Joint Modeling Extremal Sea-Levels Dependency across different French Atlantic Coast Stations

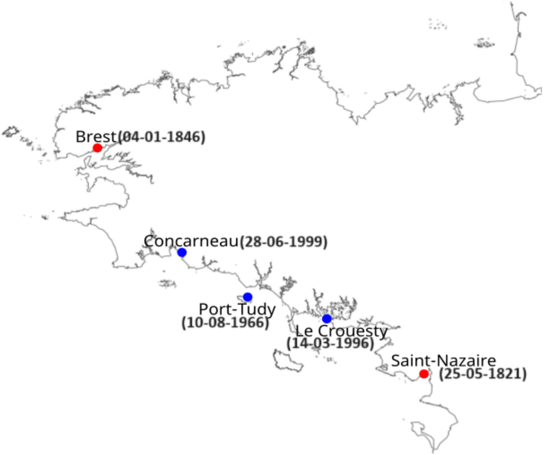
Nathan Huet[1]
Joint work with Philippe Naveau[2] and Anne Sabourin[3]

[1] Télécom Paris, Institut polytechnique de Paris, LTCI, Palaiseau. [2] Laboratoire des Sciences du Climat et l'Environnement, CNRS, Gif-sur-Yvette. [3] Université Paris Cité, CNRS, MAP5, Paris.

International Sea Level Workshop, Brest, June 2024

Context

Amount of tide-gauge data varies from station to station



Problem: inferences of high return levels with limited historical measurements suffer from massive uncertainty

SHOM data:

maximal (over a tide) **observed sea levels**

Two input stations:

- Brest, first measure in 04-01-1846;
- Saint-Nazaire, first measure in 25-05-1821;

One output station:

- Port Tudy, first measure in 10-08-1966

Notation: $(X, Y) = (X_B, X_N, Y)$ represents a sea level triplet at Brest, Saint-Nazaire and Port Tudy.

NB: exactly the same study could be done by replacing sea levels with skew surges.

Splitting

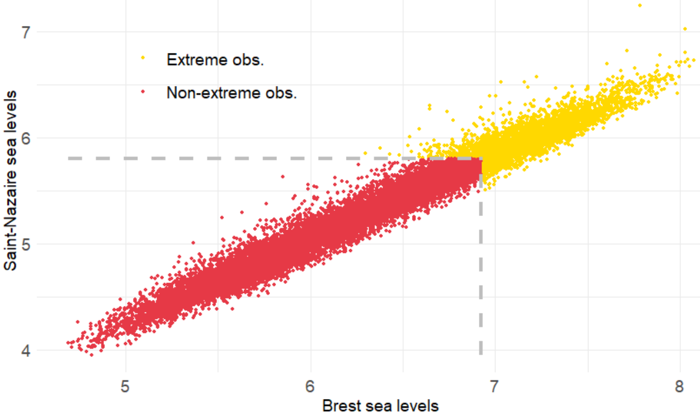
Training set: from 01-01-2000 to 31-12-2023.

Test set: from 10-08-1966 to 31-12-1999.

Only extreme observations are retained

$$X_B \geq q_B^{0.8} \text{ or } X_N \geq q_N^{0.8},$$

with $q_B^{0.8}$ and $q_N^{0.8}$ 0.8-empirical quantiles.



Two proposed approaches

- predictive approach: construction of regression predictive function
  $\Rightarrow$ **Regression On eXtreme ANglEs (ROXANE)** N.H., S. Clémençon and A. Sabourin (2024).

- modeling approach: fitting of parametric densities
  $\Rightarrow$ **Multivariate Generalized Pareto (MGP) Density Fitting**, A. Kiriliouk, H. Rootzén, J. Segers and J. L. Wadsworth (2019); J. Legrand, P. Ailliot, P. Naveau and N. Raillard (2023).

Marginal Modeling

Classic Choice: Generalized Pareto Distribution :

$$F_{\hat{\sigma},\hat{\xi}}(x) = 1 - (1 + \frac{\hat{\xi}x}{\hat{\sigma}})_{+}^{-1/\hat{\xi}}$$

$\rightsquigarrow$ empirical cdf below q and GPD cdf above q:

$$\hat{F}(x) = \begin{cases} \hat{F}_{emp}(x) & \text{if } x < q \\ (1 - (1 - \hat{F}_{emp}(q))(1 + \frac{\hat{\xi}}{\hat{\sigma}}(x - q))^{-1/\hat{\xi}}) & \text{if } x \geq q \end{cases}$$

Regression On eXtreme ANglEs

N.H., S. Clémençon and A. Sabourin (2024): Regression predictive model in extreme regions

Fréchet transformation: $1/(1-\hat{F}(x))$

Main result of the paper:
a regression function $\hat{g}$ can be optimally constructed using only the **angle** of the input variable in extreme regions

ROXANE: Rationale

Reminder: $X \in RV_{-\alpha}(\mathbb{R}^d)$ if $\lim_{t\to+\infty} b(t)\mathbb{P}(t^{-1}X \in B) = \mu(B)$

Working Assumption *(Conditional Regular Variation)*

$$\lim_{t\to+\infty} b(t)\mathbb{P}(t^{-1}X \in A, Y \in C) = \mu(A \times C)$$

Consequence: existence of (X_∞, Y_∞) s.t.

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

ROXANE: Rationale

$\rightsquigarrow$ an optimal regression function can be constructed predicting Y_∞ using only the angle of X_∞, *i.e.* using only $X_\infty / \|X_\infty\|$.

intuition: the exponent measure μ decomposes as

$$\mu(rB \times C) = r^{-\alpha}\Phi(B \times C),$$

then Y_∞ is independent of $\|X_\infty\|$.

$\Rightarrow$ all the information in X_∞ to predict Y_∞ is contained in $\Theta_\infty = X_\infty / \|X_\infty\|$.

NB: statistical guarantees related to the finite-distance estimator of the Bayes regression function in the article.

ROXANE algorithm

In practice, how does it works?

approximate decomposition of Y:

$$Y \approx \|X\| \times \underbrace{Y/\|(X_B, X_N, Y)\|}_{\text{unknown}}$$

$\Rightarrow Y/\|(X_B, X_N, Y)\|$ optimally estimated using only $X/\|X\|$

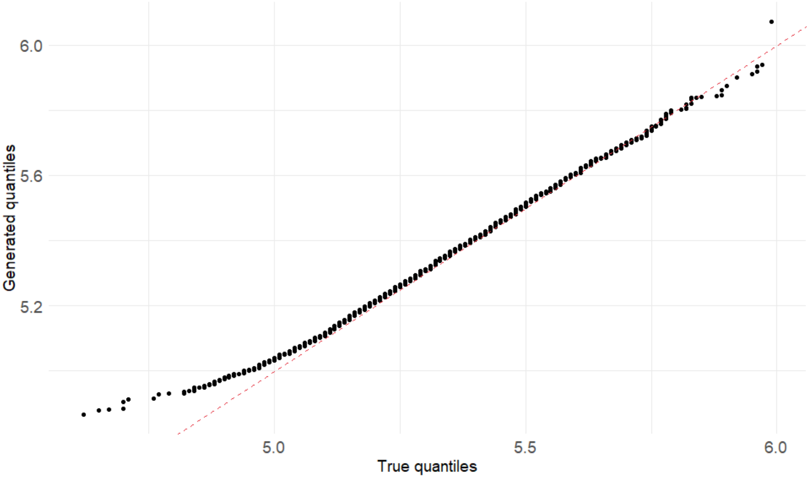
$$\hat{g}(X/\|X\|) \approx Y/\|(X_B, X_N, Y)\|$$

$\Rightarrow \hat{Y} \approx \|X\| \times \hat{g}(X/\|X\|)$

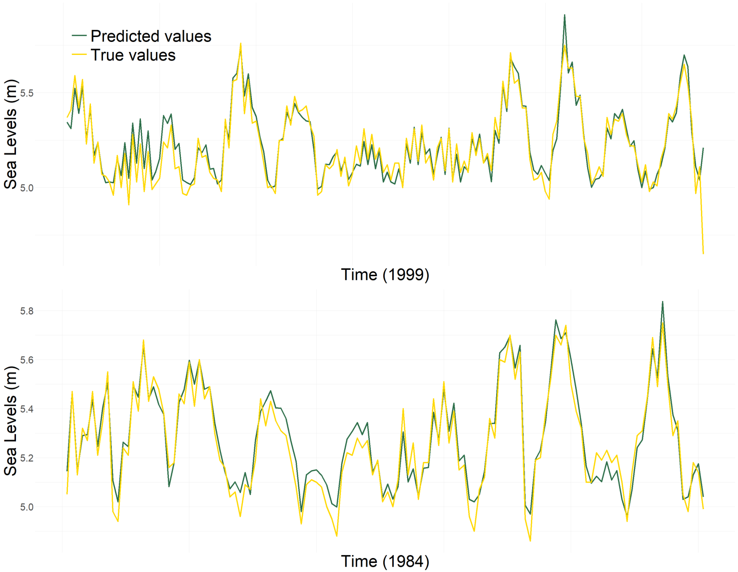
Advantages of this method:

- dimension reduction;
- adapted to high dimensional problems.

QQ-plot true values *vs* OLS ROXANE predictions



Reconstruction via OLS ROXANE 1999 and 1984



Multivariate Generalized Pareto Distribution

a classic RV assumption of $X \sim F$

$$F^n(a_n x + b_n) \to G(x),$$

with G a GEV distribution (Weibull in our case)

$$\Rightarrow \mathscr{L}\left(\frac{X - b_n}{a_n} \mid X \nleq b_n\right) \to H$$

where H is following a **MGP distribution**.

$\Rightarrow$ same extremes than for our study

Multivariate Generalized Pareto Distribution

if H follows a MGPD, then

$$\mathbb{P}(H_j \leq h_j \mid H_j > 0) = (1 + \frac{\xi_j h_j}{\sigma_j})_+^{-1/\xi_j}$$

Exponential scale transformation: $-\log(1 - F_{\hat{\sigma},\hat{\xi}}(x))$

Multivariate Generalized Pareto Distribution

Theorem 7 in H. Rootzén, J. Segers and J. L. Wadsworth (2018):

$$\begin{pmatrix} X_B \\ X_N \\ Y \end{pmatrix} = E + \begin{pmatrix} T_B \\ T_N \\ T_Y \end{pmatrix} - \max(T_B, T_N, T_Y)$$

with E a unit exponential random variable, independent of T.

$$\text{model for } \begin{pmatrix} T_B \\ T_N \\ T_Y \end{pmatrix} \Rightarrow \text{model for } \begin{pmatrix} X_B \\ X_N \\ Y \end{pmatrix}$$

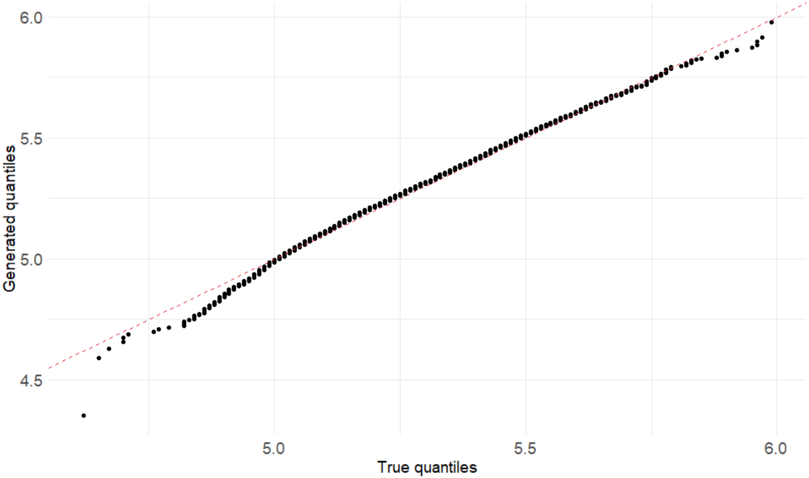
Predictive procedure via MGP dens.

1. density candidates for T;
2. fit of density candidates to (X_B, X_N, Y);
3. generate 100 values for each Y of the test set by reject sampling, according to each (X_B, X_N);
4. predict each Y by Monte Carlo averaging of the 100 generated values.

Advantages of the method:

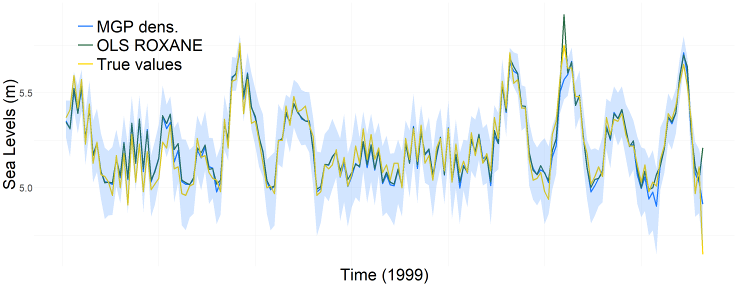
- sampling of new extreme data;
- confidence intervals on the prediction.

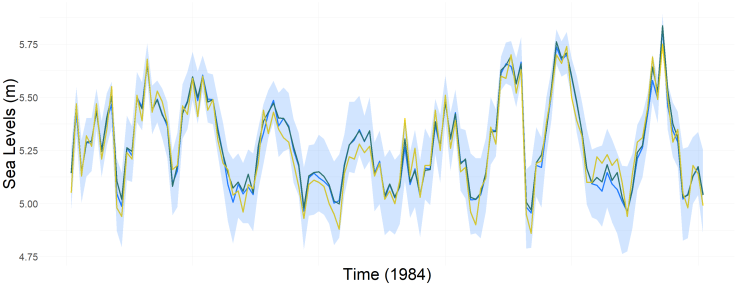
QQ-plot true values *vs* MGP dens. predictions



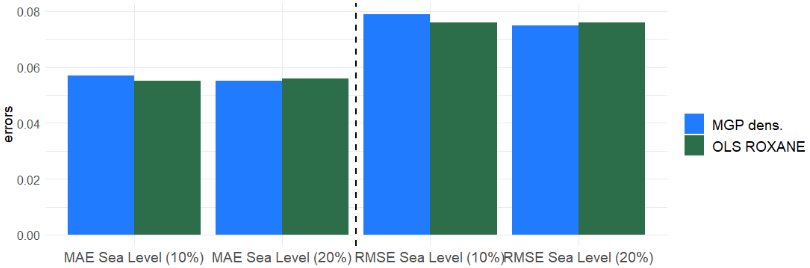
Global 0.95-coverage probability of the MGP dens. method : **0.94**.

Reconstruction 1999 and 1984

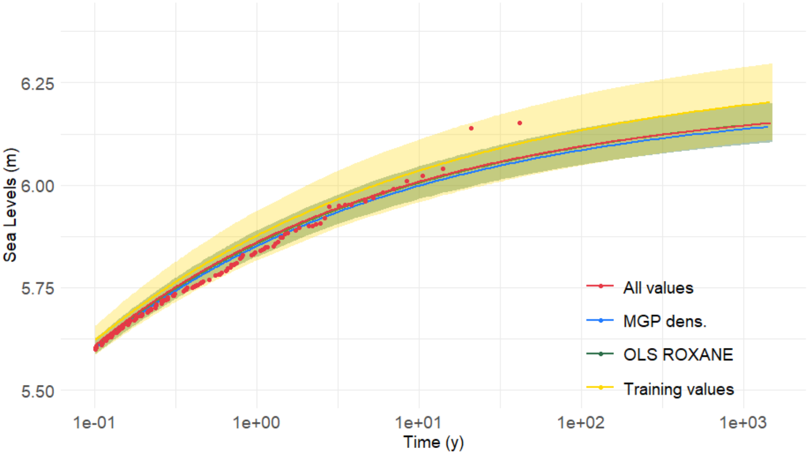




Comparison of the two methods



Return period



EST. PARAM./DATA	ALL OBS.	TRAINING SET	MGP DENS.	OLS ROXANE
$\hat{\sigma}$	0.225	0.211	0.220	0.220
$\hat{\xi}$	-0.233	-0.208	-0.229	-0.228

On going/Remaining Work

- adjust the model for the smallest extremes;
- include other variables in the models;
- convolution method for return period inference via skew surges.

On going/Remaining Work

- adjust the model for the smallest extremes;
- include other variables in the models;
- convolution method for return period inference via skew surges.

Thank you for your attention !

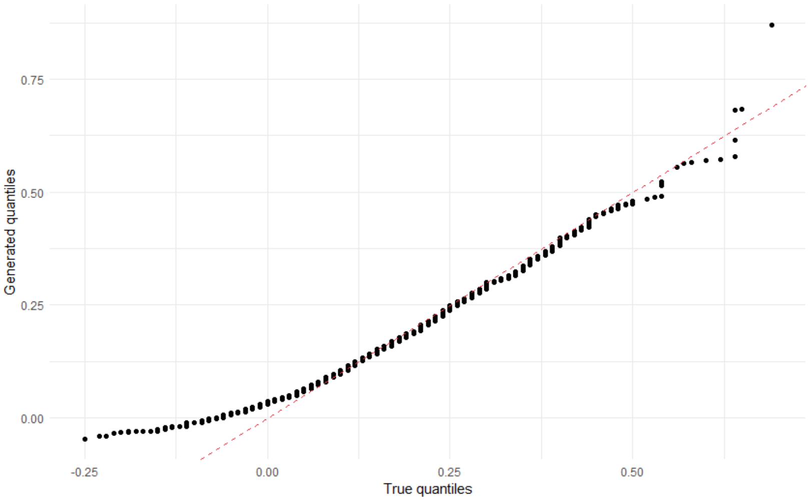
Bibliography

- N. H. and S. Clémençon and A. Sabourin (2024) *On Regression in Extreme Regions*, arXiv:2303.03084.
- A. Kiriliouk, H. Rootzén, J. Segers and J. L. Wadsworth (2019) *Peaks Over Thresholds Modeling With Multivariate Generalized Pareto Distributions*, Technometrics, 61:1, 123-135.
- J. Legrand, P. Ailliot, P. Naveau and N. Raillard (2023) *Joint stochastic simulation of extreme coastal and offshore significant wave heights*, The Annals of Applied Statistics 17 (4), 3363-3383.
- H. Rootzén, J. Segers and J. L. Wadsworth (2018) *Multivariate generalized Pareto distributions: Parametrizations, representations, and properties*, Journal of Multivariate Analysis, 165:117–131.

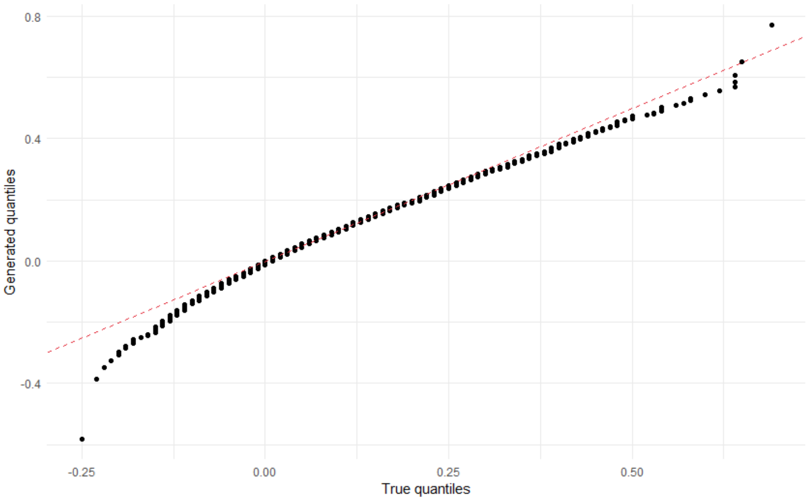
Appendix

Skew surges results

QQ-plot true values *vs* OLS ROXANE predictions



QQ-plot true values *vs* MGP dens. predictions



Reconstruction 1999 and 1984