Extreme value statistics born out of domains of attraction

Cláudia Neves

(J. Einmahl, J. El-Methni, A. Ferreira, S. Girard, L. de Haan, P. Jonathan, A. Klein Tank, E. Konzen, C. Zhou)

Energy Forecasting Innovation Conference
24 May 2022

Engineering and Physical Sciences Research Council

Man can believe the impossible, but man can never believe the improbable...

[Oscar Wilde, Intentions, 1891]

For many phenomena records must be broken in the future, so if a design is based on the worst case of the past, then we are note really prepared for the future.

[Preface of the EVT & Applications conf. proceedings, May 1993]

Outline

Motivation

Extreme Value Theory (EVT)

- The goal of EVT is to define, characterise and estimate hallmark features of extreme events and of the possible dependence between them.
- We refer to an extreme event as an event that is so rare that may have never been observed in the past.

Two related estimation problems:

1. that of a probability of an extreme event;

Motivation

Extreme Value Theory (EVT)

- The goal of EVT is to define, characterise and estimate hallmark features of extreme events and of the possible dependence between them.
- We refer to an extreme event as an event that is so rare that may have never been observed in the past.

Two related estimation problems:

1. that of a probability of an extreme event;
2. estimation of a large value is exceeded with some pre-assigned probability (near zero)

In 2009, the UK was battered by heavy rainfall with the record high of 316.4mm in Cumbria.

Two related questions

Question 1

- What are the chances of a repeat of such an extreme event?

Two related questions

Question 1

- What are the chances of a repeat of such an extreme event?
- It was determined that is was an event with a 1/500 occurrence probability (in a given year).

In 2015, Storm Desmond ravaged the UK, with heavy rain and strong winds

(Source: Christopher Furlong/Getty Images)

Two related questions

Question 2

- On December 5th, 2015, Storm Desmond broke the UK's 24-hour rainfall record with 341.4mm of rainfall in Cumbria
- How high must be a flooding barrier in order to withstand a 1 in 10,000-year event, i.e. an event with occurrence probability of $1/10,000$ in a given year?

Firstly, tail probability estimation

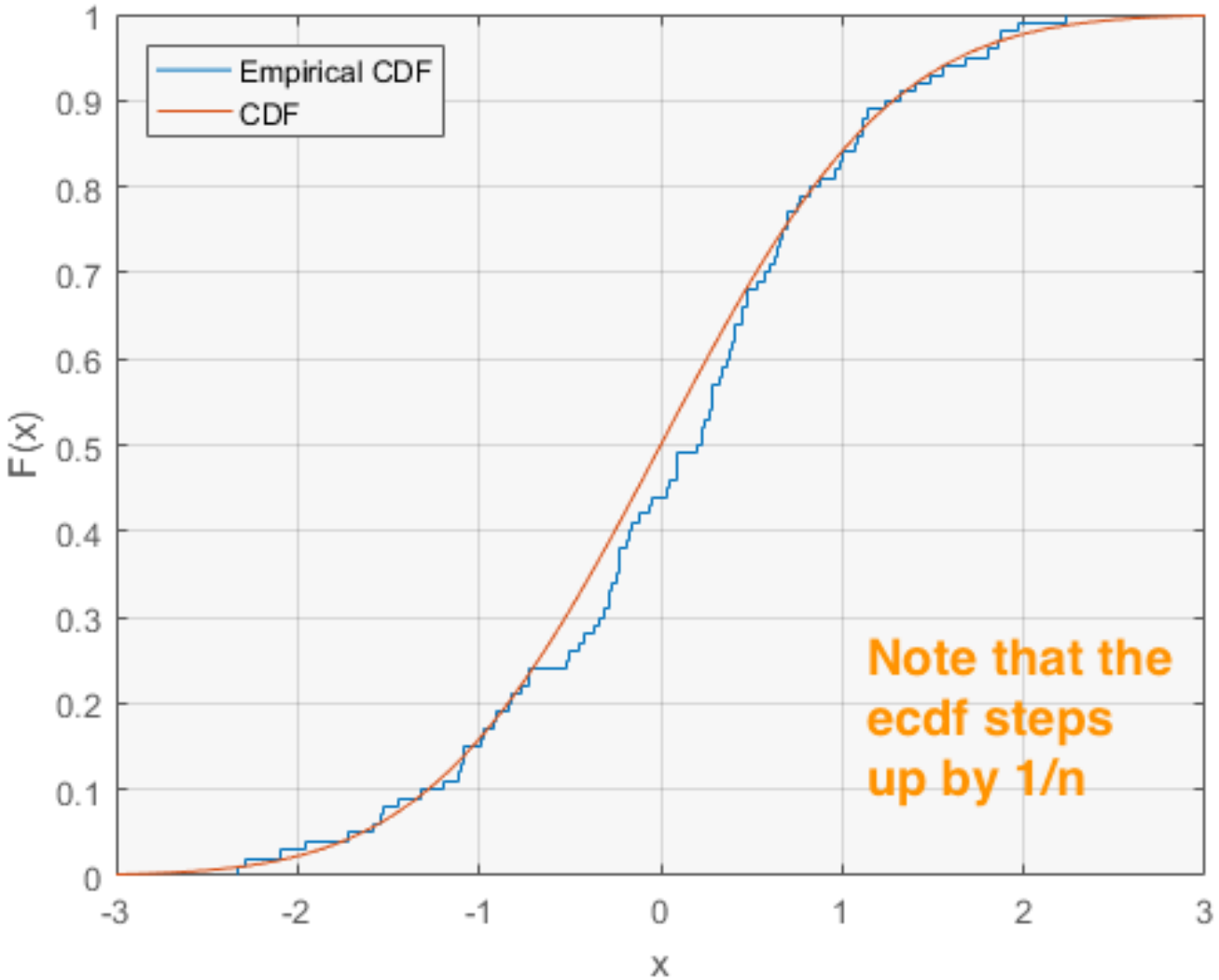
- We need to formulate this first question into an estimation problem.
- We wish to estimate the exceedance (or tail) probability $P(X > x) = 1 - F(x)$ of an already large observation x.

Empirical distribution function (edf):

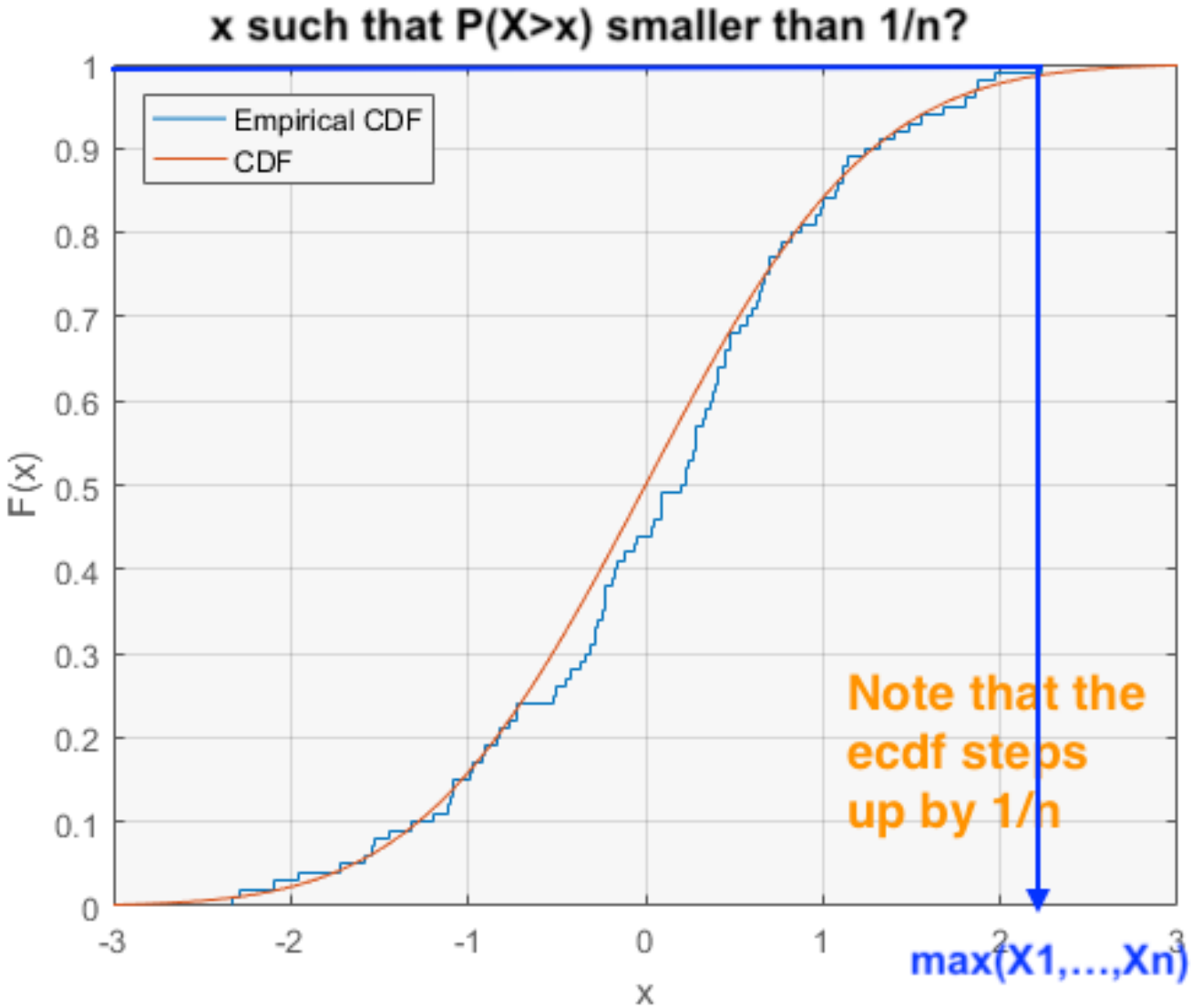
If $(X_1, X_2, \ldots, X_n)$ consists of random a sample of n days of rainfall distributed according to F, a consistent estimator for $F(x)$ is the empirical distribution function:

$$1 - F_n(x) = \frac{\#\{X_i \text{ greater than } x\}}{n} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i > x\}}$$

Firstly, tail probability estimation



Secondly, via inverse (quantile) function



Extreme Value Theory (EVT)

The answer is, of course, in EVT!

This sort of estimation problems are tractable within the framework of extreme value theory.

- Notably, this is due to the Extreme Value theorem, analogous to the Central Limit theorem.

Extreme Value (types) Theorem

Fisher & Tippett (1928), Gnedenko (1943), de Haan (1970)

If there exist constants $a_n > 0, b_n\,(n = 1, 2, \ldots)$ such that

$$\lim_{n\to\infty} P\Big\{\frac{\max(X_1, \ldots, X_n) - b_n}{a_n} \le x\Big\} = \lim_{n\to\infty} F^n(a_n x + b_n) = G(x),$$

G non-degenerate, for all continuity points x, then

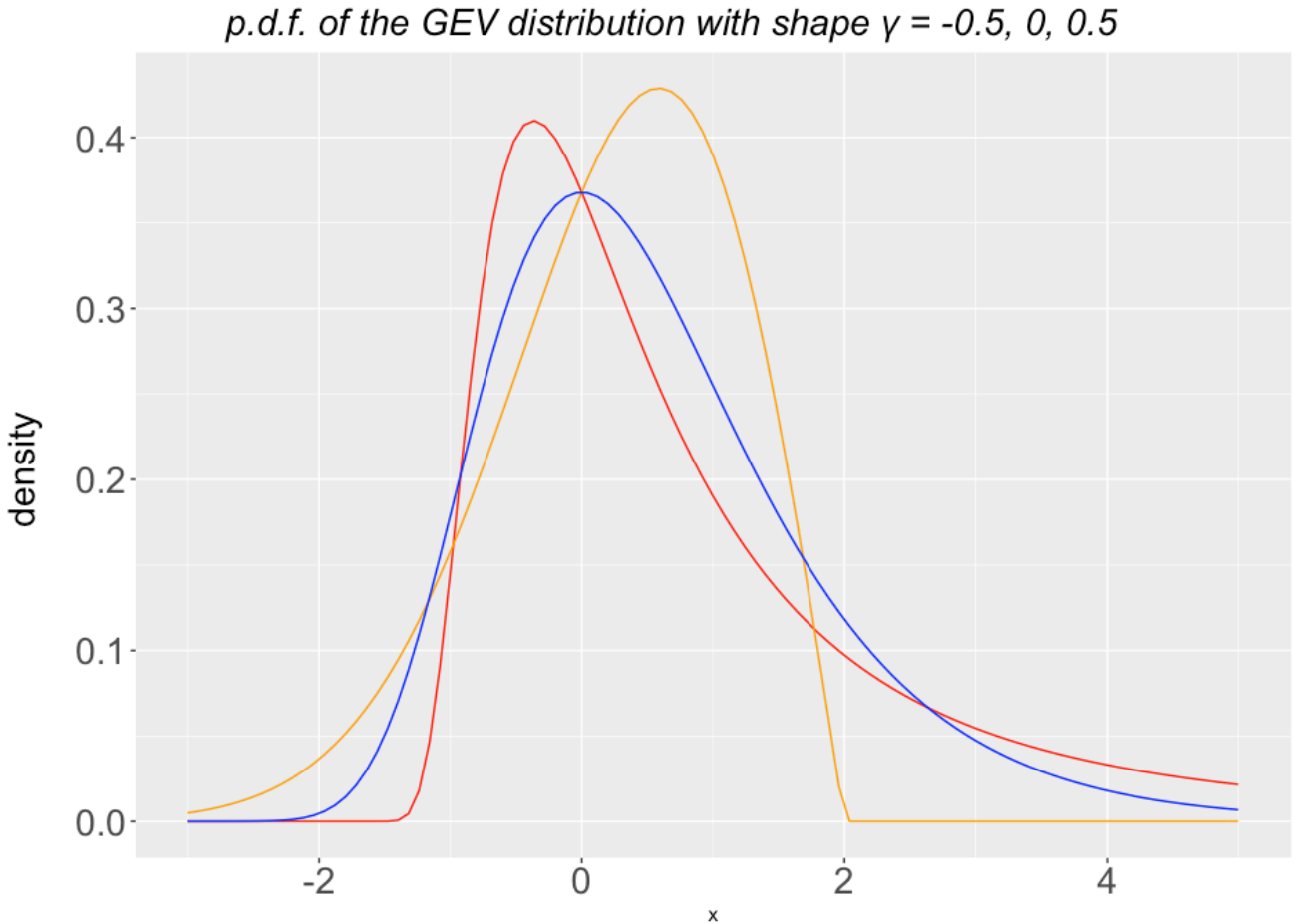
$$G(x) = G_\gamma(x) := \exp\big\{-(1+\gamma x)^{-1/\gamma}\big\},$$

for some $\gamma \in \mathbb{R}$ and all x with $1 + \gamma x > 0$.

G_γ is the Generalised Extreme Value (GEV) distribution

This introduces the extreme value index $\gamma \in \mathbb{R}$

Extreme value distributions



Probability density function $dG_\gamma(x)/dx$ for: $\gamma = 1/2$ (Fréchet; *red*), $\gamma = 0$ (Gumbel; *blue*) and $\gamma = -1/2$ (Weibull; *orange*)

Extreme value distributions

These graphs of probability densities of G_γ suggest that:

- If $\gamma > 0$, then $G_\gamma^{\leftarrow}(1) = \infty$.
- If $\gamma = 0$, then $G_\gamma^{\leftarrow}(1) = \infty$.
- If $\gamma < 0$, then $G_\gamma^{\leftarrow}(1) = -\frac{1}{\gamma}$, meaning that no observations beyond $-1/\gamma$ are possible.

Domain of attraction

A distribution function F is in the domain of attraction of G_γ for some $\gamma \in \mathbb{R}$ if the normalised maximum $X_{n:n}$ of independent observables $X_1, X_2, \ldots, X_n, \ldots$ from this F converges to the GEV distribution function G_γ, as $n \to \infty$.

Notation: $F \in \mathcal{D}(G_\gamma)$

Recall the graphs of probability densities of G_γ.

We have a similar behaviour for F:

- If $\gamma > 0$, then $F^{\leftarrow}(1) = \infty$.
- If $\gamma < 0$, then $F^{\leftarrow}(1) < \infty$, meaning all observations are bounded from above.
- If $\gamma = 0$, then $F^{\leftarrow}(1) = \infty$ or $F^{\leftarrow}(1) < \infty$.

An equivalent formulation of the EVT for exceedances

- $F \in \mathcal{D}(G_\gamma)$ for some $\gamma \in \mathbb{R}$ *iff*

$$\lim_{n\to\infty} -n \log F(a_n x + b_n) = -\log G_\gamma(x) = (1+\gamma x)^{-1/\gamma},$$

all $x \in \mathbb{R}$ such that $1+\gamma x > 0$.

- With $a(u) := a_{[u]}$ and $b(u) := b_{[u]}$,

$$\lim_{u\to\infty} u\Big(1 - F\big(a(u)x + b(u)\big)\Big) = (1+\gamma x)^{-1/\gamma}, \qquad (1)$$

for some $\gamma \in \mathbb{R}$ and all x such that $x > 0$ and $1+\gamma x > 0$.

Take $u = n/k \to \infty$ in the extreme value condition (1).

Peaks over Threshold (POT) approach

Taking $u = n/k \to \infty$ in the extreme value condition (1):

$$\frac{n}{k}\left(1 - F\left(a\left(\frac{n}{k}\right)x + b\left(\frac{n}{k}\right)\right)\right) \approx (1+\gamma x)^{-1/\gamma}.$$

Writing $t := a\left(\frac{n}{k}\right)x + b\left(\frac{n}{k}\right)$, we get

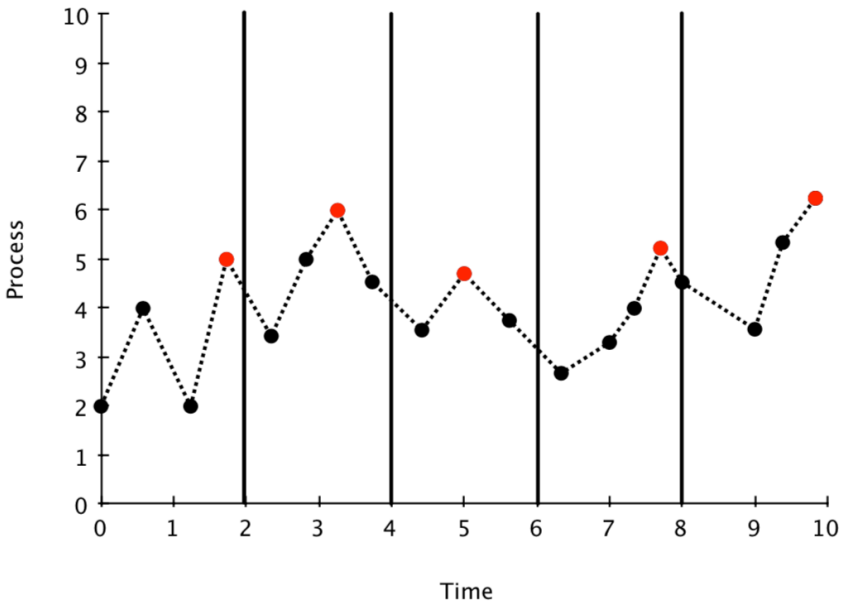
$$1 - F(t) \approx \frac{k}{n}\left(1 + \gamma\,\frac{t - b(n/k)}{a(n/k)}\right)^{-1/\gamma}, \qquad \text{as } t \uparrow F^{\leftarrow}(1).$$

This approximation is valid for any t large and can even be used for $t > X_{n:n}$.

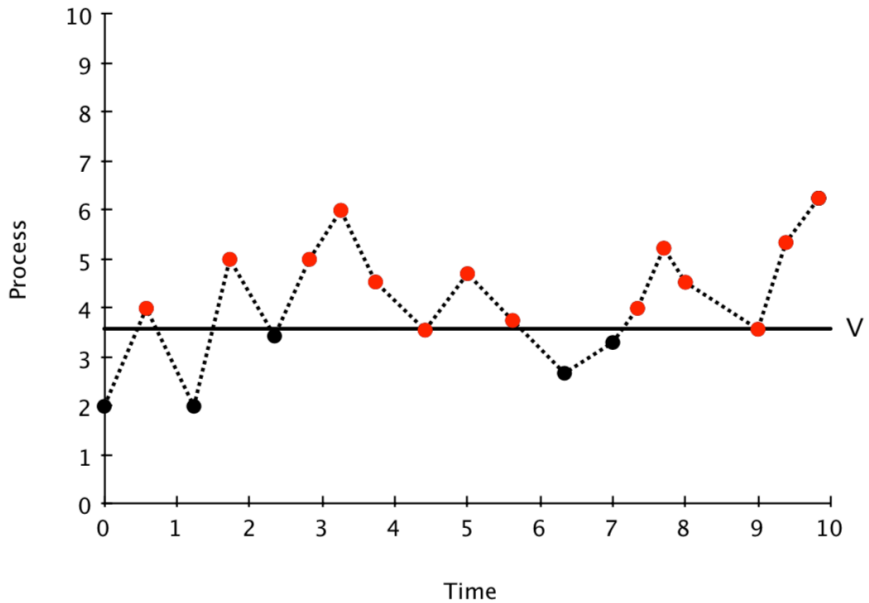
➠ Extrapolation beyond the range of the available observations

Block Maxima *versus* Peaks over Threshold

BM→GEV



POT→GPD



Estimation

Typically we wish to estimate γ, $a(n/k)$ and $b(n/k)$:

- Consider intermediate (and extreme) order statistics $X_{n:n} \geq X_{n-1:n} \geq \ldots \geq X_{n-k:n}$, where $k = k(n) \in (0, n]$, such that $k \to \infty$ and $k/n \to 0$, as $n \to \infty$.
- Estimators for the above *unknowns* are often asymptotically linear functionals of these order statistics through $\int_{X_{n-k:n}}^{x^*} \psi(s/X_{n-k:n}) dF_n(s)$, where $F_n(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_{i:n} \leq x\}}$

Semiparametric estimators for an extreme quantile

The aim now is to estimate x_p such that $p = p_n < 1/n$.
Let $k = k(n)$ be an intermediate sequence and suppose $np_n = o(k)$.

- $F \in \mathcal{D}(G_\gamma)$ for some $\gamma \in \mathbb{R}$,

$$\hat{x}_{p_n} := \hat{b}\left(\frac{n}{k}\right) + \hat{a}\left(\frac{n}{k}\right) \frac{\left(\frac{k}{np_n}\right)^{\hat{\gamma}} - 1}{\hat{\gamma}}$$

- $F \in \mathcal{D}(G_\gamma)$ for some $\gamma \in \mathbb{R}$, including $\gamma = 0$ and $F^{\leftarrow}(1) < \infty$. Set $\gamma_+ = \max(0, \gamma)$,

$$\hat{x}_{p_n} := X_{n,n} + \left(\frac{k}{np_n}\right)^{\hat{\gamma}_+} \left\{ X_{n-k,n} - \frac{2^{\hat{\gamma}_+}}{\log 2} \int_{1/2}^{1} s^{\hat{\gamma}_+} X_{n-[2ks],n} \, \frac{ds}{s} \right\}$$

(El-Methni, Girard and CN, 2022+)

Semiparametric estimators for an extreme quantile

The aim now is to estimate x_p such that $p = p_n < 1/n$.
Let $k = k(n)$ be an intermediate sequence and suppose $np_n = o(k)$.

- $F \in \mathcal{D}(G_\gamma)$ for some $\gamma \in \mathbb{R}$,

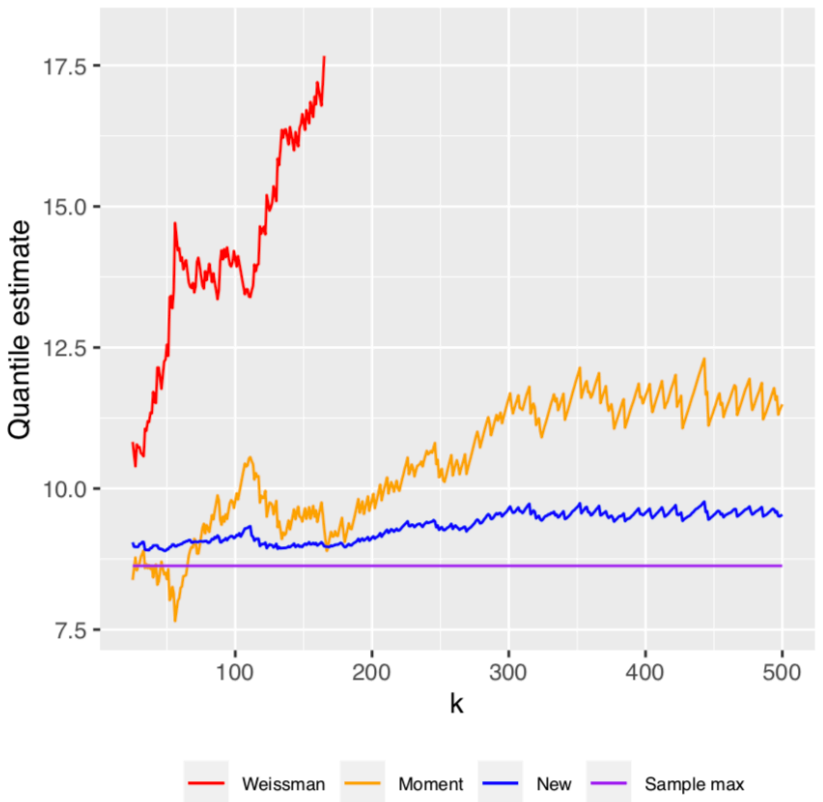
$$\hat{x}_{p_n} := \hat{b}\left(\frac{n}{k}\right) + \hat{a}\left(\frac{n}{k}\right) \frac{\left(\frac{k}{np_n}\right)^{\hat{\gamma}} - 1}{\hat{\gamma}}$$

- General right endpoint estimator w.r.t. $\gamma \leq 0$ and $F^{\leftarrow}(1) < \infty$,

$$\hat{x}_0 := X_{n,n} + X_{n-k,n} - \frac{1}{\log 2} \sum_{i=0}^{k-1} \log\left(1 + \frac{1}{k+i}\right) X_{n-k-i,n}$$

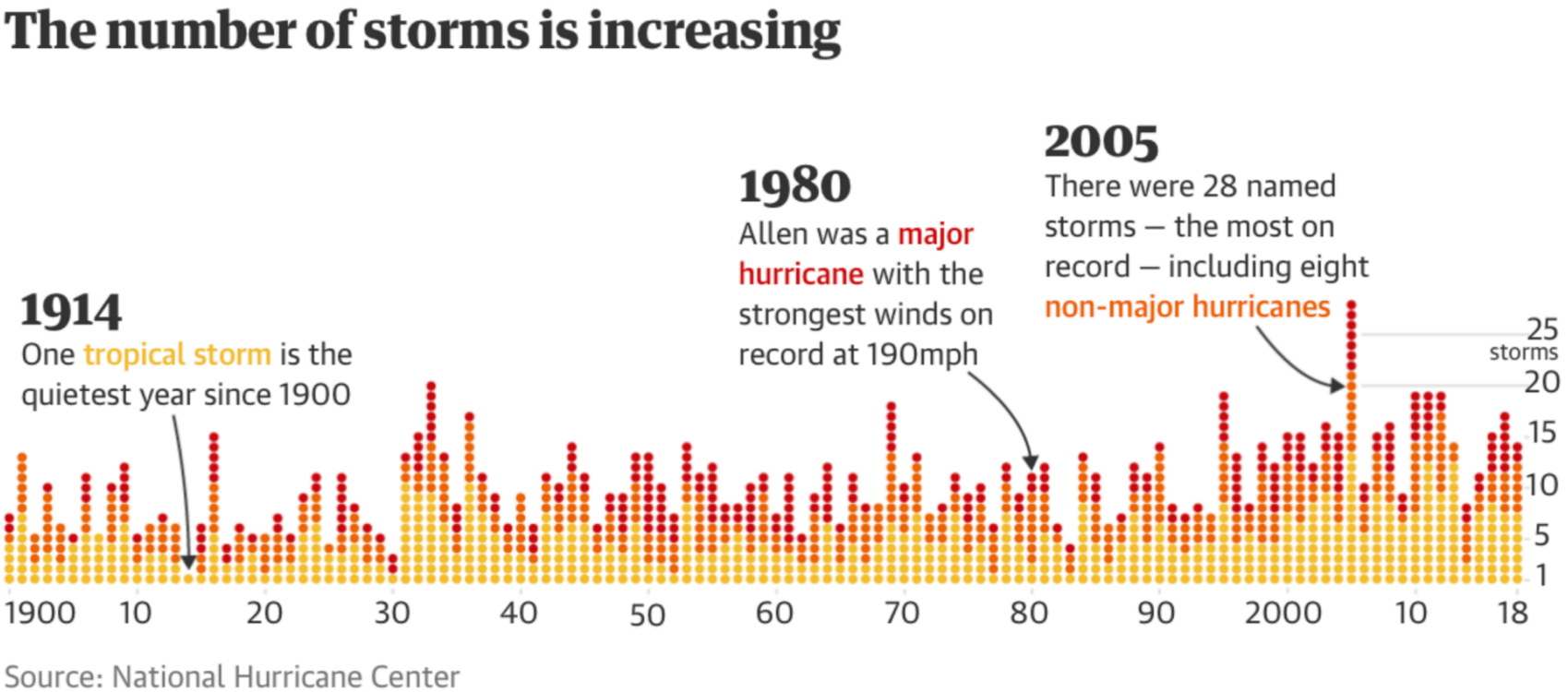
(Fraga Alves and CN, 2014)

Boeing 747 taxiway deviations at Ted Stevens Anchorage International Airport



Data measurements (in ft) collected from 9/24/2000 to 9/27/2001 ($n = 4900$)

Part 2 - Motivation



Space-time scedasis

Consider independent random vectors $(X_{i,1}, X_{i,2}, \ldots X_{i,m})_{i=1,2,\ldots}$.

- For $i = 1, \ldots, n$, $j = 1, \ldots, m$, the marginal distribution functions $F_{i,j}(x) = P\{X_{i,j} \leq x\}$ are tail equivalent, i.e., we assume that there exists a distribution function $F_0 \in \mathcal{D}(G_\gamma)$ such that for all i and j

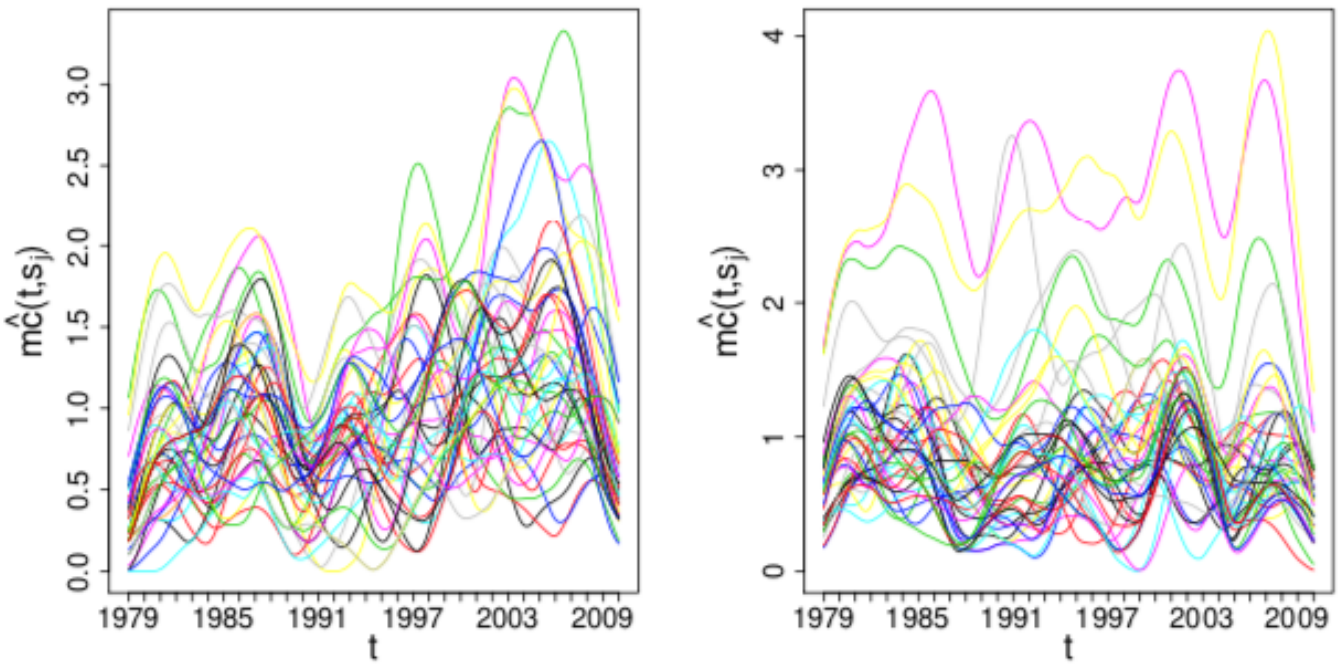
$$\lim_{x \uparrow x_0} \frac{P\{X_{i,j} > x\}}{P\{X_0 > x\}} = c\Big(\frac{i}{n}, j\Big),$$

  where $x_0 = F_0^{\leftarrow}(1)$ and $c(\cdot, j)$ is a positive continuous function for each $j = 1, 2, \ldots, m$.

- The scedasis c is made unique through $\sum_{j=1}^m C_j(1) = 1$, where

$$C_j(t) := \int_0^t \frac{c(u,j)}{m}\, du$$

Little (temporal) scedasis $c(\cdot, j)$

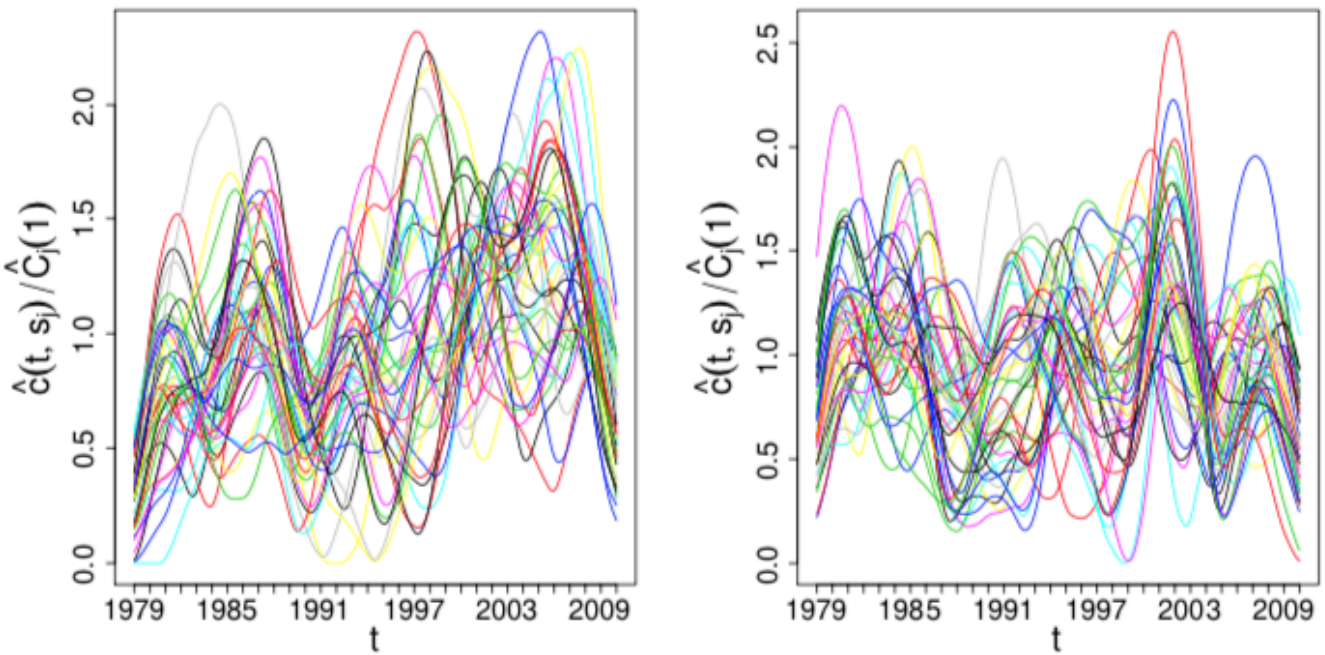


Daily rainfall data, reanalysis data for $m = 44$ stations across the UK: warm season (left) and cold season (right). Estimation with $k = 2000$.

Interpretation of space-time trend

- For all $i = 1, \ldots, n$, $j = 1, \ldots, m$, there is a unique (random) threshold $X_{N-k:N}$;
- At location $j = 1, \ldots, m$, the integrated scedasis $C_j(t)$ gives the accumulated frequency of threshold exceedances up to time $t = i/n$;
- For all i, j, the EVI $\gamma \in \mathbb{R}$, determines how extremes amplify.

Normalised (temporal) scedasis $c(\cdot, j)$



Daily rainfall data (reanalysis data) for $m = 44$ stations across the UK: warm season (left) and cold season (right).Estimation with $k = 2000$.

Scedasis embedding in threshold selection

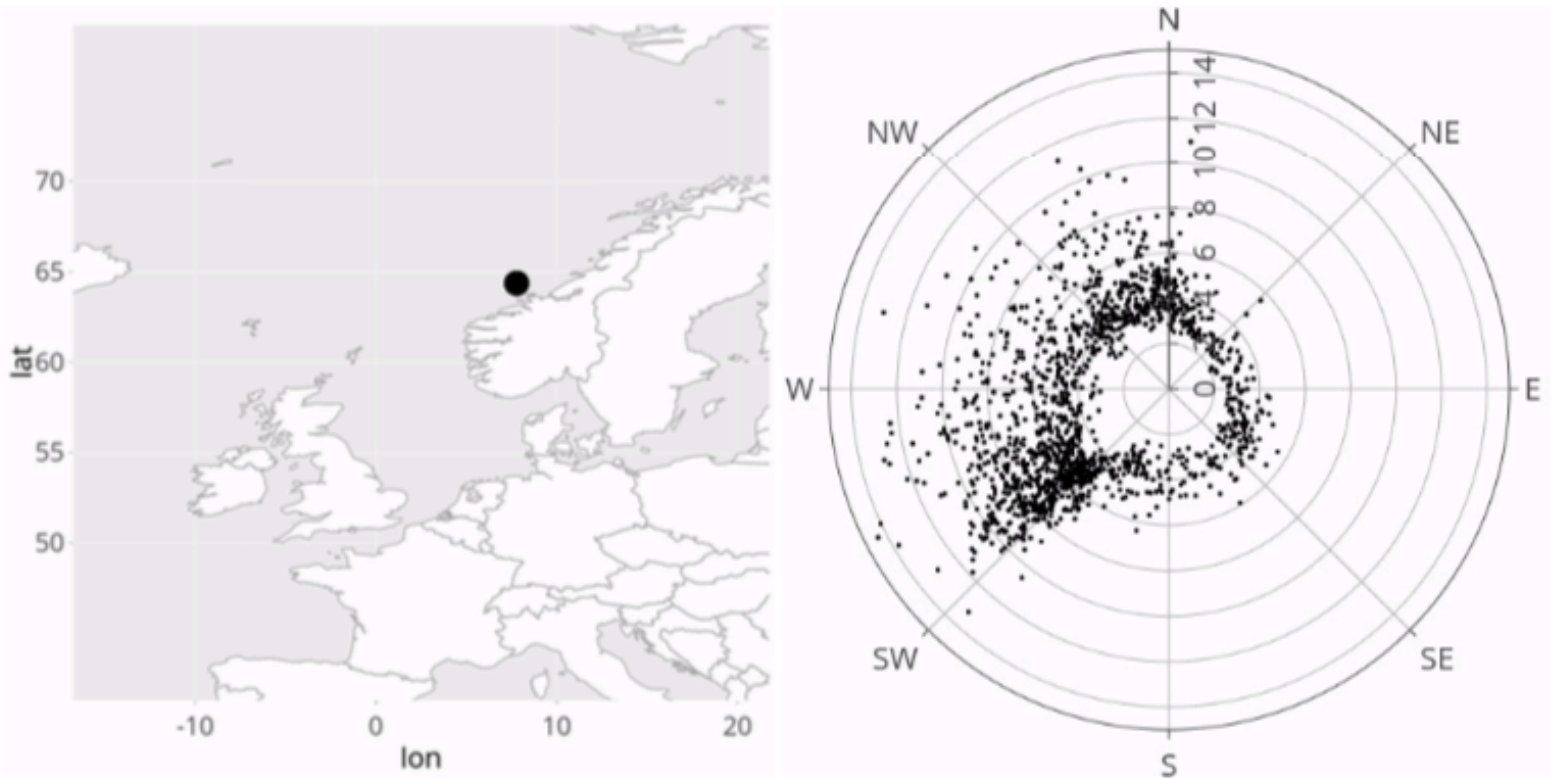
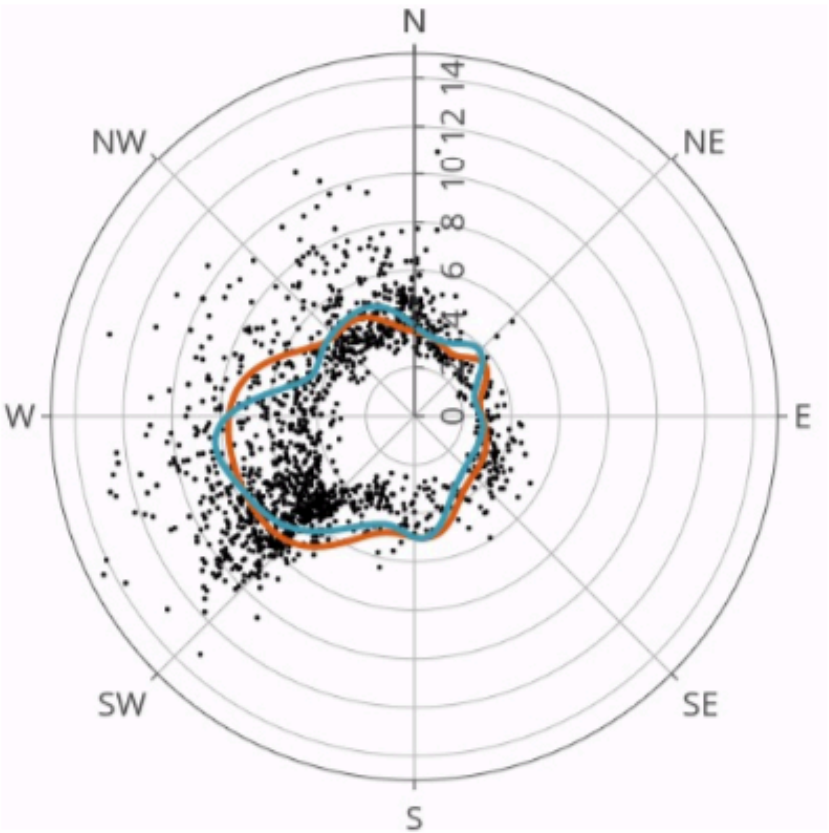


FIGURE 1 Left: map showing the North Sea location. Right: polar scatter-plot of H_S^{sp} on the direction θ from which waves propagate; radial scale of H_S^{sp} is in metres.

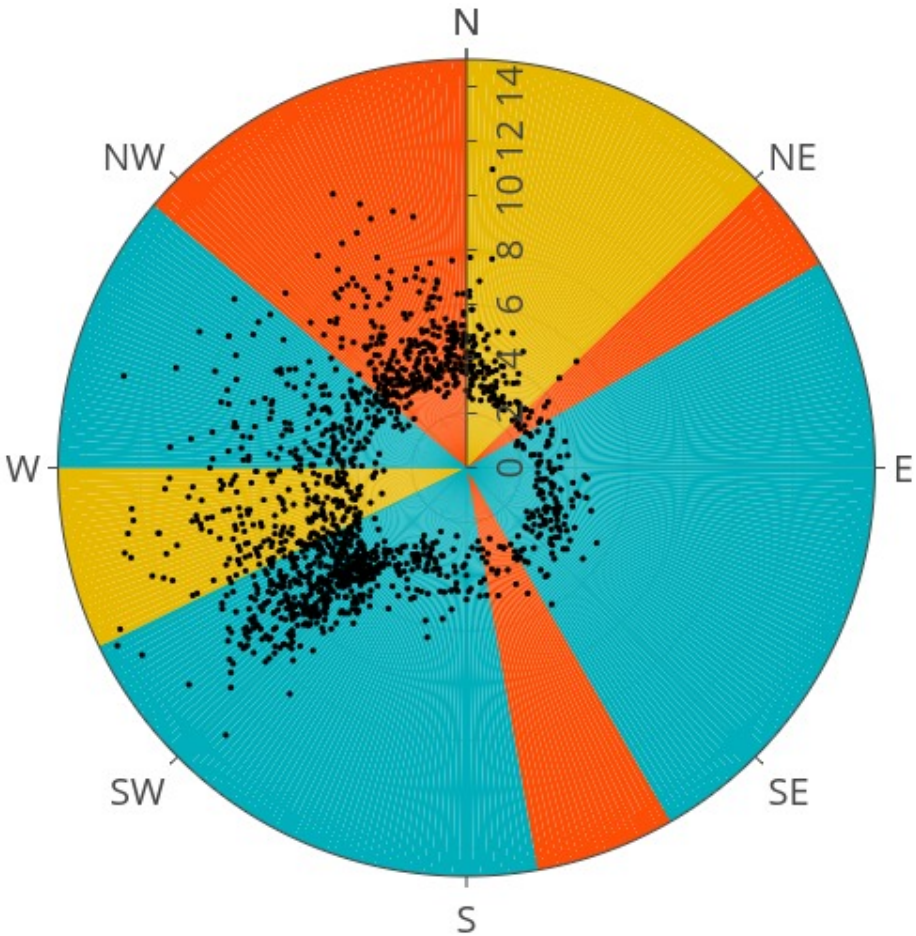
Scedasis embedding in threshold selection

$$X^{(h)}_{n-[c(\theta)k],n} - X_{n-k,n} = a_\theta\Big(\frac{n}{c(\theta)k}\Big)\frac{(c(\theta))^{-\gamma(\theta)}-1}{\gamma(\theta)} + o_p\Big(a_\theta\Big(\frac{n}{c(\theta)k}\Big)\Big),$$

uniformly in θ, with $c(\theta) > 0$ such that $\int_{\mathcal{S}}(1/|\mathcal{S}|)\,c(\theta)\,d\theta = 1$. The interpretation of $c(\theta)$ is that of the rate of extremes per directional window of width $h > 0$.



Scedasis embedding in threshold selection



Polar plot with hypotheses testing decisions at the $\alpha = 5\%$ significance level.

Orange: $\gamma(\theta) = 0, x_0 = \infty$ Yellow: $\gamma(\theta) = 0, x_0 < \infty$ Blue: $\gamma(\theta) < 0$.

Integrated scedasis – basis for inference

Theorem 3 (Einmahl, Ferreira, de Haan, CN and Zhou, 2022):
With a Skorokhod construction, for some $\gamma \in \mathbb{R}$,

$$\max_{1\le j\le m} \sup_{0<t\le 1} \Bigg| \sqrt{k}\Big(\widehat{C}_j(1) - C_j(t)\Big) - \Big\{ W_j\big(1, C_j(t)\big) - C_j(t) \sum_{r=1}^{m} W_r\big(1, C_r(1)\big) \Big\} \Bigg| \xrightarrow[n\to\infty]{P} 0,$$

where $\big(W_1(1,\cdot), \ldots, W_m(1,\cdot)\big)$ is a Gaussian vector-valued random field, with each $W_j(1,\cdot)$ being a standard Wiener process such that

$$Cov\Big(W_{j_1}\big(1, C_{j_1}(t_1)\big), W_{j_2}\big(1, C_{j_2}(t_2)\big)\Big) = \frac{1}{m} \int_0^{t_1\wedge t_2} R_{j_1,j_2}\big(c(u,j_1),\, c(u,j_2)\big)\, du.$$

Two tests

1. $H_{0,j}: C_j(t) = tC_j(1)$ for $0 \le t \le 1$, i.e., the scedasis $c(\cdot, j)$ is constant over time. Under $H_{0,j}$, the limit in distribution of the process $\{\sqrt{k}(\widehat{C}_j(t) - t\widehat{C}_j(1)\}_{0\le t\le 1}$ is essentially Brownian bridge.
2. $H_0: C_j(1) = \frac{1}{m}$ for all $j = 1, 2, \ldots, m$, i.e., the total scedasis is constant over the various locations. We perform the test by checking whether the limit vector (in distribution) of

$$\left(\sqrt{k}\left(\widehat{C}_1(1) - \frac{1}{m}\right), \sqrt{k}\left(\widehat{C}_2(1) - \frac{1}{m}\right), \ldots, \sqrt{k}\left(\widehat{C}_m(1) - \frac{1}{m}\right)\right)$$

has mean zero. This will be done via an adapted χ^2-test through appropriate quadratic form of this asymptotically normal random vector.

Application to extreme rainfall

The data:

- The considered data amounts to daily precipitation totals from three regions in North-West Germany, Bremen, Niedersachsen and Hamburg, with available observations from 1931 to 2014;
- There are in total $m = 49$ stations spanning $n = 84$ years;
- Two seasons were considered separately, a cold (winter) season from November to March and a warm (summer) season from May to September;
- Tally of $N = 49 \times 3561$ for winter and $N = 49 \times 3552$ for summer.

Detecting a temporal trend in extremes

Test 1 $\quad H_{0,j} : C_j(t)/C_j(1) = t$ for $0 \leq t \leq 1$

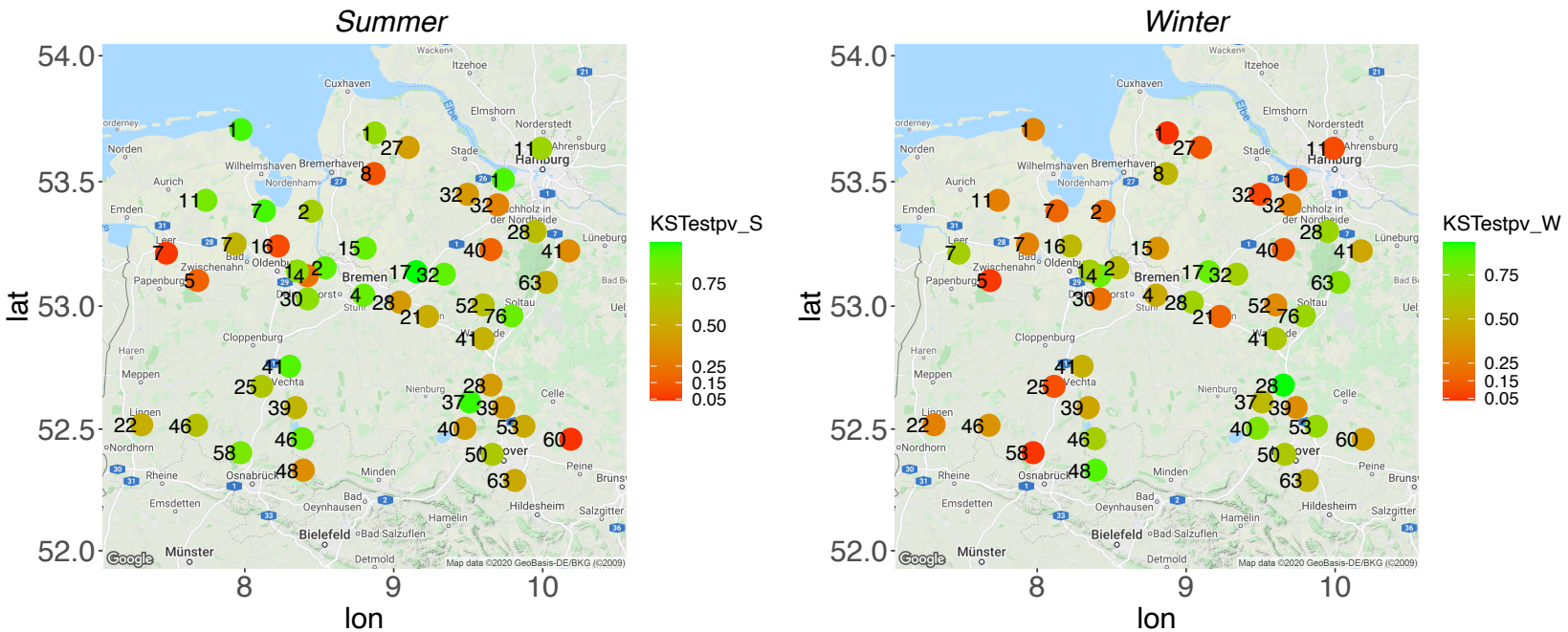
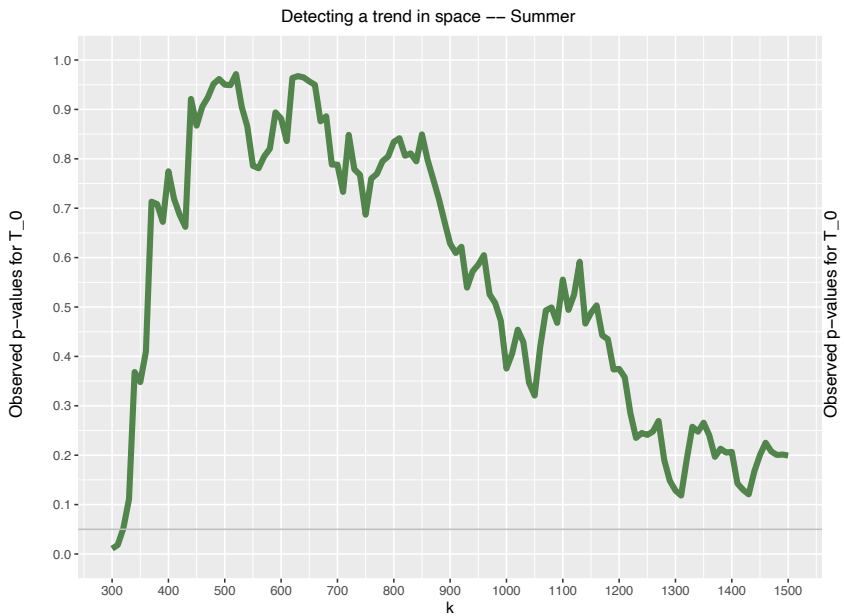
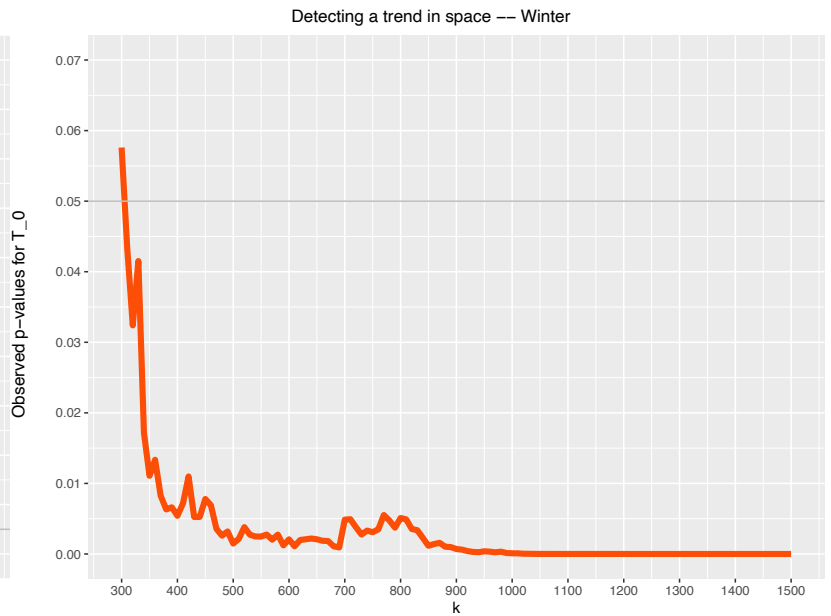


Figure: Tests for homogeneity over time drawing on $k = 1000$. Bonferroni's correction for the nominal level of the test $\alpha^* = 5\%/49 \approx 0.1\%$. Overall, p-values soar in the summer and plunge in the winter at many locations.

Test 2 – Evidence of a trend across space?





Thank you for listening!

References

For a full list of references, please visit:
`https://www.claudianet.co.uk/publications`