SMSS First Year Graduate Student Seminar

Whitney Huang

✉: wkhuang@clemson.edu
🖱: https://whitneyhuang83.github.io/

School of
MATHEMATICAL AND STATISTICAL SCIENCES
Clemson® University

February 6, 2024

Agenda

About Myself

My Research

Some General Advice

Who am I?

- **Fifth year** Assistant Professor of Applied Statistics and Data Science

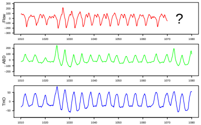
- Born in Laramie, Wyoming, grew up in Taiwan

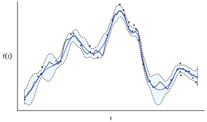
- Got a B.S. in Mechanical Engineering, switched to Statistics in graduate school

- Got a Ph.D. in Statistics, 2017 at Purdue; did a SAMSI/CANSSI Postdoc before moving to Clemson

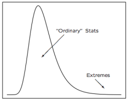
Overview of My Research

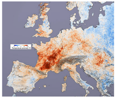
- ## Spatio-Temporal Statistics

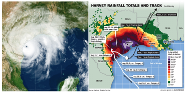




- ## Extreme Value Analysis





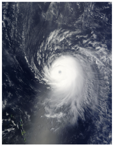


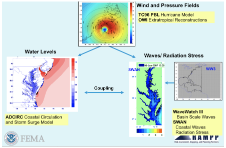
Source: NASA (Left); National Weather Service (Right)

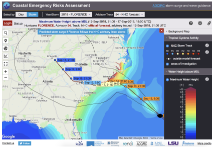
- ## Surrogate Modeling of Computer Experiments

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

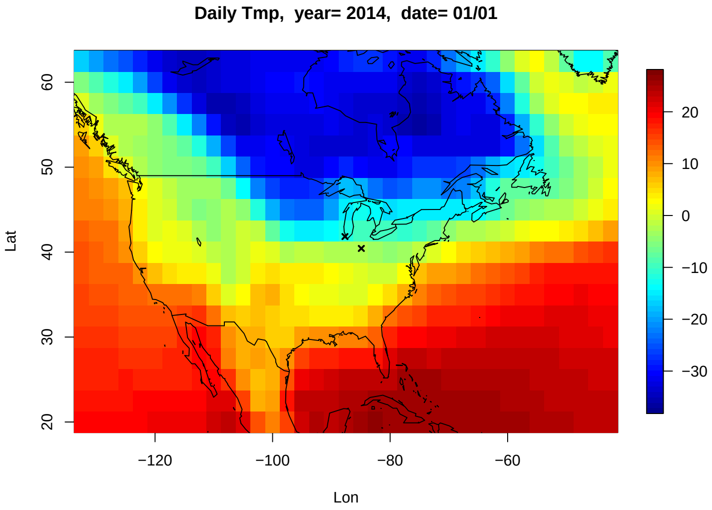
Input $\longrightarrow$ Model $\longrightarrow$ Output



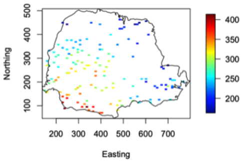


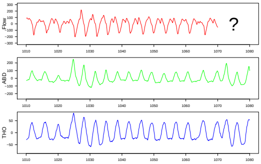


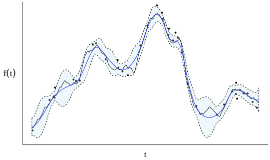
Spatio-Temporal Data



Statistical Modeling of Spatio-temporal Processes







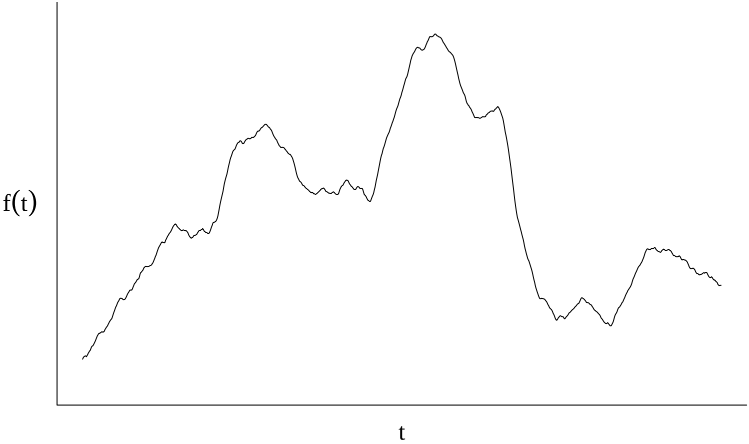
Some research questions:

- Prediction: spatial interpolation; time series forecasting
- Explanation: regression in space, time, or both space and time (and possibly other covariates)
- Uncertainty Analysis: quantify the uncertainties associated with spatio-temporal predictions

The foundation for the tasks mentioned above is function estimation

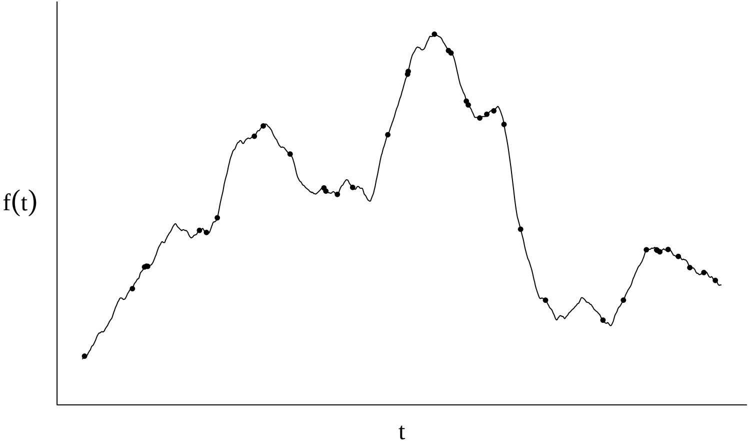
Gaussian Processes

Function Estimation



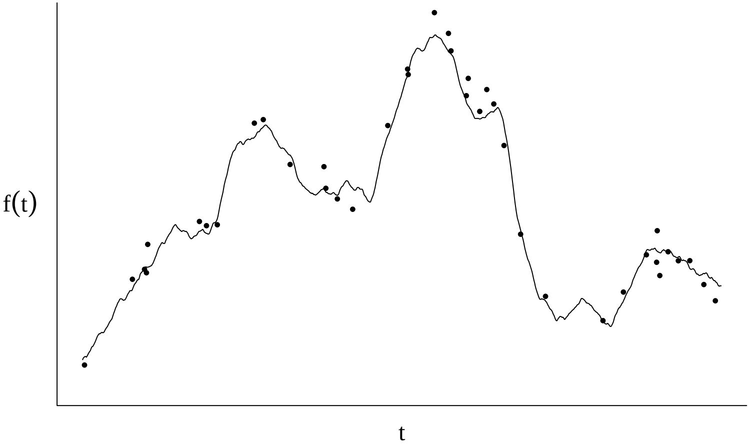
Consider a function $f(t), t \in \mathcal{T}$

Function Estimation



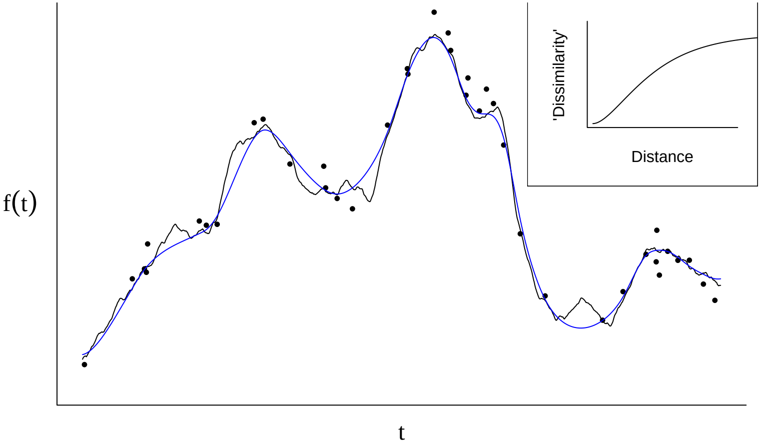
Consider $f(t)$, observed incompletely $\{f(t_i)\}_{i=1}^n$

Function Estimation



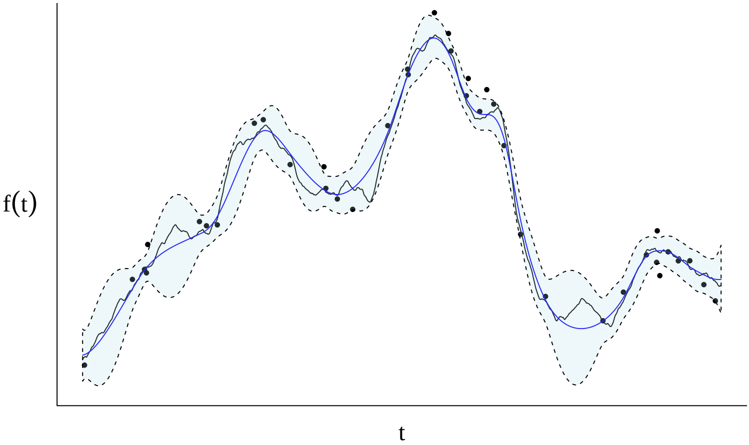
Consider $f(t)$, observed incompletely $\{f(t_i)\}_{i=1}^{n}$, and with noise $\{\varepsilon_i\}_{i=1}^{n}$

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")

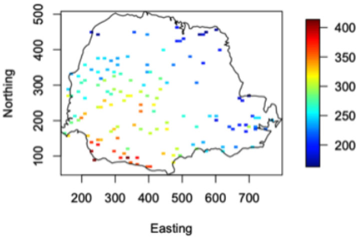
Spatial Interpolation via GP Model

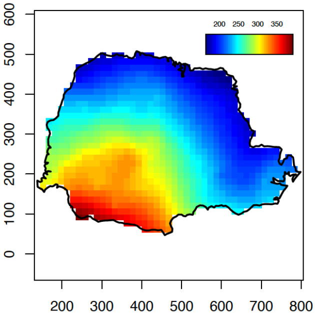
The "generic" spatial model

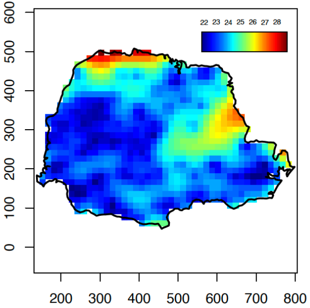
$$Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \eta(\boldsymbol{s}),\ \boldsymbol{s} \in \mathcal{S},$$

where

- $\mu(\boldsymbol{s})$ 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 $c_{\boldsymbol{\theta}}(h)$

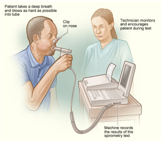


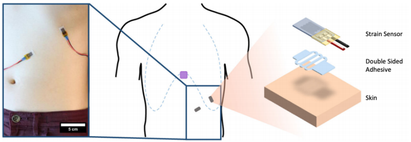


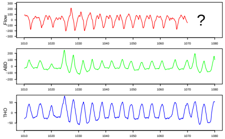


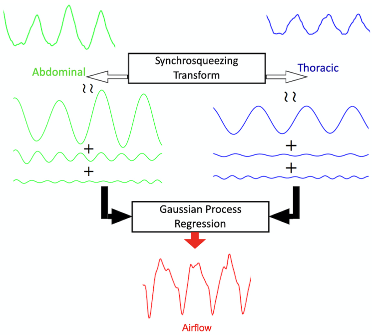
High-Frequency Physiological Waveform Signal Analysis

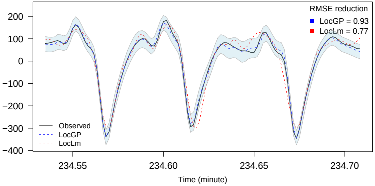
[joint with Wu (NYU/Duke), Chung (Eli Lilly), Wang (Clemson), and Mandel (UPenn)]

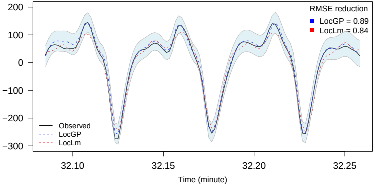




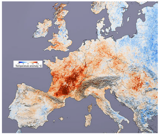


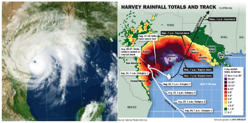




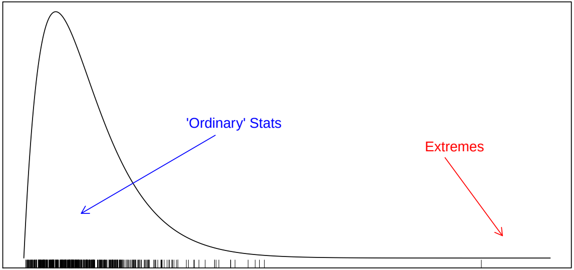


Extreme Value Analysis



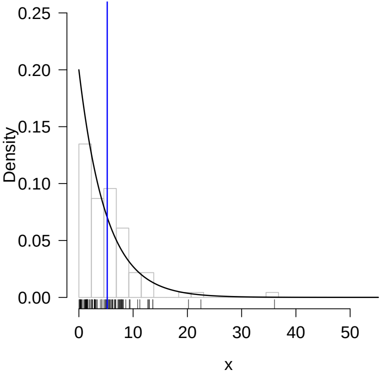
Source: NASA (Left); National Weather Service (Right)

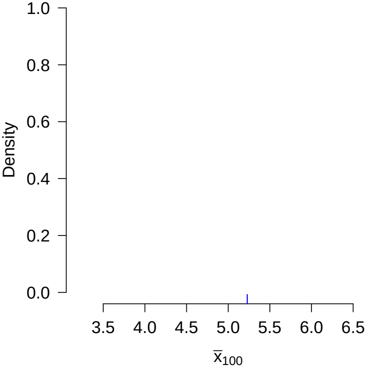


	Target	Theory	Distribution
Ordinary Stats	bulk distribution	CLT	Normal
Extreme Stats	tail distribution(s)	?	?

Central Limit Theorem Demonstration

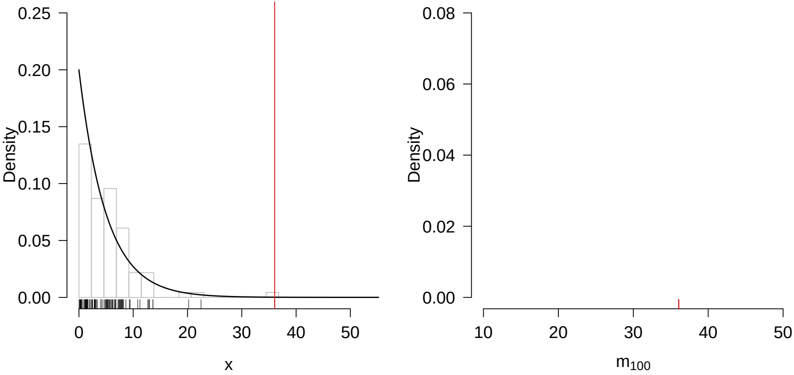
1. Generate 100 random numbers $(n = 100)$ from an Exponential distribution
2. Compute the sample mean of these 100 random numbers
3. Repeat this process 120 times



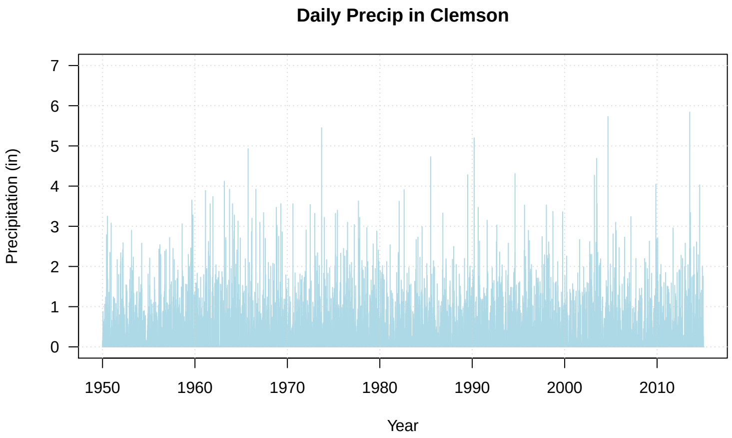


Demo: Distribution of the Sample Maximum

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



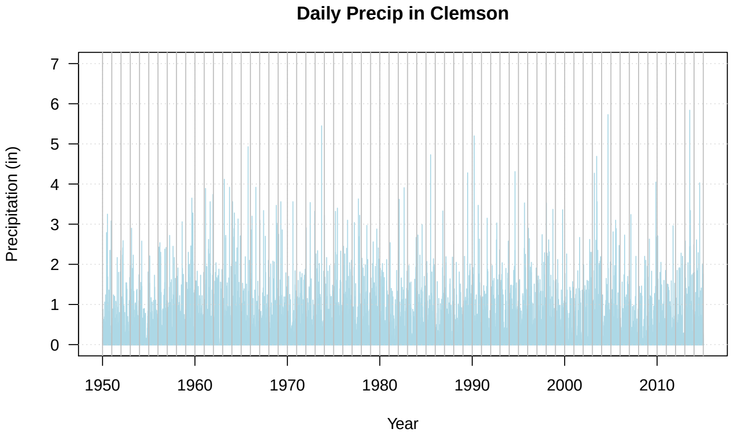
Estimating Clemson Daily Precipitation Extremes



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

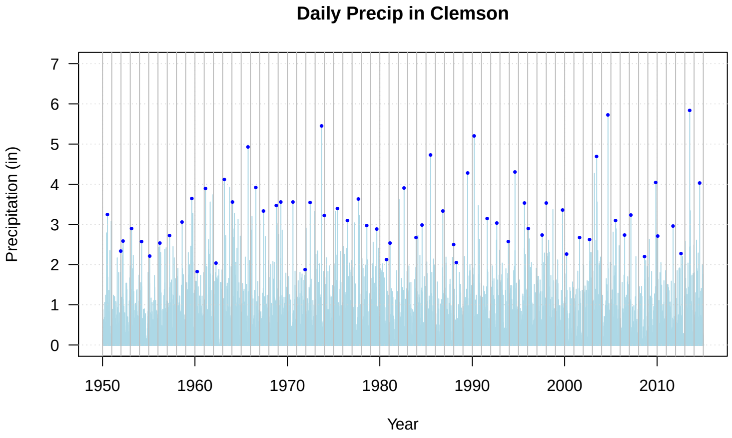
Block Maxima Method (Gumbel 1958)

1. Determine the block size and extract the block maxima



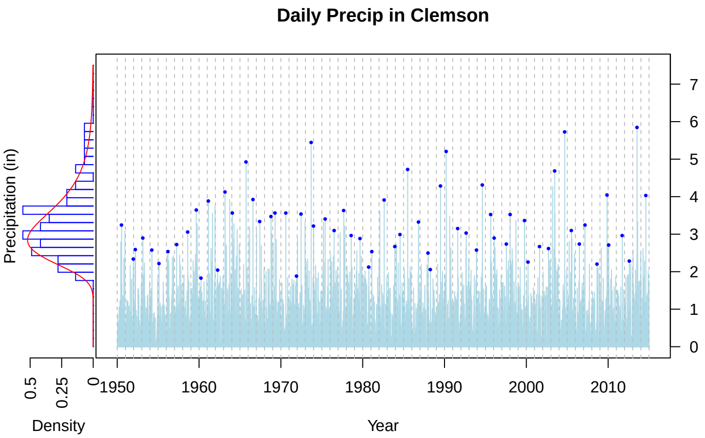
Block Maxima Method (Gumbel 1958)

1. Determine the block size and extract the block maxima



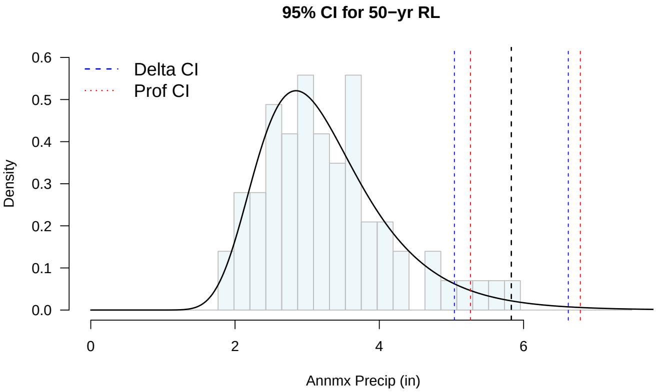
Block Maxima Method (Gumbel 1958)

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



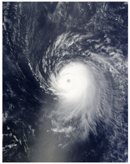
Point estimate: $5.68(\text{inches}/24\text{hrs})$.
Interval estimate: $[4.63, 6.72]$(Delta method CI)
$[4.98, 7.47]$ (Profile likelihood CI)

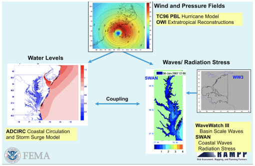
Some Research Questions

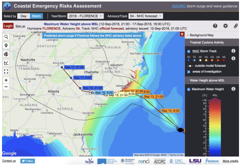
- How to overcome the "data poor" situation when modeling extremes? [**H.**, Nychka, Zhang, 2019]
- How to model extremes when the process of interest involves several variables? [**H.**, Monahan, Zwiers, 2021]
- How extremes vary in space? How extremes may change in future climate conditions? [**H.**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

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

Input → Model → Output





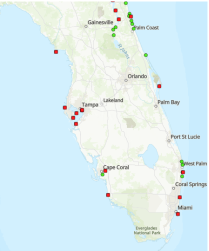


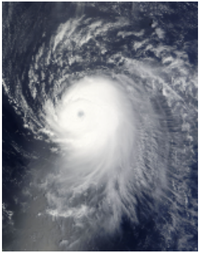
Estimating Extreme Surges: Physical-Statistical Approach

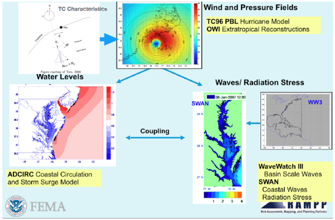
$\boldsymbol{x} \in \mathcal{X}$ $\qquad$ $\eta : \mathcal{X} \mapsto \mathcal{Y}$ $\qquad$ $y = \eta(\boldsymbol{x})$

Input → Model → Output

Variable / Data	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





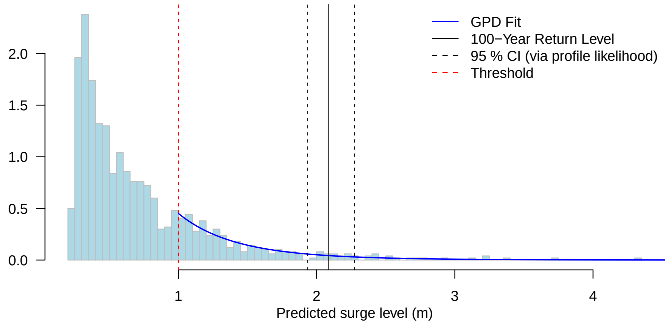


Courtesy of Gangai (Dewberry) & Danforth (FEMA)

Estimating Extreme Surges: 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



Research Group Members

- Eva Murphy, PhD, Aug, 2023 "*A Framework for Statistical 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 "*A Contribution to the Statistical Analysis of Climate-Wildfire Interaction in Northern California*." Current Position: Principal Research Statistician at Northern California Institute for Research and Education
- Emily Tidwell, MS, May 2021 "*A Combined Physical-Statistical Approach for Estimating r-year Storm Surge Depth*." Current Position: Dynetics, Huntsville, Alabama
- Andrew Bellucco, MS, Dec. 2019 "*Estimating Financial Risk*." Current Position: Senior Discovery Analyst at Credit Karma, Charlotte, NC
- Jiyun (Joyce) Huang, PhD Candidate (expected graduation date Aug. 2024), "*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*"
- Peiying Li, MS (expected graduation date May 2024) "*Statistics and Deep Learning Methods for Modeling Financial Time Series*."
- Actively looking for motivated PhD/MS students interested in a data science career in academic, government, or industry sectors

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,...)

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/

- 2024 Southern Regional Council on Statistics Summer Research Conference
  https://whitneyhuang83.github.io/SRCOS-SRC2024.html

June 2-5, 2024
Clemson University, Clemson, South Carolina, USA

Topics: Statistics and AI, Experimental Designs and UQ, Precision Medicine, Environmental Statistics, High Dimensional/Highly Structured Data, Statistics Education

Confirmed Speakers:

Bhramar Mukherjee, UMich	Simon Sheather, UKY	Steve Marron, UNC
Mikyoung Jun, UH	David Edwards, VCU	Ashwin Pananjady, GT
Lulu Kang, UMass	Zhengwu Zhang, UNC	Ana-Maria Staicu, NCSU
David Hitchcock, UofSC	Anirban Bhattacharya, TAMU	Staci Hepler, WFU
Ben Seiyon Lee, GMU	Li-Hsiang Lin, GSU	Ray Bai, UofSC
Edward Boone, VCU	Lu Wang, Umich	Nikki Freeman, UNC

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!