MODELING TRENDS IN SPATIAL EXTREMES AND THEIR CAUSAL DETERMINATION

Richard L. Smith

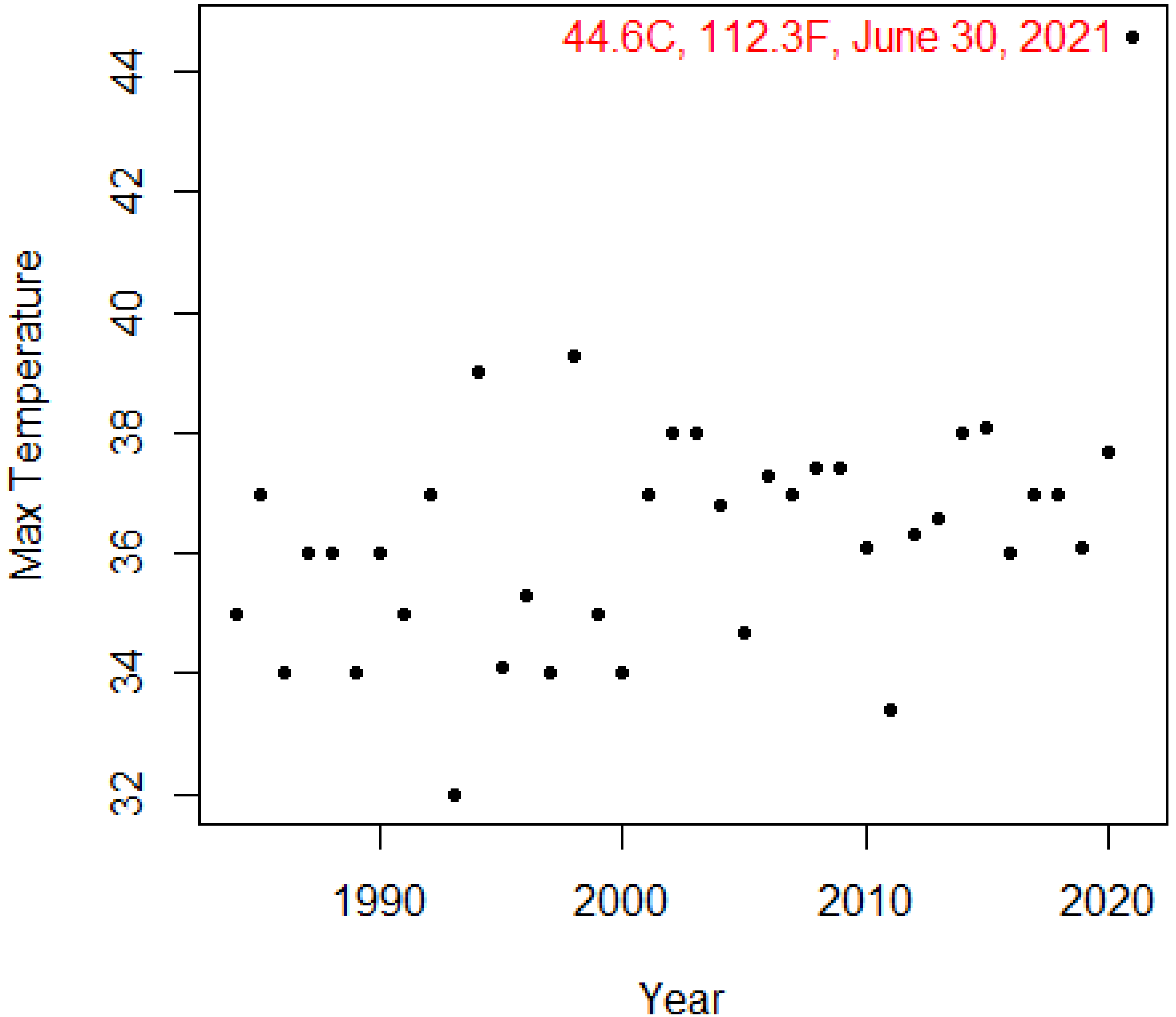
Banff Workshop on "Combining Causal Inference and Extreme Value Theory in the Study of Climate Extremes and their Causes"

University of British Columbia Okanagan

June 29, 2022 (updated July 1)

THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Annual Maximum Daily Maximum Temperatures in Kelowna, BC, 1984-2021



Generalized Extreme Value (GEV) Distribution

- $G(y) = \exp\left\{-\left(1+\xi\frac{y-\mu}{\sigma}\right)^{-1/\xi}\right\}$ defined when $1+\xi\frac{y-\mu}{\sigma} > 0$

- Shape parameter ξ
  - $\xi > 0$: long-tailed, upper bound ∞, ultimately $1 - G(y) \sim cy^{-1/\xi}$ for some $c > 0$
  - $\xi \to 0$: exponential tail, Gumbel distribution
  - $\xi < 0$: short-tailed, endpoint at $\mu - \frac{\sigma}{\xi}$

- Estimation: MLE, PWM, Bayes

- Kelowna data: fit to 1984–2020, use this to determine "how extreme" was 2021

- MLE shows $\hat{\xi} = -0.42$, estimated endpoint 39.8

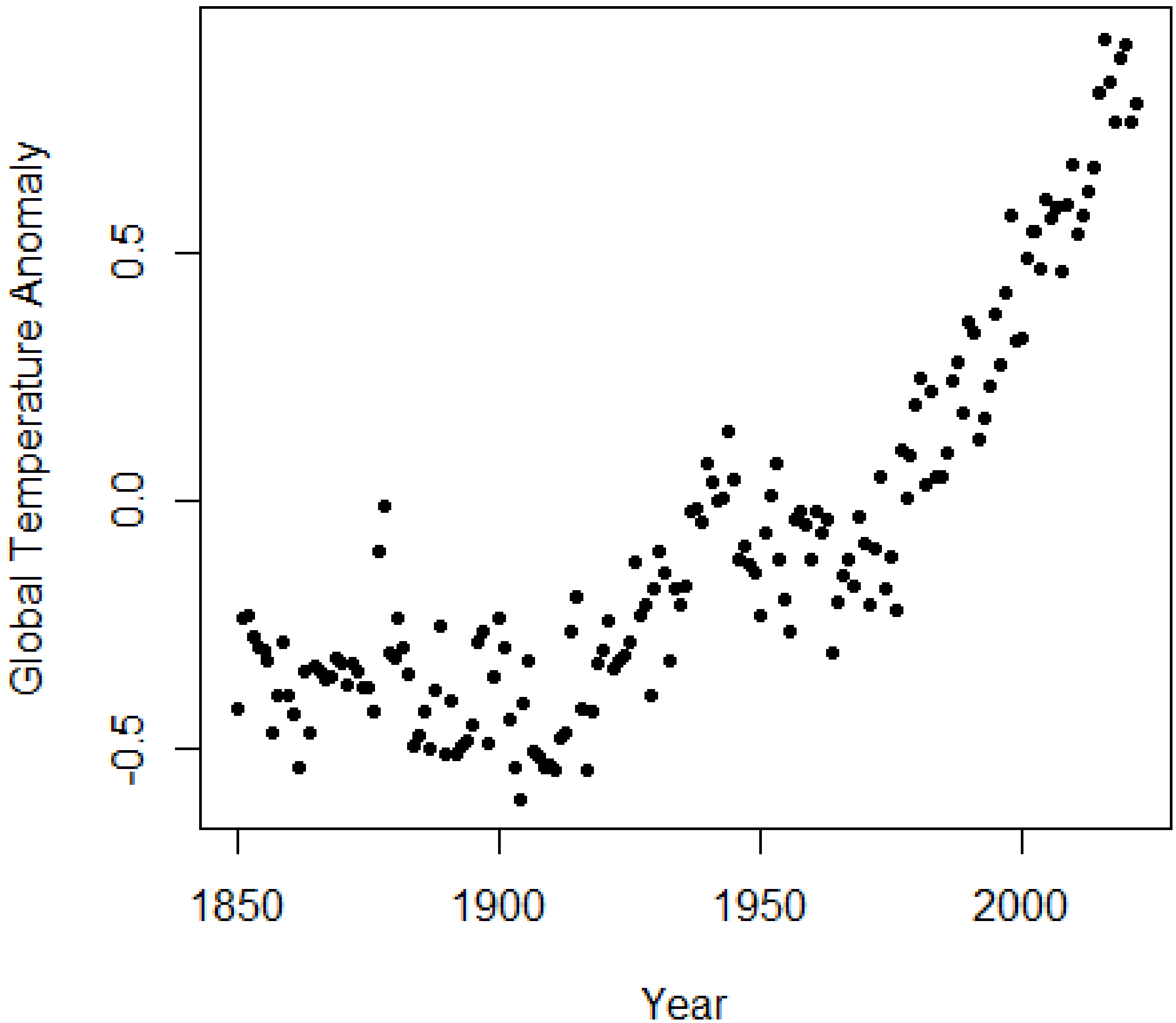
Parameter	Estimate	SE	t value	p value
mu	35.7119	0.3098	115.2624	0
log sigma	0.5477	0.1305	4.1962	0
xi	-0.4203	0.1007	-4.1718	0

Adding a Trend

- $G_t(y) = \exp\left\{ -\left(1 + \xi \frac{y - \mu_t}{\sigma}\right)^{-1/\xi} \right\}$
- $\mu_t = \beta_0 + \beta_1 x_t$
- Here, we take x_t to be global mean surface temperature for year t (HadCRUT5)
- Statistically significant against no-trend model ($p \approx 0.03$)
- $\hat{\xi} = -0.29$, estimated endpoint for 2021 is 41.6

Parameter	Estimate	SE	t value	p value
beta0	34.3983	0.6427	53.5175	0
logsig	0.4088	0.1292	3.1632	0.0016
xi	-0.2879	0.1101	-2.6139	0.009
beta1	2.6448	1.1503	2.2992	0.0215

HadCRUT5 Global Temperature Means



Better Way to Estimate Exceedance Probability

- Bayesian predictive analysis

- If $p_{\text{crit}}(\mu, \sigma, \xi) = 1 - \exp\left\{-\left(1 + \xi \frac{y_{\text{crit}} - \mu}{\sigma}\right)^{-1/\xi}\right\}$, compute

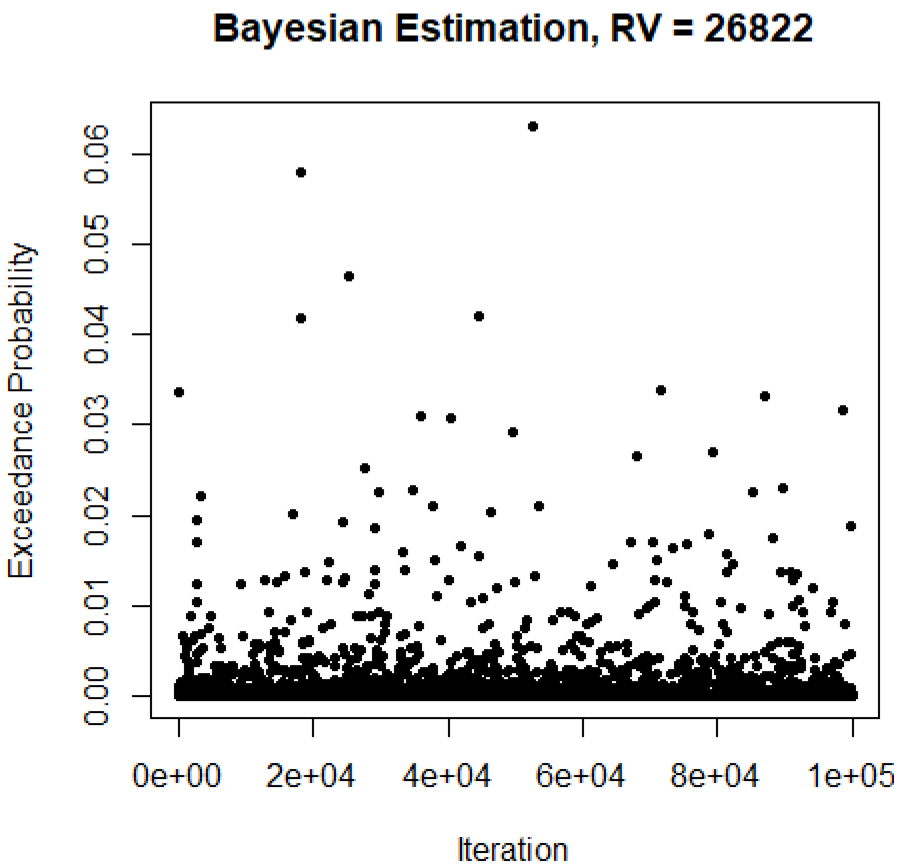
$$\pi(p_{\text{crit}}|\mathbf{Y}) = \int\int\int p_{\text{crit}}(\mu, \sigma, \xi) \cdot \pi(\mu, \sigma, \xi|\mathbf{Y}) d\mu d\sigma d\xi$$

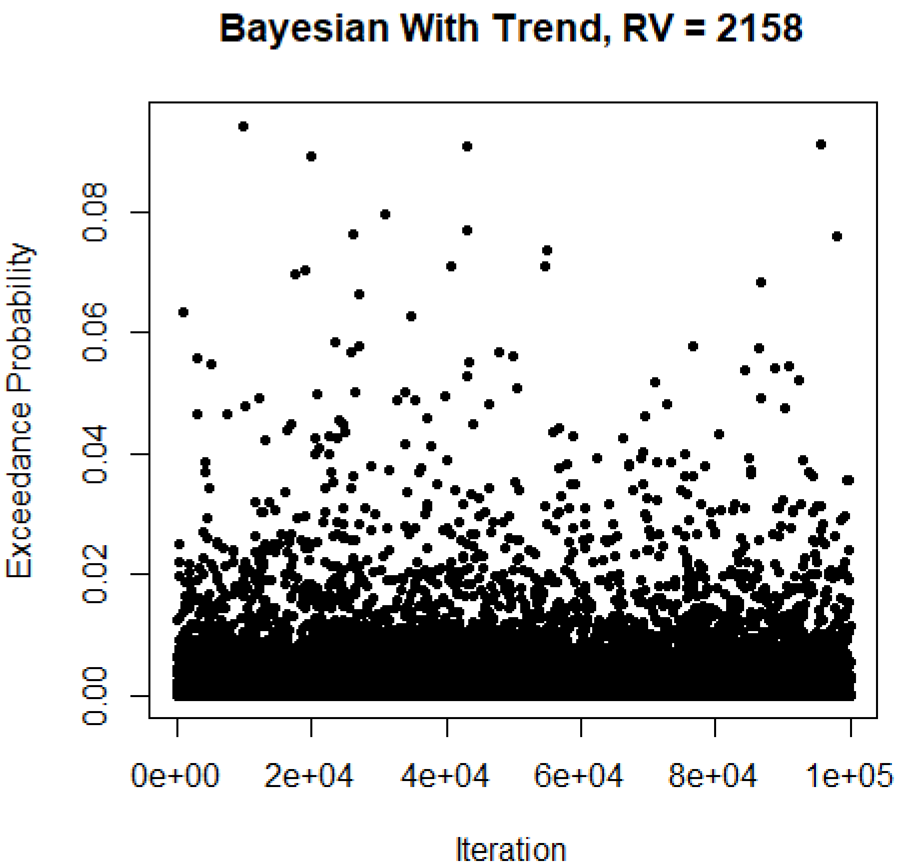
- In practice: use MCMC to produce simulation from posterior

- Adaptive Metropolis (Haario et al., 2001) works very well

- The posterior mean of p_{crit} is > 0, but the posterior distribution still has a large atom at 0

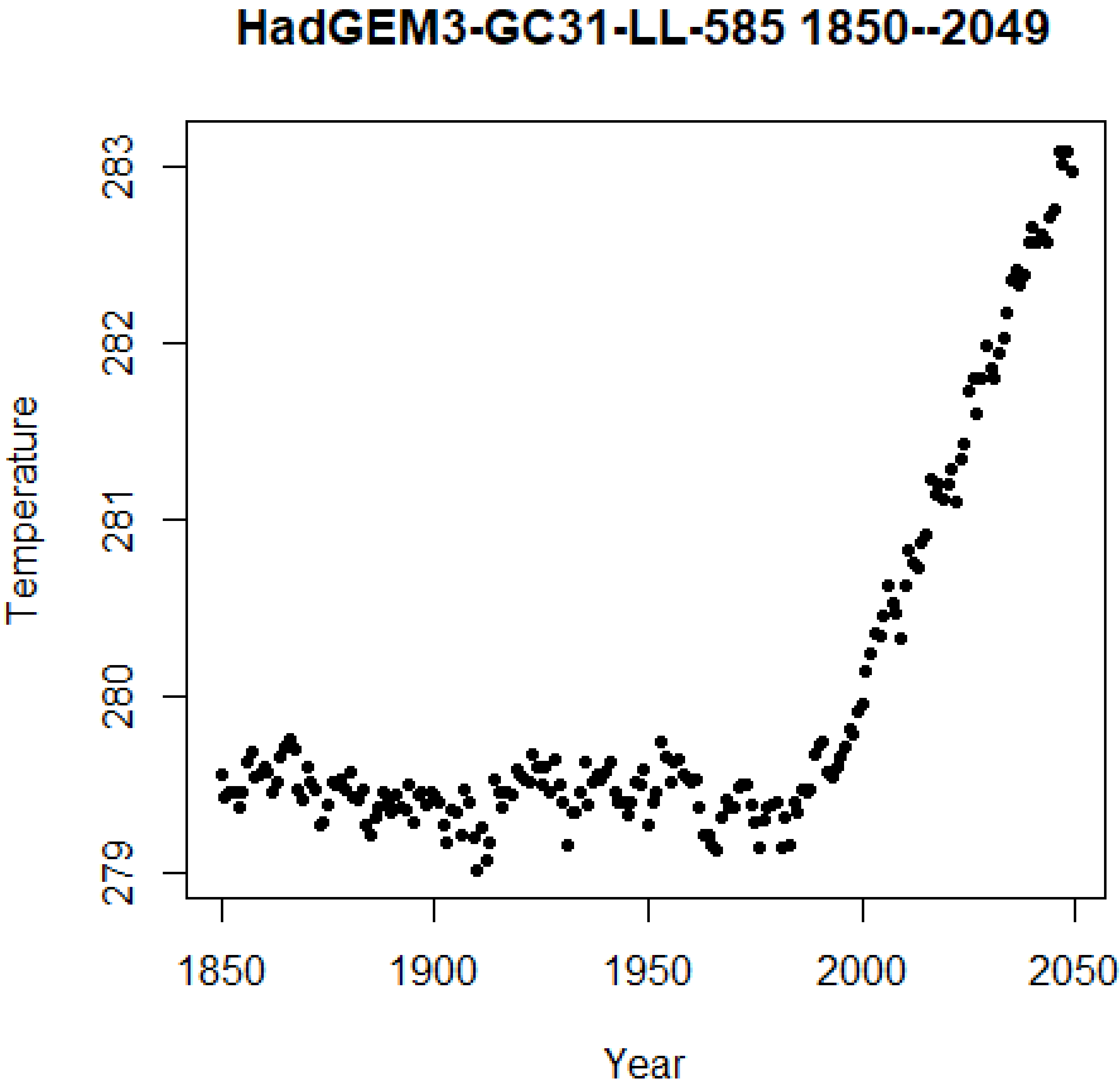
- $1/p_{\text{crit}}$ is called the *return value* (RV) for y_{crit}, but it's still questionable whether to treat the posterior mean of p_{crit} as a point estimate


Bayesian Estimation, RV = 26822
Exceedance Probability
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Iteration


Bayesian With Trend, RV = 2158
Exceedance Probability
0.00
0.02
0.04
0.06
0.08
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Iteration

- This style of analysis shows how we can relate the distribution of a local variable of interest to that of a global variable (which, for convenience, we have taken to be global mean temperature) for which the causality question has been widely studied

- However to get quantitative estimates, we need to look at an actual climate model (or several of them)


HadGEM3-GC31-LL-585 1850--2049
Temperature
283
282
281
280
279
1850
1900
1950
2000
2050
Year

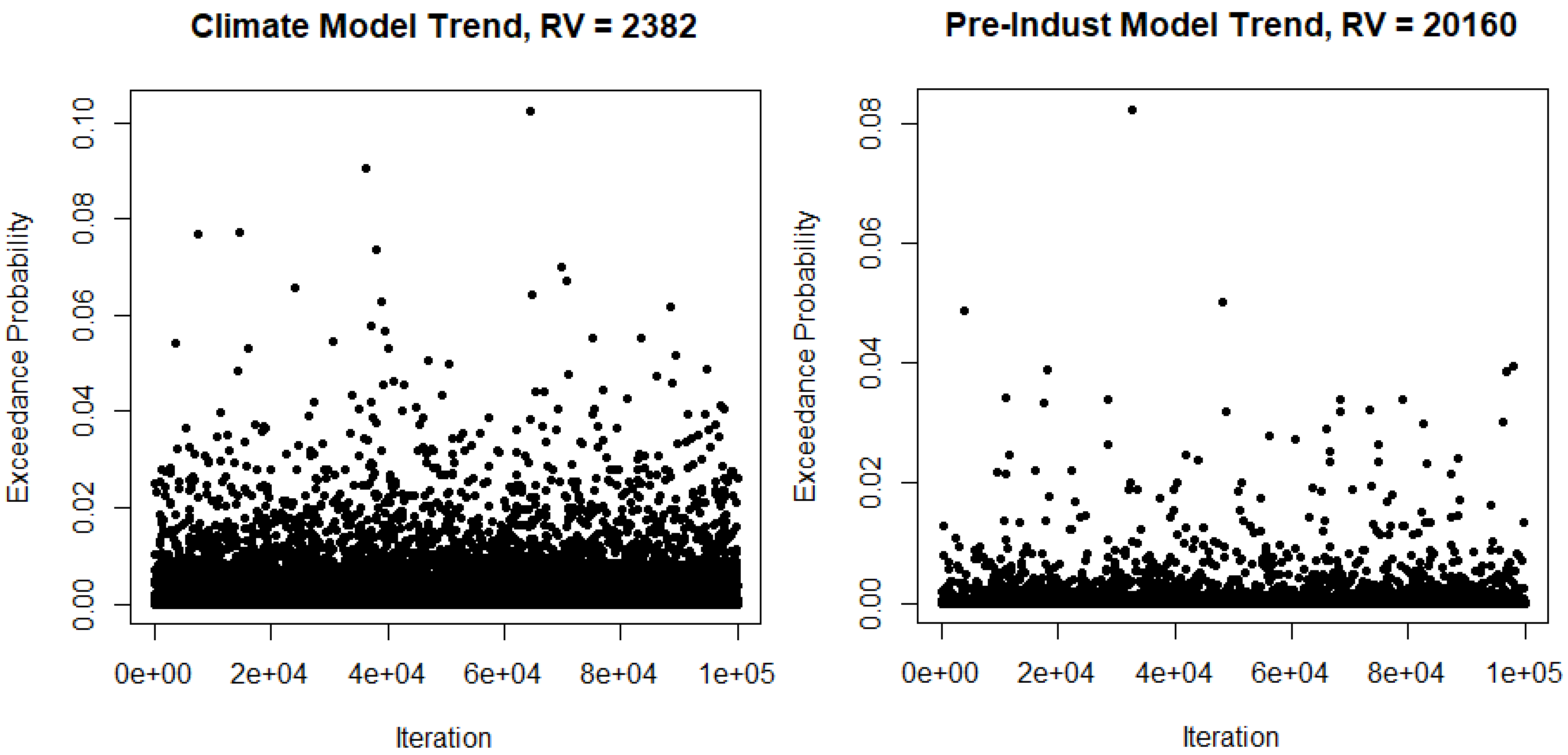
- I have repeated the previous analyses with global mean temperature from HadGEM3 in place of observational data

- For a counterfactual run, I used data from 1884–1920 instead of 1984–2020

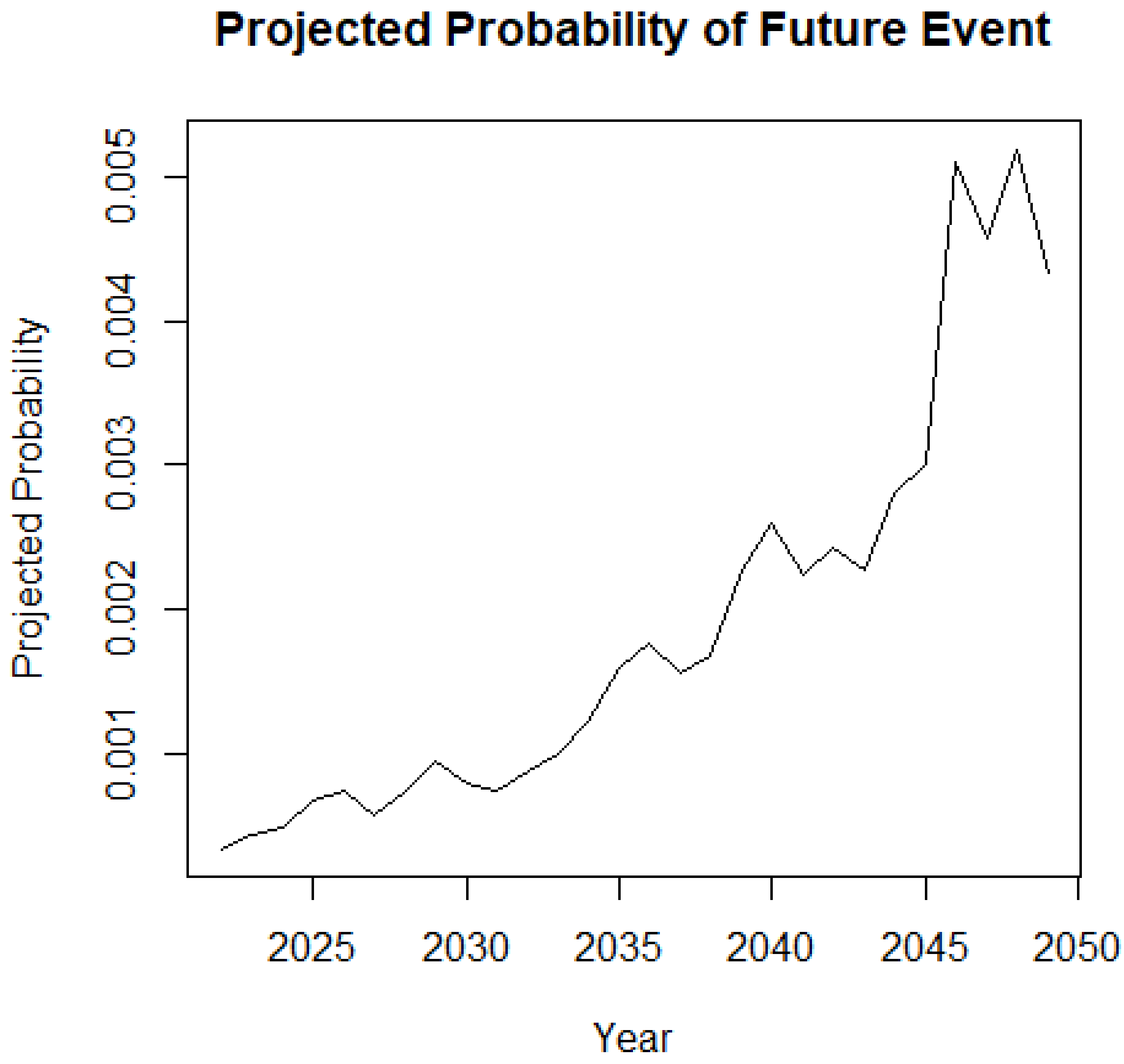
- This isn't the ideal way to do it — better to use a "natural forcings" run that includes solar fluctations and volcanic eruptions.

- A caveat about this approach: the model runs started in 2015 so the data post-2014 are projections, not based on actual temperatures

- The quoted RVs are volatile, but they give an idea how the estimated probabilities vary with the different models


Climate Model Trend, RV = 2382
Exceedance Probability
0.00
0.02
0.04
0.06
0.08
0.10
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Iteration
Pre-Indust Model Trend, RV = 20160
Exceedance Probability
0.00
0.02
0.04
0.06
0.08
0e+00
2e+04
4e+04
6e+04
8e+04
1e+05
Iteration

- Finally, we can use the climate model projections combined with the GEV model to estimate future probabilities of exceeding this high temperature


Projected Probability of Future Event
Projected Probability
0.001
0.002
0.003
0.004
0.005
2025
2030
2035
2040
2045
2050
Year

Rapid attribution analysis of the extraordinary heatwave on the Pacific Coast of the US and Canada June 2021.

Contributors

Sjoukje Y. Philip[1], Sarah F. Kew[1], Geert Jan van Oldenborgh[1,19], Wenchang Yang[2], Gabriel A. Vecchi[2,3], Faron S. Anslow[4], Sihan Li[5], Sonia I. Seneviratne[6], Linh N. Luu[1], Julie Arrighi[7,8,9], Roop Singh[7], Maarten van Aalst[7,8,10], Mathias Hauser[6], Dominik L. Schumacher[6], Carolina Pereira Marghidan[8], Kristie L Ebi[11], Rémy Bonnet[12], Robert Vautard[12], Jordis Tradowsky[13,14], Dim Coumou[1, 15], Flavio Lehner[16,17], Michael Wehner[18], Chris Rodell[20], Roland Stull[20], Rosie Howard[20], Nathan Gillett[21], Friederike E L Otto[5]

Published online, July 2021

WWA Analysis of the Pacific Northwest Heatwave

Main findings

- Based on observations and modeling, the occurrence of a heatwave with maximum daily temperatures (TXx) as observed in the area 45–52 °N, 119–123 °W, was virtually impossible without human-caused climate change.
- The observed temperatures were so extreme that they lie far outside the range of historically observed temperatures. This makes it hard to quantify with confidence how rare the event was. In the most realistic statistical analysis the event is estimated to be about a 1 in 1000 year event in today's climate.

Smmary of Their Method

- Data on TXx (annual max daily temperature) over 45–52$^{\text{O}}$N, 119–123$^{\text{O}}$W
- Model extremes as a function of GMST (GEV distribution)
- $\mu_t = \beta_0 + \beta_1 \text{GMST}_t,\ \sigma_t = \sigma,\ \xi_t = \xi$
- Compare 2021 with late nineteenth century (GMST 1.2$^{\text{O}}$C lower than 2021) or projected future events (GMST 0.8$^{\text{O}}$C higher than 2021)

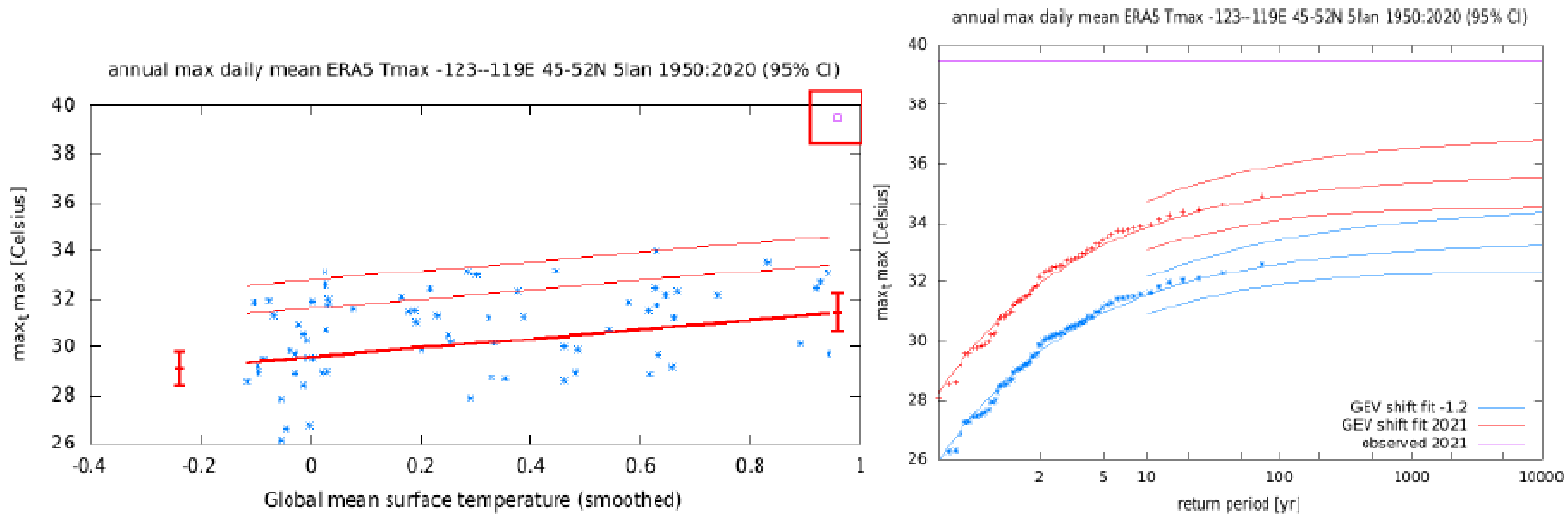


Figure 6. GEV fit with constant scale and shape parameters, and location parameter shifting proportional to GMST of the index series. No information from 2021 is included in the fit. Left: the observed TXx as a function of the smoothed GMST. The thick red line denotes the location parameter, the thin red lines the 6 and 40-yr return times. The June 2021 observation is highlighted with the red box and is not included in this fit. Right: Return time plots for the climate of 2021 (red) and a climate with GMST 1.2 °C cooler (blue). The past observations are shown twice: once shifted up to the current climate and once shifted down to the climate of the late nineteenth century. Based on ERA5 extended with operational ECMWF analyses for June 2021.

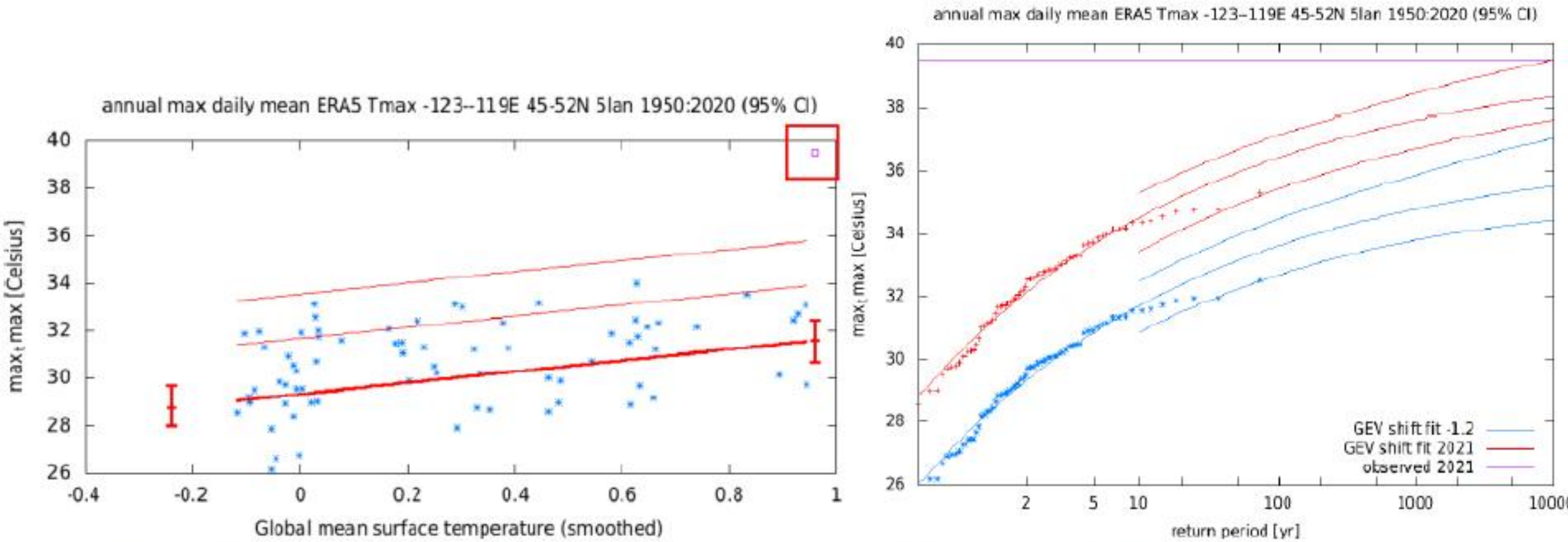


Figure 7. As Figure 6 but demanding the 2021 event is possible in the fitted GEV function, i.e., the upper bound is higher than the value observed in 2021.

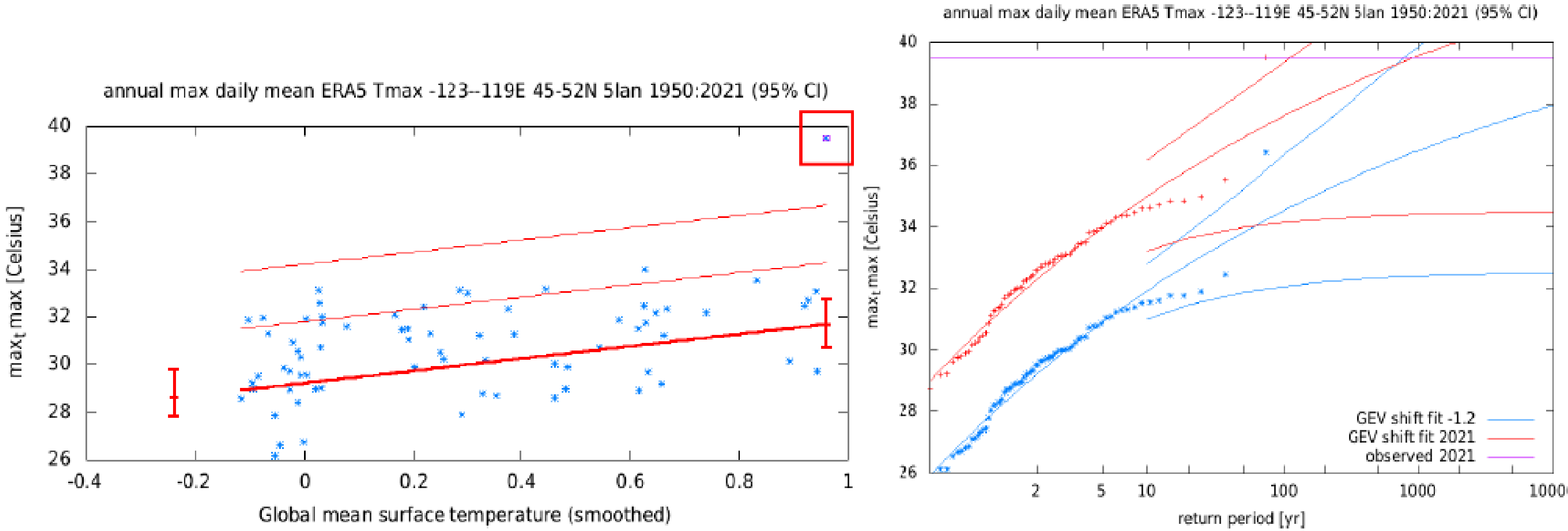


Figure 8. As for Figure 6 but including data from the 2021 heatwave into the fit.

How Extreme was Hurricane Harvey?

- Hurricane Harvey hit the Houston area at the end of August 2017

- Very excessive precipitations led to major flooding

- Meteorologically, associated with a stalling of the storm system just off the Gulf coast, but recent work by Kossin and others has suggested such events are becoming more common overall

- Statistically, questions about (a) just how extreme an event this was, (b) whether such events will become more common in the future

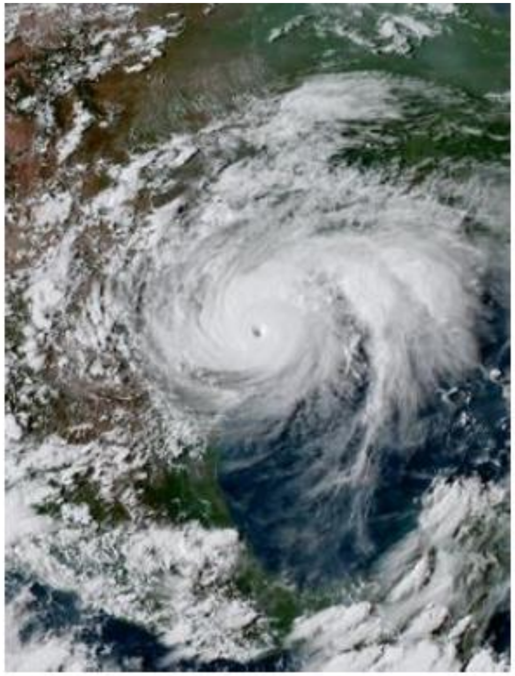


Photo Credits: NASA, CNN, Wikipedia, National Geographic

Statistical Methodology

- Annual maxima follow GEV:

$$\Pr\{Y_t \le y\} = \exp\left[-\left\{1 + \xi\left(\frac{y - \eta_t}{\tau_t}\right)\right\}_+^{-1/\xi}\right].$$

- Assume η_t and $\log\tau_t$ are linear functions of SST_t (Gulf of Mexico annual mean SST in year t) and $CO2_t$ (global mean CO_2 in year t).

- AIC chooses model:

$$\begin{aligned}\eta_t &= \theta_1 + \theta_4 SST_t + \theta_5 CO2_t,\\ \log\tau_t &= \theta_2 + \theta_6 SST_t,\\ \xi &= \theta_3.\end{aligned}$$

Parameter Estimates

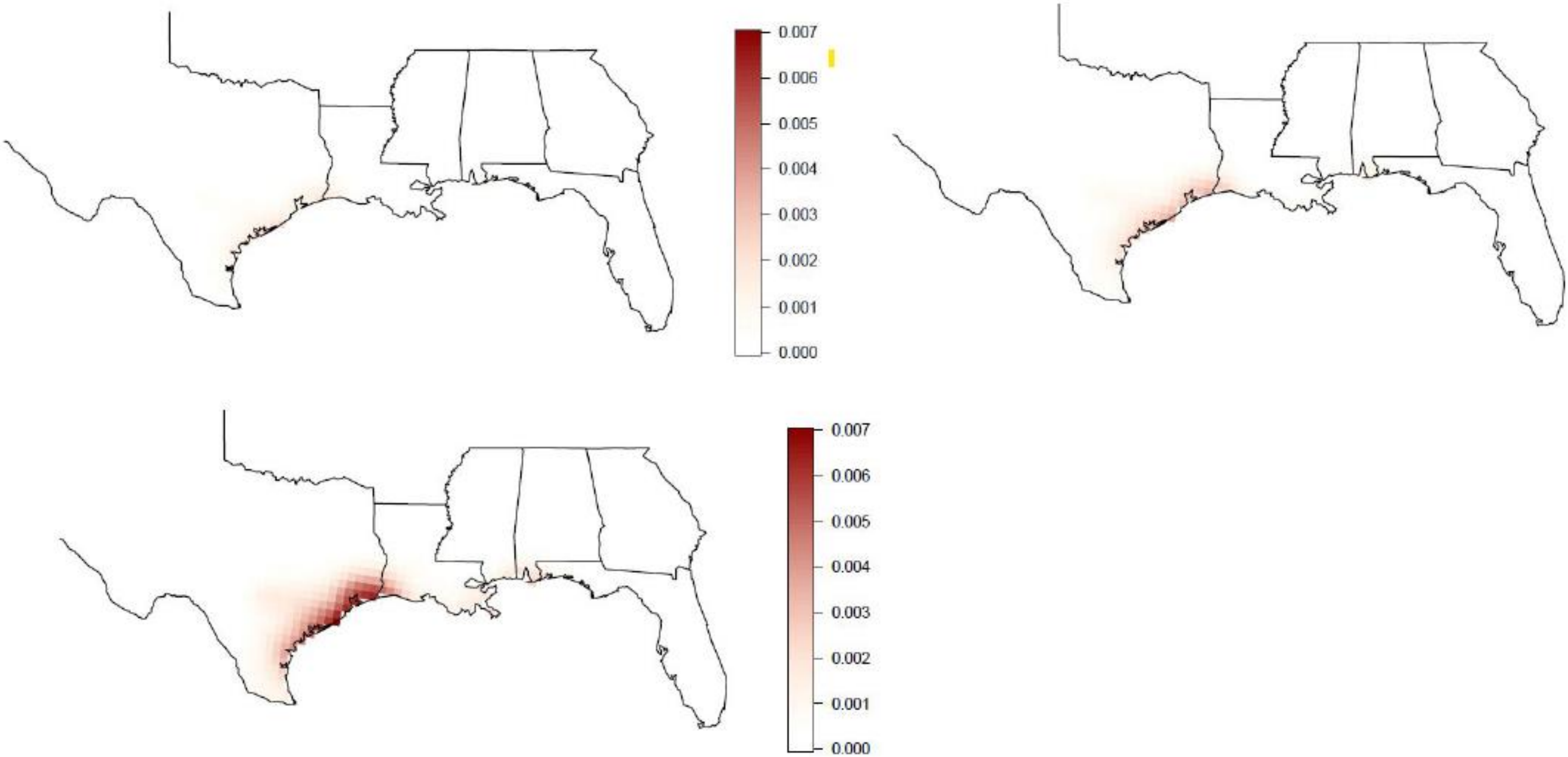
Parameter	Estimate	Standard error	t-statistic	p-value
θ_1	4.70	0.29	16.22	0.00
θ_2	0.56	0.13	4.25	0.00
θ_3	0.15	0.09	1.64	0.10
θ_4	3.06	1.49	2.06	0.04
θ_5	1.95	0.82	2.36	0.018
θ_6	1.24	0.50	2.48	0.013

How Can We Extend This to a Spatial Field?

- Full "detection and attribution" not so far attempted, but this follows Russell *et al.* (*Environmetrics*, 2020)
- Precipitation data, 326 stations in 6 states bordering Gulf
- Model $\eta_t(\mathrm{s})$, $\tau_t(\mathrm{s})$, $\xi_t(\mathrm{s})$ in year t at station s:

$$
\begin{aligned}
\eta_t(\mathrm{s}) &= \theta_1(\mathrm{s}) + \theta_2(\mathrm{s}) SST_t, \\
\log \tau_t(\mathrm{s}) &= \theta_3(\mathrm{s}) + \theta_4(\mathrm{s}) SST_t, \\
\xi_t(\mathrm{s}) &= \theta_5(\mathrm{s}),
\end{aligned}
$$

- $\boldsymbol{\theta}(\mathrm{s}) = \begin{pmatrix} \theta_1(\mathrm{s}) & \ldots & \theta_5(\mathrm{s}) \end{pmatrix}^T$ modeled as a 5-dim spatial process based on *co-regionalization* (Wackernagel and many others)
- Two-stage estimation procedure allows also for spatial correlation among individual measurements



Estimated probability that the annual maximum seven-day rainfall event exceeds 70 cm. under three scenarios: low SST (top left); high SST (top right); 2017 SST (bottom). From Russell *et al.* (2020)

Multivariate Spatial Extreme Value Analysis of Reconstructed Coastal Sea-Level Time-Series

Brian White[1] Richard L. Smith[2] Brian Blanton[3] Rick Luettich[4]

May 23, 2022

[1]Ph.D. Student|Department of Statistics & Operations Research|UNC

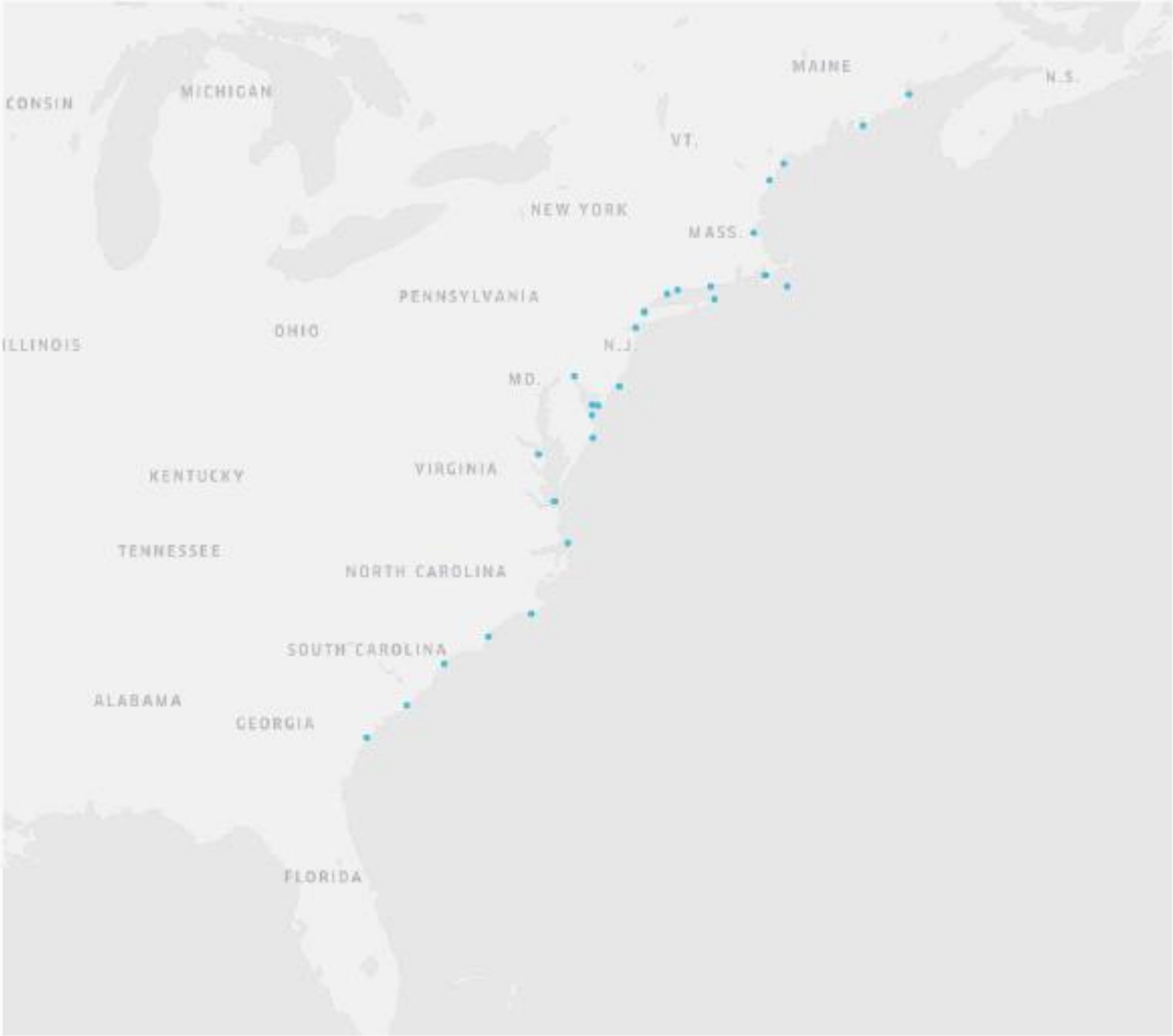
[2]Distinguished Professor|Department of Statistics & Operations Research|UNC

[3]Director of Environmental Initiatives|Renaissance Computing Institute

[4]Distinguished Professor|Department of Earth, Marine & Environmental Sciences|UNC

1. Research aim: To understand how sea-level extremes along the U.S. East Coast have varied over space and time in the past and to make predictions for how these patterns will evolve in the future under different climate change projections.
2. Tools: Oceanography, climate-science, statistical methodology (i.e. extreme value theory. spatial statistics)
3. Data:
   (a) Hourly sea-level time-series taken from NOAA observation stations along the U.S. East Coast over a 40-year period.
   (b) Corresponding model-generated (ADCIRC) reconstruction of historic sea-level time-series.

NOAA Sea-Level Observation Stations



Modeling Questions

1. How do the GEV parameters and r-year return levels depend upon their spatial location?
2. How should the NOAA data be used to validate the ADCIRC reconstruction?
3. How should one incorporate global climatic information?

Idea (Russell et al. 2019): Multivariate spatial extreme value model fit by a 2-stage inference procedure.

1. a. Independently model the yearly (detided daily mean) sea-level maxima at each station using the GEV distribution.
   b. Perform inference via MLE.
2. a. Model the MLE output from stage 1 as a multi-dimensional Gaussian process with measurement error.
   b. Perform inference via MLE.

The output of stage 2 can then be used to spatially interpolate the GEV parameters and return-levels along the coastline via Kriging.

Methods: Latent Process with Measurement Error

Let $Y(\mathbf{s})$ be the yearly (detided daily mean) maximum sea-level at location $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^2$ and assume

$$Y(\mathbf{s}) \sim GEV(\mu(\mathbf{s}), \sigma(\mathbf{s}), \xi(\mathbf{s}))$$

To characterize how the sea-level extremes vary spatially define, at both observed and unobserved locations, the latent Gaussian process

$$\boldsymbol{\theta}(\mathbf{s}) = \boldsymbol{\beta} + \boldsymbol{\eta}(\mathbf{s}) \tag{4}$$

for $\boldsymbol{\theta}(\mathbf{s}) := \left(\mu(\mathbf{s}), \log(\sigma(\mathbf{s})), \xi(\mathbf{s})\right)^t$.

Here $\boldsymbol{\beta}$ is a vector of mean parameter values over $\mathcal{D}$ and $\boldsymbol{\eta}(\mathbf{s})$ a vector of spatially correlated random effects.

Methods: Latent Process with Measurement Error

The spatially correlated random effects are defined by the relation

$$\boldsymbol{\eta}(\mathbf{s}) := A\boldsymbol{\delta}(\mathbf{s}) \tag{5}$$

where A is a lower-triangular matrix and $\boldsymbol{\delta}(\mathbf{s})$ is a vector of independent second-order stationary Gaussian processes with mean 0 and covariance function

$$Cov(\delta_i(\mathbf{s}), \delta_i(\mathbf{s}')) = \exp\left(\frac{-||\mathbf{s} - \mathbf{s}'||}{\rho_i}\right) \tag{6}$$

for $\mathbf{s}, \mathbf{s}' \in \mathcal{D}$ where $\rho_i > 0$ is the range parameter.

Methods: Latent Process with Measurement Error

For NOAA station $l \in \{1, ..., 26\}$, let $\hat{\boldsymbol{\theta}}(\mathbf{s}_l)$ be the point-wise MLE for the GEV distribution associated with $Y(\mathbf{s}_l)$. We assume that

$$\hat{\boldsymbol{\theta}}(\mathbf{s}_l) = \boldsymbol{\theta}(\mathbf{s}_l) + \epsilon(\mathbf{s}_l) \tag{7}$$

where $\epsilon(\mathbf{s}_l)$ is estimation error that is independent of $\boldsymbol{\eta}$.

Thus, the latent process with measurement error at station l is

$$\begin{aligned} \hat{\boldsymbol{\theta}}(\mathbf{s}_l) &= \boldsymbol{\theta}(\mathbf{s}_l) + \epsilon(\mathbf{s}_l) \\ &= \boldsymbol{\beta} + \boldsymbol{\eta}(\mathbf{s}_l) + \epsilon(\mathbf{s}_l) \\ &= \boldsymbol{\beta} + A\boldsymbol{\delta}(\mathbf{s}_l) + \epsilon(\mathbf{s}_l) \end{aligned} \tag{8}$$

Further, assume that

$$\boldsymbol{\epsilon} := \left(\epsilon(\mathbf{s}_1), \ldots, \epsilon(\mathbf{s}_{26})\right)^t \sim \mathcal{N}_{78}(\mathbf{0}, W) \tag{11}$$

where W is unknown and estimated via a regularized non-parametric bootstrap procedure:

$$W_{tap} := W_{bs} \circ T_{tap}(\lambda) \tag{12}$$

where W_{bs} is the non-parametric bootstrap estimate of W and $T_{tap}(\lambda)$ is a taper matrix with range parameter $\lambda > 0$

Thus,

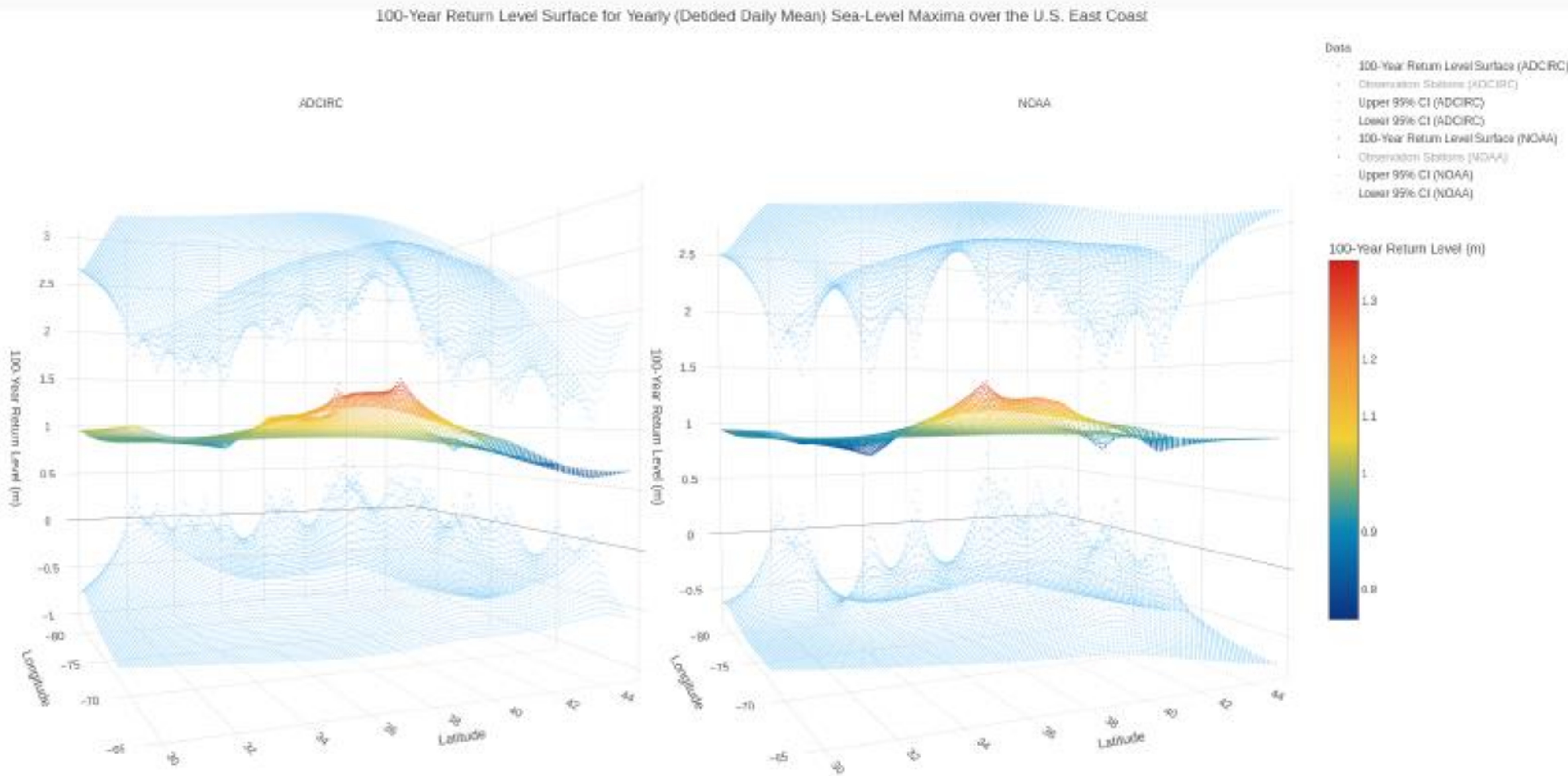
$$\hat{\mathbf{\Theta}} = \mathbf{\Theta} + \epsilon \tag{14}$$

and hence

$$\hat{\mathbf{\Theta}} \sim \mathcal{N}_{78}(\mathbf{1}_{26} \otimes \beta, \mathbf{\Sigma}_{A,\rho} + W_{tap}) \tag{15}$$

Therefore, given $\hat{\mathbf{\Theta}}$ (i.e. the output from the 1st stage of inference) and W_{tap}, we can obtain $\hat{\beta}$, $\hat{\rho}$ and $\hat{A}$ via MLE.

Preliminary Results: 100-Year Return Level Surface for Yearly (Detided Daily Mean) Sea-Level Maxima



Summary and Conclusions

- Univariate time series:
  - GEV model allows global variables as covariates; attribution of local extreme events may then follow from studies of global attribution
  - Identifying the right global variable is a challenge; not clear to me that GMST is the way to go
  - The zero-probability problem is alleviated by Bayesian analyses but it's not going away
  - Other important issues such as how to integrate observational with model data and how to combine different models
  - Maybe need to think of a different formulation...

- Spatio-temporal datasets:
  - Spatial dependence at two levels — model μ, σ, ξ as spatial processes but still need to consider spatial dependence of the errors
  - The W-matrix approach is one way to do it but there are others (e.g. max-stable processes)
  - Spatial-temporal analysis allow us to combine different variables in one analysis (may avoid zero-probability problems) and also addresses selection effects