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

June 11, 2026

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:

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

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

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

$\rightsquigarrow$ 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 $\Big\{(X_1, Y_1), ..., (X_n, Y_n)\Big\}$ 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 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: $\left\{(X_{(1)}, Y_{(1)}), ..., (X_{(n)}, Y_{(n)})\right\}$ 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?

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

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

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

Risk decomposition

what can we expect?

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

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

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

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.

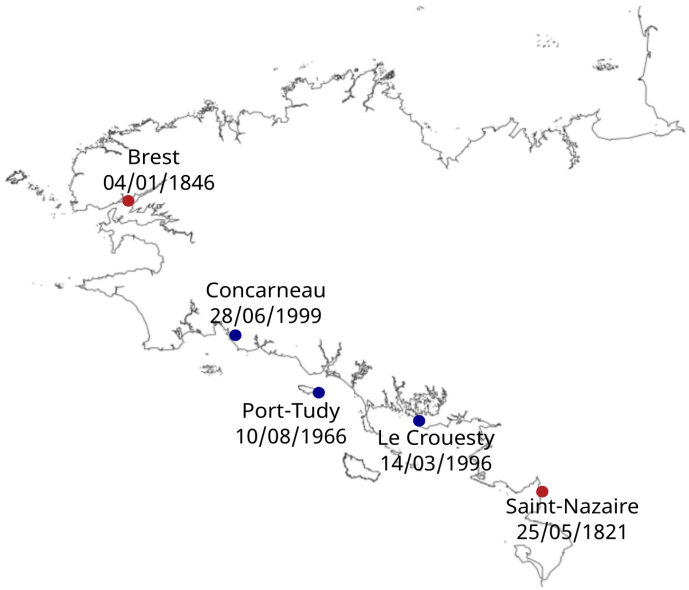
An application to the prediction of extreme sea levels

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

[N. Huet et al., 2026]

Prediction of extreme sea levels

sea levels data (SHOM)



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

Output station: Port-Tudy (10/08/1966 - 31/12/2000 - 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

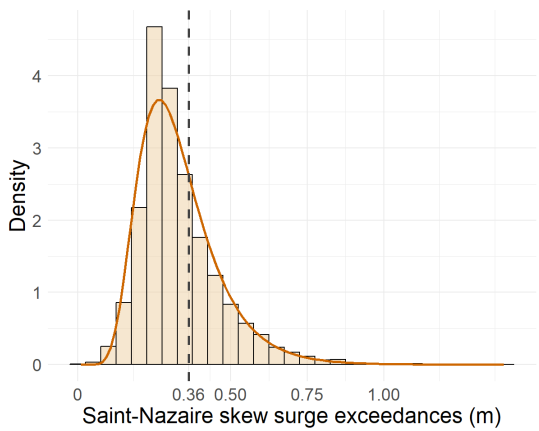
Marginal modeling

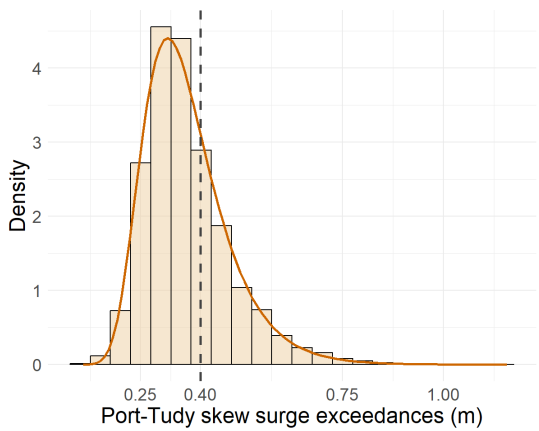
common to both procedures

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.



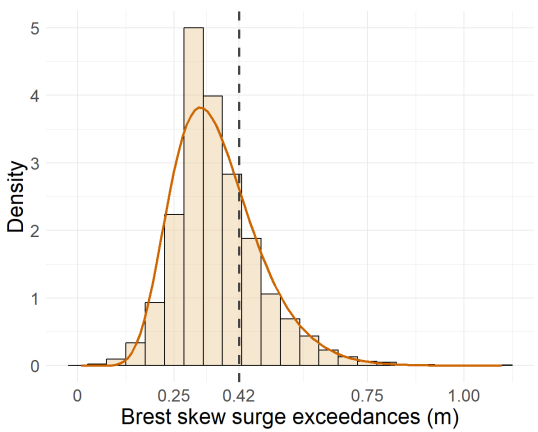


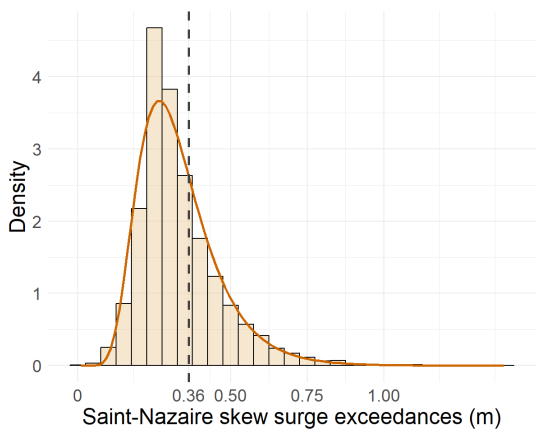
Threshold Selection

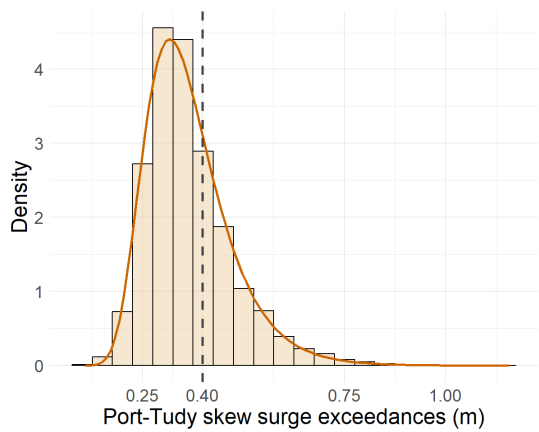
EGPD behaves as GPD in the right-tail

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

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



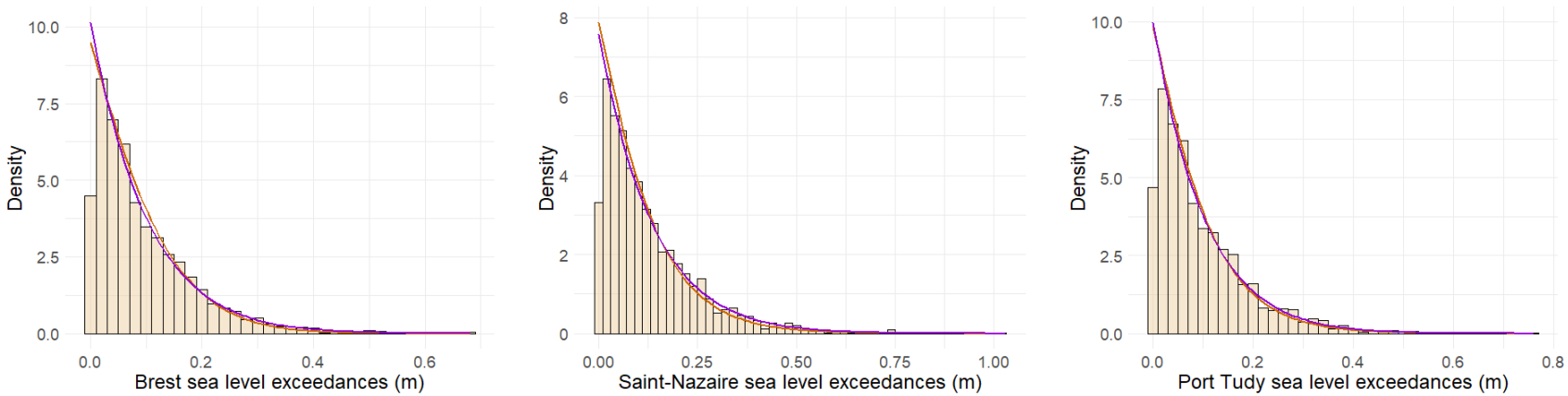




Visual validity

EGPD *vs* GPD

- Fit of a GP distribution above the selected threshold



Multivariate procedures

nonparametric vs *parametric*

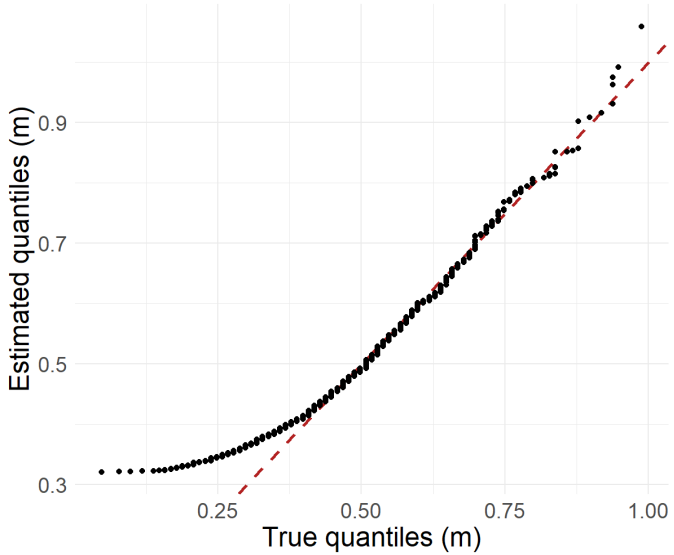
ROXANE procedure:

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.

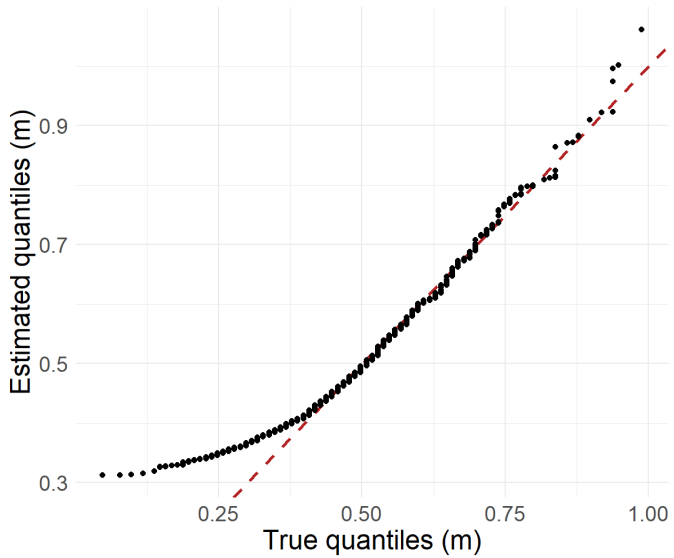
Multivariate Generalized Pareto (MGP) modeling:

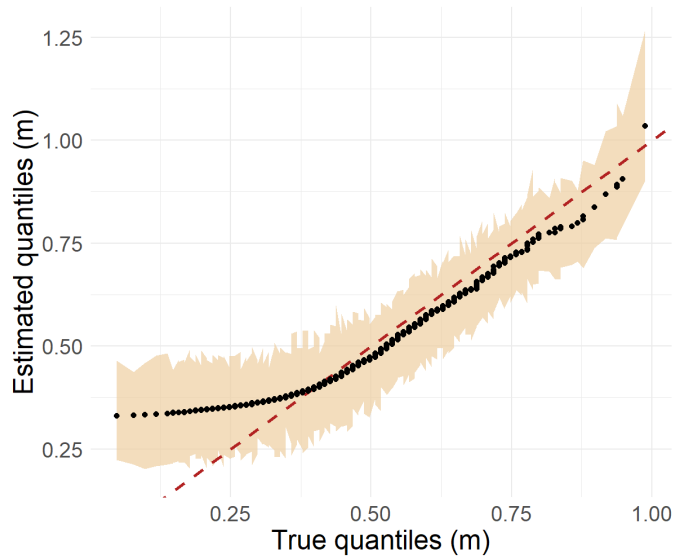
1. procedure in [A. Kiriliouk et al., 2019] to deduce a well-fitted density;
2. conditional sampling given the values at the input stations;
3. predictions *via* Monte-Carlo average of the conditionally generated values.

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



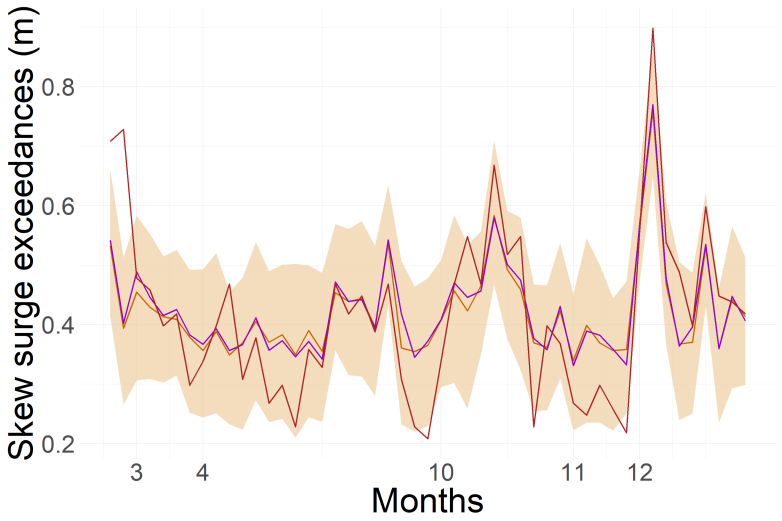
- **ROXANE OLS** (Upper-left)
- **ROXANE RF** (Bottom-left)
- **MGP** (Bottom-right)

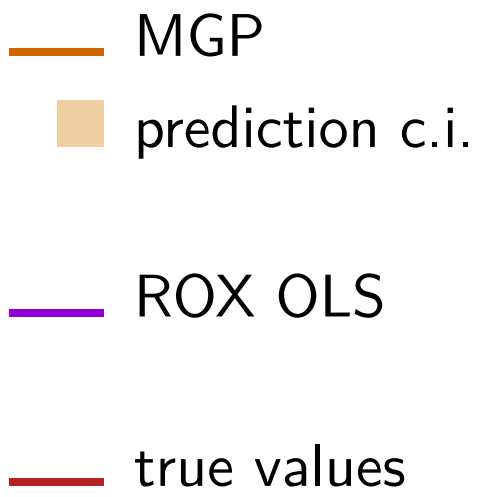


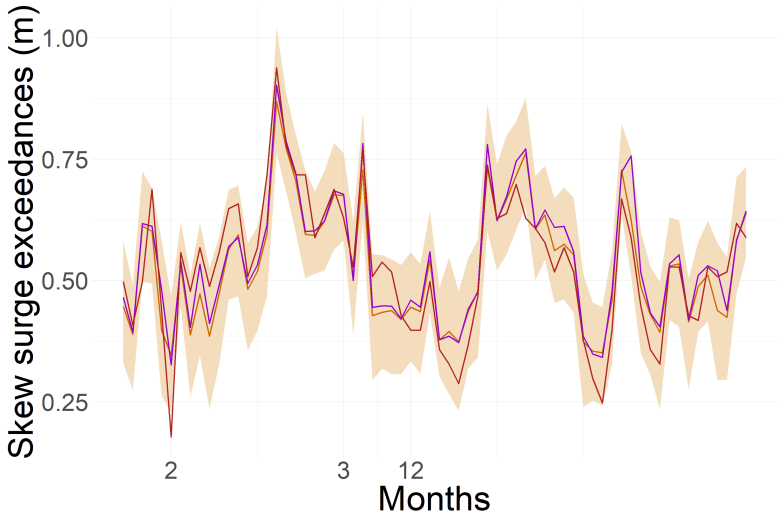


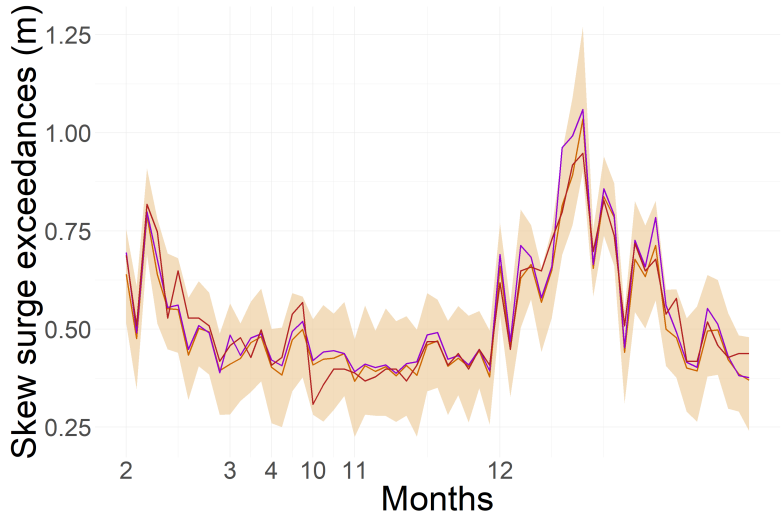
Time series prediction

of extreme skew surges for 1978, 1979, and 1989



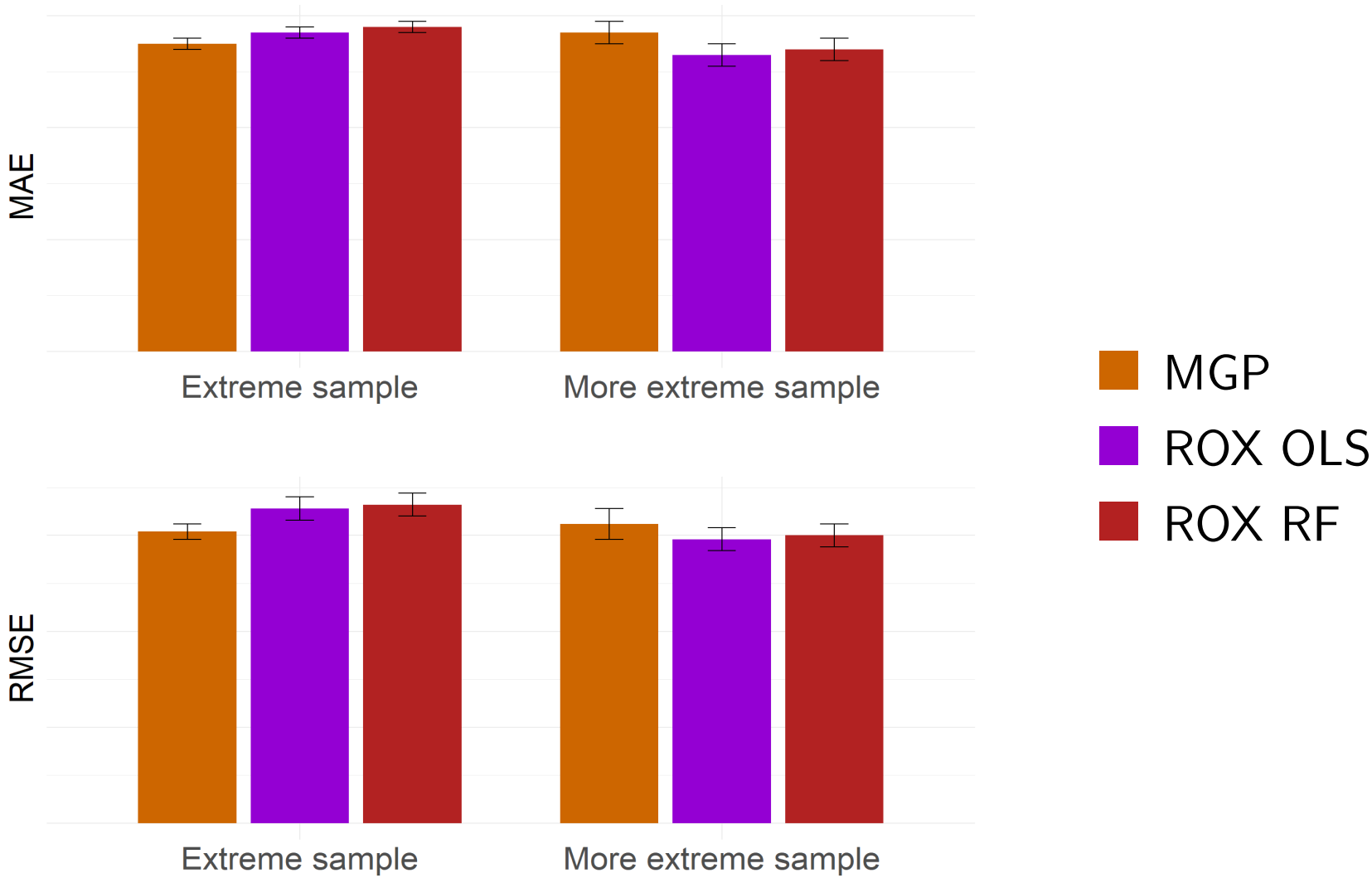






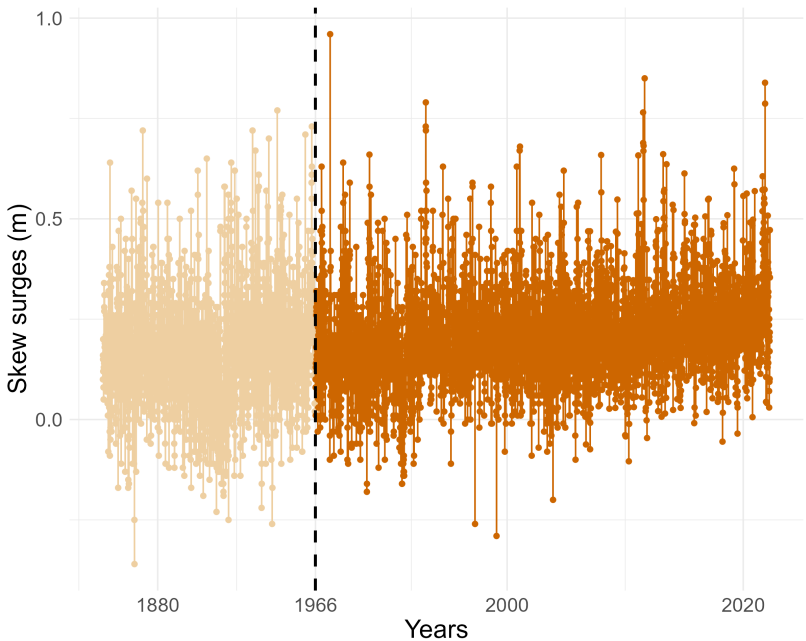
Model Errors

Mean Absolute Error/Root Mean Square Error

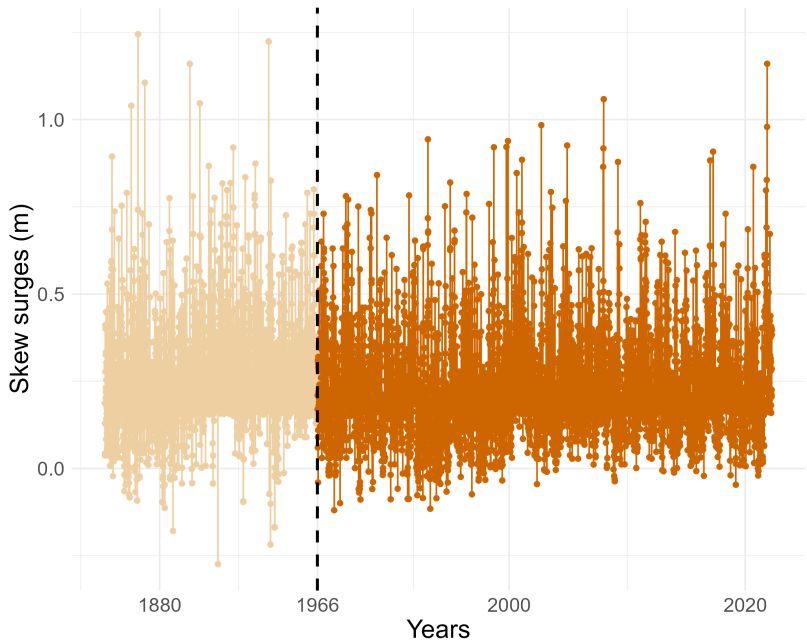


Reconstruction of the time series

the data



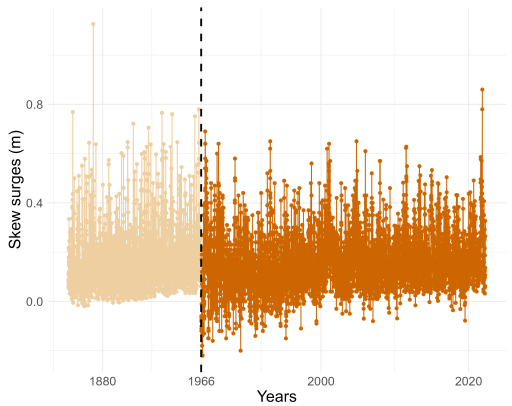
Brest skew surge time series.



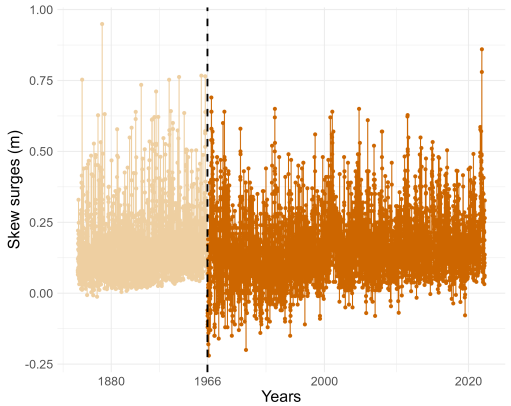
Saint-Nazaire skew surge time series.

Reconstruction of the time series

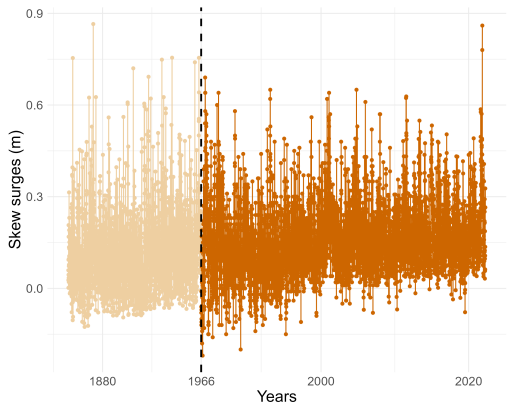
the predictions



ROX OLS



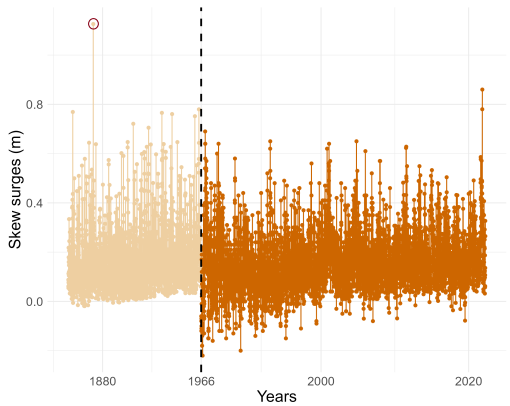
ROX RF



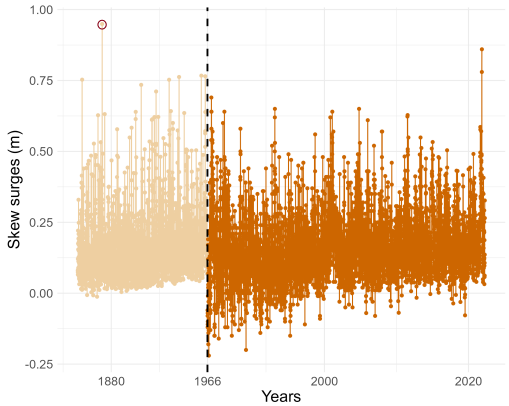
MGP

Reconstruction of the time series

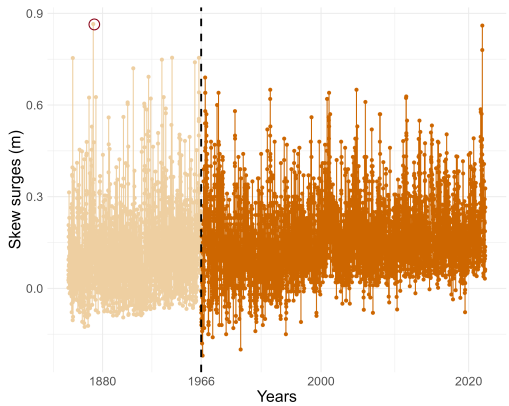
the predictions



ROX OLS



ROX RF



MGP

L'ouragan

La nuit du 31 décembre au 1er janvier a été marquée par une affreuse tempête qui a sevi sur tout le littoral du Morbihan. Plus d'une couverture à été éventrée, des cheminées sont tombées, quatres navires, dont un chargé de blé, ont été jetés sur la côte de l'île d'Arz. La mer chassée par la violence du vent a atteint une hauteur exceptionnelle dans notre golfe. L'eau du port de Vannes avait débordé, inondant le rez de chaussée de la maison de M. Dubois, négociant sur l'ancienne place du marché au froment, la cour de M. Charles Vincent et la place de la Poissonnerie. On raconte qu'à Kerbourbon près des trois sapins la rupture d'une digue a submergé un champ garni d'une abondante récolte de plantes fourragères. Le débordement se serait aussi fait sentir très fortement dans la commune de Séné.

Perspectives

- adjust the model by including meteorological variables;
- analysis of our method for improving inference on long return periods;
- use Generative AI methods to tackle this problem: how could we use all the observations at other stations?
- develop a simpler model for practionners.

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!