Extreme Value Analysis for Climate Research

Whitney Huang

CLEMSON UNIVERSITY

whuang@samsi.info

Clemson Math Club, January 17, 2020

Who am I?

- **First year** Assistant Professor of Applied Statistics
- Born in Laramie, Wyoming, grew up in Taiwan

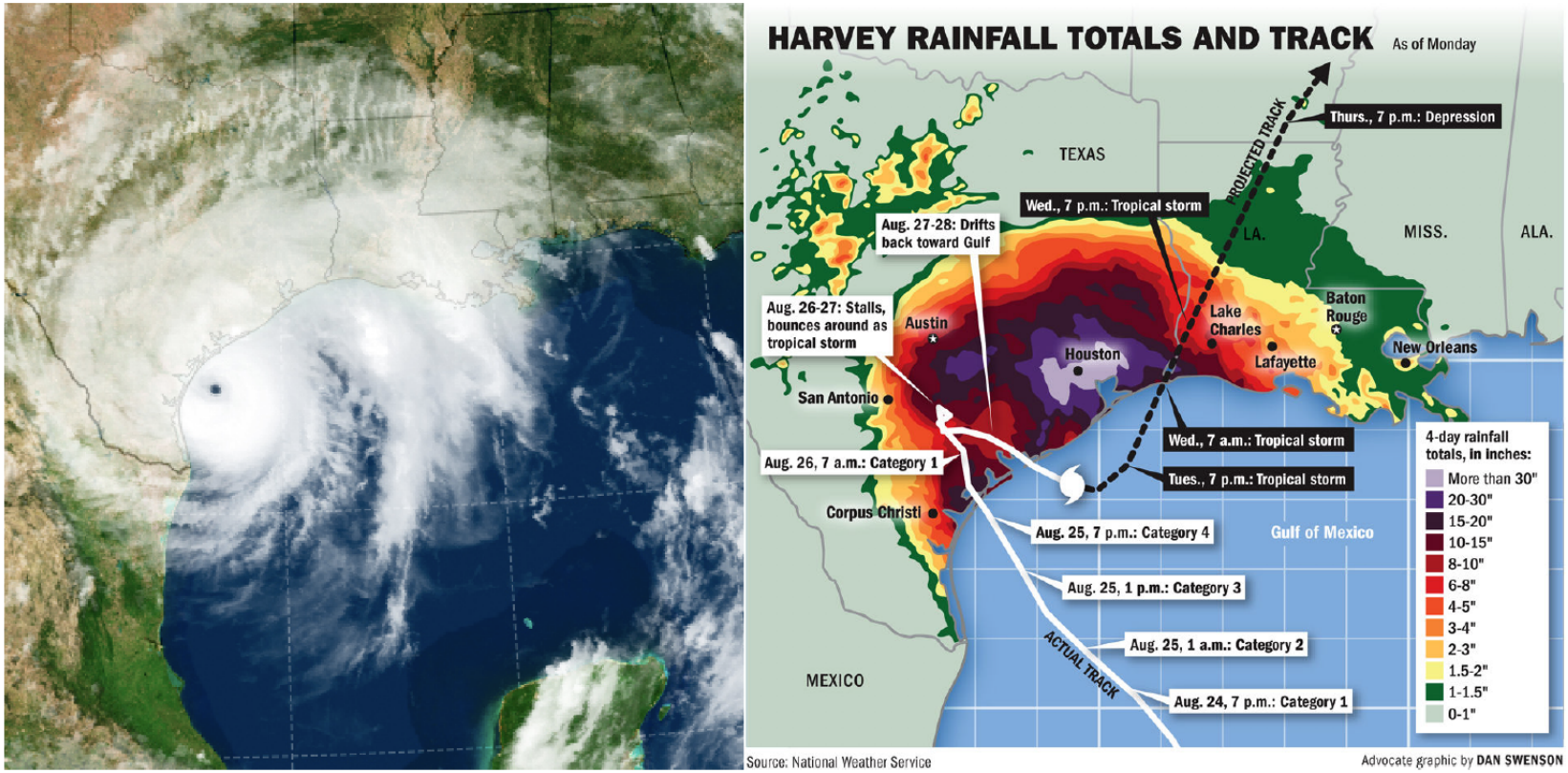
- Got a B.S. in Mechanical Engineering, switched to Statistics in graduate school
- Ph.D. in Statistics, 2017, Purdue; SAMSI/CANSSI Postdoc 2017-2019

PURDUE

samsi
NSF Duke NCSU UNC

University of Victoria

Extreme Rainfall During Hurricane Harvey



Source: NASA (Left); National Weather Service (Right)

- *"A storm forces Houston, the limitless city, to consider its limits"* – The New York Times (8.31.17)

Scientific Questions

- How to estimate the magnitude of extreme events (e.g. 100-year rainfall)?
- How extremes vary in space?
- How extremes may change in future climate conditions?

Digression: Central Limit Theorem

What is Central Limit Theorem (CLT)?

The **sampling distribution** of the **mean** will become approximately **normally distributed** as the **sample size becomes larger**, **irrespective of the shape of the population distribution!**

Let $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} F$ with $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \text{Var}[X_i] < \infty$. Then $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{d} \text{N}(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$.

▸ Proof

What is Central Limit Theorem (CLT)?

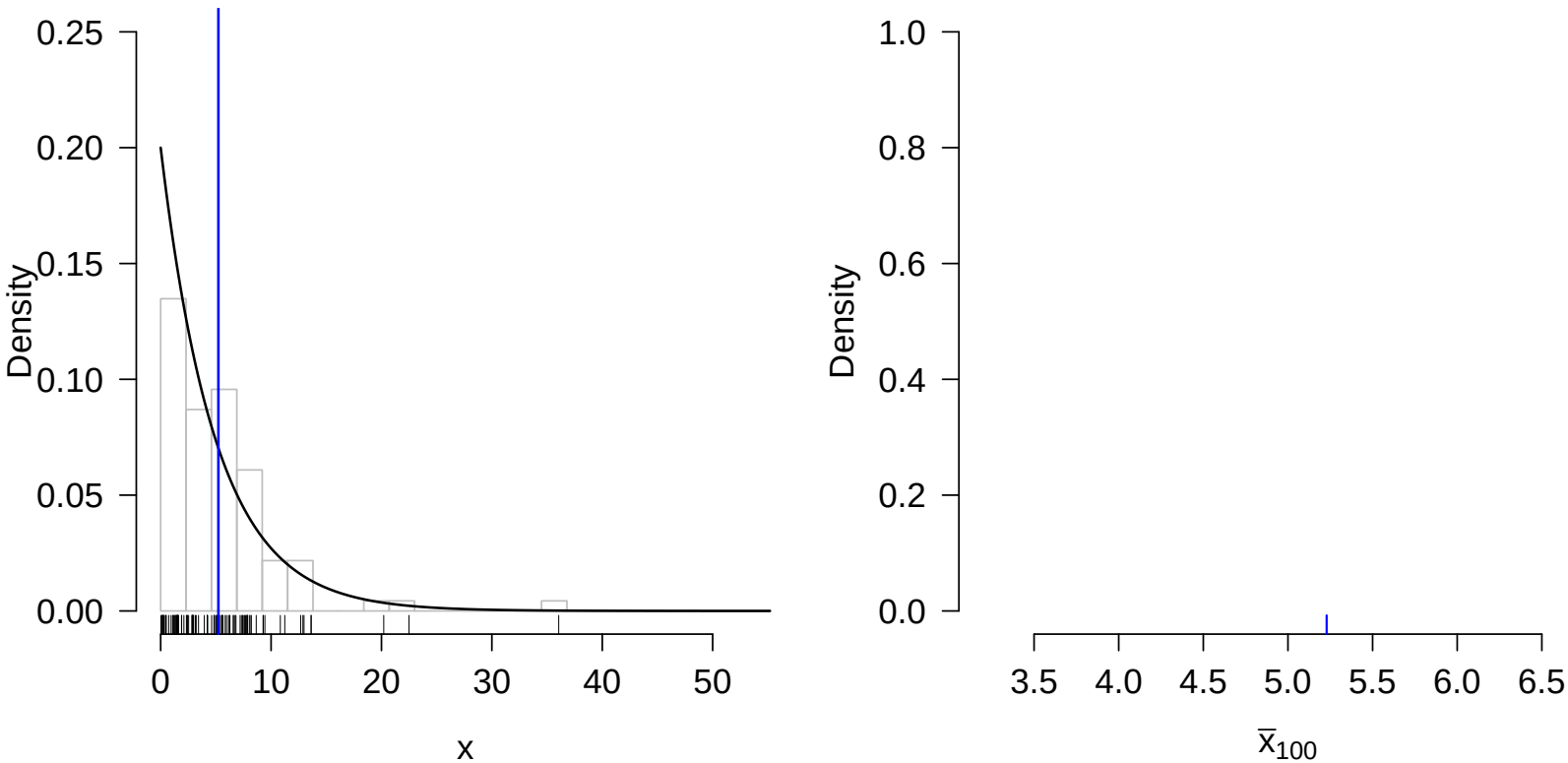
The **sampling distribution** of the **mean** will become approximately **normally distributed** as the **sample size becomes larger**, **irrespective of the shape of the population distribution!**

Let $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} F$ with $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \text{Var}[X_i] < \infty$. Then $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \xrightarrow{d} \text{N}(\mu, \frac{\sigma^2}{n})$ as $n \to \infty$.

▸ Proof

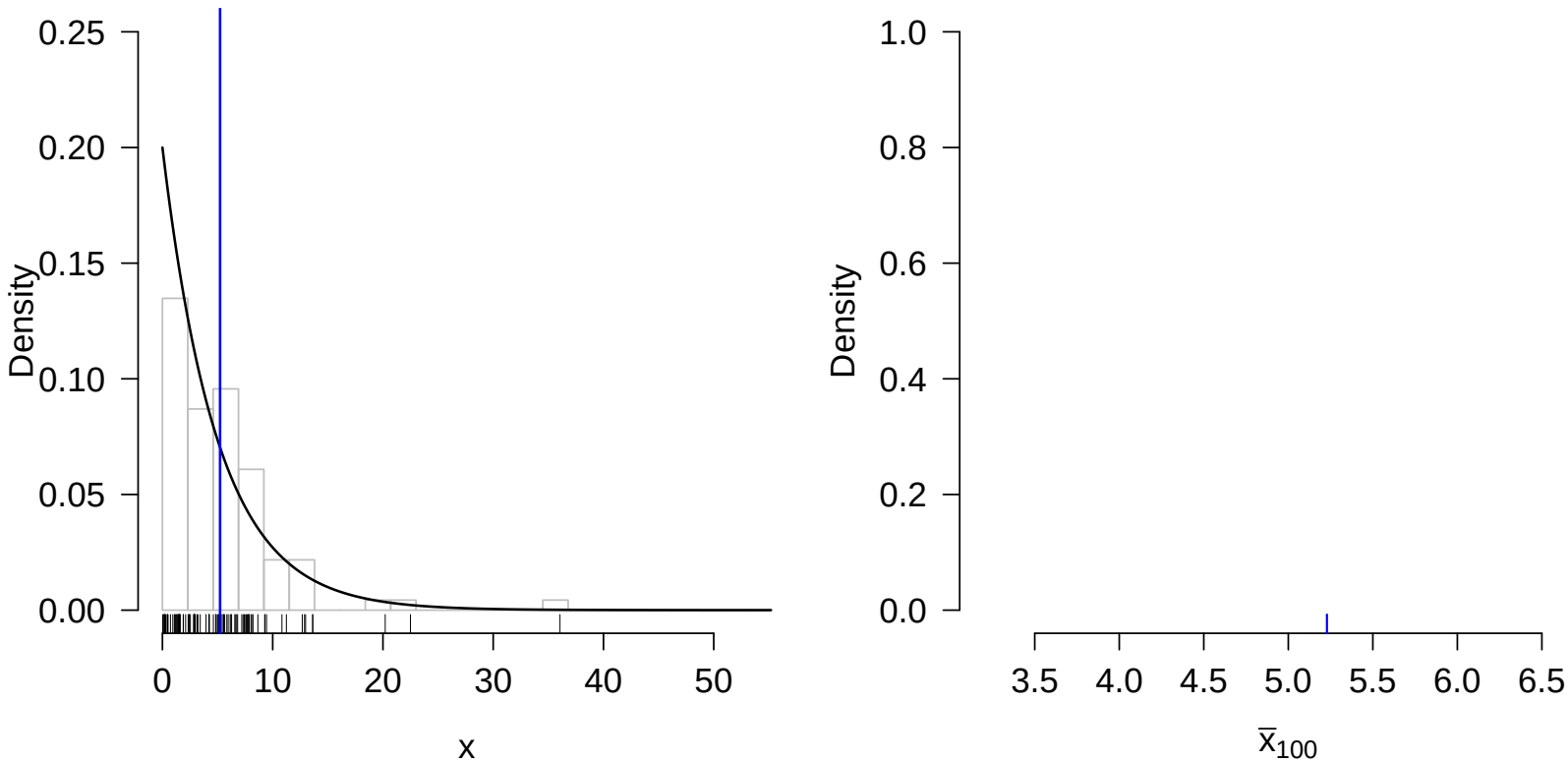
CLT in Action

1. Generate 100 random numbers $(n = 100)$ from an Exponential distribution
2. Compute the sample mean of these 100 random numbers
3. Repeat this process 120 times



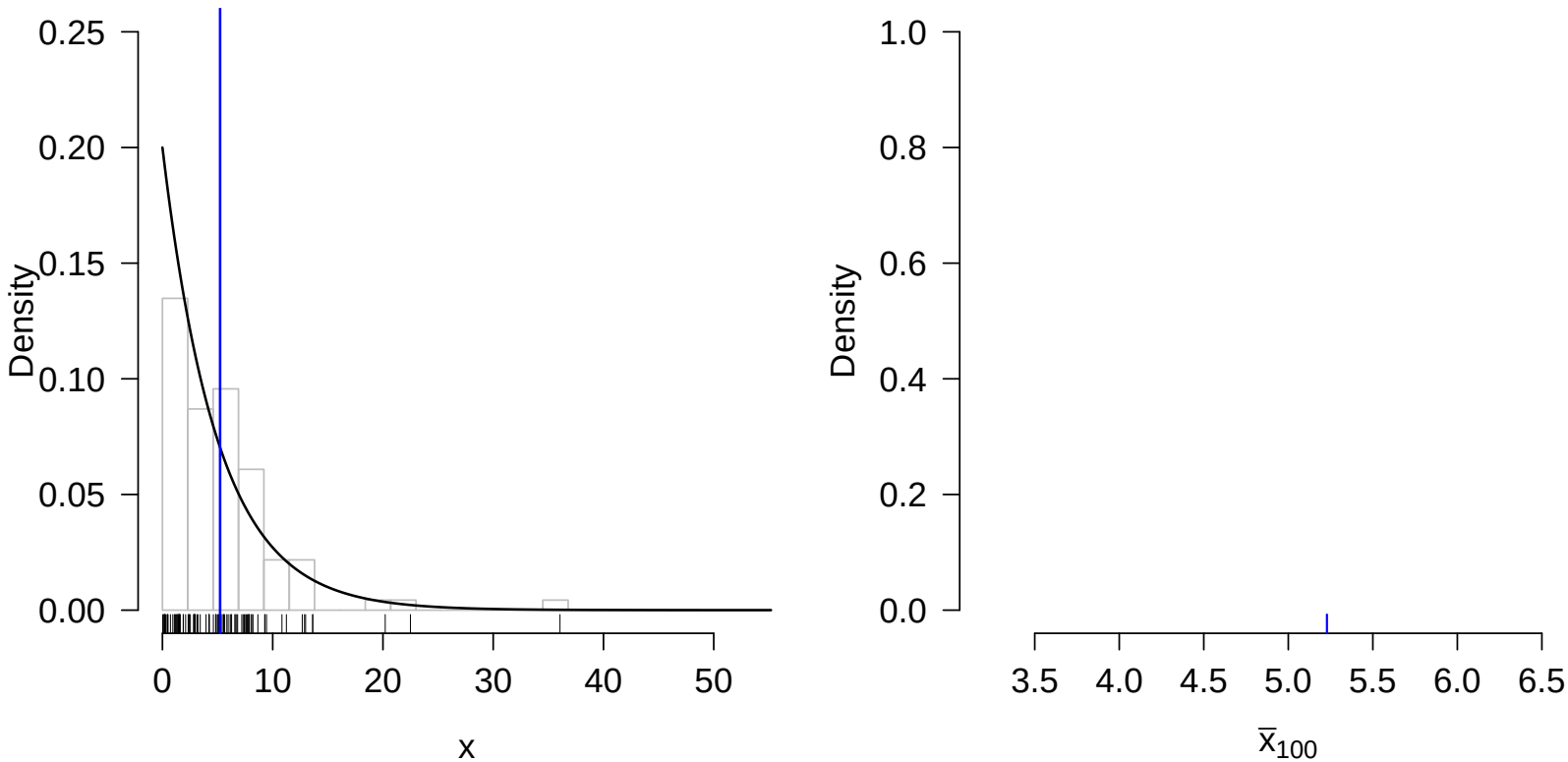
CLT in Action

1. Generate 100 random numbers $(n = 100)$ from an Exponential distribution
2. Compute the sample mean of these 100 random numbers
3. Repeat this process 120 times



CLT in Action

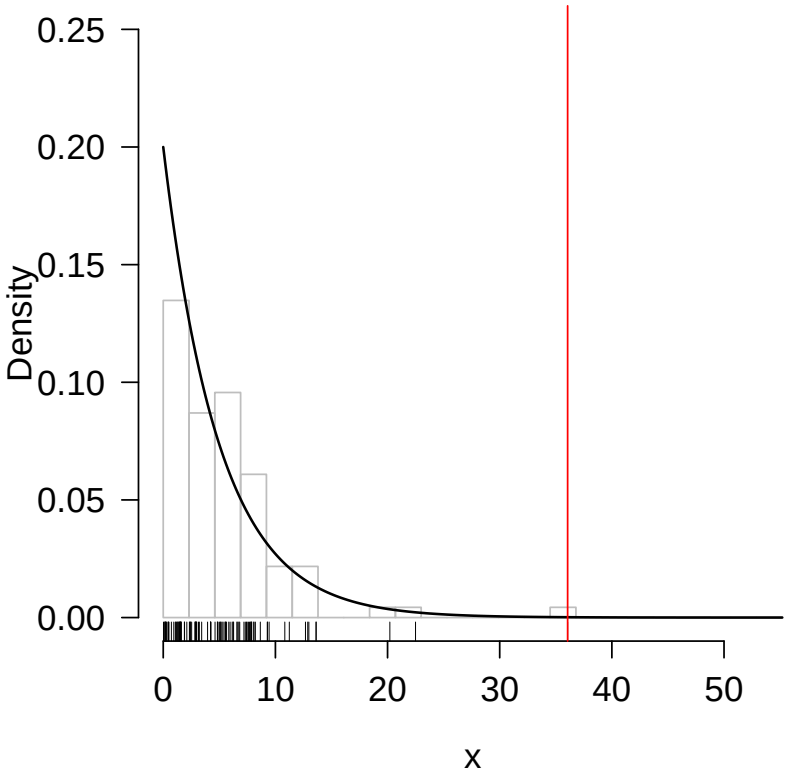
1. Generate 100 random numbers $(n = 100)$ from an Exponential distribution
2. Compute the sample mean of these 100 random numbers
3. Repeat this process 120 times

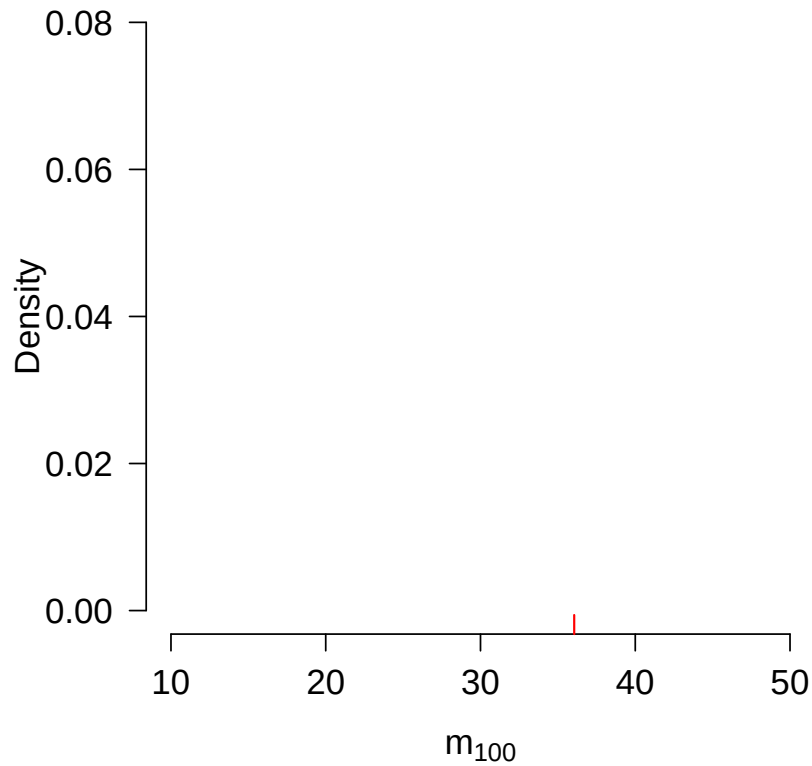


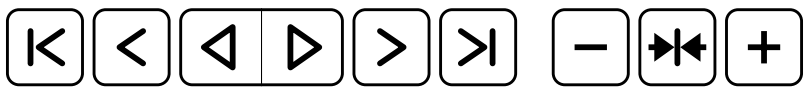
Can we find an analog of CLT for the unusually large values?

Distribution of the Sample Maximum

1. Generate 100 random numbers $(n = 100)$ from an Exponential distribution
2. Compute the sample maximum of these 100 random numbers
3. Repeat this process 120 times







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 $n \to \infty$, if

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

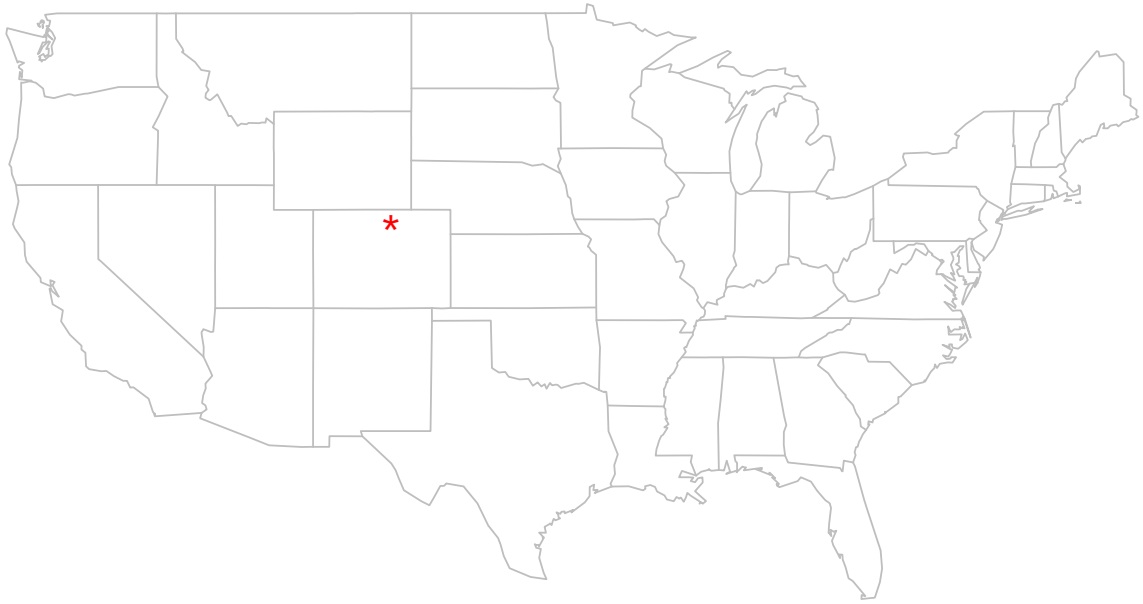
then G must be the same type of the following form:

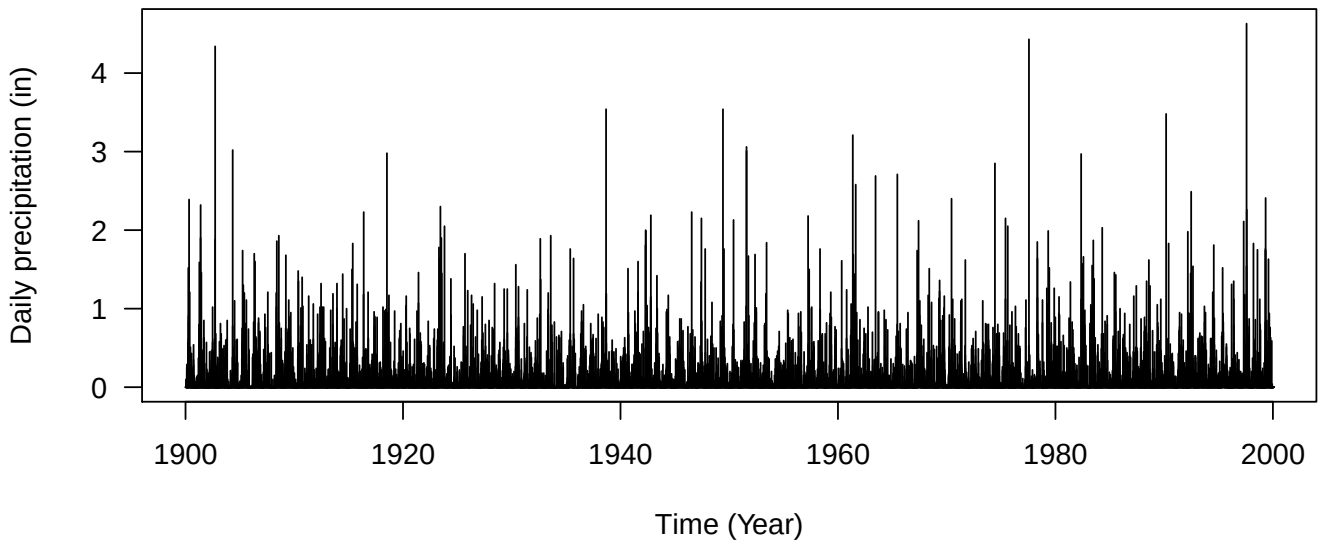
$$\mathsf{G}(x; \mu, \sigma, \xi) = \exp\left\{ -\left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_+^{\frac{-1}{\xi}} \right\}$$

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

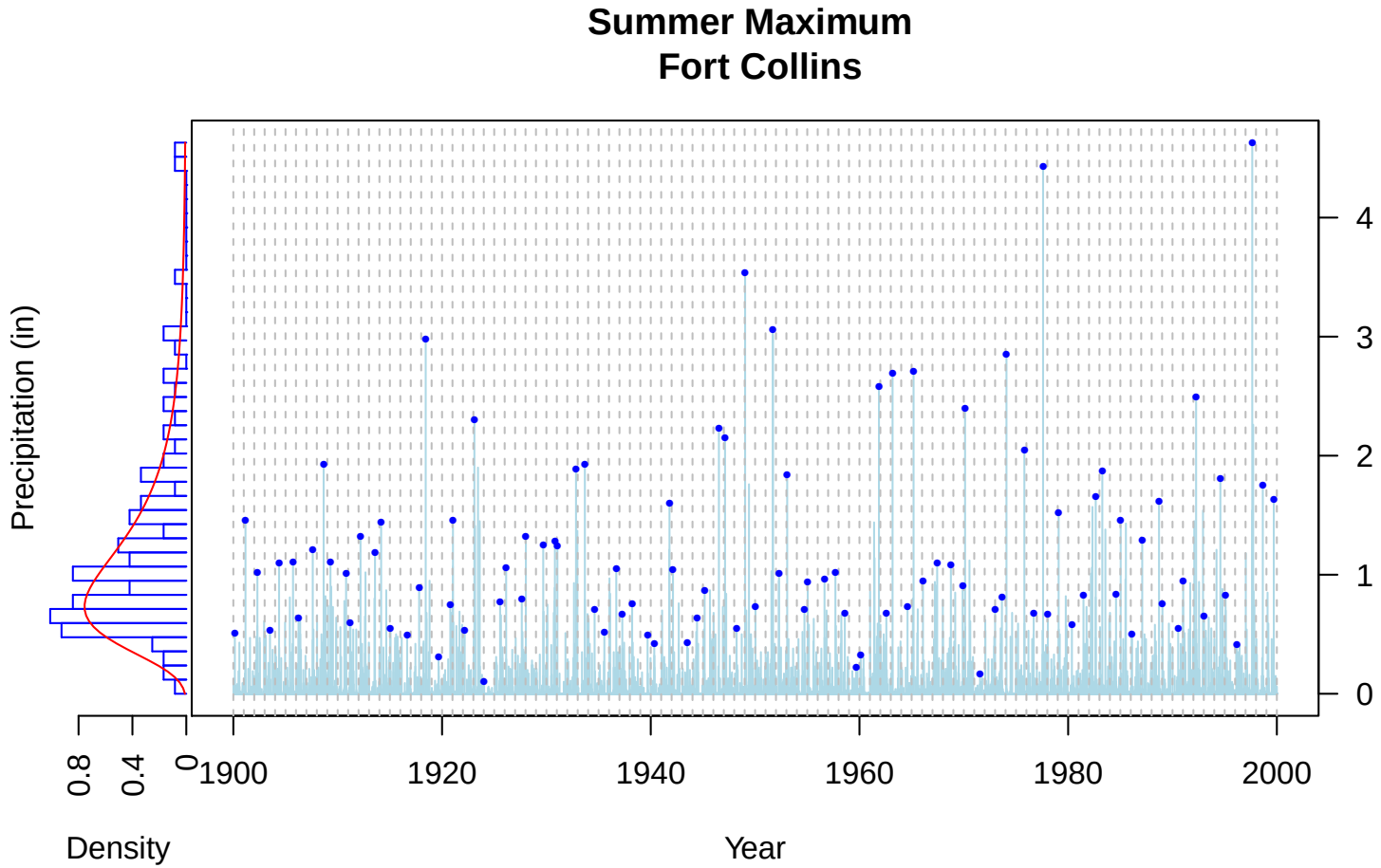
- μ and σ are location and scale parameters
- ξ is a shape parameter determining the rate of tail decay, with
  - $\xi > 0$ giving the heavy-tailed case (Fréchet)
  - $\xi = 0$ giving the light-tailed case (Gumbel)
  - $\xi < 0$ giving the bounded-tailed case (reversed Weibull)

Fort Collins Daily Precipitation





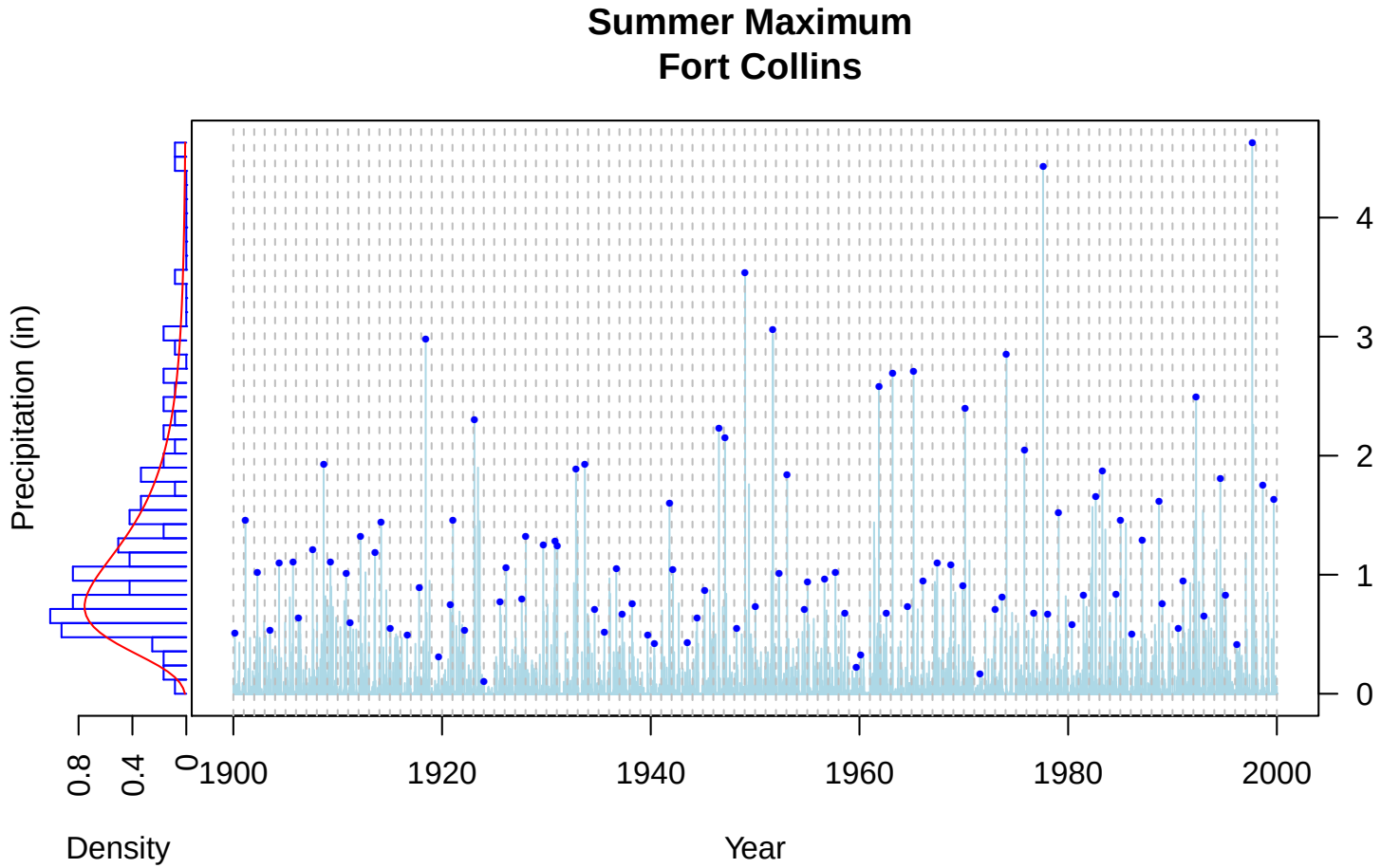
Estimating Extremes Using Block Maxima (Gumbel 1958)



Which distribution to use for annual maxima?

$\Rightarrow$ generalized extreme value distribution (GEV(μ, σ, ξ))

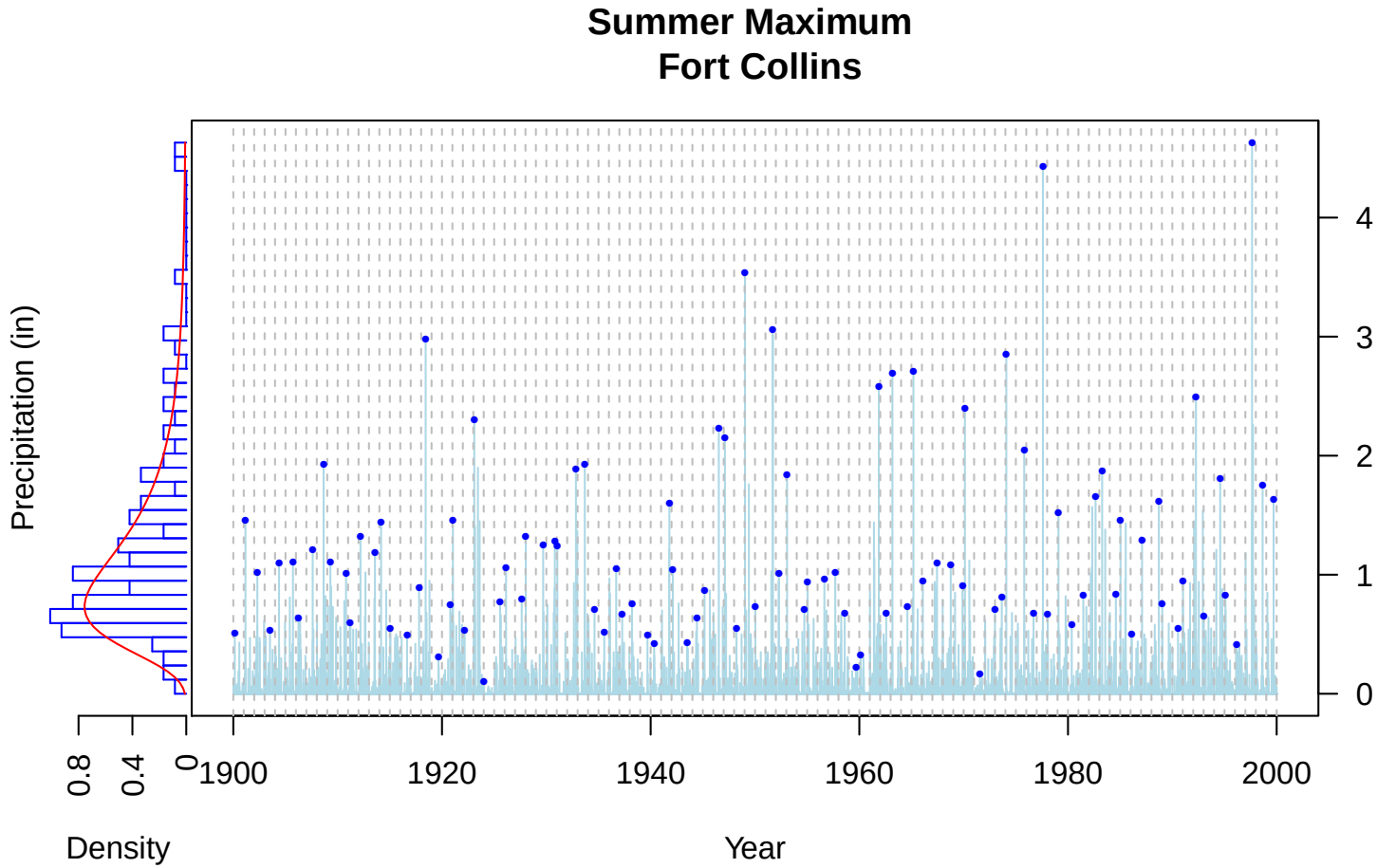
Estimating Extremes Using Block Maxima (Gumbel 1958)



Which distribution to use for annual maxima?

$\Rightarrow$ generalized extreme value distribution (GEV(μ, σ, ξ))

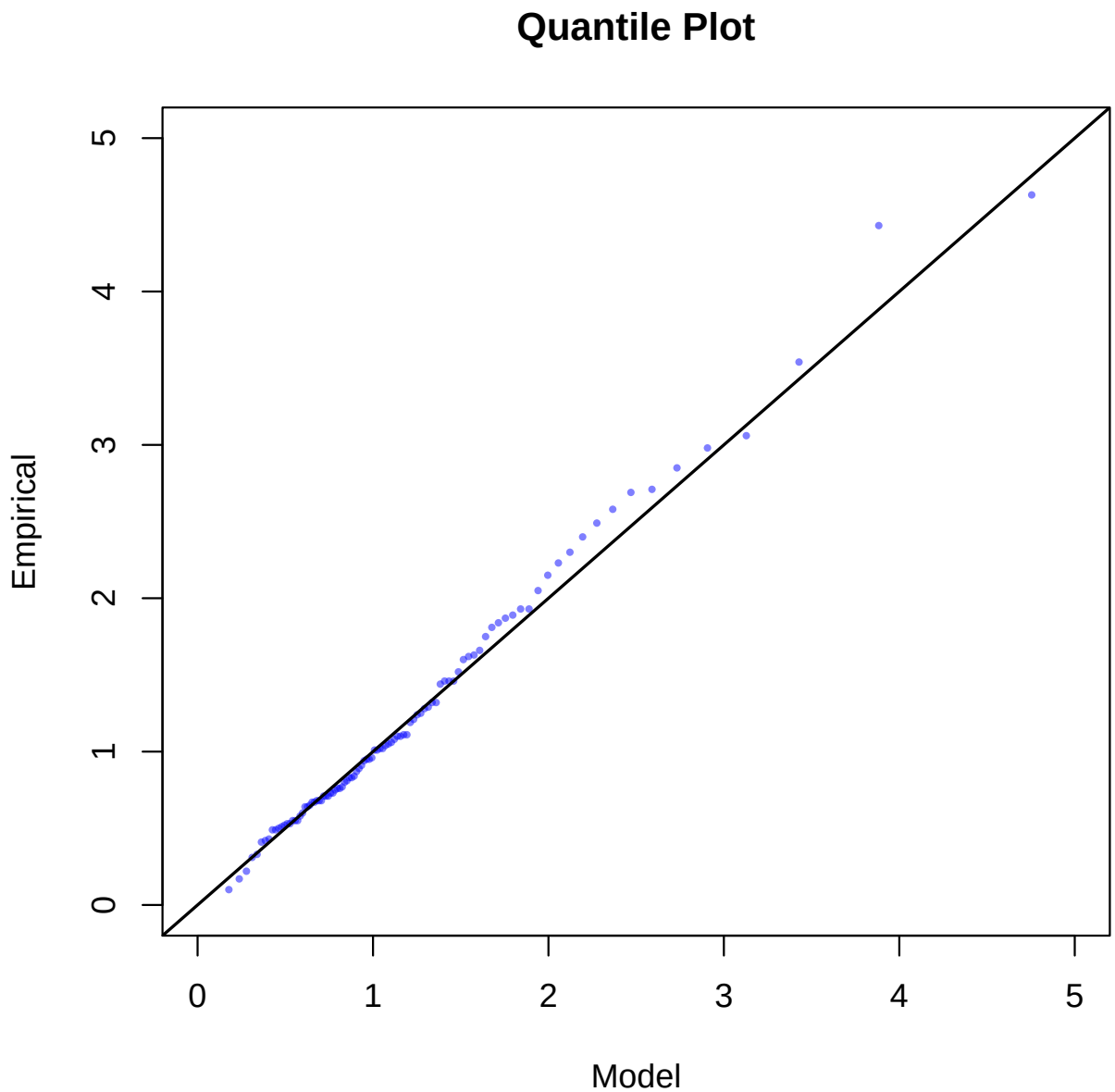
Estimating Extremes Using Block Maxima (Gumbel 1958)



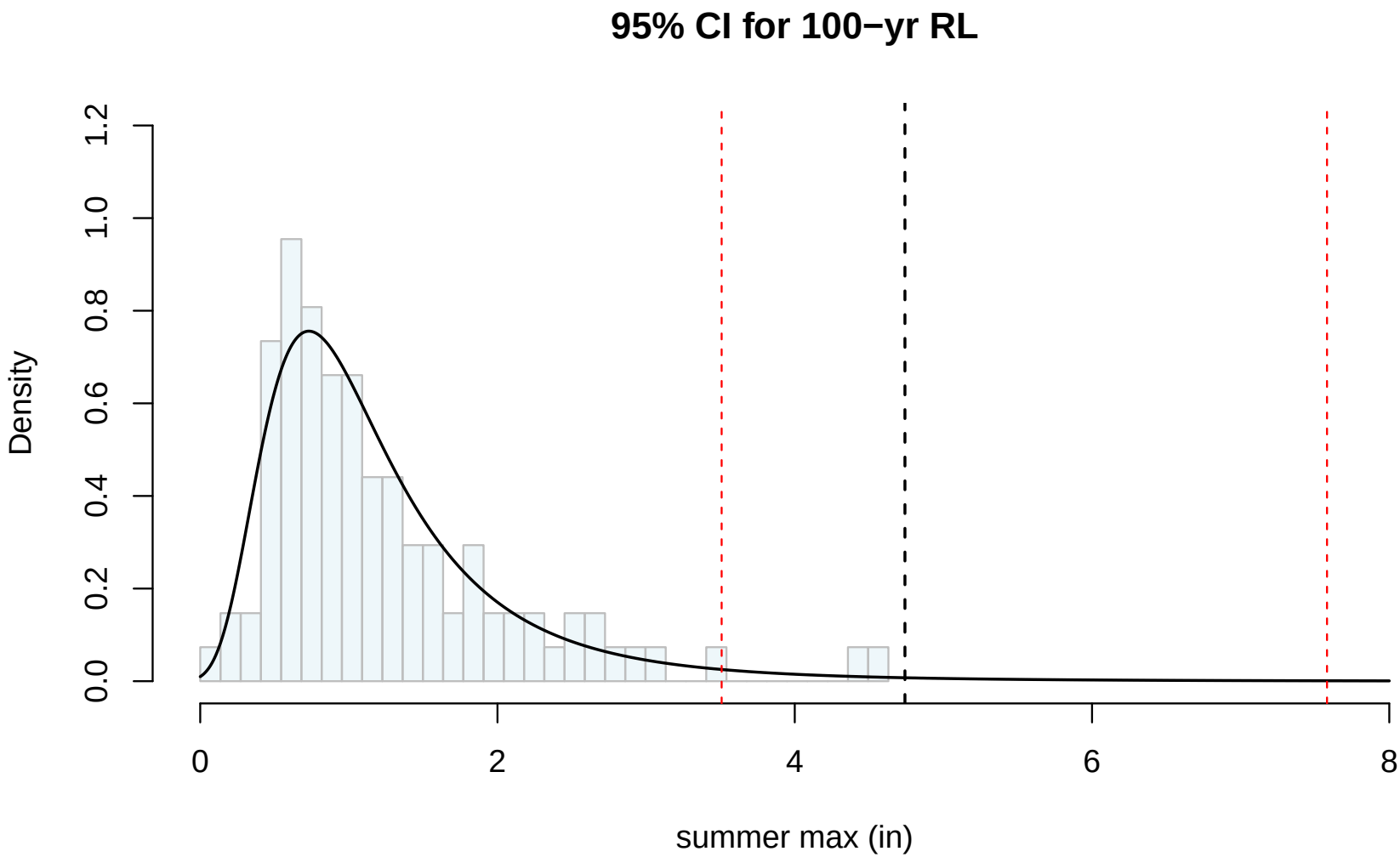
Which distribution to use for annual maxima?

$\Rightarrow$ generalized extreme value distribution (GEV(μ, σ, ξ))

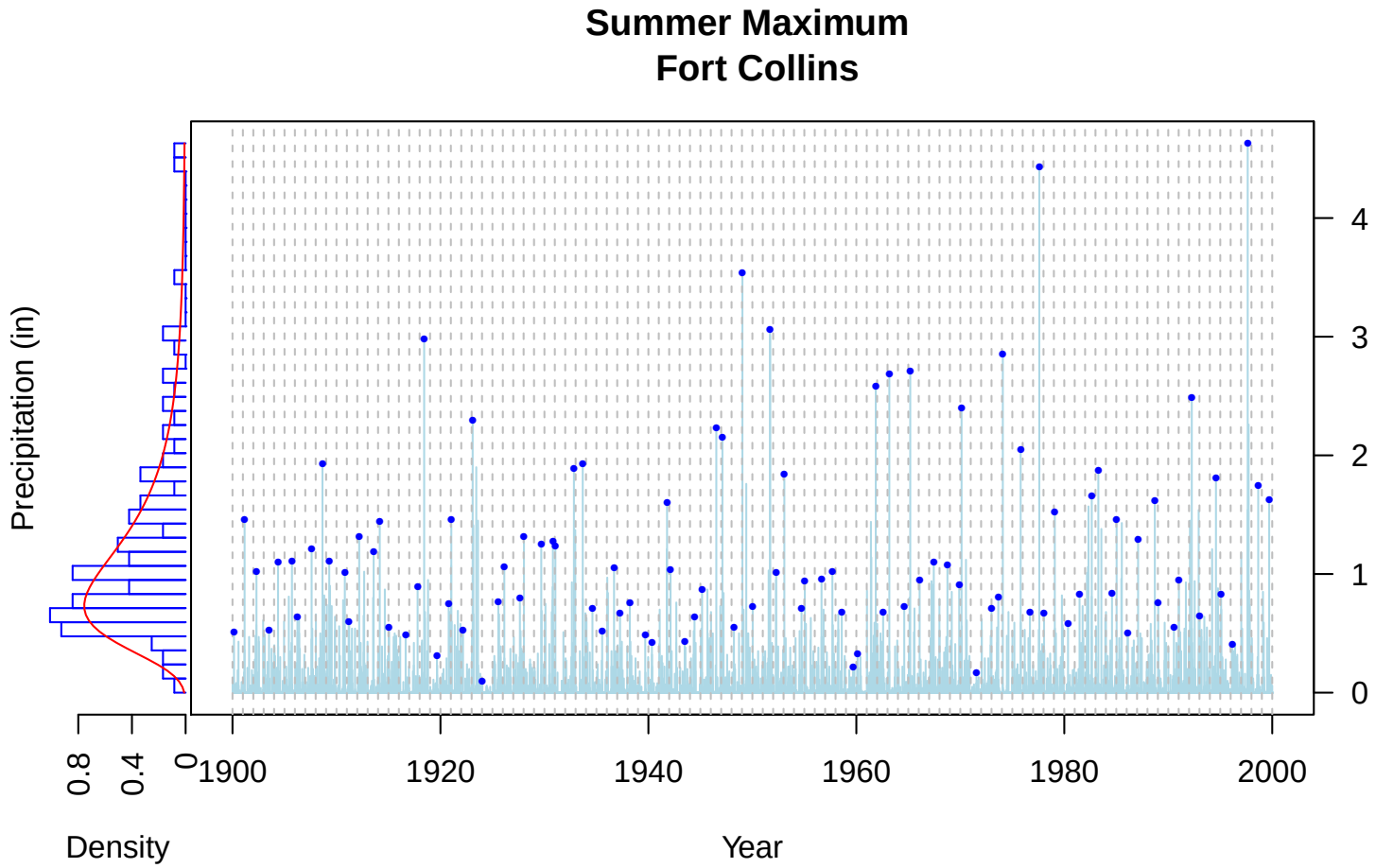
GEV Fit Diagnostics



Inference for 100-Year Event



Recall the Block Maxima Method

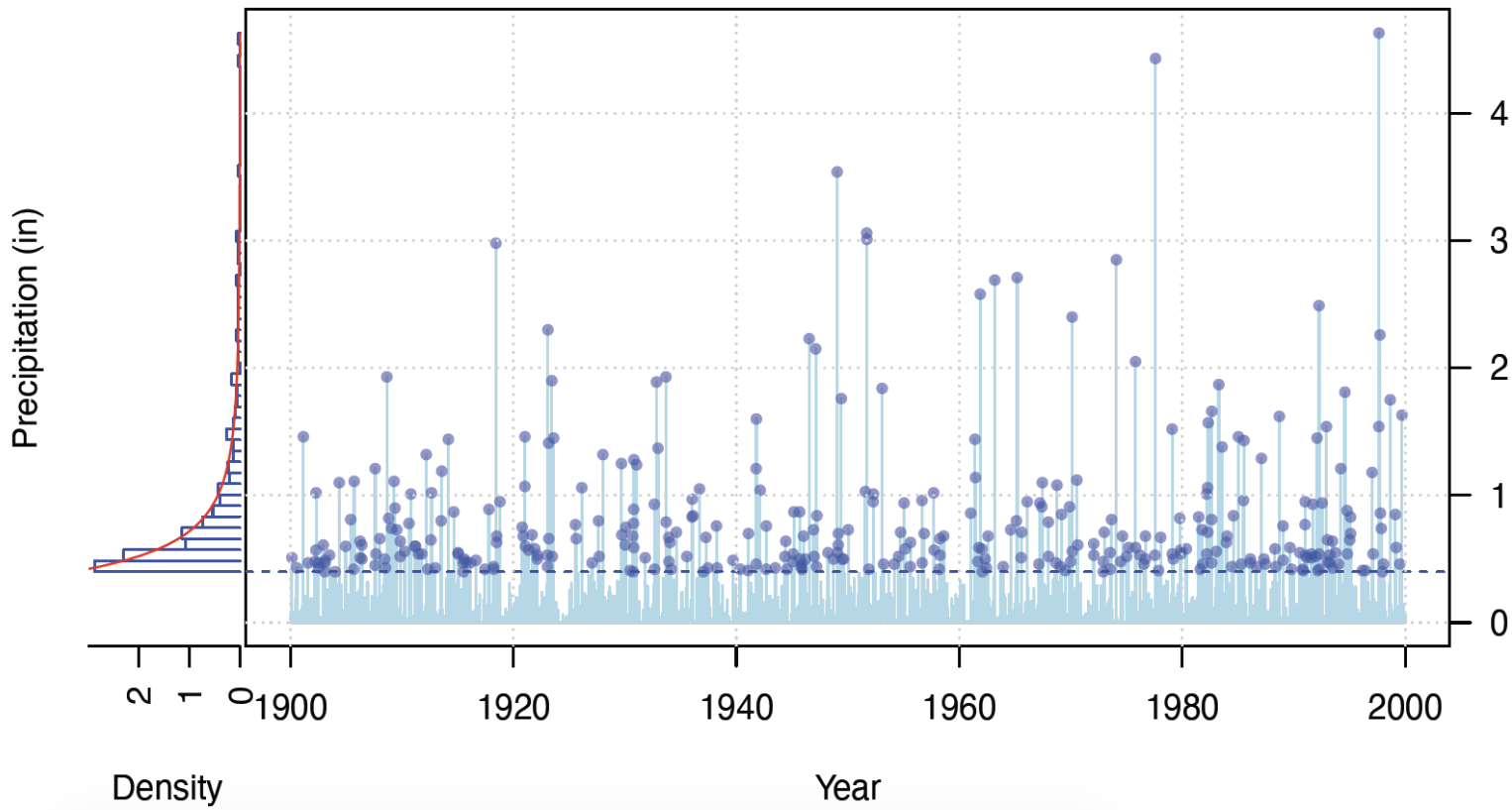


Question: Can we use data more efficiently?

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

Theorem: GEV for block maxima $\Rightarrow$ generalized Pareto distribution (GPD) for excesses over high threshold

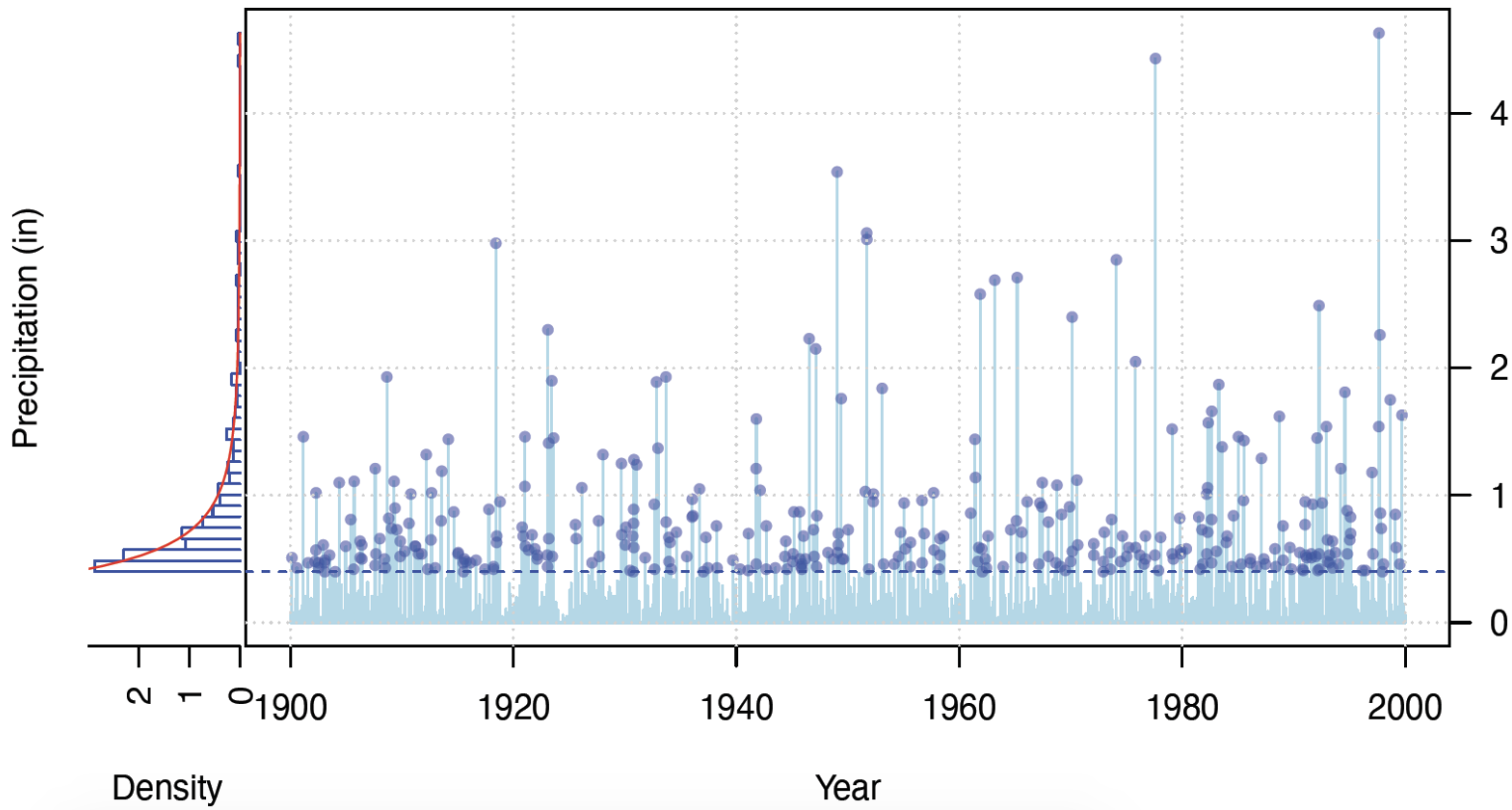
▸ Pickands–Balkema–de Haan Theorem



▸ How to choose the threshold?

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

Theorem: GEV for block maxima $\Rightarrow$ generalized Pareto distribution (GPD) for excesses over high threshold ▸ Pickands–Balkema–de Haan Theorem

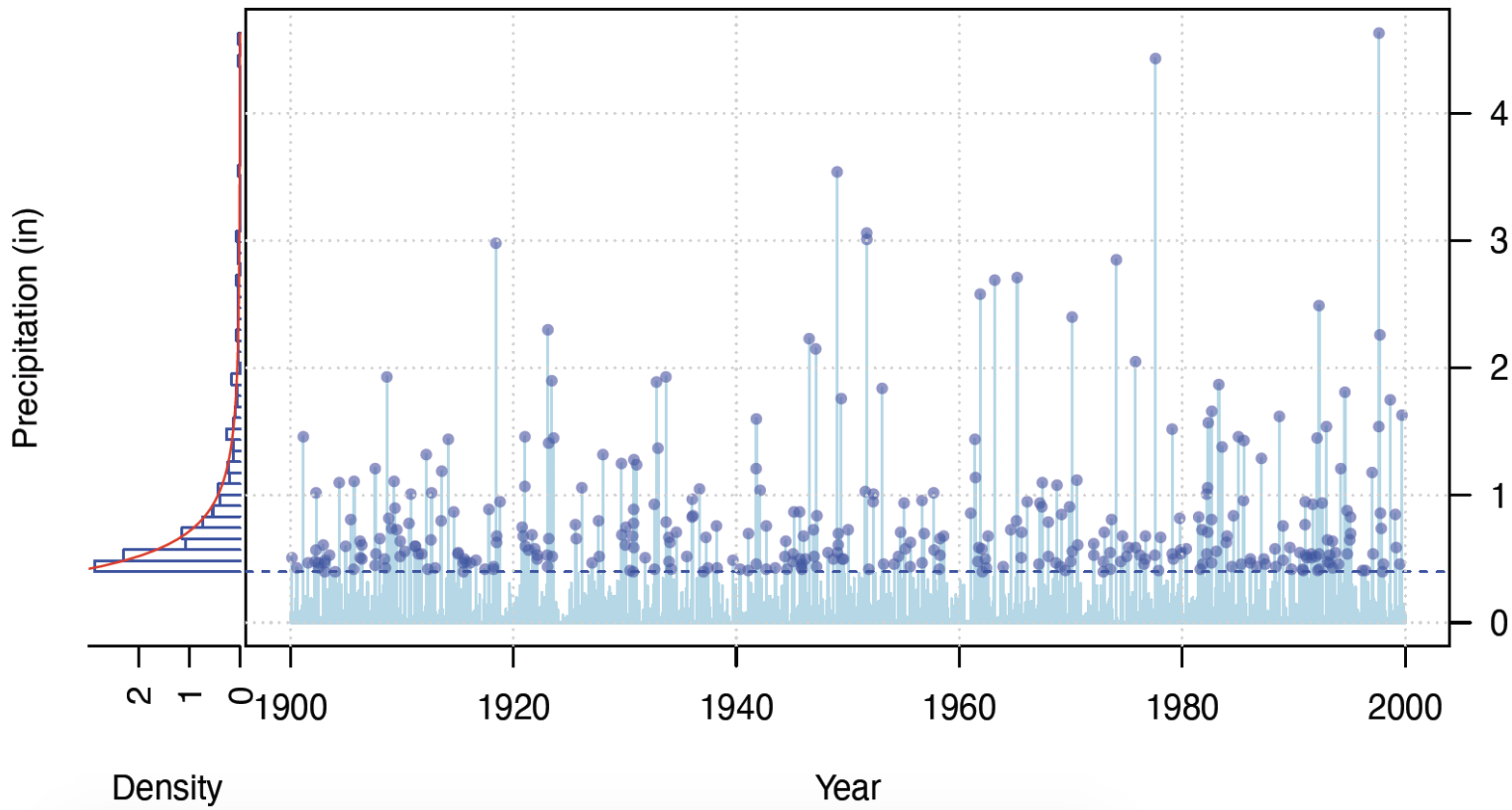


▸ How to choose the threshold?

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

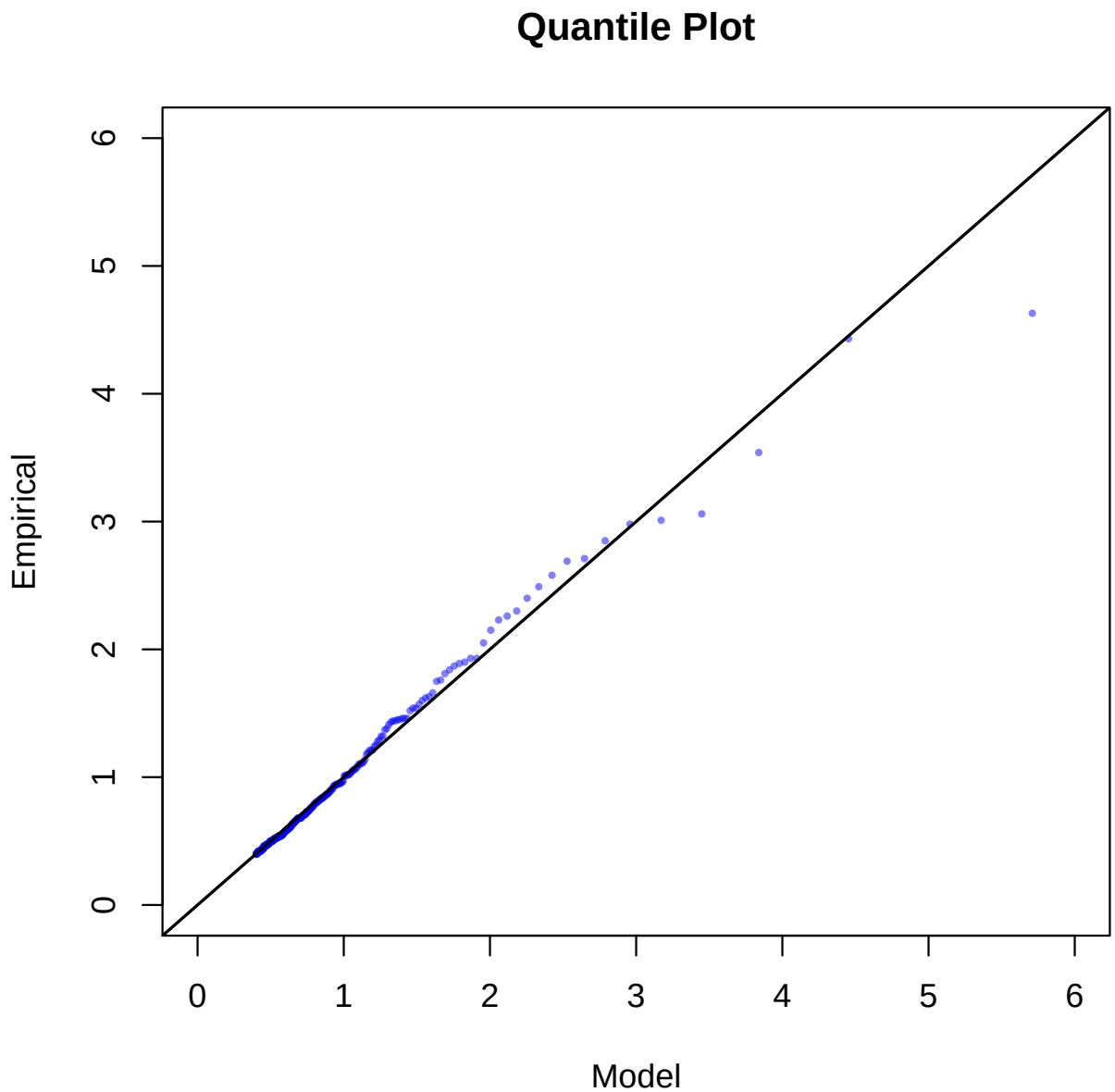
Theorem: GEV for block maxima $\Rightarrow$ generalized Pareto distribution (GPD) for excesses over high threshold

▸ Pickands–Balkema–de Haan Theorem

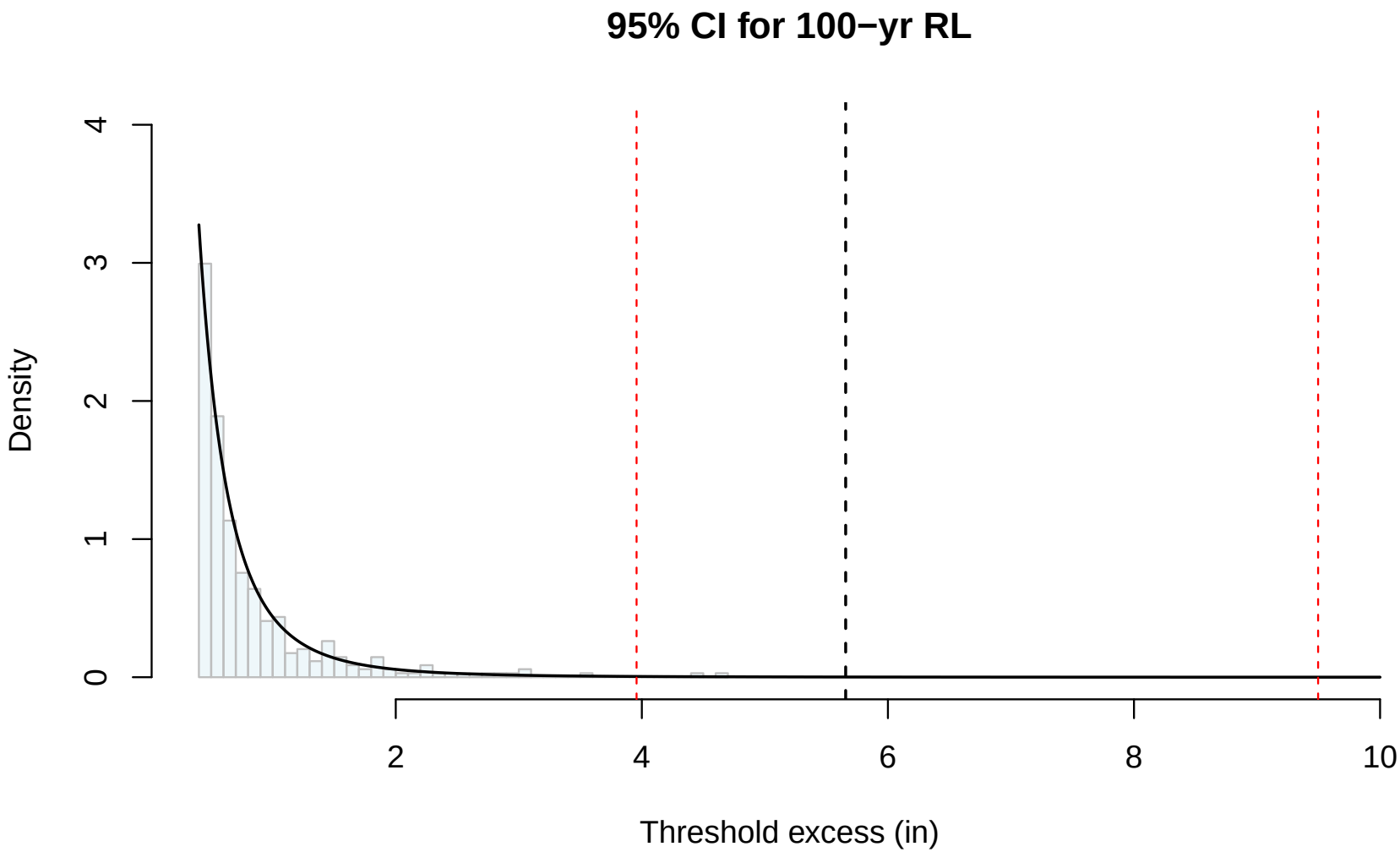


▸ How to choose the threshold?

GPD Fit Diagnostics



Inference for 100-Year Event



Summary & Discussion

- Climate extremes can have large impacts on both human society and environmental systems
- Extreme value theory provides a framework to model extreme values
  - GEV for fitting block maxima
  - GPD for fitting threshold exceedances
  - Return level for communicating risk
- Practical Issues: seasonality, temporal dependence, non-stationarity, ...

Summary & Discussion

- Climate extremes can have large impacts on both human society and environmental systems
- Extreme value theory provides a framework to model extreme values
  - GEV for fitting block maxima
  - GPD for fitting threshold exceedances
  - Return level for communicating risk
- Practical Issues: seasonality, temporal dependence, non-stationarity, ...

How to React Me?

- **Email:** wkhuang@clemson.edu
- **Office**: O-221 Martin Hall

Go Tigers!

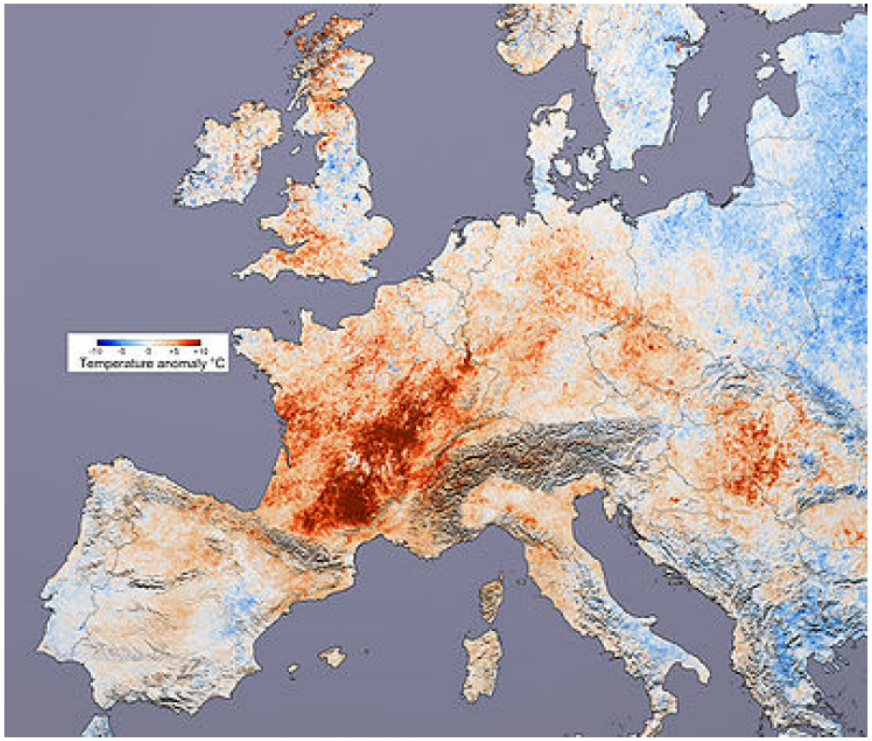
How to React Me?

- **Email:** wkhuang@clemson.edu
- **Office**: O-221 Martin Hall

Go Tigers!

Backup Slides

Environmental Extremes: Heatwaves, storm surges, etc.



- **Heat wave**: The 2003 European heat wave led to the hottest summer on record in Europe since 1540 that resulted in at least 30,000 deaths

- **Storm Surge**: Hurricane Katrina produced the highest storm surge ever recorded (27.8 feet) on the U.S. coast

Sketch of a Proof of CLT

Let's consider the case where $\mu = 0$ and $\sigma^2 = 1$. When will use the **characteristics function** $\varphi_X(t) = \mathbb{E}[e^{itX}]$ for this proof. Let $Z_n = \frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}$. What we need to show is that $\varphi_{Z_n}(t) \overset{n\to\infty}{\longrightarrow} e^{-\frac{t^2}{2}}$.

$$\varphi_{Z_n}(t) = \mathbb{E}[e^{it\{(\sum_{i=1}^{n} X_i)/\sqrt{n}\}}] = \mathbb{E}[\prod_{i=1}^{n} e^{it(X_i/\sqrt{n})}]$$

$$= \left[\mathbb{E}[e^{it(X_i/\sqrt{n})}]\right]^n = \left[\mathbb{E}[1 + it(\frac{X_i}{\sqrt{n}}) + 1/2(\frac{itX_i}{\sqrt{n}})^2 + o[(\frac{itX_i}{\sqrt{n}})^2]\right]^n$$

$$= \left[1 - \frac{t^2}{2n} + o(\frac{-t^2}{2n})\right]^n \Rightarrow \lim_{n\to\infty} \varphi_{Z_n}(t) \to e^{\frac{-t^2}{2}} \quad \text{☺}$$

Back

Pickands–Balkema–de Haan Theorem (1974, 1975)

If $M_n = \max_{1 \le i \le n}\{X_i\} \approx \text{GEV}(\mu, \sigma, \xi)$, then, for a "large" u (i.e., $u \to x_F = \sup\{x : F(x) < 1\}$), $F_u = \mathbb{P}(X - u < y | X > u)$ is well approximated by the generalized Pareto distribution (GPD). That is:

$$F_u(y) \xrightarrow{d} H_{\tilde{\sigma},\xi}(y) \qquad u \to x_F$$

where

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

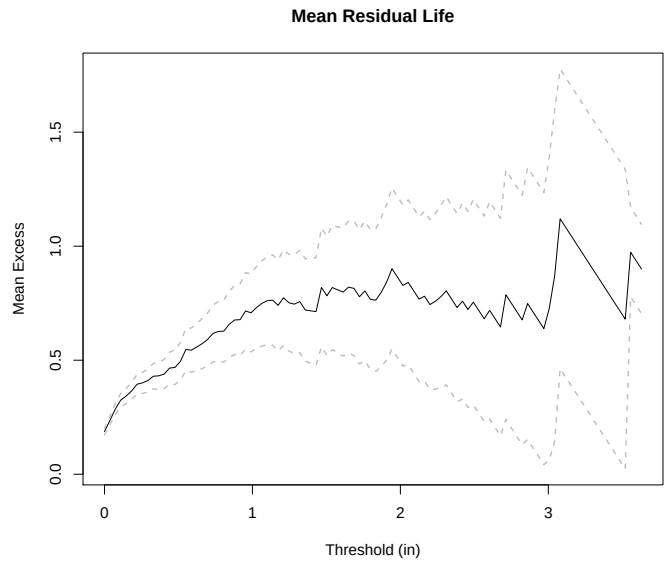
and $\tilde{\sigma} = \sigma + \xi(u - \mu)$

Back

How to Choose the Threshold?

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



Task: To choose a u_0 s.t. the Mean Residual Life curve behaves linearly $\forall u > u_0$

Back