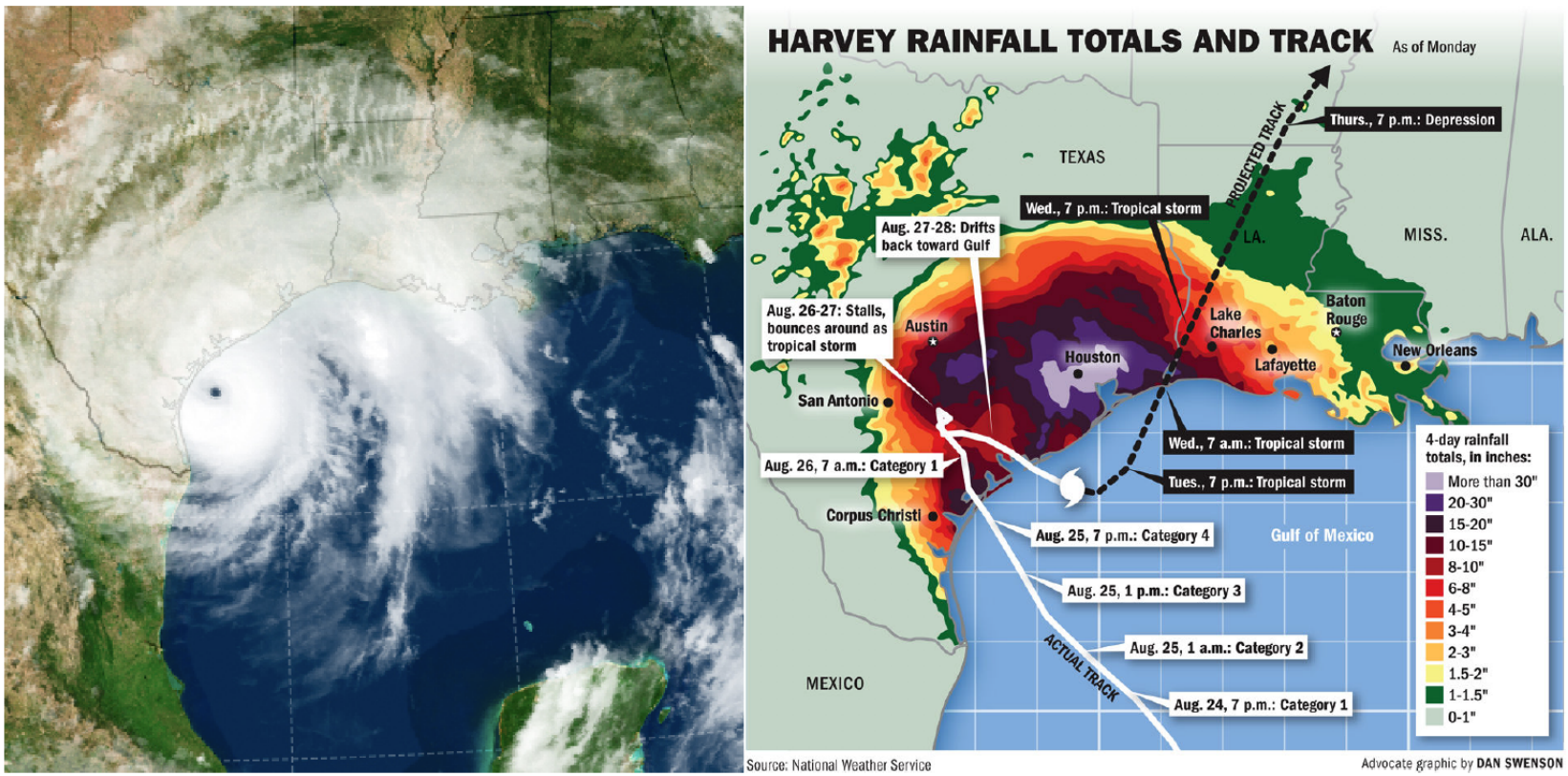
Estimating Precipitation Extremes using Log-Histospline

Whitney Huang[1]
Joint work with Doug Nychka[2] and Hao Zhang[3]

Clemson[1], Colorado School of Mines[2], Purdue University[3]

University of Georgia, August 29, 2019

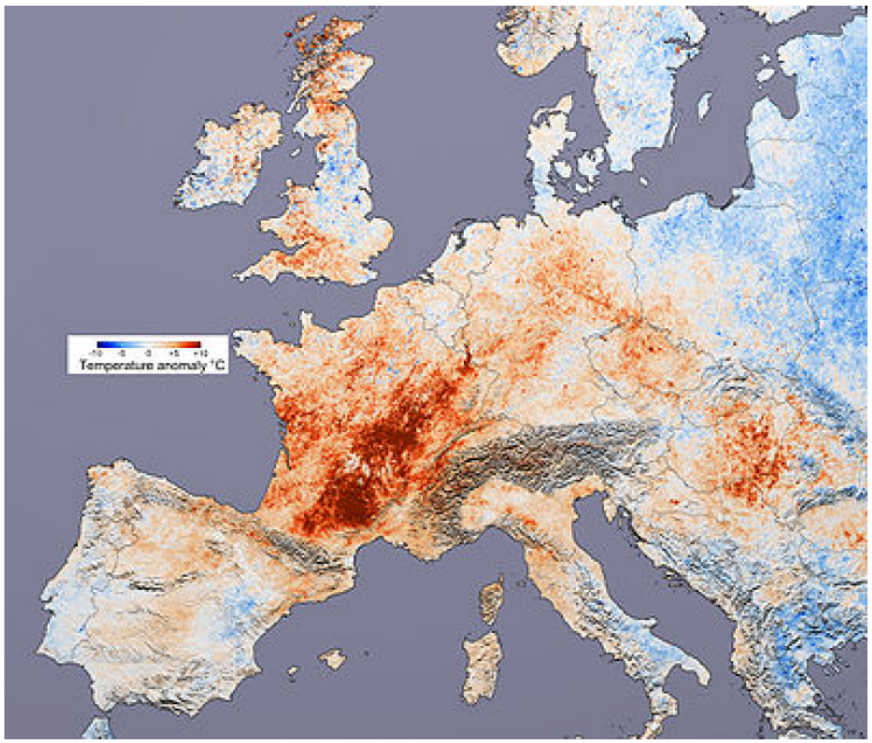
Extreme rainfall during Hurricane Harvey



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)

Environmental extremes: Heatwaves, storm surges, etc.



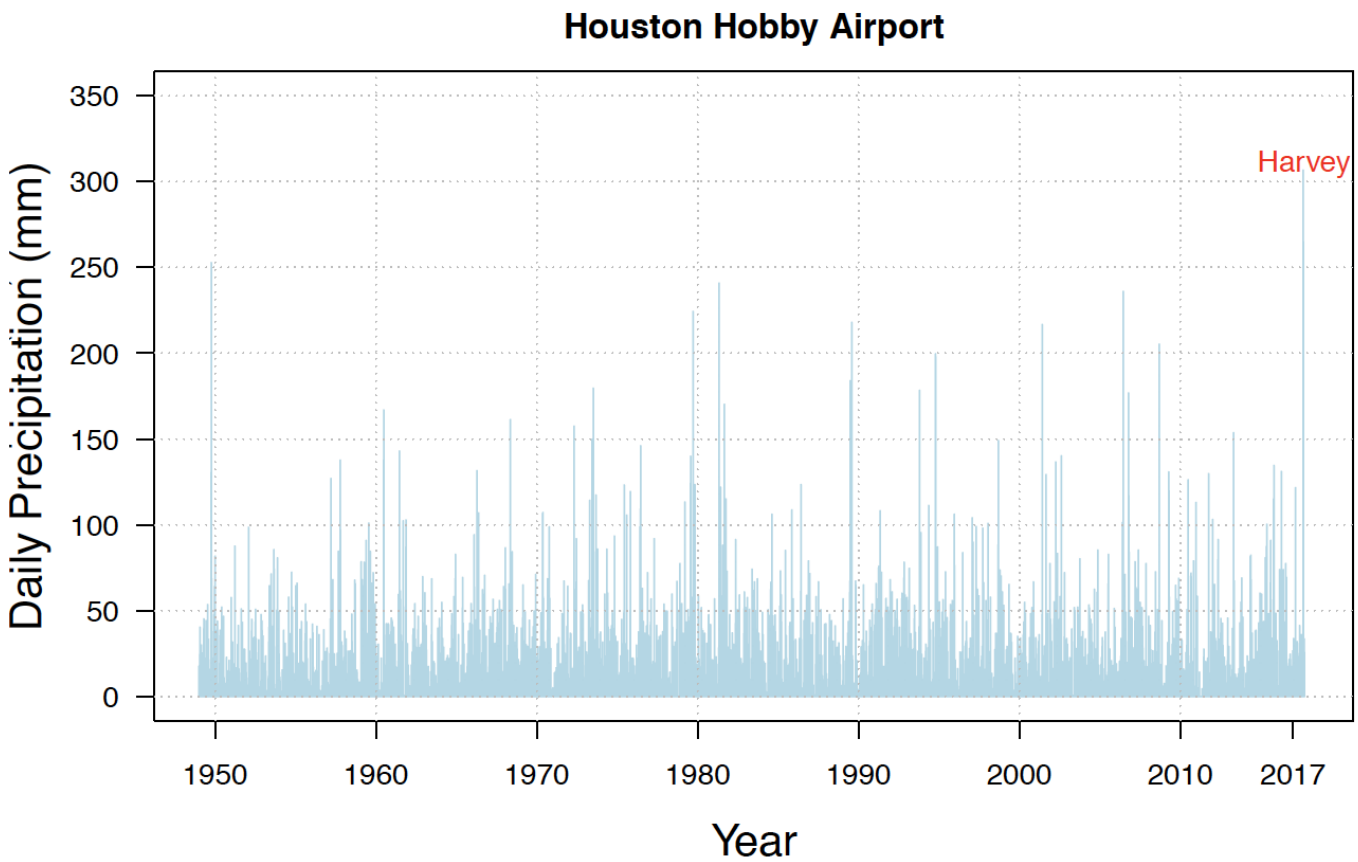
- **Heat wave**: The 2003 European heat wave led to the hottest summer on record in Europe since 1540 that resulted in at least 30,000 deaths

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

Big picture of the talk

- Scientific question: How to estimate the magnitude of extreme events (e.g. 100-year rainfall)
- Extreme value theory provides a very useful framework for estimating extremes but requires selecting a small fraction of the "large" observations
- Log-Histospline: combine **data transformation** and **spline smoothing** to estimate the precipitation extremes using the **full range** of the observations

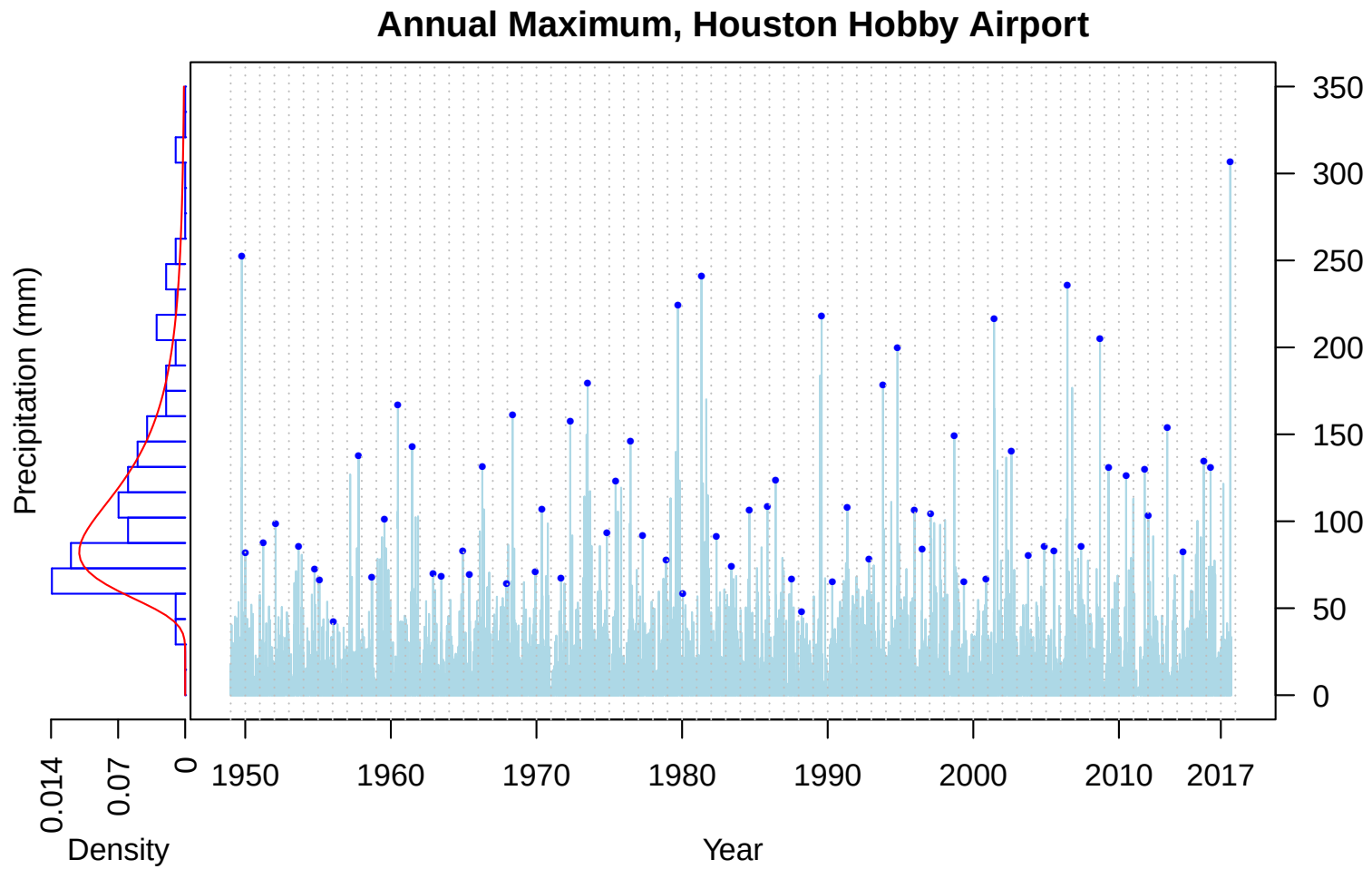
Part I: Extreme Value Analysis



Data source: Global Historical Climatology Network

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

Estimating extremes using block maxima (Gumbel 1958)



Which distribution to use for annual maxima?

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

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

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

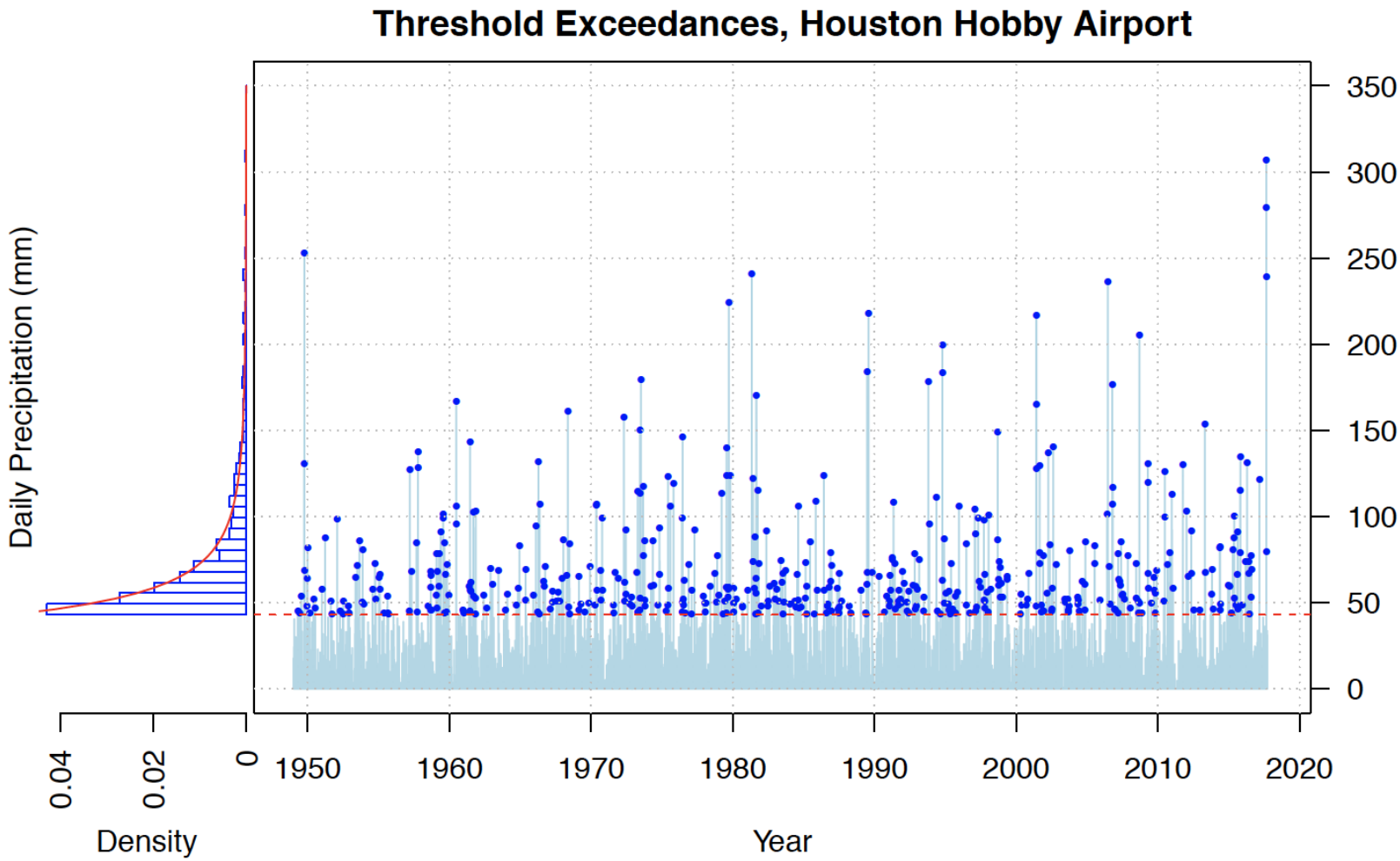
then G must be the same type of the following form:

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

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)

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



Which distribution to use for threshold exceedances?

$\Rightarrow$ generalized Pareto distribution ($\mathrm{GPD}_u(\tilde{\sigma}, \xi)$)

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

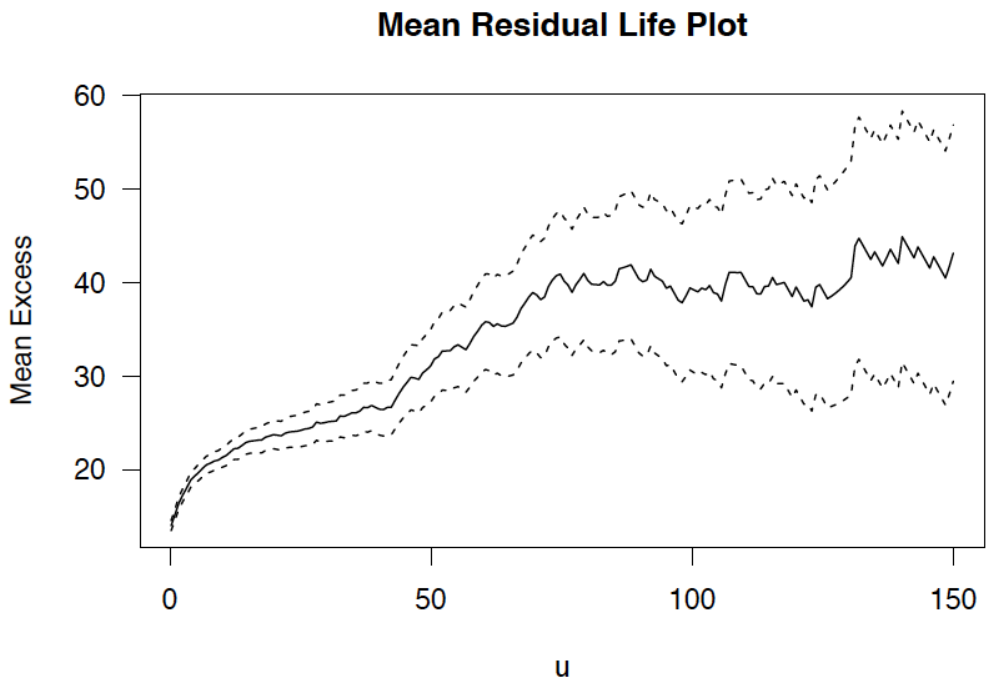
$$H_{\tilde{\sigma},\xi}(y) = \begin{cases} 1 - (1 + \xi y/\tilde{\sigma})^{-1/\xi} & \xi \neq 0; \\ \\ 1 - \exp(-y/\tilde{\sigma}) & \xi = 0. \end{cases}$$

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

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

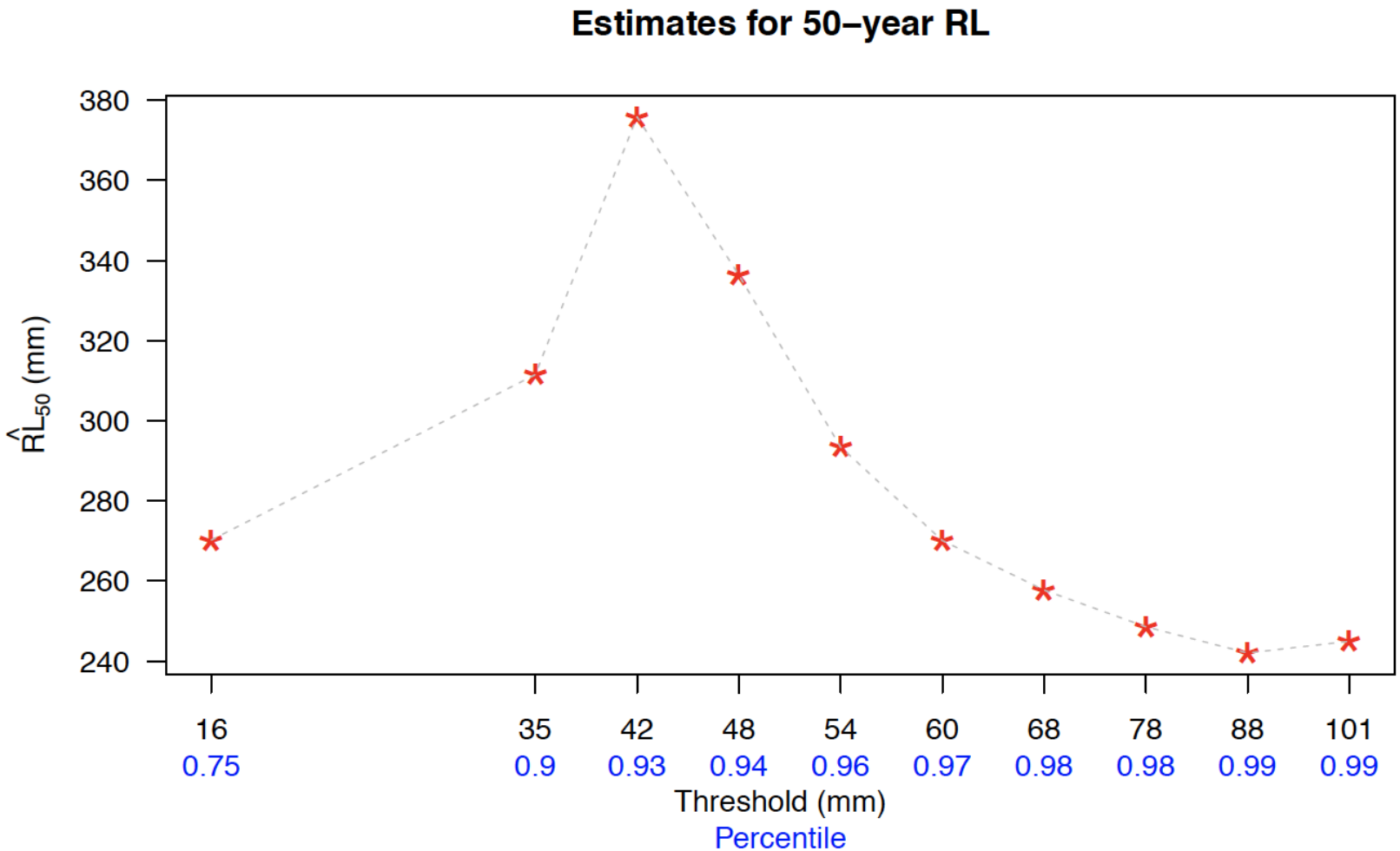
Bias–variance tradeoff:

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



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

Return level estimate is sensitive to the chosen threshold ☹



Part II: Log-Histospline

Huang, W. K., Nychka, D. W., & Zhang, H.
Estimating precipitation extremes using the log–histospline.
Environmetrics, e2543, 2019

Motivation: To develop an alternative to POT method

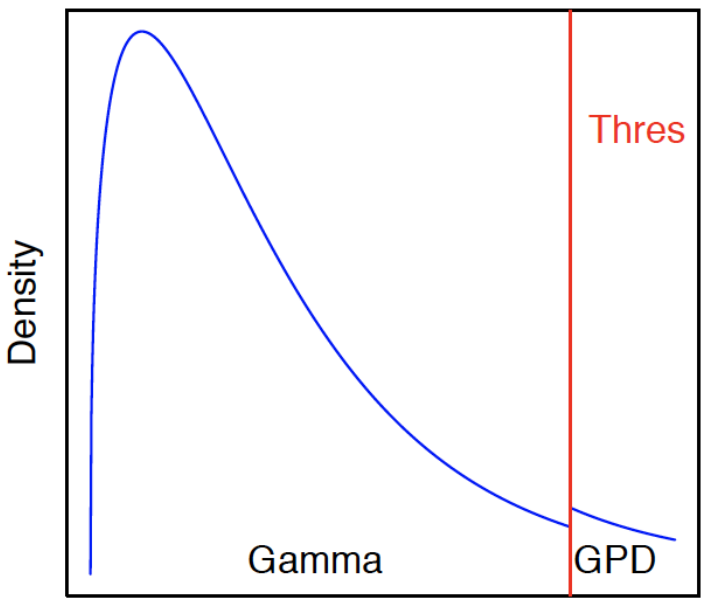
- Avoid the threshold selection
- Incorporate prior knowledge on the rainfall tail distributions
  - Polynomial upper tail behavior (i.e., $\xi > 0$)
  - Bounded lower tail

$\Rightarrow$ Model the full range of the distribution (i.e., $(0, \infty)$) while accounting for polynomial upper tail behavior $\xi > 0$ and bounded lower tail

Some existing methods for estimating extremes while modeling the full range of the distribution

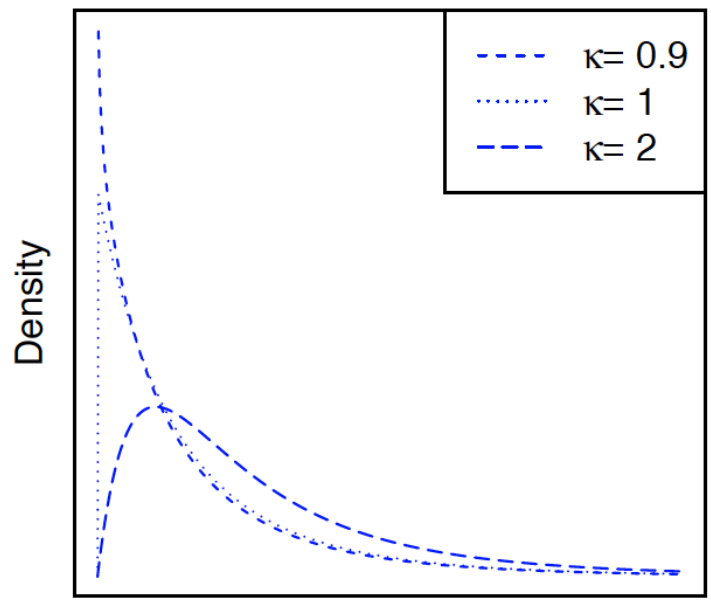
- Mixture models [e.g., Frigessi et al., 2002]

Parametric/Non-Parametric bulk+ GPD tail



- EGPD [e.g., Naveau et al., 2016]

$Y = H^{-1}_{\sigma,\xi}(G^{-1}(U))$ (e.g., $G(u) = u^{\kappa}, \quad \kappa > 0$)



Main 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 model the log-density $g(x) = \log f(x)$

- Working on log-density makes life easier

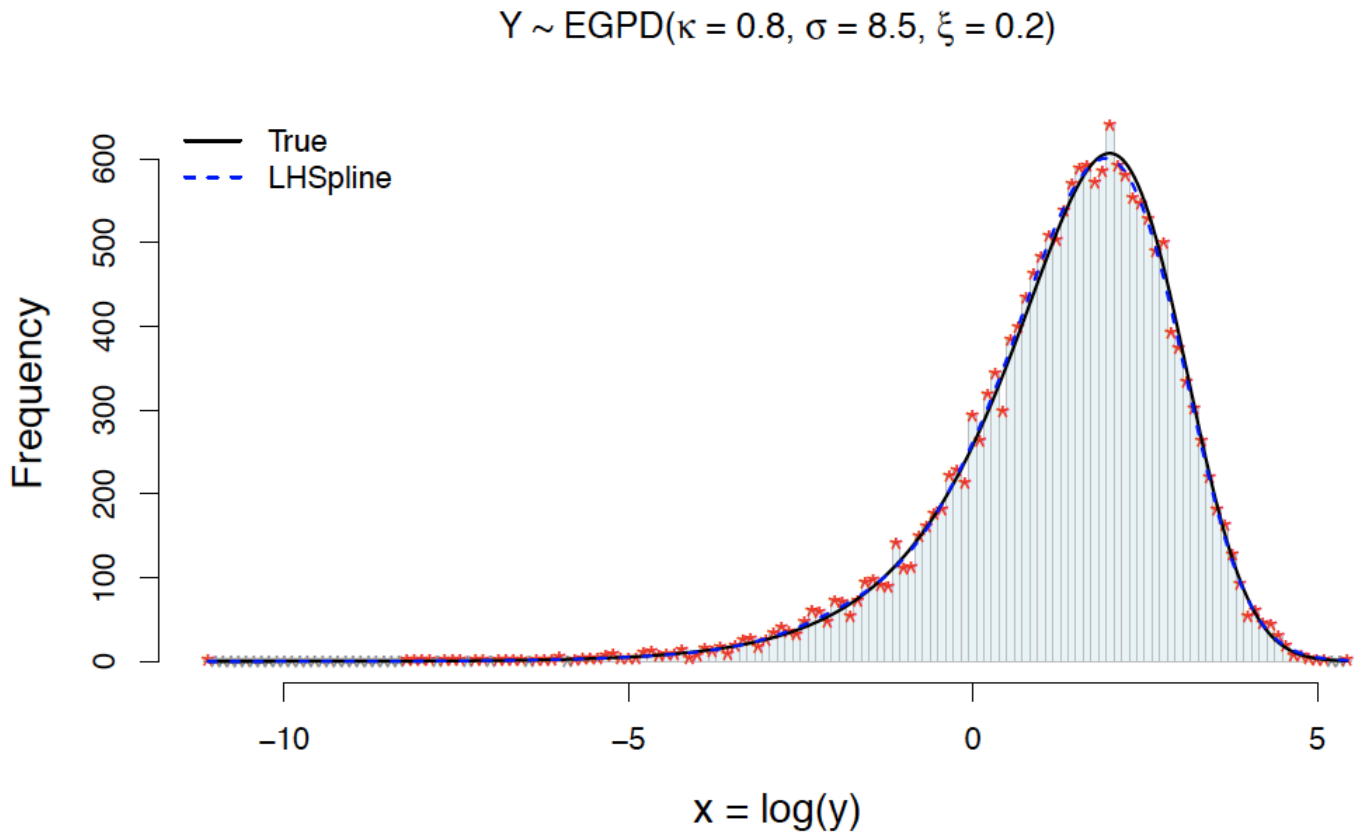
  $\Rightarrow$ remove the positivity constraint

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

Log-Histospline (LHSpline): Smooth the histogram by fitting a (penalized) Possion regression [Efron & Tibshirani 1996]



Assumption: $Z(x_j) \overset{ind}{\sim} \text{Poi}(\tilde{f}(x_j)), j = 1, \cdots, N$, where $Z(x_j)$ is the j$^{\text{th}}$ bin count. We use a **penalized Possion regression** to estimate $g(x), x \in (-\infty, \infty)$, the log-density

Estimating (log)-density by maximizing penalized likelihood

[Good and Gaskins, 1971, R. Tapia, 1978, Silverman, 1982]

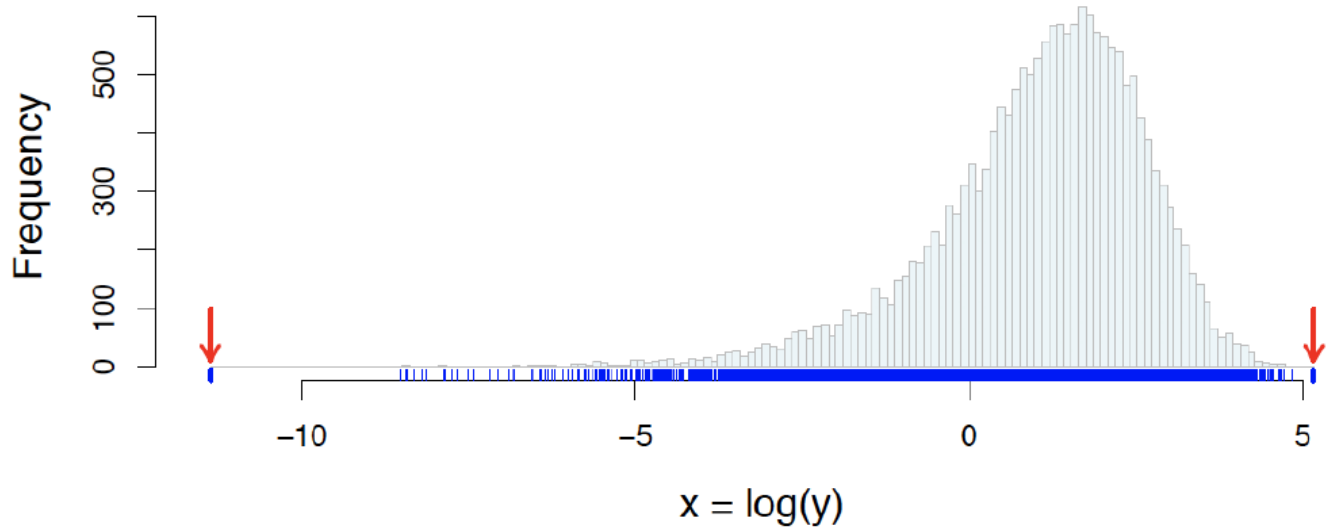
Penalized negative log likelihood:

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

where λ is the smoothing parameter

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

Bias correction for removing edge effects



These "bumps" causes overestimation in extreme upper quantiles

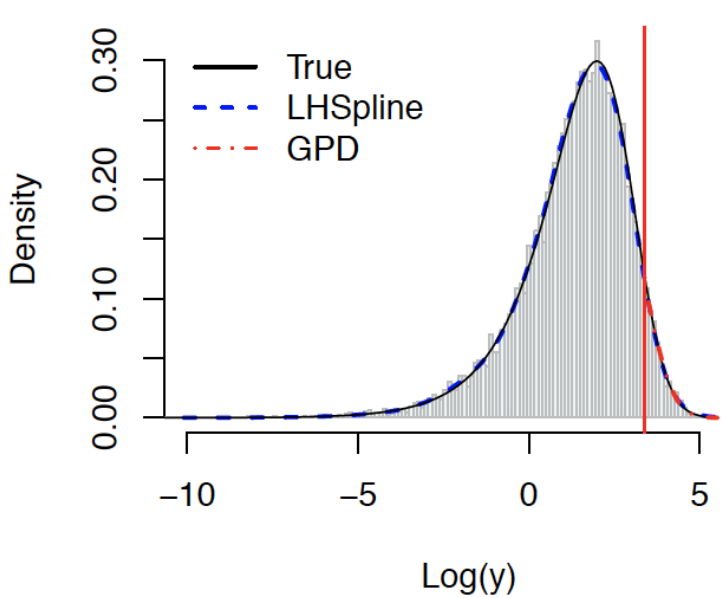
- Extend the histogram beyond the data range by exploiting the equivalence of density and intensity estimation for Poisson processes, and
  1. **Bootstrap for bias correction**: Use the "parametric" bootstrap to estimate/correct the bias
  2. **Adjust the smoothing parameter** λ: The λ chosen by CV maybe be too "large" to explore the tail features

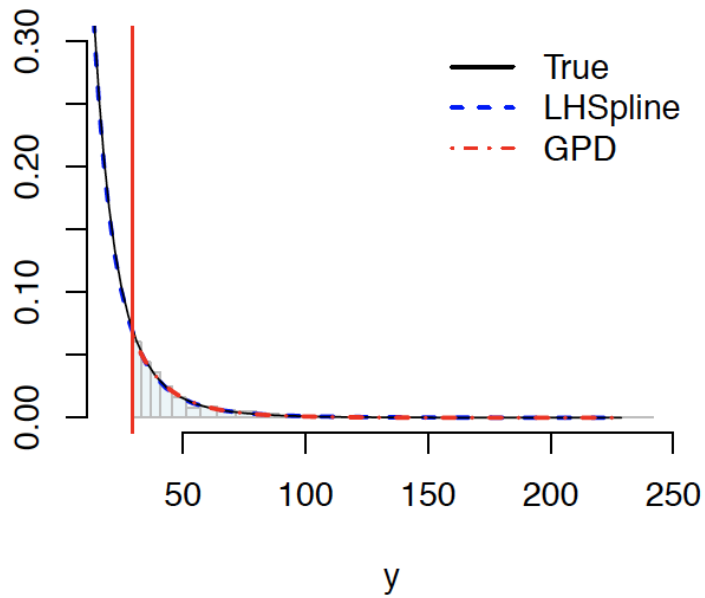
Simulation Study

Set-up:

- Generate 100 data sets each with 50 years ($n = 18{,}250$) of observations that follows EGPD($\kappa = 0.8, \sigma = 8.5, \xi = 0.2$)

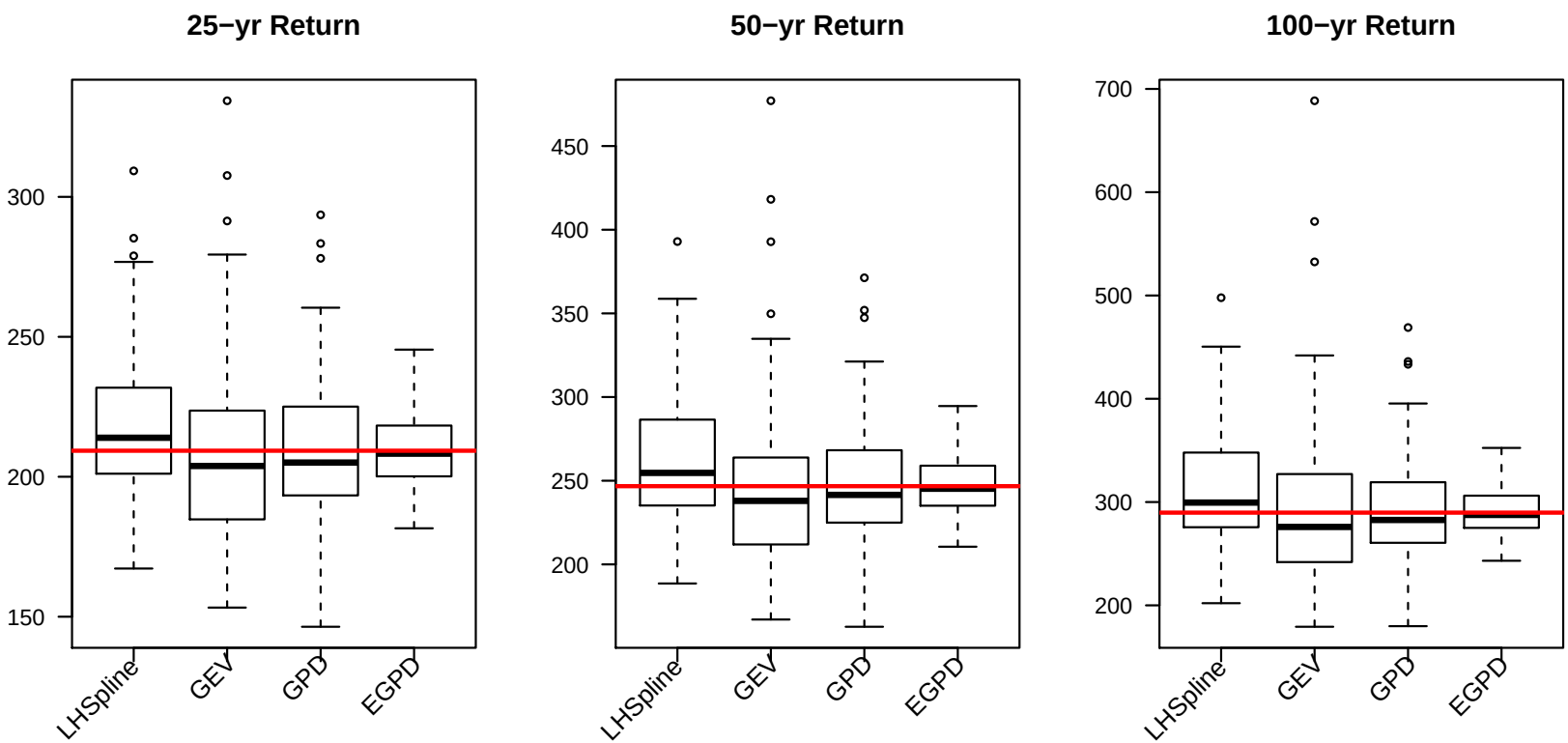
- Estimate return levels using LHSpline and compare the estimator performance with the GEV/GPD





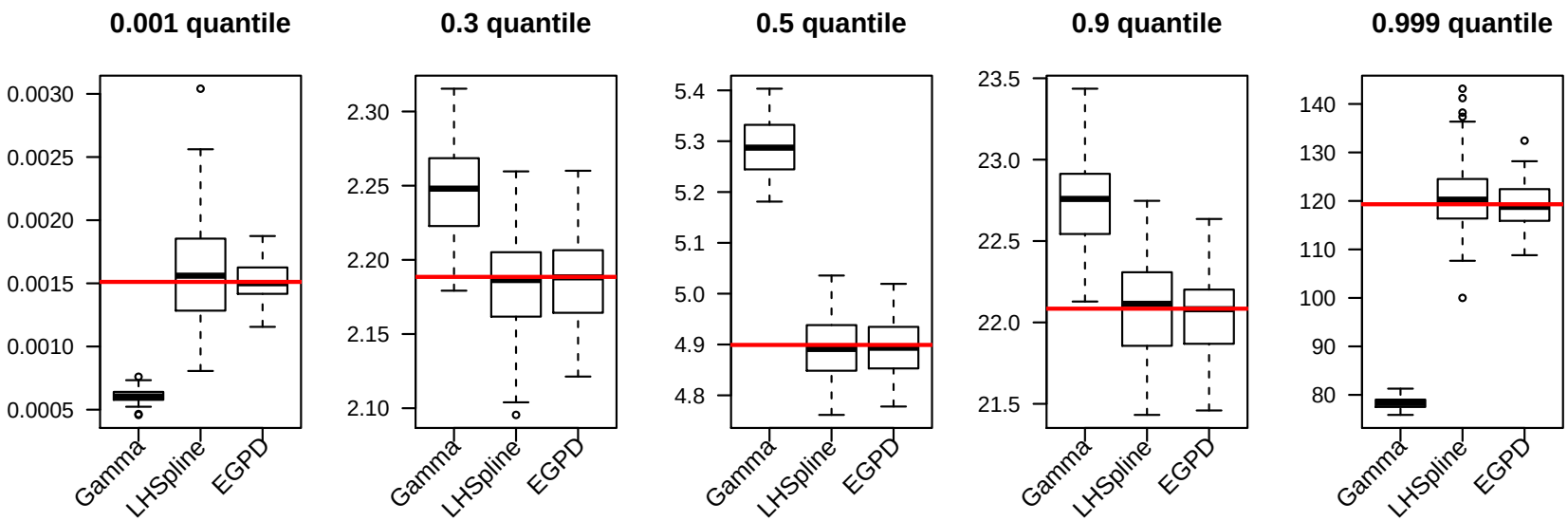
Return level estimates

- LHSpline Fit Log-Histospline to the full range of the log-transformed data
- GEV: Fit GEV to "annual maxima"
- GPD: Fit GPD to threshold exceedances
- EGPD: Fit EGPD to the full range of the data ("Oracle")



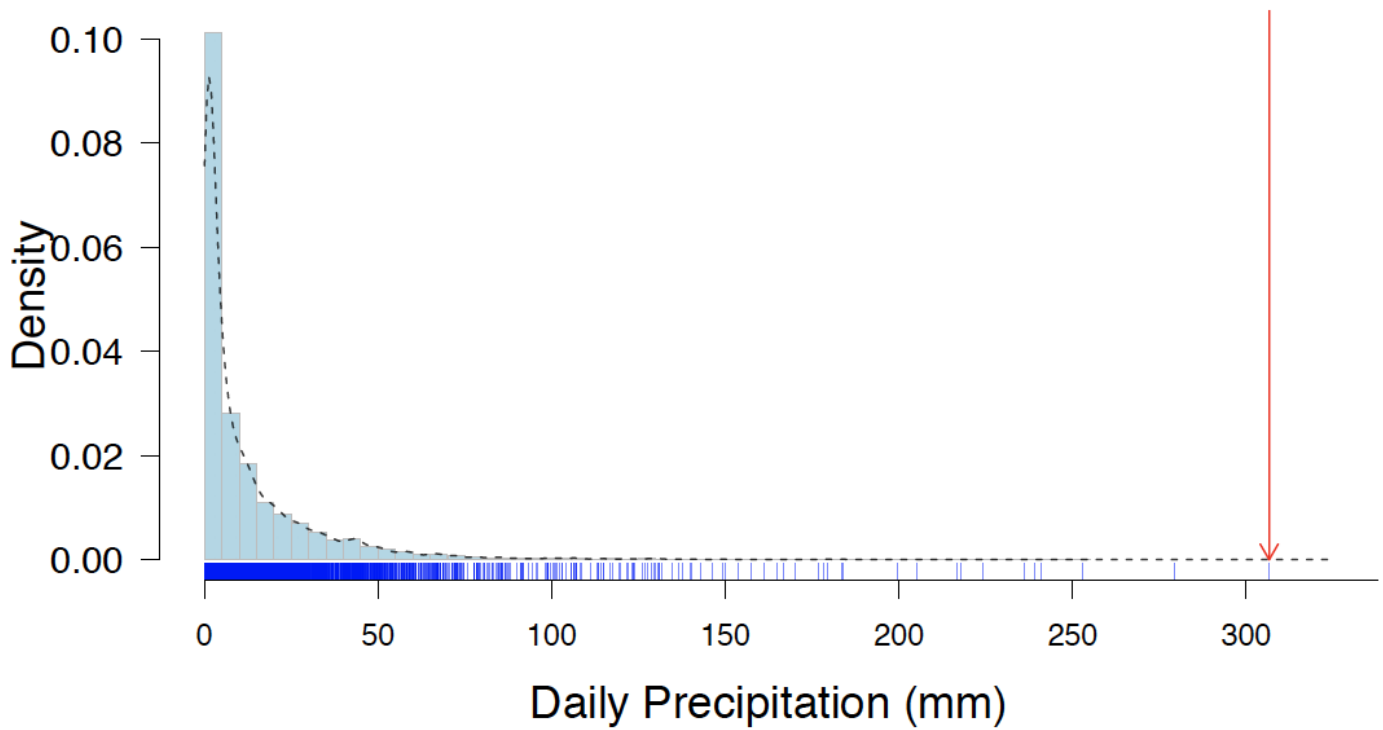
Quantile estimates

- LHSpline Fit Log-Histospline to the full range of the log-transformed data
- Gamma: Fit Gamma to the full range of the data [Wilks, 2011]
- EGPD: Fit EGPD to the full range of the data ("Oracle")

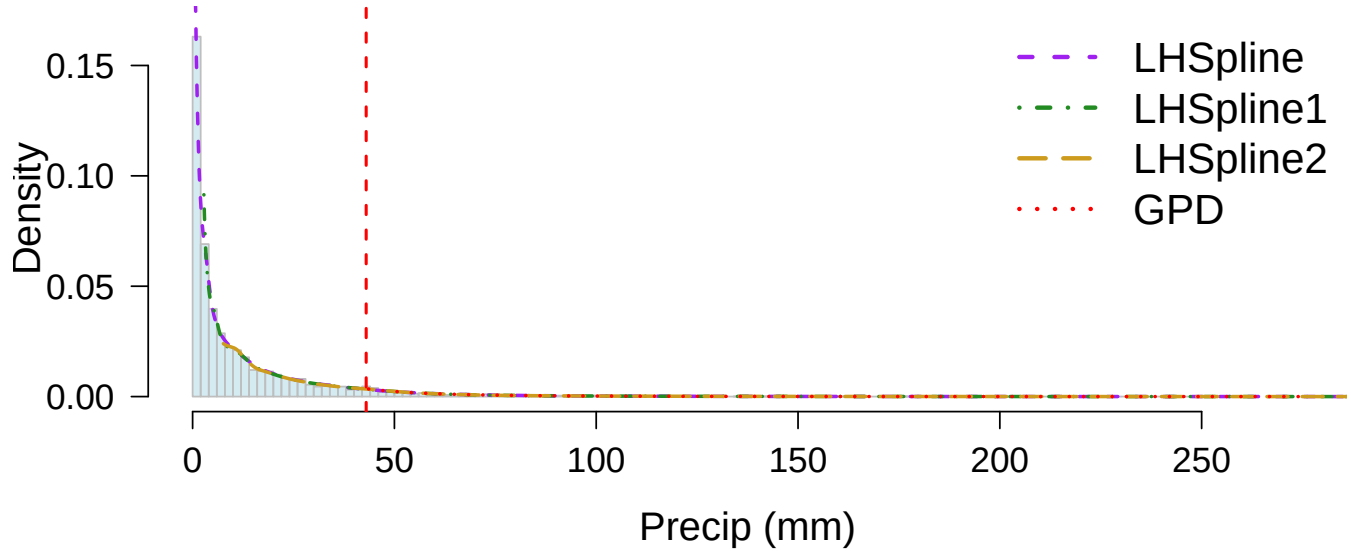


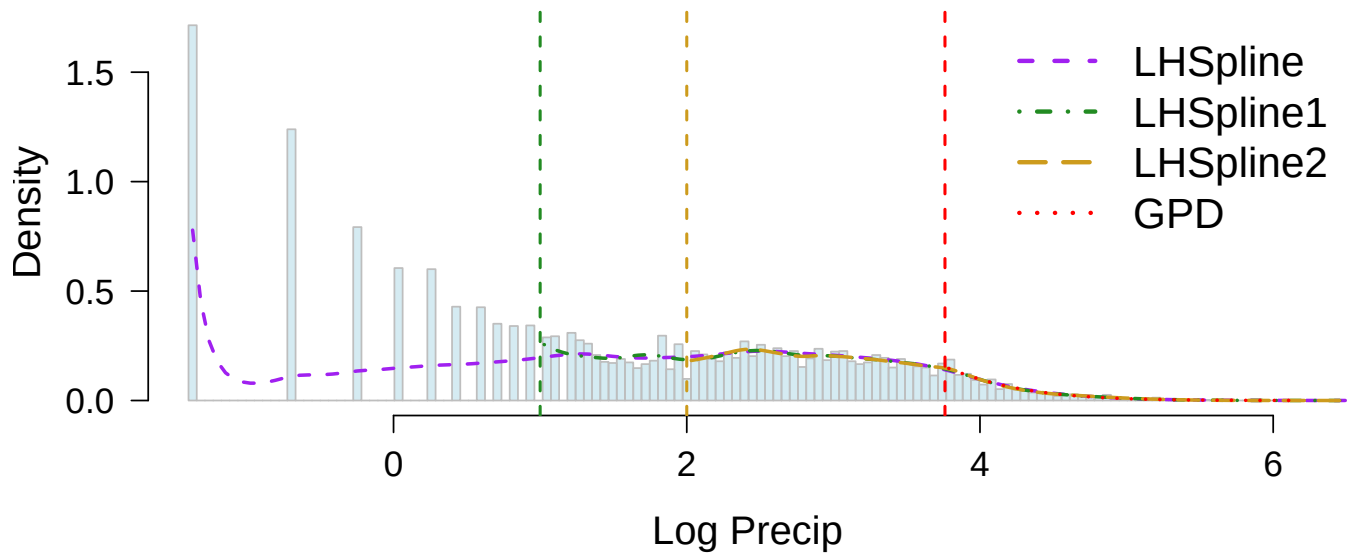
Part III: Houston Precipitation Extremes Revisited

- Use the daily rainfall data (wet day only $\sim 28\%$) from 1949 - 2016 for the model fitting
- **Question**: *How unusual was the event of 306.58 mm on 26th August, 2017?*

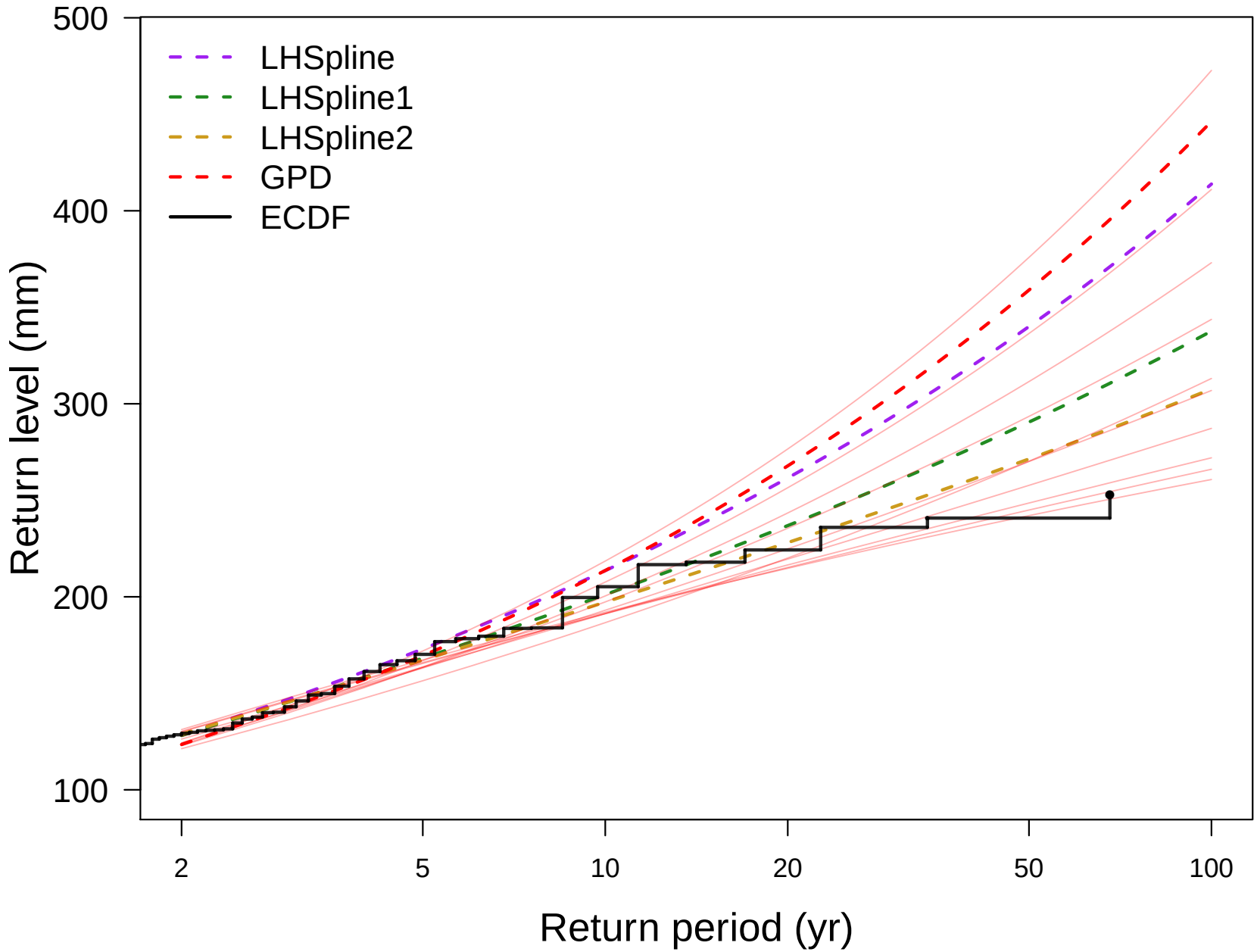


LHSpline fits to Houston Hobby Airport rainfall data





Return level estimates



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

Method	GPD	LHSpline1	LHSpline2
Estimate (years)	30.5	64.0	98.0
90% CI Lower limit	17.0	19.1	22.0
90% CI Upper limit	73.3	172.4	345.7

Estimation Uncertainty

- LHSpline: Via conditional simulation to obtain Bayesian "confidence interval" [Wahba, 1990, Green & Silverman 1994]

- GPD: Via profile likelihood [Coles, 2001]

Summary and discussion

LHSpline

+: **Does not require the threshold selection** when modeling the tail distribution

+: Applicable to **heavy-tailed distribution**: streamflow, financial return, ...

-: Only applicable to heavy-tailed distribution

-: LHSpline requires some tunings

Ongoing work

- **Nonstationary extension** to account for both seasonality and long term trend
- **Spatial extension** by using the methods from functional data analysis to facilitate the spatial smoothing of the log-densities in a spatial region to compute return level maps

Thank you for your attention!

Summary and discussion

LHSpline

+: **Does not require the threshold selection** when modeling the tail distribution

+: Applicable to **heavy-tailed distribution**: streamflow, financial return, ...

-: Only applicable to heavy-tailed distribution

-: LHSpline requires some tunings

Ongoing work

- **Nonstationary extension** to account for both seasonality and long term trend
- **Spatial extension** by using the methods from functional data analysis to facilitate the spatial smoothing of the log-densities in a spatial region to compute return level maps

Thank you for your attention!

Summary and discussion

LHSpline

+: **Does not require the threshold selection** when modeling the tail distribution

+: Applicable to **heavy-tailed distribution**: streamflow, financial return, ...

-: Only applicable to heavy-tailed distribution

-: LHSpline requires some tunings

Ongoing work

- **Nonstationary extension** to account for both seasonality and long term trend
- **Spatial extension** by using the methods from functional data analysis to facilitate the spatial smoothing of the log-densities in a spatial region to compute return level maps

Thank you for your attention!

Summary and discussion

LHSpline

+: **Does not require the threshold selection** when modeling the tail distribution

+: Applicable to **heavy-tailed distribution**: streamflow, financial return, ...

-: Only applicable to heavy-tailed distribution

-: LHSpline requires some tunings

Ongoing work

- **Nonstationary extension** to account for both seasonality and long term trend
- **Spatial extension** by using the methods from functional data analysis to facilitate the spatial smoothing of the log-densities in a spatial region to compute return level maps

Thank you for your attention!

Summary and discussion

LHSpline

+: **Does not require the threshold selection** when modeling the tail distribution

+: Applicable to **heavy-tailed distribution**: streamflow, financial return, ...

-: Only applicable to heavy-tailed distribution

-: LHSpline requires some tunings

Ongoing work

- **Nonstationary extension** to account for both seasonality and long term trend
- **Spatial extension** by using the methods from functional data analysis to facilitate the spatial smoothing of the log-densities in a spatial region to compute return level maps

Thank you for your attention!

Summary and discussion

LHSpline

+: **Does not require the threshold selection** when modeling the tail distribution

+: Applicable to **heavy-tailed distribution**: streamflow, financial return, ...

-: Only applicable to heavy-tailed distribution

-: LHSpline requires some tunings

Ongoing work

- **Nonstationary extension** to account for both seasonality and long term trend
- **Spatial extension** by using the methods from functional data analysis to facilitate the spatial smoothing of the log-densities in a spatial region to compute return level maps

Thank you for your attention!

Summary and discussion

LHSpline

+: **Does not require the threshold selection** when modeling the tail distribution

+: Applicable to **heavy-tailed distribution**: streamflow, financial return, ...

-: Only applicable to heavy-tailed distribution

-: LHSpline requires some tunings

Ongoing work

- **Nonstationary extension** to account for both seasonality and long term trend
- **Spatial extension** by using the methods from functional data analysis to facilitate the spatial smoothing of the log-densities in a spatial region to compute return level maps

Thank you for your attention!

Summary and discussion

LHSpline

+: **Does not require the threshold selection** when modeling the tail distribution

+: Applicable to **heavy-tailed distribution**: streamflow, financial return, ...

-: Only applicable to heavy-tailed distribution

-: LHSpline requires some tunings

Ongoing work

- **Nonstationary extension** to account for both seasonality and long term trend
- **Spatial extension** by using the methods from functional data analysis to facilitate the spatial smoothing of the log-densities in a spatial region to compute return level maps

Thank you for your attention!