SMSS First Year Graduate Student Seminar

Whitney Huang

✉: wkhuang@clemson.edu

☜: https://whitneyhuang83.github.io/

School of
MATHEMATICAL AND
STATISTICAL SCIENCES
Clemson® University

November 12, 2024

Agenda

My Background

My Research

Some General Advice

About Me

- **Current Role**: Sixth-year Assistant Professor of Statistics. Primarily teach in the Data Science and Analytics (DSA) program, plus a Time Series course (MATH 8090 or 4070)

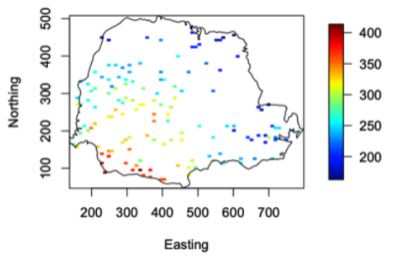
- **Background**: Born in Laramie, Wyoming, and raised in Taiwan

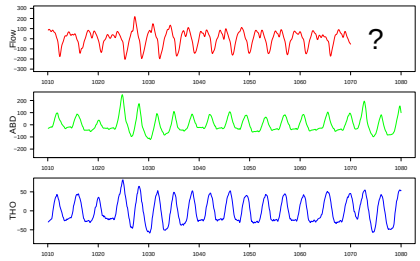
- **Academic Path**: Started with a B.S. in Mechanical Engineering and transitioned to Statistics for graduate studies

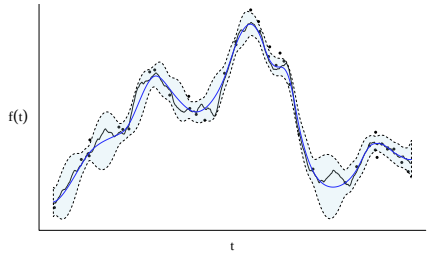
- **Doctorate**: Ph.D. in Statistics, 2017, from PURDUE UNIVERSITY. Completed a samsi NSF Duke NCSU UNC / University of Victoria / CANSSI INCASS postdoc before joining Clemson

Overview of My Research

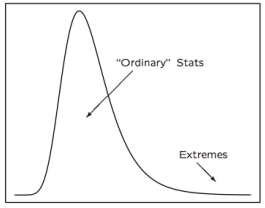
- ## Spatio-Temporal Statistics

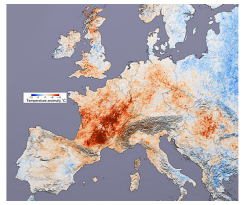


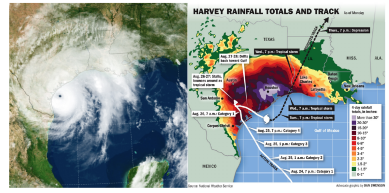




- ## Extreme Value Analysis





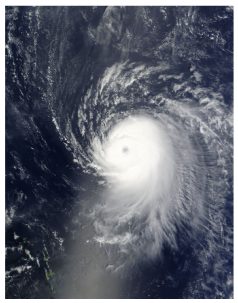


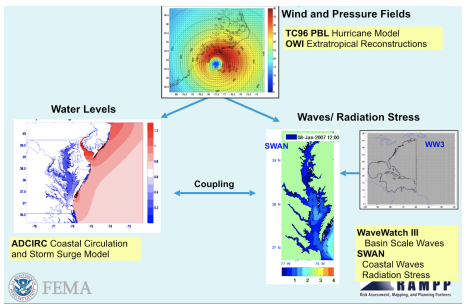
Source: NASA (Left); National Weather Service (Right)

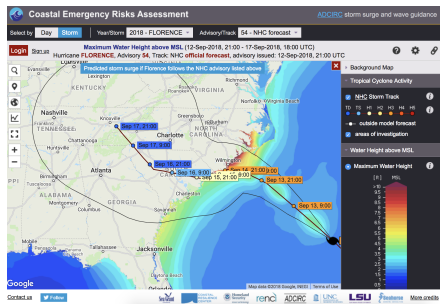
- ## Surrogate Modeling of Computer Experiments

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

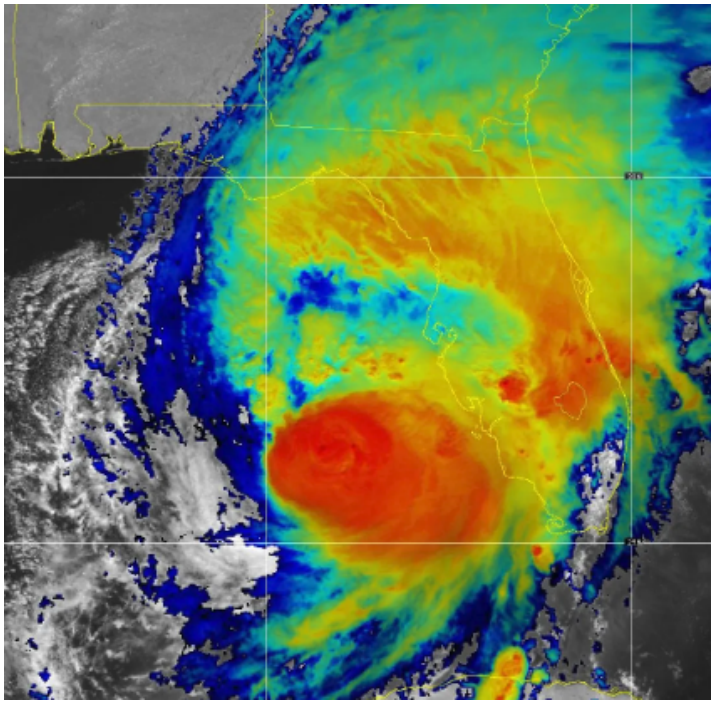
Input → Model → Output

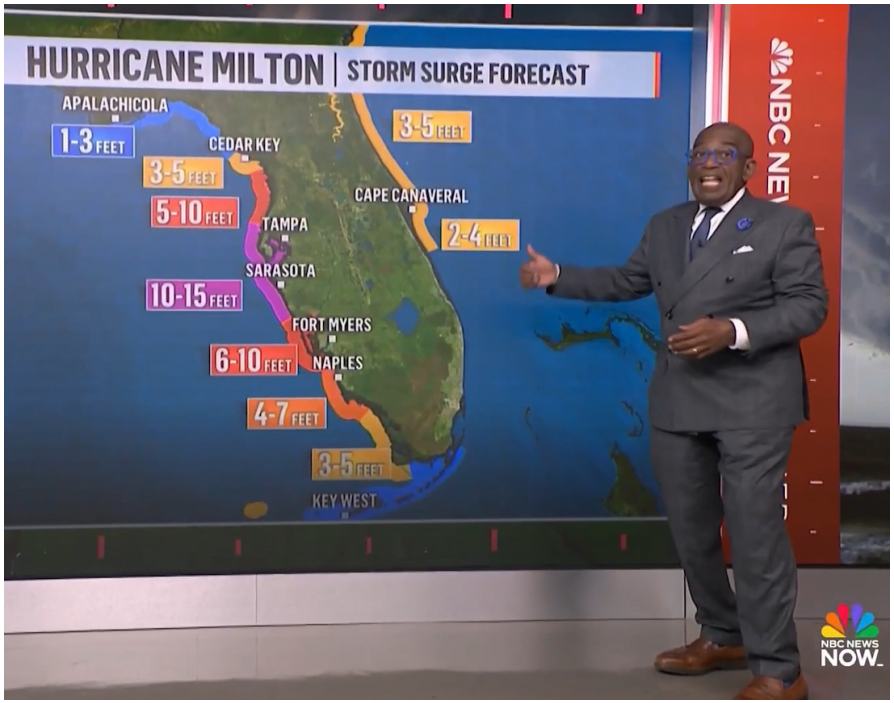






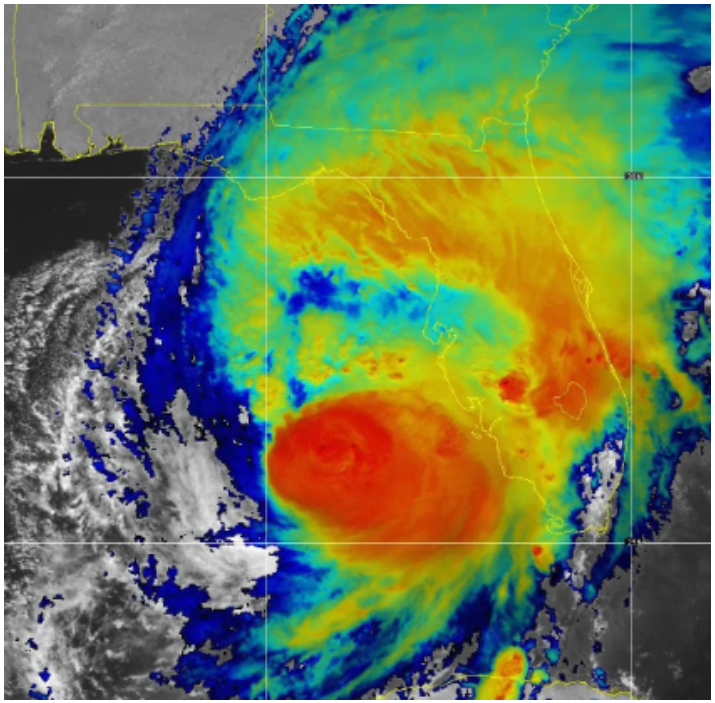
Hurricane and Storm Surge

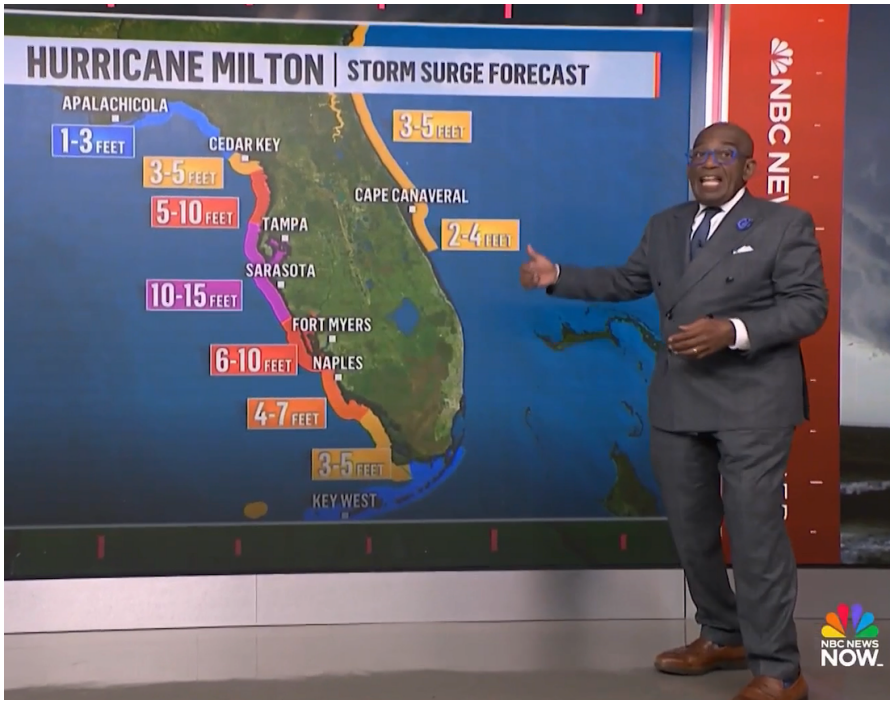




Storm surge is typically the most devastating part of a hurricane; therefore, it is crucial to accurately quantify storm surge risk

Hurricane and Storm Surge

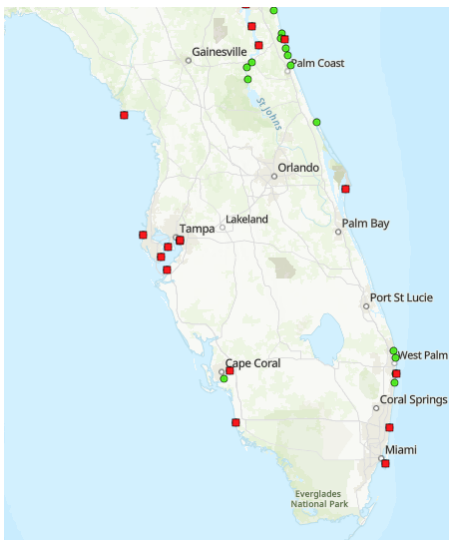


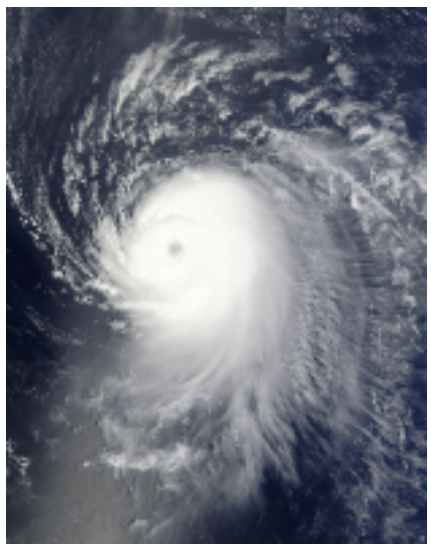


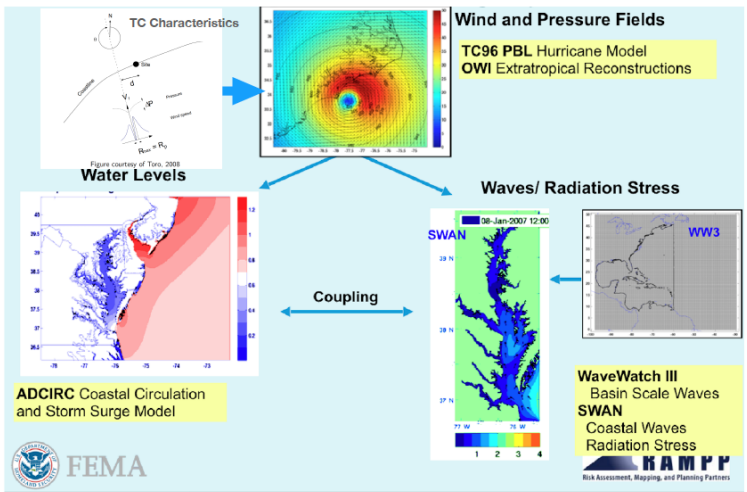
Storm surge is typically the most devastating part of a hurricane; therefore, it is crucial to accurately quantify storm surge risk

Quantifying Storm Surge Risk: Data Sources

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



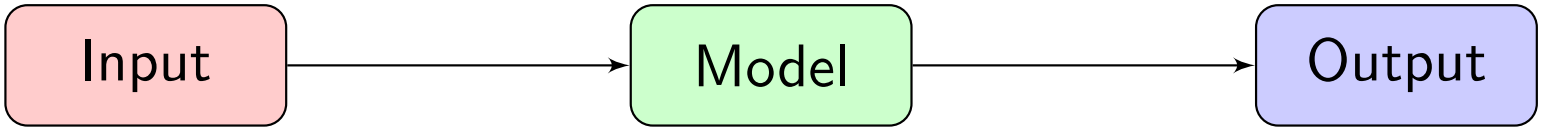




Courtesy of Gangai (Dewberry) & Danforth (FEMA)

Estimating Extreme Surges: Physical-Statistical Approach

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



TC Characteristics

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

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

Computer Model

- Simulate high fidelity surge response
- computationally extensive

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

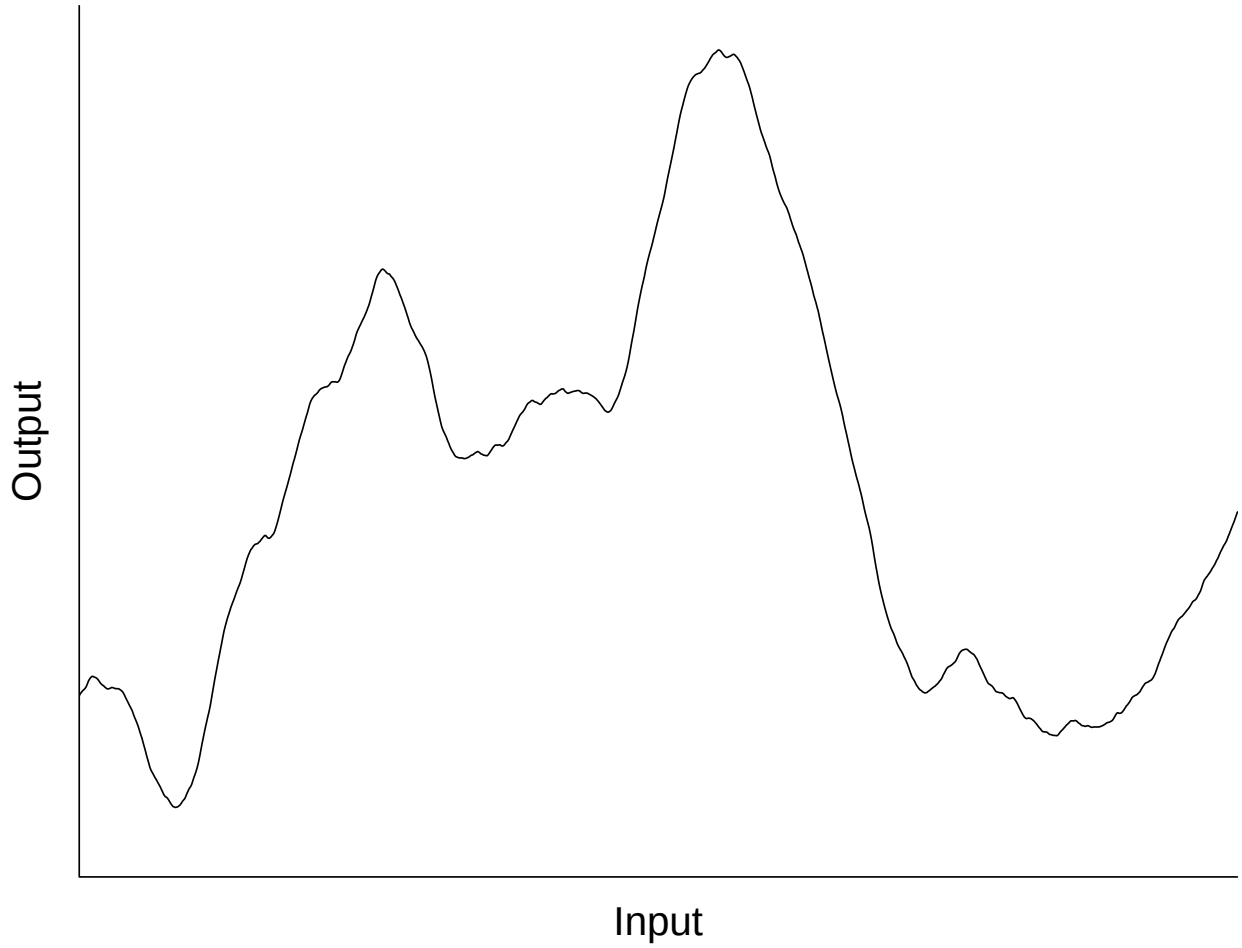
Surge Level

- simulate synthetic storms
- generate surge response for risk analysis

Task: Estimating y_r

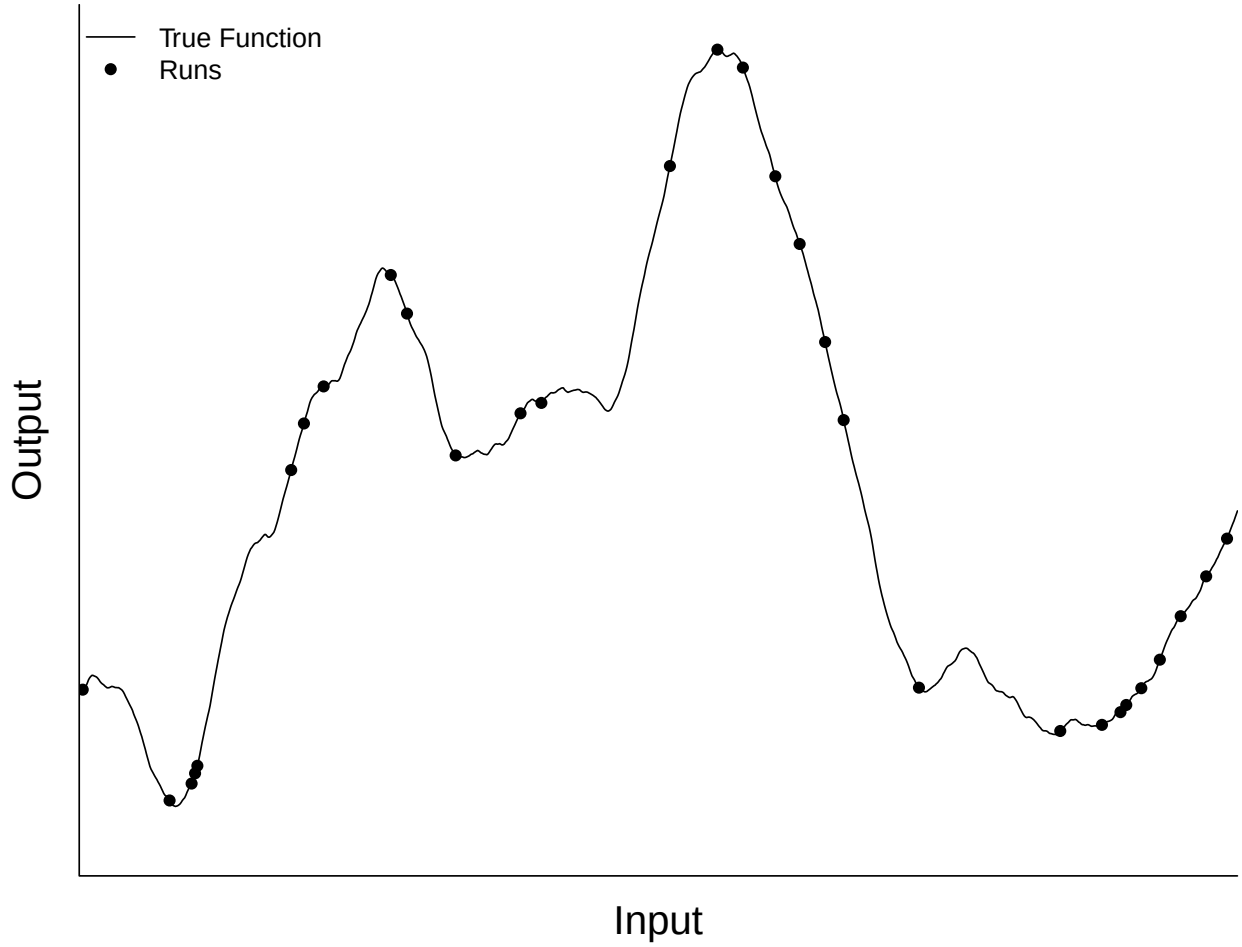
Statisticians build **statistical models**, which are mathematical representations that describe how **data** is generated from the underlying **process**

$\eta : \mathcal{X} \mapsto \mathcal{Y} \Rightarrow$ True Input-Output Relationship



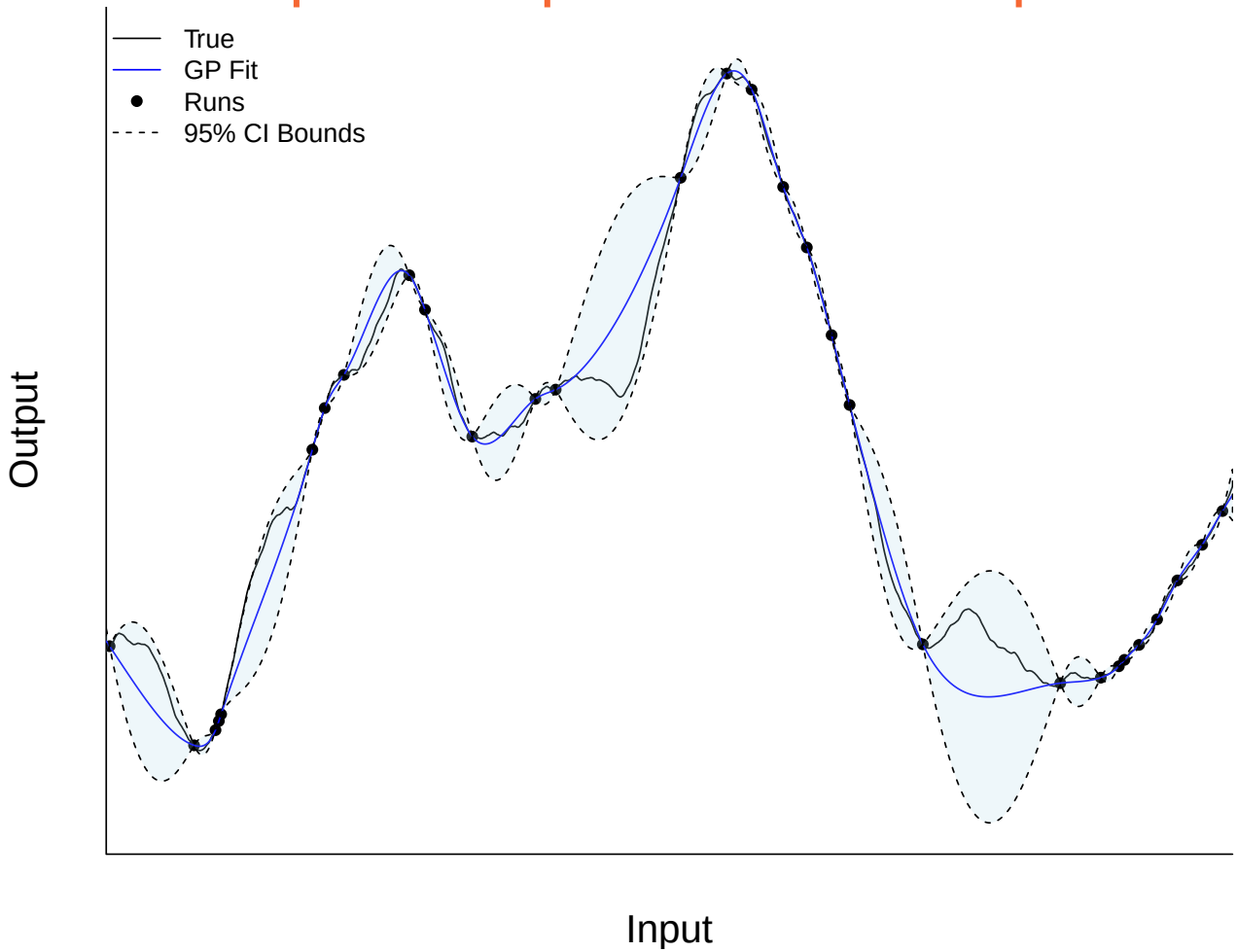
Such a function is informed by physics ☺ (fluid dynamics, shallow water equations...)

(Incomplete) $\eta(x)$ from Computer Model



Running the computer model is time-consuming, so model runs are limited ☹. $\Rightarrow$ **We need a model to fill in the missing pieces**

Estimated Input-Output Relationship and its Uncertainty

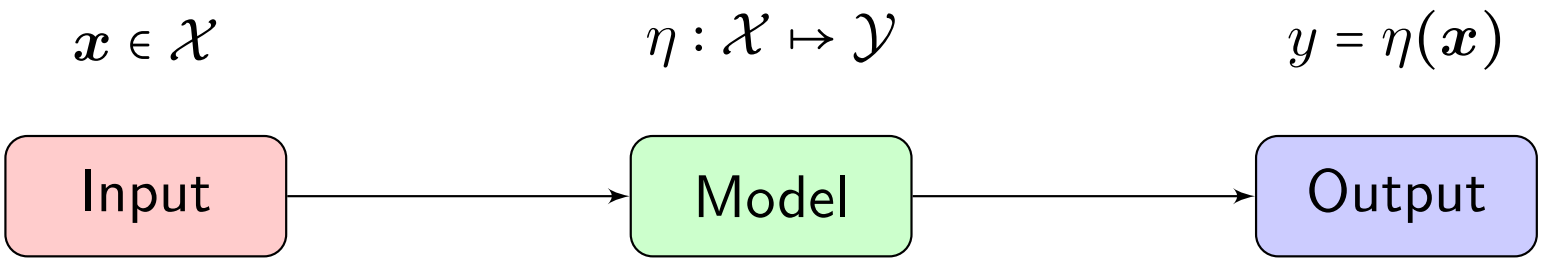


- Function estimation achieved by Gaussian process (GP):

$$\eta(\boldsymbol{x}) \sim \mathrm{GP}\left(m(\boldsymbol{x}), K(\boldsymbol{x}, \boldsymbol{x}')\right), \quad \boldsymbol{x} \in \mathcal{X}$$

- GP offers an "optimal" estimate $\hat{\eta}(\boldsymbol{x})$ with "localized" uncertainty (error bars)

Finishing the Workflow of Estimating Extreme Surge



TC Characteristics
Estimating $f(\boldsymbol{x})$✓

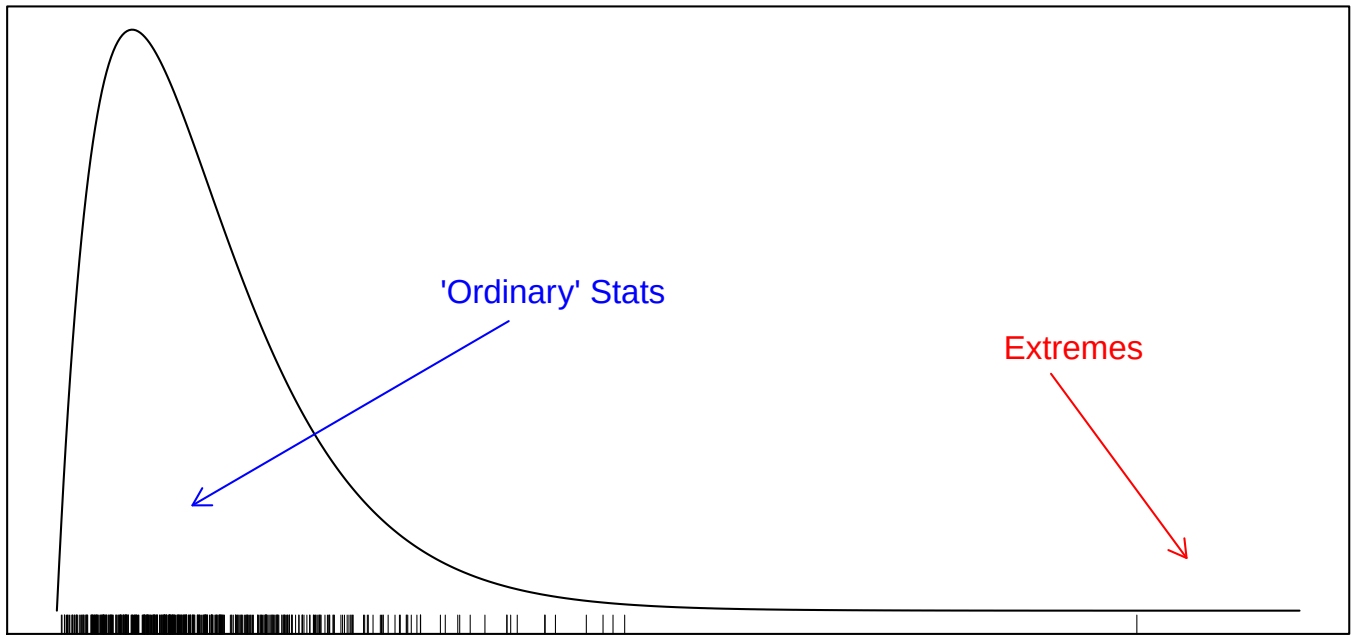
Computer Model
Estimating $\eta(\boldsymbol{x})$✓

Surge Level
Task: Estimating y_r

Workflow:

- Simulate $\boldsymbol{x}$ (synthetic storms) from $\hat{f}(\boldsymbol{x})$ (estimated probability distribution of TC characteristics)
- Input $\boldsymbol{x}$ into $\hat{\eta}(\boldsymbol{x})$ (estimated computer model input-output relationship) to obtain $[Y]$ (storm surge response distribution)
- Estimate extremes (e.g., one-in-100 surge level) of $[Y]$ via extreme value analysis

Extreme Value Analysis[1]



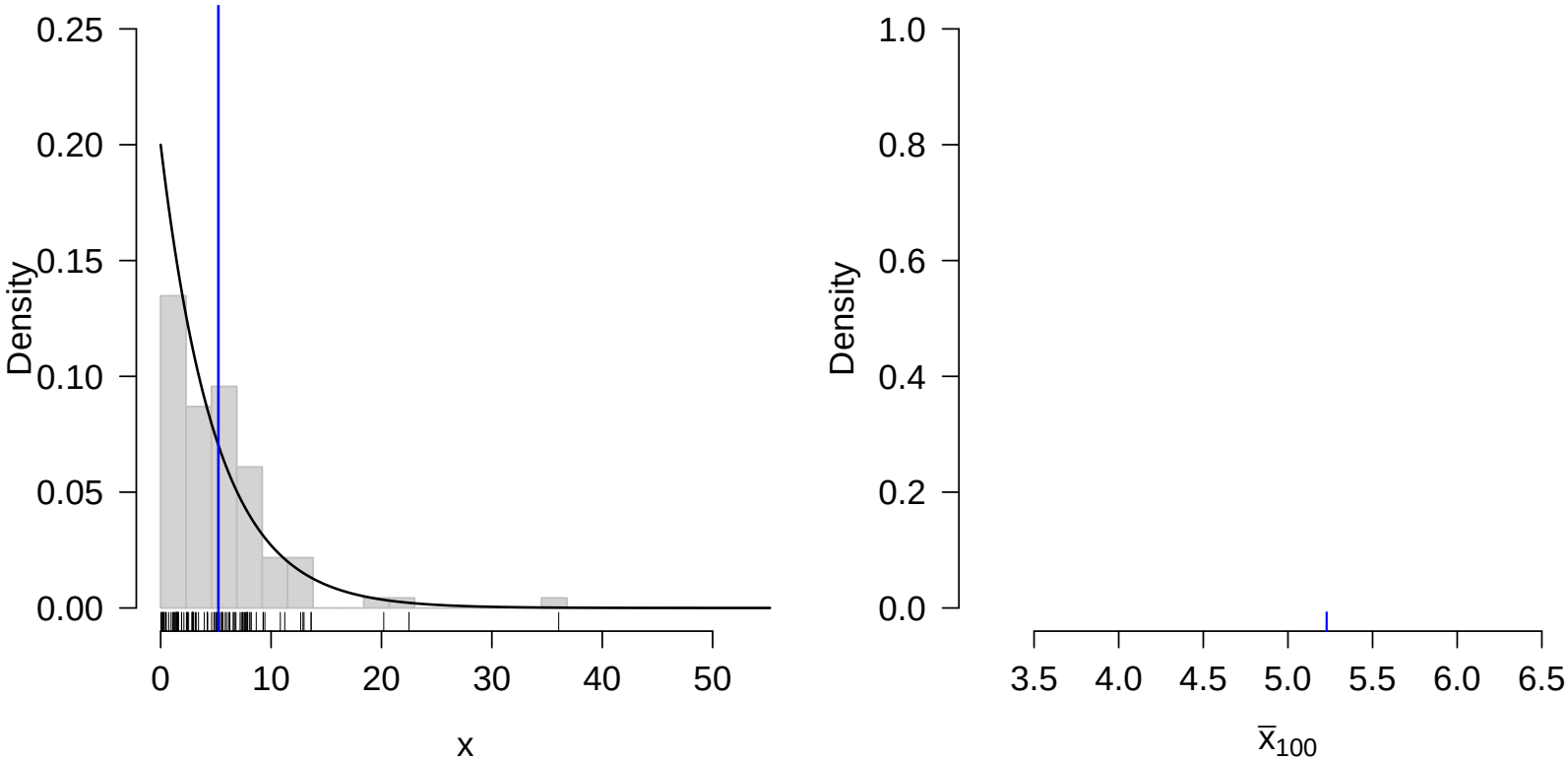
	Target	Theory	Distribution
Ordinary Stats	bulk distribution	CLT	Normal
Extreme Stats	tail distribution(s)	?	?

Let's examine the distribution of the **sample maximum** to learn about extremes

[1] I have delivered several short courses on this topic nationally and internationally, e.g., https://whitneyhuang83.github.io/EVA/Slides/WhitneyHuang_ShortCourse2_Longer_V2.pdf

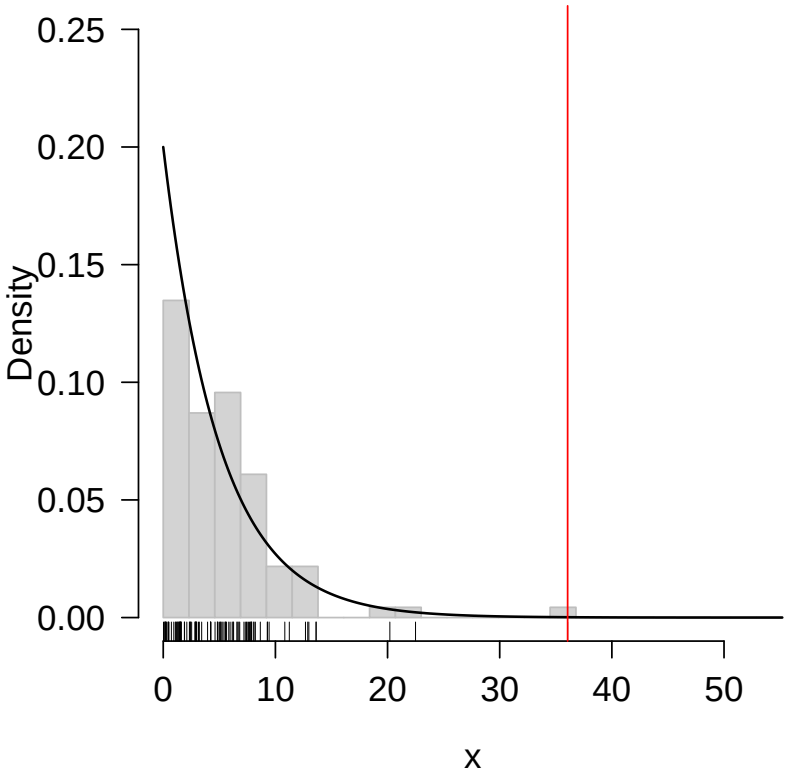
Central Limit Theorem Demonstration

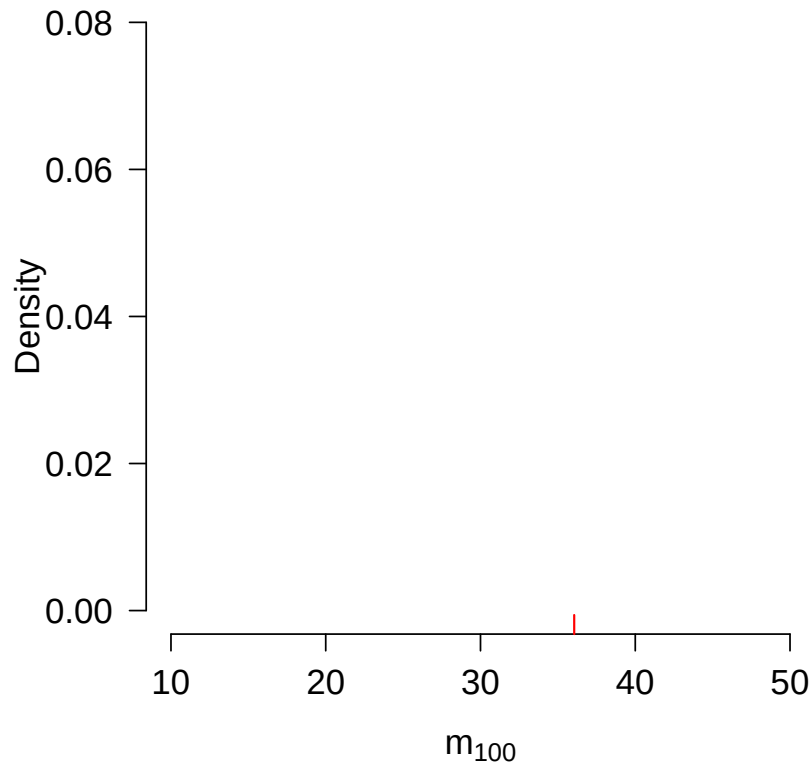
1. Generate 100 random numbers ($n = 100$) from an Exponential distribution
2. Compute the sample mean of these 100 random numbers
3. Repeat this process 120 times

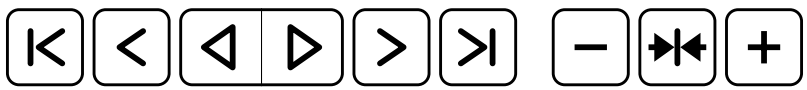


Demo: Distribution of the Sample Maximum

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







Extremal Types Theorem [Fisher–Tippett 1928, Gnedenko 1943]

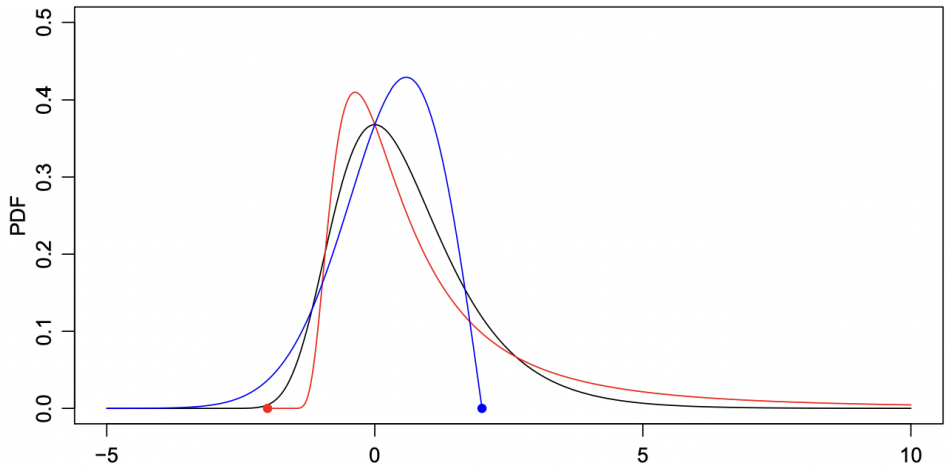
Define $M_n = \max\{X_1, \cdots, X_n\}$ where $X_1, \cdots, X_n \overset{\text{i.i.d.}}{\sim} F$. If $\exists\, a_n > 0$ and $b_n \in \mathbb{R}$ such that, as $n \to \infty$, if

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

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 ($\texttt{GEV}(\mu, \sigma, \xi)$), where μ and σ are location and scale parameters, and ξ is the shape parameter



- $\xi > 0$: Fréchet (heavy-tail)
- $\xi = 0$: Gumbel (light-tail)
- $\xi < 0$: reversed Weibull (short-tail)

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., $u \to x_F = \sup\{x : F(x) < 1\}$),

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

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

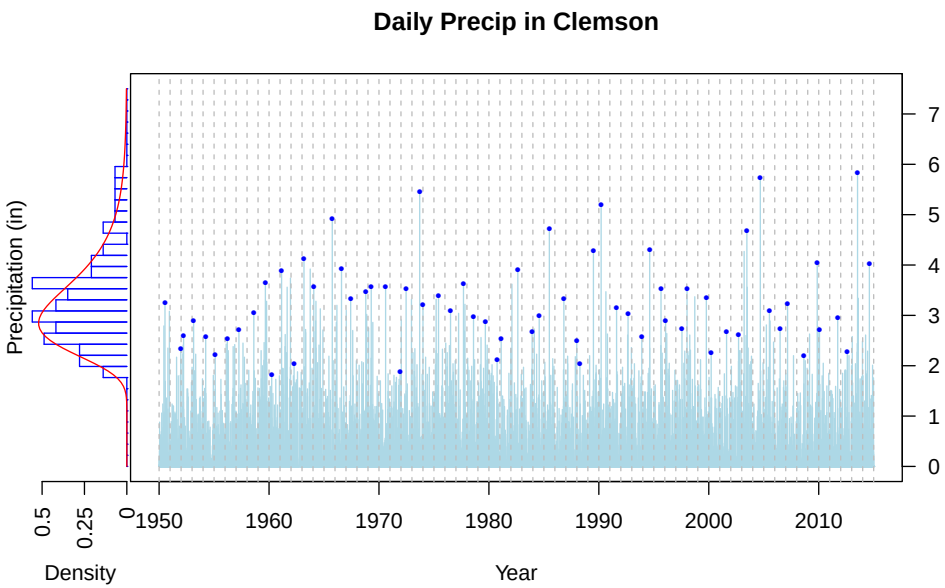
$$F_u(y) \xrightarrow{d} H_{\tilde{\sigma},\xi}(y) \qquad u \to x_F$$

where

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

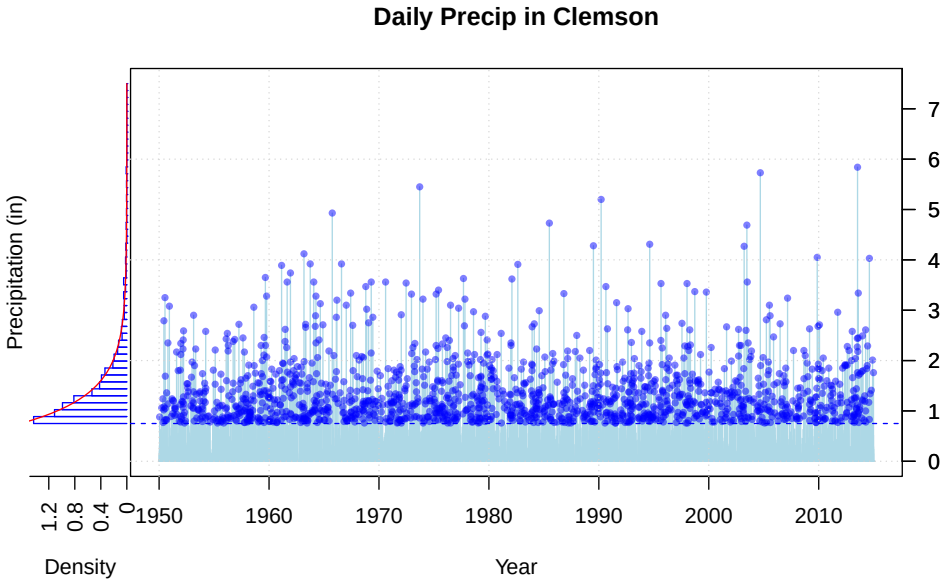
and $\tilde{\sigma} = \sigma + \xi(u - \mu)$

Univaraite Extreme Estimation: Two Main Approaches



- **Block Maxima**: Fit a generalized extreme value (GEV) distribution to block maxima (given block size big enough)

$$Y_{(n)} \overset{n\to\infty}{\sim} \text{GEV}(\mu, \sigma, \xi)$$



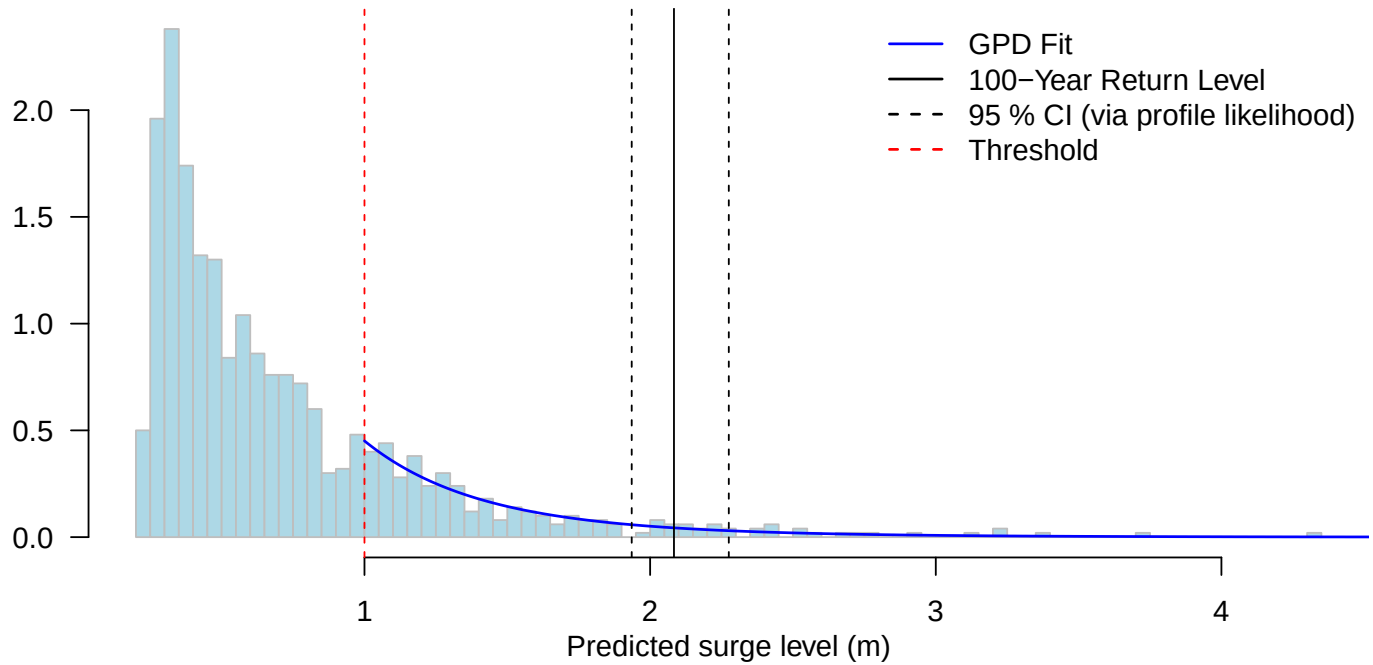
- **Threshold Exceedances**: Fit a generalized Pareto distribution (GPD) to exceedances over (a sufficiently high) threshold

$$Y - u | Y > u \overset{u\to y_F}{\sim} \text{GPD}(\tilde{\sigma}, \xi)$$

Estimating Extreme Surges: Extreme Value Analysis

We employed the threshold exceedances 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



Some General Research Questions

- How to overcome the "data poor" situation when modeling extremes? [**H.**, Nychka, Zhang, 2019]
- How to model extremes when the process of interest involves several variables? [**H.**, Monahan, Zwiers, 2021]
- How extremes vary in space? How extremes may change in future climate conditions? [**H.** et al., 2016]; [**H.** et al., 2019]

Research Group Members

- **Eva Murphy**, PhD, Aug, 2023 "*Modeling of Wind Speed and Wind Direction*." Current Position: Postdoc at Wake Forest University
- **Kanon Kamronnaher**, PhD, Dec. 2023 "*Estimating Financial and Environmental Risk*"
- **Adam Diaz**, MS, May 2022 "*Analysis of Climate-Wildfire Interaction*." Current Position: Principal Research Statistician at Northern California Institute for Research and Education
- **Emily Tidwell**, MS, May 2021 "*Physical-Statistical Approach for Estimating r-year Storm Surge*." Current Position: Dynetics, Huntsville, Alabama
- **Andrew Bellucco**, MS, Dec. 2019 "*Estimating Financial Risk*." Current Position: Senior Discovery Analyst at Credit Karma, Charlotte, NC
- **Peiying Li, MS, May 2024** "*Modeling Financial Time Series*." Current Position: Banking industrial in China

Current Members:

- **Jiyun (Joyce) Huang**, PhD Candidate (expected graduation date Aug. 2025), "*New Models and Methods for Estimating Rainfall Intensity-Duration-Frequency Curve*"
- **Katherine Kreuser**, MS, May 2022; PhD Candidate (expected graduation date Aug. 2025), "*Uncertainty Quantification for Rare Geophysical Extremes and Dynamical Engineering Applications*"
- Actively looking for motivated PhD/MS students interested in a environmental data science career in academic, government, or industry sectors. Feel free to email me or stop by (I am usually in the office unless traveling)

Some General Advice

"Survival Guide" for Graduate Students

- "How to Succeed in Graduate School" by Marie desJardins
  http://www.ai.sri.com/~marie/papers/advice-summary.html
- "Notes on the PhD Degree" by Douglas Comer
  https://www.cs.purdue.edu/homes/dec/essay.phd.html
- "Writing and Presenting your Thesis or Dissertation" by S. Joseph Levine
  http://www.learnerassociates.net/dissthes/

Some Useful Computing Skills

- R (https://www.r-project.org), RStudio (https://www.rstudio.com), and R markdown (http://rmarkdown.rstudio.com)
- LaTeX and Overleaf (https://www.overleaf.com/project)
- Python and jupyter (https://jupyter.org/)
- GitHub
- High-performance computing CCIT

Attend Seminars!

- Graduate Student Seminars
  http://siam.people.clemson.edu/gss/schedule.php
- Research Seminars (Statistics, Analysis, OR, ADM, Computational Math) & School Colloquia
- Many One World Seminar Series (e.g., Extremes, Spatial and spatio-temporal Point processes and beyond, Approximate Bayesian Computation, Probability, PDF, Mathematical Game Theory, Optimization, Mathematics of Machine Learning,...)

Join Us for the 75th Joint Clemson-UGA Colloquium

Speaker: Xiao-Li Meng
Whipple V. N. Jones Professor of Statistics, Harvard University

Date: 4/9, 2025
There will be two talks: one in Clemson U the other one in Clemson outdoor lab.

This event will be sponsored by

NISS | National Institute of Statistical Sciences

Attend Workshops!

- NSF-CBMS Regional Research Conferences in the Mathematical Sciences
  https://www.cbmsweb.org/regional-conferences/
- NSF funded mathematical sciences institutes
  https://mathinstitutes.org/
- Specialized workshops in your area

Attend Conferences! (and Give a Talk)

- Joint Mathematics Meetings (JMM): January
- Society for Industrial and Applied Mathematics (SIAM) Annual Meeting: July
- Institute for Operations Research and the Management Sciences (INFORMS) Annual Meeting: October or November
- Joint Statistical Meetings (JSM): Late July or early August
- Conference on Neural Information Processing Systems (NIPS): December

How to Reach Me?

- **Websites** 🖱: https://whitneyhuang83.github.io/
- **Email** ✉: wkhuang@clemson.edu
- **Office**: O-221 Martin Hall

Go Tigers!

Scan me!

Slides:
whitneyhuang83.github.io/Talks/2024Fall_SMSSFirstYr.pdf