Spatial analysis of extreme rainfalls In the Cévennes-Vivarais region

using a nearest neighbor approach

Caroline Bernard-Michel[(1)], Laurent Gardes[(1)], Stéphane Girard[(1)], Gilles Molinié[(2)]

[(1)] INRIA Rhône-Alpes, Team Mistis,
655 avenue de l'Europe, Montbonnot,
38334 Saint-Ismier Cedex, France.
E-mail: {Laurent.Gardes, Stephane.Girard}@inrialpes.fr
http://mistis.inrialpes.fr

[(2)] Université Joseph Fourier de Grenoble,
Laboratoire d'Etude des Transferts en Hydrologie et Environnement.
E-mail: Gilles.Molinie@hmg.inpg.fr
http://www.lthe.hmg.inpg.fr

This work is supported by the French Research Agency (ANR) through its VMC2007 program (Vulnérabilité: Milieux, Climats).

1. Introduction

The problem.

- Estimation of extreme quantiles associated to a random variable Y.
- Some covariate x is recorded simultaneously with Y.
- The extreme quantile of Y given x depends on x, and is thus refered to as a conditional extreme quantile.

Our approach: Combination of a nearest neighbor approach with extreme-value methods.

Motivating application: Estimation of return periods associated to extreme rainfalls as a function of the geographical location (x is a three-dimensional covariate involving the longitude, latitude and altitude).

Framework. E a metric space associated to a metric d.

- **Model**: The conditional tail quantile function of Y given $x \in E$ is, for all $\alpha \in (0,1]$,

$$q(\alpha, x) = F^{\leftarrow}(1-\alpha, x) = \sup\{y > 0,\ F(y,x) \leq 1-\alpha\} = \alpha^{-\gamma(x)} \ell(\alpha^{-1}, x),$$

  where

  - $\gamma(x)$ is the conditional tail-index, an unknown positive function of the covariate x,
  - for x fixed, $\ell(.,x)$ is a slowly-varying function, *i.e.* for $v > 0$,

$$\lim_{y\to\infty} \frac{\ell(vy,x)}{\ell(y,x)} = 1.$$

  Thus, the conditional distribution of Y given x is heavy-tailed.

- **Data**: A sample $(Y_1, x_1), \ldots, (Y_n, x_n)$ iid from the above model where the design points $x_1, \ldots, x_n$ are non random points in E.

Goals. For a given $t \in E$, estimate

- the conditional tail-index $\gamma(t)$,
- the conditional extreme quantiles $q(\alpha_{n,t}, t)$ where $\alpha_{n,t} \to 0$ as $n \to \infty$.

2. Nearest neighbor estimators ...

- **Number of nearest neighbors**: $m_{n,t}$ a positive sequence tending to ∞,
- **Selected observations**: $\{Z_i,\ i = 1, \ldots, m_{n,t}\}$ the response variables Y_i associated to the $m_{n,t}$ nearest covariates $x_i^\star$ of t.
- **Corresponding order statistics**: $Z_{1,m_{n,t}} \leq \ldots \leq Z_{m_{n,t},m_{n,t}}$,
- **Intermediate sequence**: $k_{n,t} \to \infty$ and $k_{n,t}/m_{n,t} \to 0$,
- **Conditional tail-index estimators:** A weighted sum of the rescaled log-spacings between the largest selected observations

$$\hat{\gamma}_n(t,a,\lambda) = \sum_{i=1}^{k_{n,t}} i \log\left(\frac{Z_{m_{n,t}-i+1,m_{n,t}}}{Z_{m_{n,t}-i,m_{n,t}}}\right) p\left(i/k_{n,t},a,\lambda\right) \Big/ \sum_{i=1}^{k_{n,t}} p\left(i/k_{n,t},a,\lambda\right)$$

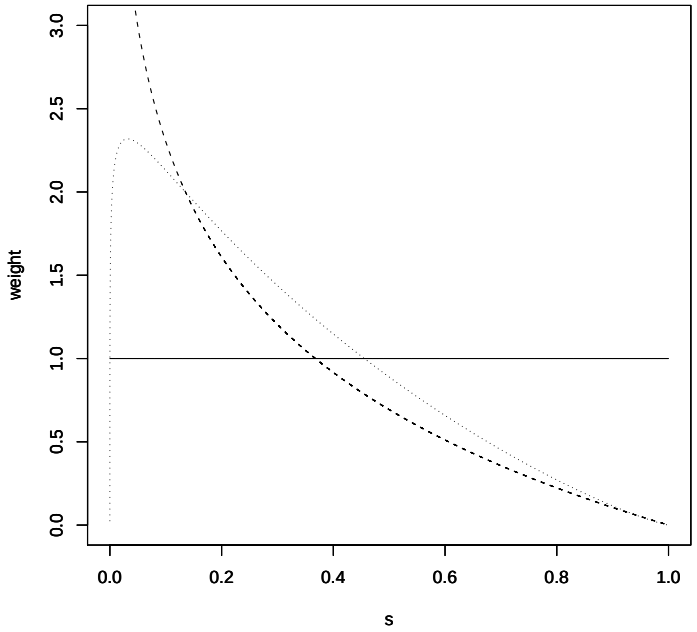
- **Conditional extreme quantile estimators:** A Weissman type estimator

$$\hat{q}(\alpha_{n,t},t) = Z_{m_{n,t}-k_{n,t}+1,m_{n,t}} \left(\frac{k_{n,t}}{m_{n,t}\alpha_{n,t}}\right)^{\hat{\gamma}_n(t,a,\lambda)}$$

... with log-gamma weights

For all $s \in (0,1]$, $a \geq 1$, $\lambda \in (0,1]$.

$$p(s,a,\lambda) = \frac{\lambda^{-a}}{\Gamma(a)} s^{1/\lambda - 1} (-\log(s))^{a-1}.$$



$a = 1$, $\lambda = 1$: full line,
$a > 1$, $\lambda = 1$: dashed line,
$a > 1$, $\lambda > 1$: dotted line.

3. Asymptotic results

Assumptions on the conditional distribution

- The slowly varying function $\ell(.,t)$ is **normalized**, *i.e.* its Karamata representation can be written as

$$\ell(\alpha^{-1},t) = c(t)\exp\left\{\int_1^{\alpha^{-1}} \frac{\Delta(v,t)}{v} dv\right\},$$

- The function $|\Delta(.,t)|$ is **regularly varying** with index $\rho(t) < 0$,
- The function $|\Delta(.,t)|$ is **ultimately decreasing**.

Controlling the regularity. The largest oscillation of the log-quantile function with respect to its second variable is defined for all $\beta \in (0, 1/2)$ as

$$\omega_n(\beta) = \sup\left\{\left|\log\left(\frac{q(\alpha,x)}{q(\alpha,x')}\right)\right| \ , \alpha \in (\beta, 1-\beta) \ , (x,x') \in \{t, x_1^\star, \ldots, x_{m_{n,t}}^\star\}^2\right\}.$$

Asymptotic normality

Theorem. If, for some $\delta > 0$, $k_{n,t}^{1/2}\Delta(m_{n,t}/k_{n,t}, t) \to \xi(t) \in \mathbb{R}$ and $k_{n,t}^2\omega_n(m_{n,t}^{-(1+\delta)}) \to 0$ then

$$k_{n,t}^{1/2}\left(\hat{\gamma}_n(t,a,\lambda) - \gamma(t) - \Delta\left(\frac{m_{n,t}}{k_{n,t}}, t\right)\mathcal{AB}(a,\lambda,\rho(t))\right)$$

converges in distribution to a $\mathcal{N}\left(0, \gamma^2(t)\mathcal{AV}(a,\lambda)\right)$ random variable, with

$$\mathcal{AB}(a,\lambda,\rho(t)) = (1-\lambda\rho(t))^{-a} \text{ and } \mathcal{AV}(a,\lambda) = \frac{\Gamma(2a-1)}{\lambda\Gamma^2(a)}(2-\lambda)^{1-2a}.$$

If, moreover, $\alpha_{n,t} < k_{n,t}/m_{n,t}$ then

$$\frac{k_{n,t}^{1/2}}{\log\left(\frac{k_{n,t}}{m_{n,t}\alpha_{n,t}}\right)}\left(\log\left(\frac{\hat{q}(\alpha_{n,t},t)}{q(\alpha_{n,t},t)}\right) - \log\left(\frac{k_{n,t}}{m_{n,t}\alpha_{n,t}}\right)\Delta\left(\frac{m_{n,t}}{k_{n,t}}, t\right)\mathcal{AB}(a,\lambda,\rho(t))\right)$$

has the same Gaussian limiting distribution.

Remark 1. If ℓ does not depend on the covariate, condition $k_{n,t}^2 \omega_n(m_{n,t}^{-(1+\delta)}) \to 0$ reduces to a regularity condition on the tail-index:

$$k_{n,t}^2 \log(m_{n,t}) \sup_{(x,x') \in \{t, x_1^\star, \ldots, x_{m_{n,t}}^\star\}^2} |\gamma(x) - \gamma(x')| \to 0 \text{ as } n \to \infty.$$

Remark 2.

- The asymptotic variance is lower bounded: $\mathcal{AV}(a, \lambda) \geq 1$ for all $a \geq 1$ and $\lambda \in (0, 1]$.
- The asymptotic bias $\mathcal{AB}(a, \lambda, \rho(t))$ is an increasing function of $\rho(t)$.

Since the second-order parameter $\rho(t)$ is unknown in practice, the comparison of asymptotic bias associated to different log-gamma weights is difficult.

$\Longrightarrow$ Introduction of the *mean-squared bias* defined as:

$$\mathcal{MSB}(a, \lambda) = \int_{-\infty}^{0} \mathcal{AB}^2(a, \lambda, \rho) d\rho = \frac{1}{\lambda(2a - 1)}.$$

4. Choice of log-gamma parameters

Nearest-neighbor Hill estimator: $(a = \lambda = 1)$

$$\hat{\gamma}_n^{\text{H}}(t) = \hat{\gamma}_n(t,1,1) = \frac{1}{k_{n,t}} \sum_{i=1}^{k_{n,t}} i \log \left(\frac{Z_{m_{n,t}-i+1,m_{n,t}}}{Z_{m_{n,t}-i,m_{n,t}}} \right) .$$

Same expression as in Hill (1975). $\mathcal{MSB}(1,1) = 1$ and $\mathcal{AV}(1,1) = 1$ (optimal variance estimator).

Nearest neighbor Zipf estimator: $(a = 2$ and $\lambda = 1)$.

$$\hat{\gamma}_n^{\text{Z}}(t) = \hat{\gamma}_n(t,2,1) = \sum_{i=1}^{k_{n,t}} i \log(k_{n,t}/i) \log \left(\frac{Z_{m_{n,t}-i+1,m_{n,t}}}{Z_{m_{n,t}-i,m_{n,t}}} \right) \Bigg/ \sum_{i=1}^{k_{n,t}} \log(k_{n,t}/i) .$$

Similar to the Zipf estimator proposed by Kratz *et al.* (1996) and Schultze *et al.* (1996). $\mathcal{MSB}(2,1) = 1/3$ and $\mathcal{AV}(2,1) = 2$.

Controlling the asymptotic mean-squared error.

- The asymptotic mean-squared error is defined as

$$\mathcal{AMSE}(a,\lambda) = \Delta^2\left(\frac{m_{n,t}}{k_{n,t}}, t\right)\mathcal{MSB}(a,\lambda) + \frac{\gamma^2(t)\mathcal{AV}(a,\lambda)}{k_{n,t}},$$

  but it cannot be evaluated since the function Δ is unknown.

- Introducing $\pi(a,\lambda) = \mathcal{MSB}(a,\lambda)\mathcal{AV}(a,\lambda)$, we obtain an **upper bound**

$$\mathcal{AMSE}(a,\lambda) \leq \frac{\pi(a,\lambda)}{k_{n,t}}\left\{\xi^2(t) + \gamma^2(t)(2a_{\max} - 1) + o(1)\right\},$$

  for all $\lambda \in (0,1]$ and $a \in [1, a_{\max}]$.

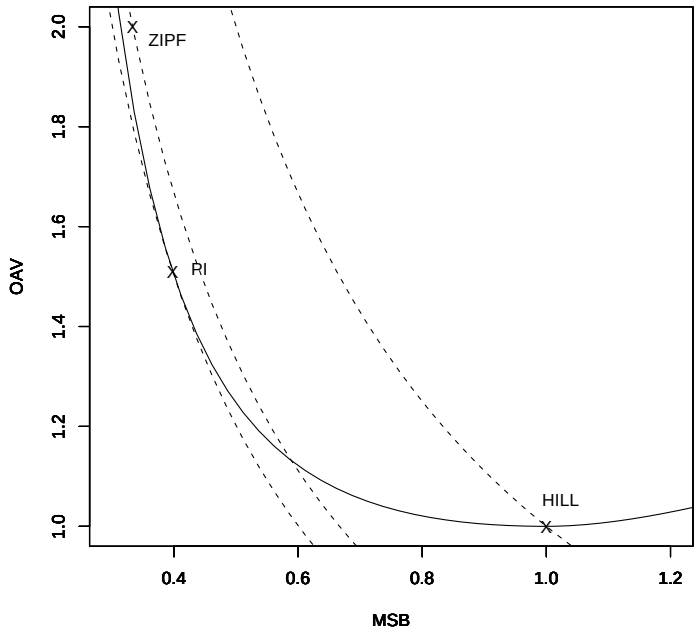
- Considering the log-gamma parameters defined as:

$$(a_\pi, \lambda_\pi) = \arg\min_{a,\lambda} \pi(a,\lambda),$$

  yields a new estimator $\hat{\gamma}_n^\pi(t,a,\lambda) = \hat{\gamma}_n(t, a_\pi, \lambda_\pi)$ with $\mathcal{MSB}(a_\pi,\lambda_\pi) \approx 0.40$ and $\mathcal{AV}(a_\pi,\lambda_\pi) \approx 1.51$.

Optimal asymptotic variance estimators. The optimal asymptotic variance is defined as the smallest variance that can be reached for a fixed bias:

$$\mathcal{OAV}(b) = \min_{a,\lambda} \mathcal{AV}(a,\lambda) \text{ under the constraint } \mathcal{MSB}(a,\lambda) = b.$$

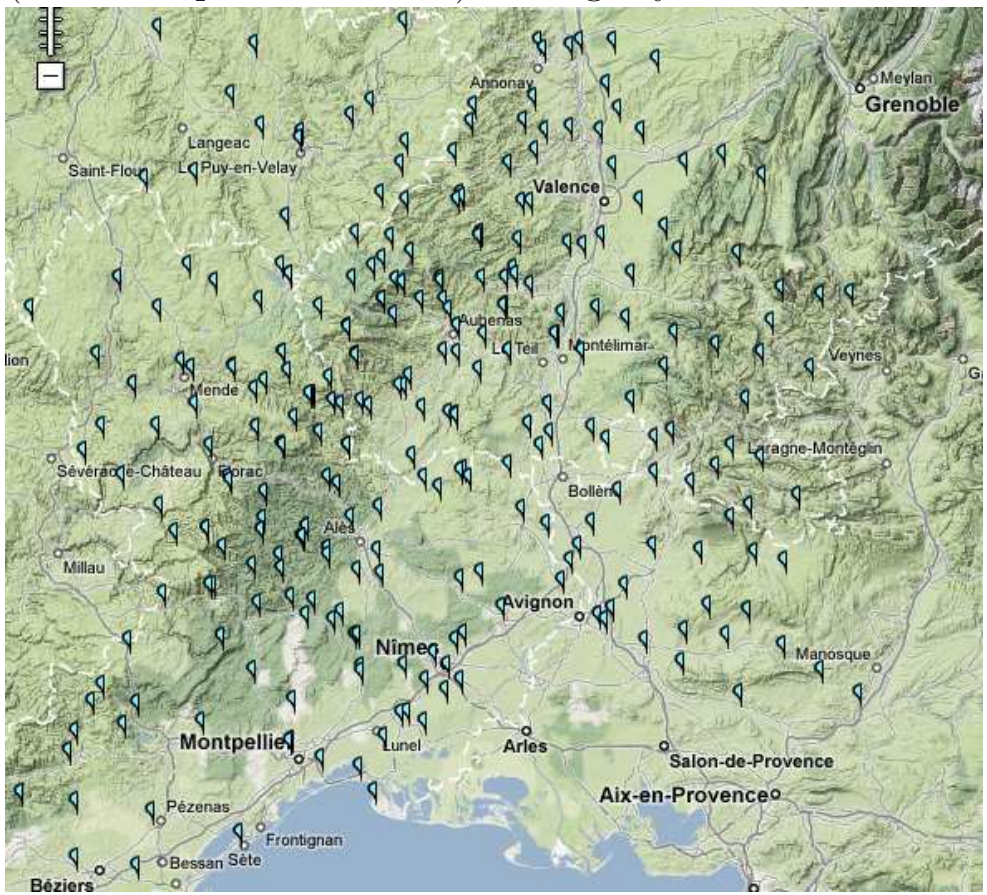


full line: $\mathcal{OAV}$ as a function of $\mathcal{MSB}$,
dashed lines: level curves of $\pi = \mathcal{AV} \times \mathcal{MSB}$,
points: positions of the estimators $\hat{\gamma}_n^{\text{H}}$ (HILL), $\hat{\gamma}_n^{\text{Z}}$ (ZIPF) and $\hat{\gamma}_n^{\pi}$ (PI).

Zipf estimator is not optimal. An estimator with same bias ($\mathcal{MSB} = 1/3$) and smaller variance ($\mathcal{AV} \approx 1.85$) can be found in the log-gamma family.

5. An application to rainfall data

$n = 264,056$ hourly rainfall observations at 142 stations in the Cévennes-Vivarais region (southern part of France) during 7 years.



Y is the hourly rainfall,
$x = (x_1, x_2, x_3)$ is a three-dimensional covariate such that x_1 is the longitude, x_2 is the latitude and x_3 is the altitude.

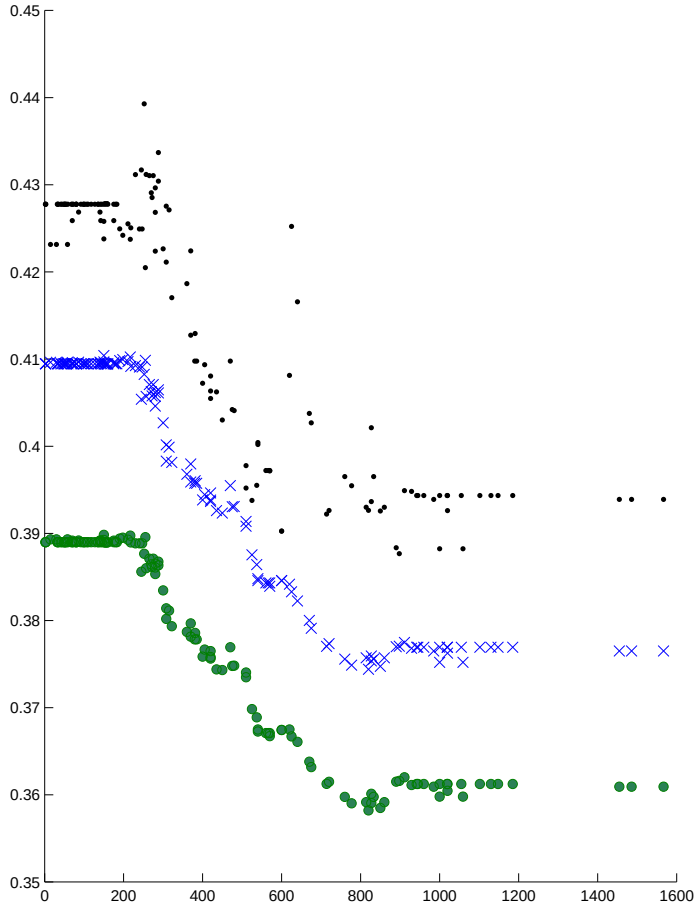
Selection of the hyperparameters.

- $m_{n,t}$ and $k_{n,t}$ are assumed to be independent of t, they are thus denoted by h and k respectively.
- They are selected by minimizing dissimilarity measure between the estimators:

$$(\hat{k}, \hat{m}) = \underset{k_{n,t}, m_{n,t}}{\arg\min} \max_{t \in S} \mathcal{D}(\hat{\gamma}_n^{\text{H}}(t), \hat{\gamma}_n^{\text{Z}}(t), \hat{\gamma}_n^{\pi}(t))$$

  with $\mathcal{D}(u_1, u_2, u_3) \overset{def}{=} \max\{|u_1 - u_2|, |u_2 - u_3|, |u_3 - u_1|\}$ and where $S = \{(x_{1,j}, x_{2,j}, x_{3,j}),\ j = 1, \ldots, 142\}$ is the set of coordinates of the raingauge stations.
- This heuristics is sometimes used in functional estimation and relies on the idea that, for a properly chosen pair $(\hat{k}, \hat{m})$, all three estimates should approximatively give the same tail index.
- This procedure yields $\hat{m}/n = 55\%$ and $\hat{k}/\hat{m} = 5.5\%$.

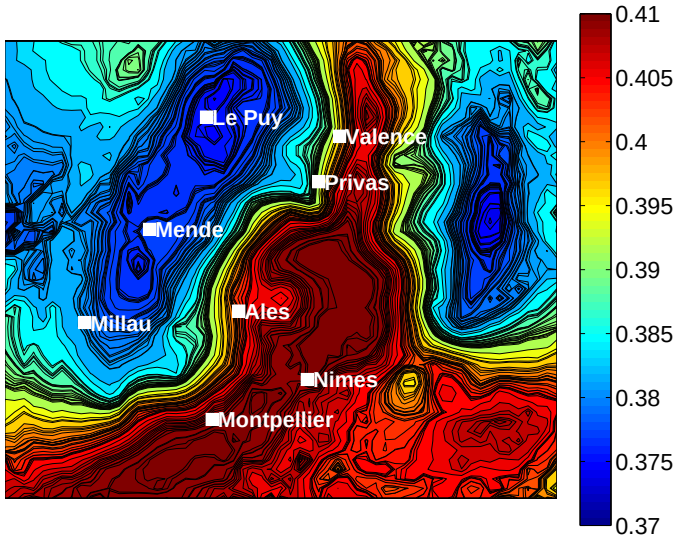
Estimated tail-index as a function of the altitude: $\hat{\gamma}_n^{\mathrm{H}}(\cdots)$, $\hat{\gamma}_n^{\pi}(\times\times\times)$ and $\hat{\gamma}_n^{\mathrm{Z}}(\bullet\bullet\bullet)$.



The shapes of the three curves representing the estimated tail index as a function of the altitude are qualitatively the same.

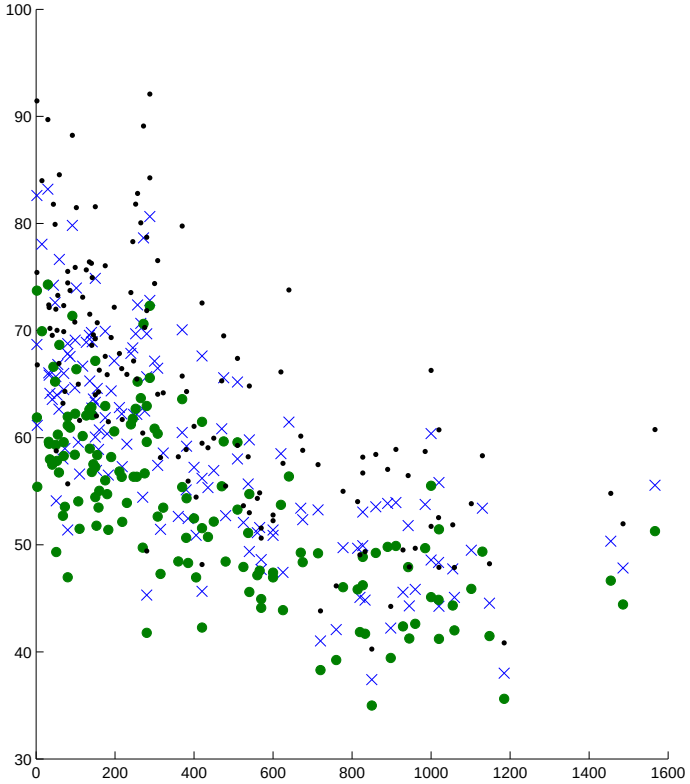
The tail index is a decreasing function of the altitude till $x_3 = 800$ meters and is constant for altitudes ranging from 800 and 1600 meters. This phenomena can be interpreted since extreme hourly rainfalls are more likely to occur in the plains than in the mountains.

Estimated tail-index $\hat{\gamma}_n^{\pi}$ as a function of the longitude and latitude.



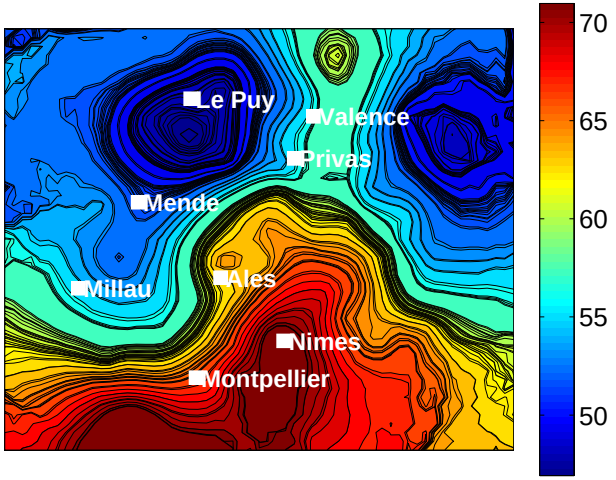
Heaviest tails are obtained in the plains (Rhône valley and Mediterranean coast): Flat areas are the more efficient in capturing the solar energy which is in turn available to involve deep convective clouds.

Estimated 10 years- return level as a function of the altitude:
$\hat{q}_n^{\mathrm{H}}(\cdots)$, $\hat{q}_n^{\pi}(\times\ \times\ \times)$ and $\hat{q}_n^{\mathrm{z}}(\bullet\ \bullet\ \bullet)$.



The considered return level is globally decreasing with the altitude. However, the observed variability indicates that altitude is not the unique factor.

Estimated 10 years- return level $\hat{q}_n^{\pi}$ as a function of the longitude and latitude.



Valence area of Rhône Valley does not suffer from high return levels whereas the southern part does. The enhancement of extreme rainfall rates could be due to the supply of warm and moist air by northward low level winds over the Mediterranean sea.

Postdoctoral position proposal

Spatial analysis of extreme rainfalls in the Cévennes-Vivarais region

- **Location:** Mistis, Rhône-Alpes Research Unit of INRIA, located near Grenoble and Lyon. The Unit includes more than 500 people, within 25 research teams and 10 support services.
- **Length:** 12 months extendable.
- **Monthly Salary after taxes:** around 1900 euros (medical insurance included).
- **Scientific context:** collaboration between Mistis & Laboratoire d'Etude des Transferts en Hydrologie et Environnement, supported by the French Research Agency (ANR) through its VMC2007 program (Vulnérabilité: Milieux, Climats).
- **Contact:** {`Laurent.Gardes, Stephane.Girard`}`@inrialpes.fr`

Bibliography

- L. Gardes, S. Girard and A. Lekina. Functional nonparametric estimation of conditional extreme quantiles, *Journal of Multivariate Analysis*, **101**, 419–433, 2010.
- L. Gardes and S. Girard. Conditional extremes from heavy-tailed distributions: An application to the estimation of extreme rainfall return levels, *Extremes*, **13**, 177–204, 2010.
- A. Daouia, L. Gardes, S. Girard and A. Lekina. Kernel estimators of extreme level curves, *Test*, **20**, 311–333, 2011.
- L. Gardes and S. Girard. Functional kernel estimators of large conditional quantiles, *Electronic Journal of Statistics*, **6**, 1715–1744, 2012.
- A. Daouia, L. Gardes, and S. Girard. On kernel smoothing for extremal quantile regression, *Bernoulli*, **19**, 2557–2589, 2013.
- J. El Methni, L. Gardes and S. Girard. Nonparametric estimation of extreme risks from conditional heavy-tailed distributions, *Scandinavian Journal of Statistics*, to appear, 2014.