Première analyse des pluies extrêmes dans la région Cévennes-Vivarais

Caroline Bernard-Michel, Laurent Gardes et Stéphane Girard

INRIA Rhône-Alpes

Novembre 2008

Problématique

Etude des pluies extrêmes

- Cévennes Vivarais : entre mer et montagnes
- Précipitations intenses donnant lieu à de violentes crues : Nîmes 1988, Vaison la Romaine 1992 ...
- Prévoir et comprendre ces évènements extrêmes
  - Cartographie des temps de retour des hauteurs de pluies extrêmes ou des niveaux de retour
  - Comprendre les phénomènes pluvieux à différentes échelles, temporelles, spatiales ou spatio-temporelles.

Plan

- Les données Cévennes-Vivarais
- Rappels
  - Théorie des valeurs extrêmes
  - Géostatistique
- Application
  - Pas de temps horaire
  - Pas de temps journalier
- Conclusions
- Discussion sur la prise en compte de la composante temporelle

Plan

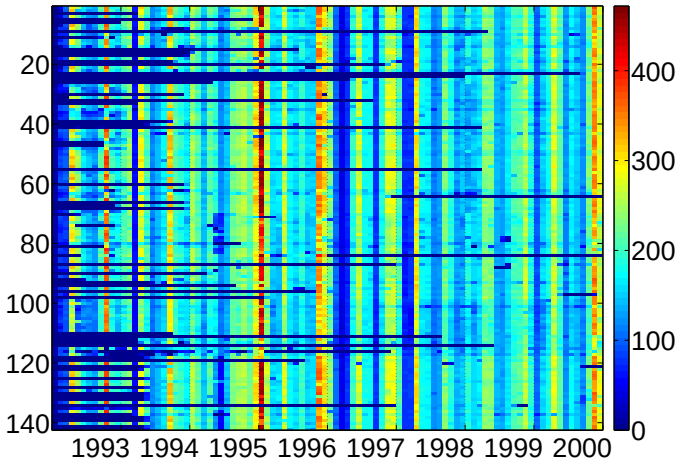
- Les données Cévennes-Vivarais
- Rappels
  - Théorie des valeurs extrêmes
  - Géostatistique
- Application
  - Pas de temps horaire
  - Pas de temps journalier
- Conclusions
- Discussion sur la prise en compte de la composante temporelle

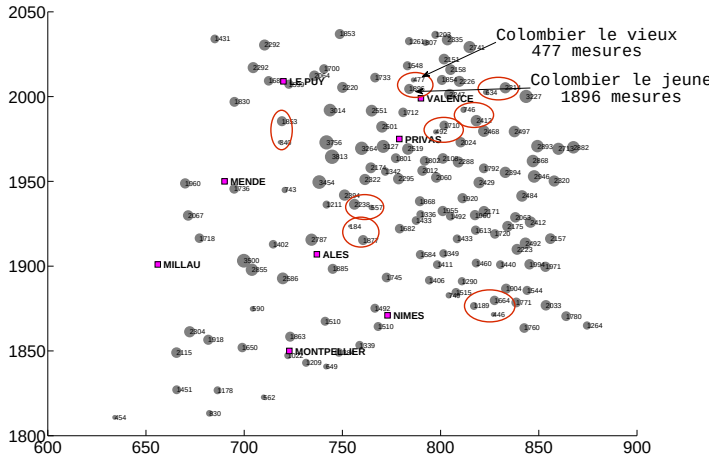
Données denses...

- Données horaires
- Environ une station tous les 10 km

mais hétérogènes...

- De 1972 à 1992 : données évènementielles, 300 stations.
- Entre 1993 à 2000 : mesures en continue, 140 stations
- Après 2000, mesures automnales, 300 stations
- Choix de la base la plus homogène possible : entre 1993 et 2000. Seulement 126 stations avec plus de 1000 mesures horaires.





Statistiques générales

En général

Données Na	Valeurs positives
7.28%	2.65%

Au pas de temps horaire (en mm)

Moyenne	Ecart-Type	$P_{25\%}$	$P_{50\%}$	$P_{75\%}$	$P_{99\%}$	$P_{99.9\%}$
3.04	3.44	1.2	2	3.5	17.2	36.8

Au pas de temps journalier (en mm)

Moyenne	Ecart-Type	$P_{25\%}$	$P_{50\%}$	$P_{75\%}$	$P_{99\%}$	$P_{99.9\%}$
11.67	17.13	2	5.5	14.2	83	161.34

Plan

- Les données Cévennes-Vivarais
- Méthodes : rappels
  - Théorie des valeurs extrêmes
  - Géostatistique
- Application
  - Pas de temps horaire
  - Pas de temps journalier
- Conclusions
- Discussion sur la prise en compte de la composante temporelle

Théorie des valeurs extrêmes

Notations - hypothèses

- X la variable aléatoire modélisant les hauteurs de pluie horaires (mm) en une station de mesures
- $\{X_1, \ldots, X_n\}$ l'échantillon de données horaires dont on dispose
- $X_{1,n} \leq X_{2,n} \leq \ldots X_{n,n}$ l'échantillon ordonné.
- Hypothèse : les données sont indépendantes et identiquement distribuées.

Deux problèmes :

- Calculer la probabilité d'observer une hauteur de pluie extrême $p = \mathbb{P}(X \geq h)$ avec $h \geq X_{n,n}$. Plus souvent, cette probabilité est exprimée en temps de retour $T = 1/p$.
- Calculer la hauteur de pluie h qui est atteinte ou dépassée une seule fois sur T heures avec $T > n$, c'est à dire résoudre $1/T = \mathbb{P}(X \geq h)$. C'est ce qu'on appelle un niveau de retour.

Théorie des valeurs extrêmes

Modéliser la fonction de survie

Modéliser la fonction de survie $\bar{F}(x) = \mathbb{P}(X \geq x) = 1 - F(x)$ où F est la fonction de répartition de X. On ne cherche pas à la modéliser dans son ensemble, mais uniquement en queue de distribution, c'est à dire quand $X \geq X_{n,n}$.

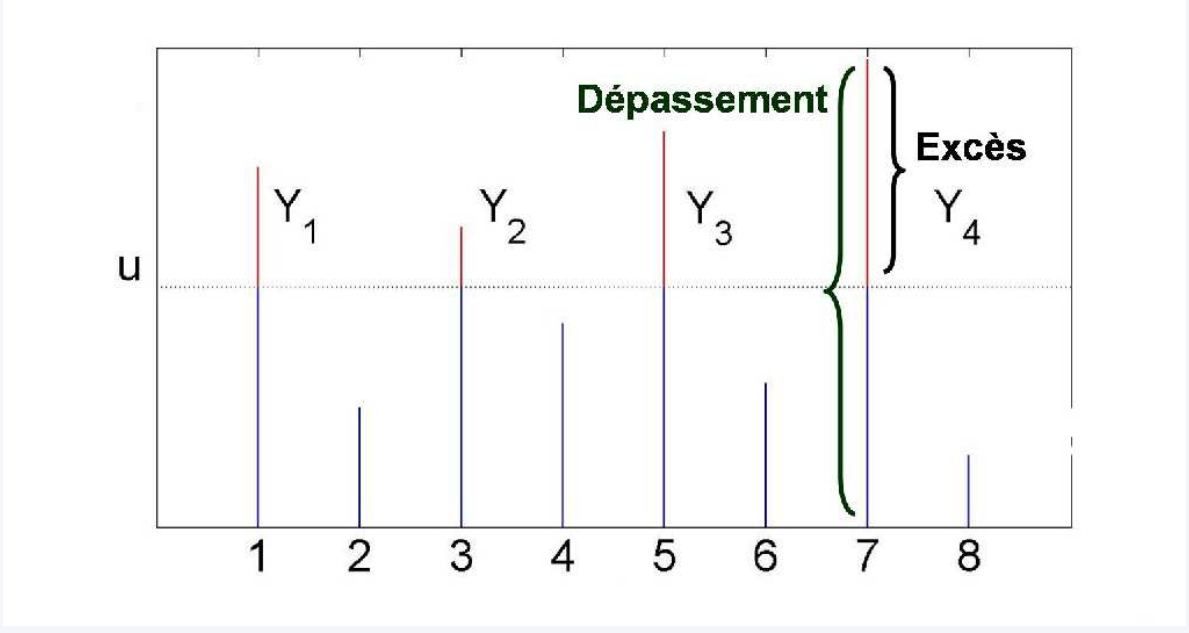
Deux approches

- Découpage des données en blocs, dont les maxima sont supposés distribués selon une loi de la famille GEV (Generalized Extreme Value). Approche typiquement utilisée par les hydrologues, qui supposent de plus que F appartient au domaine de Gumbel.
- Approche envisagée ici : modéliser la distribution des valeurs dépassant un seuil donné. On l'appelle la méthode POT (Peaks-over-threshold).

Théorie des valeurs extrêmes

Etude des excès

Dans cette approche, on se fixe un seuil u. On définit alors un excès Y de la variable X au dessus du seuil u par $Y = X - u$ quand $X > u$. On appelle dépassements les valeurs de X au dessus du seuil u.



Théorie des valeurs extrêmes

Etude des excès

La fonction de survie d'un excès au dessus de u est donnée pour $y > 0$ par :

$$\begin{aligned}\bar{F}_u(y) &= \mathbb{P}(Y > y) = \mathbb{P}(X - u > y | X > u) \\ &= \frac{\mathbb{P}(X > u + y, X > u)}{\mathbb{P}(X > u)} = \frac{\bar{F}(u + y)}{\bar{F}(u)}\end{aligned}$$

Lorsque le seuil est grand, on peut approcher cette quantité par la fonction de survie d'une loi de Pareto Généralisée (GPD) donnée par :

$$\bar{G}_{\gamma,\sigma} = (1 + \gamma\frac{y}{\sigma})^{-\frac{1}{\gamma}} \text{ si } \gamma \neq 0 \quad (1)$$

$$= exp(-\frac{y}{\sigma}) \text{ sinon} \quad (2)$$

Son ensemble de définition est $\mathbb{R}^+$ si $\gamma \geq 0$ ou $[0, -\frac{\sigma}{\gamma}[$ si $\gamma < 0$.

Domaine d'attraction

La loi GPD dépend de deux paramètres :

- $\gamma \in \mathbb{R}$ est le paramètre de forme,
- $\sigma > 0$ est le paramètre d'échelle (gradex).

On distingue trois types de lois selon la valeur du paramètre de forme γ :

- Si $\gamma > 0$, on dit que F appartient au domaine d'attraction de Fréchet (queue lourde),
- si $\gamma = 0$, on dit que F appartient au domaine d'attraction de Gumbel (queue légère),
- si $\gamma < 0$, on dit que F appartient au domaine d'attraction de Weibull (queue finie).

Temps de retour

Connaissant la loi des excès, le temps de retour associé à une hauteur de pluie x se déduit facilement. Comme on a :

$$\mathbb{P}(X > x | X > u) = [1 + \gamma \frac{x - u}{\sigma}]^{-\frac{1}{\gamma}}$$

Alors

$$\mathbb{P}(X > x) = \xi_u [1 + \gamma \frac{x - u}{\sigma}]^{-\frac{1}{\gamma}} \text{ si } \gamma \neq 0$$
$$= \xi_u exp(-\frac{x - u}{\sigma}) \text{ sinon}$$

avec $\xi_u = \mathbb{P}(X > u)$

Niveau de retour

De même, le niveau de retour x_T qui est dépassé en moyenne toutes les T heures est solution de :

$$\xi_u[1+\gamma(\frac{x_T-u}{\sigma})]^{-\frac{1}{\gamma}} = \frac{1}{T}$$

On en déduit :

$$\begin{aligned} x_T &= u + \frac{\sigma}{\gamma}[(T\xi_u)^\gamma - 1] \text{ si } \gamma \neq 0 \\ &= u + \sigma \log(T\xi_u) \text{ sinon} \end{aligned}$$

Maximum de vraisemblance

- $y_1, \ldots, y_k$ les k excès observés au dessus du seuil u.
- Pour $\gamma \neq 0$, la log-vraisemblance s'écrit :

$$l(\gamma, \sigma) = -k \log \sigma - (1 + \frac{1}{\gamma}) \sum_{i=1}^{k} \log(1 + \gamma \frac{y_i}{\sigma})$$

si $(1 + \sigma^{-1} \gamma y_i) > 0$ pour $i = 1, \ldots, k$

$= -\infty$ sinon

- Pour $\gamma = 0$, elle s'écrit :

$$l(\gamma, \sigma) = -k \log \sigma - \sigma^{-1} \sum_{i=1}^{k} y_i$$

- Maximisation de la log-vraisemblance par des méthodes numériques.

Estimateur de Hill

Lorsqu'on se restreint au domaine de Fréchet, on a la caractérisation :

$$\bar{F}(x) = x^{-\frac{1}{\gamma}} \ell(x) \tag{3}$$

avec $\gamma > 0$ et l une fonction à variations lentes, c'est à dire que pour $t > 1$

$$\lim_{u \to \infty} \frac{\ell(tu)}{\ell(u)} = 1 \tag{4}$$

Ce qui conduit en posant $\xi_u = \mathbb{P}(X > u)$ et pour $x > u$ et $p \leqslant \xi_u$

$$\bar{F}(x) \simeq \xi_u (\frac{x}{u})^{-\frac{1}{\gamma}} \tag{5}$$

$$\bar{F}^{-1}(p) \simeq u (\frac{p}{\xi_u})^{-\gamma} \tag{6}$$

On en déduit

$$\log \bar{F}^{-1}(p) - \log(u) \simeq \gamma \log(\frac{\xi_u}{p}) \tag{7}$$

Estimateur de Hill

En posant $\xi_u = k/n$ et en choisissant plusieurs valeurs de $p = i/n, i = 1, \ldots, k$, on obtient

$$\log \bar{F}^{-1}(\frac{i}{n}) - \log(u) \simeq \gamma \log(\frac{k}{i}) \tag{8}$$

ou encore en estimant les fonctions de survie par leur équivalent empirique :

$$\log X_{n-i,n} - \log(u) \simeq \gamma \log(\frac{k}{i}) \tag{9}$$

- Diagramme de Hill : tracer $\log X_{n-i,n} - \log(u)$ en fonction de $\log(k/i)$
- Estimateur de Hill= pente du diagramme de Hill

$$\hat{\gamma}(k) = \frac{1}{k+1} \sum_{i=1}^{k} (\log X_{n-i,n} - \log(u)) \tag{10}$$

Plan

- Les données Cévennes-Vivarais
- Méthodes : rappels
  - Théorie des valeurs extrêmes
  - Géostatistique
- Application
  - Pas de temps horaire
  - Pas de temps journalier
- Conclusions
- Discussion sur la prise en compte de la composante temporelle

Le variogramme

- La variable régionalisée $z(x)$ en tout point du champs étudié est la réalisation d'une fonction aléatoire $Z(x)$
- Etude restreinte aux deux premiers moments
- Inférence difficile $\Longrightarrow$ hypothèses de stationnarité (ou intrinsèque)
- On suppose donc que $Z(x)$ est une fonction aléatoire intrinsèque sans dérive :

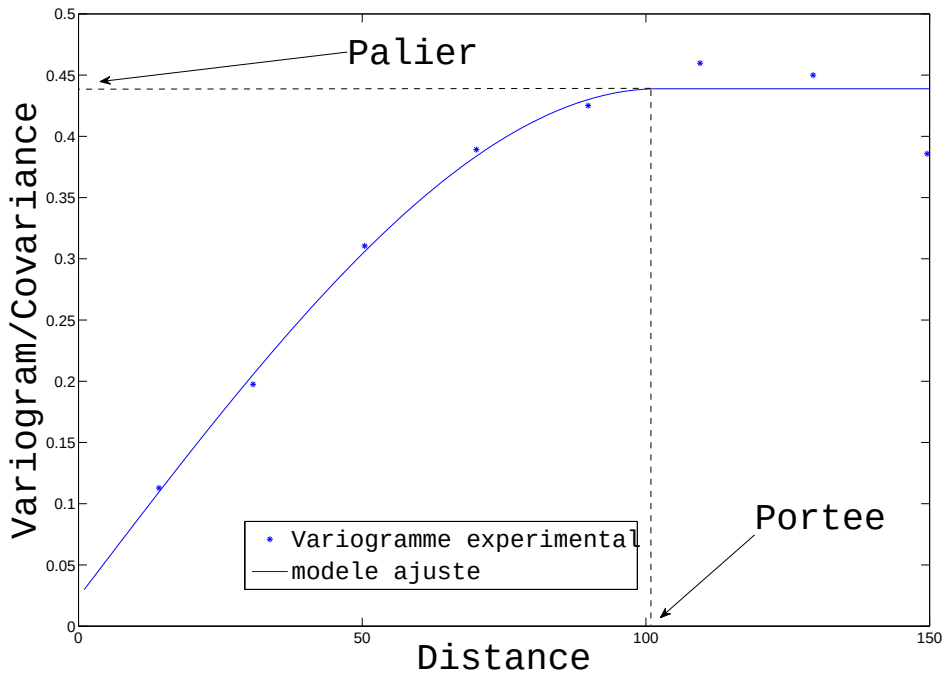
$$\forall x, x+h \in D, \begin{cases} E[Z(x+h) - Z(x)] = 0 \\ var[Z(x+h) - Z(x)] = 2\gamma(h) \end{cases} \quad (11)$$

- $\gamma(h)$ est la variogramme

Variogramme expérimental

- $z(x_1), \ldots, z(x_n), x_\alpha \in \mathcal{D}$, données expérimentales
- $$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{N(h)} [z(x_\alpha) - z(x_\beta)]^2 \qquad (12)$$

où $N(h) = \{(\alpha, \beta) \text{ tel que } x_\alpha - x_\beta = h\}$ et $|N(h)|$ est le nombre de paires distinctes de l'ensemble $N(h)$



Krigeage : 4 étapes

- Contrainte de linéarité : on estime la valeur en un point x_0 par une combinaison linéaire des valeurs aux sites voisins $x_1, \ldots, x_n$.

$$\hat{Z}(x_0) = \sum_{\alpha=1}^{n} \lambda_\alpha Z(x_\alpha) \tag{13}$$

- Contrainte d'autorisation : toute combinaison linéaire des $Z(x_i), i \in \{1, \ldots, n\}$ possède une espérance et une variance finies. Dans le cadre stationnaire, cette condition est toujours vérifiée. Dans le cadre intrinsèque, on doit imposer que la somme des poids soit 1.

$$\sum_{\alpha=1}^{n} \lambda_\alpha = 1 \tag{14}$$

- Contrainte de non biais :

$$E(\hat{Z}(x_0) - Z(x_0)) = 0 \tag{15}$$

  elle conduit à imposer que la somme des poids soit 1.
- Contrainte d'optimalité : on cherche les poids qui minimisent la variance d'estimation $Var(\hat{Z}(x_0) - Z(x_0))$
- Problème de minimisation sous contrainte : système de krigeage.

Plan

- Les données Cévennes-Vivarais
- Rappels
  - Théorie des valeurs extrêmes
  - Géostatistique
- Application
  - Pas de temps horaire
  - Pas de temps journalier
- Conclusions
- Discussion sur la prise en compte de la composante temporelle

Choix du seuil

- Seuil ou pourcentage d'excès
- Compromis entre nombre de mesures suffisant et validité du modèle

Tests d'adéquation

- Tests du Chi2, Anderson-Darling...
- Tracer la pvaleur en fonction du pourcentage d'excès retenus

Stabilité des estimations

- Tracer en fonction du pourcentage d'excès les estimations des paramètres de forme et d'échelle

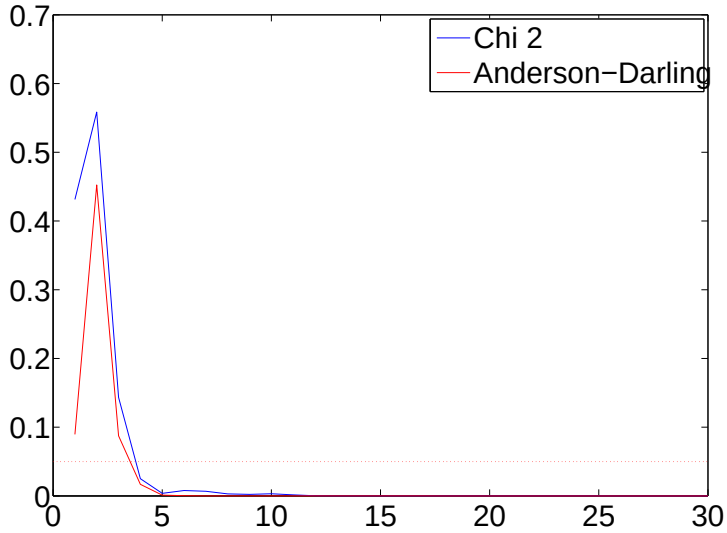
Cohérence des estimations

- Cohérence des estimations par maximum de vraisemblance et par Hill en fonction du pourcentage d'excès.

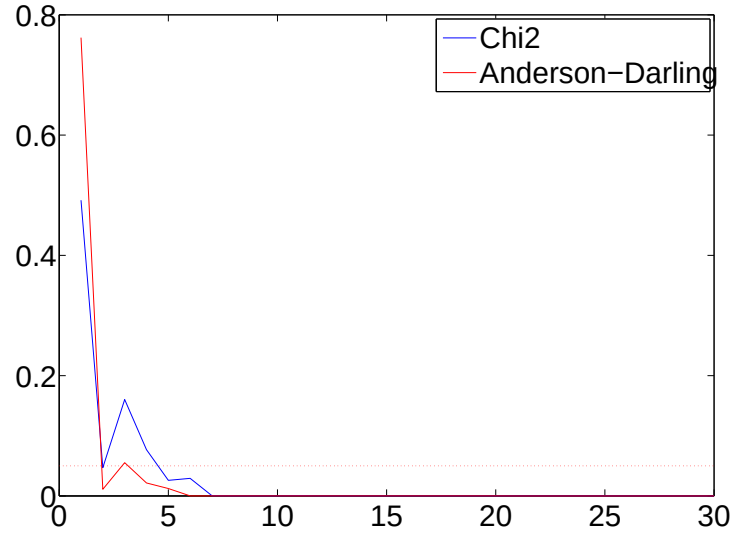
Choix du seuil

Test du Chi 2

- Exemple sur les 2 stations les mieux informées
- Test d'adéquation des $Z_i = \log(Y_i/u)$ à la loi exponentielle de paramètre $\hat{\gamma}$ (ici par Hill)
- Pvaleur en fonction du seuil
- En moyenne sur les 126 stations, seuil fixé à 7% d'excès



Branas



Marzan-L'Abbaye

Choix du seuil

- Seuil ou pourcentage d'excès
- Compromis entre nombre de mesures suffisant et validité du modèle

Tests d'adéquation

- Tests du Chi2, Anderson-Darling...
- Tracer la pvaleur en fonction du pourcentage d'excès retenus

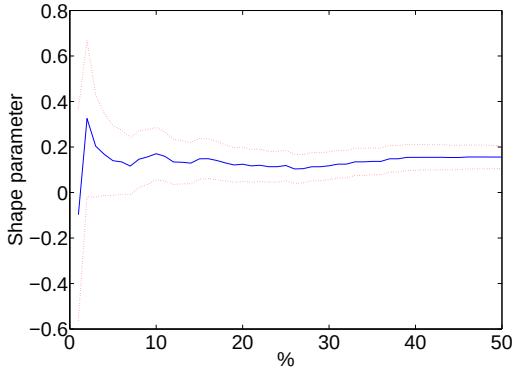
Stabilité des estimations

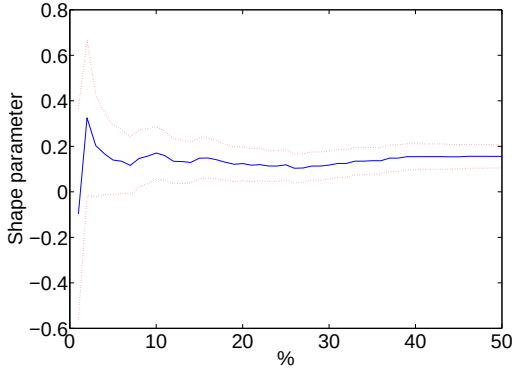
- Tracer en fonction du pourcentage d'excès les estimations des paramètres de forme et d'échelle

Cohérence des estimations

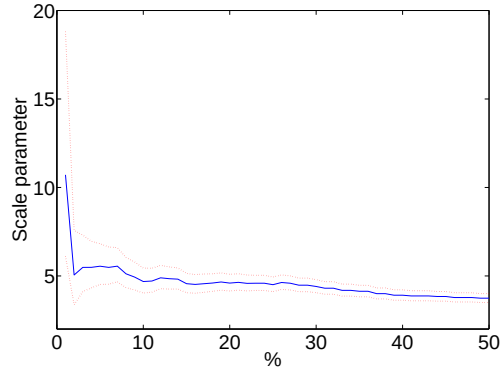
- Cohérence des estimations par maximum de vraisemblance et par Hill en fonction du pourcentage d'excès.

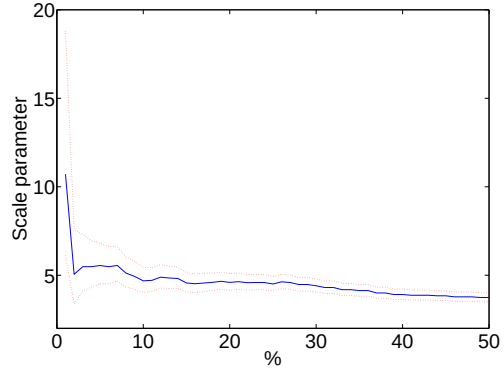
Choix du seuil : Evolution des paramètres en fonction du seuil



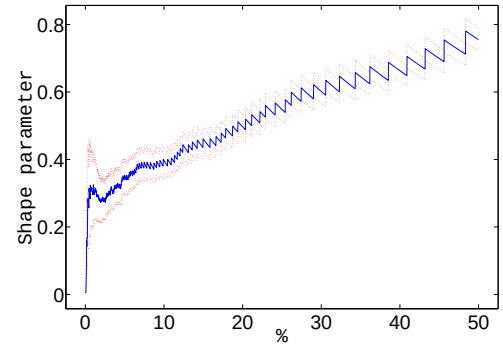


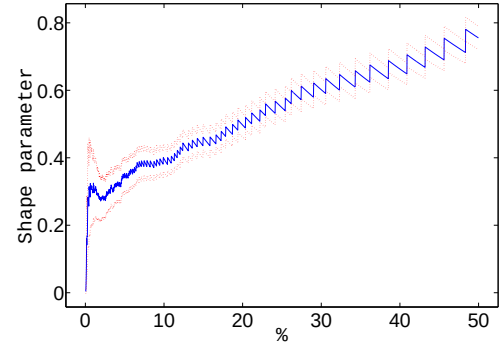
γ (ML)





σ (ML)





γ (Hill)

Choix du seuil

- Seuil ou pourcentage d'excès
- Compromis entre nombre de mesures suffisant et validité du modèle

Tests d'adéquation

- Tests du Chi2, Anderson-Darling...
- Tracer la pvaleur en fonction du pourcentage d'excès retenus

Stabilité des estimations

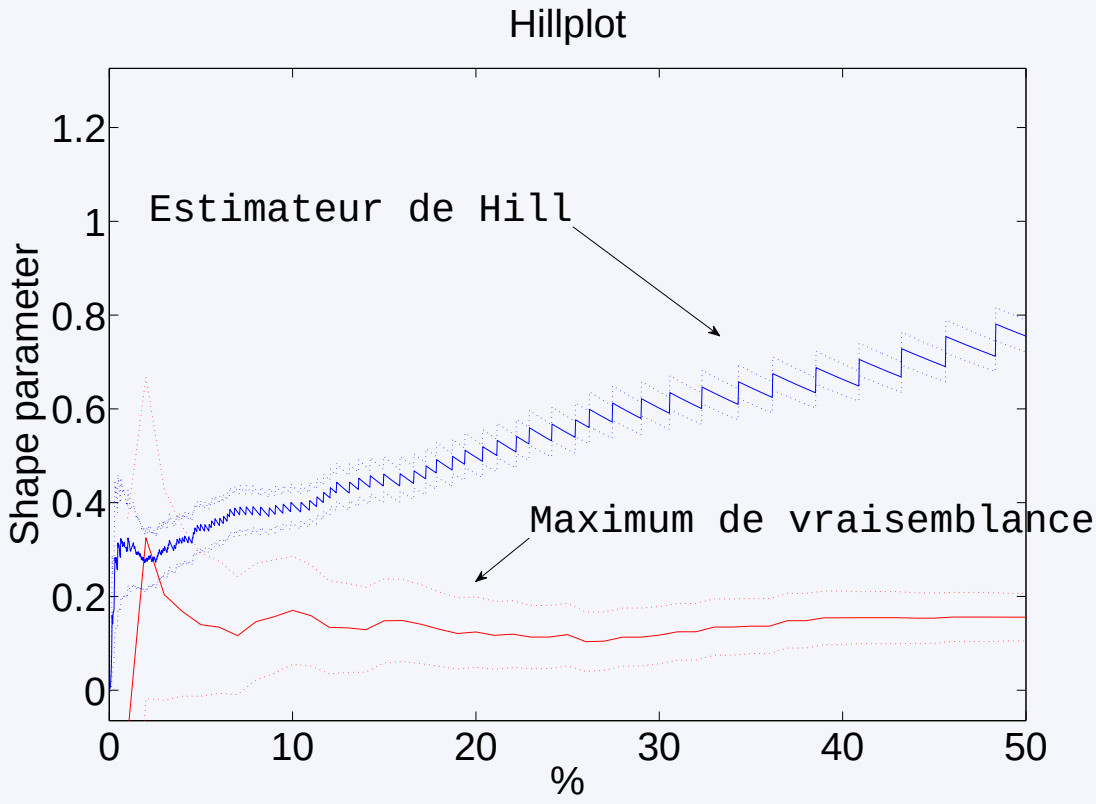
- Tracer en fonction du pourcentage d'excès les estimations des paramètres de forme et d'échelle

Cohérence des estimations

- Cohérence des estimations par maximum de vraisemblance et par Hill en fonction du pourcentage d'excès.

Choix du seuil

Comparaison des estimations par maximum de vraisemblance et par Hill



Domaine d'attraction : Fréchet ?

Diagramme de Hill

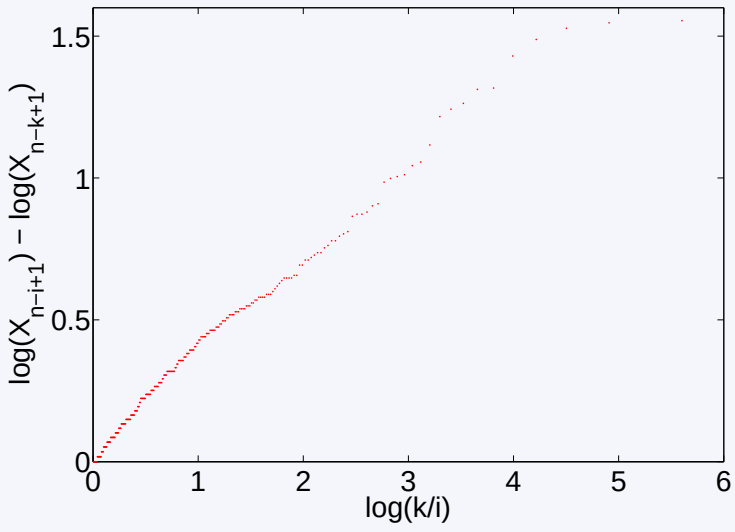
Tracer $\log X_{n-i,n} - \log(u)$ en fonction de $\log(k/i)$.

Estimation du paramètre de forme

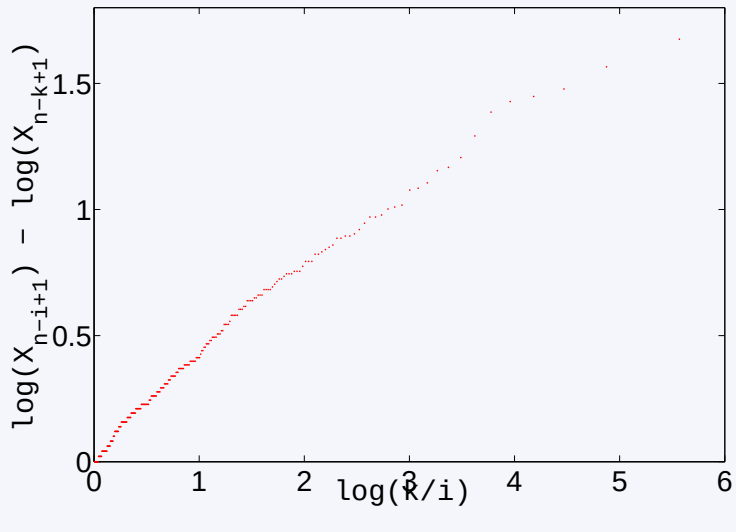
Estimation du paramètre de forme par maximum de vraisemblance et par Hill

Domaine d'attraction

Diagramme de Hill



Barnas



Marzan-L'Abbaye

Choix du domaine d'attraction

Diagramme de Hill

Tracer $\log X_{n-i,n} - \log(u)$ en fonction de $\log(k/i)$.

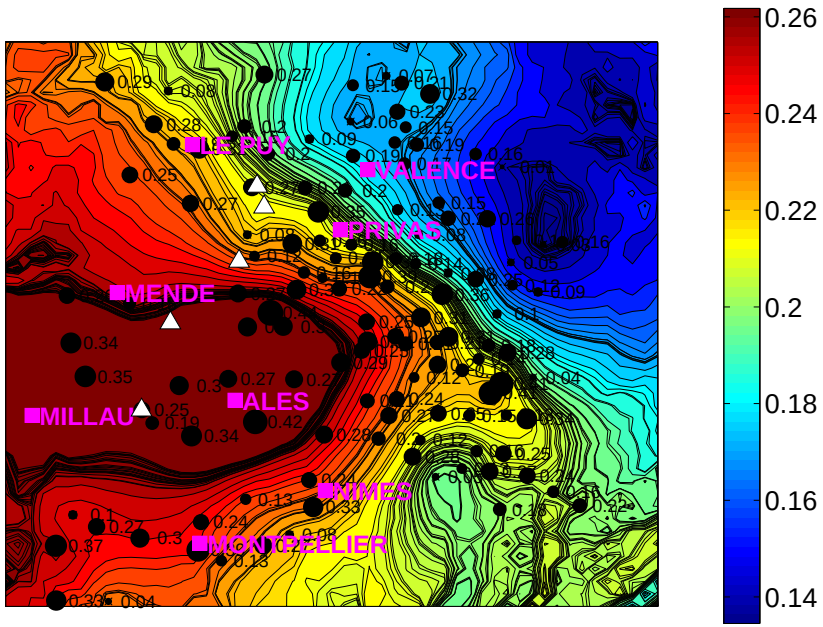
Estimation du paramètre de forme

Estimation du paramètre de forme par maximum de vraisemblance et par Hill

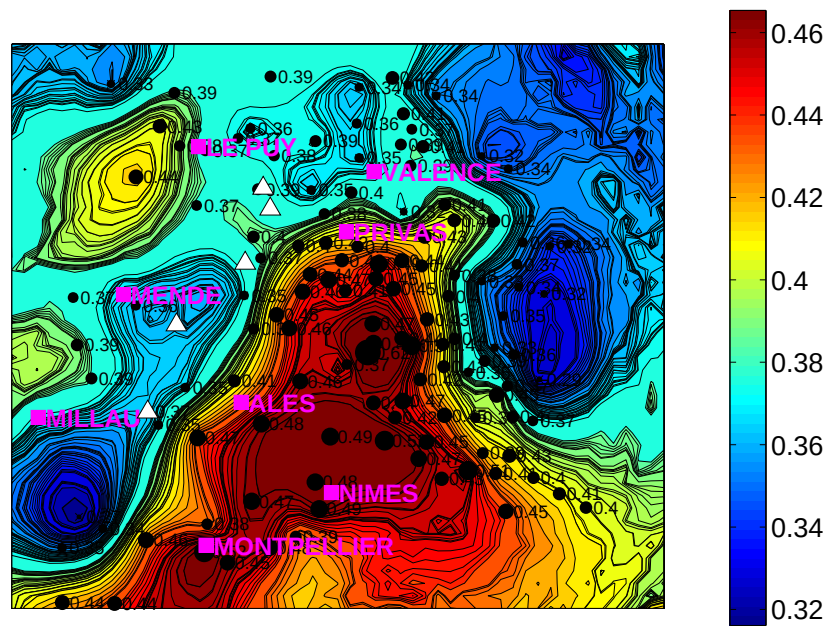
Domaine d'attraction

Estimation du paramètre de forme

- Paramètre de forme positif pour les deux méthodes
- Cartes très différentes pour les deux méthodes
- Hill : paramètre de forme lié au relief.
- Approche par Hill : plus réaliste.



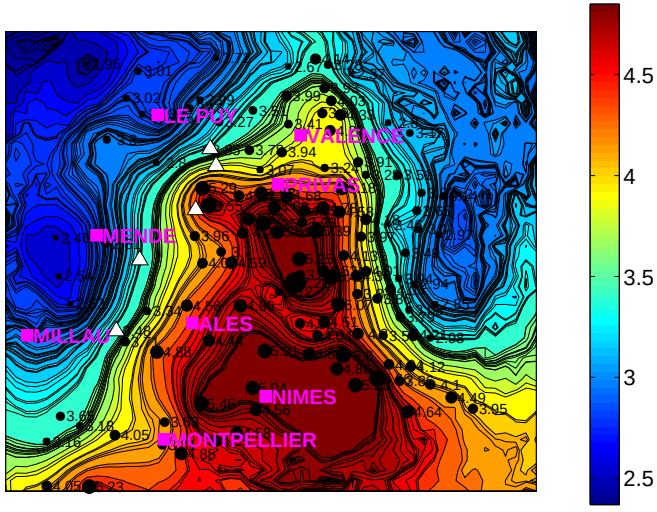
γ (ML)



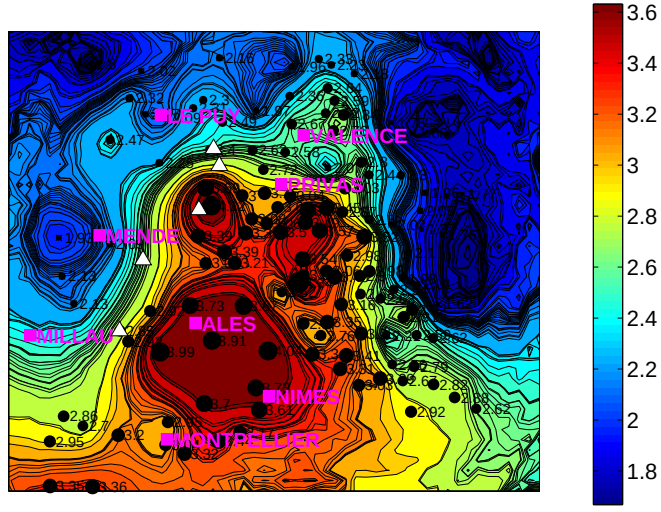
γ (Hill)

Estimation du paramètre d'échelle

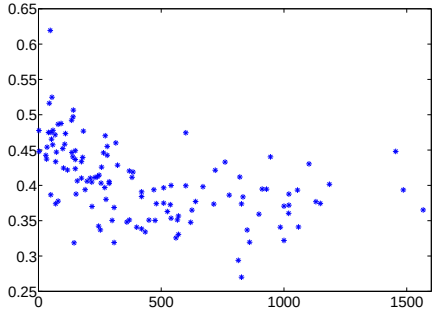
- Cartes similaires par maximum de vraisemblance et Hill
- Paramètre d'echelle lié au relief



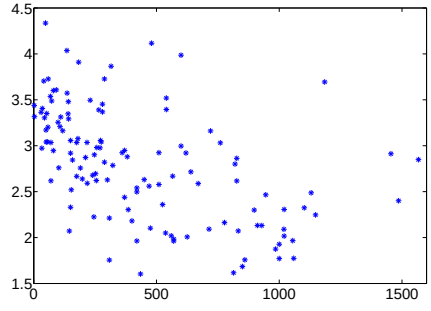
ML



Hill

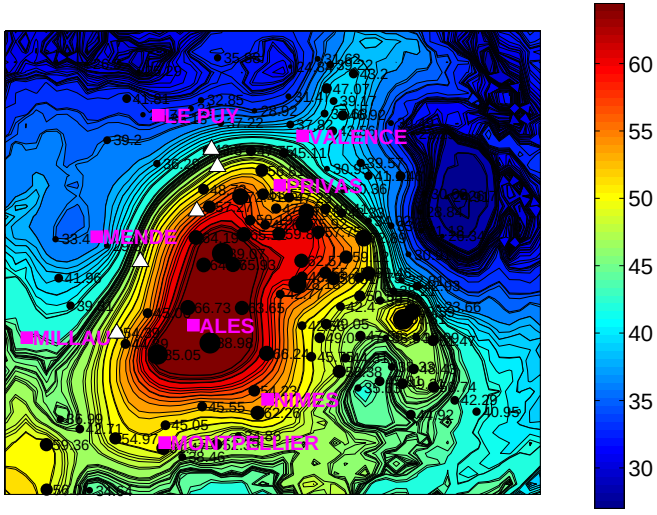


γ versus altitude

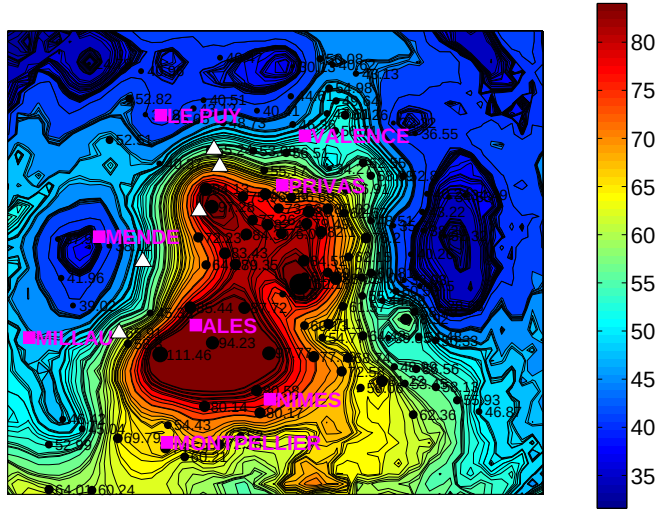


σ versus altitude

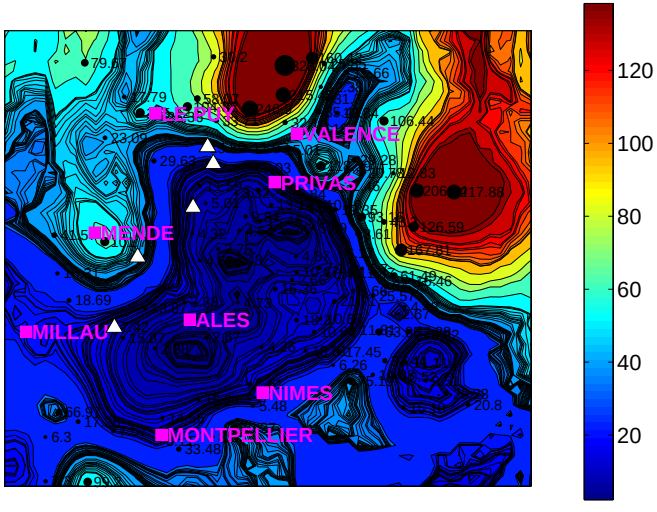
Niveaux de retour et périodes de retour



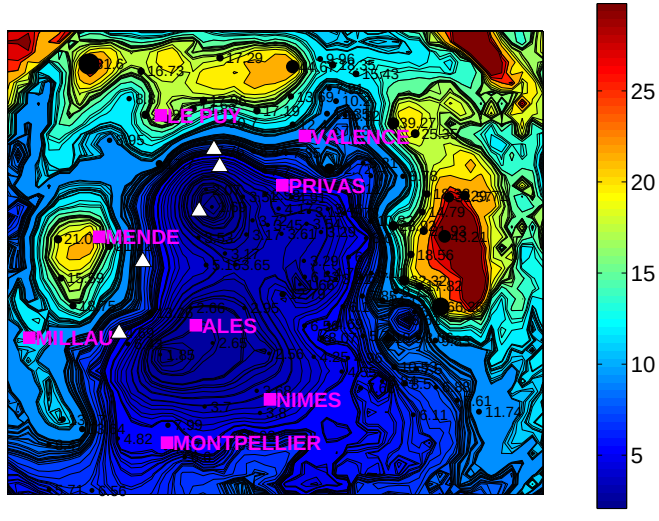
Niveau de retour (ML, 10 ans)



Niveau de retour (Hill, 10 ans)



Temps de retour (ML, 50mm)

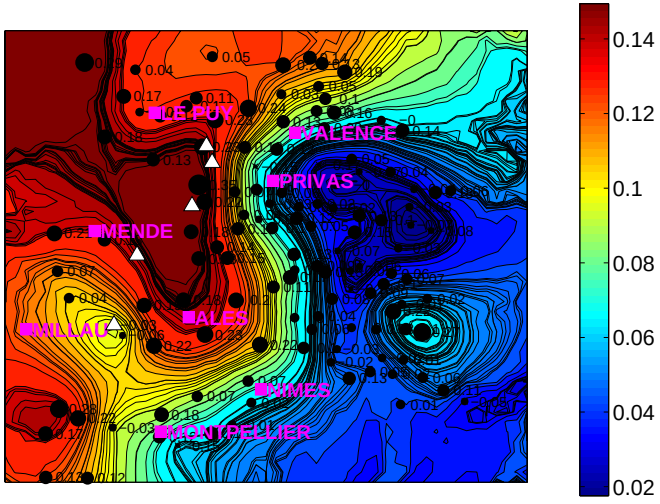


Temps de retour (Hill, 50mm)

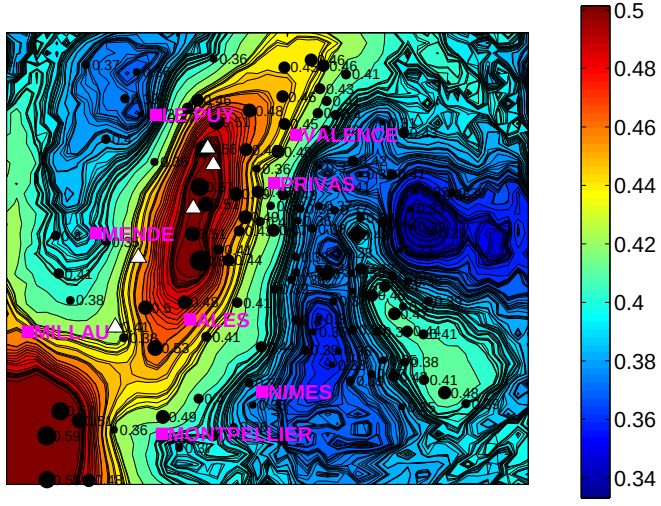
Plan

- Les données Cévennes-Vivarais
- Rappels
  - Théorie des valeurs extrêmes
  - Géostatistique
- Application
  - Pas de temps horaire
  - Pas de temps journalier
- Conclusions
- Discussion sur la prise en compte de la composante temporelle

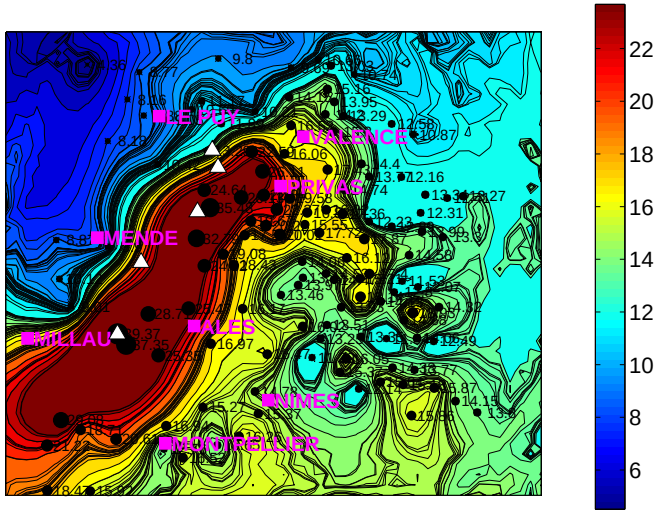
Paramètres de forme et d'échelle



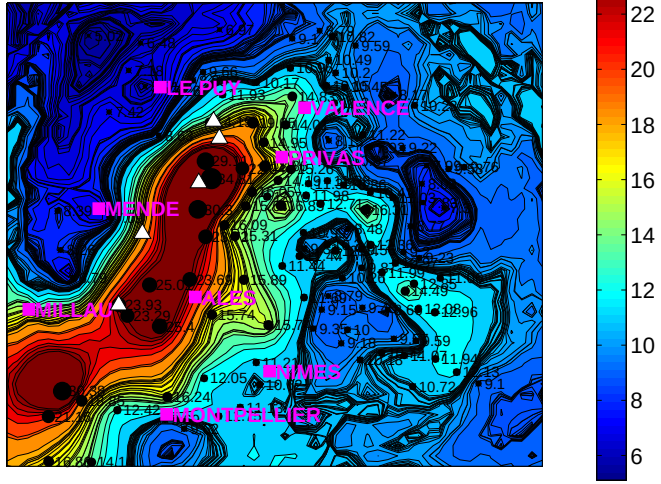
Paramètre de forme (ML)



Paramètre de forme (Hill)

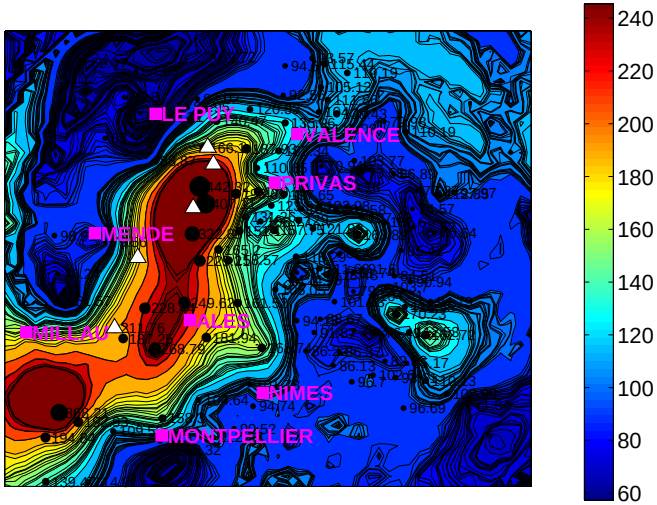


Paramètre d'échelle (ML)

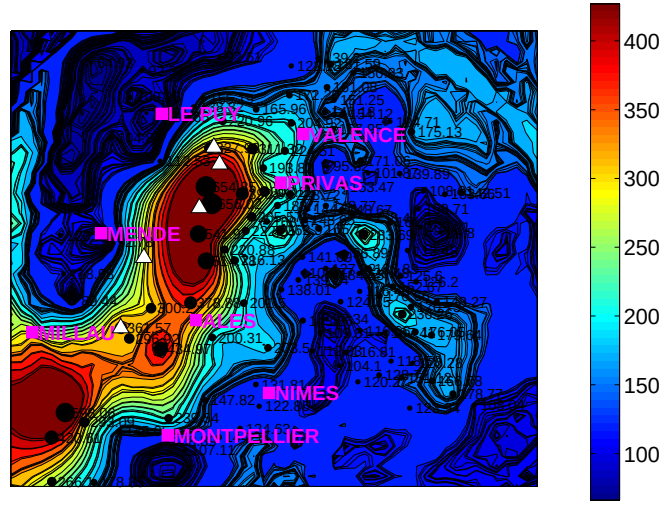


Paramètre d'échelle (Hill)

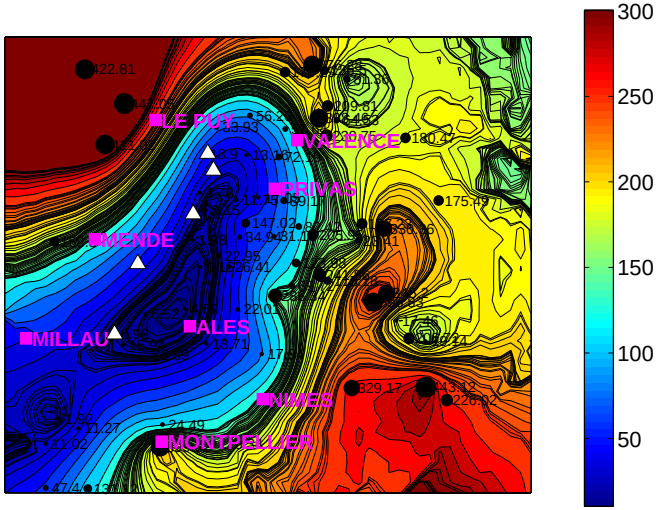
Niveaux et temps de retour



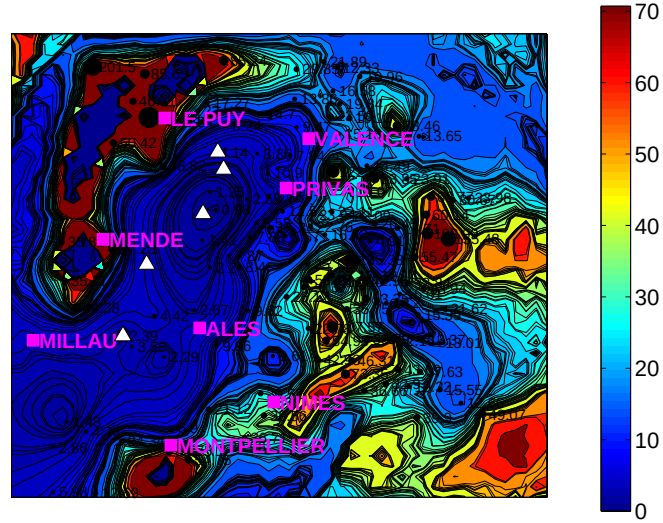
Niveau de retour (ML, 10 ans)



Niveau de retour (Hill, 10 ans)



Temps de retour (ML, 200mm)



Temps de retour (Hill, 200mm)

Plan

- Les données Cévennes-Vivarais
- Rappels
  - Théorie des valeurs extrêmes
  - Géostatistique
- Application
  - Pas de temps horaire
  - Pas de temps journalier
- Conclusions
- Discussion sur la prise en compte de la composante temporelle

Conclusions

Synthèse

- Domaine d'attraction = Fréchet et non Gumbel !
- Incertitude plus forte par Maximum de vraisemblance, résultats moins réalistes
- Choix du seuil et du modèle très important !
- Changement d'échelle= conclusions très différentes.
  - Pas de temps horaire : pluies intenses en plaine !
  - Pas de temps journalier : pluies intenses en altitude, mais seulement dans les Cévennes !

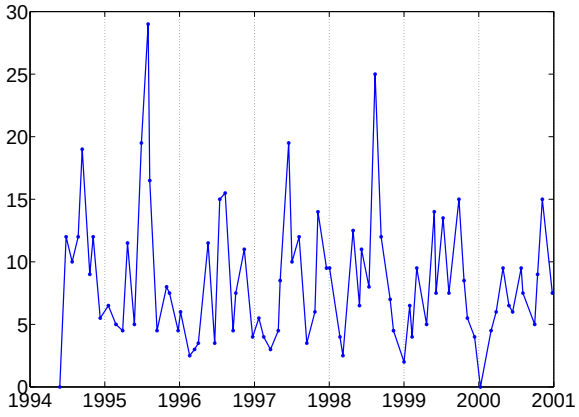
A étudier

- Etude des changements d'échelle nécessite modèle temporel
- Hypothèse d'indépendance et de même loi pour les mesures FAUSSE
- Modèle spatial ? Spatio-temporel ?

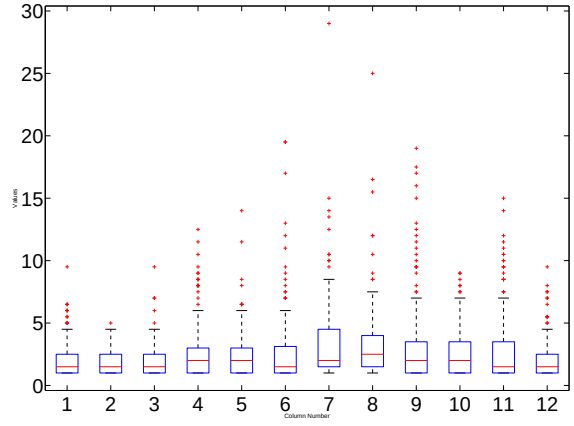
Plan

- Les données Cévennes-Vivarais
- Rappels
  - Théorie des valeurs extrêmes
  - Géostatistique
- Application
  - Pas de temps horaire
  - Pas de temps journalier
- Conclusions
- Discussion sur la prise en compte de la composante temporelle

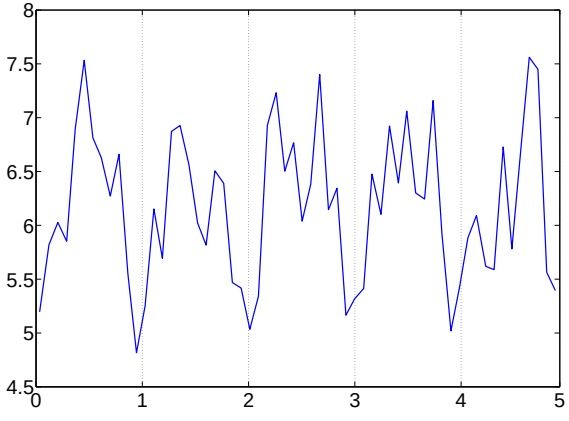
Tendance, saisonnalité, corrélation ?



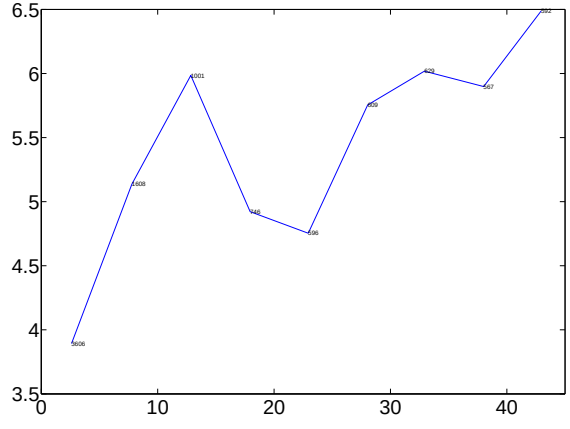
Maxima mensuels



Boxplot



variogramme
pas d'un mois



variogramme
pas de 5 heures

Prise en compte de la saisonnalité

Prise en compte dans les paramètres de la loi GPD

$$Y(t) \backsim GPD(\gamma(t), \sigma(t)) \tag{16}$$

avec par exemple $\gamma(t) = exp(\beta_0 + \beta_1 t)$ et $\sigma(t) = exp(\beta_3 + \beta_4 t)$

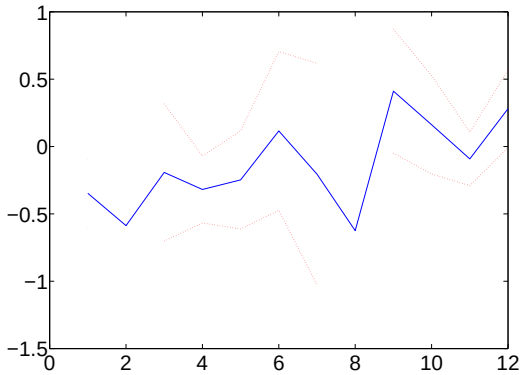
Découpage en saisons

- On découpe l'année en S saisons $S_1, \ldots, S_S$
- Stationnarité sur chacune des saisons
- Modèle de mélange

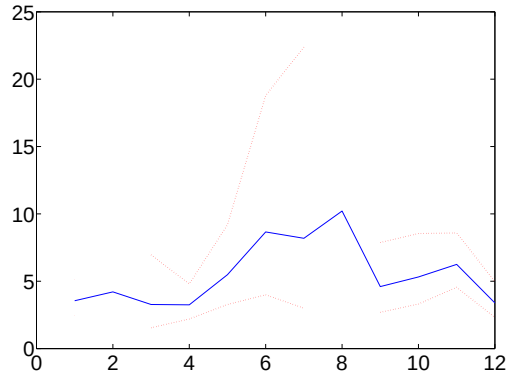
$$\mathbb{P}(X > s) = \sum_{k=1}^{S} \mathbb{P}(X > s | X \in S_k) P(X \in S_k) \tag{17}$$

Prise en compte dans les paramètres

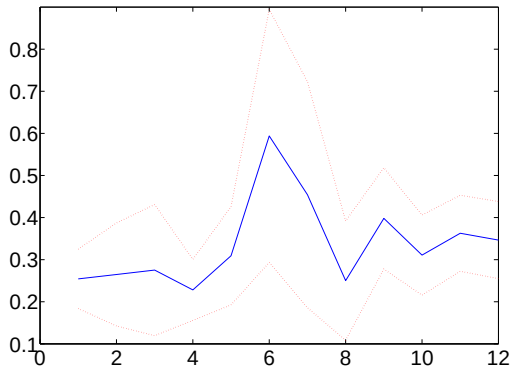
- Le nombre de paramètres à estimer croît
- Quelle modèle pour les paramètres ?
- Choix du seuil ?



γ par mois (ML)



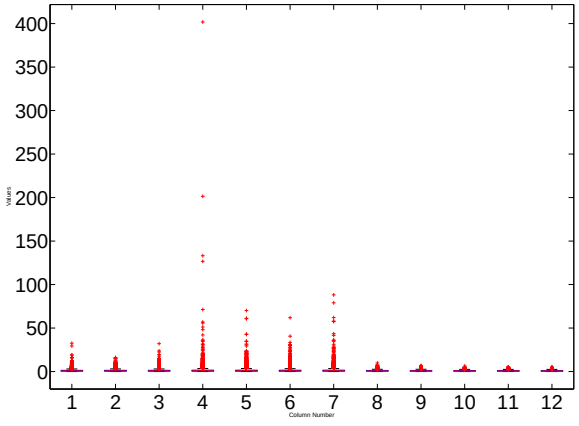
σ par mois (ML)



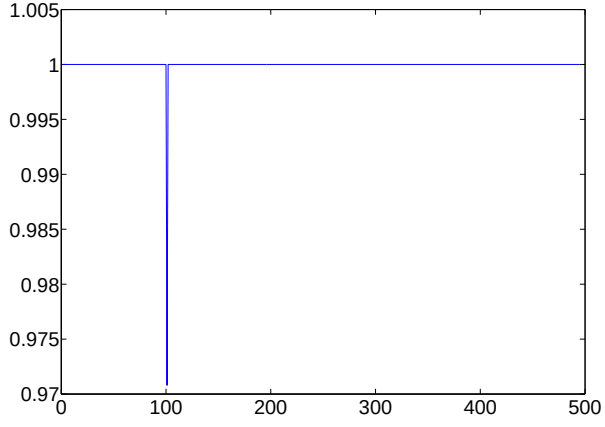
γ par mois (Hill)

Découpage en saisons

- Regroupement des mesures mois par mois
- Calcul des excès par mois
- Choix du nombre de saisons S
- Calcul des $N = C_{12}^{S}$ configurations de saisons possibles
- Pour les N configurations, test non paramétrique de Kruskal Wallis entre les mois de chaque saison
- Minimisation du coût : 1 - min(pvaleurs)



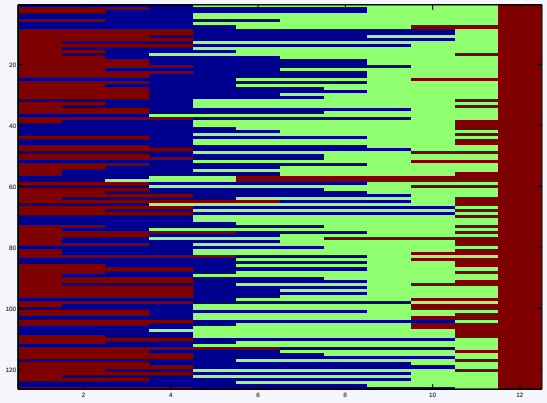
Simulation (Student)



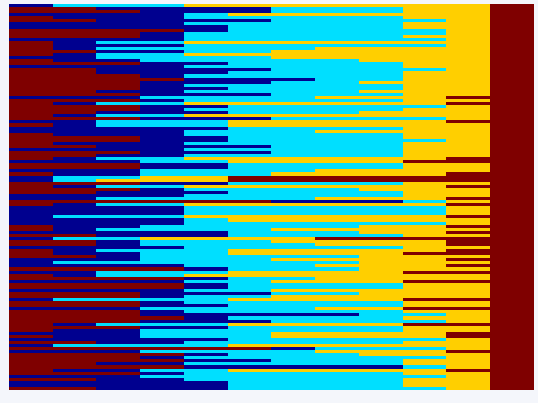
Coût

Sur l'ensemble des stations

- En général, les saisons ne sont pas parfaitement délimitées
- Pas la même saison pour toutes les stations mais certaine cohérence
- Comment prendre en compte l'incertitude sur la saison ?



3 saisons



4 saisons