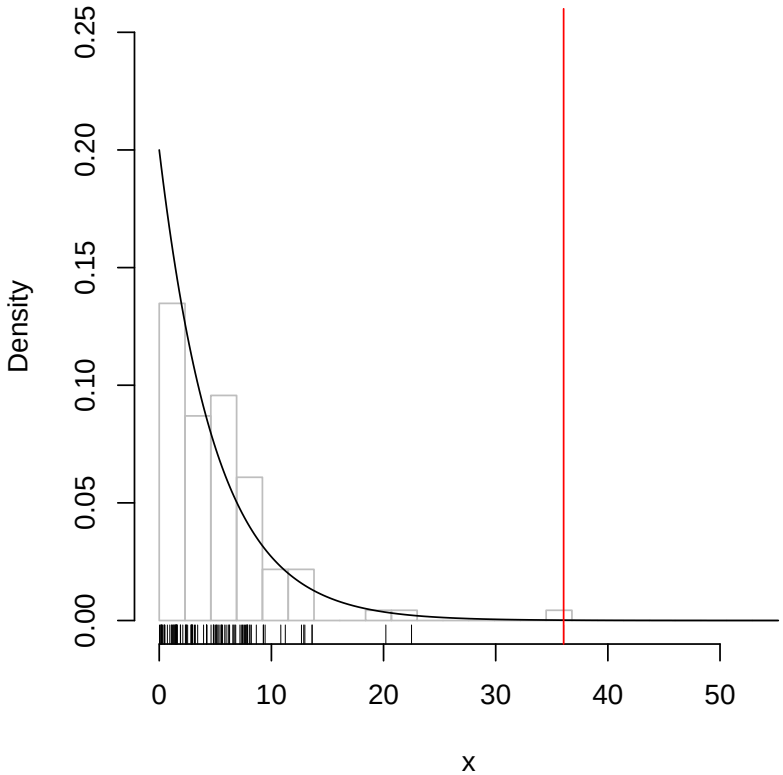


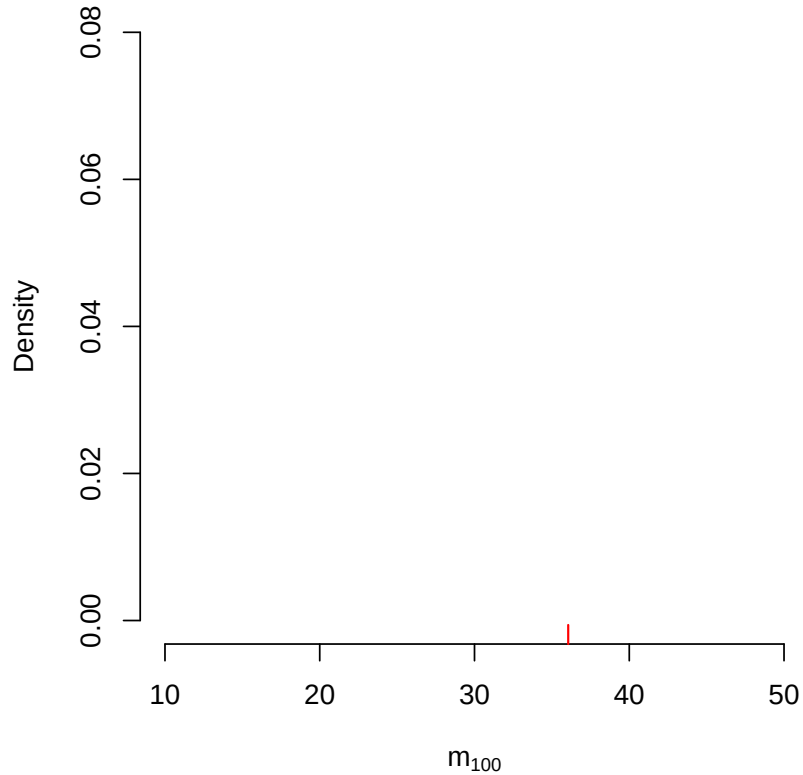
Which distribution to use for annual maxima?

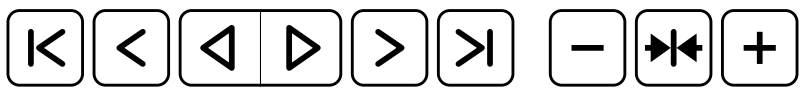
CLT: Normal distribution for sample means



Which distribution to use for sample (block) maxima?

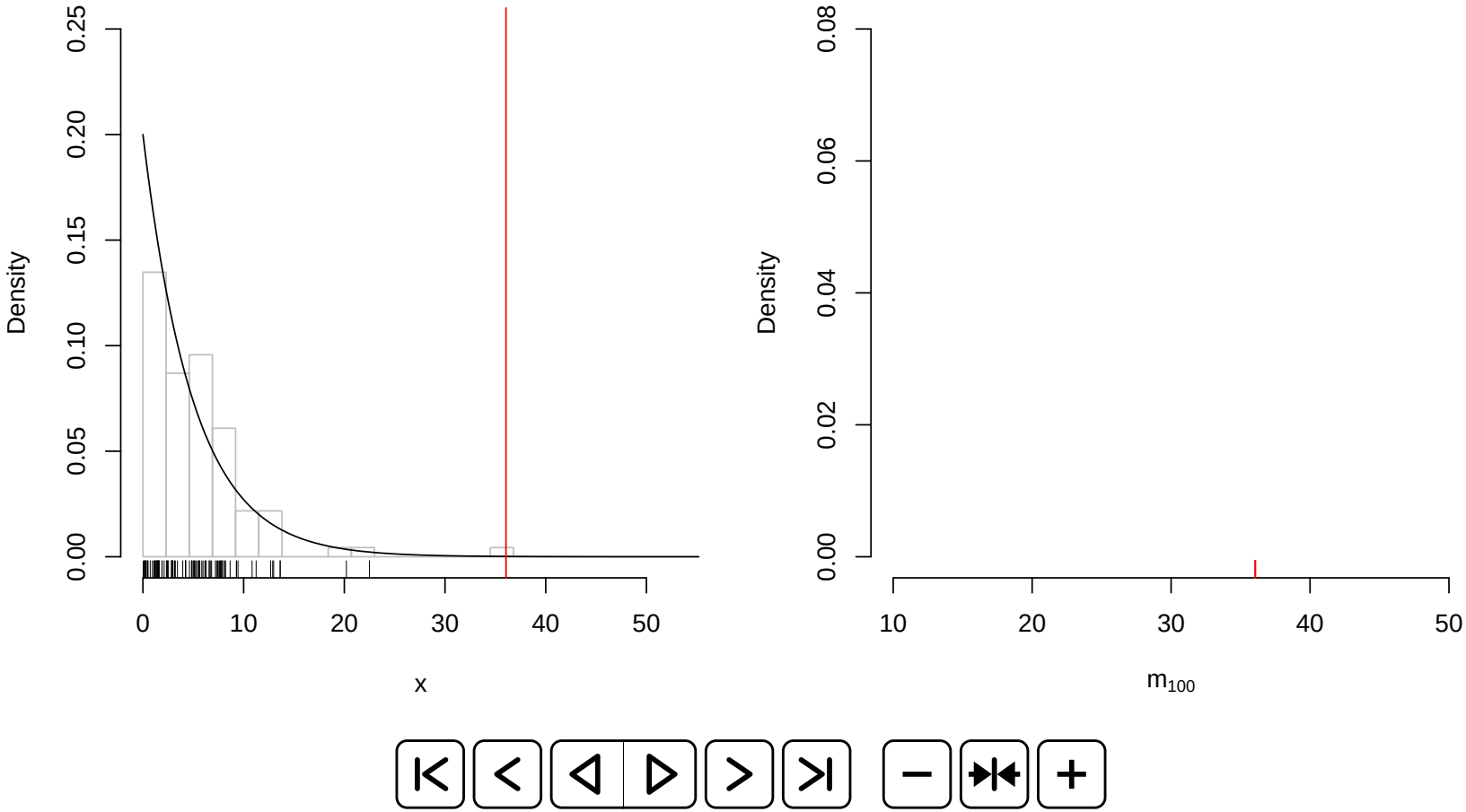






Use the Generalized Extreme Value (GEV) distribution

Which distribution to use for sample (block) maxima?



Use the Generalized Extreme Value (GEV) distribution

Extremal types theorem (Fisher–Tippett 1928, Gnedenko 1943)

Define $M_n = \max\{X_1, \cdots, X_n\}$ where $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$. If $\exists\, a_n > 0$ and $b_n \in \mathbb{R}$ such that, as $n \to \infty$

$$\mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \xrightarrow{d} \mathsf{G}(x)$$

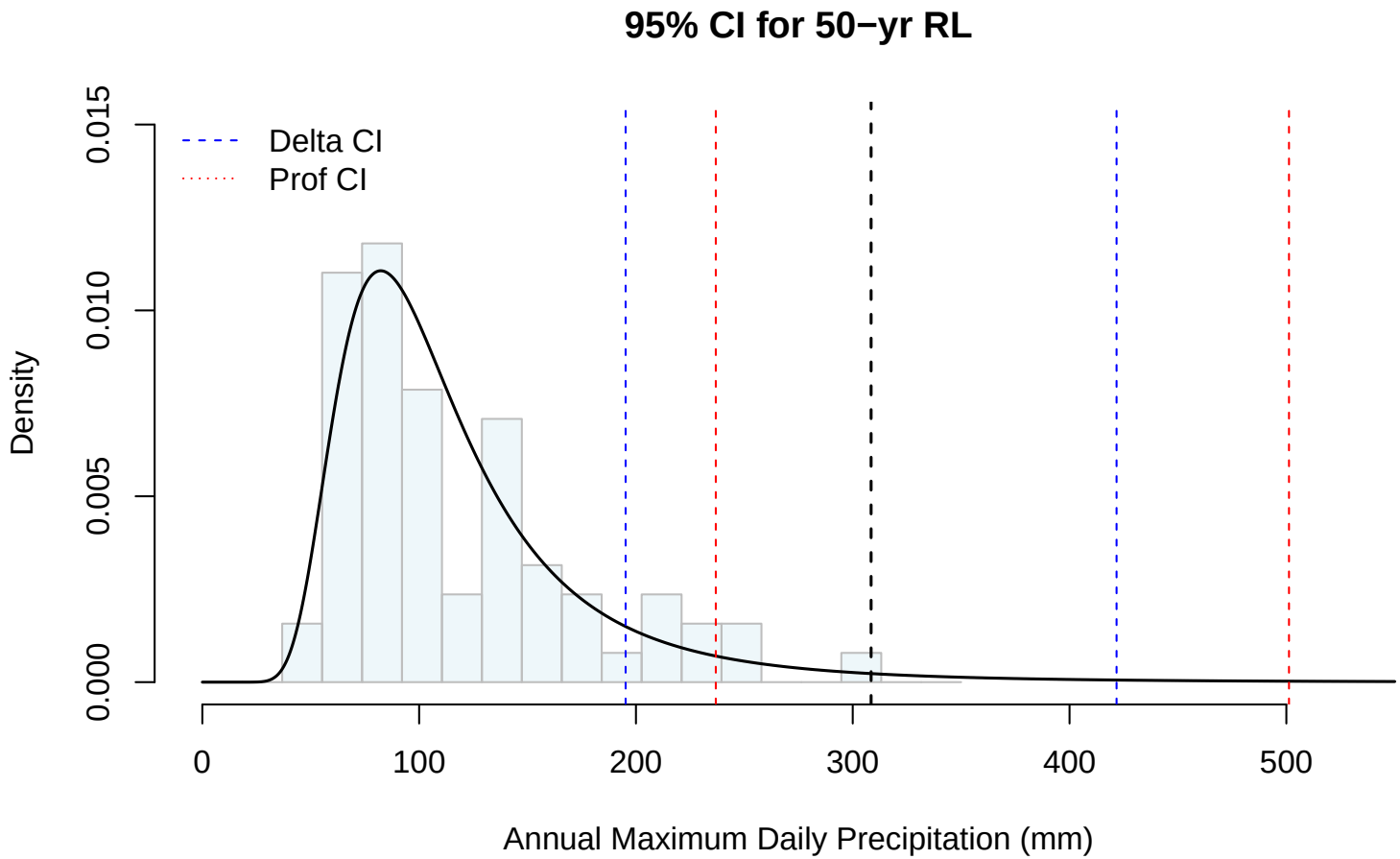
then G must be the same type of the following form:

$$\mathsf{G}(x; \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{\frac{-1}{\xi}} \right\}$$

where $x_+ = \max(x, 0)$ and $G(x)$ is the distribution function of the generalized extreme value distribution (GEV)

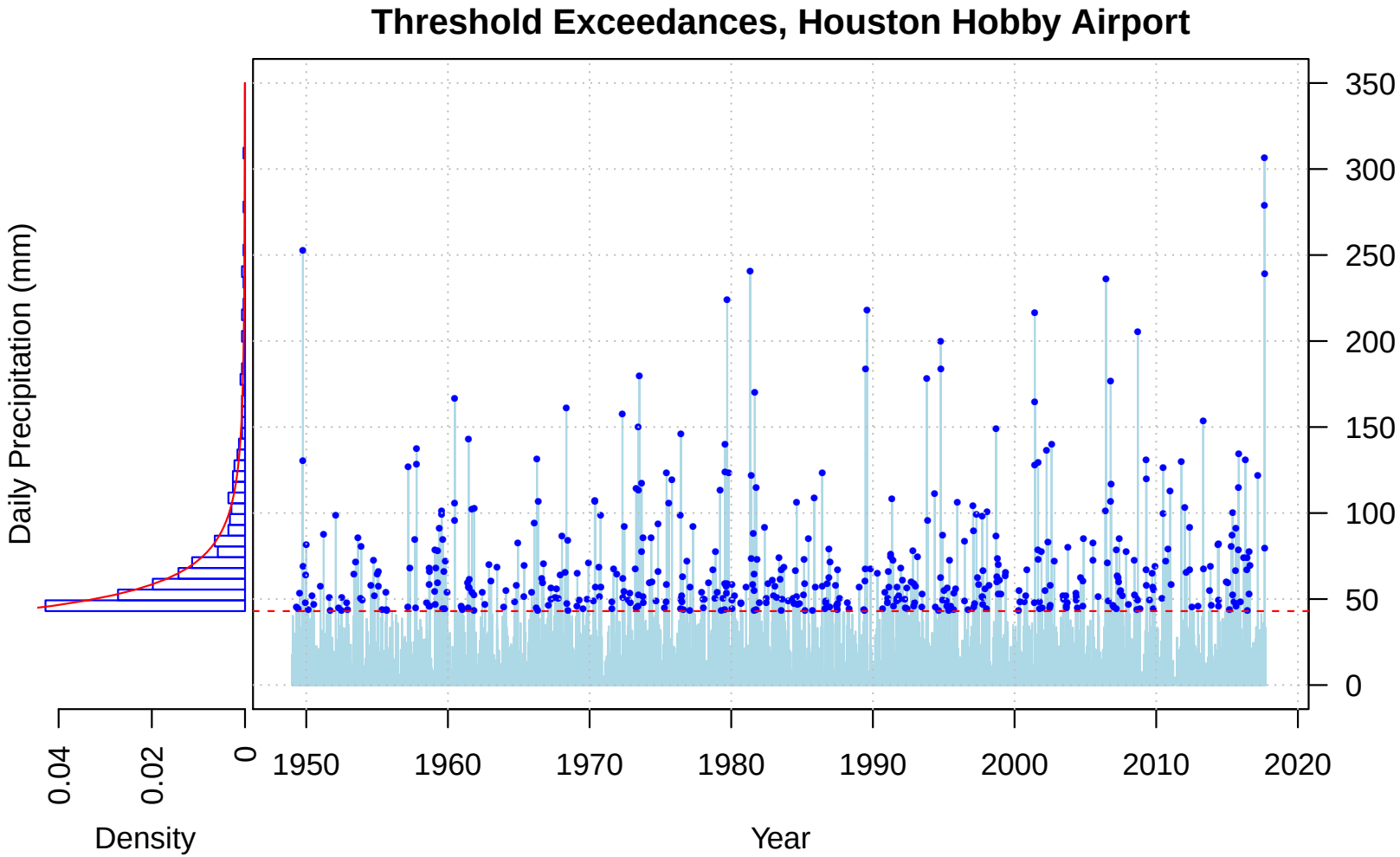
- μ and σ are location and scale parameters
- ξ is a shape parameter determining the rate of tail decay, with
  - $\xi > 0$ giving the heavy-tailed case (Fréchet)
  - $\xi = 0$ giving the light-tailed case (Gumbel)
  - $\xi < 0$ giving the bounded-tailed case (reversed Weibull)

Inference for 50-year return level: Block maxima method



$\hat{\mu} = 89.37(4.74), \hat{\sigma} = 34.12(3.88), \hat{\xi} = 0.24(0.11)$

Peaks–over–threshold (POT) method [Davison & Smith 1990]



Which distribution to use for threshold exceedances?

$\Rightarrow$ Generalized Pareto distribution ($\text{GPD}_u(\tilde{\sigma}, \xi)$)

Pickands (1975)–Balkema–de Haan (1974) theorem

If $M_n = \max\{X_1, \cdots, X_n\} \approx \text{GEV}(\mu, \sigma, \xi)$, then, for a large u,

$$\mathbb{P}(X > u) \approx \frac{1}{n}\left[1 + \xi\left(\frac{u-\mu}{\sigma}\right)\right]^{\frac{-1}{\xi}}$$

and $F_u = \mathbb{P}(X - u < y | X > u)$ is well approximated by the generalized Pareto distribution (GPD). That is:

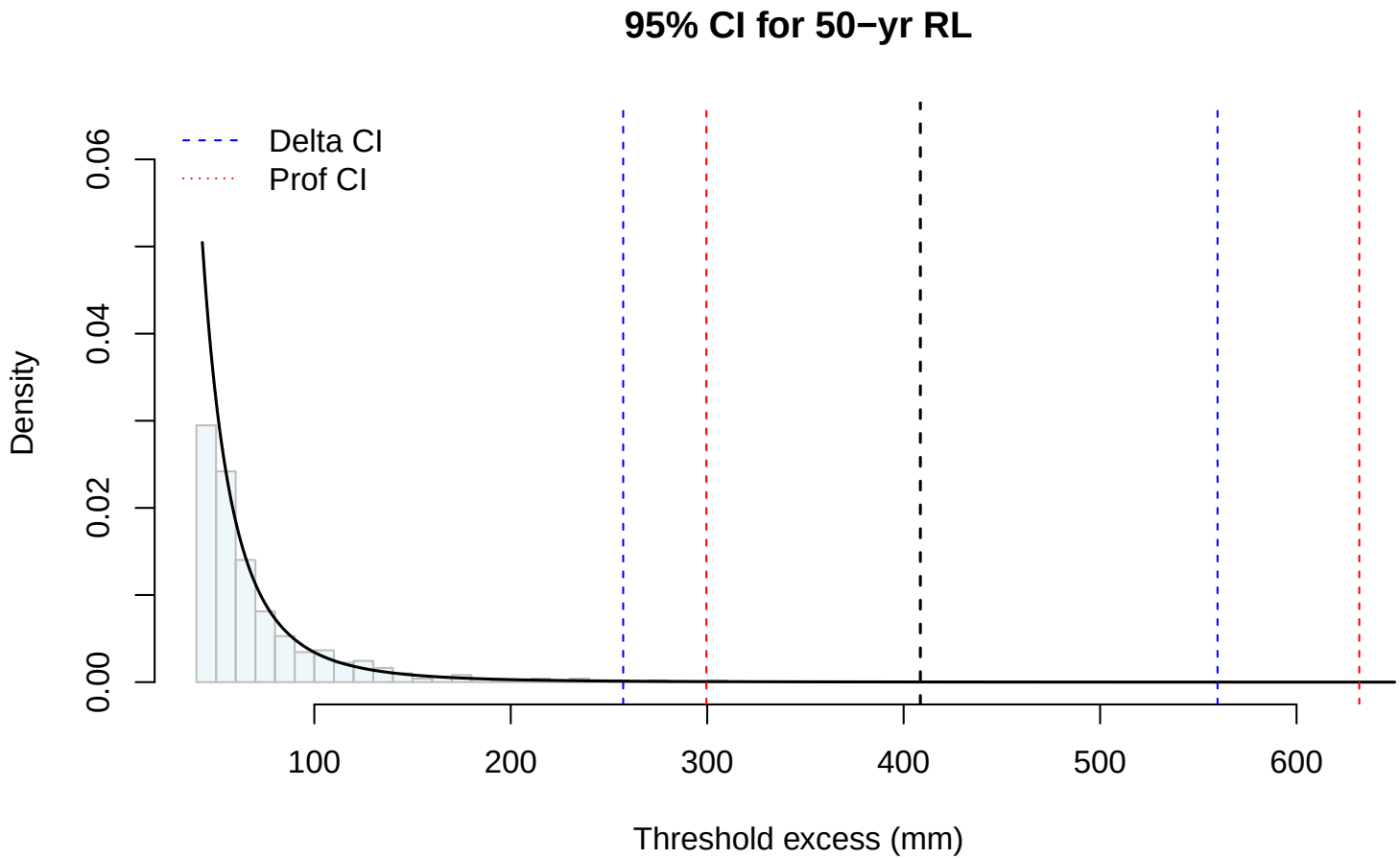
$$F_u(y) \xrightarrow{d} \mathsf{H}(y) \qquad u \to u_F$$

where

$$\mathsf{H}(y) = \begin{cases} 1 - (1 + \xi y/\tilde{\sigma})^{-1/\xi} & \xi \neq 0; \\ \\ 1 - \exp(-y/\tilde{\sigma}) & \xi = 0. \end{cases}$$

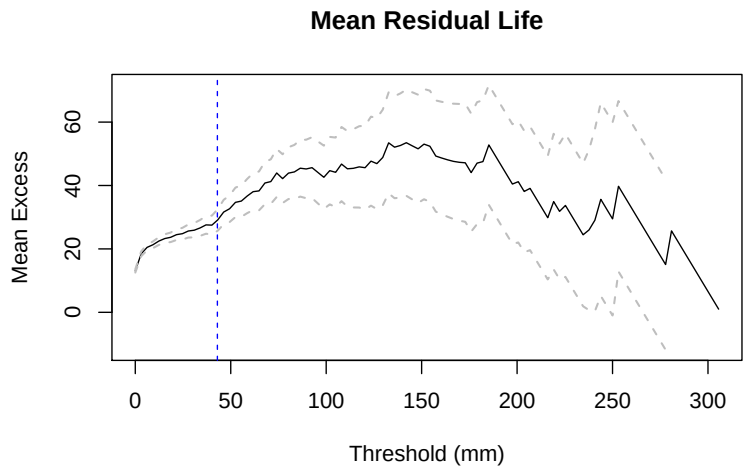
and $\tilde{\sigma} = \sigma + \xi(u - \mu)$

Inference for 50-year return level: POT method



$u = 43, \hat{\sigma}_u = 19.81(1.56), \hat{\xi} = 0.34(0.07)$

Fit GPD to threshold exceedances: How to choose an "appropriate" u?



Bias–variance tradeoff:

- Threshold too low $\Rightarrow$ bias because of the model asymptotics being invalid
- Threshold too high $\Rightarrow$ variance is large due to few data points

It is not easy to choose the "right" threshold but the estimates might be sensitive to the chosen threshold

Part III: Estimating Precipitation Extremes using the Log-Histospline [H., Nychka, Zhang (2018+), Environmetrics]

Motivation: To develop an alternative to POT method

Wish List

- Bypass the threshold selection
  - No need to select the "correct" threshold
  - Make use of data more efficiently by including the bulk of data
- Incorporate prior information on the tail of rain distributions
  - Polynomial (right) tail behavior

Model the full range of the distribution while accounting for GPD tail behavior with $\xi > 0$

Basic idea: Model the log density as a natural cubic spline

Let Y be a non-negative random variable and $X = \log(Y)$ with density $f(x), x \in \mathbb{R}$. We assume $f(x) = e^{g(x)}$

- Working on log density makes life easier

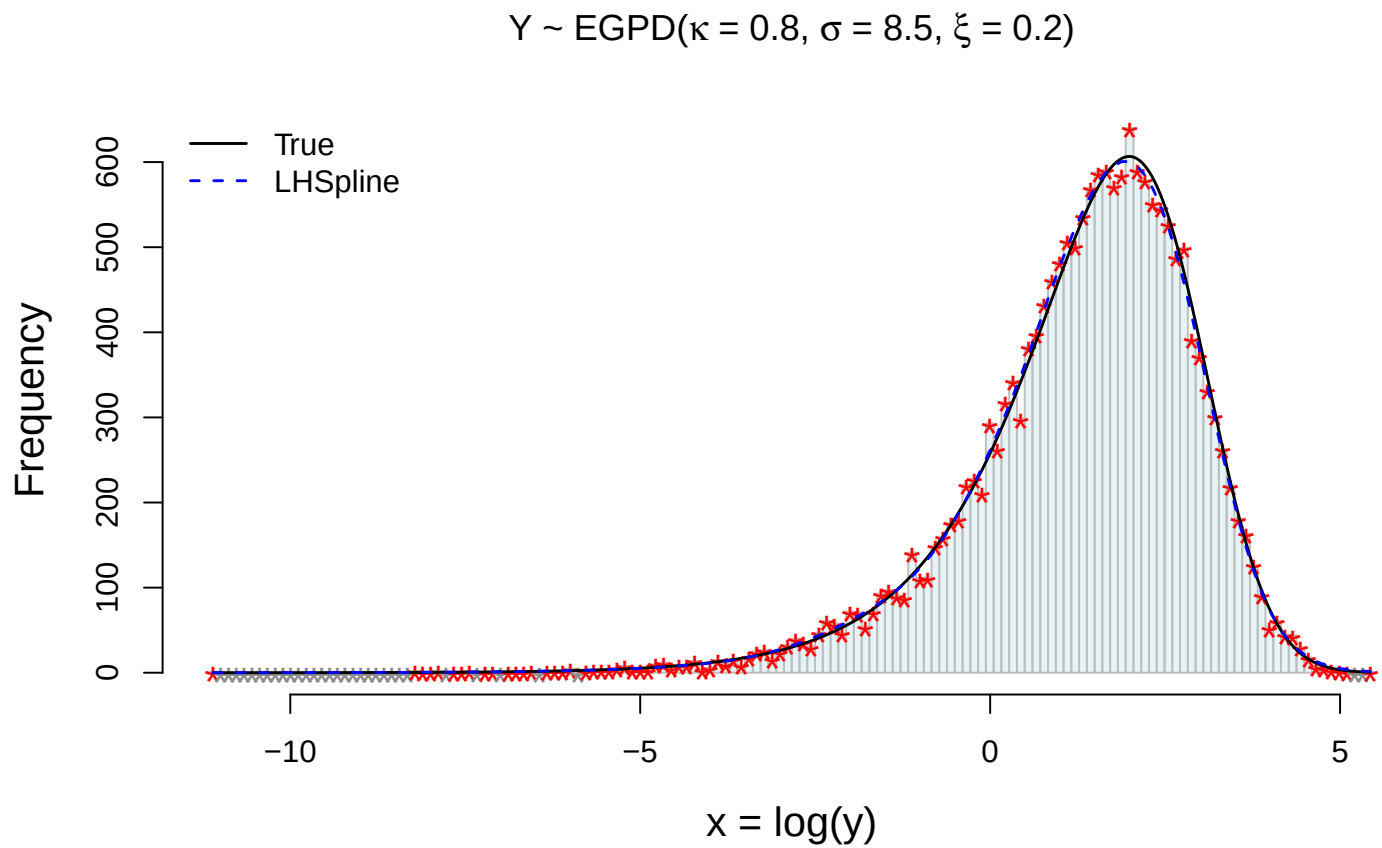
  $\Rightarrow$ remove the positivity constraint

- Estimate $g(x)$ as a flexible (cubic) spline function
  - $g(x)$ is a cubic polynomial for $(x_k, x_{k+1}]$, $k = 1, \cdots, K$
  - $g(x)$ extrapolate linearly

- Unlogged variable Y will have a polynomial tail i.e. $\xi > 0$

  linear tail in $g \rightarrow$ exponential tail in $X \rightarrow$ polynomial tail in Y

Fit density to histogram bin counts using Possion regression



Fit a (penalized) Poisson regression to the finely binned histogram counts

Density estimation: penalized Poisson regression

Penalized negative log likelihood:

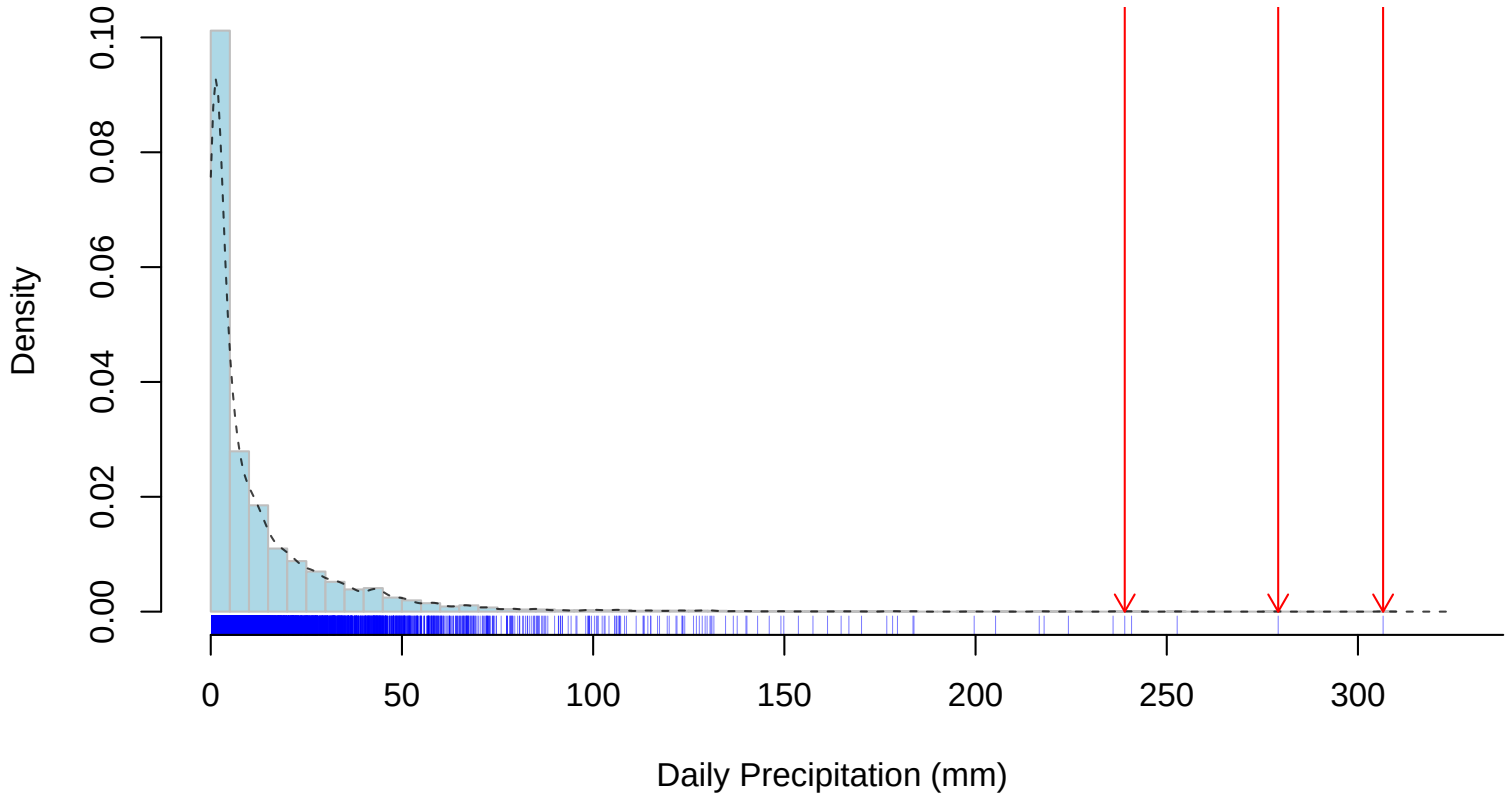
$$-\sum_{j=1}^{N}\{g_j z_j - e^{g_j} - \log(z_j!)\} + \lambda\left(\int_{x\in\mathbb{R}} \left(g''(x)\right)^2\, dx\right)$$

where λ is the smoothing parameter

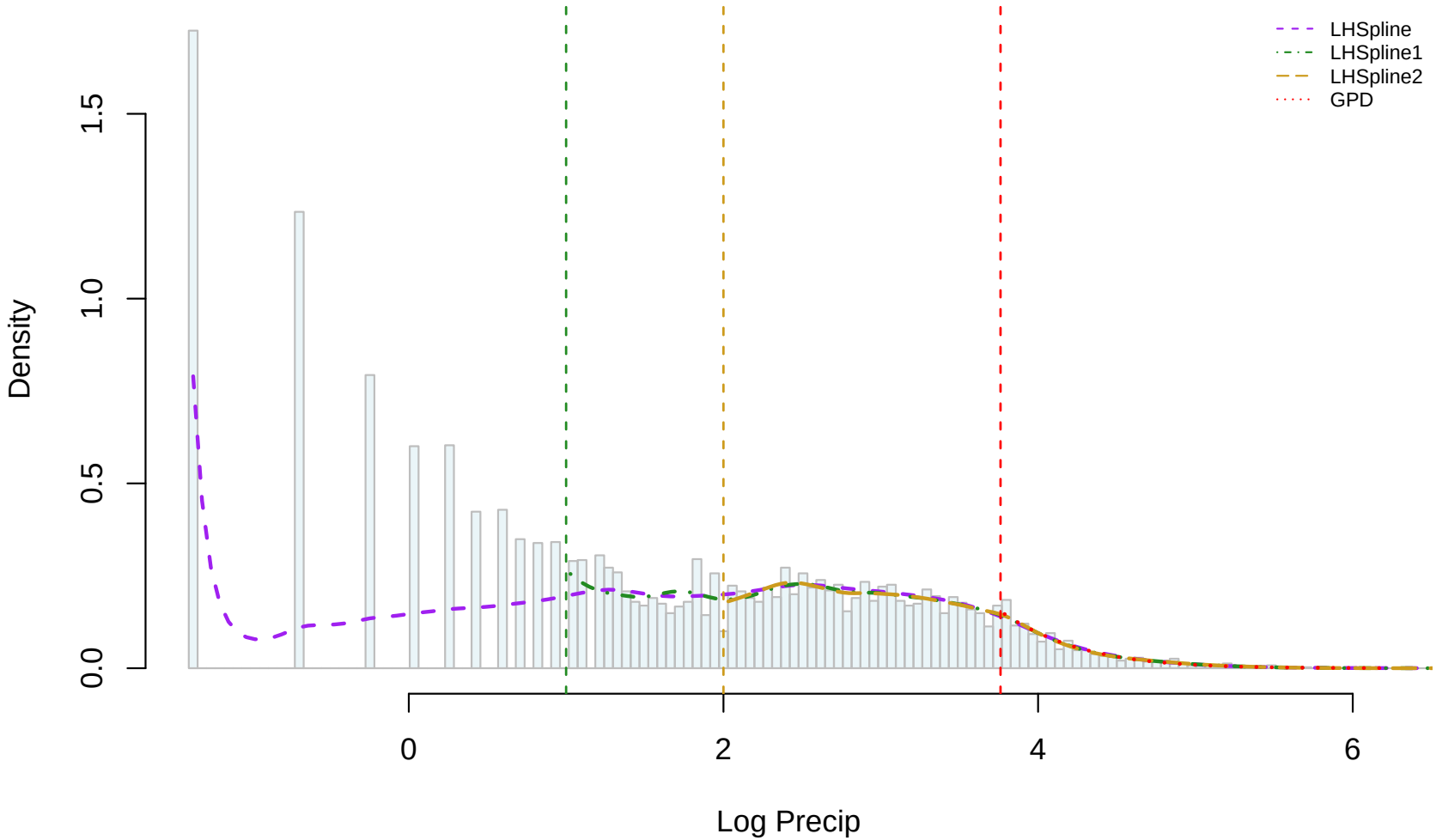
- The minimizer $\hat{g}_\lambda(x)$ is a natural cubic spline
- λ is chosen by cross validation (CV)
- Bayesian interpretation: $\hat{g}$ is the posterior mode with the prior proportional to $\alpha \int g''^2$

Part IV: Houston Precipitation Extremes

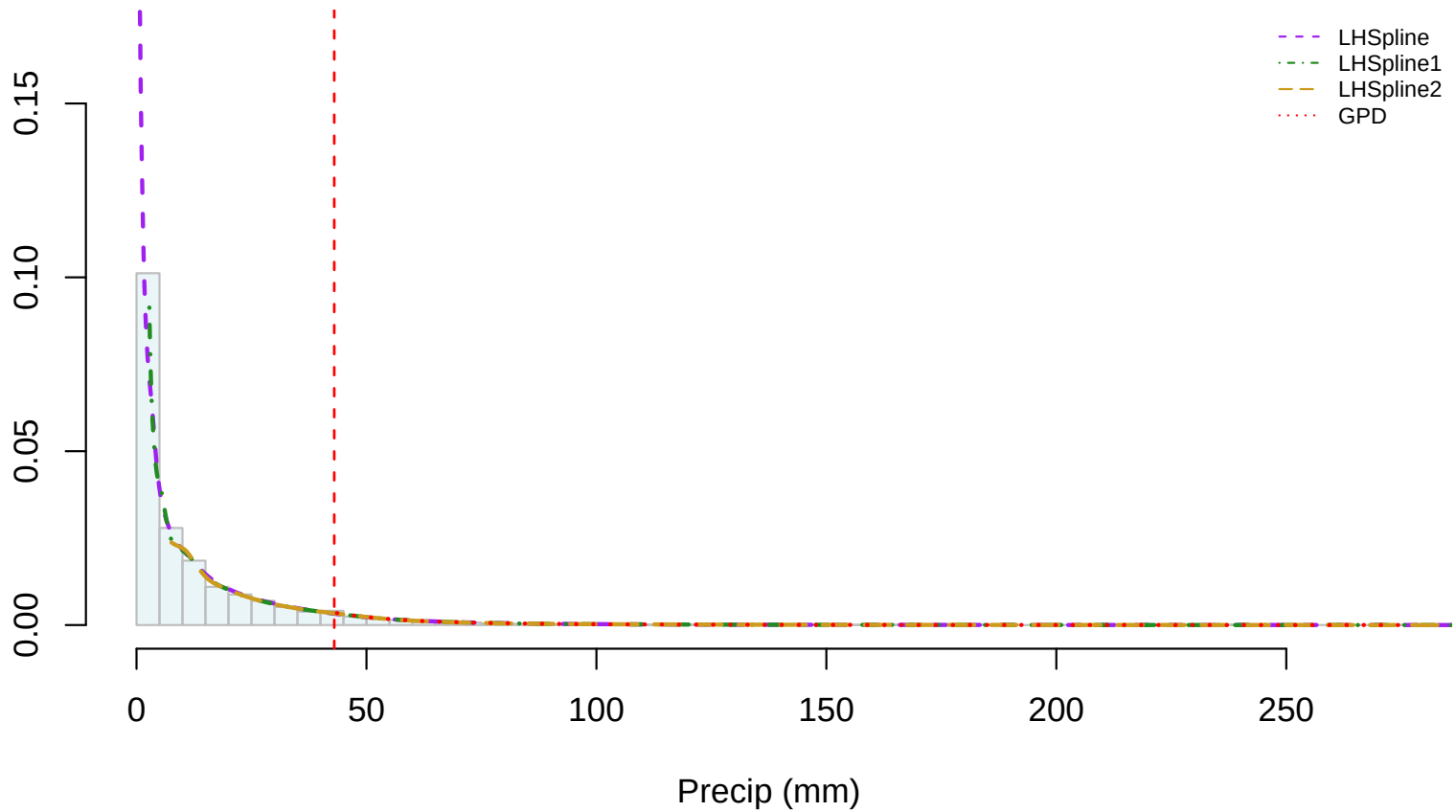
Houston Hobby Airport daily rainfall data



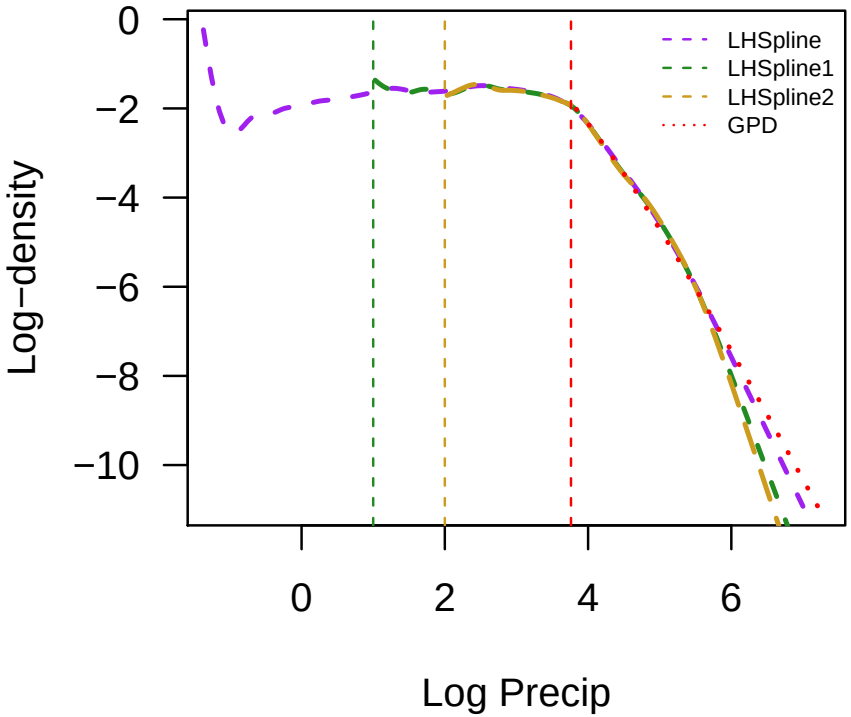
LHSpline fit on the log scale

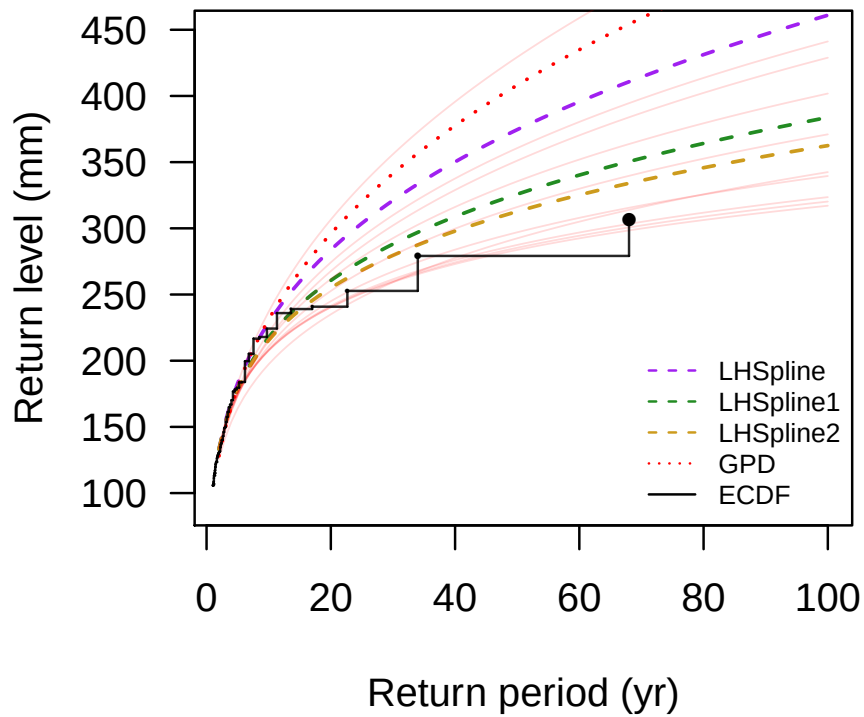


LHSpline fit: Original scale

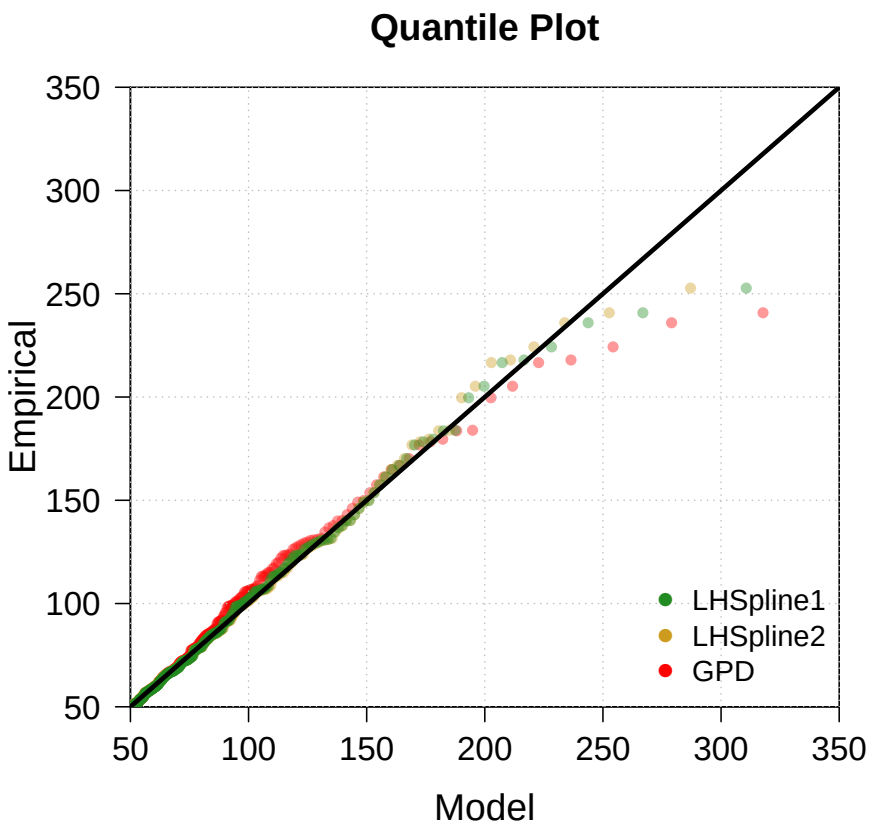


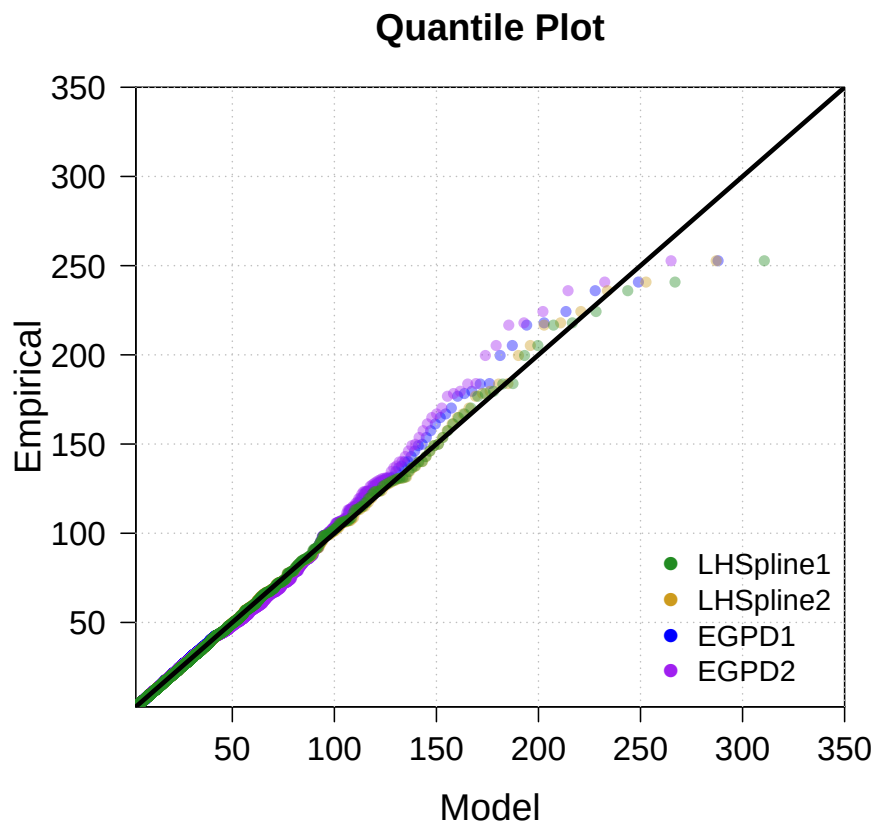
LHSpline fit: Return Levels





QQ plots


Quantile Plot
Empirical
Model
50
100
150
200
250
300
350
LHSpline1
LHSpline2
GPD


Quantile Plot
Empirical
Model
50
100
150
200
250
300
350
LHSpline1
LHSpline2
EGPD1
EGPD2

How unusual was the event of 306.58 mm on 26th August, 2017?

Method	POT	LHSpline	LHSpline1	LHSpline2
Estimate (years)	30.5	34.8	64.0	98.0
90% CI Lower limit	17.0	15.2	19.1	22.0
90% CI Upper limit	73.3	73.2	172.4	345.7

Summary and discussion

- A brief introduction to extreme value analysis with an climate application (estimating rainfall extremes)

- Log-Histospline:
  - + : **Does not require the threshold selection** when modeling the tail (and the bulk) distribution
  - + : Applicable to **heavy-tailed distribution**: streamflow, financial return, ...
  - -: Only applicable to heavy-tailed distributions
  - -: Inference is not straightforward

- Things not Addressed:
  - Modeling temporal, spatial (and spatio-temporal) dependence of extremes
  - Nonstationarity of extremes

For Further Reading on Extreme Value Analysis

- **Books**:
  1. Coles (2001): *An introduction to statistical modeling of extreme values*
  2. de Haan and Ferreira (2006): *Extreme Value Theory: an Introduction*
  3. Yan and Dey (2015): *Extreme Value Modeling and Risk Analysis*
  4. Embrechts, Klüppelberg, and Mikosch (1997): *Modelling extremal events for insurance and finance*
- **Review Papers**:
  1. Cooley et al. (2012): *A survey of spatial extremes: Measuring spatial dependence and modeling spatial effects*, *Revstat*,10 (1), 135-165
  2. Davison, Padoan, and Ribatet (2012): "Statistical modeling of spatial extremes", *Statistical science*, 27 (2), 161-186
  3. Davison and Huser (2015): "Statistics of extremes", *Annual Review of Statistics and its Application* 2, 203-235
- **R packages**: `ismev`, `extRemes`, `SpatialExtremes`, $\cdots$