Some (Useful) Statistical Tools for Environmental Research

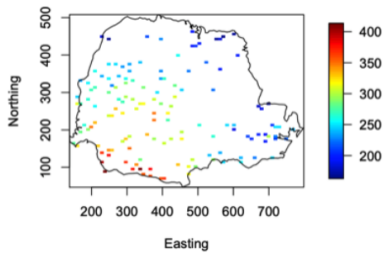
Whitney Huang
Assistant Professor of Applied Statistics and Data Science

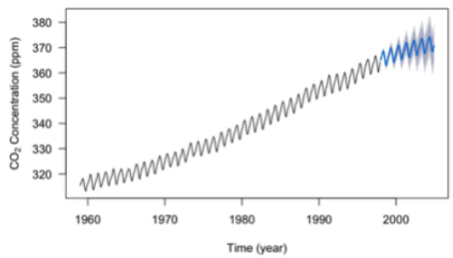
wkhuang@clemson.edu

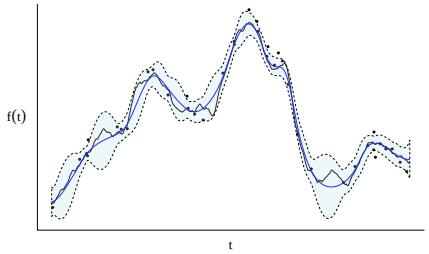
Computational Sciences and Engineering Division
May 8, 2023

Overview of My Research

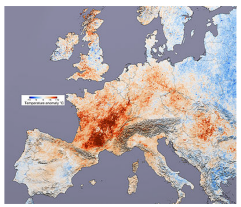
- ## Spatio-Temporal Statistics

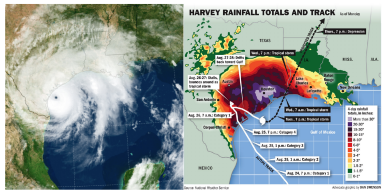




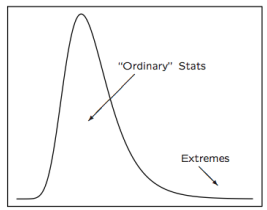


- ## Extreme Value Analysis



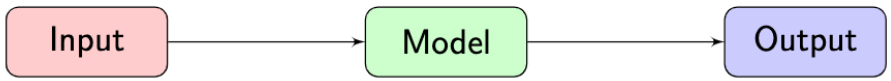


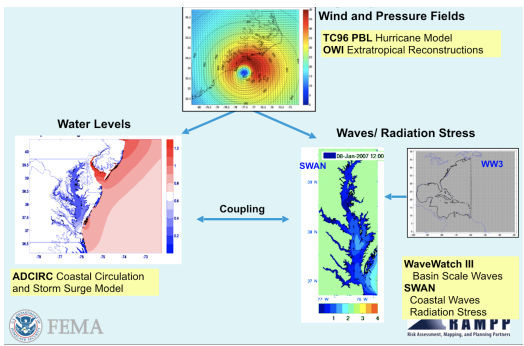
Source: NASA (Left); National Weather Service (Right)

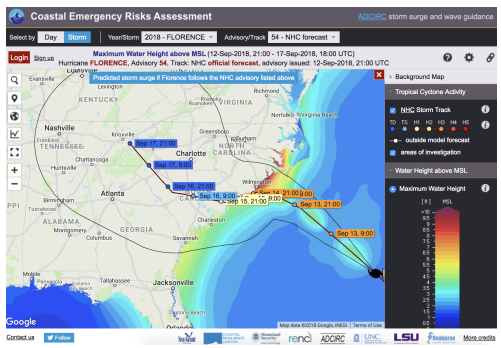


- ## Surrogate Modeling of Computer Experiments

$$\boldsymbol{x} \in \mathcal{X} \qquad \eta : \mathcal{X} \mapsto \mathcal{Y} \qquad y = \eta(\boldsymbol{x})$$

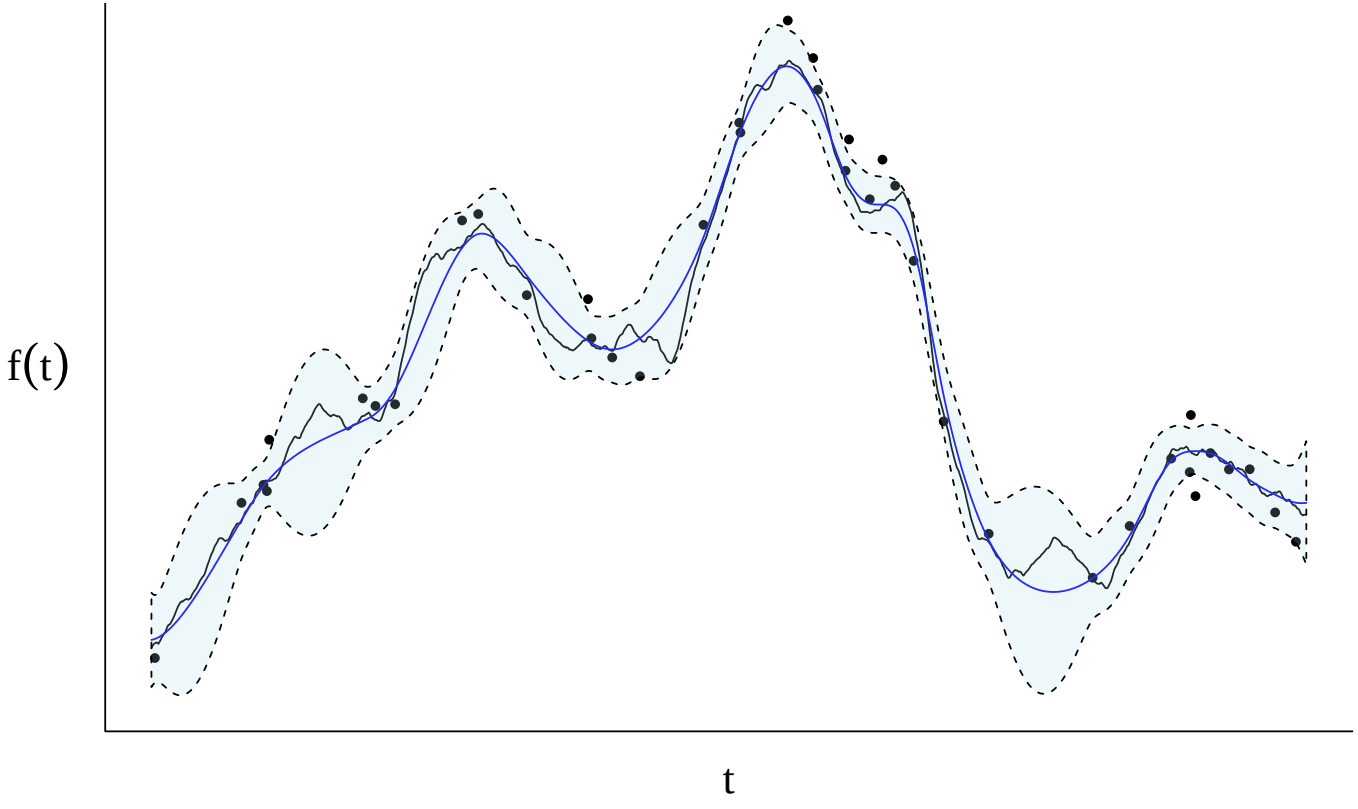




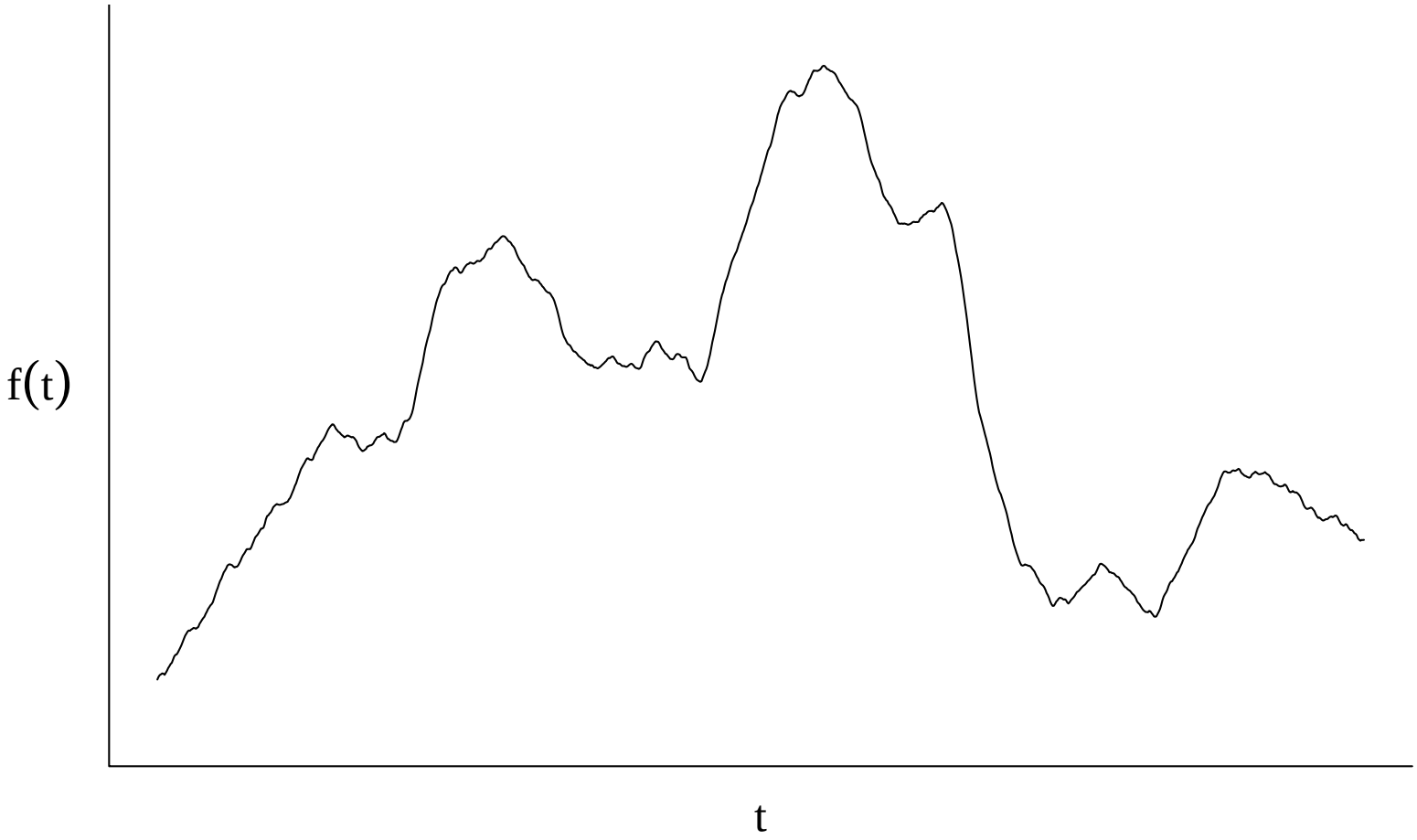


Gaussian Processes

$$f(t) \sim \mathrm{GP}(m(t), K(t, t')), \quad t \in \mathcal{T}$$

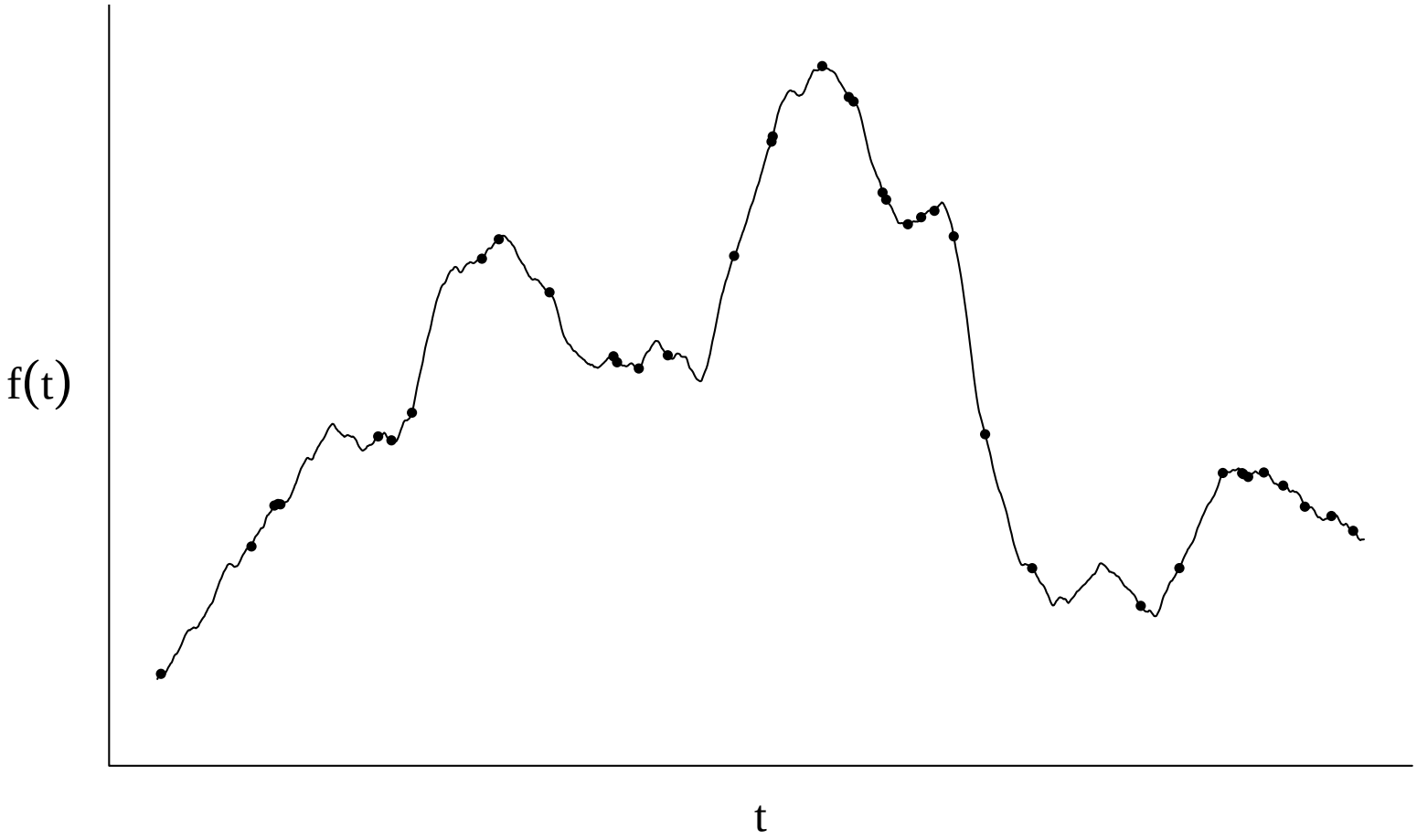


Function Estimation



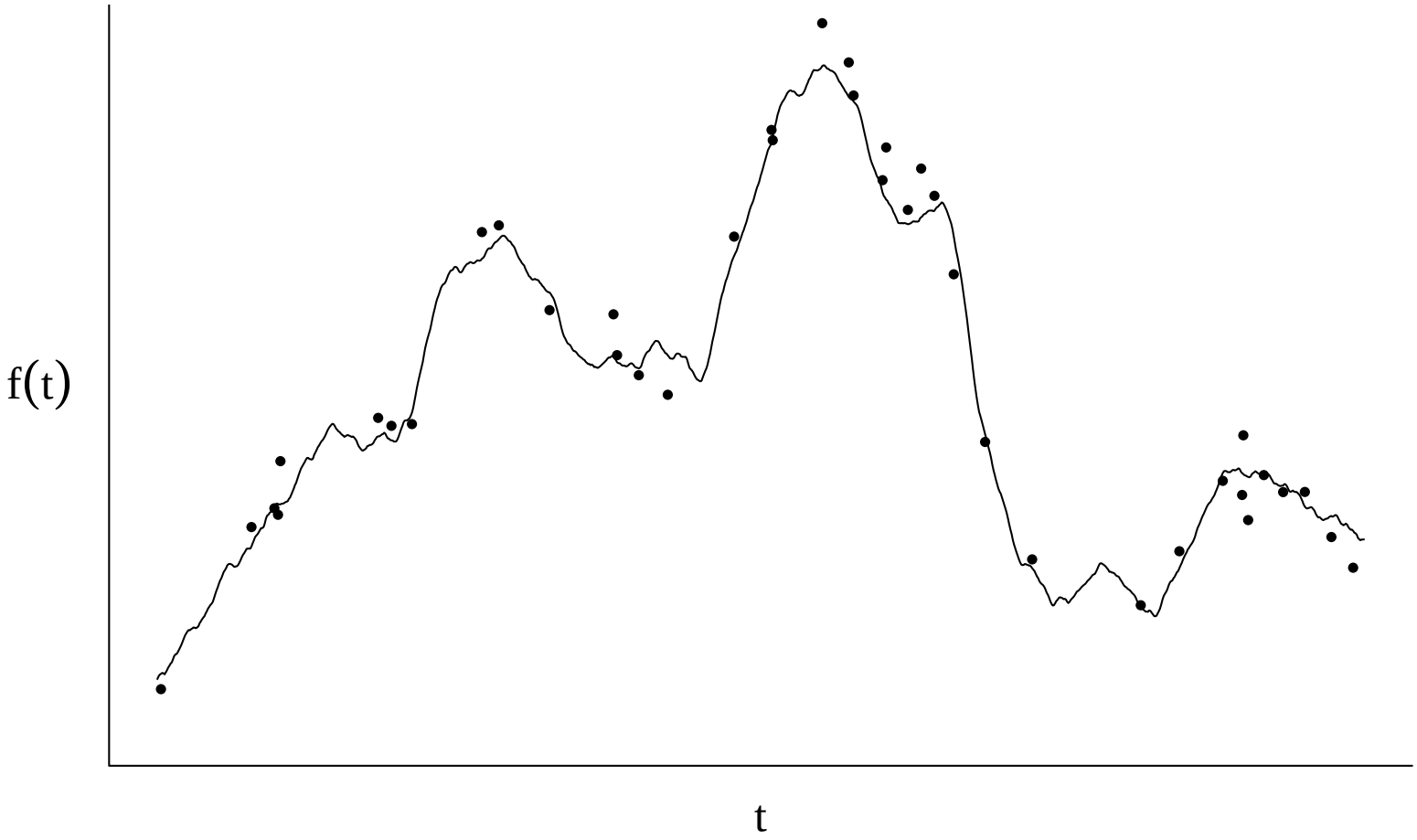
Consider a function, say $f(t)$

Function Estimation



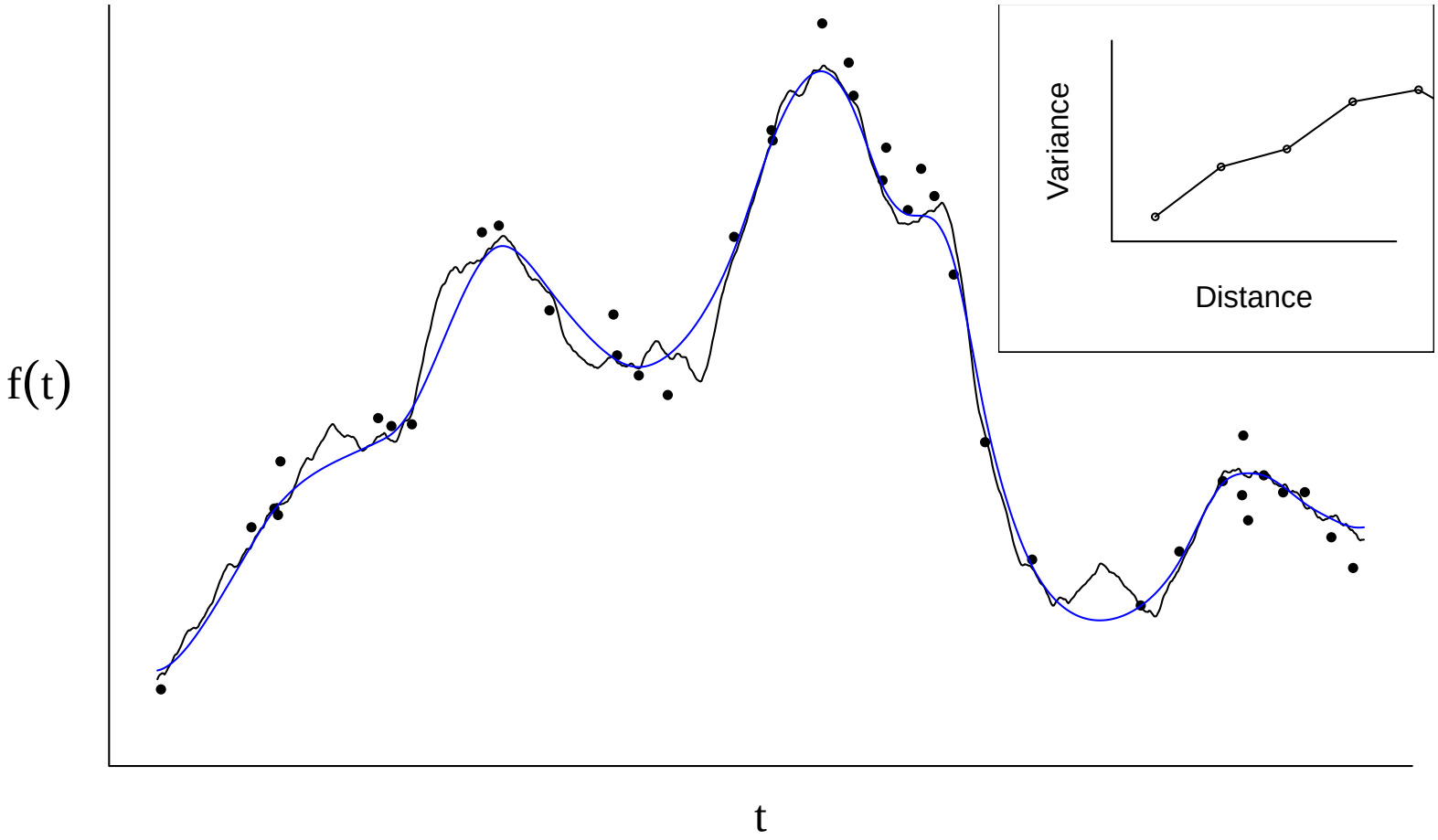
Consider $f(t)$, observed incompletely $\{f(t_i)\}_{i=1}^{n}$

Function Estimation



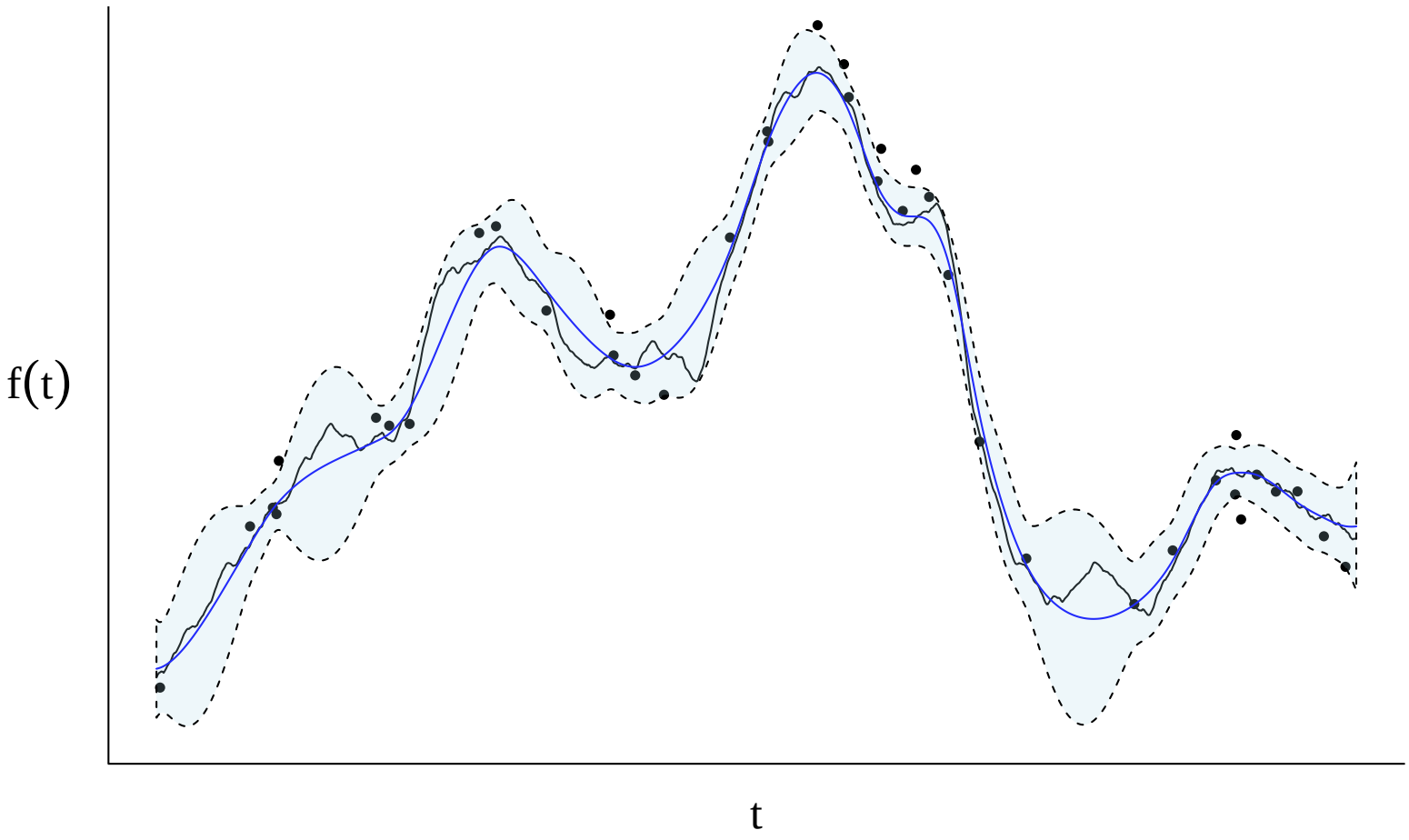
Consider $f(t)$, observed incompletely $\{f(t_i)\}_{i=1}^{n}$, and with noise

Gaussian Processes (GPs): Function Estimators



Main idea: exploit the inter-point correlation to estimate $f(t)$

GPs: Probabilistic Function Estimators



GP provides an optimal estimate $\hat{f}(t)$ along with "localized" uncertainty quantification ("error bars")

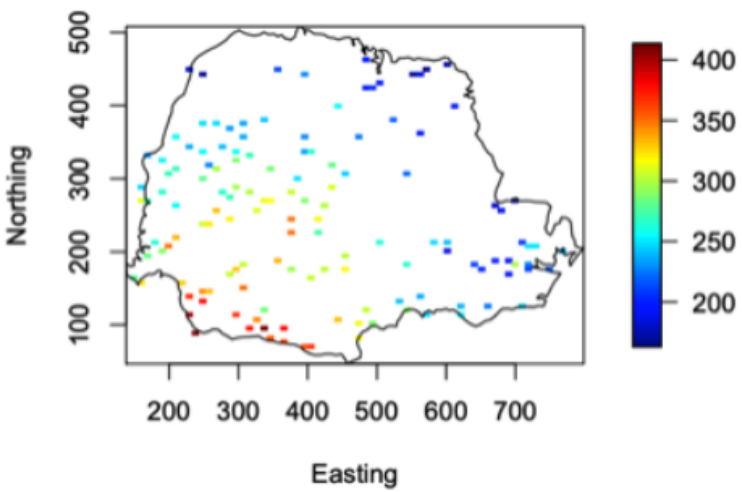
Spatial Interpolation via GP Model

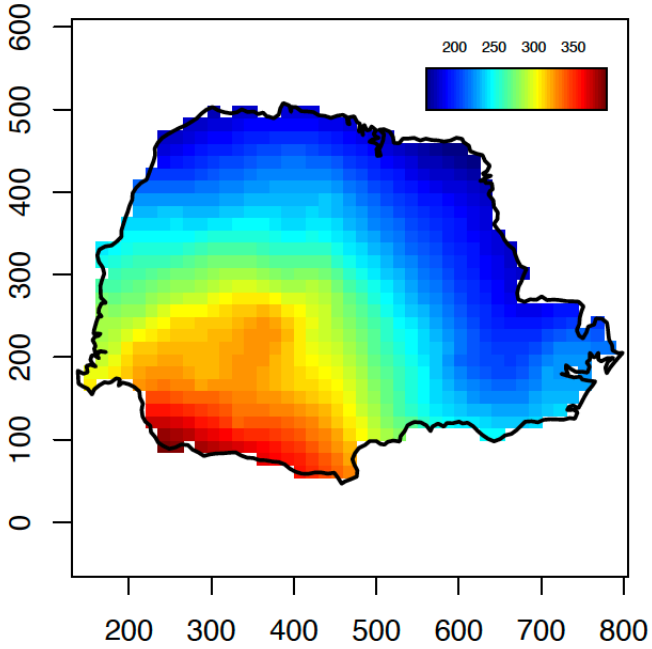
The "generic" spatial model

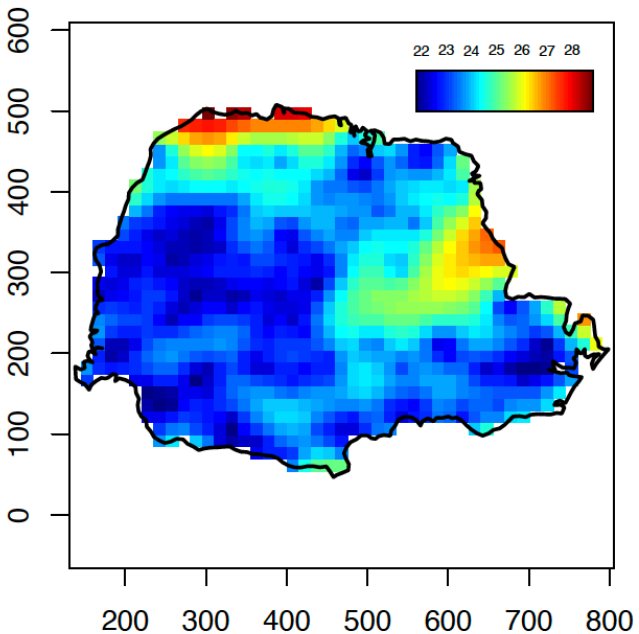
$$Y(\boldsymbol{s}) = m(\boldsymbol{s}) + \eta(\boldsymbol{s}),\ \boldsymbol{s} \in \mathcal{S},$$

where

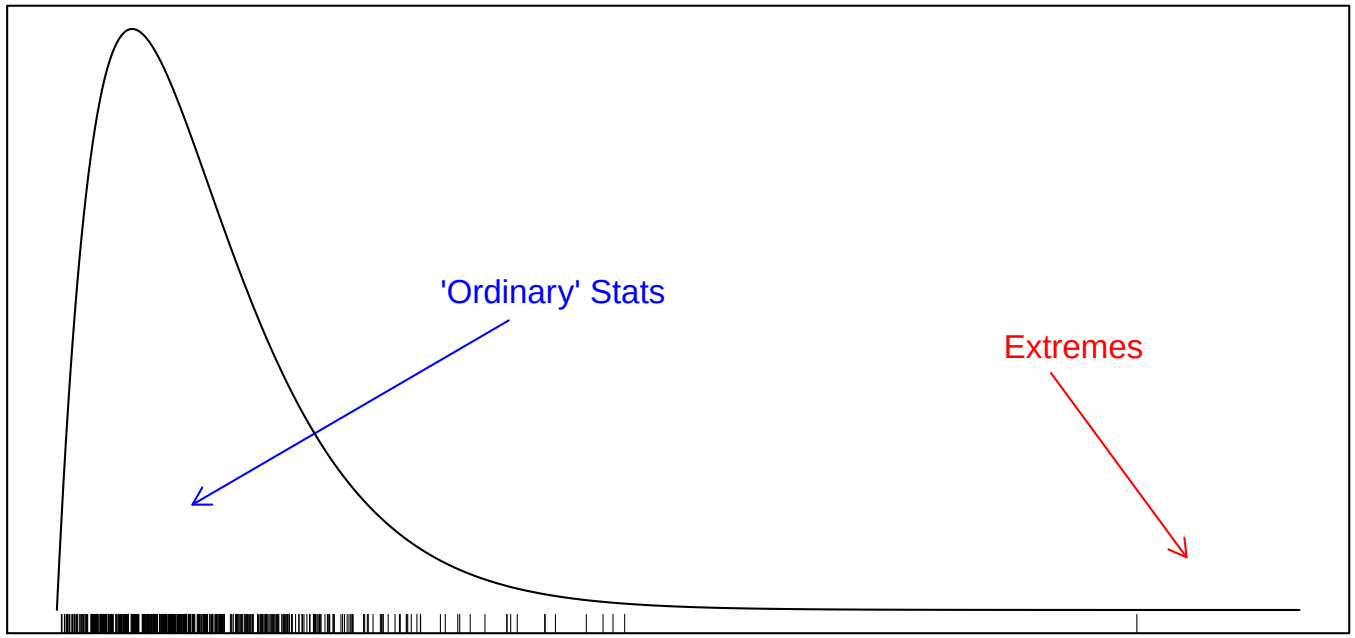
- $m(\boldsymbol{s})$ is the mean function, e.g., $\boldsymbol{X}(\boldsymbol{s})^T\boldsymbol{\beta}$
- $\eta(\boldsymbol{s})$ is a GP with mean zero and covaraince function $c_{\boldsymbol{\theta}}(h)$







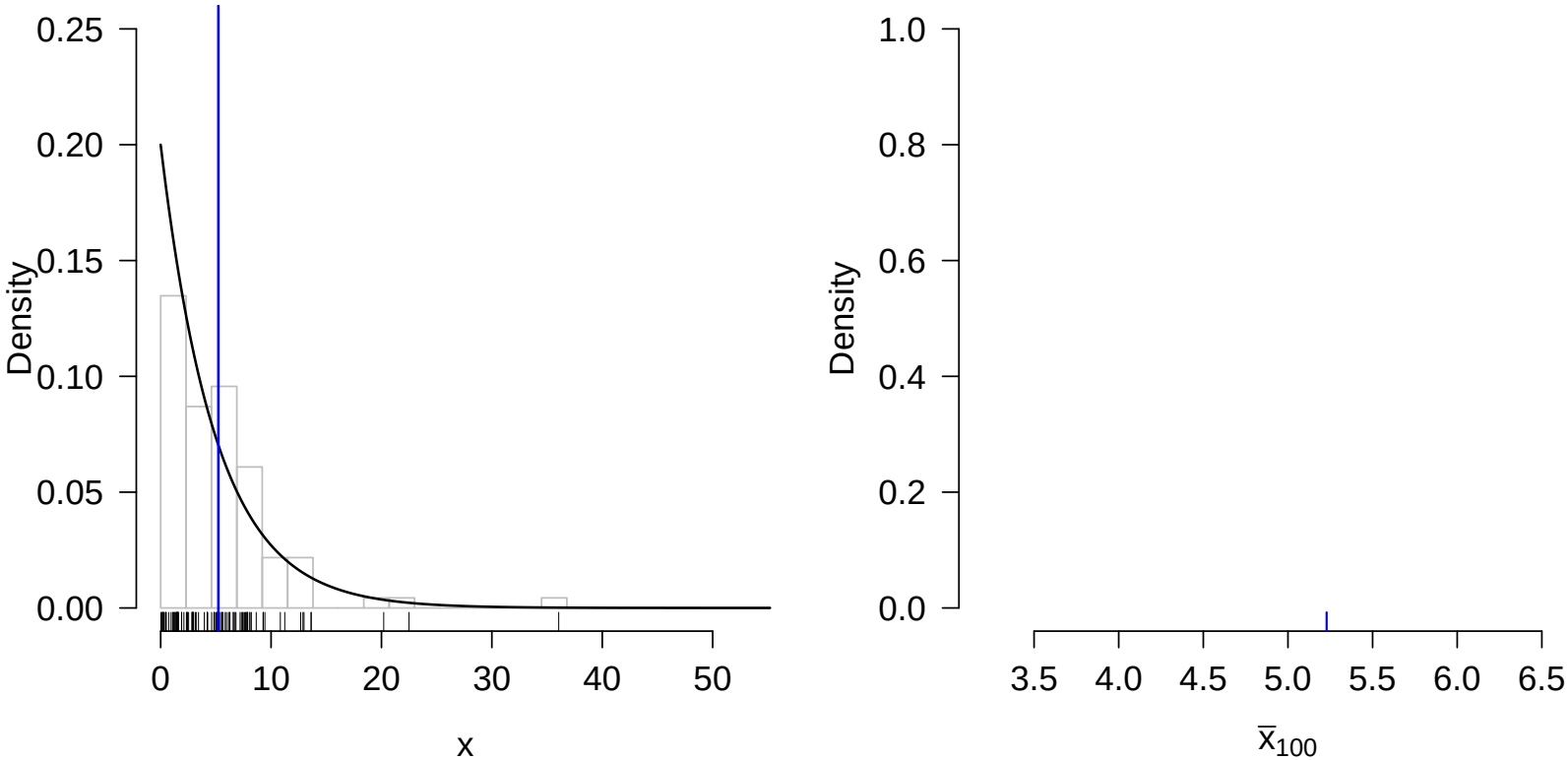
Extreme Value Analysis



	Target	Theory	Distribution
Ordinary Stats	bulk distribution	CLT	Normal
Extreme Stats	tail distribution(s)	?	?

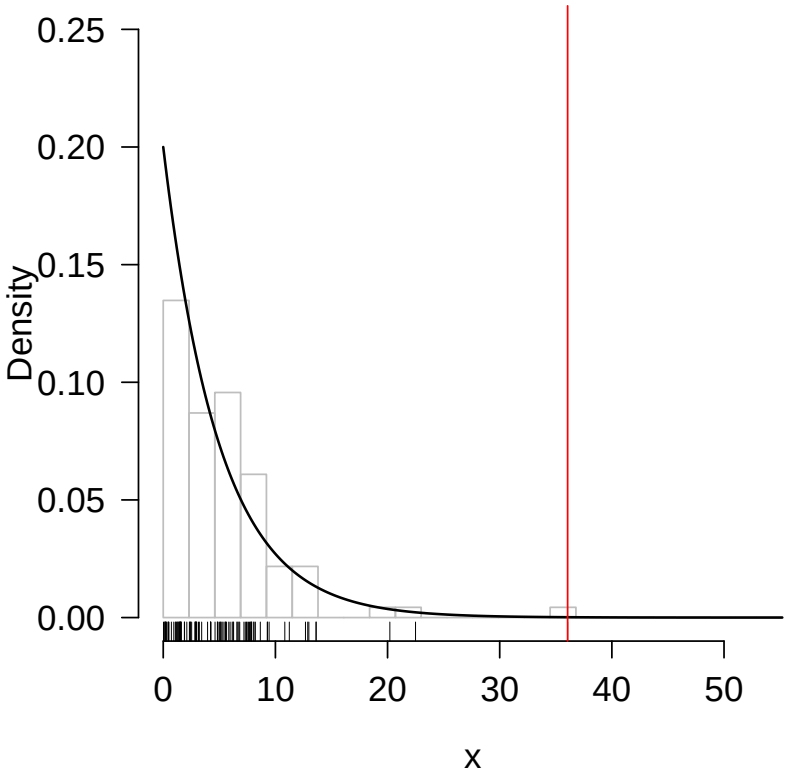
Central Limit Theorem Demonstration

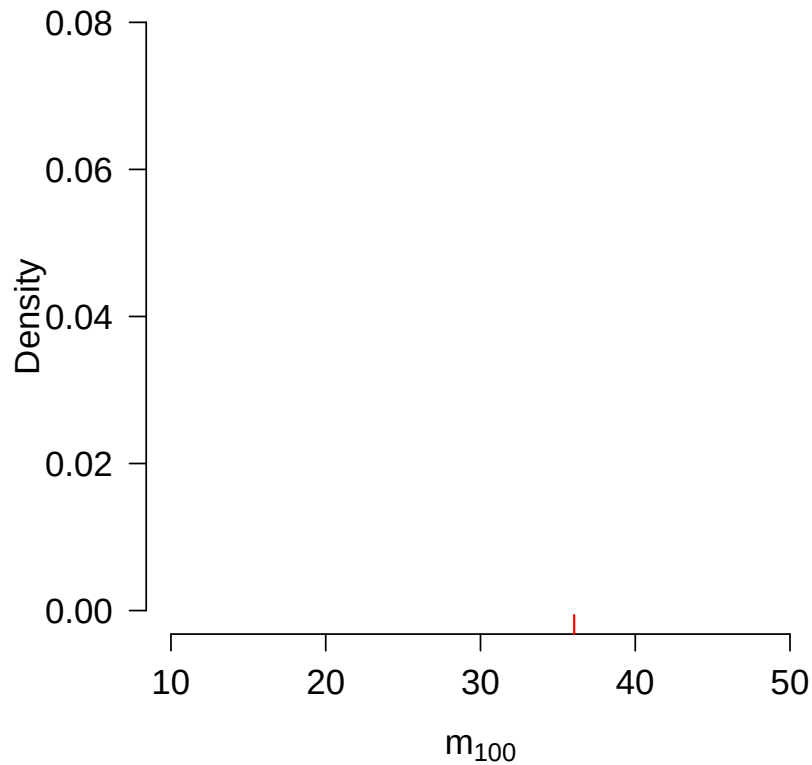
1. Generate 100 random numbers $(n = 100)$ from an Exponential distribution
2. Compute the sample mean of these 100 random numbers
3. Repeat this process 120 times

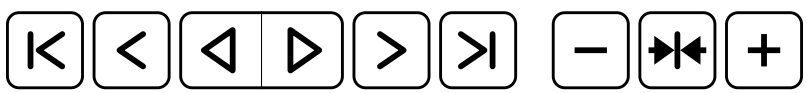


Demo: Distribution of the Sample Maximum

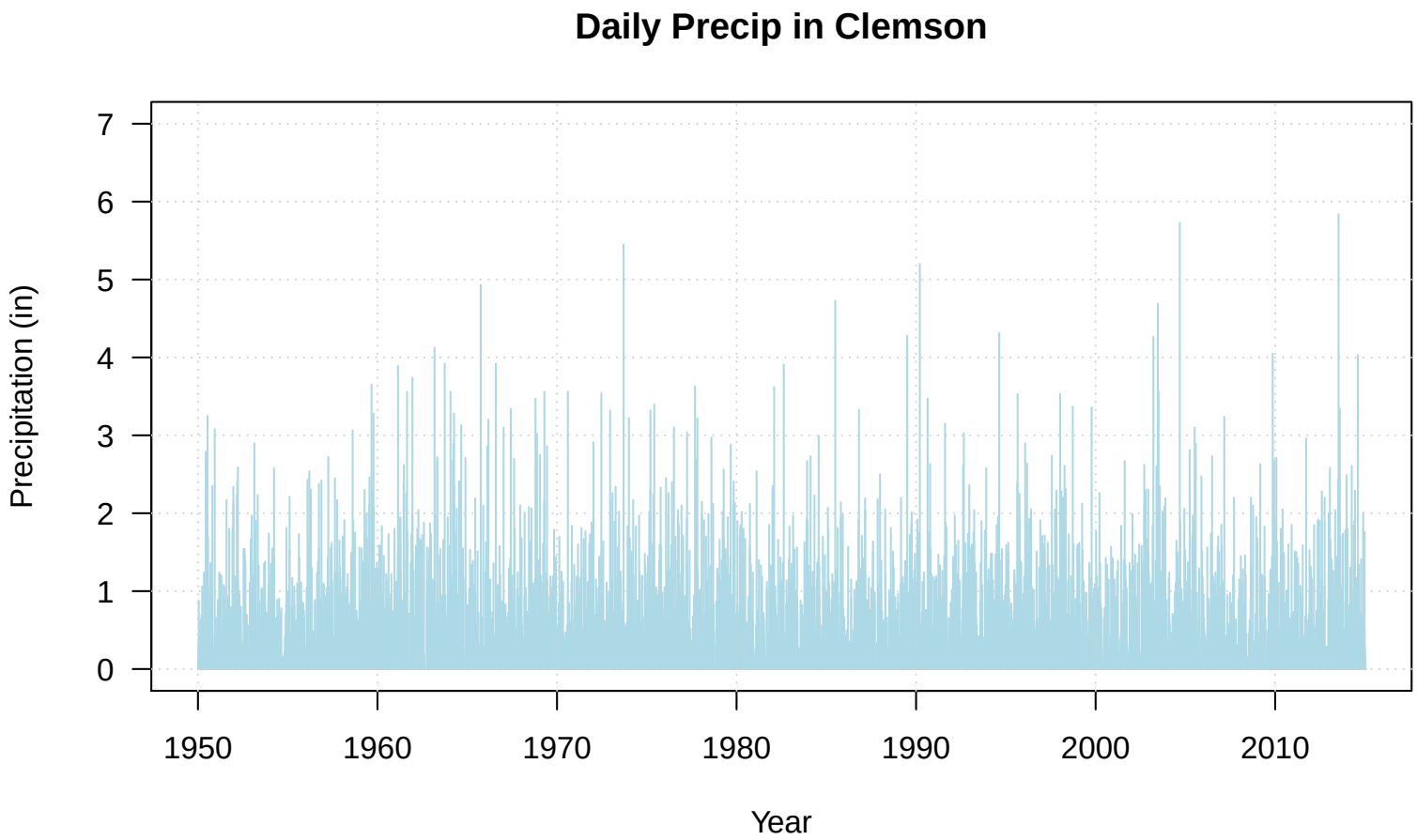
1. Generate 100 random numbers $(n = 100)$ from an Exponential distribution
2. Compute the sample maximum of these 100 random numbers
3. Repeat this process 120 times







Estimating Clemson[1] Daily Precipitation[2] Extremes



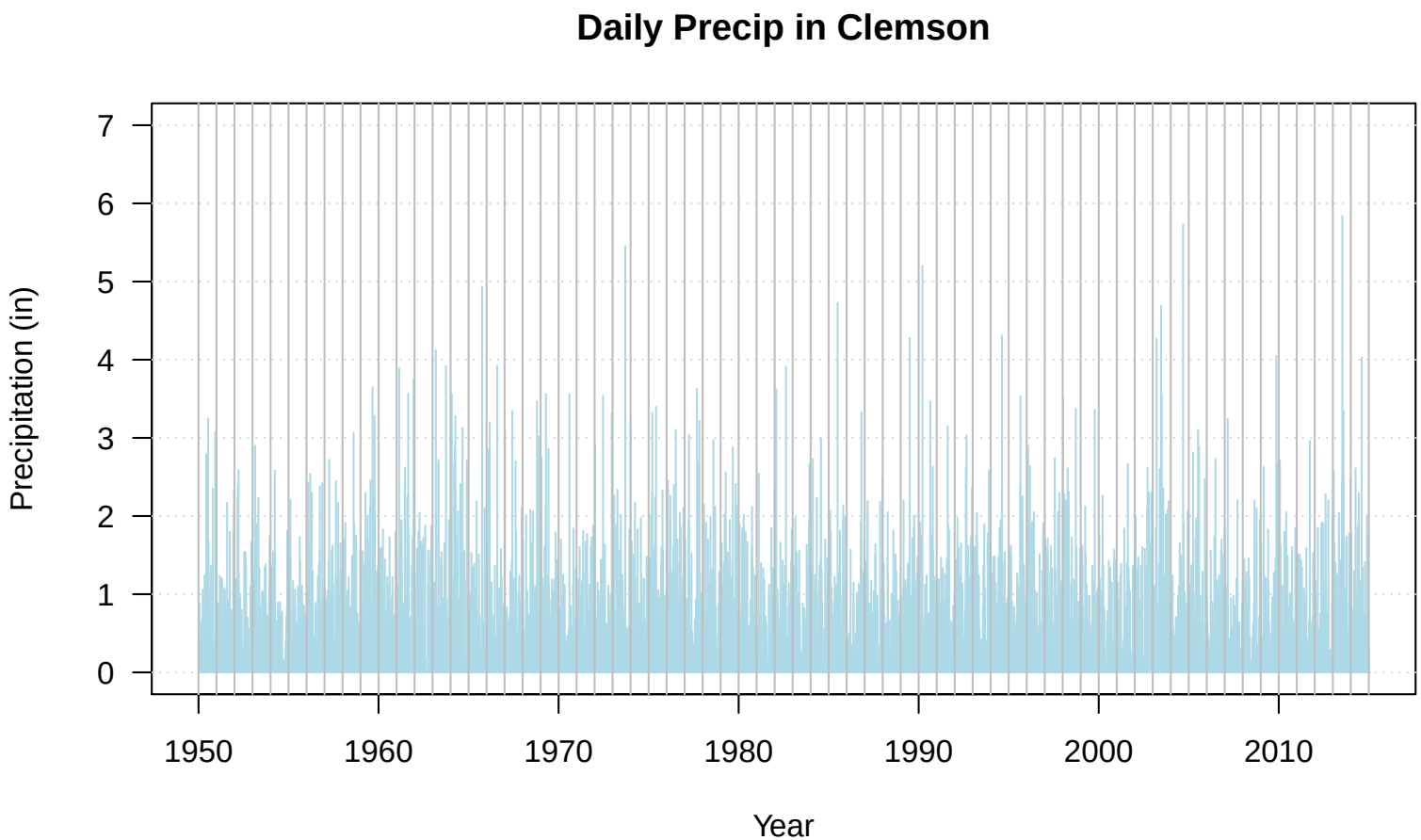
We may want to estimate, say the 50-year return level

[1] Near LaMaster dairy center

[2] Data Source: U.S. Historical Climatology Network (USHCN)

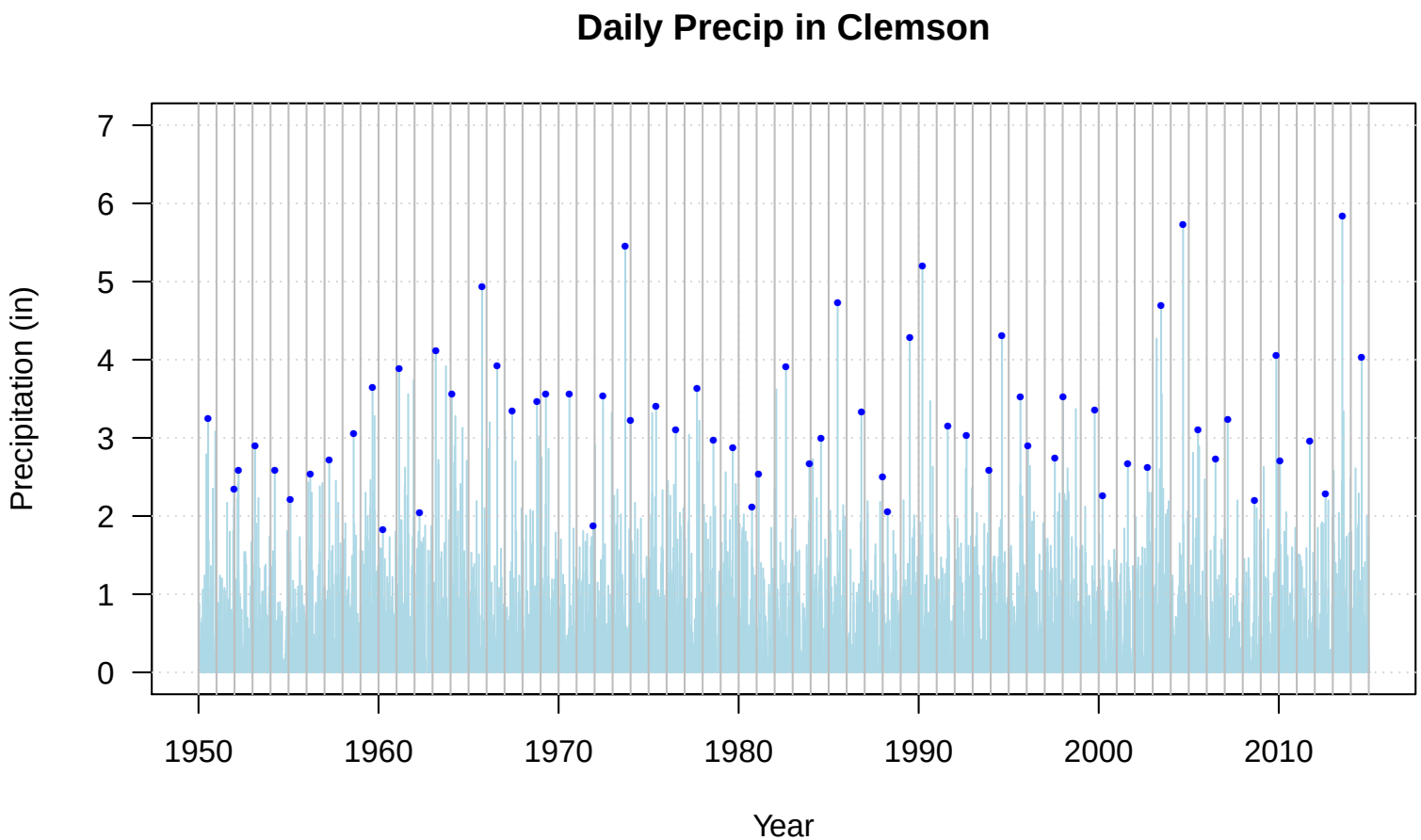
Block Maxima Method (Gumbel 1958)

1. Determine the block size and extract the block maxima



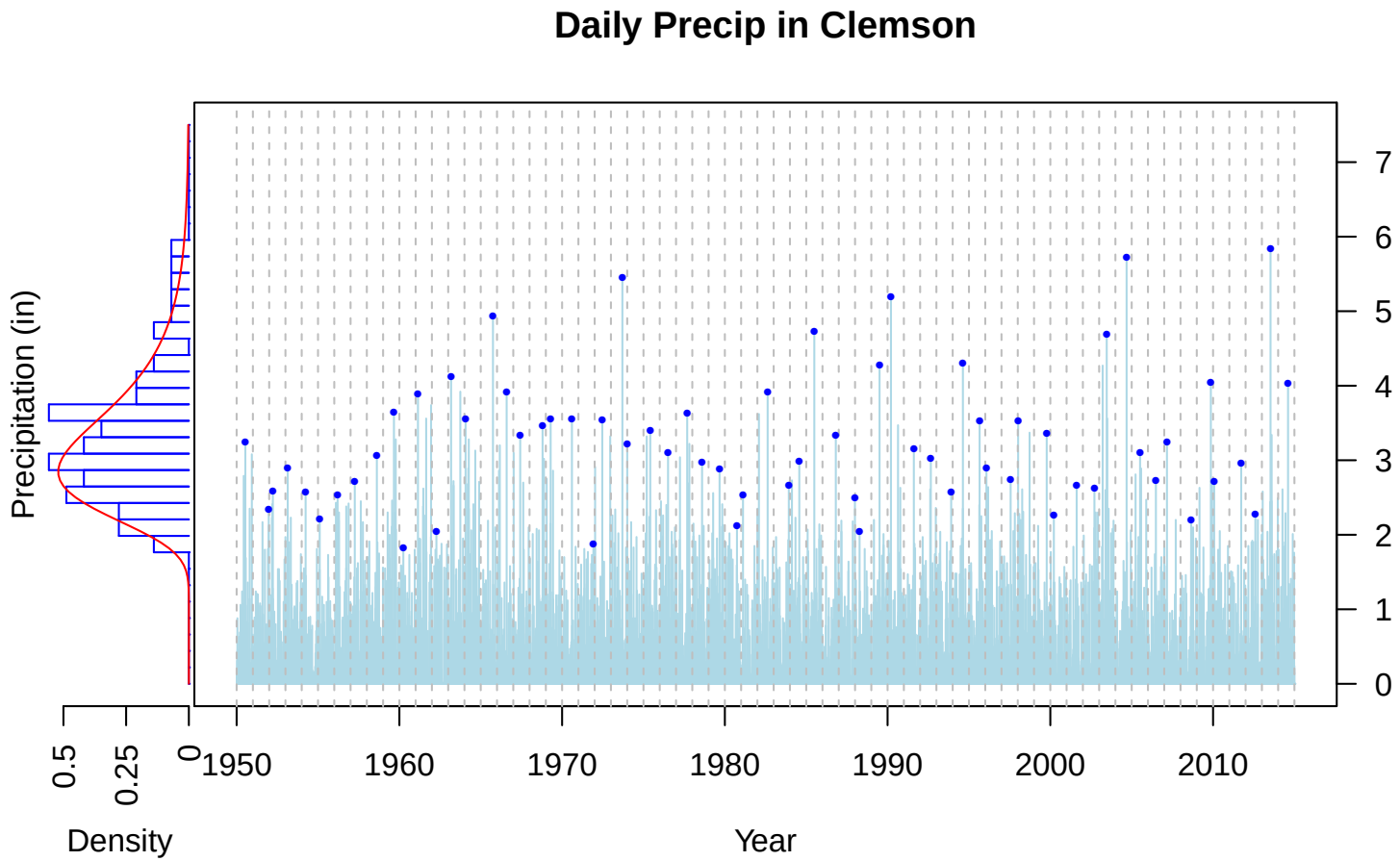
Block Maxima Method (Gumbel 1958)

1. Determine the block size and extract the block maxima



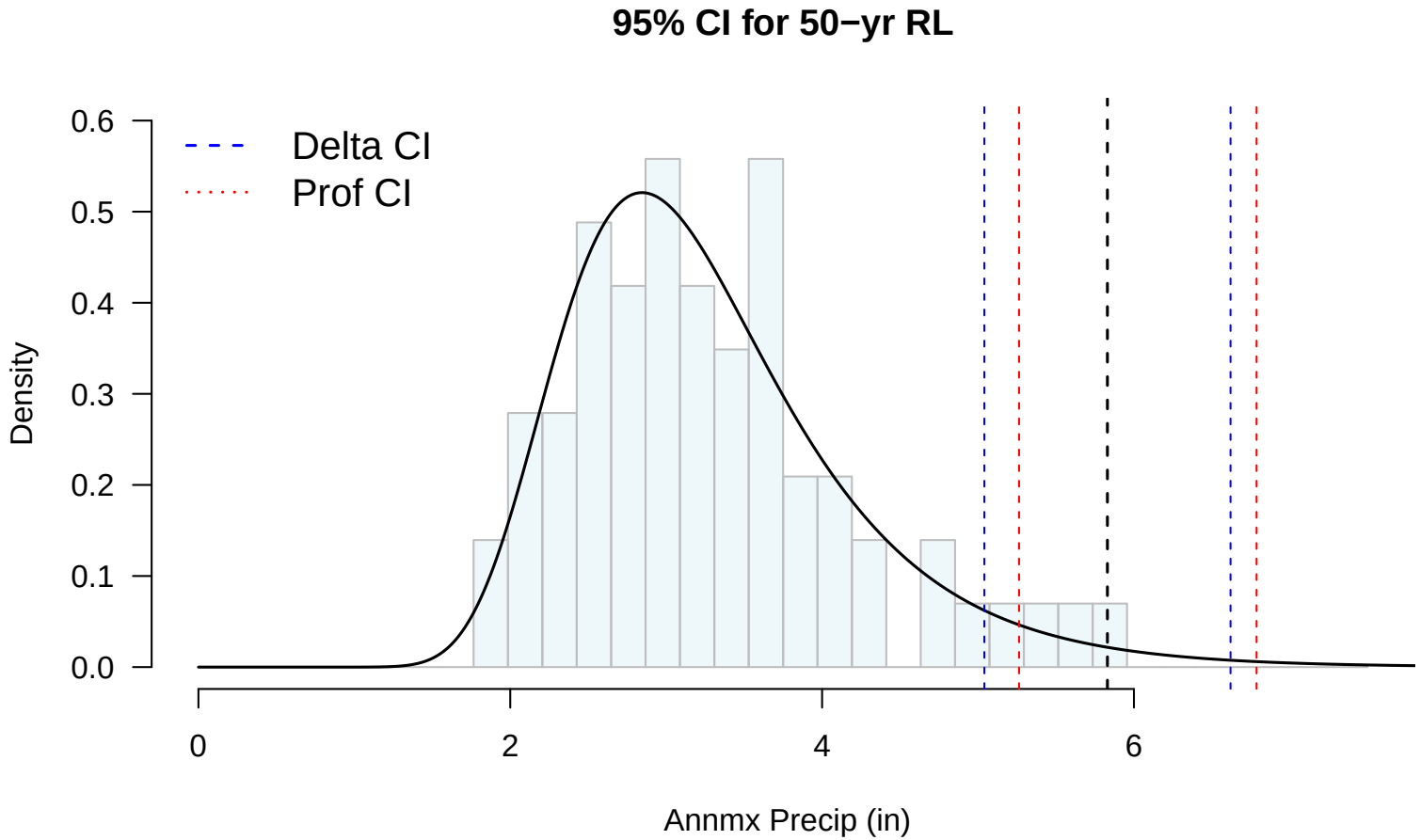
Block Maxima Method (Gumbel 1958)

2. Fit the Generalized extreme value (GEV) distribution to the maxima and assess the fit



Block Maxima Method (Gumbel 1958)

3. Perform inference for return levels or exceedance probabilitie



Point estimate: 5.68.

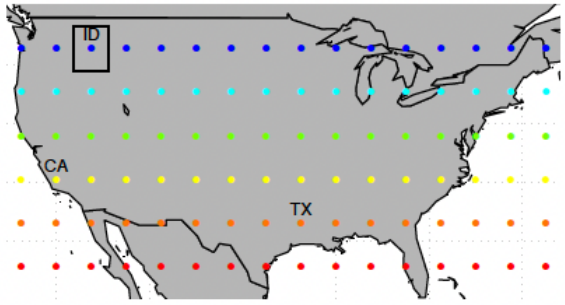
Interval estimate: $[4.63, 6.72]$ and $[4.98, 7.47]$

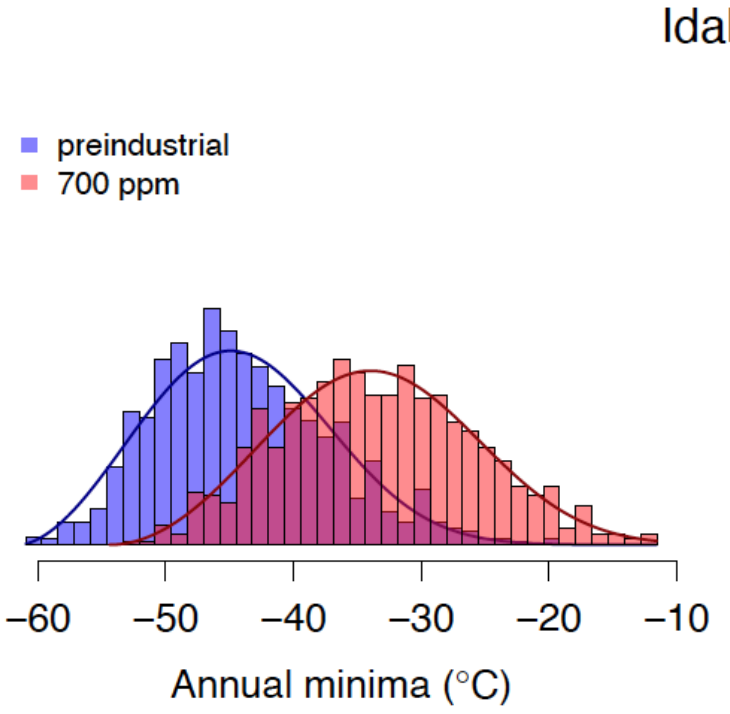
Some Research Questions

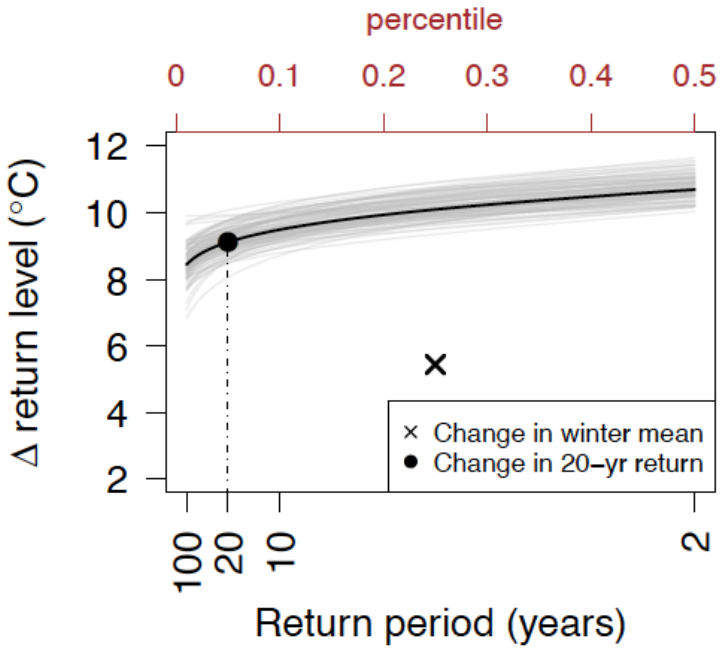
- How extremes vary in space? How extremes may change in future climate conditions?
- How to model extremes when the process of interest involves several variables (i.e., compound extremes)?
- How to combine data from different sources to infer extremes?
- How to leverage physical knowledge to better model extremes?

Estimating Changes in Temperature Extremes

Data \ Climate State	Present	Future
Model output	✓	✓
"Observations"	✓	?







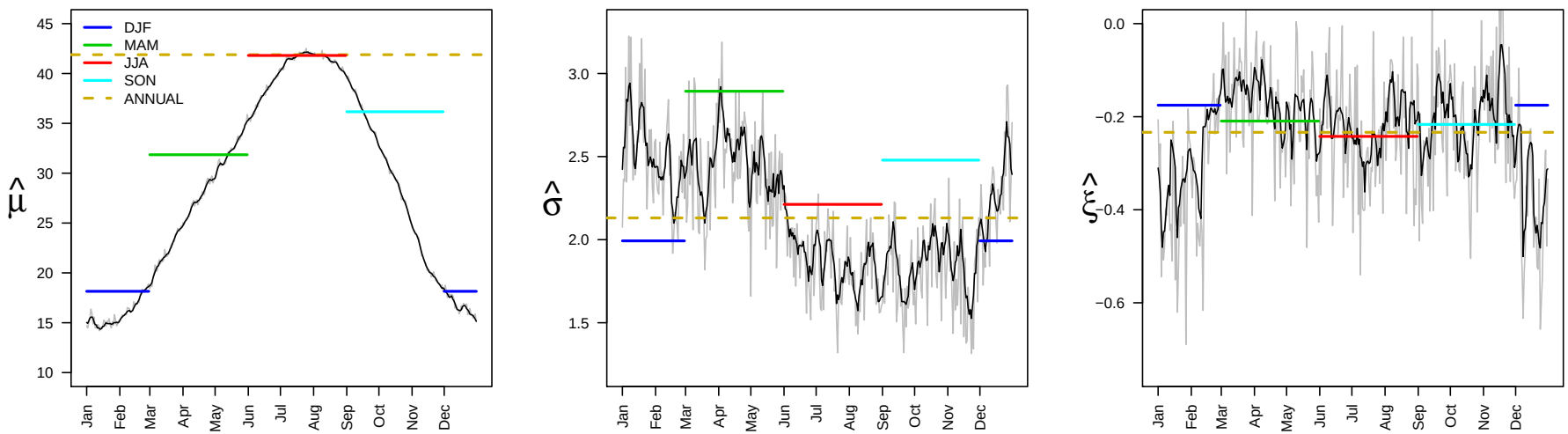
Modeling Seasonality in Extremes

With 1000 years of daily output $T_{i,j},\ i = 1, \cdots, 365,\ j = 1, \cdots, 1000$, we can explicitly incorporate seasonality by modeling extremes for each day:

$$M_{i,k} = \max_{j=(k-1)\times b+1}^{k\times b} T_{i,j} \sim \text{GEV}(\mu_i, \sigma_i, \xi_i)$$

where b is the block size

Example: "IL" pixel, T_{max}, $b = 25$



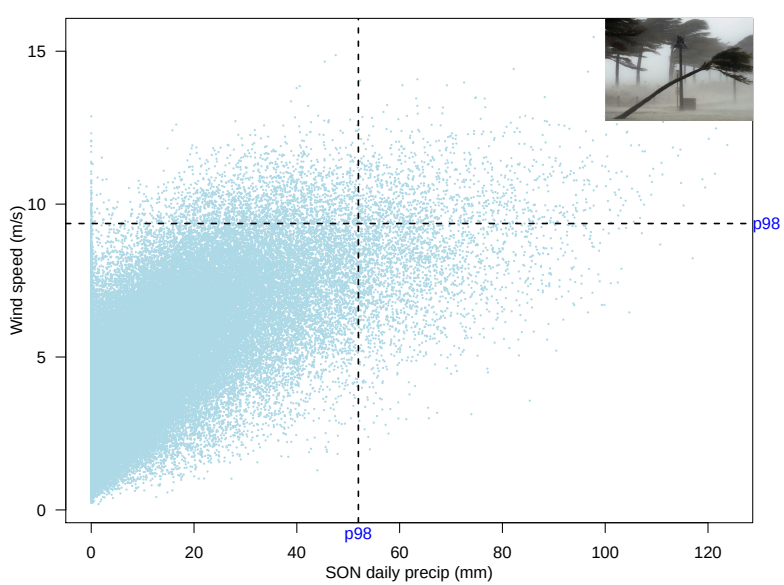
CanRCM4 Large Initial Condition Ensemble

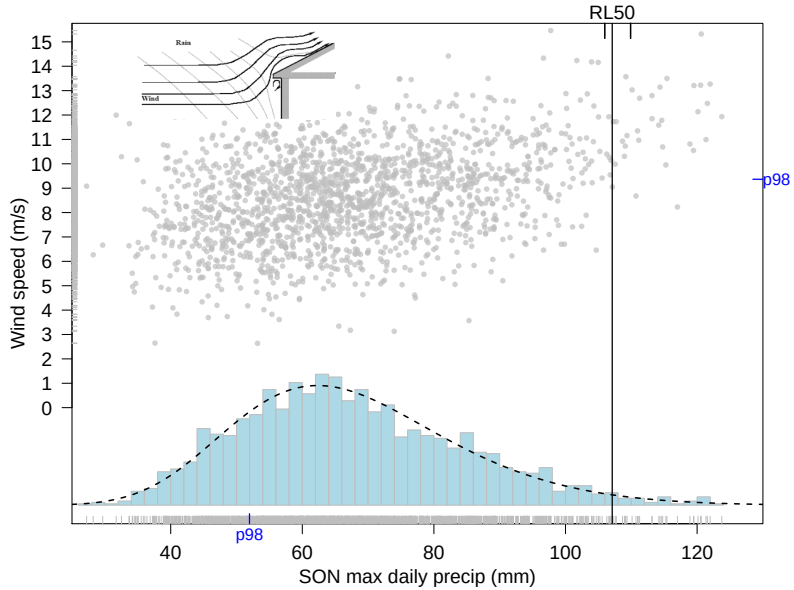
- 35-member initial-condition ensemble
- Using output from 1950-1999 with CMIP5 historical forcings
- North American region, 0.44° horizontal grid ($\sim$ 50 km). We will show the results from a "Vancouver" (NW) grid cell

Each run in ensemble produces (nearly) statistically independent realizations of climate system, which allows us to:

- provide more accurate estimates in climate extremes
- assess how well statistical procedures work

Concurrent Wind and Precipitation Extremes



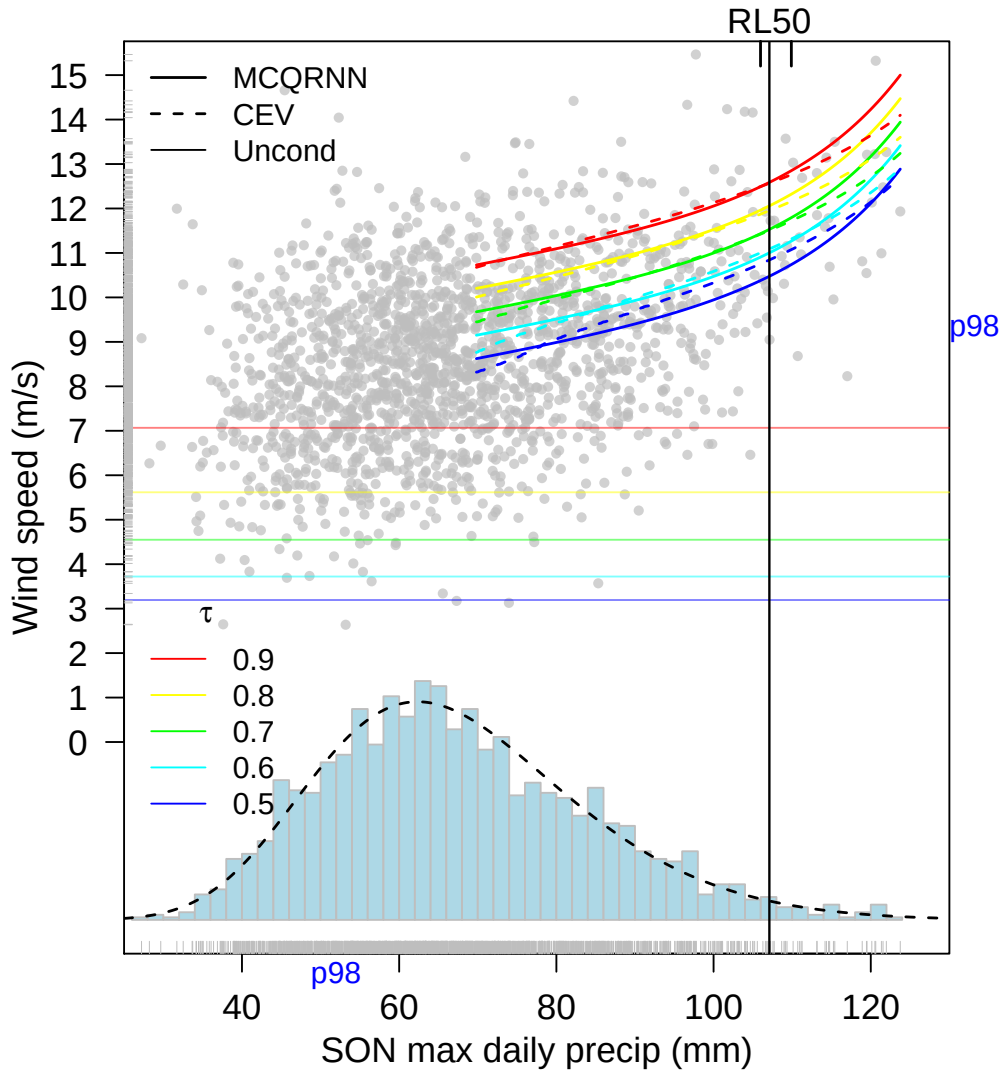


- Most (climate) literature focus on *estimating the occurrence probability of an concurrent extreme event*
- Here we would like to estimate the "**tail distribution**" via a conditional approach

Estimating Conditional Quantiles

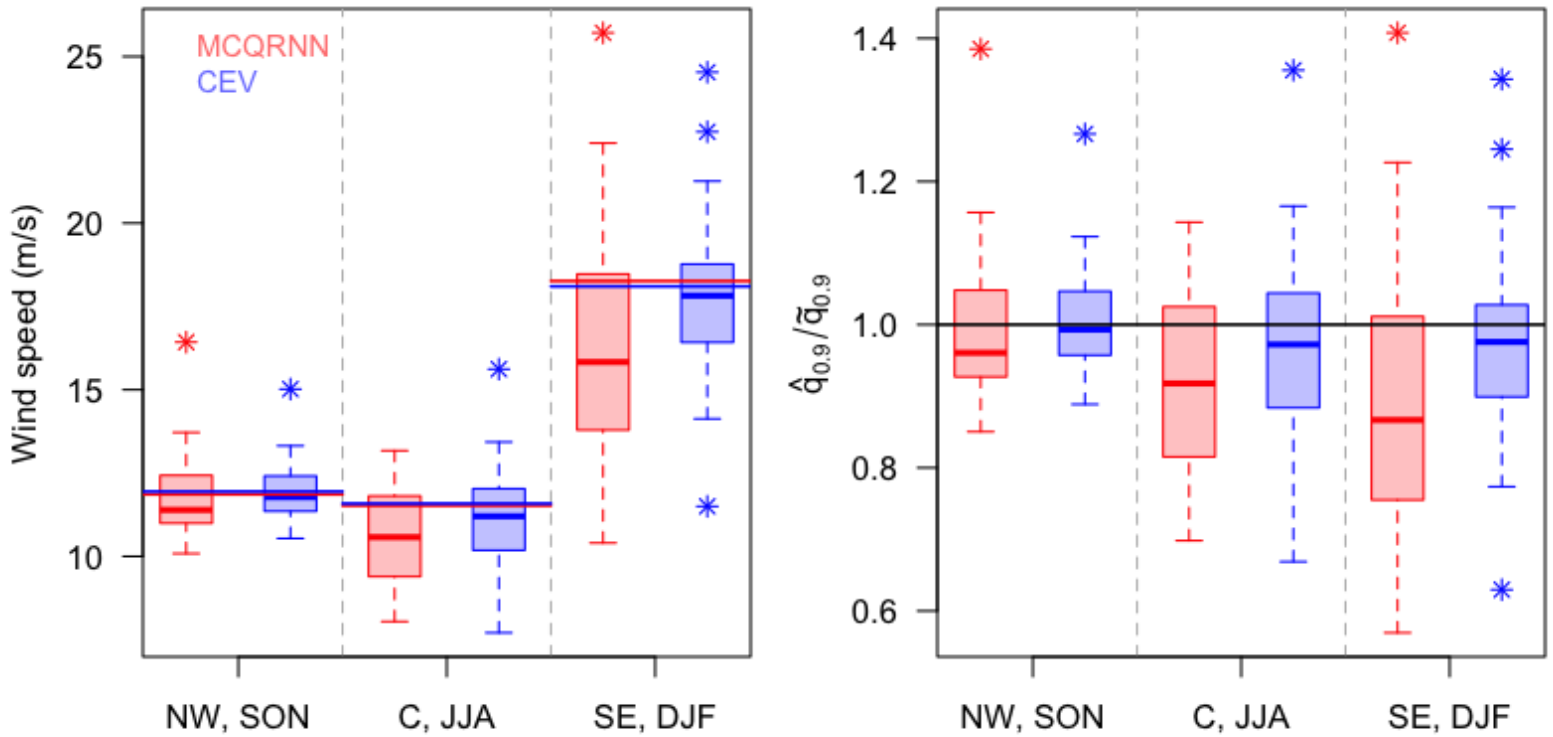
To model $[Y|X \text{ "large"}]$, we explore two methods:

- Monotone Composite Quantile Regression Neural Network (MCQRNN) [Cannon, 2018]
- Conditional extreme value (CEV) models [Heffernan & Tawn, 04]



Assessing Estimation Performance Via Large Ensemble

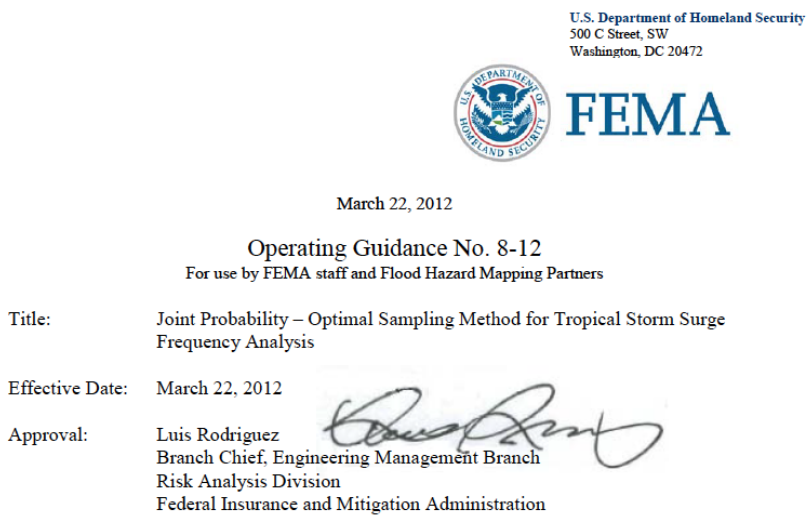
- Treat the fitted conditional quantile function at $\tau = 0.9$ using all 35 ensemble members as the "truth"
- Assess the model performance by fitting CEV and MCQRNN for each individual ensemble member



Quantifying Storm Surge Risk

- Flood insurance: The Federal Emergency Management Agency (FEMA) requires estimates of the magnitude of surges in terms of 10-, 50-, 100-, and 500-year return levels for coastal areas to determine insurance rates, infrastructure design, and emergency planning

- Flood mitigation: To assess whether coastal nuclear plants meet 1 in 10,000 year flood protection criteria.



U.S. Department of Homeland Security
500 C Street, SW
Washington, DC 20472

FEMA

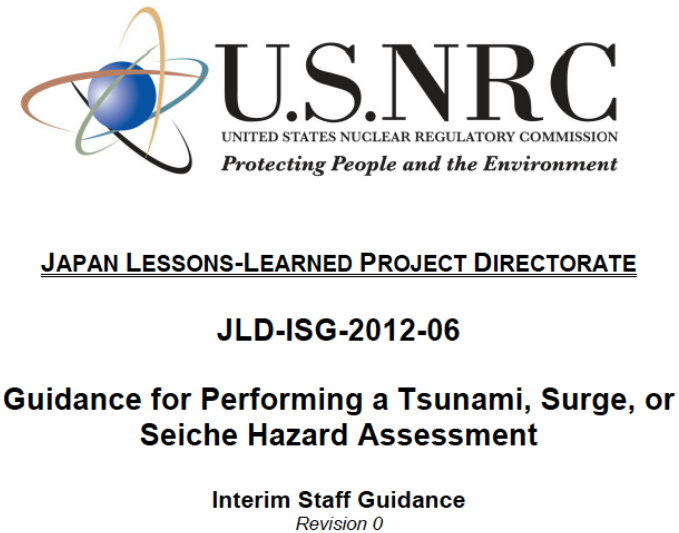
March 22, 2012

Operating Guidance No. 8-12
For use by FEMA staff and Flood Hazard Mapping Partners

Title: Joint Probability – Optimal Sampling Method for Tropical Storm Surge Frequency Analysis

Effective Date: March 22, 2012

Approval: Luis Rodriguez
Branch Chief, Engineering Management Branch
Risk Analysis Division
Federal Insurance and Mitigation Administration



U.S.NRC
UNITED STATES NUCLEAR REGULATORY COMMISSION
Protecting People and the Environment

JAPAN LESSONS-LEARNED PROJECT DIRECTORATE

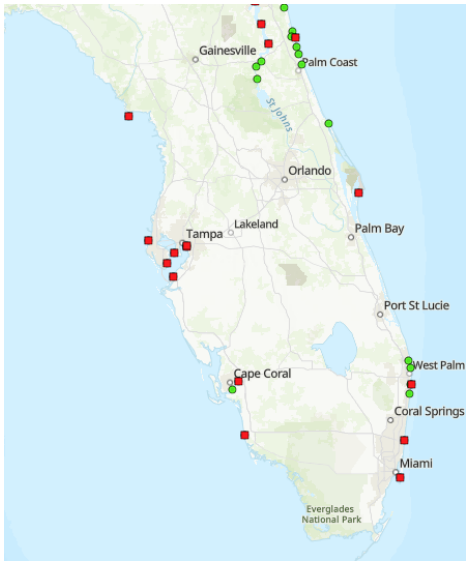
JLD-ISG-2012-06

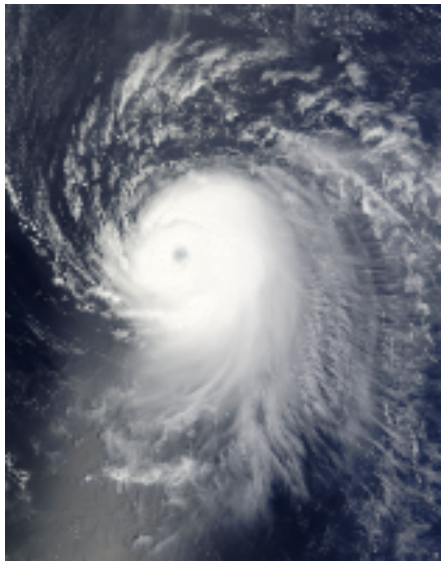
Guidance for Performing a Tsunami, Surge, or Seiche Hazard Assessment

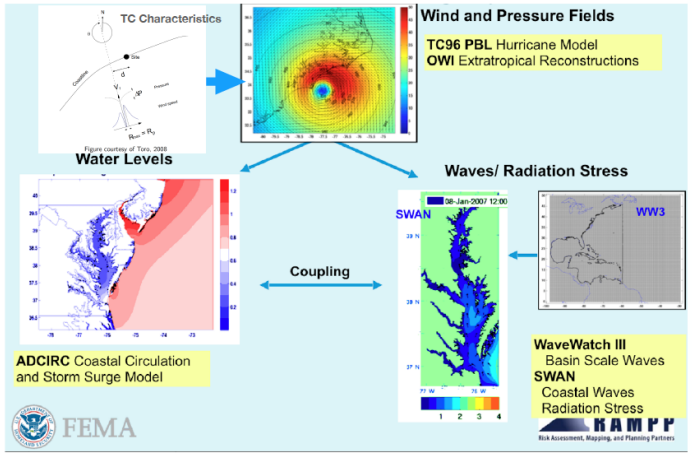
Interim Staff Guidance
Revision 0

Quantifying Storm Surge Risk: Data Sources

Data \ Variable	y ("Output")	$\boldsymbol{x}$ ("input")
Observation	Observed storm surges $\Rightarrow$ Very limited in space and time	Storm characteristics $\Rightarrow$ Limited but well observed
Simulation	Simulated storm surge responses	"Synthetic" storm characteristics

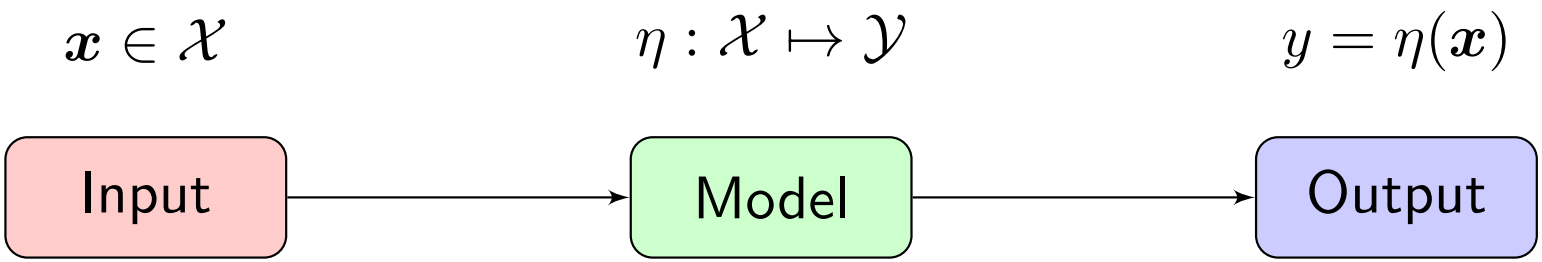






Courtesy of Gangai (Dewberry) & Danforth (FEMA)

Estimating Extreme Surges: Physical-Statistical Approach



TC Characteristics

- Records are more complete than surge levels
- Input to simulate storm surge levels

Task: Estimating $f(\boldsymbol{x})$
$\Rightarrow$ Density Estimation

Computer Model

- Simulate high fidelity surge response
- computationally extensive

Task: Estimating $\eta(\boldsymbol{x})$
$\Rightarrow$ GP ☺

Surge Level

- simulate synthetic storms
- generate surge response for risk analysis

Task: Estimating y_r
$\Rightarrow$ EVA ☺

Estimating Input Distribution $f(x)$: Data

- IBTrACS contains global tropical cyclone best-track data from 1851 - 2018 [Knapp et al., 2010]
- Only included "strong" storms (those with pressure deficit $\geq$ 13mb) from 1950[3] that made landfall within 100 km of the coastline of the study region[4] $\Rightarrow$ 28 storms were included in out Southwest Florida case study

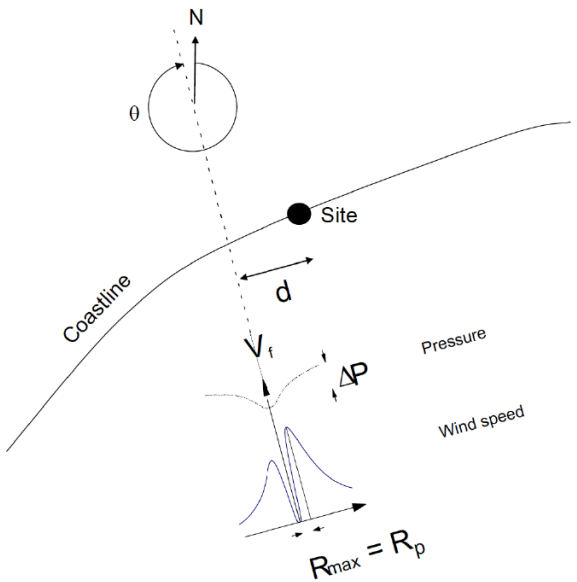
Pressure deficit Δp

Radius to max wind speed R_{max}

Forward speed V_f

Heading angle θ

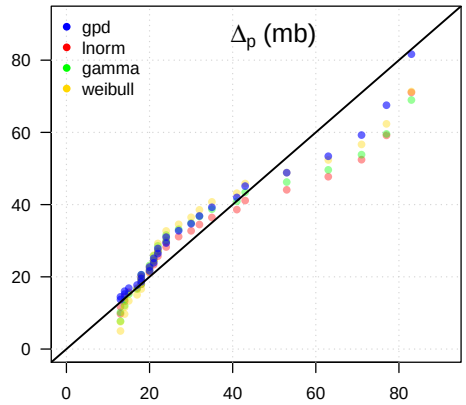
Distance from landfall d

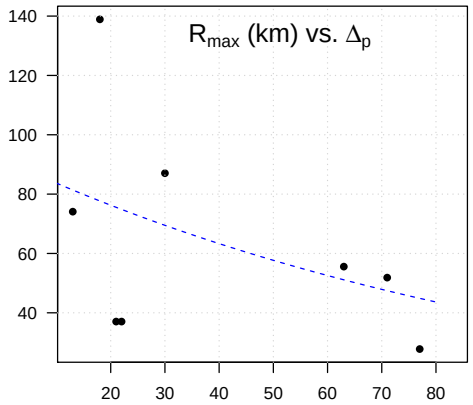


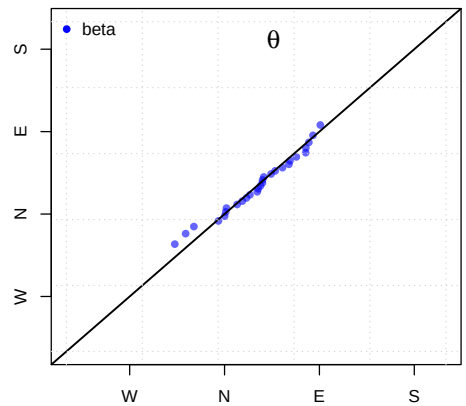
[3] due to data quality issue

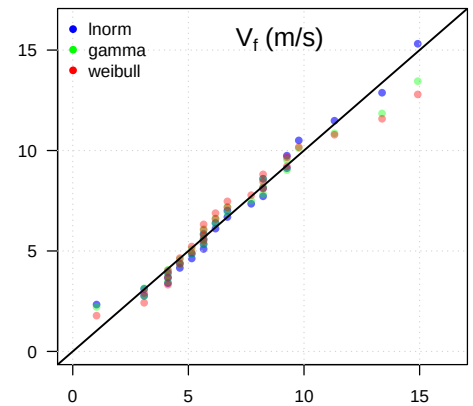
[4] surge levels generated by the excluded storms are negligible

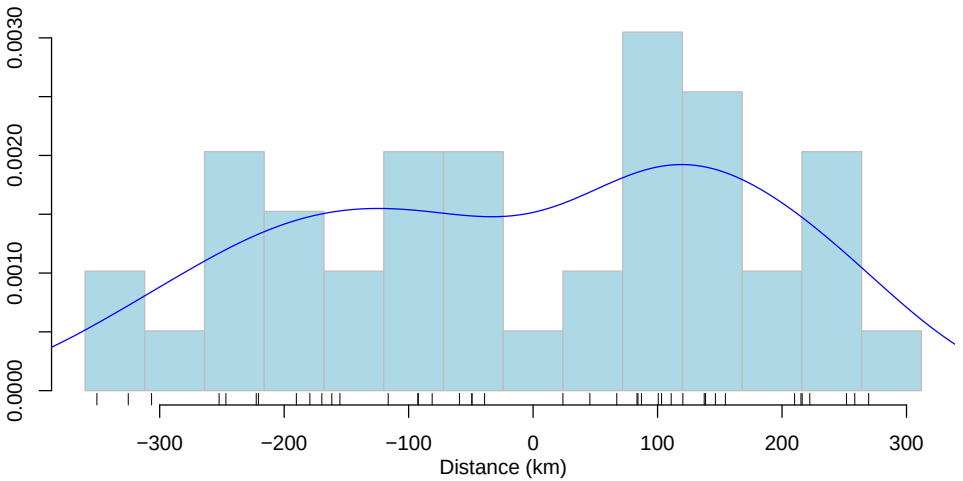
Estimating Input Distribution: Model & Results







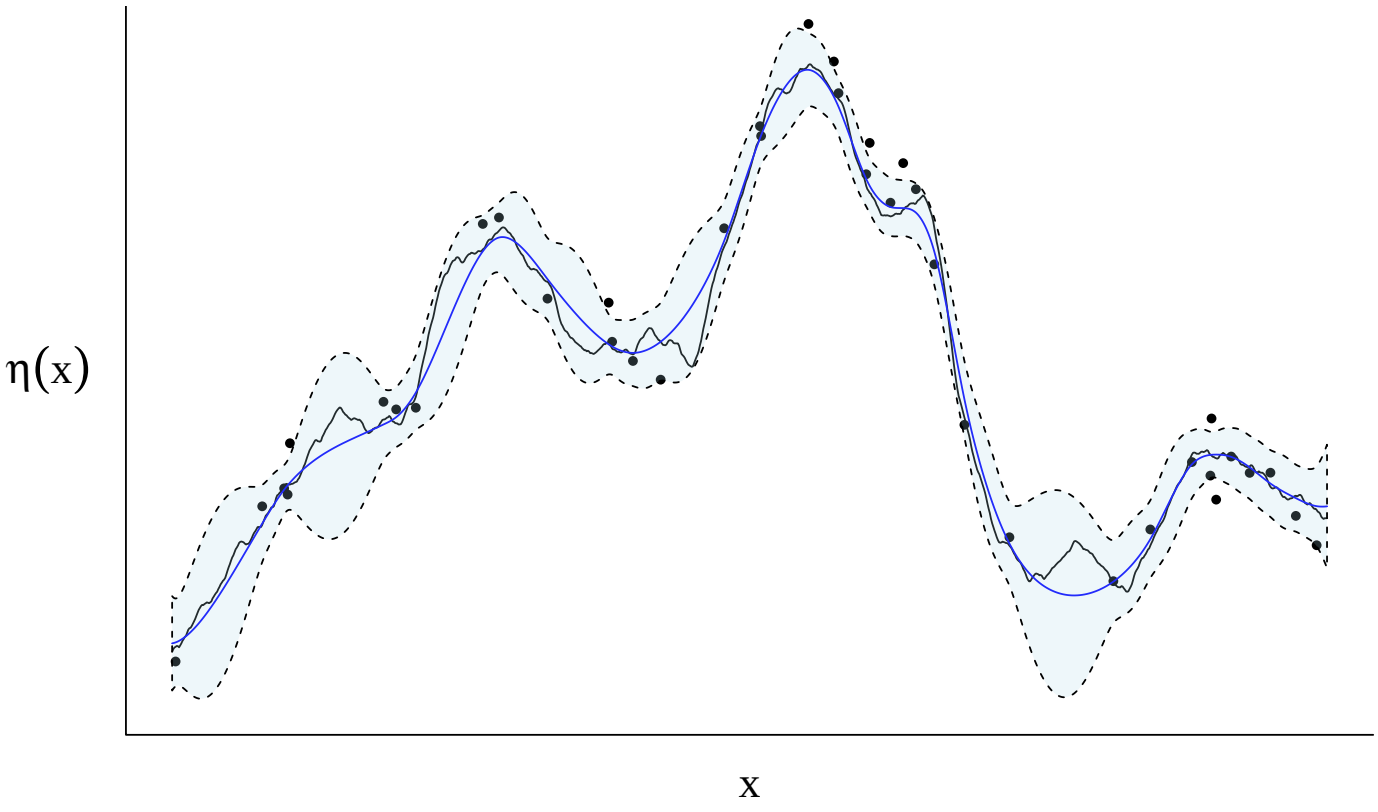




- Estimating $f(\boldsymbol{x}) = f(\Delta p, R_{max}, V_f, \theta, d)$ based on $\{\boldsymbol{x}_{o,i}\}_{i=1}^{28}$
  - Δp: generalized Pareto
  - $R_{max}|\Delta p$: log-normal
  - V_f: log-normal
  - θ: beta (shifted and re-scaled to $[0, 1]$)
- Distance d was not described well by a parametric distribution. We preform a kernel density estimation

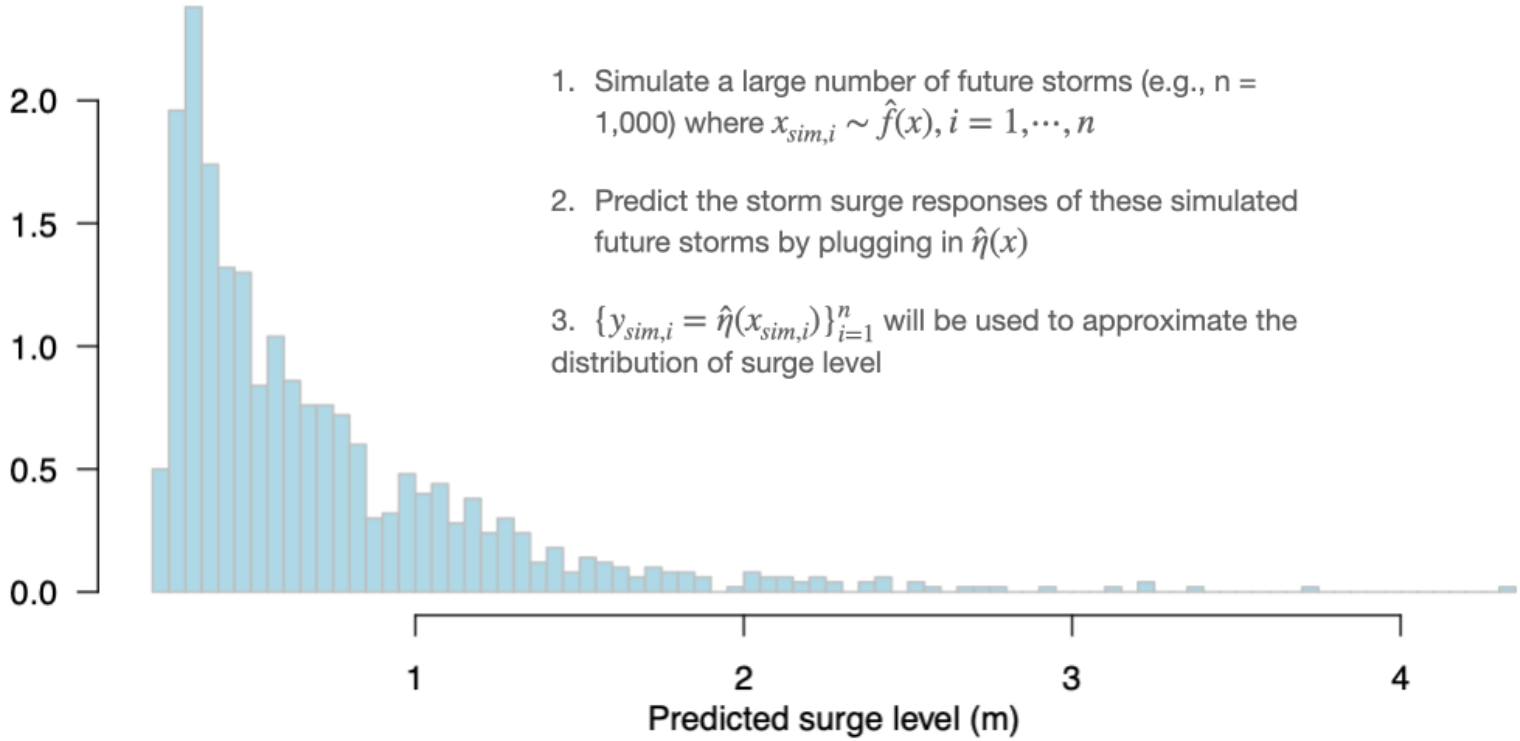
Estimating $\eta(\cdot)$: GP Emulator

$$y = \eta(\boldsymbol{x}) + \epsilon, \quad \eta(\cdot) \sim \mathsf{GP}(m(\cdot), K_\phi(\cdot,\cdot)), \quad \epsilon \sim N(0, \tau^2)$$



Generating Synthetic Storms and their Surge Levels

Combining $\hat{f}(\boldsymbol{x})$ and $\hat{\eta}(\boldsymbol{x})$ to generate more storm events

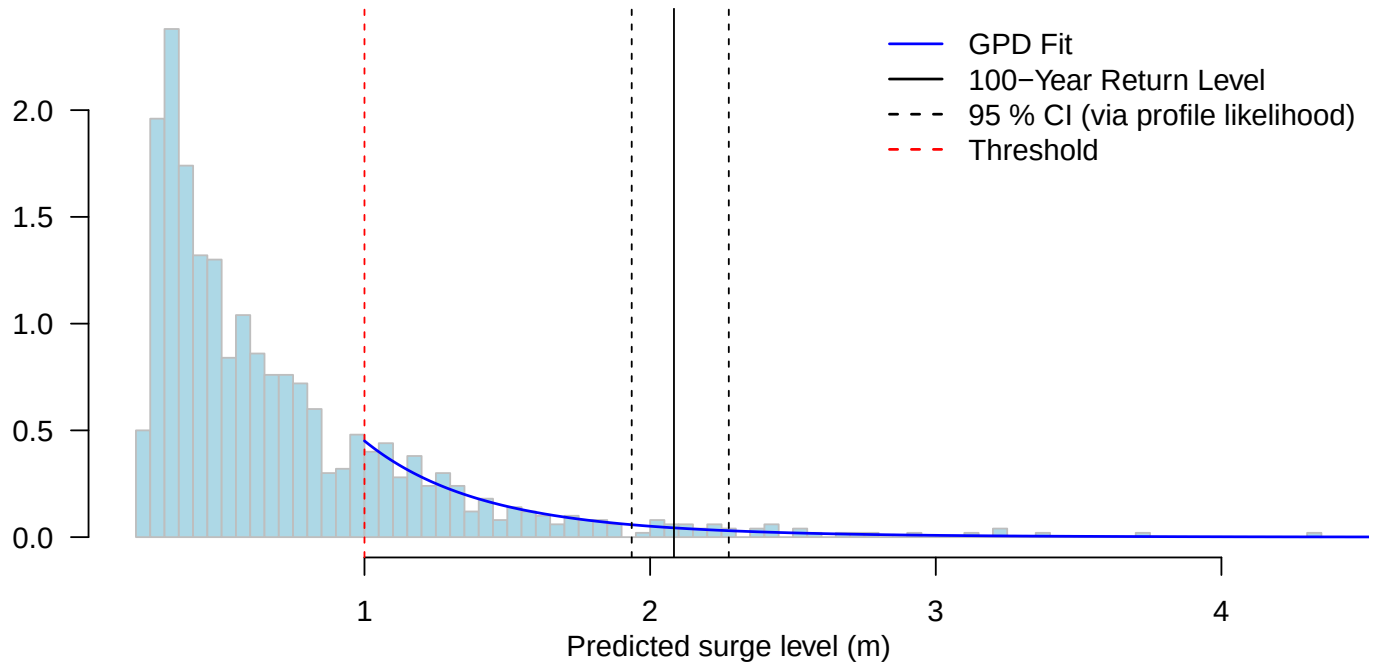


We are going to use these synthetic surge responses to estimate extreme surges

Estimating Extreme Surges: Extreme Value Analysis

We employed the peaks-over-threshold method [Davison and Smith, 1990] to estimate the r-year return levels

- Assuming upper tail follow a generalized Pareto distribution (GPD)
- Using profile likelihood method to construct confidence interval (CI), which gives asymmetric interval



Interdisciplinary Workshop on Weather/Climate Extremes

- **When/Where**: May 16-18, 2023 Clemson SC

- 2 Short courses ("Introduction to statistical methods for climate extremes" by Eric Gilleland and "The Toolkit for Extreme Climate Analysis" by Travis O'Brien), research talks, discussion sessions, poster session

- More information can be found on the website: https://whitneyhuang83.github.io/WCE2023