Regression for extremes : an application to the prediction of sea levels

Nathan Huet
Department of Environmental Sciences, Informatics and Statistics
Ca' Foscari University of Venice

November 25, 2026

About

21 - 24 **PhD in Statistics**, Telecom Paris, Institut Polytechnique de Paris, *Statistical learning for multivariate and functional extremes*, under the supervision of Stephan Clémençon and Anne Sabourin.

24 - 26 **Postdoctoral researcher**, University Ca' Foscari of Venice, DAIS, *Statistical Methods for Data-centric Environmental Studies*, under the supervision of Ilaria Prosdocimi.

Next year **Postdoctoral researcher**, University of Padua, Geosciences Department, *Analytical derivation of limit statistical distributions for extreme precipitation induced by different atmospheric processes*, under the supervision of Francesco Marra.

Study of Extreme Values

Why? model, predict, understand, anticipate, or manage extreme phenomena such as heavy precipitation, devastating floods, stock market crashes...



Flood in Netherlands, 1953 (photo from *Watersnoodmuseum*).

Extreme Value Theory

Focus: observations outside the mass center of the distribution

Usual assumptions on X a random element

- convergence in distribution of maxima, *i.e.*

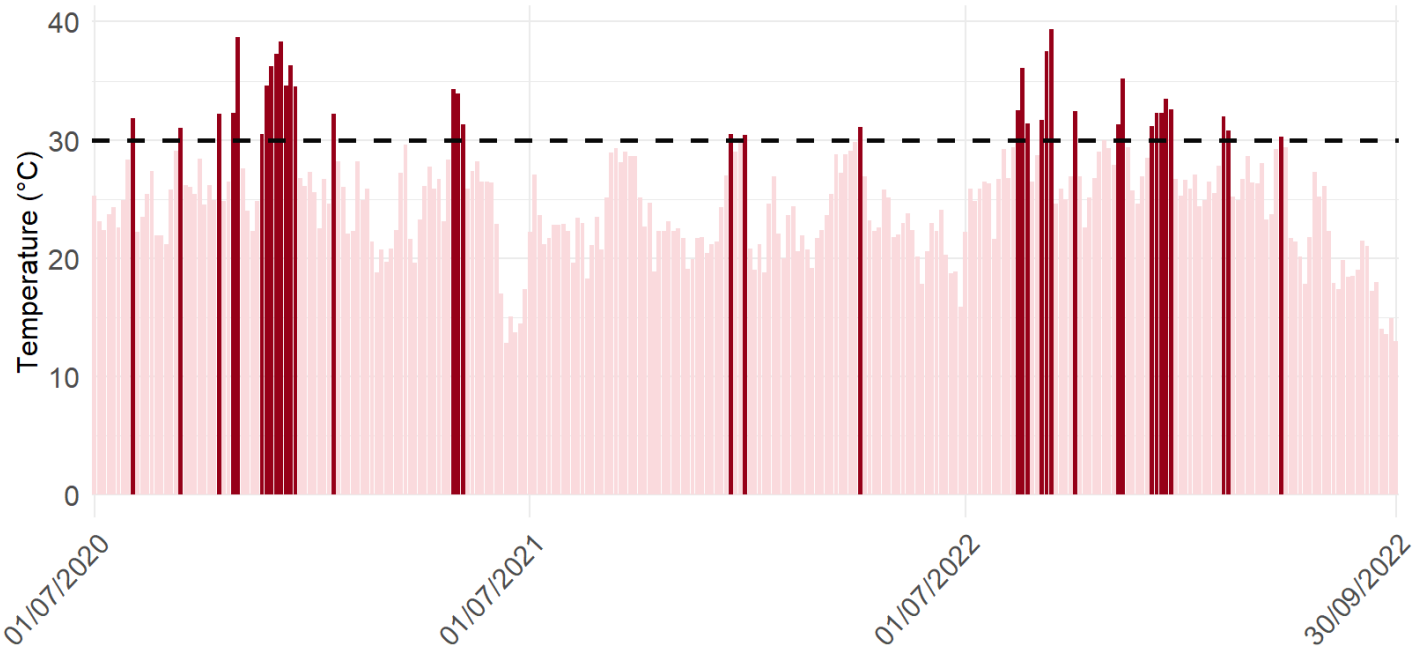
$$\lim_{n \to +\infty} \mathscr{L}\left(\frac{\max_{i=1}^{n} X_i - b_n}{a_n} \right) = \mathscr{L}(Z),$$

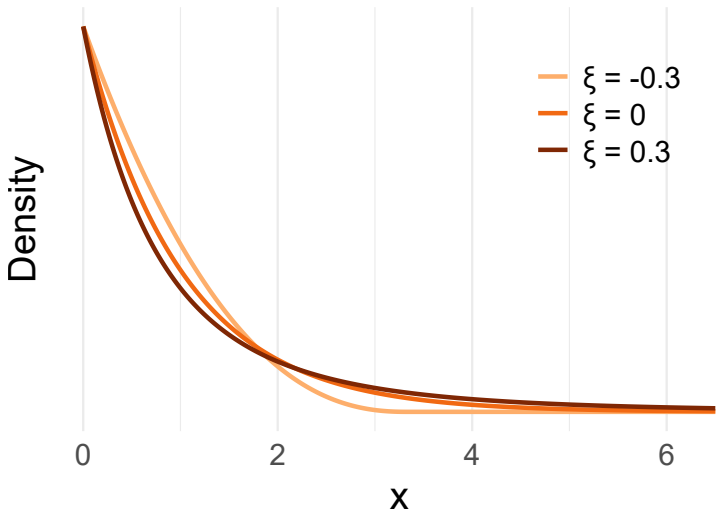
  with $X_i \overset{i.i.d.}{\sim} X$.

- convergence in distribution of excesses, *i.e.*

$$\lim_{t \to +\infty} \mathscr{L}(X/t \mid \|X\| \geq t) = \mathscr{L}(X_\infty).$$

Peaks-over-Threshold





Usual distribution to model exceedances: the Generalised Pareto distribution with cdf

$$F_{GP}(x;\sigma,\xi) = 1 - \left(1 + \frac{\xi x}{\sigma}\right)_+^{-1/\xi},$$

$\sigma > 0,\ \xi \in \mathbb{R}.$

Main research focus

Statistical Learning $\rightarrow$ **Extreme Values**: classical statistical methods focus on the bulk of the distribution
$\Rightarrow$ they need some adaptation to perform well in extreme regions

Theory $\rightarrow$ **Application**: very powerful tools with many guarantees are being developed, while applied work often relies on basic, and sometimes outdated, methods
$\Rightarrow$ some work is needed to bridge the gap between theorists and practitioners

Outline of the presentation: Regression $\rightarrow$ Extreme values $\rightarrow$ Application to sea level prediction

Regression for extremes

joint work with Stephan Clémençon (Télécom Paris) and Anne Sabourin (Centre Borelli)

[S. Clémençon et al., 2025]

Goal and Motivation

Goal. for $(X, Y) \in \mathbb{R}^d \times [-M, M]$ input/output random pair, find f s.t. $f(X) \approx Y$ given that $\|X\|$ is large

Risk decomposition:

$$R(f) = \mathbb{P}(\|X\| \leq t)\mathbb{E}\Big[(Y - f(X))^2 \mid \|X\| \leq t\Big] + \underbrace{\mathbb{P}(\|X\| \geq t)}_{\ll 1, \text{ if } t \gg 1} \mathbb{E}\Big[(Y - f(X))^2 \mid \|X\| \geq t\Big]$$

$\Rightarrow$ Extremes are negligible in standard Empirical Risk Minimisation

$\Rightarrow$ focus on the minimisation of the *Conditional Risk*

$$R_t(f) := \mathbb{E}\Big[(Y - f(X))^2 \mid \|X\| \geq t\Big].$$

Beyond observed data

$$R_t(f) := \mathbb{E}\Big[(Y - f(X))^2 \mid \|X\| \geq t\Big].$$

 minimiser of R_t depends on t

$\Rightarrow$ no performance guarantees in more distant regions (for $t' > t$).

$\Rightarrow$ focus on the minimisation of the *Asymptotic Risk*

$$R_\infty(f) := \limsup_{t \to +\infty} R_t(f) = \limsup_{t \to +\infty} \mathbb{E}[(Y - f(X))^2 \mid \|X\| \geq t].$$

 Regular variation assumption

Regular Variation of $X \in \mathbb{R}^d$

Peaks-over-threshold assumption

$X \in RV(\mathbb{R}^d)$ if there exist a regularly varying function b with index $\alpha > 0$ (*i.e.*, $b(tx)/b(t) \underset{t\to+\infty}{\longrightarrow} x^\alpha$) and a nonzero Borel measure μ on $\mathbb{R}^d\backslash\{0\}$, finite on all Borelian sets bounded away from zero s.t.

$$\lim_{t\to+\infty} b(t)\mathbb{P}(X/t \in A) = \mu(A), \qquad \text{(vague convergence)}$$

for all Borelian sets A bounded away from zero and s.t. $\mu(\partial A) = 0$.

$\Leftrightarrow$ there exists a limit random variable X_∞ s.t.

$$\lim_{t\to+\infty} \mathscr{L}(X/t \mid \|X\| \geq t) = \mathscr{L}(X_\infty)$$

$\Leftrightarrow$ there exist a limit radius R_∞ and limit angle Θ_∞ s.t.

$$\lim_{t\to+\infty} \mathscr{L}(X/\|X\|, \|X\|/t \mid \|X\| \geq t) = \mathscr{L}(\Theta_\infty, R_\infty)$$

$$X_\infty = R_\infty.\Theta_\infty \text{ and } R_\infty \perp\!\!\!\perp \Theta_\infty$$

Regular Variation w.r.t. some component

⚠ in the classic regular variation, extremality is measured on the whole vector

Regular Variation w.r.t. the covariates.

$$\lim_{t \to +\infty} b(t)\mathbb{P}(X/t \in A, Y \in C) = \mu(A \times C),$$

for all $C \in \mathscr{B}([-M, M])$ and $A \in \mathscr{B}(\mathbb{R}^d)$ bounded away from zero s.t. $\mu(\partial(A \times C)) = 0$.

◦ adaption of the classic assumption **to measure the extremality according to some component** (here the input variable).

Important example

Predicting a missing component in a regularly varying vector

Let $Z = (Z_1, ..., Z_{d+1}) \in RV(\mathbb{R}^{d+1})$. Under classic extreme-value assumptions on the density of Z, the pair (X, Y), defined as

$$X = (Z_1, ..., Z_d) \quad \text{and} \quad Y = Z_{d+1}/\|Z\|_p,$$

meets our assumptions.

$\Rightarrow$ **our framework is well-suited for predicting Z_{d+1} based on $Z_1, ..., Z_d$ given that $\|(Z_1, ..., Z_d)\|_p$ is large**

NB back to original scale through

$$Y = \frac{Z_{d+1}}{\|Z\|_p} \quad \Longleftrightarrow \quad Z_{d+1} = \frac{Y\|X\|_p}{(1 - |Y|^p)^{1/p}}.$$

Consequences

of regular variation w.r.t. X

- Existence of $(R_\infty, \Theta_\infty, Y_\infty)$ s.t.

$$\mathscr{L}(t^{-1}X, Y \mid \|X\| \geq t) \underset{t\to+\infty}{\longrightarrow} \mathscr{L}(R_\infty.\Theta_\infty, Y_\infty)$$

with $R_\infty \perp\!\!\!\perp \Theta_\infty, Y_\infty$

$\rightsquigarrow$ Θ_∞ conveys all the information in $X_\infty = R_\infty.\Theta_\infty$ to predict Y_∞, *i.e.*,

$$f^*_\infty(X_\infty) = \mathbb{E}[Y_\infty \mid X_\infty] = \mathbb{E}[Y_\infty \mid \Theta_\infty]$$

We say that f^*_∞ is an *angular function*.

Propagation of this property to finite-distance extreme regions?

Propagation of the angular property

Notation: $\theta(x) = x/\|x\|$ and $\Theta = X/\|X\|$.

Proposition*(angular minimiser at finite-distance)*.
With existence of densities and regularity conditions:

Convergence of minima: $\inf_f R_t(f) \underset{t\to+\infty}{\longrightarrow} \inf_f R_\infty(f)$.

Angular minimiser: $\inf_f R_\infty(f) = R_\infty(f^*_\infty)$,
with $f^*_\infty(x) = f^*_\infty(\theta(x))$.

Consequence: $\inf_h R_t(h \circ \theta) \underset{t\to+\infty}{\longrightarrow} \inf_f R_\infty(f)$.

$\Rightarrow$ suggests replacing the former minimisation problem with

$$\min_h R_t(h \circ \theta).$$

Benefits: extrapolation property + dimension reduction

ROXANE algorithm

Regression On eXtreme ANglEs

Input sample $\Big\{(X_1, Y_1), ..., (X_n, Y_n)\Big\}$ of input/output pairs; a class of angular regression functions $\mathscr{H}$; number $k \leq n$ of extreme observations.

Truncation keep the k 'largest' observations $\Big\{(X_{(1)}, Y_{(1)}), ..., (X_{(k)}, Y_{(k)}))\Big\}$.

Extreme ERM solve the minimisation problem

$$\min_{h \in \mathscr{H}} \frac{1}{k} \sum_{i=1}^{k} \Big(Y_{(i)} - h(\theta(X_{(i)}))\Big)^2.$$

Output angular prediction function $\hat{h} \circ \theta$ for new examples such that $\|X\| \geq \|X_{(k)}\|$.

Statistical Guarantees

Empirical Risk Minimisation

Ordered sample: $\left\{(X_{(1)}, Y_{(1)}), ..., (X_{(n)}, Y_{(n)})\right\}$ such that $\|X_{(1)}\| \geq \|X_{(2)}\| \geq ....$

$\rightsquigarrow$ *Empirical Conditional Risk* associated with the k largest obs.

$$\hat{R}_{n,k}(h \circ \theta) := \frac{1}{k} \sum_{i=1}^{n} \Big(Y_i - h(\theta(X_i))\Big)^2 \mathbb{1}\{\|X_i\| \geq \|X_{(k)}\|\}$$

$$= \frac{1}{k} \sum_{i=1}^{k} \Big(Y_{(i)} - h(\theta(X_{(i)}))\Big)^2.$$

$\rightsquigarrow$ $\hat{h}_{\theta,k}$ solution of $\min_{h \in \mathscr{H}} \hat{R}_{n,k}(h \circ \theta)$ over a class $\mathscr{H}$

NB $\|X_{(k)}\|$ is the empirical version of the quantile $t_{n,k}$ s.t.

$$\mathbb{P}(\|X\| \geq t_{n,k}) = k/n.$$

Risk decomposition

what can we expect?

$$R_\infty(\hat{h}_{\theta,k} \circ \theta) - \inf_f R_\infty(f) \leq \underbrace{\left(\inf_{h \in \mathscr{H}} R_{t_{n,k}}(h \circ \theta) - \inf_f R_{t_{n,k}}(f)\right)}_{\text{model bias}}$$

$$+ \underbrace{2 \sup_{h \in \mathscr{H}} |R_{t_{n,k}}(h \circ \theta) - R_\infty(h \circ \theta)|}_{\text{extreme bias 1}} + \underbrace{\left(\inf_f R_{t_{n,k}}(f) - \inf_f R_\infty(f)\right)}_{\text{extreme bias 2: } \underset{n,k \to +\infty}{\longrightarrow} 0}$$

$$+ \underbrace{2 \sup_{h \in \mathscr{H}} |\hat{R}_{n,k}(h \circ \theta) - R_{t_{n,k}}(h \circ \theta)|}_{\text{stochastic error}}$$

Uniform Statistical Guarantees

a concentration bound + a negligible bias

Assumption(*VC-class*): $\mathscr{H} \subset \mathscr{C}^0(\mathbb{S}, \mathbb{R})$ with VC-dimension $V_{\mathscr{H}} < +\infty$, uniformly bounded

Theorem(*Statistical Guarantees*).
Control of stochastic error: With large probability:

$$\sup_{h \in \mathscr{H}} \left| \hat{R}_{n,k}(h \circ \theta) - R_{t_{n,k}}(h \circ \theta) \right| \leq C/\sqrt{k} + O(1/k).$$

Control of extreme bias 1: Under a mild additional assumption, we have:

$$\sup_{h \in \mathscr{H}} \left| R_{t_{n,k}}(h \circ \theta) - R_{\infty}(h \circ \theta) \right| \underset{n,k \to +\infty}{\longrightarrow} 0.$$

Tools: VC-bound + Bernstein's type inequality.

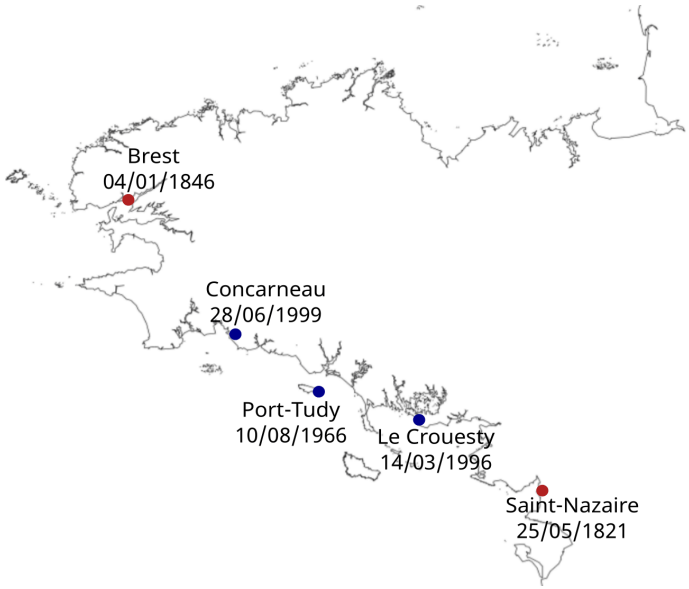
An application to the prediction of extreme sea levels

joint work with Philippe Naveau (LSCE) and
Anne Sabourin (Centre Borelli)

[N. Huet et al., 2026]

Prediction of extreme sea levels

Motivation



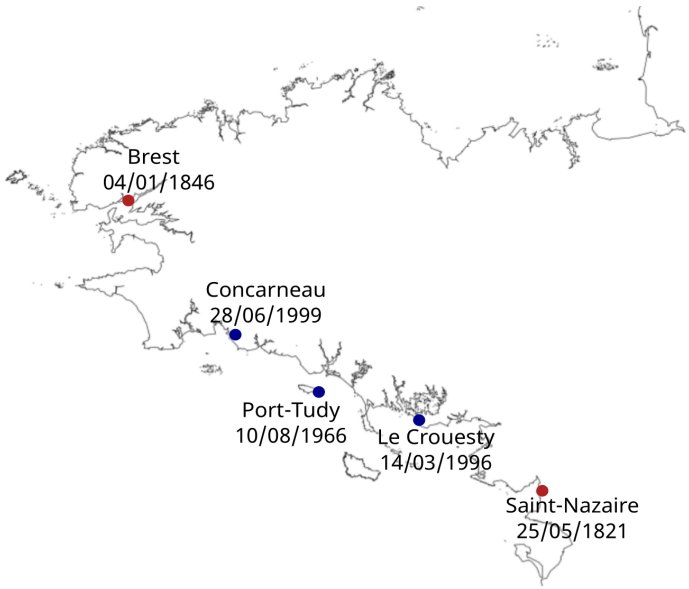
Motivation: recent tide gauges have very short records $\Rightarrow$ inferring 100-year return value at a station with about 20 years of data is subject to large uncertainty

(extreme) observations at nearby stations are strongly dependent + some stations have hundred years of measurements

$\Rightarrow$ we can use these long records to reconstruct the past time series to improve the inference of return levels

Prediction of extreme sea levels

sea level data (SHOM)



Inputs: Brest (X_B) and Saint-Nazaire (X_N)

Output: Port-Tudy (Y)

Training: 01/01/2000 - 31/12/2023
Validation: 10/08/1966 - 31/12/1999
Extreme observations: (X_B, X_N, Y) given that $\left\{X_B \geq t_B \text{ or } X_N \geq t_N\right\}$ with t_B, t_N large thresholds

comparison of ROXANE to a parametric method

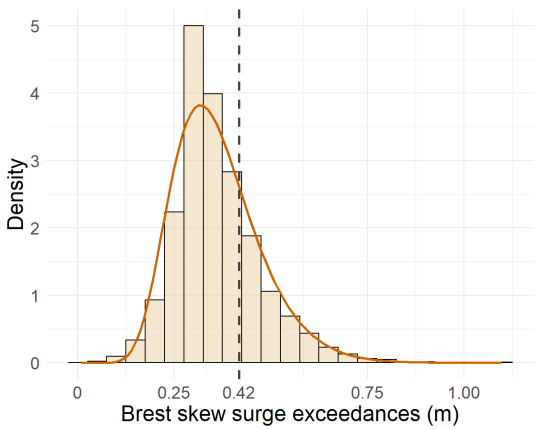
Marginal modelling

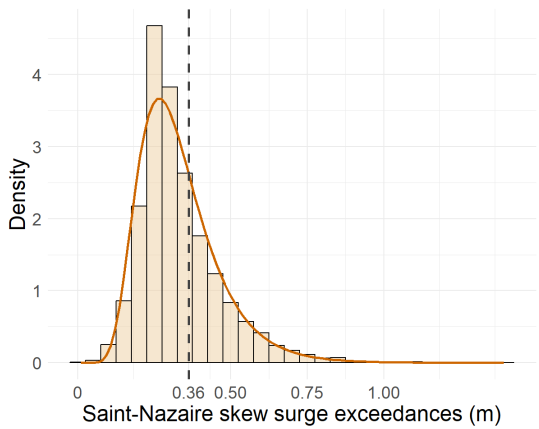
common to both procedures

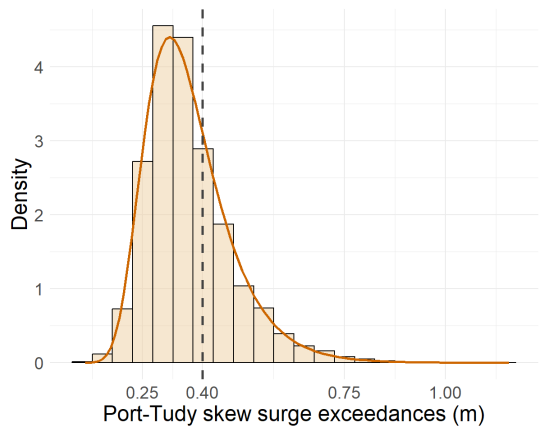
Margins are modelled by an Extended Generalised Pareto Distribution (EGPD) with cdf

$$F_{EGP}(x;\sigma,\xi,\kappa) = \left(1-\left(1+\frac{\xi x}{\sigma}\right)^{-1/\xi}\right)^{\kappa}$$

- Generalised Pareto (GP) behaviour in the right-tail;
- κ parameter controls the lower-tail behavior.







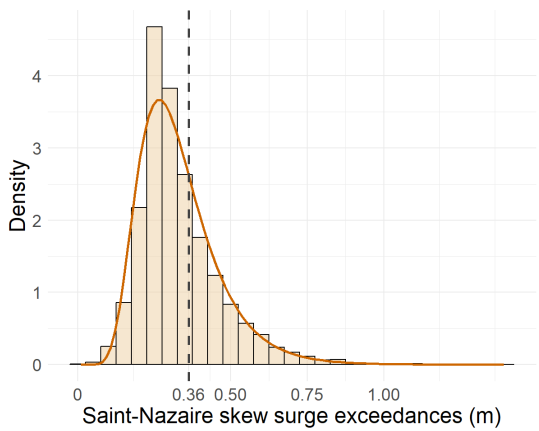
Threshold Selection

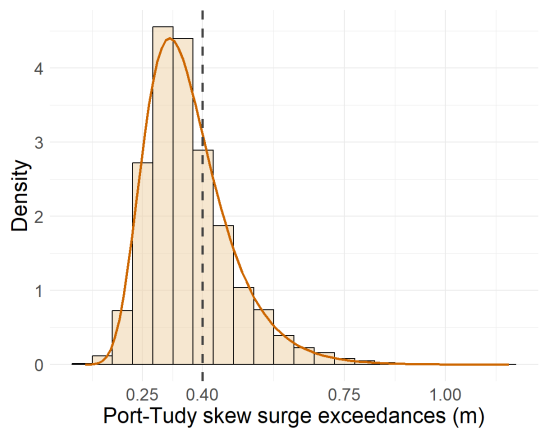
EGPD behaves as GPD in the right-tail

$+$ GP density strictly convex for $\xi > -1/2$

$\rightsquigarrow$ **selected threshold** $\boldsymbol{t}$ lowest points above which the fitted densities are convex, *i.e.*, largest zeros of $d^3 F_{EGP}(x; \sigma, \xi, \kappa)/dx^3$.



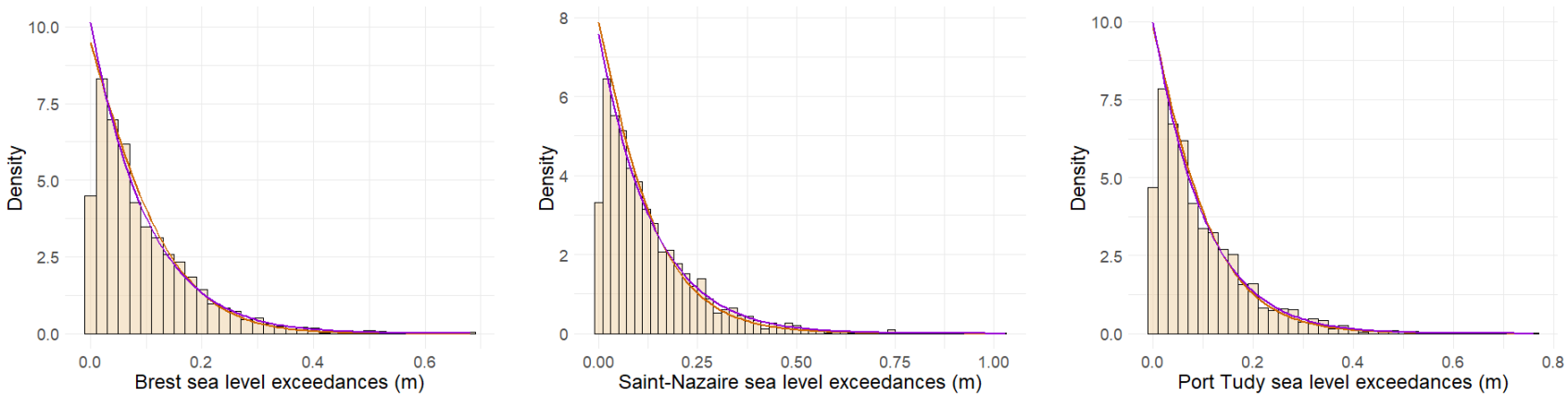




Visual validity

EGPD *vs* GPD

- Fit of a GP distribution above the selected threshold



Multivariate procedures

nonparametric vs *parametric*

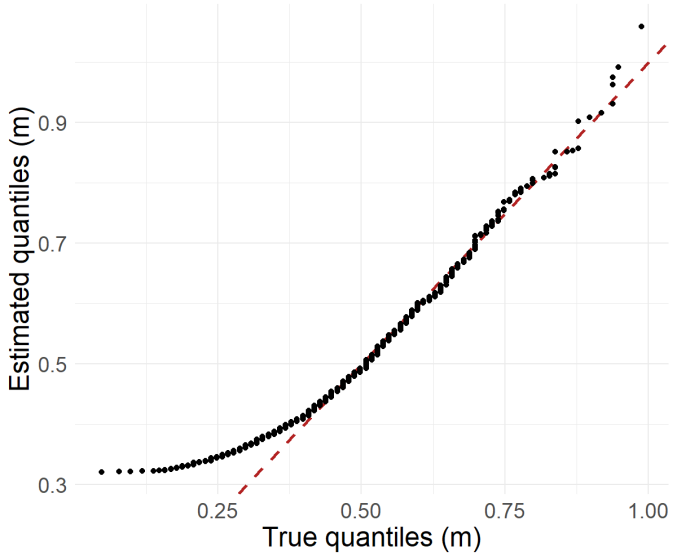
ROXANE procedure:

1. Pareto marginal transformation (to satisfy regular variation condition);
2. "angular" transformation as in the "Important example" (to fit our framework);
3. predictions *via* predictive function estimated by OLS or RF.

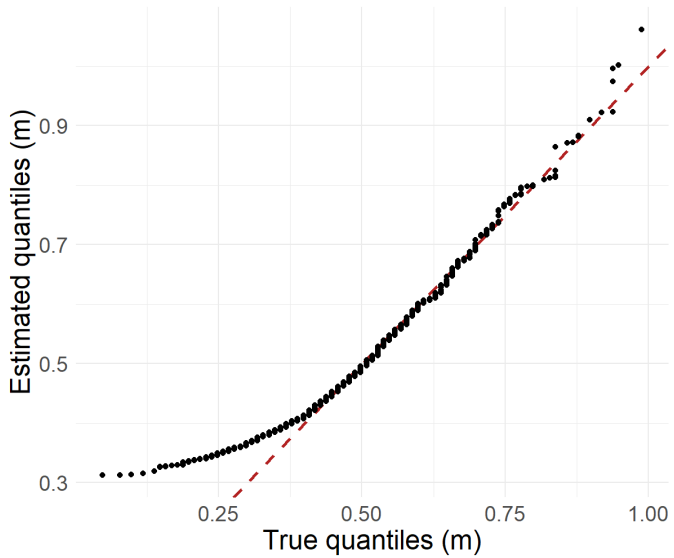
Multivariate Generalised Pareto (MGP) modelling:

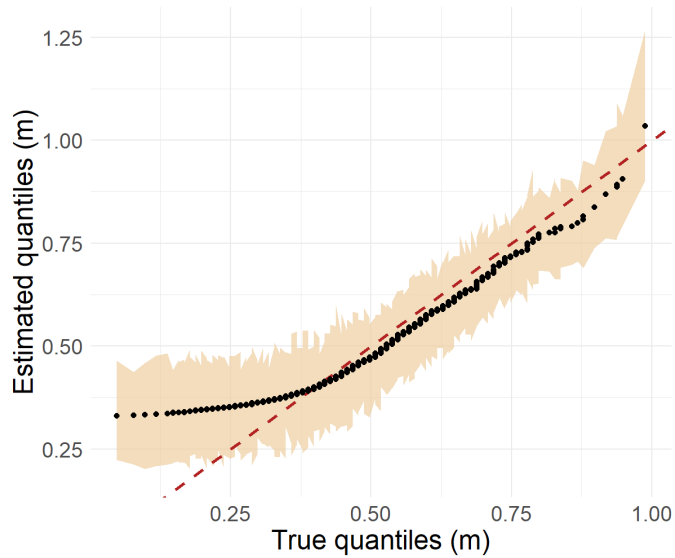
1. procedure in [A. Kiriliouk et al., 2019] to deduce a well-fitted density;
2. conditional sampling given the values at the input stations;
3. predictions *via* Monte-Carlo average of the conditionally generated values.

QQ-plots of the true values *vs* the estimated ones



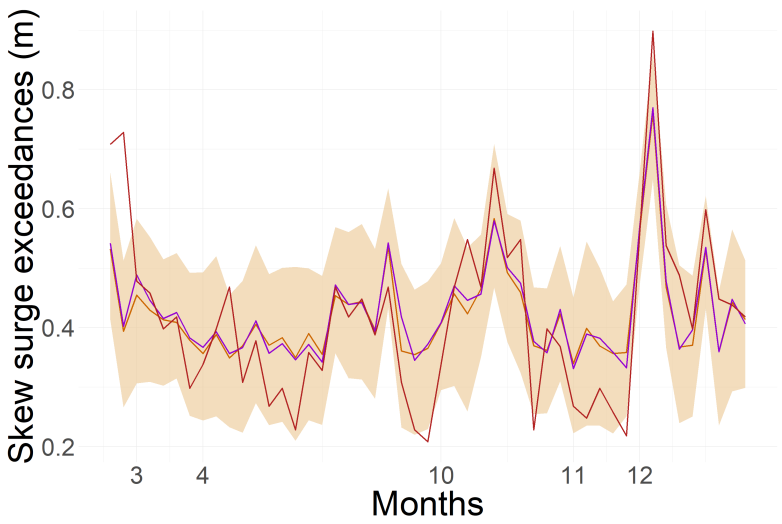
- **ROXANE OLS** (Upper-left)
- **ROXANE RF** (Bottom-left)
- **MGP** (Bottom-right)

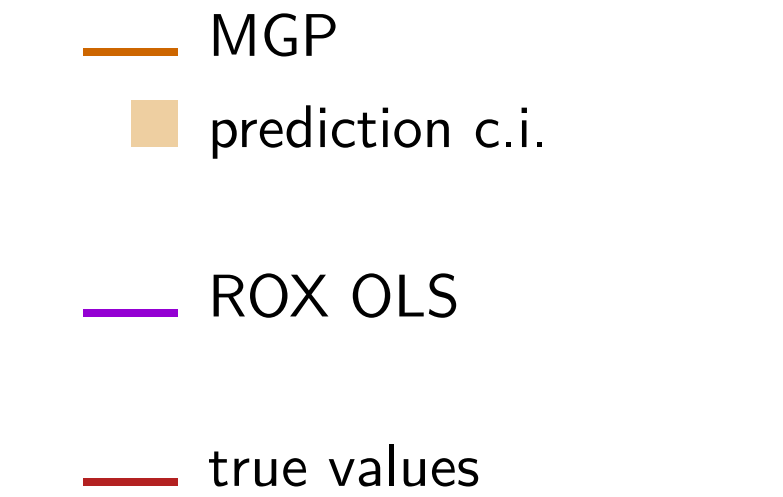


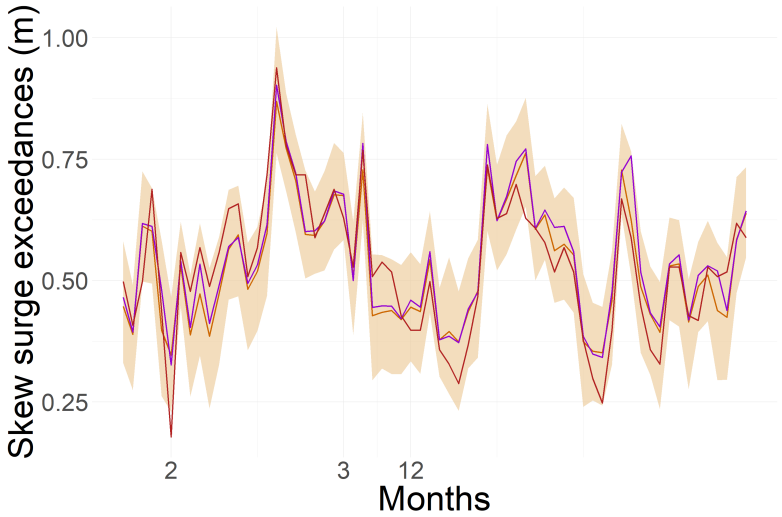


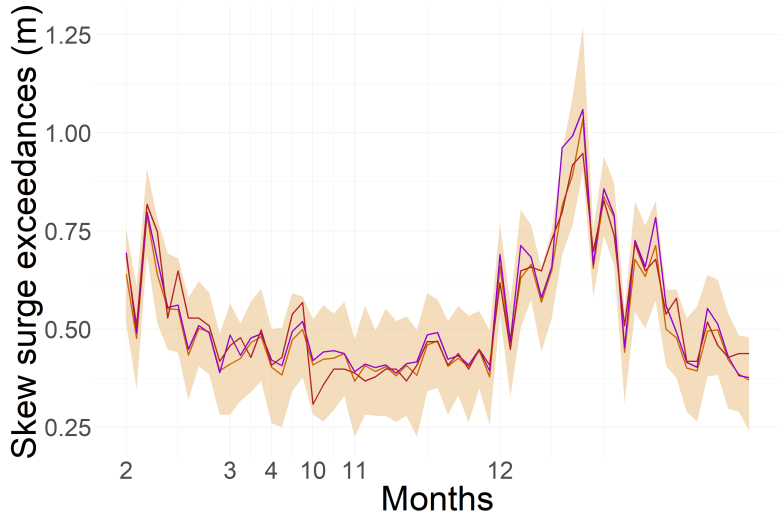
Time series prediction

of extreme skew surges for 1978, 1979, and 1989



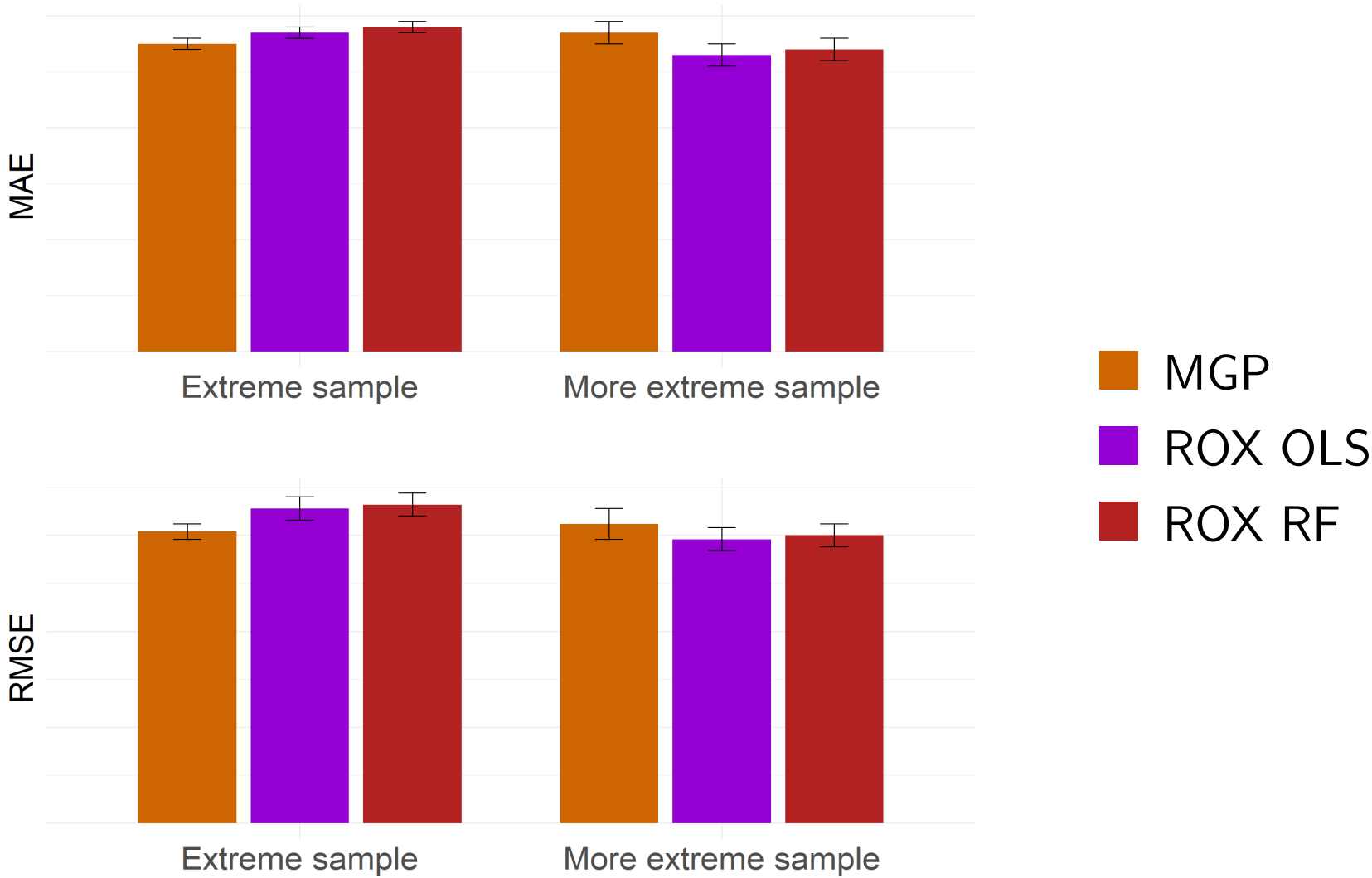






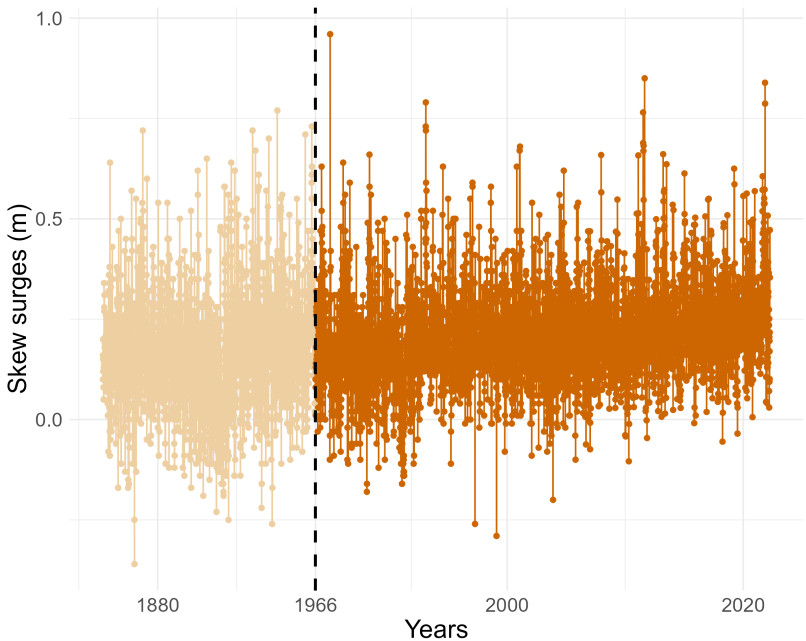
Model Errors

Mean Absolute Error/Root Mean Square Error

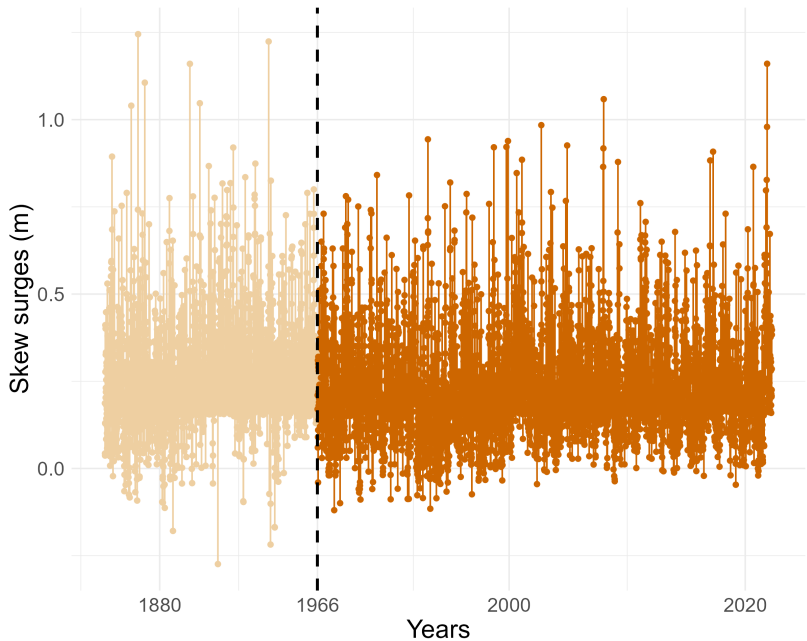


Reconstruction of the time series

the data



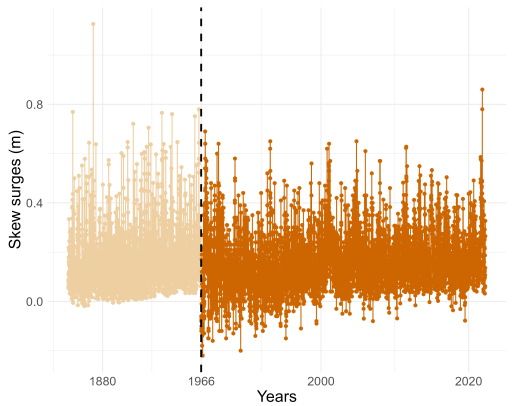
Brest skew surge time series.



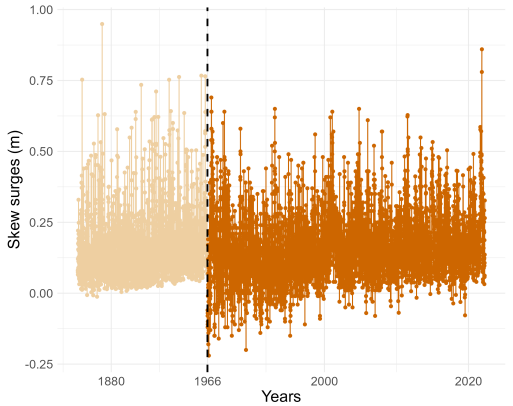
Saint-Nazaire skew surge time series.

Reconstruction of the time series

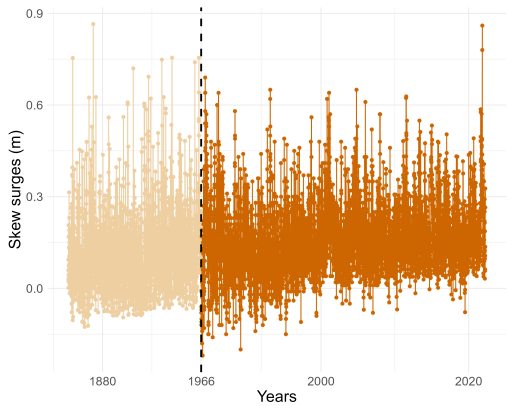
the predictions



ROX OLS



ROX RF



MGP

Perspectives

- adjust the model by including meteorological variables;
- analysis of our method for improving inference on long return periods;
- develop a simpler model for practitioners.

References

- S. Clémençon, N. Huet and A. Sabourin, *On Regression in Extreme Regions*, Electronic Journal of Statistics, 2025;
- N. Huet, P. Naveau and Anne Sabourin, *Multi-site modelling and reconstruction of past extreme skew surges along the French Atlantic coast*, Journal of the Royal Statistical Society Series C: Applied Statistics, 2026;
- A. Kiriliouk, H. Rootzén, J. Segers and J. Wadsworth, *Peaks over thresholds modeling with multivariate generalized Pareto distributions*, Technometrics, 2019;

Thank you for your attention!