Estimation of extreme quantiles from heavy-tailed distributions with neural networks

Michaël Allouche   Stéphane Girard   Emmanuel Gobet

February 2025

Motivations

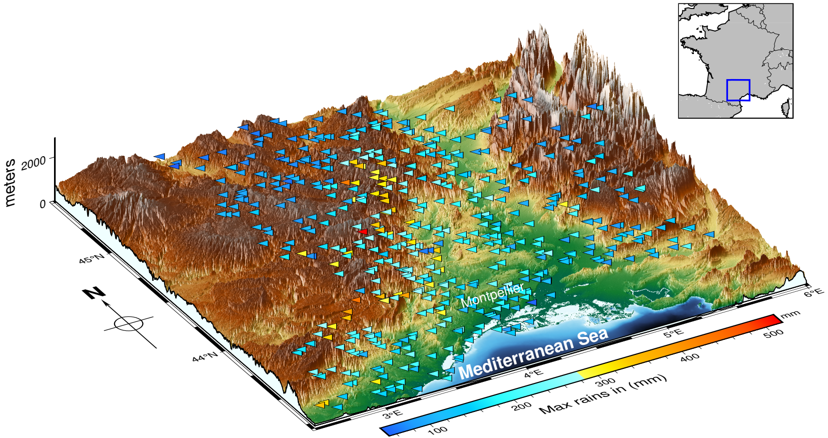


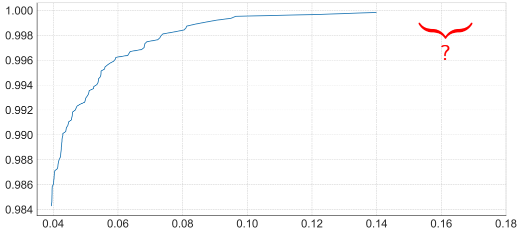
Figure: Historical (1958-2000) daily rainfall maxima in mm per station in the Cévennes-Vivarais region of France.

How to compute extreme return levels at ungauged locations:

- Extrapolation (above the sample maxima at each station),
- Interpolation (where there is no measure).

Extrapolation problem

Let $X_1, \ldots, X_n$ be an i.i.d sample from an unknown cdf F, with an associated quantile function q. Let us denote by $X_{1,n} \leq \cdots \leq X_{n,n}$ the order statistics.



Empirical c.d.f estimated on the negative returns of the Paris stock index (CAC 40)

Objective. NN estimation of $q(1-\alpha_n)$ such that $n\alpha_n \to 0$ as $n \to \infty$

Challenges.

- $q(1-\alpha_n)$ is larger than the sample maxima $X_{n,n}$ (with high proba).
- Single-layer NN not able to simulate around the max [Allouche et al., 2022].

Statistical framework

Focusing on heavy-tailed distributions ($F \in \mathrm{MDA}(\text{Fréchet})$), the tail quantile function $q(1-1/t), \forall t > 1$, is **regularly varying** with tail index $\gamma > 0$ and $q(1-1/t) = t^{\gamma} L(t)$ with

$$L(zt)/L(t) \to 1 \text{ as } t \to \infty, \forall z > 0$$

Idea. Choose an intermediate sequence δ_n s.t. $k_n := \lfloor n\delta_n \rfloor \to \infty, n \to \infty$,

$$\begin{aligned} \log q(1-\alpha_n) - \log q(1-\delta_n) &= \gamma \log(\delta_n/\alpha_n) + \varphi\big(\log(\delta_n/\alpha_n), \log(1/\delta_n)\big) \\ &=: f\big(\log(\delta_n/\alpha_n), \log(1/\delta_n)\big) \end{aligned}$$

with

$$(x_1 > 0, x_2 > 0) \mapsto \varphi(x_1, x_2) := \log\left(\frac{L(\exp(x_1 + x_2))}{L(\exp(x_2))}\right)$$

Unknown quantities.

1. Intermediate quantile $q(1-\delta_n)$
2. Tail index γ
3. Log-spacing function $\varphi(\cdot, \cdot)$

Weissman. [Weissman, 1978]

1. $X_{n-k+1,n}$
2. $\hat{\gamma}(k)$ [Hill, 1975]
3. 0

Bias correction (second order)

Second order condition. There exist $\gamma > 0, \rho_2 \leq 0$ and a function A_2 with $A_2(t) \to 0$ as $t \to \infty$ s.t. for all $z \geq 1$

$$\log\left(\frac{L(zt)}{L(t)}\right) = A_2(t)\int_1^z z_2^{\rho_2-1}\,\mathrm{d}z_2 + o(A_2(t)), \qquad \text{as } t \to \infty$$

Ignoring the $o(\cdot)$ term and assuming (Hall-Welsh model)

$$A_2(t) = \gamma\beta_2 t^{\rho_2}$$

with $\beta_2 \neq 0$ and $\rho_2 < 0$, give a parametric approximation of $\varphi(x_1, x_2)$ as

$$\begin{aligned}\widetilde{\varphi}_\theta(x_1, x_2) &= \gamma\beta_2 \exp(\rho_2 x_2)(\exp(\rho_2 x_1) - 1)/\rho_2 \\ &= \gamma\beta_2\Big(\sigma^{\mathtt{E}}(\rho_2(x_1 + x_2)) - \sigma^{\mathtt{E}}(\rho_2 x_2)\Big)/\rho_2,\end{aligned}$$

with $\theta = (\gamma, \rho_2, \beta_2)$ and where $\sigma^{\mathtt{E}}(x) = \mathbb{1}_{\{x\geq 0\}}x + \mathbb{1}_{\{x<0\}}(\exp(x) - 1)$ is the **eLU** function.

Bias correction (J-th order)

***J*-th order condition.** There exist $\gamma > 0$, and $\forall j \in \{2, \dots, J\}$, $\rho_j \leq 0$ and functions A_j with $A_j(t) \to 0$ as $t \to \infty$ s.t. for all $z \geq 1$

$$\log\left(\frac{L(zt)}{L(t)}\right) = \sum_{j=2}^{J} \prod_{\ell=2}^{j} A_\ell(t) R_j(z) + o\left(\prod_{j=2}^{J} A_j(t)\right) \quad \text{as } t \to \infty, \qquad (1)$$

$$R_j(z) = \int_1^z z_2^{\rho_2 - 1} \int_1^{z_2} z_3^{\rho_3 - 1} \dots \int_1^{z_{j-1}} z_j^{\rho_j - 1} \, \mathrm{d}z_j \dots \mathrm{d}z_3 \, \mathrm{d}z_2.$$

Proposition

Assume the J-th order condition holds with $A_j(t) = c_j t^{\rho_j}$, where $c_j \neq 0$ and $\rho_j < 0$ for $j \in \{2, \dots, J\}$. Then, for all $x_1 > 0$ and $x_2 > 0$

$$\varphi(x_1, x_2) = \sum_{i=1}^{J(J-1)/2} w_i^{(1)} \left(\sigma^{\mathtt{E}}\left(w_i^{(2)} x_1 + w_i^{(3)} x_2\right) - \sigma^{\mathtt{E}}(w_i^{(4)} x_2)\right) + o\,(\dots)$$

with $w_i^{(1)} \in \mathbb{R}$, $w_i^{(2)} < 0$, $w_i^{(3)} < 0$, $w_i^{(4)} < 0$, $\forall i \in \{1, \dots, J(J-1)/2\}$.

Results

NN approximation and estimation. For $k \in \{2, \ldots, n-1\}$,

$$\tilde{q}_{\tilde{\phi}}^{\mathtt{NN}_J}(1-\alpha_n; 1-k/n) := q(1-k/n) \exp\left(\tilde{f}_{\tilde{\phi}}^{\mathtt{NN}_J}(\log(k/(n\alpha_n)), \log(n/k))\right)$$

$$\hat{q}_{\hat{\phi}}^{\mathtt{NN}_J}(1-\alpha_n; 1-k/n) := X_{n-k+1,n} \exp\left(\tilde{f}_{\hat{\phi}}^{\mathtt{NN}_J}(\log(k/(n\alpha_n)), \log(n/k))\right)$$

where $\tilde{f}_{\hat{\phi}}^{\mathtt{NN}_J}(x_1, x_2) := \hat{w}_0 x_1 + \tilde{\varphi}_{\hat{\phi}}^{\mathtt{NN}_J}(x_1, x_2)$, with $\hat{w}_0 > 0$.

Theorem

Assume conditions of the Proposition hold with $\bar{\rho}_J := \rho_2 + \cdots + \rho_J$. *Then, there exists a* **one hidden-layer** *feedforward neural network approximation with* $J(J-1)$ *neurons and* $2J(J-1)$ *parameters such that*

$$\inf_{\tilde{\phi} \in \mathbf{\Phi}} \left| \log q(1-\alpha_n) - \log \tilde{q}_{\tilde{\phi}}^{\mathtt{NN}_J}(1-\alpha_n; 1-\delta_n) \right| = \mathcal{O}\left({\alpha_n}^{-\bar{\rho}_J}\right),$$

as $\alpha_n \to 0$ *and* $\delta_n/\alpha_n \to \infty$.

Experiments - Simulated data

- Simulate $n_R = 500$ replications of n =500 samples $X_1, \ldots, X_n$ from 7 heavy-tailed distributions parametrized by (γ, ρ_2).
- Compute the log-spacings $\hat{S}_{i,k} := \log X_{n-i+1,n} - \log X_{n-k+1,n}$ with $i \in \{1 \ldots, k-1\}$, $k \in \{2, \ldots, n-1\}$
- Fit the approximation $\tilde{f}^{\mathrm{NN}_J}_{\hat{\phi}}(\log(k/i), \log(n/k))$ by training the neural network in a regression framework $J \in [2, 5] \Leftrightarrow [2, 10]$ neurons
- Estimate at extreme quantile level $1 - \alpha_n = 1 - 1/(2n)$ and compare the RMedSE with competitors [Gomes and Pestana, 2007, Allouche et al., 2023]

Burr	NN	W	RW	CW	CH	CH_p	PRB_p	$\mathrm{CH}_{p^\star}$	$\mathrm{PRB}_{p^\star}$
$\gamma = 1$									
$\rho_2 = -1/8$	**0.3133**	-	0.8625	-	-	-	-	-	-
$\rho_2 = -1/4$	**0.1962**	-	0.5423	-	-	-	-	-	0.6617
$\rho_2 = -1/2$	0.2142	-	0.3291	-	0.0949	0.1021	0.1488	**0.0874**	0.1185
$\rho_2 = -1$	0.1877	-	0.2438	**0.1289**	0.4120	0.3737	0.3761	0.3658	0.4261
$\rho_2 = -2$	**0.1432**	0.2065	0.1488	0.2115	0.3394	0.3384	0.2893	0.2933	0.3058

RMedSE associated with eight estimators on five Burr distributions. The best result is emphasized in **bold**. Results larger than 1 are not displayed. More results in [Allouche et al., 2024].

Illustration

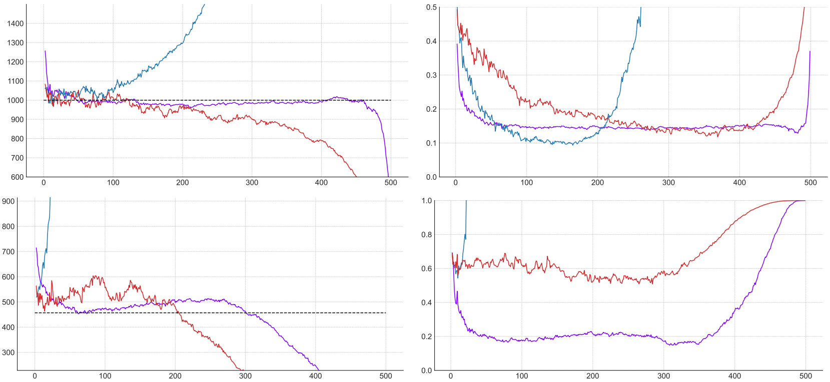
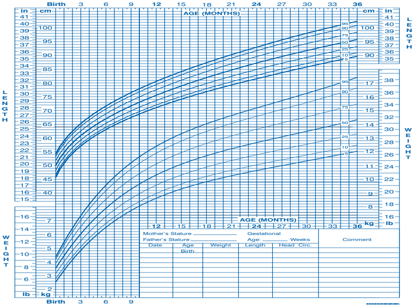


Figure: Illustration on a Burr distribution with $\gamma = 1$ and $\rho \in \{-2, -1/4\}$ (from top to bottom). Median of the estimators (left panel) of the extreme quantile (black dashed line) and RMedSE (right panel), as functions of $k \in \{2, \ldots, n-1\}$, associated with W (blue), RW (red), NN (purple).

Extension to conditional extrapolation

Suppose now X is a r.v. associated with an explanatory random vector $Y \in \mathcal{Y} \subset \mathbb{R}^{d_y}, d_y \geq 1$. Denote the conditional c.d.f by $F(\cdot \mid y)$ and the conditional quantile function by $q(\cdot \mid y)$.



Boys growth chart - source

Goal: NN estimation of $q(1-\alpha_n \mid y)$ such that $n\alpha_n \to 0$ for all y.

Challenges.

- Same as in the non-conditional framework.
- $q(1-\delta_n \mid y)$ can no longer be estimated by an order statistic.

A) Conditional Extrapolation Neural Network

Similarly to the non-conditional case, the conditional tail quantile function is supposed to be **regularly varying** $q(1-1/t \mid y) = t^{\gamma(y)} L(t \mid y)$, with conditional tail index $\gamma(y) > 0$ ($q(1-1/\cdot \mid y) \in \mathcal{RV}_{\gamma(y)}$) with

$$L(zt \mid y)/L(t \mid y) = 1 \text{ as } t \to \infty, \forall z > 0 \, (L(\cdot \mid y) \in \mathcal{RV}_0)$$

Idea. Same method as before but now all the $2J(J-1)+1$ parameters $\{(w_i^{(1)}, w_i^{(2)}, w_i^{(3)}, w_i^{(4)}), i \in \{1, \ldots, J(J-1)/2\}\}$ and γ depend of the covariate and have to be approximated by appropriate NNs:

$$\tilde{f}_{\tilde{\phi}}^{\mathtt{NN}_J}(x_1, x_2 \mid y) = \tilde{w}_{\tilde{\theta}^{(0)}}^{\mathtt{NN}}(y) x_1 + \tilde{\varphi}_{\tilde{\theta}}^{\mathtt{NN}_J}(x_1, x_2 \mid y)$$

includes $2J(J-1)+1$ deep ReLU NNs with ReLU activation functions $\sigma^{\mathtt{R}}(x) = \max(x, 0)$.

An approximation result

Theorem

Suppose the **conditional** *extensions of the assumptions of the Proposition hold, with all functions* $\{w_i^{(1)}(\cdot), \ldots, w_i^{(4)}(\cdot)\}$, $i \in \{1, \ldots, J(J-1)/2\}$, *and* $\gamma(\cdot)$ *are* **continuous** *on the compact set* $\mathcal{Y} \subset \mathbb{R}^{d_y}$.
Then, there exists a conditional **deep** *feedforward NN approximation with* $\mathcal{O}(J^2)$ *sub-networks composed by fixed* $\mathcal{O}(d_y)$ *neurons in each of the hidden layers with a minimum depth of magnitude* $\simeq {\alpha_n}^{\bar{\rho}_{\sup}/2}$ *such that*

$$\inf_{\tilde{\phi}\in\Phi} \sup_{y\in\mathcal{Y}} \left| \log q(1-\alpha_n \mid y) - \log \tilde{q}^{\mathrm{NN}_J}_{\tilde{\phi}}(1-\alpha_n; 1-\delta_n \mid y) \right| = \mathcal{O}\left({\alpha_n}^{-\bar{\rho}_{\sup}}\right),$$

as $\alpha_n \to 0$ *and* $\delta_n/\alpha_n \to \infty$ *as* $n \to \infty$, *where* $\bar{\rho}_{\sup} = \sup_{y\in\mathcal{Y}} \bar{\rho}_J(y)$ *with* $\bar{\rho}_J(y) = \rho_2(y) + \cdots + \rho_J(y)$.

Conditional neural network estimator.

$$\hat{q}^{\mathrm{NN}_J}_{\hat{\phi}}(1-\alpha_n; 1-\delta_n \mid y) := \hat{q}(1-\delta_n \mid y) \exp\left(\tilde{f}^{\mathrm{NN}_J}_{\hat{\phi}}\left(\log\left(\delta_n/\alpha_n\right), \log\left(1/\delta_n\right) \mid y\right)\right)$$

B) Location Dispersion Neural Network

Location-dispersion regression model [Van Keilegom and Wang, 2010] :

$$X = a(Y) + b(Y)Z,$$

where $a : \mathcal{Y} \to \mathbb{R}$ and $b : \mathcal{Y} \to \mathbb{R}^+$ are defined respectively as the **location** and the **dispersion** functions while $Z \in \mathbb{R}$ is a heavy-tailed r.v. with tail index γ and quantile function $q_Z(\cdot)$.

Idea. Consider three levels of quantiles $0 < \alpha_n < \delta_n < \tau_n < 1$. Since

$$q(1 - \alpha_n \mid y) = a(y) + b(y) q_Z(1 - \alpha_n)$$

the following combination of quantiles

$$\frac{q(1 - \alpha_n \mid y) - q(1 - \delta_n \mid y)}{q(1 - \delta_n \mid y) - q(1 - \tau_n \mid y)} = g\big(\log(\delta_n/\alpha_n), \log(1/\delta_n), \log(\delta_n/\tau_n)\big)$$

is independent of the covariate y. One can apply the non conditional approximation method to g instead of f.

An approximation result

Theorem

Assume the location-dispersion model and conditions of Proposition hold for. Suppose $a(\cdot)$ and $b(\cdot)$ are continuous functions on $\mathcal{Y}$ and that $b(\cdot)$ is bounded from below by a positive constant. Then, there exists a **one hidden-layer** *NN approximation such that*

$$\inf_{\tilde{\phi}\in\Phi} \sup_{y\in\mathcal{Y}} \left| \log q(1-\alpha_n \mid y) - \log \tilde{q}_{\tilde{\phi}}^{\mathrm{NN}_J}(1-\alpha_n; 1-\delta_n, 1-\tau_n \mid y) \right|$$
$$= \mathcal{O}({\alpha_n}^{-\bar{\rho}_J}) + \mathcal{O}({\tau_n}^{-\bar{\rho}_J-\gamma}{\delta_n}^{\gamma})$$

with $\alpha_n \to 0$, $\delta_n/\tau_n \to 0$ *and* $\delta_n/\alpha_n \to \infty$ *as* $n \to \infty$.

Conditional neural network estimator.

$$\hat{q}_{\hat{\phi}}^{\mathrm{NN}_J}(1-\alpha_n; 1-\delta_n, 1-\tau_n \mid y) = \hat{q}(1-\delta_n \mid y)$$
$$+ \left(\hat{q}(1-\delta_n \mid y) - \hat{q}(1-\tau_n \mid y)\right) \tilde{g}_{\hat{\phi}}^{\mathrm{NN}_J}\left(\log(\delta_n/\alpha_n), \log(1/\delta_n), \log(\delta_n/\tau_n)\right)$$

Experiments - Real data

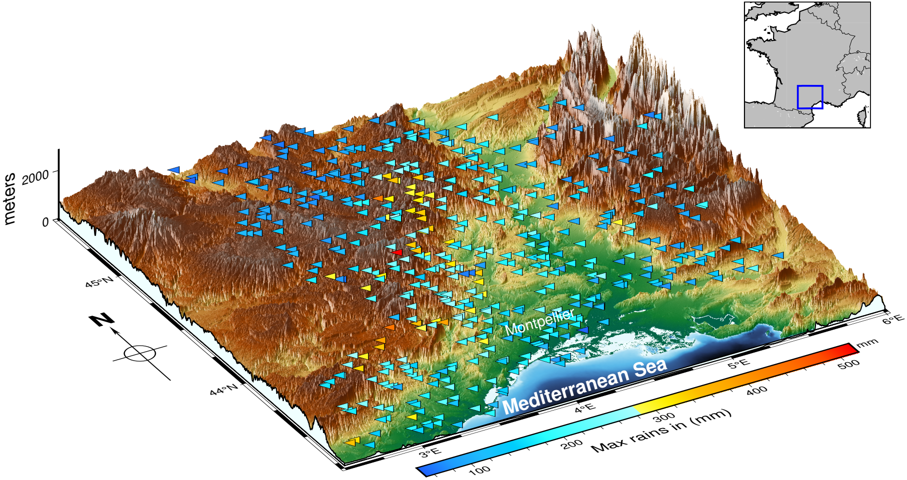


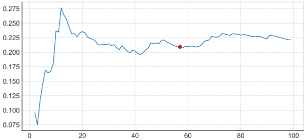
Figure: Historical (1958-2000) daily rainfall maxima in mm per station in the Cévennes-Vivarais region of France.

Experiments - Real data

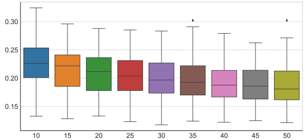
Experimental design

- Data: $n_D = 15{,}706$ rainfalls $X(Y) \in \mathbb{R}$ at $n_S = 524$ stations in the Cévennes-Vivarais region given a covariate $Y \in \mathbb{R}^3$ (long., lat., alt.)
- Estimate the intermediate conditional quantile by K-NN.
  - fix n_K neighbors and apply K-NN on Y using the Mahalanobis distance $\sqrt{(Y_t - Y_{t'})^\top \Sigma^{-1} (Y_t - Y_{t'})}, (t, t') \in \{1, \ldots, n_S\}^2$,
  - merge the historical values to the $n_K - 1$ closest stations, leading to $n_o = n_D \times n_K$ observations assumed i.i.d within each neighborhood.
- Train the NNs with the highest unique historical values $\left\{X^{(n_o - i + 1, n_o)}(Y_t), i \in \{2, \ldots, n_h\}, n_h \in \{2, \ldots, n_o\}\right\}$ for all $t \in \{1, \ldots, n_S\}$.
- Estimate the quantiles at level $1 - \alpha_n = 1 - 1/n_o$ and compare with all maximum order statistics $X^{(n_o, n_o)}(Y_t)$ for all $t \in \{1, \ldots, n_S\}$.

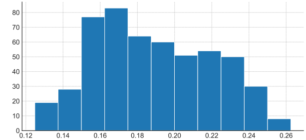
Influence of hyperparameters n_K, n_h



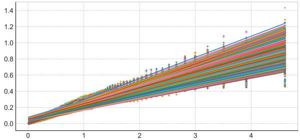
(a)



(b)



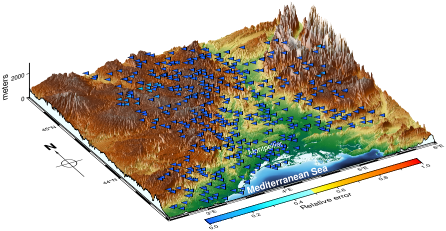
(c)



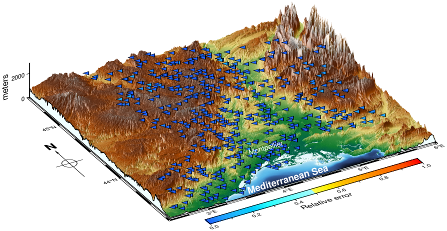
(d)

Figure: Illustrations on real data. (a) Example of Hill estimate as a function of $k \in \{2, \ldots, n_h - 1\}$, within the neighborhood of a given station with $n_h = 100$ and $n_K = 45$. The selected $k^\star$ is depicted by the red circle. (b) Box-plots of estimated $\hat{\gamma}$'s as functions of n_K with $n_h = 100$. (c) Histogram of estimated $\hat{\gamma}$'s for all stations $t \in \{1, \ldots, n_S\}$ with $n_K = 45$ and $n_h = 100$. (d) quantile-quantile plot $\log(n_h/i) \mapsto \log(X^{(n_o - i+1, n_o)}(Y_t)) - \log(X^{(n_o - n_h+1, n_o)}(Y_t))$, $t \in \{1, \ldots, n_S\}$, $i \in \{1, \ldots, n_h - 1\}$ with $n_h = 100$ and $n_K = 45$.

Estimation of conditional extreme quantile



(a) CENN (RMedSE=0.0047)
$\sim 2,000$ parameters, $2.5\,10^6$ data



(b) LDNN (RMedSE=0.0022)
~ 10 parameters, $82\,10^6$ data

Figure: Estimation of the conditional extreme quantile at order $1 - \alpha_n = 1 - 1/n_o$ at each station. Squared relative error associated with the CENN (a) and LDNN (b) models.

Spatial interpolation

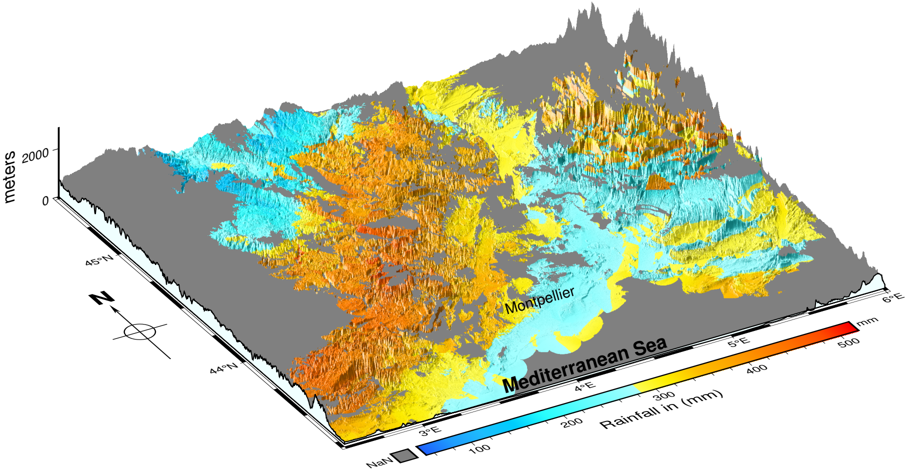


Figure: Spatial interpolation by the CENN quantile estimator at order $1 - \alpha_n = 1 - 1/n_o$. The gray region corresponds to an area where no interpolation is performed.

Conclusion

- Propose a NN architecture as a natural bias reduced extreme quantile estimator,
- Prove uniform convergence rates of the NN approximation error in extreme quantile estimation in both non-conditional and conditional settings,
- Outperform other competitors in hard heavy-tailed simulations,
- Illustrate the conditional extrapolation on real data.
- Extension to the estimation of more general risk measures [Allouche et al., 2025].

References I

Allouche, M., El Methni, J., and Girard, S. (2023).
A refined Weissman estimator for extreme quantiles.
Extremes, 26(3):545–572.

Allouche, M., Girard, S., and Gobet, E. (2022).
EV-GAN: Simulation of extreme events with ReLU neural networks.
Journal of Machine Learning Research, 23(150):1–39.

Allouche, M., Girard, S., and Gobet, E. (2024).
Estimation of extreme quantiles from heavy-tailed distributions with neural networks.
Statistics and Computing, 34:12.

Allouche, M., Girard, S., and Gobet, E. (2025).
Learning extreme expected shortfall and conditional tail moments with neural networks. application to cryptocurrency data.
Neural Networks, 182:106903.

References II

Gomes, M. and Pestana, D. (2007).
A sturdy reduced-bias extreme quantile (VaR) estimator.
Journal of the American Statistical Association, 102(477):280–292.

Hill, B. M. (1975).
A simple general approach to inference about the tail of a distribution.
The Annals of Statistics, 3(5):1163–1174.

Van Keilegom, I. and Wang, L. (2010).
Semiparametric modeling and estimation of heteroscedasticity in regression analysis of cross-sectional data.
Electronic Journal of Statistics, 4:133–160.

Weissman, I. (1978).
Estimation of parameters and large quantiles based on the k largest observations.
Journal of the American Statistical Association, 73(364):812–815.