Università
Ca'Foscari
Venezia

Statistical Learning for Extremes : an application to the prediction of extreme sea levels

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

May 13, 2025

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, *i.e.* in the tail of the distribution

Usual assumptions on X a random element

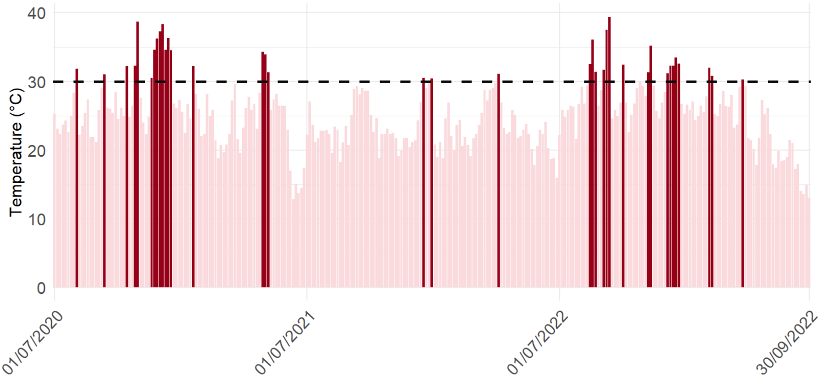
- convergence in distribution of maxima, *i.e.*

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

  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



Focus in my work : observations exceeding a high threshold

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

PoT 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.

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

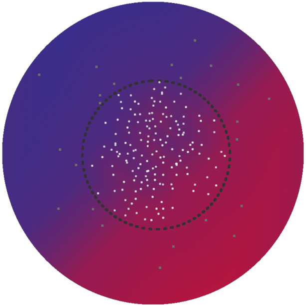
$$\boxed{X_\infty = R_\infty.\Theta_\infty \text{ and } R_\infty \perp\!\!\!\perp \Theta_\infty}$$

Question

Focus in my thesis: How to obtain guarantees for **Extreme Values** through **Statistical Learning** methods?

Statistical learning for extremes?

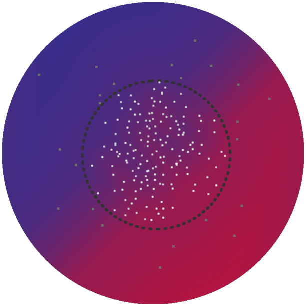
◦ classic algorithms and concentration results focus on **the bulk of the distribution** (under boundedness or sub-Gaussianity assumptions)



⚠ **'normal' behavior**
$\neq$
extreme behavior

Statistical learning for extremes?

◦ classic algorithms and concentration results focus on **the bulk of the distribution** (under boundedness or sub-Gaussianity assumptions)



⚠ **'normal' behavior** $\neq$ **extreme behavior**

$\Rightarrow$ **classic statistical learning methods need adaptation to perform well in extreme regions**

Statistical learning for extremes in the literature

still fresh...

supervised learning	**Classification** [Jalalzai et al.,2018] [Clémençon et al.,2023]; **Regression** [Huet et al.,2024] [Buritica and Engelke,2024]
functional data analysis	**functional PCA** [Kokoszka and Xiong,2018], [Kokoszka and Kulik,2023] [Kim and Kokoszka,2024],[Huet et al.,2024]
miscellanea	**Dimension reduction** [Goix et al.,2016] [Cooley and Thibaud,2019] [Drees and Sabourin,2021]; **Anomaly detection** [Goix et al.,2017] [Chiapino et al.,2020]; **Clustering** [Janßen and Wan,2020] [Vignotto et al.,2021]; **Quantile regression** [Velthoen et al.,2023] [Gnecco et al.,2023]; **Cross validation** [Aghbalou et al.,2022]; **Graphical models** [Engelke and Hitz,2020]
concentration	[Boucheron and Thomas,2012][Goix et al.,2015] [Lhaut and Segers,2021][Lhaut et al.,2022]

Regression for extremes

joint work with Stephan Clémençon and Anne Sabourin

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:

$$R(f) = \mathbb{P}(\|X\| \leq t)\mathbb{E}\Big[(Y - f(X))^2 \mid \|X\| \leq t\Big] + \mathbb{P}(\|X\| \geq t)\,\mathbb{E}\Big[(Y - f(X))^2 \mid \|X\| \geq t\Big]$$

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 Minimization

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 Minimization
$\Rightarrow$ focus on the minimization of the *Conditional Risk*

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

Beyond observed data

⚠ minimizer of R_t depends on t

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

Beyond observed data

⚠ minimizer of R_t depends on t

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

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

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

Beyond observed data

⚠ minimizer of R_t depends on t

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

$\Rightarrow$ focus on the minimization 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 w.r.t. some component

Regular Variation w.r.t. some component

Appropriate regularity/stability condition?

Reminder: $X \in RV(\mathbb{R}^d)$ if $\lim_{t\to+\infty} b(t)\mathbb{P}(X/t \in \cdot) = \mu$.

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.

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 $\boxed{R_\infty \perp\!\!\!\perp \Theta_\infty, Y_\infty}$

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 $\boxed{R_\infty \perp\!\!\!\perp \Theta_\infty, Y_\infty}$

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

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

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 minimizer 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 minimizer: $\inf_f R_\infty(f) = R_\infty(f_\infty^*)$,
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 minimization problem with

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

Benefits: extrapolation property + dimension reduction

ROXANE algorithm

to handle regression in extreme regions

Input sample $\left\{(X_1, Y_1), \ldots, (X_n, Y_n)\right\}$ 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 $\left\{(X_{(1)}, Y_{(1)}), \ldots, (X_{(k)}, Y_{(k)}))\right\}$.

Extreme ERM solve the minimization 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 Minimization

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

$\rightsquigarrow$ *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?

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

$$+\ 2 \sup_{h\in\mathscr{H}} |R_{t_{n,k}}(h \circ \theta) - R_\infty(h \circ \theta)| + \left(\inf_f R_{t_{n,k}}(f) - \inf_f R_\infty(f)\right)$$

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

Risk decomposition

what can we expect?

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

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

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

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:

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

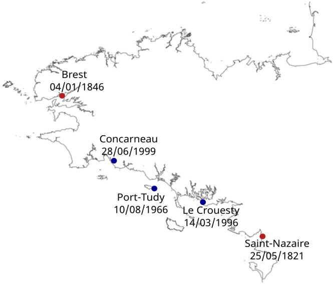
Tools: VC-bound + Bernstein's type inequality.

An application to the prediction of extreme sea levels

joint work with Philippe Naveau and Anne Sabourin

Prediction of extreme sea levels

sea levels data (SHOM)



Goal: predict sea levels Y at some output tide gauges (•) given extreme sea levels $X = (X_B, X_N)$ measured at nearby input stations (•).

Output station: Port-Tudy (10/08/1966 - 31/12/2023)

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

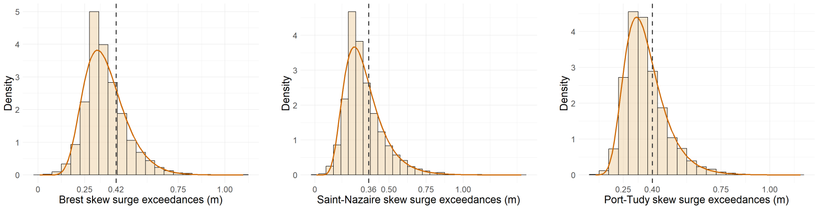
Marginal modeling

common to both procedures

Margins are modeled by an Extended Generalized Pareto distribution with cdf

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

- Generalized Pareto behavior in the right-tail;
- κ parameter controls the lower-tail behavior.

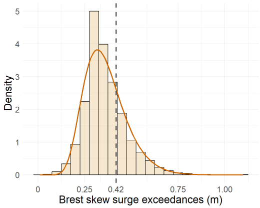


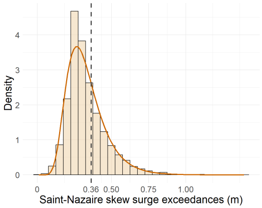
Threshold Selection

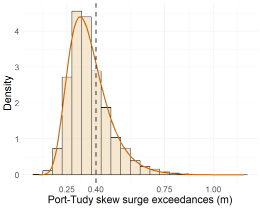
EGPD behaves as GPD in the right-tail

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

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



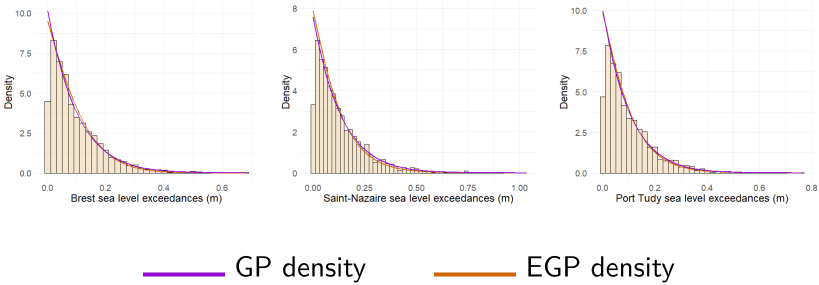




Visual validity

EGPD *vs* GPD

◦ Fit of a GP distribution above the selected threshold



Multivariate procedures

nonparametric vs *parametric*

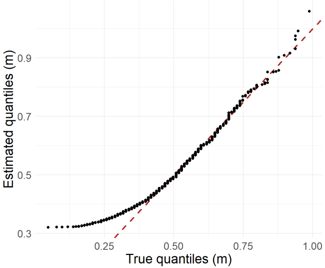
ROXANE procedure:

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.

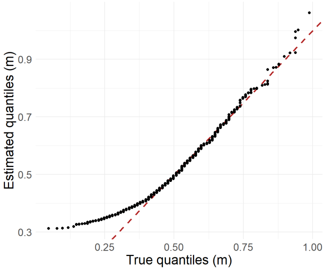
Multivariate Generalized Pareto (MGP) modeling:

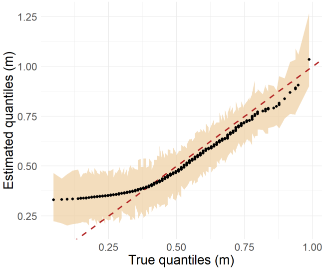
1. procedure in [Kiriliouk et al., 2019] to deduce a well-fitted density;
2. conditional sampling given the values at the input stations;
3. predictions *via* Monte-Carlo average of the conditionally generated values.

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



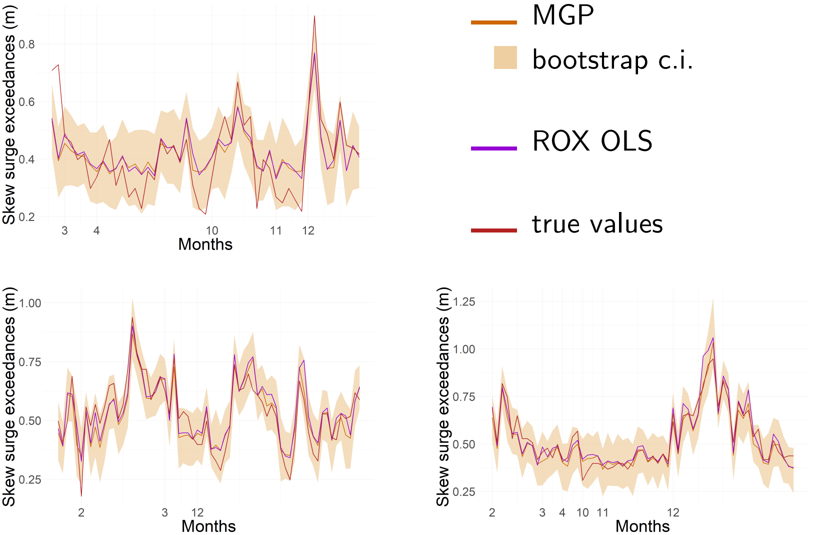
- **ROXANE OLS** (Upper-left)
- **ROXANE RF** (Bottom-left)
- **MGP** (Bottom-right)





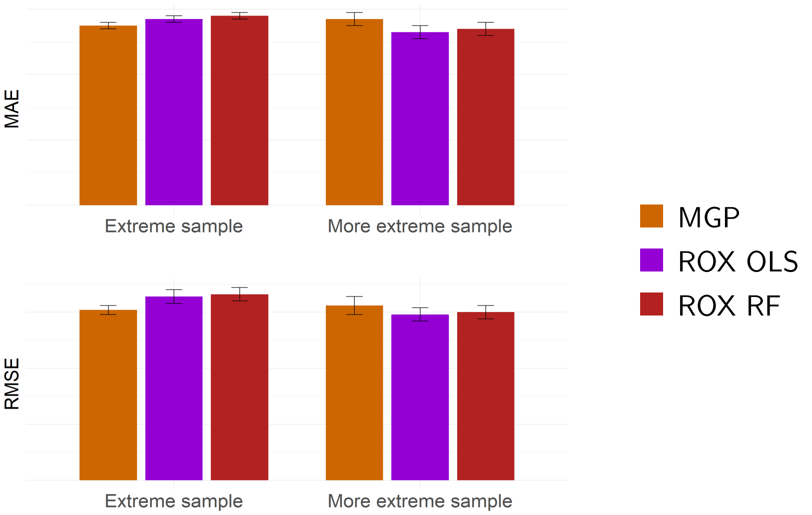
Time series prediction

of extreme skew surges for 1978, 1979, and 1989



Model Errors

Mean Absolute Error/Root Mean Square Error



Perspectives

Regression for extremes

- relaxation of assumptions (in particular the regular variation);
- statistical guarantees for the empirical marginal standardization in the ROXANE algorithm.

Modeling and Reconstruction of Extreme Sea Levels

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

References

- A. Aghbalou, P. Bertail, F. Portier and A. Sabourin, *Cross-validation on extreme regions*, Extremes, 2024;
- S. Boucheron and M. Thomas, *Concentration inequalities for order statistics*, Electron. Commun. Probab., 2012;
- G. Buritica and S. Engelke, *Progression: an extrapolation principle for regression*, arXiv, 2024;
- M. Chiapino, S. Clémençon, V. Feuillard and A. Sabourin, *A multivariate extreme value theory approach to anomaly clustering and visualization*, Computational Statistics, 2020;
- S. Clémençon, N. Huet, and A. Sabourin, *On regression in extreme regions*, arXiv, 2024;
- S. Clémençon, H. Jalalzai, S. Lhaut, A. Sabourin and J. Segers, *Concentration bounds for the empirical angular measure with statistical learning applications*, Bernoulli, 2023;

References

- D. Cooley and E. Thibaud, *Decompositions of dependence for high-dimensional extremes*, Biometrika, 2019;
- H. Drees and A. Sabourin, *Principal component analysis for multivariate extremes*, Electron. J. Statist., 2021;
- S. Engelke and A. S. Hitz, *Graphical models for extremes*, J. R. Statist. Soc. B, 2020;
- N. Gnecco, J. Peters, S. Engelke and N. Pfister, *Boosted control function : Distribution generalization and invariance in confounded models*, arXiv, 2023;
- N. Goix, A. Sabourin and S. Clémençon, *Sparse representation of multivariate extremes with applciations to anomaly detection*, AISTATS, 2016;
- H. Jalalzai, S. Clémençon and A. Sabourin, *On binary classification in extreme regions*, NIPS, 2018;

References

- A. Janßen and P. Wan, *k-means clustering of extremes*, Electron. J. Statist., 2020;
- A. Kiriliouk, H. Rootzen, J. Segers and J. L. Wadsworth, *Peaks over thresholds modeling with multivariate generalized Pareto distributions*, Technometrics, 2019;
- S. Lhaut, A. Sabourin and J. Segers, *Uniform concentration bounds for frequencies of rare events*, Statistics & Probability Letters, 2022;
- J. Velthoen, C. Dombry, J. J. Cai and S. Engelke, *Gradient boosting for extreme quantile regression*, Extremes, 2023;
- E. Vignotto, S. Engelke and J. Zscheischler, *Clustering bivariate dependencies of compound precipitation and wind extremes over Great Britain*, Weather and Climate Extremes, 2021.

Thank you for your attention!