Projet ANR-08-RISK-03-01

Prédétermination des valeurs extrêmes de pluies et de crues (EXTRAFLO)

Programme RISKNAT 2008

Tâche III : Inter-comparaison des méthodes probabilistes

Rapport III.1 « *Comparaison des méthodes locales pour l'estimation des pluies extrêmes*»

Date : Septembre 2012

Rapport réalisé par :
(1) Météo-France, Direction de la Climatologie

Avec la participation de :
(2) Irstea, Centre d'Aix-en-Provence, OHAX
(3) EDF/DTG

Auteurs :
J.M. Veysseire [1], J.M. Soubeyroux [1], P. Arnaud [2], F. Garavaglia [3], F. Borchi [1], R. Fantin [1]

METEO FRANCE
Toujours un temps d'avance

irstea

EDF

Sommaire

1 Introduction

Ce rapport présente l'ensemble des résultats obtenus dans le cadre de l'action dénommée « pluie-méthodes locales » du projet ANR Extraflo entre 2010 et 1012, animée par Météo-France avec la participation de l'Irstea (Aix en Provence) et EDF/DTG.

2 Présentation des méthodes

Trois types de méthodes d'estimation de quantiles de durées de retour de pluies ont été évalués :

- les méthodes basées sur la théorie des valeurs extrêmes (loi GEV et loi de Pareto généralisée, notée GP). De plus, la loi exponentielle (EXPO), qui est un cas particulier de la loi de Pareto généralisée où la valeur du paramètre de forme est fixée à 0 a été utilisée pour évaluer l'apport d'une variation de ce paramètre suivant la série étudiée,
- une nouvelle méthode paramétrique (MEWP) développée par EDF/DTG,
- un modèle stochastique (SHYPRE).

Pour une sélection du meilleur candidat issu de la famille des lois sur les valeurs extrêmes, nous avons procédé en deux temps : d'abord, nous avons comparé la loi sur les valeurs extrêmes GEV basée sur les maxima annuels et GP utilisant les valeurs supérieures à un seuil à partir de la même méthode d'estimation de leurs paramètres (le maximum de vraisemblance); dans un deuxième temps, après avoir constaté que la loi GP était préférable à la loi GEV, nous avons recherché le meilleur estimateur des paramètres pour la loi GP en comparant trois méthodes d'estimation : la méthode du maximum de vraisemblance, la méthode des moments et la méthode des moments pondérés.
Nous avons comparé au final la meilleure estimation de la méthode GP avec les méthodes MEWP, SHYPRE et EXPO.

Les différentes méthodes sont étudiées principalement pour l'estimation centrale, en utilisant le meilleur ensemble de paramètres $\hat{\rho}$ fournis par la méthode d'estimation :

$$\hat{F}(x) = \int_0^x \frac{\partial F}{\partial x}(x|\hat{\rho})dx \qquad \textbf{(1)}$$

Nous ajouterons aussi une comparaison des méthodes à l'aide d'une distribution prédictive prenant en compte l'incertitude d'échantillonnage :

$$F_{\text{pred}}(x) = \int_0^x \int_\rho \frac{\partial F}{\partial x}(x|\rho)\, f(\rho)\, dx\, d\rho \qquad \textbf{(2)}$$

En supposant que la distribution d'échantillonnage $f(\rho)$ est normalement distribuée, nous créons un échantillon de l'ensemble des paramètres à l'aide des estimations de leurs moyennes, variances et corrélations.

2.1 Lois de valeurs extrêmes

La modélisation des valeurs extrêmes est présentée dans l'ouvrage de Coles (2001). Elle est basée sur la théorie des valeurs extrêmes qui indique que la distribution asymptotique du minimum ou du maximum d'un très grand nombre de variables aléatoires indépendantes équidistribuées est une loi GEV ou GP suivant le type d'échantillonnage (valeurs maximales annuelles ou valeurs supérieures à un seuil). Pour la loi GEV trois paramètres sont estimés : le paramètre de position μ, le paramètre d'échelle σ et le paramètre de forme ξ. Pour la loi GP, le seuil est fixé et deux paramètres sont estimés : le paramètre d'échelle σ et le paramètre de forme ξ. Dans les deux cas, le paramètre de forme est lié au comportement

de la queue de la distribution et définit trois sous-familles : la famille de Gumbel si ξ est proche de 0, celle de Fréchet si ξ est plus grand que 0 et celle de Weibull si ξ est inférieur à 0.

– Distribution généralisée des valeurs extrêmes.

La loi généralisée des valeurs extrêmes a été introduite par Jenkinson (1955). C'est une distribution à trois paramètres combinant trois distributions de valeurs extrêmes : Gumbel, Fréchet et Weibull. Les durées de retour à chaque station sont calculées en utilisant l'échantillon des valeurs maximales de chaque année. L'expression de la distribution du maximum annuel est :

$$F(x,\mu,\sigma,\xi) = \exp\left\{-\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}. \qquad \textbf{(3)}$$

– Distribution de Pareto généralisée.

La distribution de Pareto généralisée n'utilise pas les mêmes observations que la loi GEV (Pickands, 1975). Au lieu du maximum de chaque année, on utilise toutes les observations supérieures à un seuil défini séparément pour chaque station (en anglais POT : *Peaks Over Threshold*). Pour chaque station *i*, on choisit un certain nombre d'observations : après avoir classé toutes les observations on retient les N_i plus grandes. On conserve aussi toutes les observations égales au minimum des valeurs sélectionnées. On définit ainsi un seuil égal au minimum des valeurs sélectionnées diminué de 0,1 puisque 0,1 mm est la précision des mesures. Nous n'avons pas trouvé de règles dans la littérature sur le nombre N_i d'observations nécessaire et nous avons donc essayé différentes possibilités permettant de garder suffisamment d'observations même pour les stations ayant seulement 10 années de données ; nous avons finalement retenu quatre observations par an (voir section 4.1). Un autre choix consisterait à prendre un seuil unique pour toutes les stations, mais nous n'avons pas utilisé cette méthode à cause de la disparité entre les stations : il n'est pas possible d'utiliser un même seuil de définition de valeurs extrêmes pour des stations ayant des valeurs supérieures à 400 mm comme dans le Languedoc-Roussillon et pour des stations n'ayant pas de valeurs supérieures à 60 mm comme dans le centre de la France.

Soit donc une station *i* disposant de n_i observations pendant *m* années. Si les observations *X* sont classées suivant la valeur de la pluie :

- X_1 est la valeur minimale
- X_{n_i} est la valeur maximale
- $X_k - 0{,}1$ est le seuil, avec $k = n_i - 4*m + 1$. Toutes les observations supérieures au seuil sont considérées comme des valeurs extrêmes.

Alors, les observations conservées à chaque station suivent la loi :

$$F(x,\mu,\sigma,\xi) = \begin{cases} 1-(1+\dfrac{\xi(x-\mu)}{\sigma})^{-1/\xi} \text{ pour } \xi \neq 0 \\ 1-\exp\left(-\dfrac{x-\mu}{\sigma}\right) \text{pour } \xi = 0 \end{cases} \qquad \textbf{(4)}$$

– Estimateurs

On peut utiliser différents estimateurs des paramètres pour les lois GEV et GP. Les plus couramment utilisés sont la méthode des moments, des moments pondérés et du maximum de vraisemblance :

- La méthode des moments consiste à estimer les paramètres recherchés en égalisant certains moments théoriques (qui dépendent de ces paramètres) avec leurs contreparties empiriques. L'égalisation se justifie par la loi des grands nombres qui implique que l'on peut "approcher" une

espérance mathématique par une moyenne empirique. On est alors amené à résoudre un système d'équations.

- Les moments pondérés sont des statistiques analogues aux moments classiques qui en diffèrent en ce qu'ils sont calculés à l'aide de combinaisons linéaires des données ordonnées (Hosking, 1990). Comme pour la méthode des moments, on égalise les moments pondérés théoriques avec leurs analogues empiriques.
- L'estimation du maximum de vraisemblance consiste à trouver une estimation des paramètres telle que la vraisemblance d'avoir obtenu l'échantillon effectivement observé soit maximisée : soit une famille de distributions de probabilités dépendant d'un paramètre θ dont les éléments sont associés soit à une densité de probabilité (distribution continue), soit à une fonction de masse (distribution discrète), notée f_θ. On observe un échantillon de n valeurs x_1, x_2, ..., x_n de la distribution, et l'on calcule la densité de probabilité associée aux données observées : c'est une fonction de θ avec x_1, ..., x_n fixés, que l'on appelle la vraisemblance de l'échantillon $L(\theta) = f_\theta(x_1,\ldots,x_n \mid \theta)$. La méthode du maximum de vraisemblance recherche les valeurs de θ qui maximisent $L(\theta)$. On en trouve un exemple pour la loi GEV dans Prescott et Walden (1980).

Ashkar *et al.* (2007) expliquent comment estimer les paramètres de forme et d'échelle avec chaque méthode.

2.2 Méthode MEWP

La méthode MEWP (Multi Exponential Weather Pattern) a été introduite par Garavaglia *et al.* (2011). Elle est issue d'une combinaison de distributions exponentielles calées selon une classification en huit types de temps sur la France et en deux saisons. Un exemple de construction est présenté sur la Figure 1 extraite de la thèse de F Garavaglia en 2010. Les paramètres de chaque loi exponentielle sont obtenus selon la méthode du maximum de vraisemblance en utilisant les valeurs supérieures à un seuil relié au quantile 70% de la distribution des pluies en chaque station.

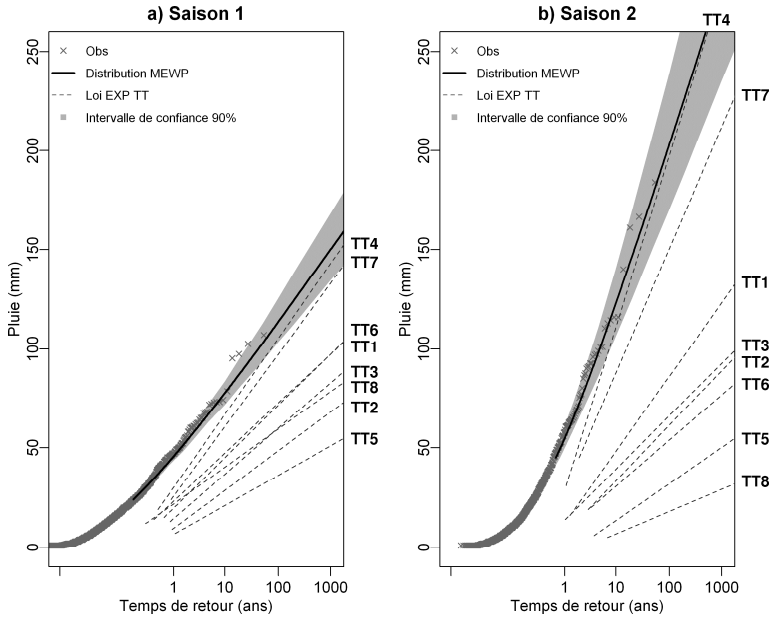


Figure 1. *Méthode MEWP (EDF-DTG), Multi Exponential Weather Pattern : combinaison de lois exponentielles avec des sous échantillon par saison (2) et type de temps (8) selon Garavaglia, 2010,*

2.3 Méthode SHYPRE

La méthode SHYPRE (Simulated HYdrographs for flood PRobability Estimation – Cernesson 1993, Arnaud 1997, Arnaud *et al.*, 2007), a été conçue pour étudier les distributions de variables hydrologiques (pluies et débits). Elle combine un modèle stochastique pour la pluie horaire avec un modèle pluie-débit (voir figure 2). L'extrapolation de la distribution de la pluie vers les grandes durées de retour est obtenue en générant beaucoup d'événements différents sur une grande période de simulation plutôt qu'en ajustant directement une distribution de probabilité théorique sur des valeurs observées. Le modèle SHYPRE est généralement initialisé avec des données horaires mais dans cette étude il a été adapté à des observations quotidiennes. Ce générateur de précipitations, testé sous différents climats (Arnaud *et al.*, 2007) a été utilisé dans cette étude dans sa version de 2009 (Cantet, 2009) avec un calage adapté sur des données journalières.

Ce générateur de pluies horaires est généralement calé à partir d'information de pluies horaires permettant une analyse des caractéristiques des hyétogrammes, en vue de leur reconstitution. En l'absence d'information horaire, le générateur peut être calé par une information journalière. Dans ce cas, certains paramètres sont fixés (car peu variables ou peu sensibles) et d'autres sont estimés à partir de variables issues de pluies journalières. Cette version, destinée à être régionalisée, est appelée SHYREG (pour SHYPRE régionalisé) : SHYREG-local si les paramètres journaliers sont déterminés à partir d'une information journalière locale (série pluviométrique) et SHYREG-régional si les paramètres ont été régionalisés.

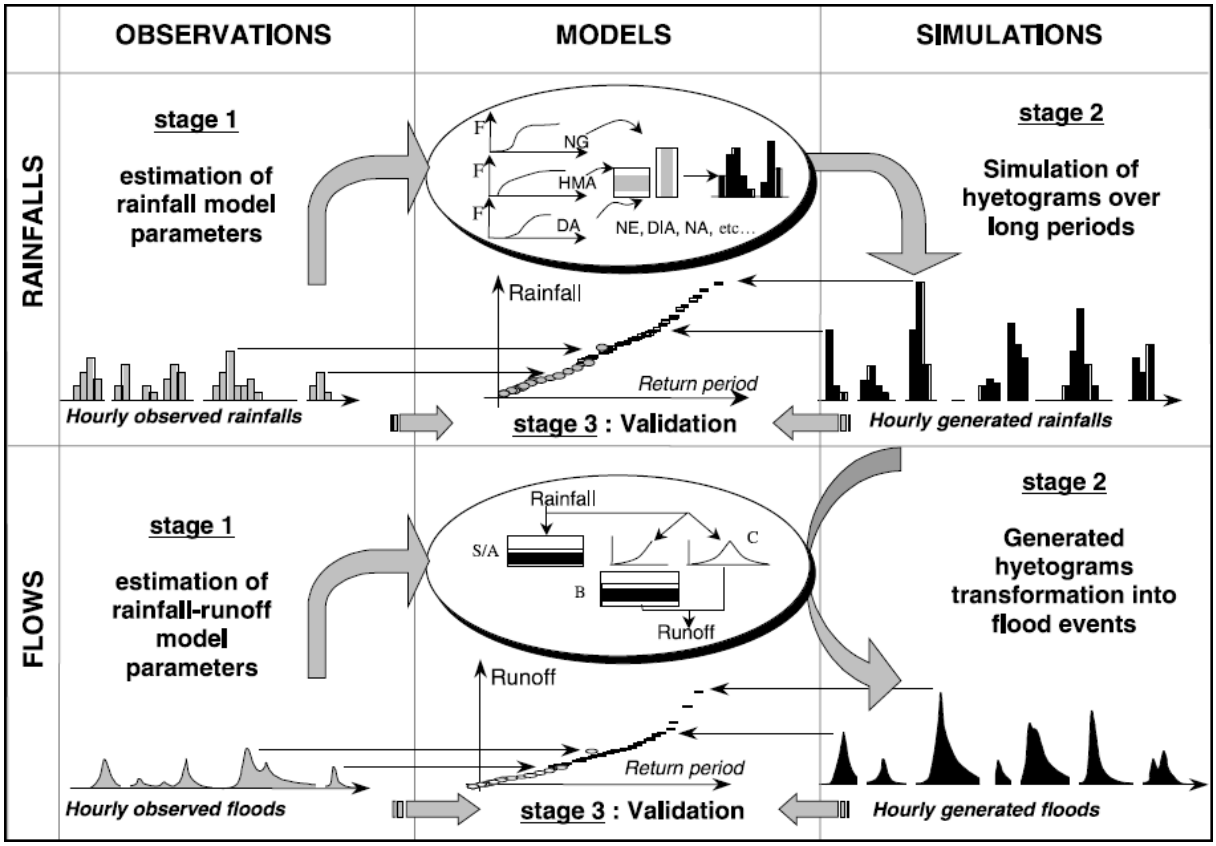


Figure 2. *Principe du modèle SHYPRE*

Localement, on peut donc déterminer ces variables journalières pour caler le générateur de pluies horaires. Ces variables journalières caractérisent les événements pluvieux normalement sélectionnés

pour être analysés par SHYPRE lorsque l'on dispose de chroniques horaires (cf. Figure 3). En l'absence de pluies horaires on retient donc uniquement les caractéristiques journalières.

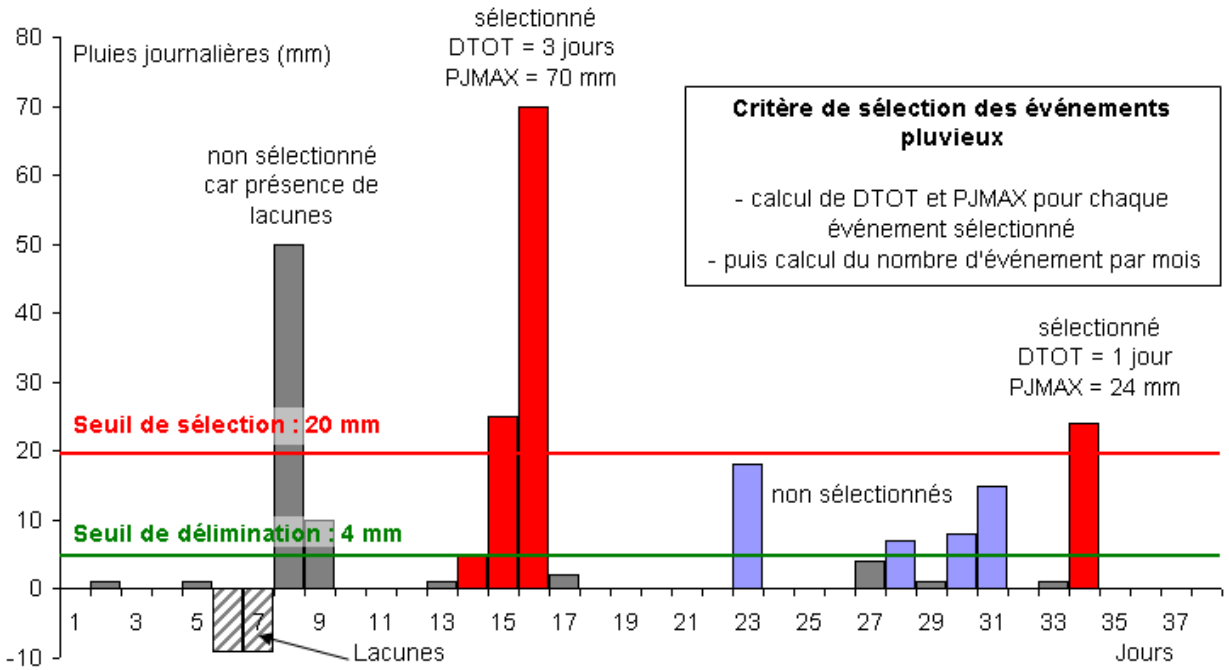


Figure 3 : critère de sélection des événements pluvieux et calcul de leur caractéristique.

La procédure mise en œuvre pour caler la méthode SHYREG est donc la suivante:

- Pour chaque mois de chaque poste disponible, on détermine le nombre d'événements pluvieux définis au sens de SHYPRE([1]). Un événement pluvieux est associé au mois de son premier jour.
- Pour chaque événement pluvieux on calcule sa durée (*DTOT* en jour) et sa pluie journalière maximale (*PJMAX* en mm).
- Pour chaque mois de chaque poste disponible, on détermine le nombre de jour en lacune.
- Un mois ayant dix jours de lacune ou plus est considéré en lacune, ainsi que tous les événements qui pourrait y être associé.

On dispose alors de caractéristiques mensuelles des chroniques de pluies : nombre d'événements pluvieux du mois (0 si le mois est considéré en lacune), la durée de chaque épisode et le pluie journalière maximale de l'événement.

On calcule alors pour chaque poste, et sur les années choisies[2] pour les différents tests d'échantillonnage, les caractéristiques suivantes :

- La moyenne des *DTOT* des événements des mois de juin à novembre : $\mu DTOT\ été$
- La moyenne des *DTOT* des événements des mois de décembre à mai : $\mu DTOT\ hiver$
- La moyenne des *PJMAX* des événements des mois de juin à novembre : $\mu PJMAX\ été$
- La moyenne des *PJMAX* des événements des mois de décembre à mai : $\mu PJMAX\ hiver$

[1] Un événement pluvieux est défini par une succession de pluies journalières supérieures à 4 mm (non bornées par des lacunes) avec la présence d'au moins une pluie journalière dépassant les 20 mm.

[2] On rappelle que par convention, l'année N est caractérisée par les données des mois de juin à décembre de l'année N et des mois de janvier à mai de l'année N+1.

- Le nombre d'événements retenus sur les mois de juin à novembre, par an : *NE été*
- Le nombre d'événements retenus sur les mois de décembre à mai, par an : *NE hiver*

Le calage du générateur de pluies est réalisé par le calcul de ces trois paramètres pour les deux saisons définies : l'été de juin à novembre et l'hiver de décembre à mai.

On peut alors simuler des chroniques de pluies horaires (séries d'événements non datés) sur les deux saisons définies. On extrait alors de ces simulations les caractéristiques des pluies horaires générées : les pluies maximales en 1, 2, 3 … 72 heures de chaque événement (*PMd*). On trace ensuite les distributions empiriques de ces caractéristiques pour en extraire certains quantiles.

On rappelle ici que les distributions de fréquences issues de SHYPRE sont des distributions empiriques associées aux caractéristiques des événements pluvieux horaires générés. C'est donc un produit de contrôle des capacités du générateur à reproduire des pluies horaires dont les caractéristiques statistiques sont proches des chroniques observées. En aucun cas ces distributions ne sont issues d'un ajustement d'une loi statistique sur les mêmes caractéristiques observées.

Les simulations effectuées correspondent à la simulation d'une centaine d'échantillons de 500 ans. La distribution moyenne des cents distributions déduites des 500 ans de simulation nous permet d'obtenir une distribution central relativement peu soumis à l'échantillonnage des simulations. Ce point a déjà été abordés dans différentes études qui montre qu'une centaine de simulations reste un minimum pour stabilité des estimations par SHYPRE (Arnaud, Lang et al. 1998; Muller 2006).

Ce générateur de précipitations, testé sous différents climats (Arnaud *et al.*, 2007) a été utilisé dans cette étude dans sa version de 2009 (Cantet, 2009) avec un calage adapté sur des données journalières.

3 Jeux de données et méthodologie

3.1 Sélection des séries climatologiques

La sélection du jeu de données pluviométriques du projet Extraflo a visé à rassembler les meilleures séries climatologiques françaises en termes de qualité (moins de 10% de valeurs manquantes, séries contrôlées et validées) et de longueur (notamment séries de plus de 50 ans). Une attention particulière a été portée aux régions méditerranéennes concernées par les pluies journalières extrêmes les plus fortes (au-delà de 500 mm en 24 h - voir le site pluiesextremes.meteo.fr). La figure 4 présente la carte des 1568 séries utilisées.

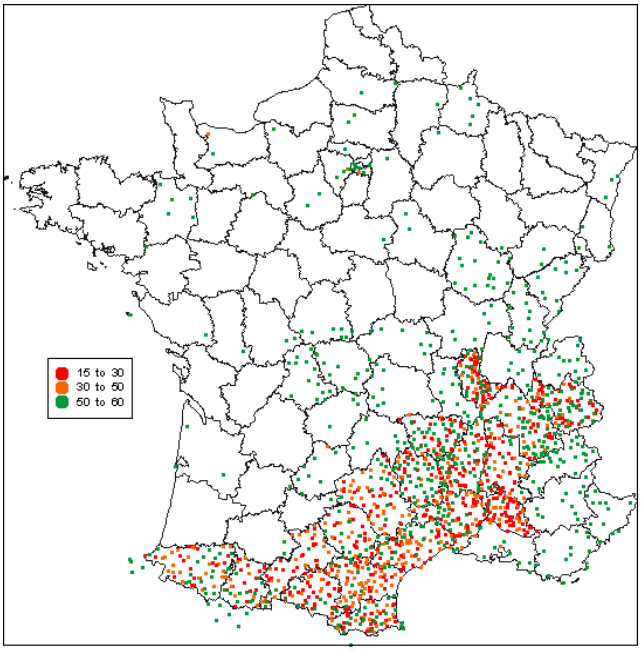


Figure 4. *Carte des 1568 stations utilisées dans le projet EXTRAFLO : les points verts signalent les séries de longueur supérieure à 50 ans, les points orange avec entre 30 et 50 ans, les points rouges avec entre 15 et 30 ans.*

Pour parvenir à cette sélection, le projet Extraflo a rassemblé un ensemble de données sur une grande partie de la France en utilisant les archives de l'EDF et de Météo-France. La sélection des séries vise à représenter les différentes régions climatiques pour les précipitations extrêmes en France (Choisnel et Payen, 1988): océanique, continentale, de montagne. La longueur des séries a été un critère déterminant dans la constitution des jeux de données. Toutes les données utilisées avaient été soumises préalablement à des contrôles de qualité particuliers, dans le cadre de leur utilisation opérationnelle. Pour ce projet, seules les séries ayant moins de 10% de données manquantes ont été sélectionnées.

L'ensemble des données a été obtenu à partir de trois ensembles de séries quotidiennes (voir *Tableau 1* t figure 4) et elles peuvent être divisées en deux classes :

- Un ensemble de longues séries de données de longueur supérieure à 50 ans (points verts sur la figure 4) : 446 séries fournies par EDF (364 séries) et Météo-France (82 séries). Les séries d'EDF ont été déjà utilisées dans une étude précédente pour la validation de la méthode SCHADEX (Garavaglia *et al.*, 2011) et ont été soumises à un contrôle complet de leur qualité. Ces séries ont des données disponibles depuis 1950 jusqu'à 2005. Ces stations sont principalement situées dans les Alpes, les Pyrénées et le Massif Central à une altitude moyenne de 620 m. Les données de Météo-France sont des séries de SQR (Séries Quotidiennes de Référence) préparées pour des études sur le changement climatique (Moisselin *et al.*, 2002). Ces séries ont été vérifiées par une méthode d'homogénéisation (Mestre, 2004) avec un test pour la détection des points de rupture : seules les

meilleures séries n'ayant pas de point de rupture important (inférieur à 10 % de la valeur moyenne mensuelle) ont été utilisées. Elles sont principalement localisées en plaine (altitude moyenne de 200 m).

- Un ensemble dense de données dans le Sud de la France avec des séries de plus de 15, 30 ou 50 années (respectivement points rouges, orange et verts sur la figure 4) : 1122 séries fournies par Météo-France à partir de la Banque de Données Climatologiques (BDClim) et sélectionnées pour leur situation dans le Sud de la France et la Région Méditerranéenne (altitude moyenne 500 m). Ces séries ont été contrôlées selon les règles du guide d'exploitation climatologique de Météo-France et sont disponibles sur le serveur climatologique en ligne "Climathèque" : http://climatheque.meteo.fr/.

Tableau 1 : *Ensembles de données*

	Période retenue	Nombre moyen d'années	Nombre de stations	Réseau
Longues séries	1948-2005	57	364	EDF
	1951-2003	52	82	Météo France (SQR)
Ensemble dense	1950-2009	35	1122	Météo-France BDClim)

Ces stations se comportent de façon très différente en ce qui concerne les précipitations extrêmes, ce dont nous pouvons rendre compte en analysant la distribution du ratio entre la moyenne des maximums annuels et le cumul annuel moyen (Penot, 2011-2014) qui illustre l'écart entre les valeurs extrêmes et moyennes. La figure 4 s'appuyant sur le jeu des 693 séries de plus de 50 ans, met en évidence la pertinence climatique de cette approche avec un zonage utilisant comme borne les ratios de 0.07 et 0.097 (resp. quantiles 70% et 90%) : valeurs fortes sur l'arc méditerranéen ; valeurs intermédiaires sur les reliefs du sud de la France, Cévennes et Alpes du Sud notamment ; valeurs plus faibles ailleurs.

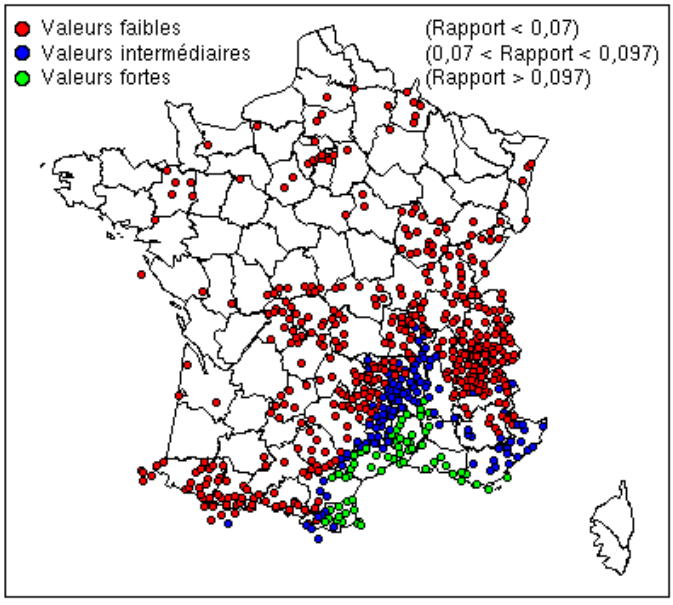


Figure 5. *Distribution du rapport entre la moyenne des maximums annuels et le cumul annuel moyen. Échantillon complet, stations de 50 années ou plus.*

3.2 Constitution des échantillons calage/validation

Nous utiliserons huit échantillonnages différents divisés en trois catégories : échantillonnage calage-validation pour vérifier la justesse des modèles, échantillonnage "échantillon 1-échantillon 2" pour

vérifier la robustesse des estimations et échantillonnage “échantillon complet-échantillon sans la valeur maximale ” pour vérifier la robustesse de la méthode vis-à vis de la valeur maximale. Pour chaque catégorie, nous disposons de différents échantillons pour tester l’impact de la longueur de la série sur les scores (voir tableau 2).

Tableau 2 : *Description des échantillons utilisés.*

	Nombre de stations
Échantillonnage Calage - Validation (Catégorie 1)	
C50V50	693
C33V66	693
Échantillonnage Échantillon 1 - Échantillon 2 (Catégorie 2)	
10 ans – 10 ans	1287
15 ans – 15 ans	1016
25 ans – 25 ans	671
Échantillonnage Échantillon complet –sans la valeur maximale (Catégorie 3)	
20 ans	1568
30 ans	1040
50 ans	693

– Echantillonnage calage - validation

Le but est ici de vérifier si les estimations calculées sur un échantillon de la station *i* appelé échantillon de calage sont proches des quantiles observés sur un autre échantillon de la même station appelé échantillon de validation (justesse de l’estimation). Pour chaque station ayant 50 années de données ou plus, nous séparons de façon aléatoire les observations en deux groupes : ceci représente 693 stations. Toutes les observations d’une même année sont soit dans le groupe de calage, soit dans le groupe de validation. Deux sortes d’échantillonnage sont utilisées pour les échantillons calibrage-validation. La première sorte utilise la moitié des années pour le calage et l’autre moitié pour la validation (C50V50). La deuxième sorte utilise le tiers des années pour le calage et deux tiers pour la validation (C33V66).
Nous avons décidé de séparer en deux groupes les années entières et pas les observations individuelles pour créer des groupes utilisables et comparables pour toutes les méthodes. En effet, toutes les méthodes n’utilisent pas les observations de la même façon : par exemple, la méthode GEV n’utilise qu’une observation par an, la méthode GP en utilise plusieurs, mais pas toutes.

– Échantillonnage échantillon1 - échantillon 2

Nous voulons vérifier si deux estimations calculées sur des échantillons différents d’une même station donnent des résultats semblables (robustesse de l’estimation). Pour cela, nous séparons de façon aléatoire les observations en deux groupes d’années : pour chaque station, chaque échantillon contient la moitié des années. Trois échantillonnages sont utilisés, à partir respectivement des stations ayant au moins 20, 30 ou 50 années de données.

– Échantillonnage complet ou sans la valeur maximale

Le but est de vérifier si la valeur maximale n’a pas un poids trop fort sur l’estimation des paramètres. Pour chaque station, nous conservons toutes les données dans un premier échantillon, et nous enlevons toutes les observations de l’année ayant la valeur maximale pour constituer le second échantillon. Ici aussi, trois échantillonnages sont utilisés, à partir respectivement des stations ayant au moins 20, 30 ou 50 années de données.

3.3 Critères de comparaison

Les méthodes d'estimation des valeurs extrêmes ont été évaluées selon des critères mis au point dans le cadre du projet et permettant de caractériser d'une part leur justesse, mesurant la capacité d'un modèle à donner une valeur proche de la valeur réelle et d'autre part leur robustesse, capacité pour un modèle à donner des estimations proches avec des échantillons différents. On trouvera les formules des scores dans le tableau 3 (fin de section 3).

Nous considérons donc deux qualités : la robustesse et la justesse. La robustesse, qui est la capacité d'un modèle à donner la même estimation sur différentes périodes de calage, est mesurée par deux scores le $SPAN_T$ et le $COVER_T$ sur des échantillons de catégorie 2 et 3. La justesse, qui est la capacité d'un modèle à donner des estimations proches de la vraie valeur, est mesurée par deux critères N_T et FF sur des échantillonnages calage-validation. Ces quatre scores qui ont été introduits par Renard *et al.* (2013) seront calculés pour deux durées de retour : $T = 10$ ans et 100 ans. T est la durée de retour théorique, c'est-à-dire l'inverse de la probabilité qu'une quantité de pluie soit dépassée pendant l'année.

D'autre part nous souhaitons vérifier si la qualité des estimateurs GP dépend du paramètre de forme. Dans ce but, nous calculons les critères sur des sous-échantillons, créés en plusieurs étapes. Tout d'abord, nous ajustons une distribution GP sur chaque station avec les méthodes ML, MM et PWM. Ensuite, nous calculons la moyenne des trois estimations du paramètre de forme ξ et répartissons les stations dans l'un des cinq groupes suivants : $\xi < -0{,}1$; $\xi \in [-0{,}1;\ 0]$; $\xi \in [0;\ 0{,}1]$; $\xi \in [0{,}1;\ 0{,}2]$; $\xi > 0{,}2$. Pour finir nous calculons les critères pour chaque groupe.

3.3.1 Robustesse

3.3.1.1 $SPAN_T$

Le critère $SPAN_T$ est utilisé pour évaluer la stabilité de l'estimation de la durée de retour T, en calculant la différence entre les estimations faites sur deux échantillons différents d'une même station. Il a été proposé par Garavaglia *et al.* (2010). Pour chaque station i et pour chaque durée de retour T, nous calculons un score positif $SPAN_T$, la valeur optimale du score étant 0.

$$SPAN_{T,i} = 2\frac{\left|\hat{q}_{T,i}(C1_i) - \hat{q}_{T,i}(C2_i)\right|}{\hat{q}_{T,i}(C1_i) + \hat{q}_{T,i}(C2_i)} \qquad \textbf{(5)}$$

Ensuite nous calculons un score global pour chaque durée de retour.

$$SPAN_T = 1 - \frac{1}{N}.\sum_{i=1}^{N} SPAN_{T,i} \qquad \textbf{(6)}$$

Théoriquement ce critère $SPAN_T$ peut être négatif si les estimations faites sur les deux échantillons sont complètement différentes. Comme il reste toujours positif dans notre étude, nous avons choisi cette formulation pour pouvoir le reporter facilement sur le même graphe que les autres scores compris entre 0 et 1.

3.3.1.2 $COVER_T$

Le critère $COVER_T$ est utilisé pour évaluer la capacité du modèle à calculer la variance des estimations. En effet, si les estimations sont comparables mais que les intervalles de confiance sont disjoints, ceci signifie que la variance des estimations est sous-estimée. La limite de ce critère est qu'il n'est pas possible de déterminer si la variance est surestimée : nous supposerons que ce n'est pas le cas.

Pour chaque station *i* et pour chaque durée de retour *T*, nous calculons un score $COVER_T$ basé sur un intervalle de confiance à 90 pour cent du quantile $\hat{q}_{T,i}$ (α = 0,1). Soient $a_{\alpha,i}$ et $b_{\alpha,i}$ les bornes de la partie commune des intervalles de confiance des deux estimations.

$$a_{\alpha,i} = \max(\hat{q}_{\alpha/2,i}(C1_i), \hat{q}_{\alpha/2,i}(C2_i)) \quad \textbf{(7)}$$

$$b_{\alpha,i} = \min(\hat{q}_{1-\alpha/2,i}(C1_i), \hat{q}_{1-\alpha/2,i}(C2_i)) \quad \textbf{(8)}$$

$$COVER_{T,i} = \frac{\mathrm{P}(a_{\alpha,i} < \hat{q}_{T,i}(C1_i) < b_{\alpha,i})\mathrm{P}(a_{\alpha,i} < \hat{q}_{T,i}(C2_i) < b_{\alpha,i})}{(1-\alpha)^2} \quad \textbf{(9)}$$

Ensuite nous calculons un score global pour chaque durée de retour.

$$COVER_T = \frac{1}{N}.\sum_{i=1}^{N} COVER_{T,i} \quad \textbf{(10)}$$

Le score $COVER_T$ est compris entre 0 et 1. 1 est le score optimal.

Les critères $COVER_T$ et $SPAN_T$ sont corrélés positivement. Si les estimations venant de deux échantillons de la même station sont très différentes, ce qui est la définition d'un $SPAN_T$ proche de 0, il est probable que l'on obtiendra un score $COVER_T$ proche de 0 même si l'estimation de la variance est correcte. Donc si deux modèles donnent des scores $COVER_T$ différents, il y a deux possibilités : si les scores $SPAN_T$ sont proches, la différence des scores $COVER_T$ peut être interprétée comme une différence dans la qualité de l'estimation de la variance. Mais si les scores $SPAN_T$ des deux modèles sont aussi différents, il est difficile d'évaluer l'impact de cette différence sur la différence observée des $COVER_T$. Donc nous n'interpréterons la différence de critère $COVER_T$ entre deux modèles que si leurs scores $SPAN_T$ sont proches.

3.3.2 Justesse

3.3.2.1 *FF*

Le critère *FF* est utilisé pour estimer la justesse de l'estimation de la probabilité de non-dépassement de la valeur maximale observée d'un échantillon de taille fixée. Il est basé sur une procédure de division d'un échantillon et a été introduit par Garçon (1995). Chaque échantillon de la station *i* est divisé en un échantillon de calage 1 de taille $N_{1,i}$ et un échantillon de validation 2 de taille $N_{2,i}$.
Soient $m_{2,i}$ la valeur maximale de l'échantillon de validation de la station *i*. Soit $F_{1,i}$ la fonction de répartition calculée sur l'échantillon de validation. Si le modèle est parfait, $F_{1,i}(m_{2,i})$ suit une distribution de Kumaraswamy de paramètres $N_{2,i}$ et 1, i.e. $K[N_{2,i},1]$ (Kumaraswamy, 1980) :

$$FF_i = \left[F_{1,i}(m_{2,i})\right]^{N_{2,i}} \quad \textbf{(11)}$$

Nous calculons alors la différence entre les FF_i et la première bissectrice. Soient (FF'_i) les (FF_i) classés en ordre croissant.

$$FF = 1 - \frac{2}{N}.\sum_{i=1}^{N}\left|FF'_i - \frac{i}{N+1}\right| \quad \textbf{(12)}$$

Le critère *FF* est compris entre 0 et 1. 1 est le score optimal.

3.3.2.2 N_T

Le critère N_T est utilisé pour vérifier si les quantiles calculés sont cohérents avec les observations de l'échantillon de validation. Soit $q_{T,i}$ le quantile estimé associé à la durée de retour *T*, et $N_{T,i}$ le nombre

d'observations de l'échantillon de validation supérieures à $q_{T,i}$. Si l'estimation est juste, $N_{T,i}$ est une réalisation d'une distribution binomiale :
Le dépassement du quantile $q_{T,i}$ est une épreuve de Bernoulli, de probabilité de succès

$$\Pr(X > q_{T,i}) = 1/T \qquad \textbf{(13)}$$

$N_{T,i}$ est le nombre de succès parmi $N_{2,i}$ essais, et suit donc une loi binomiale.

$$N_{T,i} \sim \text{Binomiale}\,(N_{2,i}, \frac{1}{T}) \qquad \textbf{(14)}$$

Soient $N'_{T,i}$ les probabilités de dépassement des $N_{T,i}$ classées en ordre croissant.

$$N_T = 1 - \frac{2}{N} . \sum_{i=1}^{N} \left| N'_{T,i} - \frac{i}{N+1} \right| \qquad \textbf{(15)}$$

Le critère N_T est compris entre 0 et 1. 1 est le score optimal.

Sous cette forme, le score N_T n'est pas adapté à la durée de retour 100 ans : du fait de son caractère discret il n'est pas possible d'obtenir un score N_T proche de 1, même avec une bonne justesse de l'estimation. En effet pour une durée de retour T grande devant $N_{2,i}$ les nombres $N_{T,i}$ seront souvent nuls du fait de la rareté de l'événement, la longueur des séries étant limitée à 50 années : un grand nombre de $N'_{T,i}$ seront égaux à $\Pr(N_{T,i} > 0) = \left(1 - \frac{1}{T}\right)^{N_{2,1}}$ et seront donc éloignés de la première bissectrice. Pour éviter cela les probabilités de dépassement $N'_{T,i}$ ont été modifiées de la façon suivante :

Pour une loi binomiale $(N_{2,i}, \frac{1}{T})$:

$f_1 = \Pr(X > N_{T,i})$

$f_2 = \Pr(X > N_{T,i} - 1)$ si $N_{T,i} \neq 0$, $\quad f_2 = 0$ sinon

$N'_{T,i} = f_2 + U(0,1) \times (f_1 - f_2)$

où $U(0,1)$ est un nombre tiré au hasard entre 0 et 1

Par exemple, si $N_{T,i} = 0$, $N'_{T,i} = U(0,1) \times \left(1 - \frac{1}{T}\right)^{N_{2,i}}$.

3.3.3 Interprétation

Ces quatre critères permettent de comparer les méthodes mais les scores globaux N_{10} et FF doivent être utilisés avec précaution. En effet, ils peuvent donner de bons résultats s'il y a autant de cas où les estimations sont sous-estimées que de cas où elles sont surestimées. Il est donc important de vérifier si on trouve le même résultat global sur des sous-échantillons, en utilisant des graphes.

Nous porterons donc les fréquences empiriques sur l'axe des x et les $N_{T,i}$ ou FF_i classés en ordre croissant sur l'axe des y. Les graphes de $N_{T,i}$ et FF_i permettent de déterminer deux propriétés des modèles : si les estimations des quantiles sont surestimées ou sous-estimées, et si les modèles sont sur-paramétrés (Garavaglia et al. 2010).[3]

[3] Dans le rapport II.1 sur la méthodologie de comparaison et les actions de comparaison III.2 à III.6, nous avons choisi de mettre en abscisse les valeurs classes des scores et en ordonnée les fréquences empiriques. Le présent rapport III.1 a été rédigé avant ce choix méthodologique. Il n'a pas été actualisé ensuite.

Si la courbe des $N_{T,i}$ est toujours au-dessous de la bissectrice, les estimations des quantiles sont sous-estimées. Il y a trop de cas dans l'échantillon de validation où les valeurs sont supérieures aux quantiles calculés avec l'échantillon de calage. Inversement, si la courbe des N_T est toujours au-dessus de la bissectrice, les estimations des quantiles sont surestimées. C'est le contraire avec le critère FF_i ; si la courbe des FF_i est toujours au-dessous de la bissectrice, la durée de retour associée à la valeur maximale est sous-estimée, et donc le quantile que nous calculons pour la durée de retour T sera surestimé. Inversement, si la courbe des FF_i est toujours au-dessus de la bissectrice, les estimations des quantiles sont sous-estimées.

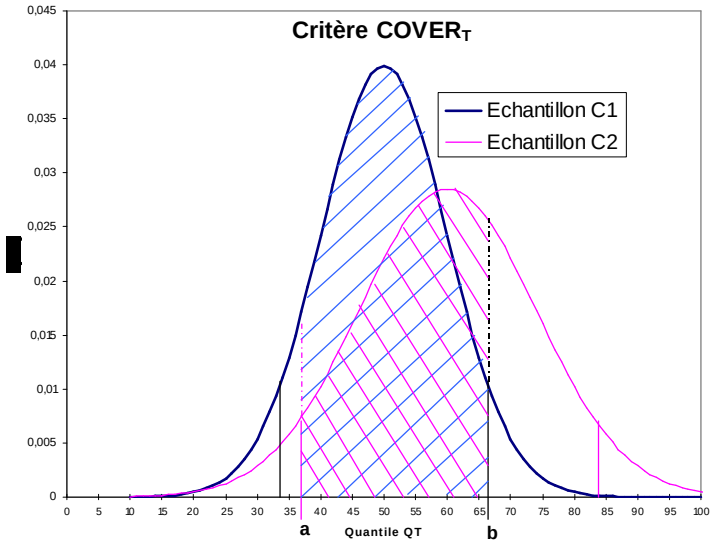
L'analyse des courbes des FF_i et des $N_{T,i}$ permet de savoir si les modèles sont surparamétrés. En effet, si la courbe est au-dessus de la bissectrice jusqu'à un certain point, et au-dessous de la bissectrice après ce point, le modèle est surparamétré. Ses prévisions dépendent trop des données de l'échantillon de calage.

Comme tous les critères ont été calibrés pour avoir leur valeur maximale égale à 1, on peut résumer les conclusions sur un diagramme en étoile où l'on porte à partir d'un point fixe les valeurs de chaque critère sur différents rayons régulièrement espacés : une méthode est d'autant meilleure que la courbe joignant ses critères est éloignée du centre.

Tableau 3 : *Scores définis dans le cadre du projet visant à comparer la justesse (FF et NT) et la robustesse (SPANT et COVERT) des lois d'estimation des valeurs extrêmes (Renard et al., 2013)*

Score	**Objectif**	**Mode de calcul**
FF	**Justesse :** capacité à fournir une estimation de la probabilité de la valeur maximale qui soit cohérente avec les observations	**Statistique** *FF* : Pour une station *i* parmi *NS*, soient $F_{1,i}$ la distribution de l'échantillon de calage, et $m_{2,i}$ le maximum de l'échantillon de validation de taille $N_{2,i}$ $FF_i = F_{1,i}(m_{2,i})$ est une réalisation de la statistique *FF* qui suit une loi de probabilité de Kumaraswamy : $K(x) = P(FF < x) = x^{N_2}$
		Report graphique : $K[FF_i]$: classement par ordre croissant des *NS* valeurs $(FF_i)^{N_2,i}$ Courbe expérimentale : en abscisse $i/(NS+1)$; ordonnée $K[FF_i]$
		Calcul du score : Score(*FF*)=1-2.Aire(surface comprise entre la courbe expérimentale et la bissectrice) $= 1 - [2/(NS+1)] \sum_{i=1}^{NS} \lvert K[FF_i] - i/(NS+1) \rvert$ Score(*FF*) compris entre 0 et 1 (1 est le score optimal)
N_T	**Justesse :** cohérence du nombre de dépassements d'un quantile avec sa période de retour de référence	**Statistique** N_T : $N_{T,i}$ le nombre de dépassement d'un quantile $\hat{q}_{T,i}$, estimé sur un échantillon *i*, de taille N_i $N_{T,i}$ est une réalisation de la statistique N_T qui suit une loi binomiale : $B(k) = P(N_T \le k) = \sum_{j=0}^{k} C_N^j (1/T)^j (1-1/T)^{N-j}$
		Report graphique : $B[N_{T,i}]$: classement par ordre croissant des *NS* valeurs $B(N_{T,i})$ Courbe expérimentale : en abscisse $i/(NS+1)$; ordonnée $B[N_{T,i}]$

Score	Objectif	Mode de calcul
		Calcul du score : Score(N_T)=1-2.Aire(surface comprise entre la courbe expérimentale et la bissectrice) $= 1 - [2/(NS+1)] \sum_{i=1}^{NS} \left| B[N_{T,i}] - i/(NS+1) \right|$ Score(N_T) compris entre 0 et 1 (1 est le score optimal)
$SPAN_T$	**Robustesse :** stabilité de l'estimation d'un quantile de crue pour deux périodes de calage différentes	**Statistique** $SPAN_T$: Soient $\hat{q}_{T,i}(C_1)$ et $\hat{q}_{T,i}(C_2)$, les estimations d'un quantile de période de retour *T*, sur deux échantillons C_1 et C_2 d'une station *i*. $SPAN_{T,i} = 2\left|\hat{q}_{T,i}(C_1) - \hat{q}_{T,i}(C_2)\right| / \left[\hat{q}_{T,i}(C_1) + \hat{q}_{T,i}(C_2)\right]$
		Report graphique : $SPAN[_{T,i}]$: classement par ordre croissant des *NS* valeurs $SPAN_{T,i}$ Courbe expérimentale : en abscisse $i/(NS+1)$; ordonnée $SPAN'_{T,i}$
		Calcul du score : Score($SPAN_T$) = 1 – Moyenne($SPAN_T$) / $MaxMoySPAN_T$ avec $MaxMoySPAN_T = \underset{i=1,N}{Max}\left(Moyenne(SPAN_T)(\text{méthode } Mi)\right)$ Score($SPAN_T$) compris entre 0 et 1 (1 est le score optimal, correspondant à une estimation identique pour les deux échantillons ; 0 correspond au cas de la moins bonne méthode testée)
$COVER_T$	**Robustesse :** stabilité de l'estimation de l'intervalle de confiance d'un quantile de crue pour deux périodes de calage différentes	**Statistique** $COVER_T$: Soient $a_{\alpha,i}$ et $b_{\alpha,i}$ les limites de l'intervalle de confiance de niveau α (en %) du quantile $\hat{q}_{T,i}$ pour une station *i*, avec deux échantillons : Échantillon C_1 : $P\left[a_{\alpha,i}(C_1) < \hat{q}_{T,i}(C_1) < b_{\alpha,i}(C_1)\right] = \alpha$ Échantillon C_2 : $P\left[a_{\alpha,i}(C_2) < \hat{q}_{T,i}(C_2) < b_{\alpha,i}(C_2)\right] = \alpha$ Le croisement des deux intervalles de confiance donne les limites : $a = Max(a_{\alpha,i}(C_1); a_{\alpha,i}(C_2)) \quad et \quad b = Min(b_{\alpha,i}(C_1); b_{\alpha,i}(C_2))$ La probabilité de recouvrement des deux intervalles vaut : $COVER_{T,i} = (1/\alpha^2)\, P\left[a < \hat{q}_{T,i}(C_1) < b\right] P\left[a < \hat{q}_{T,i}(C_2) < b\right]$

Score	Objectif	Mode de calcul
		Report graphique : *COVER*$_{[T,\,i]}$: classement par ordre croissant des *NS* valeurs *COVER*$_{T,\,i}$ Courbe expérimentale : en abscisse *i*/(*NS*+1) ; ordonnée *COVER*$_{[T,\,i]}$ 
		Calcul du score : Score(*COVER*$_T$)= Moyenne(*COVER*$_T$) $= (1/NS)\sum_{i=1}^{NS} COVER_{T,i}$ Score(*COVER*$_T$) compris entre 0 et 1 (1 est le score optimal, correspondant à un recouvrement intégral des deux intervalles)

4 Résultats

4.1 Choix du seuil pour une loi GP

La première question à examiner est celle du choix du seuil : si ce seuil est trop bas, on n'a peut-être pas atteint le domaine de validité de l'approximation asymptotique par une loi de Pareto généralisée et on va introduire un biais ; s'il est trop élevé, on risque de ne plus avoir suffisamment de valeurs dans l'échantillon et donc d'avoir une variance de l'estimation très élevée. On trouve par exemple dans Coles (2001) plusieurs façons de déterminer le seuil, les méthodes graphiques correspondantes étant disponibles dans le package R « extRemes » (Gilleland et Katz, 2005). Les durées de retour de précipitation calculées opérationnellement à Météo-France utilisent le seuil pour lequel le test du χ^2 indique le meilleur ajustement à la loi de Pareto.

Nous avons choisi ici de prendre un seuil égal à un quantile fixé. Cinq valeurs du seuil ont été testées selon leur impact en termes de justesse (score FF) : Q365, Q182, Q122, Q91 et Q73, soient les valeurs dépassées en moyenne 360, 180, 120, 91 ou 73 jours par an.

Le meilleur résultat (tableau 4) est obtenu par le quantile Q91 très légèrement meilleur respectivement que Q122 ou Q73. Pour la suite de l'étude, ce seuil correspondant à la prise en compte moyenne de quatre valeurs supérieures au seuil par année d'échantillon, sera systématiquement utilisé.

Tableau 4 : *Justesse d'une loi GPD en fonction du choix du seuil basé sur différents quantiles de 365 à 73. Échantillonnage de catégorie 1, C50V50.*

Seuils loi GPD	Q365	Q182	Q122	Q91	Q73
Score FF	0,890	0,915	0,930	0,933	0,929

4.2 Comparaison des méthodes GP et GEV

On compare à présent la justesse et la robustesse des méthodes GEV et GP en utilisant l'estimation du maximum de vraisemblance à l'aide des quatre scores : $SPAN_T$, $COVER_T$, N_T et FF.
En premier lieu, on met en évidence que les estimations GEV et GP sont extrêmement corrélées. Les estimations des quantiles de durée de retour 10 ans calculées avec l'échantillon complet des stations ayant 50 années de données ou plus ont un coefficient de corrélation supérieur de 0,99 et de 0,95 pour les durées de retour 100 ans (voir Figure 5). La différence moyenne entre les deux estimations est égale à 1,9% de la moyenne de ces estimations à 10 ans et 7% à 100 ans. Cependant les estimations GP sont plus élevées que les estimations GEV dans 76% des cas à 10 ans et 60% des cas à 100 ans. La moyenne du quantile de durée de retour 10 ans est de 92,3 mm avec les estimations GEV contre 93,7 mm avec les estimations GP et celle du quantile de durée de retour 100 ans est de 143,2 mm avec les estimations GEV contre 146,7 mm avec les estimations GP.

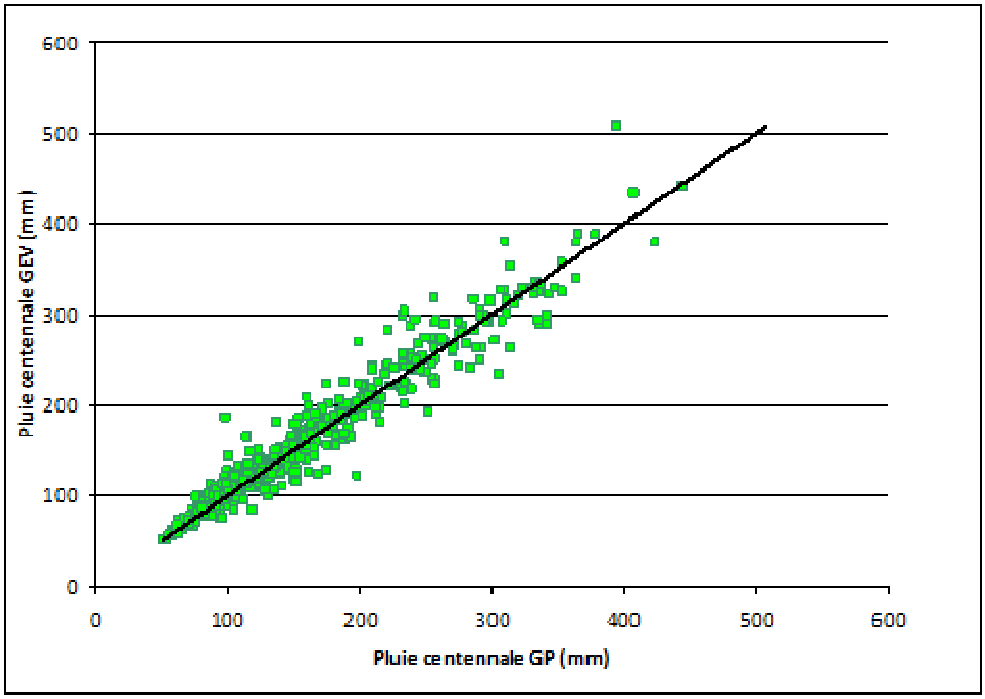


Figure 6. *Comparaison des quantiles de pluie centennale loi GP (mm) vs loi GEV (mm)*

Donc les estimations des quantiles avec GEV et GP sont très corrélées mais les estimations GEV sont en moyenne inférieures aux estimations GP. Nous allons analyser l'impact de cette différence sur la justesse des modèles.

4.2.1 Justesse

Le critère N_{10} ne montre pas de différence réelle entre les modèles GEV et GP (voir tableau 5). Ceci était attendu puisque nous avons vu précédemment que les estimations pour la durée de retour 10 ans avec GEV et GP sont très proches ; le critère N_{100} donne quant à lui une légère préférence à la loi GP. Sur les graphiques (voir figures 7 à 10), les deux courbes N_{10} restent proches et le plus souvent au-dessous de la bissectrice : les estimations du quantile de durée de retour 10 sont sous-estimées pour les deux modèles GEV et GP. La sous-estimation persiste pour le quantile de durée de retour 100 ans, davantage pour GEV que pour GP.

Tableau 5. *Comparaison entre les estimations GEV et GP basée sur les critères N10 et N100. Échantillonnage de catégorie 1*

	25 ans – 25 ans		17 ans – 33 ans	
	N_{10}	N_{100}	N_{10}	N_{100}
GEV	0,91	0,84	0,81	0,76
GP	0,90	0,88	0,80	0,83

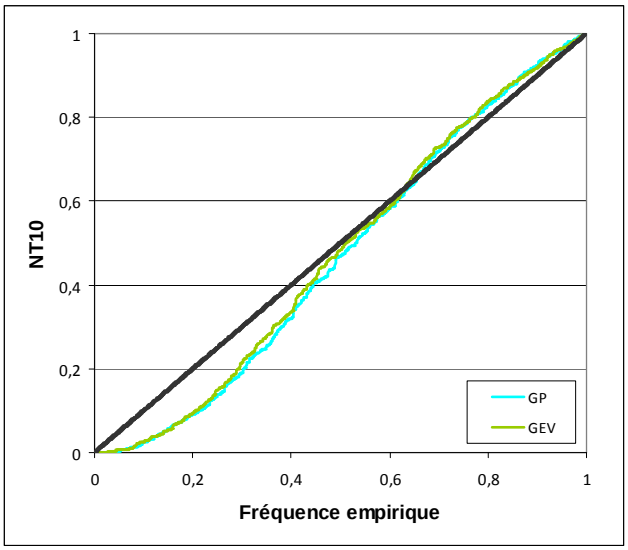


Figure 7. *Comparaison entre les estimations GEV et GP basée sur le critère N_{10}. Échantillonnage de catégorie 1, C50V50.*

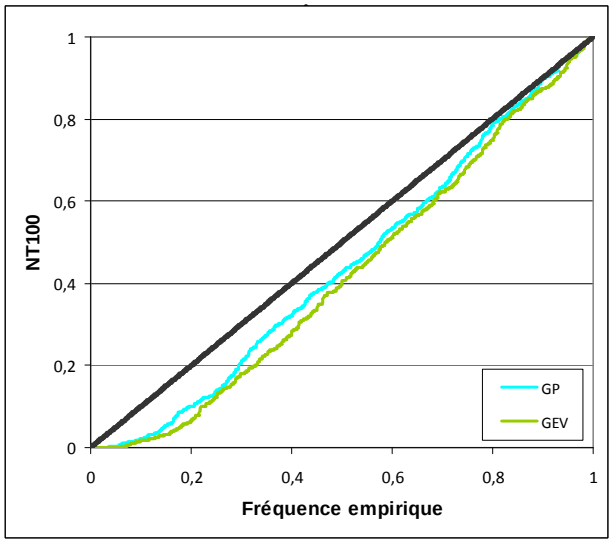


Figure 8. *Comparaison entre les estimations GEV et GP basée sur le critère N_{100}. Échantillonnage de catégorie 1, C50V50*

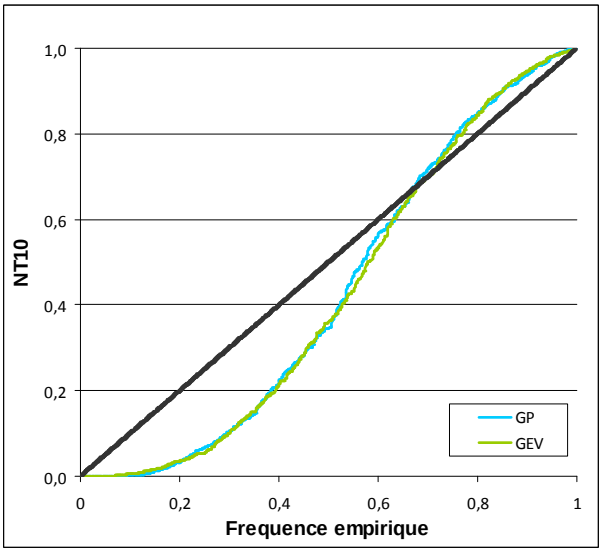


Figure 9. *Comparaison entre les estimations GEV et GP basée sur le critère N_{10}. Échantillonnage de catégorie 1, C33V66*

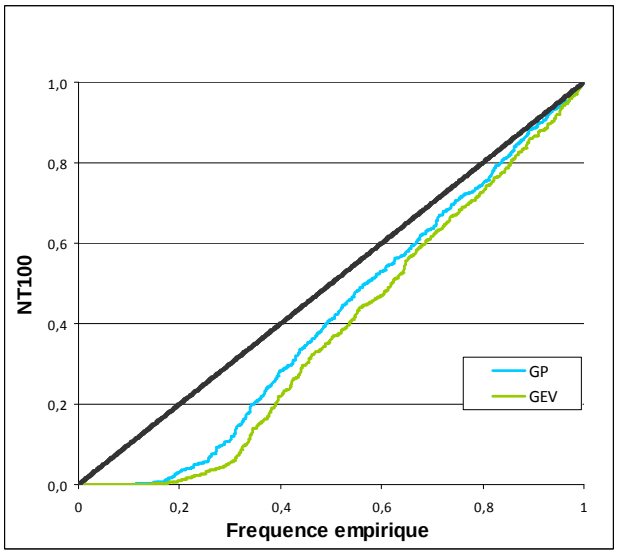


Figure 10. *Comparaison entre les estimations GEV et GP basée sur le critère N_{100}. Échantillonnage de catégorie 1, C33V66.*

La méthode GP donne de meilleurs résultats pour le score *FF*, surtout avec de petits échantillons (voir tableau 6). Quand l'échantillon de calage décroît (17 années au lieu de 25), la méthode GEV perd davantage de précision pour l'estimation des valeurs extrêmes que la méthode GP. La figure 11 confirme aussi que les estimations des quantiles avec GEV sont sous-estimées : la courbe *FF* pour GEV est toujours sous la bissectrice. Inversement, la sous-estimation des quantiles calculés avec le modèle GP est moins évidente avec le critère *FF* qu'avec les scores N_{10} et N_{100}. Comme le critère *FF* est calculé sur un échantillon de valeurs maximales, de 33 années ou plus, nous pouvons en déduire que le modèle GP sous-estime moins les quantiles lorsque la durée de retour augmente, ce que le modèle GEV ne fait pas. Donc le modèle GP donne des résultats plus fiables pour les grandes durées de retour que le modèle GEV, surtout pour de petits échantillons.

Tableau 6 : *comparaison entre les estimations GEV et GP basée sur le critère FF. Échantillonnage de catégorie 1.*

	FF	
	25 ans – 25 ans	17 ans – 33 ans
GEV	0,88	0,72
GP	0,91	0,84

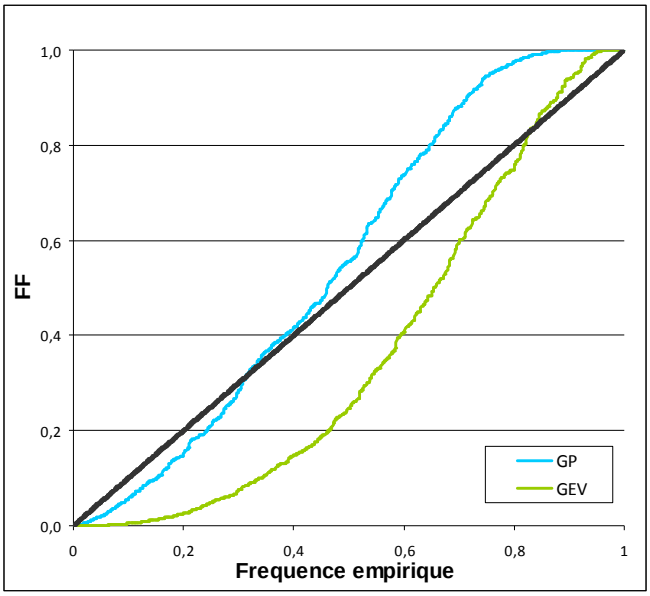


Figure 11. *Comparaison entre les estimations GEV et GP basée sur le critère FF. Échantillonnage de catégorie 1, C33V66.*

4.2.2 Robustesse

Nous n'utiliserons pas de sous-échantillons ayant moins de 20 années de données car il y a trop de cas où la méthode GEV ne peut fournir de résultats avec aussi peu d'observations. Ribereau *et al.* (2008) avaient indiqué cette limite : la méthode GEV ne fournit pas toujours une estimation correcte des quantiles, surtout pour de petites tailles d'échantillons ou des distributions à queues lourdes.

Si le $SPAN_{10}$ ne montre pas une réelle différence de robustesse entre les deux estimateurs, les quantiles de durée de retour 100 ans estimés avec l'estimateur GP sont plus robustes que ceux estimés avec l'estimateur GEV (voir tableau 7). La différence de robustesse des estimations du quantile de durée de retour 100 ans entre les méthodes GEV et GP est plus visible pour les stations ayant des estimations de quantiles élevées ou basses. Par conséquent, nous classons les stations suivant la moyenne des quatre estimations calculées (méthodes GEV et GP pour l'échantillon 1 et l'échantillon 2). Nous obtenons que pour les stations ayant un quantile de durée de retour 100 ans inférieur à 108 mm (premier quartile) ou supérieur à 180 mm (dernier quartile) avec l'échantillonnage de 50 ans de données ou plus, les $SPAN_{100}$ sont bien meilleurs avec la méthode GP que pour la méthode GEV (0,76 avec la méthode GP contre 0,69 avec la méthode GEV) tandis qu'ils sont semblables pour les stations ayant un quantile de durée de retour 100 ans entre 108 mm et 180 mm (0,74 avec la méthode GP et 0,73 avec la méthode GEV). Nous notons d'autre part que les $SPAN_{100,i}$ décroissent avec les méthodes GP et GEV : plus les estimations sont élevées, moins elles sont robustes.

Tableau 7. *Comparaison entre les estimations GEV et GP basée sur les critères SPANT et COVERT. Échantillonnage de catégorie 2.*

$SPAN_T$	15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
GEV	0,85	0,64	0,89	0,71
GP	0,84	0,67	0,88	0,75

$COVER_T$	15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
GEV	0,42	0,30	0,50	0,36
GP	0,56	0,54	0,58	0,59

La méthode GP donne des estimations plus robustes pour la durée de retour 100 ans que la méthode GEV : les $SPAN_{100}$ sont plus grands, surtout pour l'évaluation de quantiles faibles ou élevés. De plus, le critère $COVER_T$ montre que l'estimation de la variance est bien meilleure avec la méthode GP qu'avec la méthode GEV pour les deux durées de retour 10 et 100 ans.

4.2.3 Robustesse relativement à la valeur maximale

Le critère $SPAN_T$ montre que les estimations GEV sont aussi robustes que les estimations GP pour la valeur maximale de chaque station (voir tableau 8). Dans les deux cas, les estimations sont très robustes, particulièrement pour la durée de retour 10 ans, vis-à-vis de la valeur maximale : les $SPAN_{10}$ sont supérieurs à 0,94. Comme les scores $SPAN_T$ des deux modèles sont proches, nous pouvons interpréter les différences des $COVER_T$: l'estimation de la variance est largement meilleure avec les estimations GP qu'avec les estimations GEV, surtout pour la durée de retour 100 ans.

Tableau 8. *Comparaison entre les estimations GEV et GP basée sur les critères SPANT et COVERT. Échantillonnage de catégorie 3.*

$SPAN_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
GEV	0,95	0,87	0,95	0,88	0,96	0,91
GP	0,94	0,88	0,95	0,88	0,96	0,91

$COVER_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
GEV	0,67	0,57	0,69	0,59	0,70	0,61
GP	0,77	0,76	0,79	0,78	0,80	0,78

4.2.4 Conclusion sur la comparaison entre les modèles GEV et GP

Nous avons observé certaines caractéristiques constantes : les estimations GEV et GP sont plus robustes avec 25 années de données qu'avec 15 ans ; pour la durée de retour 10 ans que pour la durée de retour 100 ans ; pour des quantiles faibles que pour des quantiles élevés. Si l'on dispose de davantage de données pour le calcul des estimations, on s'attend à ce qu'elles soient meilleures. De même, on s'attend à ce qu'avec un quantile plus élevé les estimations soient moins bonnes.

Les résultats sont résumés sur les diagrammes en étoile (figure 12 pour la durée de retour 10 ans et figure 13 pour la durée de retour 100 ans). La méthode GP donne de meilleurs résultats que la méthode GEV. En fait, les critères $COVER_T$ et *FF* indiquent qu'il existe des différences importantes. Le score $COVER_T$ montre que les estimations des intervalles de confiance sont meilleures avec la méthode GP.

De plus, le score *FF* montre que la méthode GP donne une meilleure estimation de la distribution de la valeur maximale. En particulier, comme la méthode GP utilise davantage d'observations que la méthode GEV, ses estimations sont plus robustes. Par exemple, le modèle GEV doit estimer les paramètres avec seulement 17 observations dans les échantillonnages C33V66 utilisés pour calculer les scores N_{10} et *FF*. Ainsi, même si chaque observation est le maximum de son année et donne donc beaucoup d'information sur les pluies extrêmes, la précision est moins bonne qu'avec la méthode GP qui utilise quatre observations par an en moyenne. La méthode GP donne donc de meilleurs résultats particulièrement pour les grandes durées de retour et pour de petits échantillons.

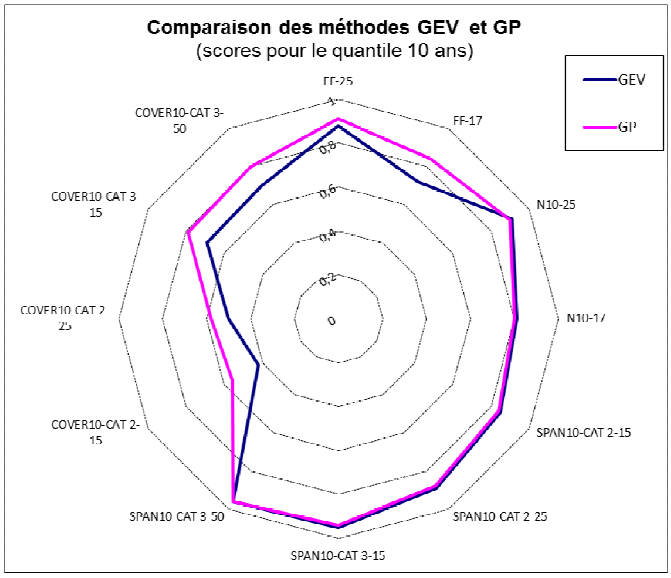


Figure 12. *Résultats des différents critères sur les estimations GEV et GP. Durée de retour 10 ans.*

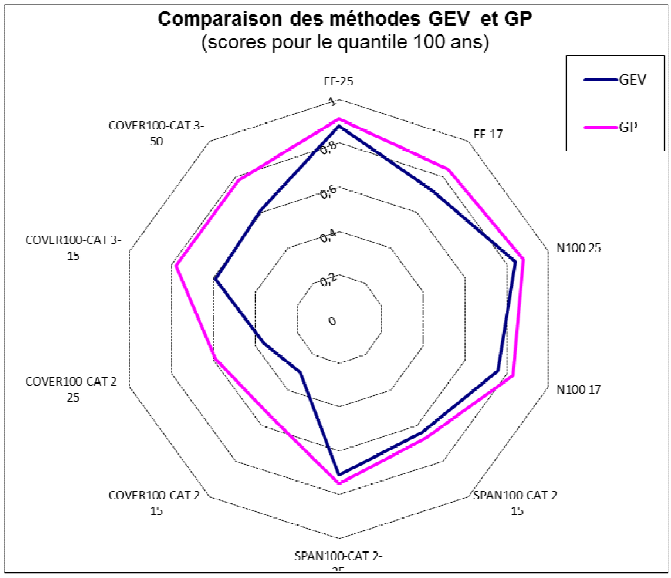


Figure 13. *Résultats des différents critères sur les estimations GEV et GP. Durée de retour 100 ans*

4.3 Comparaison des méthodes GP et Exponentielle

On cherche à analyser à présent l'apport d'une paramétrisation des lois sur les valeurs extrêmes à trois paramètres au lieu de deux. Pour cela, on compare les méthodes GP et Exponentielle (EXPO) en utilisant l'estimation du maximum de vraisemblance et les quatre scores précédents représentant la justesse et la robustesse : *FF*, N_T , $SPAN_T$ et $COVER_T$.

En premier lieu, on met en évidence que les estimations GP et EXPO sont assez différentes (plus qu'entre GP et GEV) au-delà du quantile 10 ans. Ainsi, les estimations des quantiles de durée de retour 10 ans calculées avec l'échantillon complet des stations ayant 50 années de données ou plus ont un coefficient de corrélation de 0,98 mais seulement de 0,86 pour les durées de retour 100 ans (voir figure 14). Les estimations EXPO sont inférieures aux estimations GP dans 75% des cas à 10 ans, comme à 100 ans. La moyenne du quantile de durée de retour 10 ans est de 89,0 mm avec les estimations EXPO contre 93,7 mm avec les estimations GP et celle du quantile de durée de retour 100 ans est de 123,3 mm avec les estimations EXPO contre 146,7 mm avec les estimations GP.

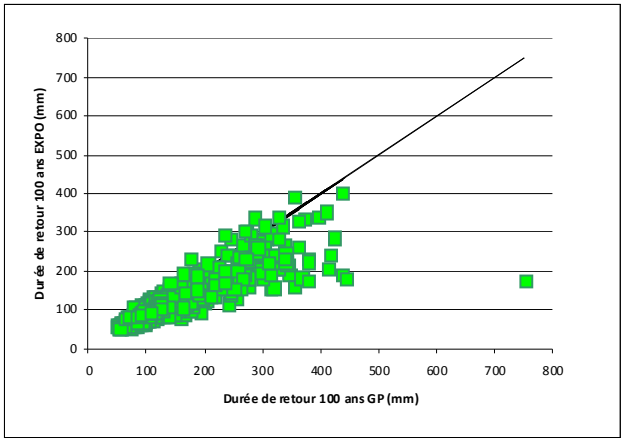


Figure 14. *Pluie centennale estimée avec les méthodes GP et EXPO. Échantillon complet, 50 années et plus.*

4.3.1 Justesse

Les critères N_{10} et N_{100} montrent des différences sensibles entre les modèles GP et EXPO (voir tableau 9). Sur les graphiques (voir figures 15 et 16), la courbe N_{10} de EXPO reste toujours en dessous de celle de GP et souvent sous la bissectrice : les estimations du quantile de durée de retour 10 ans sont sous-estimées pour les deux modèles GP et EXPO, mais beaucoup plus par EXPO. La sous-estimation s'accentue pour le quantile de durée de retour 100 ans pour les deux méthodes.

Tableau 9. *Comparaison entre les estimations GP et EXPO basée sur les critères N10 et N100. Échantillonnage de catégorie 1.*

	25 ans – 25 ans		17 ans – 33 ans	
	N_{10}	N_{100}	N_{10}	N_{100}
GP	0,90	0,88	0,80	0,83
EXPO	0,76	0,72	0,68	0,64

La méthode GP donne des résultats nettement meilleurs qu'EXPO pour le score *FF*, tant avec les échantillons de 25 ans que de 17 ans. La figure 17 confirme aussi que les estimations de la probabilité au non-dépassement des valeurs maximales sont surestimées systématiquement, avec un degré moindre pour GP : la courbe *FF* est toujours au-dessus de la bissectrice.

Tableau 10. *Comparaison entre les estimations GP et EXPO basée sur le critère FF. Échantillonnage de catégorie 1.*

	FF	
	25 ans – 25 ans	17 ans – 33 ans
GP	0,91	0,84
EXPO	0,69	0,65

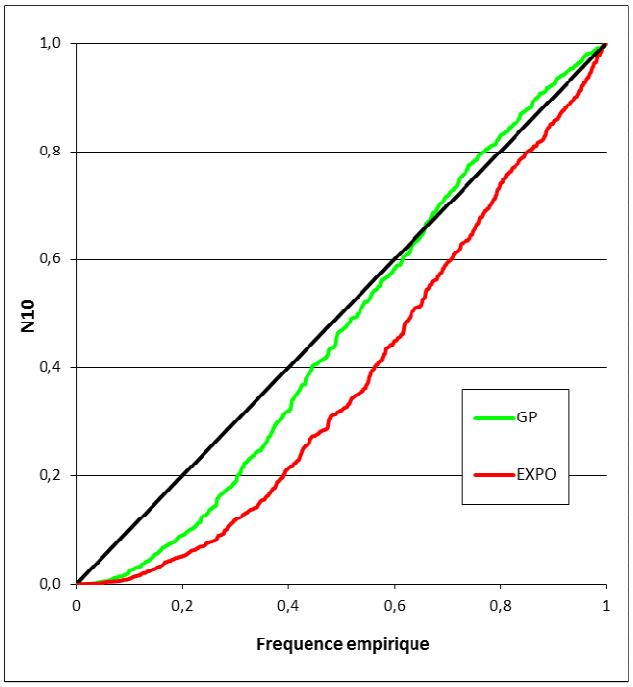


Figure 15. *Comparaison entre les estimations GP et EXPO basée sur le critère N_{10}. Échantillonnage de catégorie 1, C50V50.*

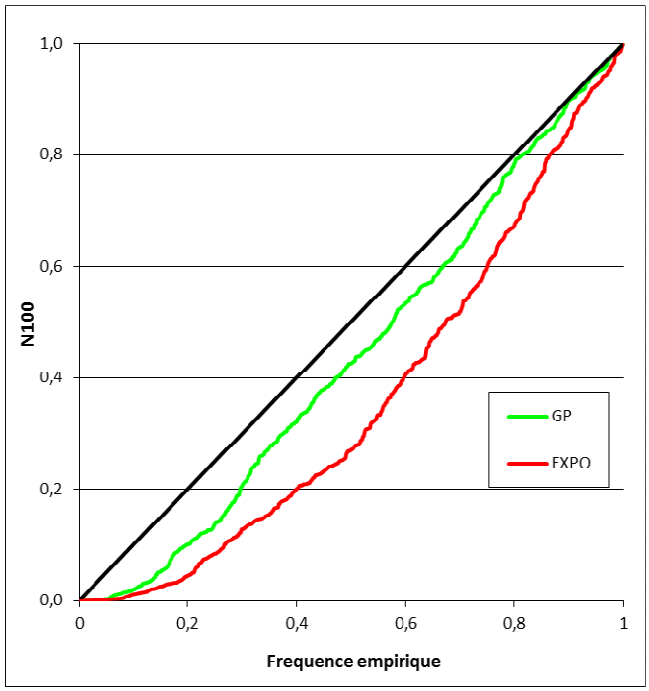


Figure 16. *Comparaison entre les estimations GP et EXPO basée sur le critère N_{100}. Échantillonnage de catégorie 1, C50V50.*

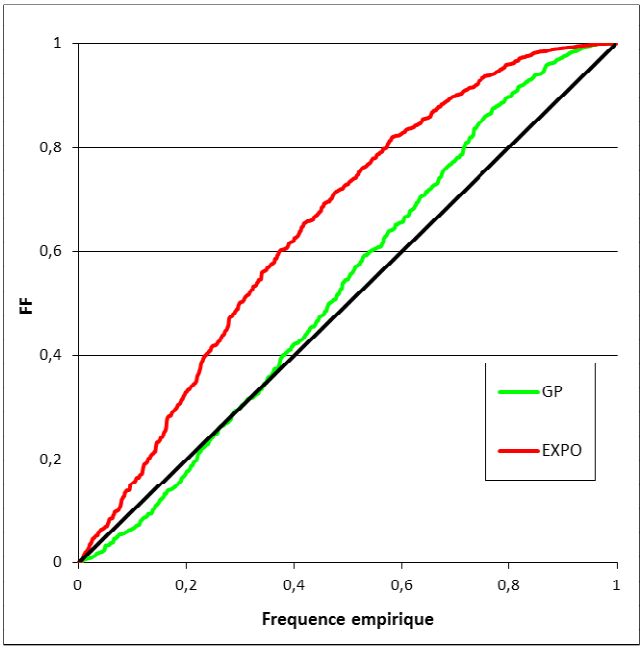


Figure 17. *Comparaison entre les estimations GP et EXPO basée sur le critère FF. Échantillonnage de catégorie 1, C50V50.*

4.3.2 Robustesse

Si le $SPAN_{10}$ montre déjà une différence importante de robustesse entre les deux méthodes, les écarts s'accentuent avec les quantiles de durée de retour 100 ans estimés. Dans tous les cas les quantiles estimées par la méthode EXPO sont beaucoup plus robustes que ceux estimés avec GP (voir tableau 11). La différence de robustesse des estimations du quantile de durée de retour 100 ans diminue légèrement lorsque la longueur des séries augmente, 25 ans au lieu de 10 ans.

Tableau 11. *Comparaison entre les estimations GP et EXPO basée sur le critère $SPAN_T$. Échantillonnage de catégorie 2.*

$SPAN_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
GP	0,80	0,59	0,84	0,67	0,88	0,75
EXPO	0,86	0,85	0,90	0,88	0,92	0,90

4.3.3 Robustesse relativement à la valeur maximale

Le critère $SPAN_T$ montre que les estimations EXPO restent toujours plus robustes que les estimations GP pour la valeur maximale de chaque station (voir tableau 12). La méthode GP gagne en robustesse et se rapproche des performances de la méthode EXPO pour les faibles durées de retour (10 ans) et les échantillons longs (25 ans)

Tableau 12. *Comparaison entre les estimations GP et EXPO basée sur le critère $SPAN_T$. Échantillonnage de catégorie 3.*

$SPAN_T$	20 ans		30 ans		50 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
GP	0,94	0,88	0,95	0,90	0,96	0,91
EXPO	0,97	0,97	0,98	0,98	0,98	0,98

4.3.4 Conclusion sur la comparaison entre les modèles GP et EXPO

Nous avons observé des caractéristiques opposées entre les estimations GP, beaucoup plus performantes en justesse et Expo beaucoup plus robustes même avec des échantillons réduits. Un des problèmes de la représentation par la loi exponentielle est le caractère hyper-exponentiel de la majorité des ajustements traités dans nos jeux de données. Ainsi, la moyenne des paramètres de forme obtenue avec la loi GP sur les 693 longues séries est de +0,09. Les résultats sont résumés sur les diagrammes en étoile (figure 18 pour la durée de retour 10 ans et figure 19 pour la durée de retour 100 ans).

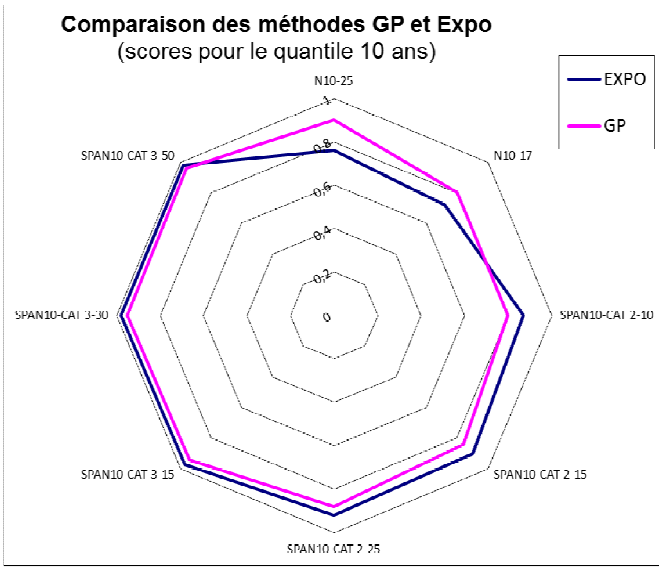


Figure 18. *Résultats des différents critères sur les estimations GEV et EXPO. Durée de retour 10 ans*

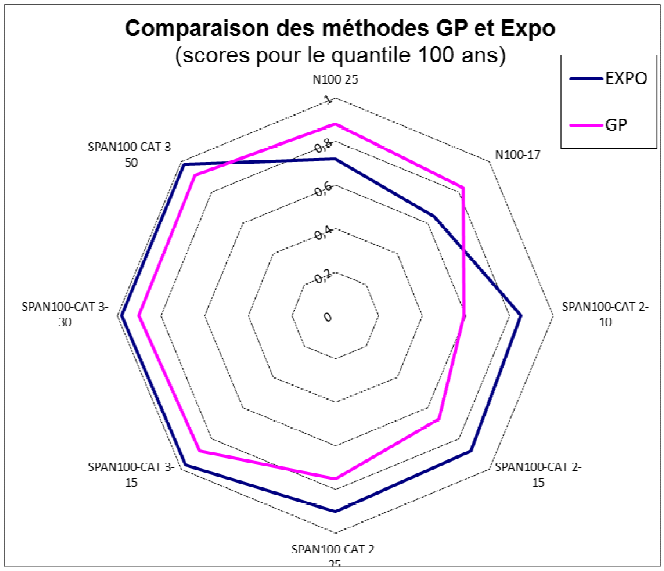


Figure 19. *Résultats des différents critères sur les estimations GEV et EXPO. Durée de retour 100 ans.*

4.4 Comparaison des méthodes d'estimation des paramètres de la loi GP

Comme nous l'avons vu ci-dessus, la méthode GP a été considérée comme légèrement préférable à la méthode GEV ; nous allons donc déterminer maintenant pour la méthode GP quel estimateur est le meilleur. La méthode GP peut être utilisée avec différentes méthodes d'estimation. Nous allons en tester trois : la méthode du maximum de vraisemblance (ML), la méthode des moments (MM) et la méthode des moments pondérés (PWM).

Les valeurs du quantile de durée de retour 100 ans données par les trois estimations sont très corrélées (voir figure 20). Les coefficients de corrélation sont plus grands que 0,99. Cependant, les estimations ML sont légèrement inférieures aux estimations données par les estimations MM et PWM. Par exemple, avec l'échantillon complet des stations ayant 50 années de données ou plus, la moyenne des estimations MM est de 139,3 mm tandis que la moyenne des estimations ML est de 144,3 mm et la moyenne des estimations PWM est de 142,9 mm.

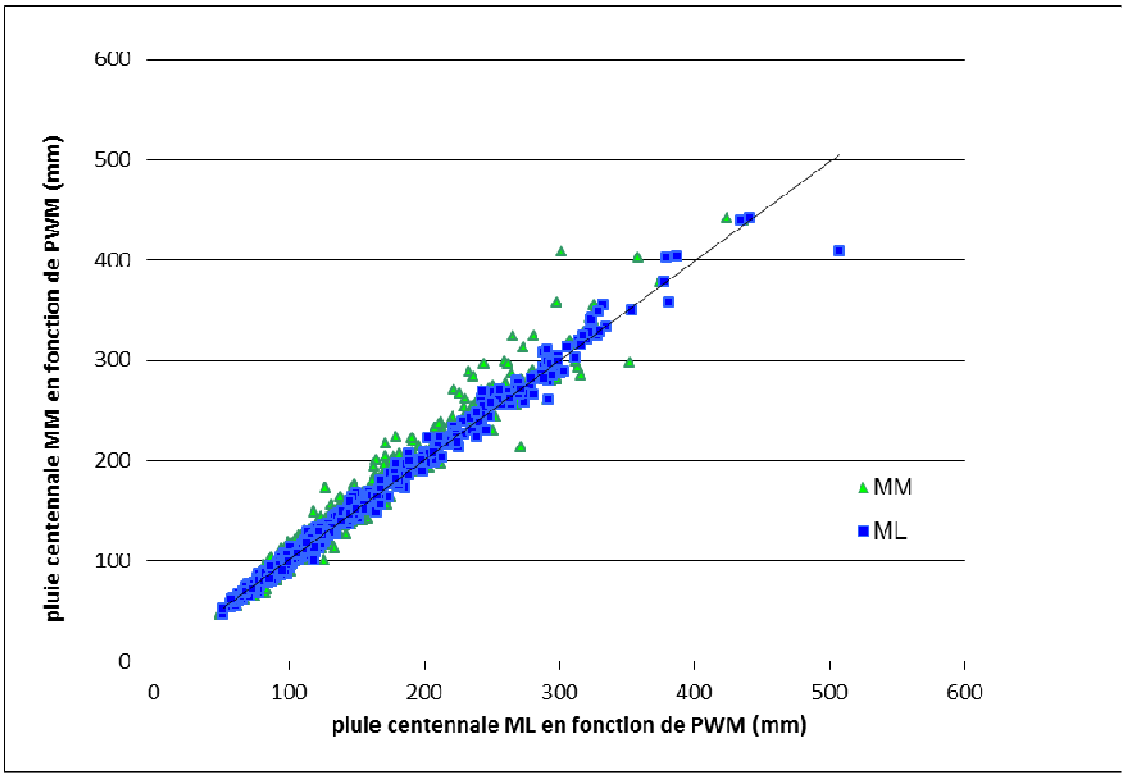


Figure 20. *Quantiles de durée de retour 100 ans estimés avec les méthodes MM, ML et PWM. Échantillon complet, 50 années et plus. Stations possédant des estimations MM, PWM et ML*

4.4.1 Justesse

Les scores N_{10}, N_{100} et *FF* montrent que l'estimateur PWM est légèrement plus juste que les estimateurs MM et ML (voir Tableau 13).

Tableau 13. ***Comparaison entre différents estimateurs du modèle GP basée sur les critères N_{10}, N_{100}. et FF. Échantillonnage de catégorie 1.***

	25 ans – 25 ans			17 ans – 33 ans		
	N_{10}	N_{100}	*FF*	N_{10}	N_{100}	*FF*
ML	0,90	0,88	0,91	0,80	0,83	0,84
MM	0,89	0,90	0,88	0,81	0,83	0,83
PWM	0,91	0,95	0,93	0,84	0,88	0,88

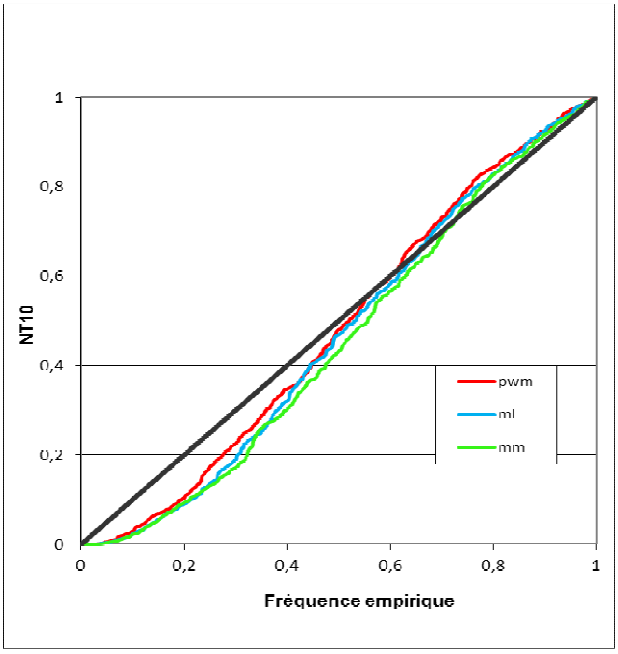


Figure 21. *Comparaison entre différents estimateurs du modèle GP basée sur le critère N_{10}. Échantillonnage de catégorie 1, C50V50*

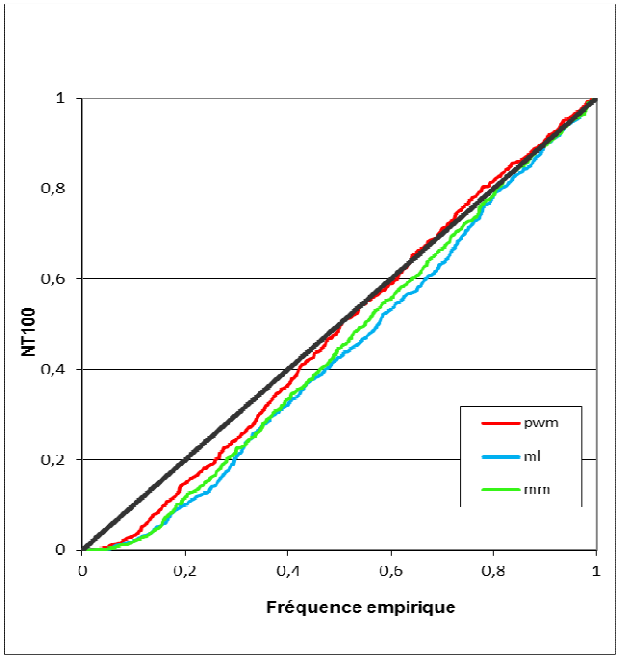


Figure 22. *Comparaison entre différents estimateurs du modèle GP basée sur le critère N_{100}. Échantillonnage de catégorie 1, C50V50*

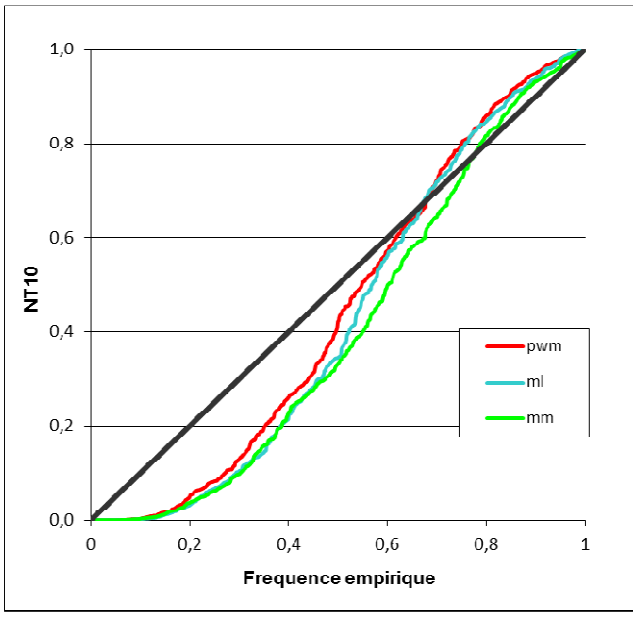


Figure 23. *Comparaison entre différents estimateurs du modèle GP basée sur le critère N_{10}. Échantillonnage de catégorie 1, C33V66*

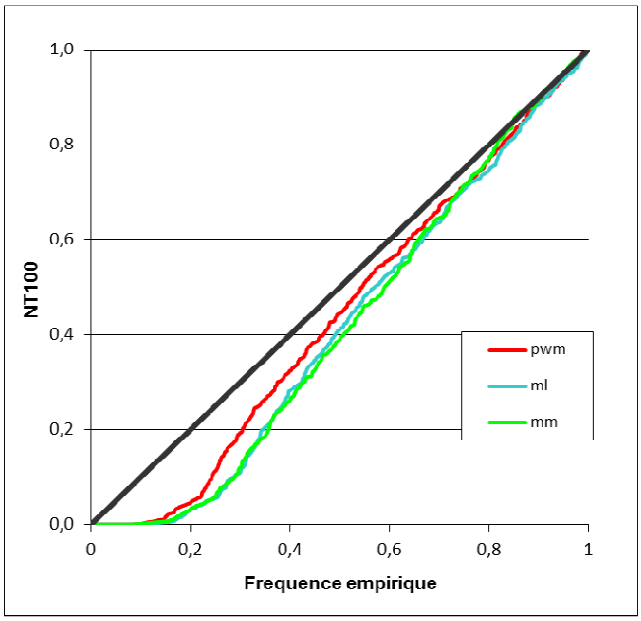


Figure 24. *Comparaison entre différents estimateurs du modèle GP basée sur le critère N_{100}. Échantillonnage de catégorie 1, C33V66*

Mais l'information principale vient de l'analyse des graphes. Les figures 21 à 24 montrent que les courbes N_{10} et N_{100} sont au-dessous de la bissectrice pour les trois estimateurs : les quantiles sont sous-estimés. On peut évaluer la sous-estimation en multipliant par un certain nombre le quantile q_T calculé avec l'échantillon C1, qui est utilisé pour évaluer la durée de retour, et en recalculant les scores. Il apparaît que les quantiles évalués avec l'estimateur PWM sont sous-estimés d'environ 5 pour cent : les

meilleurs scores sont obtenus quand le quantile estimé est multiplié par 1,05. Comme les courbes sont toujours du même côté de la bissectrice (au-dessous), les modèles ne sont pas sur-paramétrés.

La figure *25* montre que l'estimateur PWM donne une courbe FF plus proche de la bissectrice que les estimateurs MM et ML. En particulier, la méthode MM sous-estime la distribution théorique.

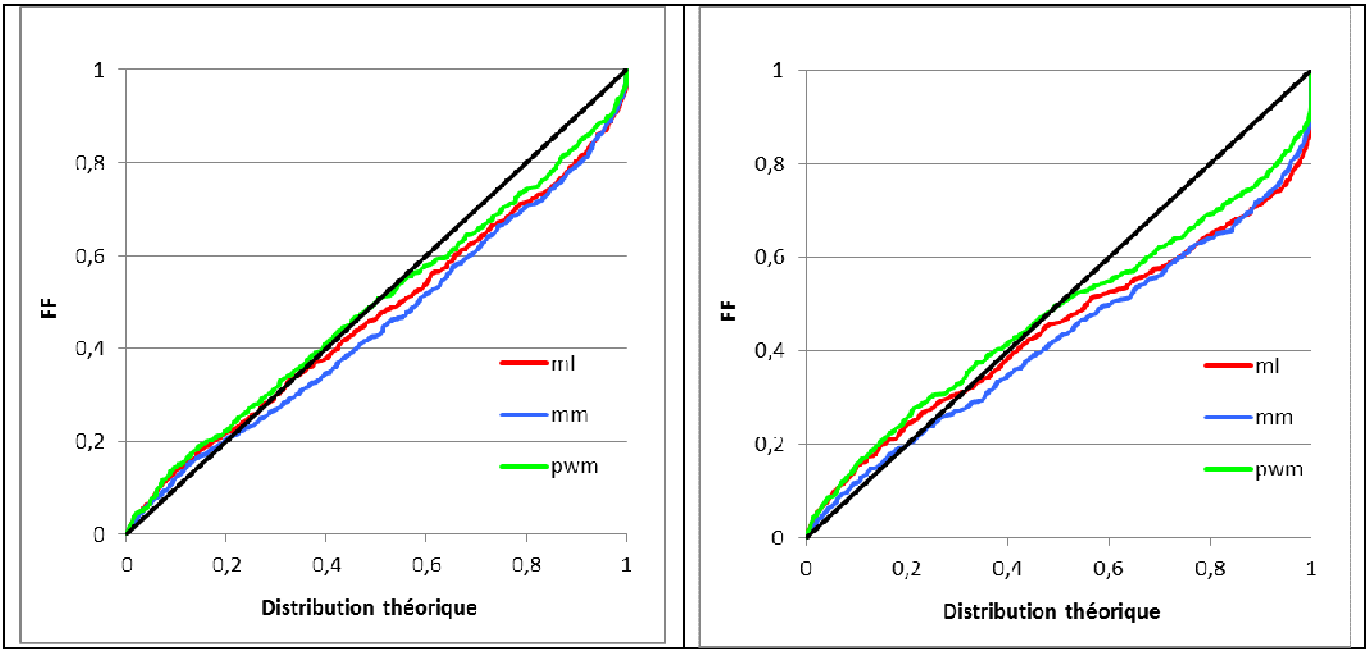


Figure 25. *Comparaison entre différents estimateurs du modèle GP basée sur le critère FF. Échantillonnage de catégorie 1, C50V50 à gauche et C33V66 à droite*

4.4.2 Robustesse

Nous pouvons observer des tendances communes. Les trois estimateurs donnent des estimations plus robustes pour les longues séries (25 ans) que pour les courtes (10 ans) et pour la durée de retour 10 ans que pour la durée de retour 100 ans (voir Tableau 14). Inversement, si l'estimation de la variance est moins précise quand la taille de l'échantillon diminue, il n'y a qu'une très légère différence entre les durées de retour 10 et 100 ans.

Tableau 14. *Comparaison entre différents estimateurs du modèle GP basée sur les critères $SPAN_T$ et $COVER_T$. Échantillonnage de catégorie 2.*

$SPAN_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
ML	0,80	0,59	0,84	0,67	0,88	0,75
MM	0,83	0,71	0,87	0,75	0,89	0,80
PWM	0,82	0,64	0,86	0,70	0,88	0,76

$COVER_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
ML	0,52	0,48	0,56	0,54	0,58	0,59
MM	0,49	0,48	0,53	0,53	0,57	0,58
PWM	0,58	0,56	0,61	0,60	0,61	0,63

Les méthodes ne sont pas toujours capables de donner une estimation de la variance des estimations, surtout avec de petits échantillons (voir Tableau 15). En effet dans le cas de la loi de Fréchet, lorsque le paramètre de forme prend de grandes valeurs, certains moments théoriques de la distribution n'existent pas, ce qui pénalise les méthodes fondées sur une estimation des moments. Dans l'échantillonnage 10 ans - 10 ans, la méthode MM ne donne pas ces estimations dans 154 cas sur 1286 et la méthode PWM dans 22. Pour pouvoir comparer les scores $COVER_T$ des trois estimateurs, nous ne comparerons les stations que lorsque les trois méthodes peuvent fournir une estimation. Ceci explique pourquoi nous n'avons pas exactement les mêmes scores dans la partie suivante de l'étude.

Tableau 15. *Comparaison du nombre de fois où le modèle GP ne donne pas d'estimation de la variance des estimations. Échantillonnage de catégorie 2.*

	10 ans – 10 ans	15 ans – 15 ans	25 ans – 25 ans
Nombre de stations	1286	1016	671
ML	0	0	0
MM	154	98	39
PWM	22	6	0

L'estimateur MM (méthode des moments) donne des scores $SPAN_T$ légèrement meilleurs pour la durée de retour 10 ans, et davantage pour une durée de retour 100 ans, en particulier pour des séries courtes. Mais les scores $COVER_T$ montrent que la méthode PWM donne des estimations légèrement meilleures des intervalles de confiance.

4.4.3 Robustesse relativement à la valeur maximale

Les trois méthodes donnent des résultats équivalents pour les scores $SPAN_{10}$ lorsque l'année contenant la valeur maximale est enlevée (voir tableau 16). Il y a cependant une légère différence, surtout entre l'estimateur MM et les estimateurs ML ou PWM. Dans presque 3% des cas, le $SPAN_{10,i}$ est plus grand que 0,10 avec l'estimateur MM, contre environ 1% des cas avec les estimateurs ML et PWM. Ainsi les estimations des quantiles sont plus dépendantes de la plus grande valeur avec l'estimateur MM. De plus, l'estimateur PWM donne des estimations plus robustes que l'estimateur ML pour la durée de retour 100 ans. Nous concluons aussi que la méthode PWM fournit de meilleures estimations de la variance des estimations. Les scores $COVER_{10}$ sont meilleurs de 4 points avec la méthode PWM qu'avec la méthode MM et les scores $COVER_{100}$ sont meilleurs de 7 points. L'estimateur PWM semble être l'estimateur le plus robuste relativement à la valeur maximale.

Tableau 16. *Comparaison entre différents estimateurs du modèle GP basée sur les critères $SPAN_T$ et $COVER_T$. Échantillonnage de catégorie 3.*

$SPAN_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
ML	0,94	0,88	0,95	0,88	0,96	0,91
MM	0,94	0,89	0,95	0,90	0,96	0,91
PWM	0,95	0,90	0,96	0,92	0,96	0,93

$COVER_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
ML	0,77	0,76	0,79	0,78	0,80	0,78
MM	0,75	0,74	0,79	0,78	0,80	0,78
PWM	0,82	0,83	0,84	0,84	0,84	0,85

4.4.4 La distribution prédictive

La distribution prédictive donne les mêmes résultats que l'estimation centrale, en termes de comparaison des méthodes d'estimation. L'estimateur PWM est plus fiable que l'estimateur MM, surtout quand la taille de l'échantillon diminue (voir Tableau 17). Mais l'estimateur MM fournit une estimation plus robuste surtout pour la durée de retour 100 ans (voir Tableau 18).

Tableau 17. *Comparaison entre différents estimateurs du modèle GP basée sur les critères N_{10} et FF. Distribution prédictive. Échantillonnage de catégorie 1.*

	25 ans – 25 ans		17 ans – 33 ans	
	N_{10}	*FF*	N_{10}	*FF*
ML	0,76	0,92	0,72	0,78
MM	0,74	0,90	0,66	0,74
PWM	0,78	0,94	0,74	0,82

Tableau 18. *Comparaison entre différents estimateurs du modèle GP basée sur le critère $SPAN_T$. Distribution prédictive. Échantillonnage de catégorie 2.*

$SPAN_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans –25 ans	
	10 ans	100 ans	10 ans	100 ans	10 ans	100 ans
ML	0,77	0,54	0,81	0,62	0,85	0,69
MM	0,80	0,61	0,82	0,65	0,85	0,71
PWM	0,78	0,56	0,82	0,63	0,85	0,69

4.4.5 Conclusion sur les différents estimateurs du modèle GP

Les résultats sont résumés sur des diagrammes en étoile (*Figure 25* pour la durée de retour 10 ans et *figure 26* pour la durée de retour 100 ans).

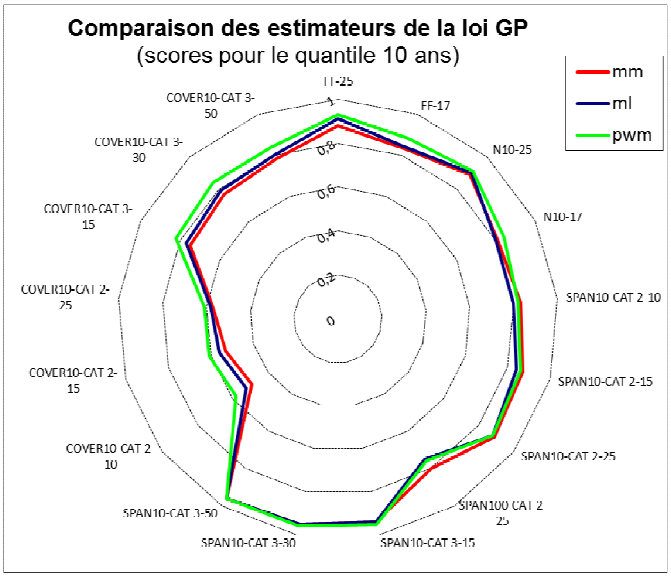


Figure 25 *Résultats des différents estimateurs du modèle GP. Durée de retour 10 ans*

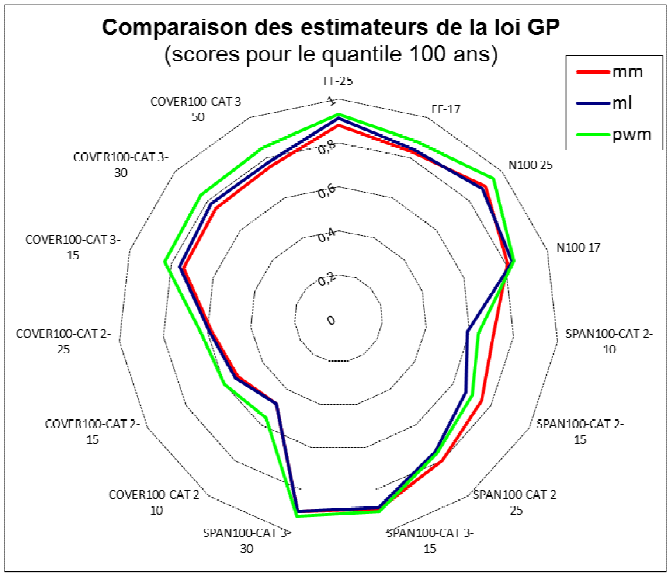


Figure 26. *Résultats des différents estimateurs du modèle GP. Durée de retour 100 ans*

Les trois estimateurs donnent des estimations proches des quantiles quand il y a beaucoup d'observations. La figure 20 a d'ailleurs montré une corrélation extrême entre les estimations fournies par les trois estimateurs. Cependant les estimateurs MM et PWM sont légèrement plus robustes que l'estimateur ML, ils ont besoin de moins d'observations et sont plus performants pour les grandes durées de retour. D'après le critère $SPAN_T$ l'estimateur MM donne des estimations plus robustes et

PWM est plus robuste relativement à la valeur maximale et plus fiable. L'application des scores montre un léger avantage de l'estimateur PWM sur l'estimateur MM. Mais la principale limite de l'estimateur MM est le nombre de stations pour lesquelles il n'est pas capable de calculer une estimation de la variance, et donc de l'intervalle de confiance du quantile quand la taille de l'échantillon diminue. Par conséquent nous avons décidé de garder l'estimateur PWM dans la partie suivante.

4.5 Comparaison des méthodes GP, SHYPRE, MEWP

4.5.1 Résultats sur l'ensemble de la zone d'étude

Les analyses des sections 4.2, 4.3 et 4.4 ont montré que le modèle GP utilisant l'estimateur PWM était la meilleure méthode paramétrique classique. Cette dernière section va consister à comparer cette méthode à deux méthodes plus originales: la méthode SHYPRE et la méthode MEWP.

Les valeurs des quantiles obtenus par les trois méthodes sont relativement bien corrélées (supérieur à 0,90) avec une liaison plus forte entre SHYPRE et MEWP et plus faible de GP avec les autres méthodes (0,90 et 0,92). En moyenne, les quantiles estimés par la méthode MEWP s'avèrent inférieurs à ceux issus des méthodes GP et SHYPRE, dont les moyennes sont très proches. Ces différences persistent sur les quantiles supérieurs et notamment pour la durée de retour centennale (voir figure 27) : pour la durée de retour 100 ans, la moyenne des estimations MEWP est de 136 mm tandis que celle de SHYPRE est de 146 mm et celle de GP est de 147 mm.

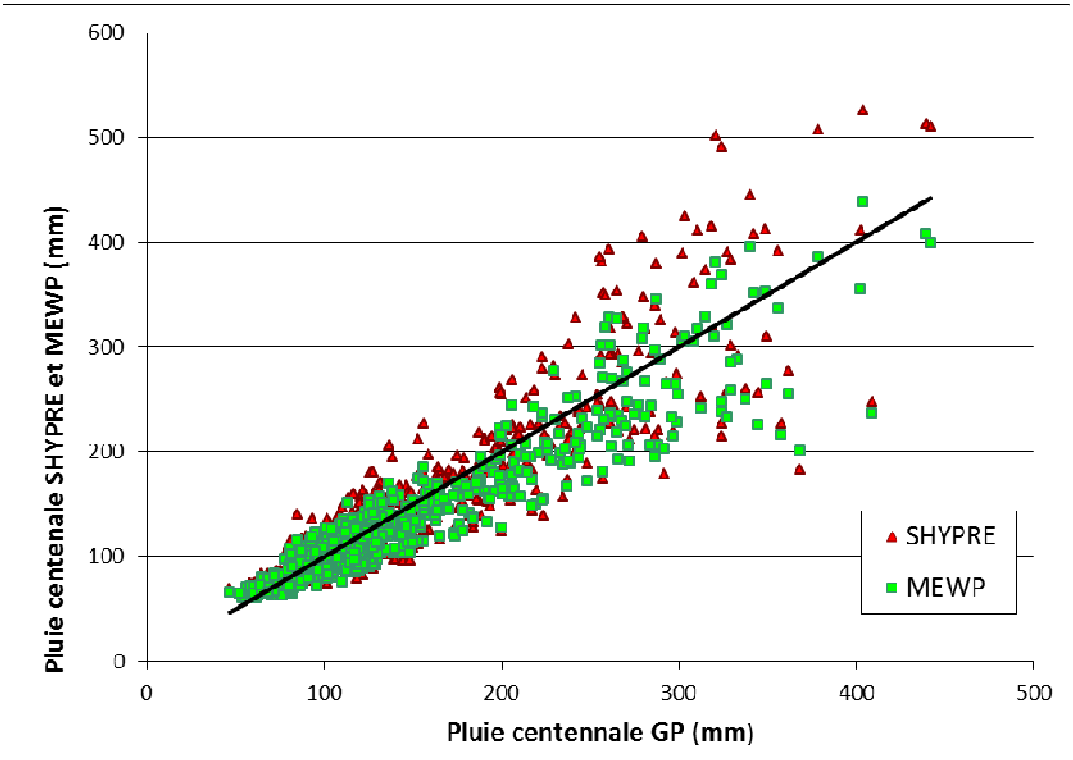


Figure 27. *Estimations des durées de retour 100 ans avec les méthodes GP, MEWP et SHYPRE : échantillon complet (693 séries de plus de 50 ans).*

La figure 27 montre également que les écarts entre les trois méthodes MEWP, SHYPRE et GP sont maximums pour les valeurs de précipitation les plus fortes, et que la méthode SHYPRE donne alors des estimations en moyenne plus élevées et plus dispersées. Par exemple, quand les trois méthodes donnent pour le quantile de durée de retour 100 ans une valeur supérieure à 200 mm (99 stations), la moyenne des estimations de SHYPRE est de 305 mm contre 286 mm et 266 mm avec les modèles GP et MEWP.

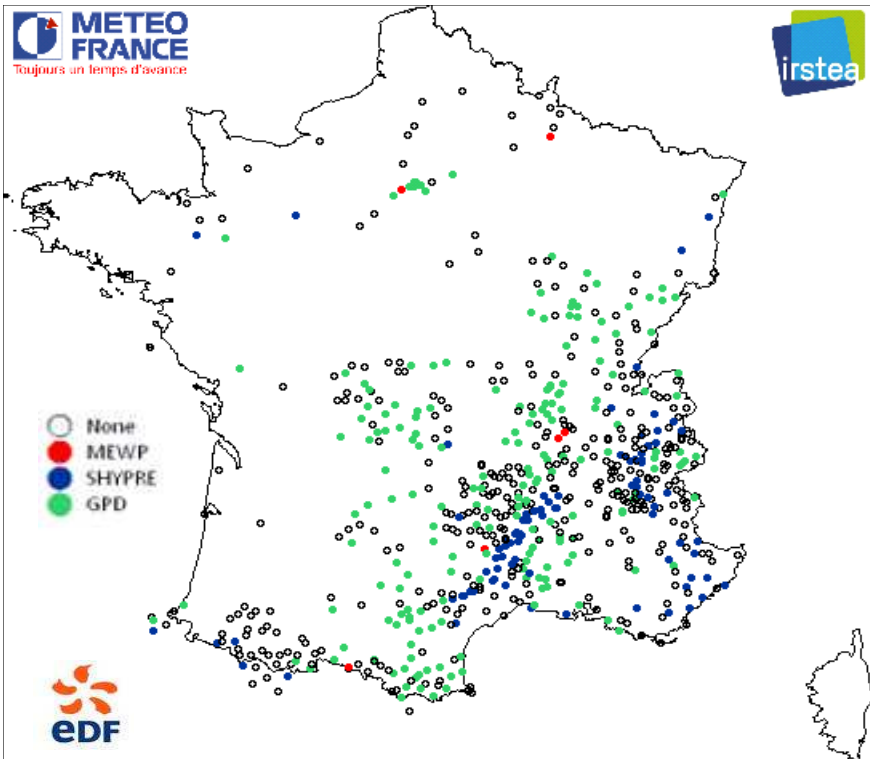


Figure 28. *Carte des différences entre les estimations de pluie de durée de retour centennale sur la France : les points rouges indiquent une estimation plus forte de MEWP (+10% par rapport aux deux autres méthodes), les points bleus de SHYPRE, les points verts de GP. Les points blancs représentent les stations où les estimations sont proches (différence inférieure à 10% entre au moins deux des estimations). Échantillon complet (693 séries de plus de 50 ans).*

Sur la figure 28 on a pointé les stations pour lesquelles une des trois méthodes précédentes est significativement supérieure aux deux autres (écart supérieur à 10%). On voit que les estimations GP sont plus souvent supérieures aux deux autres (213 stations) que celles de SHYPRE (95 stations) et que celles de MEWP ne le sont pratiquement jamais (12 stations). Mais on peut aussi mettre en évidence que les différences entre les méthodes présentent certaines structures régionales. En particulier, les estimations supérieures pour la méthode SHYPRE (qui donne les estimations les plus élevées pour les valeurs de précipitation les plus fortes) se retrouvent préférentiellement sur le relief, notamment sur les Cévennes et les Alpes.

4.5.1.1 Justesse

Les scores caractérisant la justesse (N_{10}, N_{100} et FF) sont assez proches entre eux (voir Tableau 19 et figures 29 à 33). Il y a une sous estimation par MEWP, ainsi qu'une tendance globale des méthodes à sous estimer les valeurs maximales (FF) et une légère dégradation de N_{10} et N_{100} en C33V66 pour GP. Globalement SHYPRE obtient des résultats légèrement meilleurs à MEWP et GP.

Tableau 19. *Comparaison entre les estimations SHYPRE, MEWP, GP basée sur les critères N_{10} , N_{100}. et FF. Échantillonnage de catégorie 1.*

	25 ans – 25 ans			17 ans – 33 ans		
	N_{10}	N_{100}	FF	N_{10}	N_{100}	FF
SHYPRE	0,95	0,96	0,95	0,91	0,95	0,95
MEWP	0,85	0,92	0,92	0,84	0,92	0,93
GP	0,91	0,95	0,93	0,84	0,88	0,88

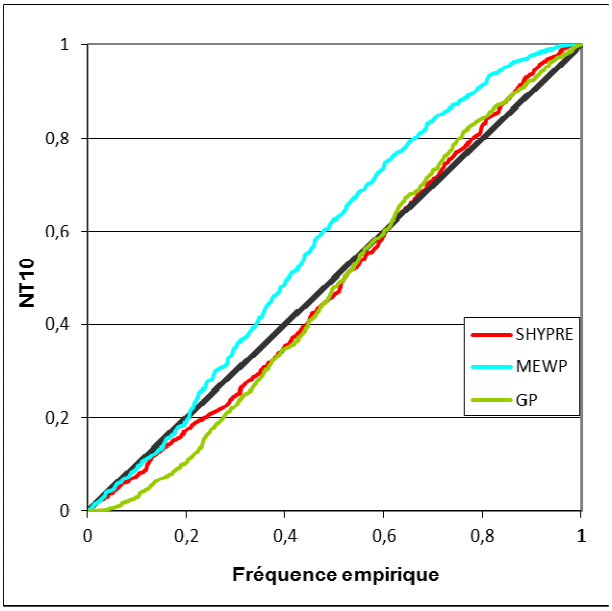


Figure 29. *Comparaison entre les estimations SHYPRE, MEWP et GP basée sur le critère N_{10}. Échantillonnage de catégorie 1, C50V50*

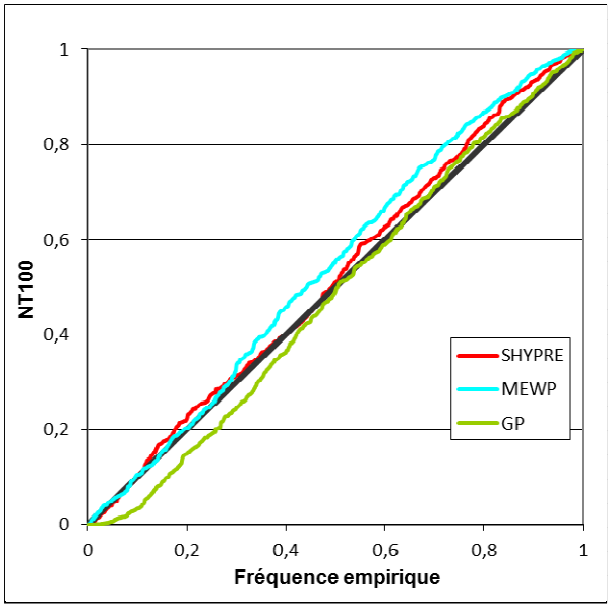


Figure 30. *Comparaison entre les estimations SHYPRE, MEWP et GP basée sur le critère N_{100}. Échantillonnage de catégorie 1, C50V50*

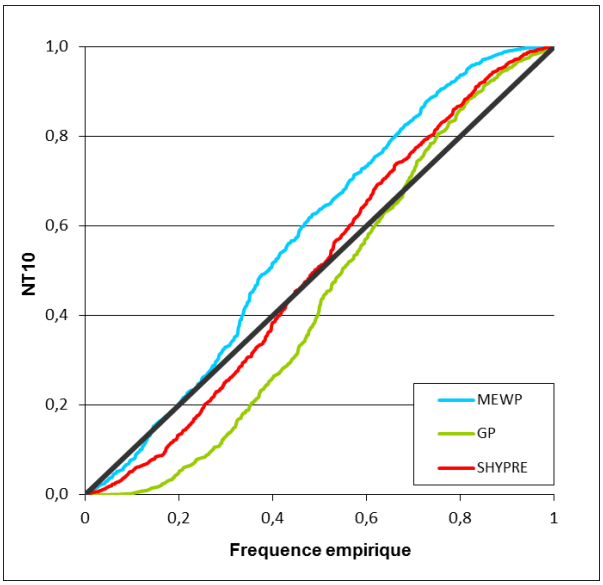


Figure 31. *Comparaison entre les estimations SHYPRE, MEWP et GP basée sur le critère N_{10}. Échantillonnage de catégorie 1, C33V66*

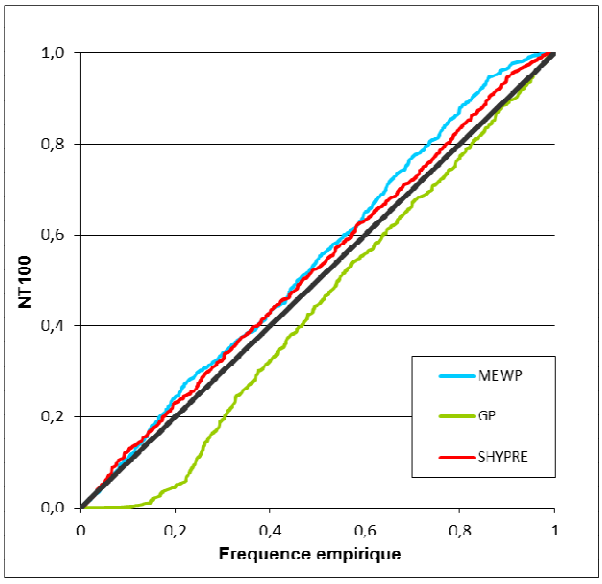


Figure 32. *Comparaison entre les estimations SHYPRE, MEWP et GP basée sur le critère N_{100}. Échantillonnage de catégorie 1, C33V66*

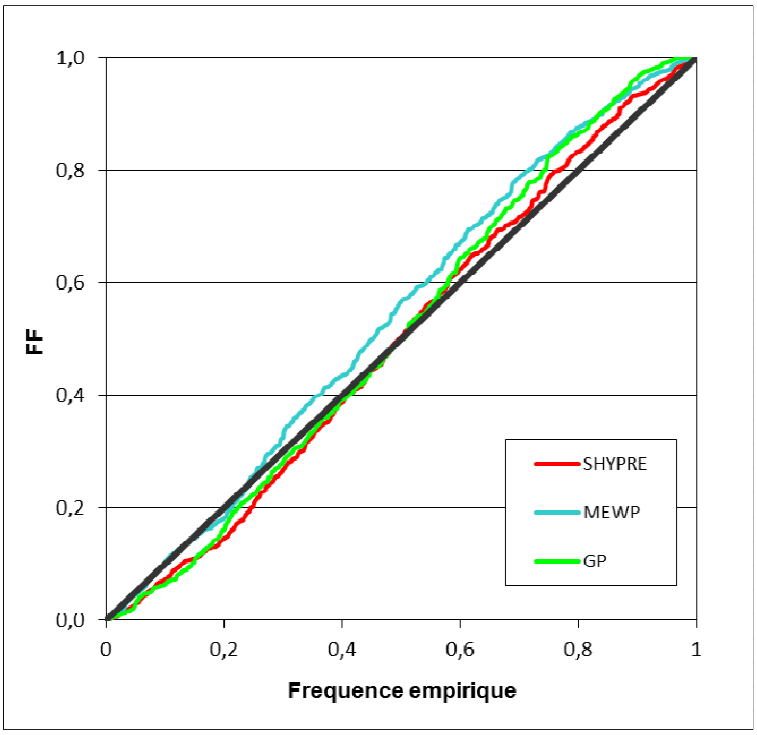


Figure 33. *Comparaison entre les estimations SHYPRE, MEWP et GP basée sur le critère FF. Échantillonnage de catégorie 1, C50V50.*

4.5.1.2 Robustesse

Le critère $SPAN_T$ montre que les modèles MEWP et SHYPRE sont plus robustes que le modèle GP. Nous ne calculerons pas le critère $COVER_T$ pour le modèle SHYPRE, le calcul étant trop long ; ce critère montre que le modèle MEWP donne des estimations plus robustes de la variance que le modèle GP, mais la différence est plus petite que pour le score $SPAN_T$ (voir Tableau 20). Comme les deux critères sont corrélés, on peut penser que la différence observée est due au manque de robustesse des estimations GP. Donc le résultat principal est la différence de robustesse évaluée avec le critère $SPAN_T$ entre la méthode GP et les méthodes MEWP et SHYPRE : les estimations du quantile de durée de retour 100 ans sont beaucoup moins robustes avec la méthode GP.

Tableau 20. *Comparaison entre les estimations GP, MEWP, SHYPRE basée sur les critères $SPAN_T$ et $COVER_T$. Échantillonnage de catégorie 2.*

$SPAN_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
GP	0,82	0,64	0,86	0,70	0,88	0,76
MEWP	0,86	0,81	0,89	0,85	0,92	0,89
SHYPRE	0,87	0,84	0,90	0,87	0,93	0,91

$COVER_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
GP	0,59	0,58	0,62	0,63	0,61	0,64
MEWP	0,60	0,62	0,62	0,63	0,64	0,63

4.5.1.3 Robustesse relativement à la valeur maximale

Le principal résultat est que la méthode GP est moins robuste vis-à-vis de la valeur maximale que les deux autres modèles, particulièrement pour la durée de retour 100 ans (voir Tableau 21). Le score $SPAN_T$ montre que les modèles MEWP et SHYPRE sont robustes de la même manière relativement à la valeur maximale et donnent d'excellents résultats : ils sont quasi insensibles à la taille de l'échantillon et à la durée de retour. Le critère $COVER_T$ confirme la différence de robustesse entre les modèles GP et MEWP pour les deux quantiles 10 et 100 ans. En conclusion, l'analyse de l'impact de la valeur maximale confirme les résultats précédents sur la robustesse des modèles : les modèles MEWP et SHYPRE sont plus robustes que le modèle GP.

Tableau 21. *Comparaison entre les estimations GP, MEWP et SHYPRE basée sur les critères $SPAN_T$ et $COVER_T$. Échantillonnage de catégorie 3.*

$SPAN_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
GP	0,95	0,90	0,96	0,92	0,96	0,93
MEWP	0,97	0,95	0,98	0,97	0,98	0,97
SHYPRE	0,98	0,97	0,98	0,98	0,99	0,98

$COVER_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans	T= 10 ans	T= 100 ans
GP	0,82	0,83	0,82	0,83	0,84	0,85
MEWP	0,90	0,88	0,92	0,89	0,94	0,91

4.5.1.4 Distribution prédictive

L'analyse avec la distribution prédictive donne les mêmes résultats, en terme de comparaison entre les méthodes GP, SHYPRE et MEWP, que l'analyse avec les estimations centrales. La principale différence entre les modèles GP et MEWP porte sur la robustesse : les estimations MEWP sont plus robustes que les estimations GP, surtout pour les grandes durées de retour (voir Tableau 22).

Tableau 22. *Comparaison entre les estimations MEWP et GP basée sur le critère $SPAN_T$. Distribution prédictive. Échantillonnage de catégorie 2.*

$SPAN_T$	10 ans – 10 ans		15 ans – 15 ans		25 ans – 25 ans	
	10 ans	100 ans	10 ans	100 ans	10 ans	100 ans
GP	0,78	0,56	0,82	0,63	0,85	0,69
MEWP	0,87	0,82	0,90	0,85	0,92	0,89

On remarque aussi que les estimations MEWP utilisant la distribution prédictive sont plus fiables que celles utilisant la distribution centrale (voir tableau 23). Ceci peut s'expliquer parce que, comme on l'a vu précédemment, le modèle MEWP sous-estime les quantiles. Lorsque nous utilisons la distribution exponentielle, le milieu de l'intervalle de confiance est supérieur à la médiane. Et le modèle prédictif donne des estimations plus proches de la médiane que les estimations par la méthode centrale. Donc pour MEWP la distribution prédictive donne de meilleurs résultats.

Tableau 23. *Comparaison entre les estimations MEWP et GP basée sur les critères N_{10} et FF. Distribution prédictive. Échantillonnage de catégorie 1.*

	25 ans – 25 ans		17 ans – 33 ans	
	N_{10}	*FF*	N_{10}	*FF*
GP	0,78	0,94	0,74	0,82
MEWP	0,77	0,97	0,79	0,95

4.5.1.5 Conclusion sur les estimations par les méthodes SHYPRE, MEWP et GP

Les résultats sont résumés sur la figure 34 pour la durée de retour 10 ans et la figure 35 pour la durée de retour 100 ans. En conclusion, si la méthode MEWP donne des estimations de quantiles souvent inférieures aux méthodes GP ou SHYPRE, elle permet d'obtenir de bons scores globaux en justesse et robustesse. La méthode SHYPRE, qui présente des estimations différentes des deux autres méthodes sur certaines zones de relief, obtient aussi de bons scores en justesse et robustesse. La loi GP présente des performances inférieures aux méthodes SHYPRE et MEWP, surtout en termes de robustesse et s'avère sensible aux effets d'échantillonnage. L'application des scores montre donc un léger avantage de l'estimation SHYPRE sur l'estimation MEWP.

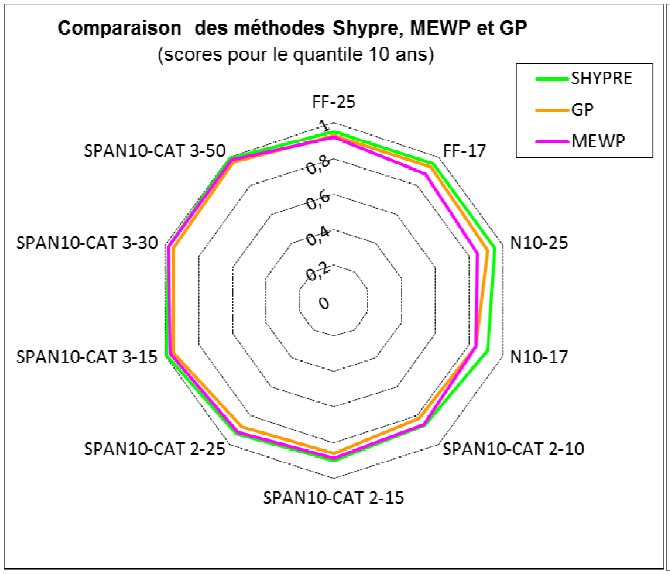


Figure 34. *Résultats pour les estimations SHYPRE, MEWP et GP. Durée de retour 10 ans*

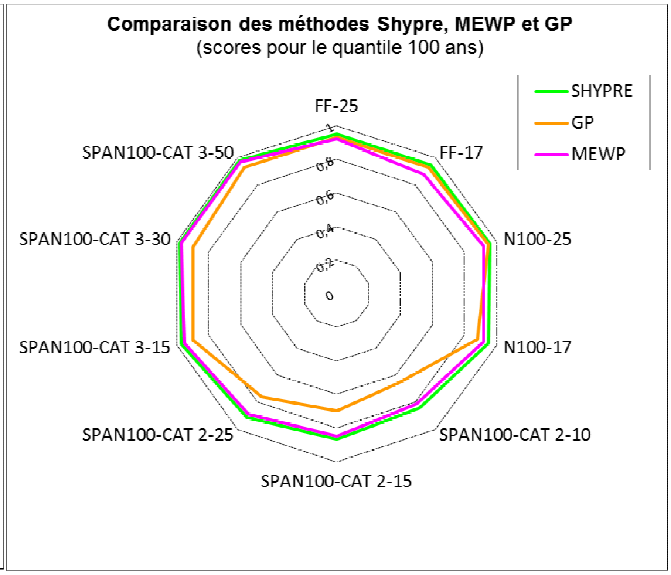


Figure 35. *Résultats pour les estimations SHYPRE, MEWP et GP. Durée de retour 100 ans*

4.6 Discussion régionale

Nous examinons ici le comportement régional des différentes méthodes étudiées précédemment SHYPRE, MEWP et GP sur l'échantillon 1 (693 longues séries de plus de 50 ans). On montre d'abord sur la Figure 35 que le découpage défini dans le §3.1 (voir Figure 4) relatif au ratio moyen entre précipitations extrêmes (PJX) et précipitations annuelles (PA) est beaucoup plus robuste qu'une approche basée directement sur les paramètres de forme Xi d'une loi GP trop influencé par les effets d'échantillonnage.

De plus, le Tableau 24 montre que ce découpage s'avère pertinent pour la distinction du caractère hyper-exponentiel des ajustements (valeurs différentes de la médiane de Xi pour les 3 sous zones) :

- la zone 1 avec un Xi médian de 0,151 correspond aux zones méditerranéennes,
- la zone 2 avec un Xi médian de 0,111 correspond aux montagnes du sud de la France,
- la zone 3 avec un Xi médian de 0,085 correspond au reste du pays.

Il permet également d'identifier des comportements différents entre les méthodes dans ces 3 zones. Cette distinction est mise en évidence à partir du comptage des séries pour lesquelles une méthode donne une estimation de pluie centennale significativement supérieure aux 2 autres (cf §4.5.1 et *Figure 28*). Ainsi la méthode SHYPRE donne des estimations en moyenne supérieures aux deux autres méthodes dans la zone 2 (ratio intermédiaire), avec une fréquence 2 fois plus fortes que sur l'ensemble de l'échantillon.

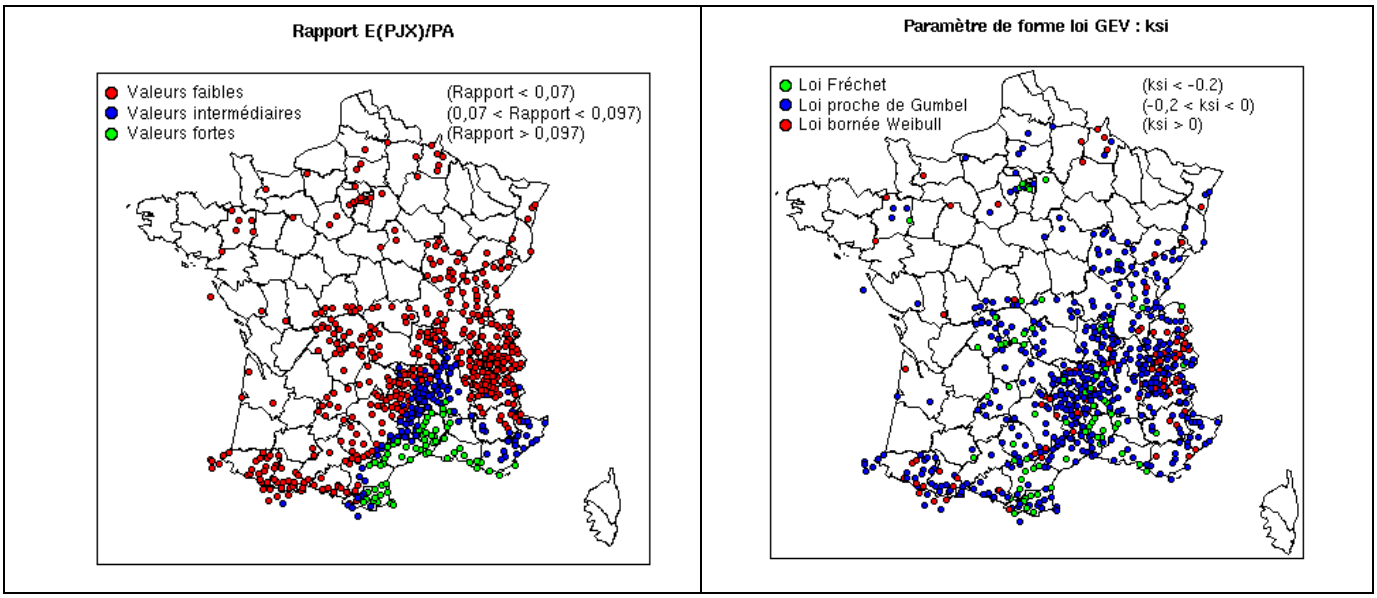


Figure 35. *Comparaison de zonages issus du ratio PJX/PA à gauche et du paramètre Xi d'une loi GP à droite, pour l'échantillonnage de catégorie 1.*

Tableau 24. *Nombre et ratio en pourcentage des stations pour lesquelles une des estimations centennales GPD, SHYPRE ou MEWP est supérieure de plus de 10% aux deux autres.*

Zones	Valeur médiane de Xi	SHYPRE Supérieur	MEWP Supérieur	GP Supérieur	Aucune
1 : Ratio PJX/PA fort (>0,097) (70 stations)	0,151	8 (11 %)	0 (0 %)	29 (41%)	33 (47 %)
2 : Ratio PJX/PA intermédiaire ([0,007 ;0.097] (138 stations)	0,111	43 (31 %)	3 (2 %)	28 (20 %)	64 (46 %)
3 : Ratio PJX/PA faible (<0,07) (485 stations)	0,085	43 (9%)	4 (1%)	144 (30%)	294 (61%)
France entière (693 stations)	0,096	94 (14 %)	7 (1 %)	201 (29 %)	391 (56 %)

L'analyse du critère de justesse FF (Tableau 25) met en évidence des performances moindres de l'ensemble des méthodes sur la zone Méditerranée, avec une dégradation plus marquée pour la loi MEWP, influencée par la combinaison de lois exponentielles peu pertinentes dans cette zone.

Tableau 25. *Comparaison entre les estimations SHYPRE, MEWP et GP basée sur le critère FF. Échantillonnage de catégorie 1, C50V50*

	FF		
	Méditerranée	Montagne Sud	Reste de la France
SHYPRE	0,88	0,96	0,94
MEWP	0,83	0,89	0,93
GP	0,88	0,95	0,92

L’analyse des figures 36 et 37 sur les zones Méditerranée et à un degré moindre Montagne Sud montre que la dégradation des scores observés est principalement due à une sous estimation des quantiles estimés par rapport aux valeurs extrêmes observées.

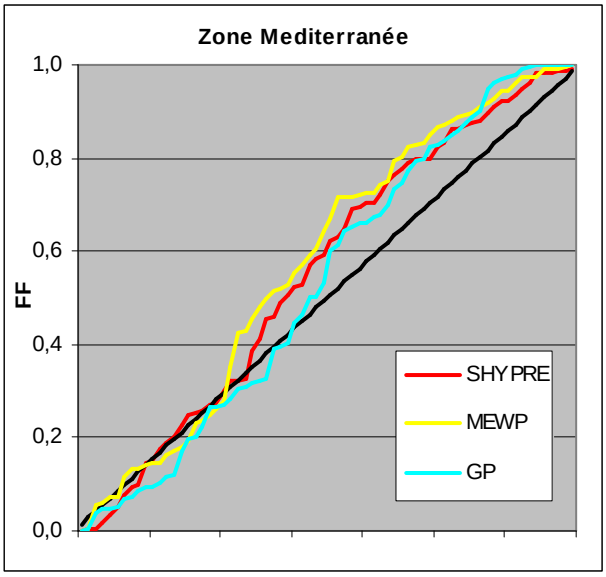


Figure 36. *Comparaison du critère FF sur les estimations GP, MEWP et SHYPRE. Zone Méditerranée. Échantillonnage de catégorie 1, C50V50*

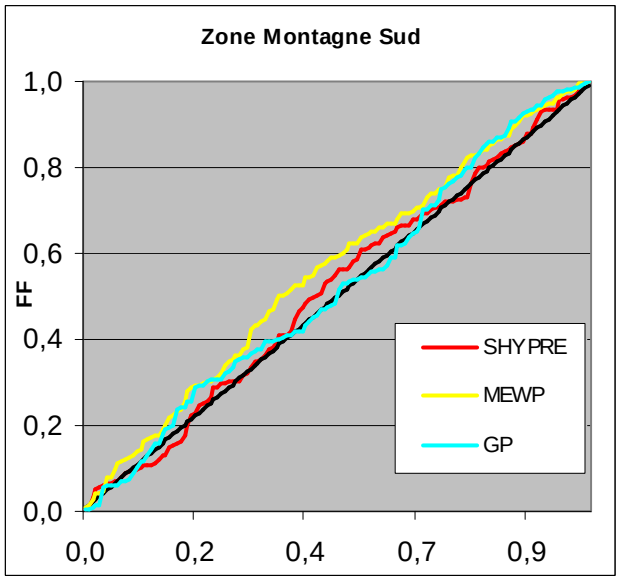


Figure 37. *Comparaison du critère FF sur les estimations GP, MEWP et SHYPRE. Zone Montagne Sud. Échantillonnage de catégorie 1, C50V50*

5 Conclusions et perspectives

Ce rapport présente une comparaison des principales méthodes d'estimation des valeurs extrêmes de pluies, à l'échelle locale, grâce à une très grande base de données quotidiennes regroupant des données provenant d'EDF et de Météo France. Nous avons comparé sept méthodes pour estimer les valeurs extrêmes : lois GEV, exponentielle, GP (avec trois estimateurs différents des paramètres), et méthodes MEWP et SHYPRE. Tout d'abord, nous vérifions que ces méthodes donnent des estimations des quantiles de durée de retour 10 ans plus proches entre elles lorsque l'on dispose de 50 années de données ou plus. Ainsi quand la taille de l'échantillon augmente, toutes les méthodes tendent vers la même estimation. Mais on a aussi constaté que ces estimations sont légèrement sous-estimées, d'environ 5 pour cent pour le modèle GP avec un estimateur PWM par exemple, et un peu plus pour le modèle MEWP. Nous notons une différence entre les estimations principalement pour les stations qui peuvent observer les plus fortes pluies : le modèle SHYPRE fournit des estimations plus élevées du quantile de durée de retour 10 ans. Mais pour la grande majorité des stations, les estimations tendent à être les mêmes et les différences entre les méthodes en robustesse et en justesse apparaissent soit quand la taille de l'échantillon décroît, soit quand la durée de retour augmente.

Pour commencer nous avons comparé les modèles paramétriques classiques (GEV et GP) utilisant l'estimateur du maximum de vraisemblance. Il en ressort que la méthode GP donne de meilleurs résultats, surtout pour les grandes périodes de retour, quand la taille de l'échantillon décroît. Nous avons alors cherché à analyser l'apport d'une paramétrisation des lois sur les valeurs extrêmes à trois paramètres au lieu de deux. Pour cela, nous avons comparé les méthodes GP et Exponentielle (EXPO) en utilisant l'estimation du maximum de vraisemblance et les quatre scores précédents. Nous avons observé des caractéristiques opposées entre les estimations de GP, beaucoup plus performantes en justesse et celles d'EXPO beaucoup plus robustes, même avec des échantillons réduits : avec moins de paramètres à estimer, les estimations sont plus stables sur des échantillons différents, mais la différence importante sur la justesse montre tout l'intérêt de pouvoir disposer d'un paramètre de forme. Enfin, nous avons complété l'étude des modèles paramétriques en comparant trois estimateurs pour le modèle GP : méthode du maximum de vraisemblance, méthode des moments et méthode des moments pondérés. L'estimateur MM fournit des estimations plus robustes et PWM est plus robuste relativement à la valeur maximale et légèrement plus fiable. Nous choisissons finalement l'estimateur PWM non à cause du petit avantage montré par les scores mais à cause du nombre de fois où l'estimateur MM ne peut fournir une estimation des quantiles.

L'étape suivante a consisté à comparer le meilleur modèle paramétrique classique, soit la loi GP avec l'estimateur PWM, à deux autres modèles : les modèles MEWP et SHYPRE. Nous avons constaté que SHYPRE apparaît comme la méthode la plus juste devant GP et MEWP. Toutes les méthodes semblent connaître des performances moindres en zone Méditerranée avec une tendance constatée à la sous-estimation, plus marquée pour la méthode MEWP pénalisée par la combinaison de lois exponentielles peu adaptés au contexte Méditerranéen des précipitations extrêmes. La méthode SHYPRE donne des estimations plus fortes que les autres méthodes dans la zone Montagne Sud sans que l'on puisse mettre en évidence de biais.
Quand on considère la robustesse, on constate que la robustesse du modèle GP est inférieure à celles de MEWP et de SHYPRE. D'une part le modèle GP fournit des estimations de quantiles moins robustes, surtout pour la durée de retour 100 ans. D'autre part, l'analyse de l'impact de la plus grande observation sur l'estimation indique que le modèle GP est là aussi moins robuste.

D'après cette étude, le modèle EXPO est robuste mais manque de justesse. Le modèle GEV est pénalisé par la petite taille des échantillons constitués uniquement de valeurs maximales annuelles. Si la

différence du modèle GP avec les modèles MEWP et SHYPRE est peu importante pour les petites durées de retour (10 ans), nous notons que les modèles MEWP et SHYPRE donnent des estimations beaucoup plus robustes pour les quantiles de durée de retour 100 ans. On rappelle que cette étude ne traite pas des très grandes durées de retour (1000 années et plus) pour lesquelles la validation reste très difficile.

Nous avons basé nos conclusions sur quatre scores décrivant la robustesse et la justesse des modèles, ainsi que sur des interprétations graphiques et des scores calculés sur des sous-échantillons. Le choix de ces critères vient de la littérature et, même s'il est toujours possible d'ajouter de l'information avec d'autres scores, nous pouvons considérer que nous décrivons aussi bien que possible les qualités des modèles. Cependant, un score supplémentaire capable de donner de l'information sur la justesse des modèles pour des durées de retour supérieures à 100 ans serait intéressant.

Cette étude donne donc un certain nombre d'indications sur la comparaison de modèles pour estimer des valeurs extrêmes, mais d'autres questions ne sont pas résolues comme la justesse exacte des estimateurs pour les grandes et très grandes durées de retour ou les effets du changement climatique sur les résultats des modèles.

6 Bibliographie

- Arnaud, P., 1997 : Modèle de prédétermination de crues basé sur la simulation. Extension de sa zone de validité, paramétrisation du modèle horaire par l'information journalière et couplage des deux pas de temps. Thèse de doctorat de l'Université Montpellier II
- Arnaud, P., Fine, J-A., Lavabre, J., 2007 : An hourly rainfall generation model applicable to all types of climate. Atmospheric Research 95, 230-242
- Ashkar, F., Nwentsa Tatsambon, C., 2007 : Revisiting some estimation methods for the generalized Pareto distribution. Journal of Hydrology, 346, 136-143
- Cantet, P., 2009: Impacts du changement climatique sur les pluies extrêmes par l'utilisation d'un générateur stochastique de pluies. Thèse de doctorat de l'Université Montpellier II, 178 p
- Cernesson, F., 1993 : Modèle simple de prédétermination des crues de fréquences courantes à rares sur petits bassins versants méditerranéens. Thèse de doctorat de l'Université Montpellier II
- Choisnel, E., et Payen, D., 1988 : Les climats de la France. La Recherche, supplément au n°201, pages 32 à 41
- Coles, S., 2001: An introduction to statistical modelling of extreme values. Springer series in statistics
- Coles, S., Perricchi, L., Sisson, S., 2003 : A fully probabilistic approach of extreme rainfall modelling. Journal of Hydrology, 273, 35-50
- Garavaglia, F. et al., 2010 : Introducing a rainfall compound distribution model based on weather patterns sub-sampling. Hydrology and Earth System Sciences, 14, p.p. 951 - p. 964.
- Garavaglia, F., 2011 : Méthode SHADEX de prédétermination des crues extrêmes. Thèse de doctorat présentée et soutenue le 23 février 2011, Université de Grenoble
- Garavaglia F., Lang M., Paquet E., Gailhard J., Garçon R., and Renard B., 2011. Reliability and robustness of rainfall compound distribution model based on weather pattern sub-sampling. Hydrol. Earth Syst. Sci., 15, 519-532
- Garçon, R., 1995 : Communication orale. Statistical and Bayesian Methods in Hydrological Sciences. A joint UNESCO International Conference in honour of Jacques Bernier, Paris, 11-13 septembre 1995.
- Gilleland, E. et Katz R.W., 2005 : Tutorial for the Extremes Toolkit: Weather and Climate Applications of Extreme Value Statistics, http://www.assessment.ucar.edu/toolkit.
- Hosking, J. R. M., 1990: L-moments: Analysis and estimation of distributions using linear combinations of order statistics. J.R. Stat. Soc. Ser, B 52, 105-124.
- Jenkinson, A-F., 1955 : The frequency distribution of the annual maximum (or minimum) of meteorological elements. Quarterly Journal of the Royal Meteorological Society, 81, 158-176
- Kumaraswany, P., 1980: a generalized probability density functions for double-bounded random processes. Journal of Hydrology, 46, 79-88
- Lang, M., Ouarda, T., Bobée, B., 1999: Towards operational guidelines for over-threshold modelling. Journal of Hydrology, 225, 103-117
- Mestre O.,2004 : Detection and correction of artificial shifts. Appl. Statist, 53, Part 3, pp. 405–425.
- Min, S.-K., Zhang, X, Zwiers, F.W., Hegeri G.-C., 2011: Human contribution to more intense precipitation extremes. Nature 470, 378-381, doi: 10.1038/nature09763

- Moisselin J.M., Schneider M., Canellas C., Mestre O., 2002: Les changements climatiques en France au 20ème siècle. Étude des longues séries homogénéisées de données de température et de précipitations. La Météorologie, n°38, août 2002, 45-56p.
- Penot D, 2011-2014. Cartographie de pluies extrêmes et application de la méthode SCHADEX en site non jaugé. Thèse en cours à EDF/DTG, Université Grenoble
- Pirazzoli P.A., 2007 : Données pour le dimensionnement des structures côtières et des ouvrages de bord de mer à longue échéance
- Pickands, J.,1975 : Statistical inference using extreme order statistics. Annals of Statistics 3: 119–131
- Prescott, P. et Walden, A. T., 1980 : Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. Biometrika 67, 723-724.
- Renard B., Kochanek K., Lang M., Garavaglia F., Paquet E., Neppel L., Najib K., Carreau J., Arnaud P., Aubert Y., Borchi F., Soubeyroux J.M., Jourdain S., Veysseire J.M., Sauquet E., Cipriani T., Auffray A., 2013. Data-based comparison of frequency analysis methods: A general framework. Water Resources Research, 49, 1-19.
- Ribereau, P., Guillou, A., Naveau, P., 2008 : Estimating return levels from maxima of non-stationary random sequences using Generalized PWM method. Nonlinear Processes in Geophysics, 15, 1033-1039.