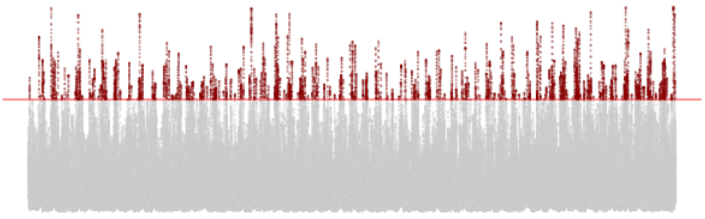
Université de Montpellier
Faculté des Sciences
M2 Statistiques et Sciences des Données

Etude des vagues extrêmes dans le golfe du Lion

HAX005X



Auteurs:
Wiam Chaoui
Anas Zakroum

Enseigné par:
Pre. Gwladys Toulemonde
Pr. Nicolas Meyer
Pr. Jean-Noël Bacro

3 mars 2022

Table des matières

Introduction

Nous disposons d'un jeu de donnée décrivant les hauteurs de vagues dans le golfe du Lion , mesurées dans une vingtaine de stations dédiées à cet effet. Nous nous intéressons dans cette étude aux stations 6, 16. Nous disposons des coordonnées géographique des stations : la longitude et la latitude.

Le jeu de donnée est très riche ; les mesures sont à échelle horaire et s'étendent sur une periode d'un demi siècle, de 1961 à 2012. Un néttoyage des données a été effectué. Le jeu de données contenait des observations dupliquées. Regardons la répartition des hauteurs de vagues enregistrées par la station 6.

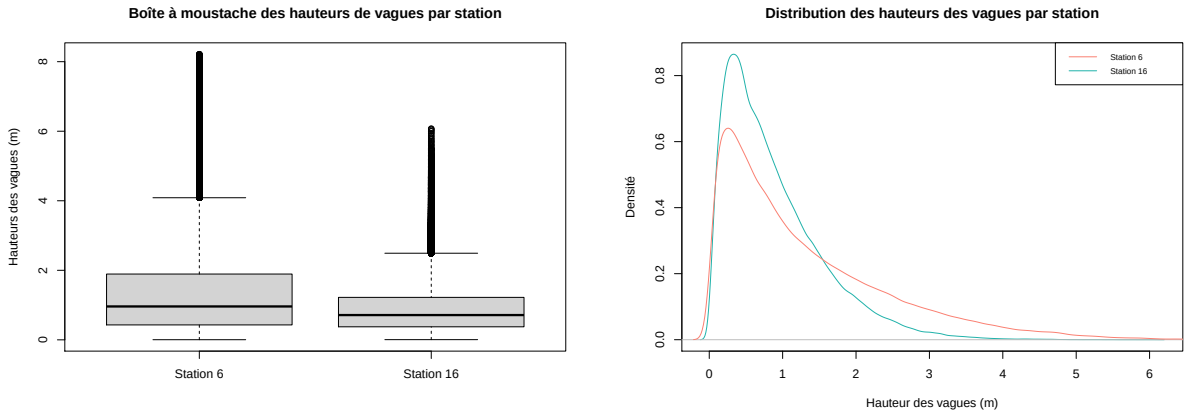


FIGURE 1 – Repartition des hauteurs des vagues enregistrées dans les stations 6 et 16.

La première observation qui ressort de la Figure 1 est que les hauteurs des vagues diffèrent significativement d'une station à une autre. Ceci peut s'expliquer par le fait que les stations sont à des longitudes et un niveau de profondeur différents.
Le boxplot nous renseigne sur les quantiles et la densité sur les modes et la distribution des données. Sur la station 6 par exemple, la grande majorité des vagues ne dépasse pas 2 mètres avec une médiane à 0.96 mètres, une moyenne à 1.31 mètres et le troisième quantile, au niveau de 1.9 mètres. Ceci nous donne une première idée sur les ordres de grandeurs et sur ce que peut être considéré comme évènement extrême pour chaque station.

Dans la suite, nous nous penchons à l'étude des extrêmes par l'intermediaire de différents modèles. Dans un premier temps nous étudierons les extrêmes de la station 6, les niveaux de retour puis la dépendance temporelle. Dans un second temps, nous adoptons une approche bivariée où nous étudierons la station 6 et une station choisie arbitrairement puis la station 6 et la station la plus proche d'elle dans une tentative d'estimer les quantiles extrêmes conditionnels.

1 Etude unidimensionnelle

Nous penchons notre étude sur une analyse des extrêmes univariée pour la station 6. Nous considérerons deux approches standards ; l'étude du maxima par bloc et l'étude des dépassements de seuil.

1.1 Maxima par bloc

Cette approche est motivée par les travaux de Fisher et Tippet [1] énoncant que le comportement asymptotique du maximum de variables aléatoires sous l'existence d'une suite normalisante est attiré vers une famille de lois limites appelées GEV. Cette famille fournit un modèle pour la distribution des maximums par bloc.
Dans l'implémentation de ce modèle, les données sont regroupées sous forme de bloc de taille équivalente. Le choix de la taille des blocs est un élément détérminant de la qualité de l'ajustement du modèle. Des blocs trops petits produiront du biais dans l'estimation et de l'extrapolation. Des blocs trops grand fournissent peu de données et par conséquent, généreront de la variance dans l'estimation.

1.1.1 Quelle taille de bloc choisir ?

Afin de choisir la taille des blocs pour notre étude, nous avons décidé de regarder des blocs d'un mois et des blocs d'un an :

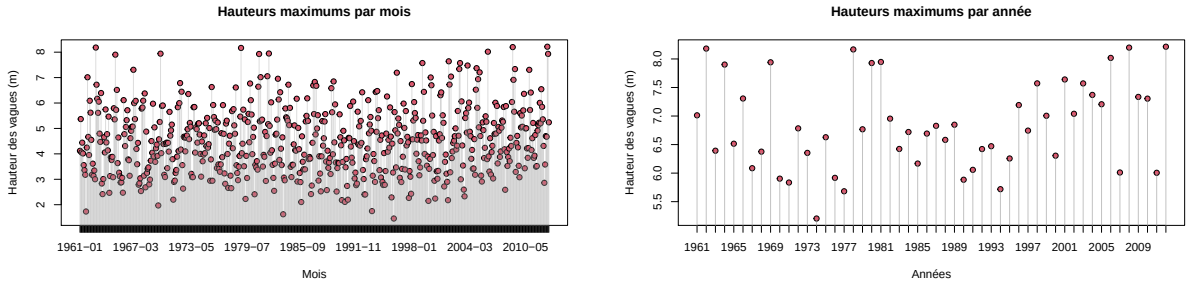


FIGURE 2 – Construction des maximums par bloc pour la station 6. Les points en rouge constituent les maximums pour chaque mois (à gauche) et pour chaque année (à droite).

Nous obtenons donc 52 mesures pour les maximums annuels et 624 mesures pour les maximums mensuels. Autant de données dans le cas des blocs mensuels fourniraient des estimations robustes sous reserve que les hypothèses centrales du théorème des valeurs extremes sont vérifiées comme l'indépendance des observations où la distribution identique des données. En effet, les vagues peuvent être différentes selon les vents et les périodes de l'année.
De plus au regard de la Figure 2, les blocs mensuels capturent également des vagues communes , entre 1 et 3 mètres.

Le modèle des blocs mensuels résulte donc, en plus, en des maximums qui ne sont pas des extrêmes, et donc nous portons notre étude sur le modèle aux blocs annuels qui pour lequel, même si la justification formelle du paradigme des valeurs extrêmes est difficile à vérifier, les hypothèses que les blocs individuels possèdent la meme distribution et de l'independance des vagues extrême d'une année à une autre semblent plausibles.
Notons enfin, qu'il est possible de procéder à des choix de tailles de bloc plus adaptés au phénomène étudié prenant en compte les périodes de l'année où la mère est plus agitée ou encore la dépendance des observations entre elles...Ces considérations résulterons donc en des modèles à la fois plus fins mais aussi plus complexes.

1.1.2 Ajustement du modèle

On suppose que les observations des maximums annuels sont indépendantes et sont issues d'une GEV. Divers travaux ont été developpés dans le but d'estimer les paramètres de GEV en partant d'une modélisation par blocs comme la méthode des L-moments [2], la méthode des moments pondérés [3] ou autres ...

Nous utilisons dans notre étude des estimations basée sur la maximisation de la vraisemblance, des travaux ont été conduits en ce qui concerne l'étude des GEV [4].
La maximisation de la vraisemblance résulte en les estimateurs suivants :

$$(\hat{\mu}, \hat{\sigma}, \hat{\gamma}) = (6.56, 0.735, -0.257) \tag{1}$$

.

La combinaison des estimateurs avec leurs écarts-types respectifs résultent en les intervalles de confiance 95% suivants :

TABLE 1 – Intervalles de confiance des estimateurs du maximum de vraisemblance.

	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\gamma}$
$IC_{95\%}$	[6.3264, 6.7939]	[0.5596, 0.9103]	[-0.5299, 0.0146]

Une meilleure précision de l'intervalle de confiance peut être atteinte en utilisant la vraisemblance profilée, nous obtenons un intervalle très semblable au premier ;

$$[-0.5429, 0.0167].$$

Commentaire

L'estimateur $\hat{\gamma}$ est négatif avec un intervalle de confiance qui arrive à peine à zero. Les données suggèrent de manière plutôt forte, la possibilité de la présence d'une queue de distribution bornée correspondant à une loi de Weibull.

1.1.3 Validité du modèle et diagnostic

Nous procédons à un diagnostic visuel de la validité du modèle ajusté.

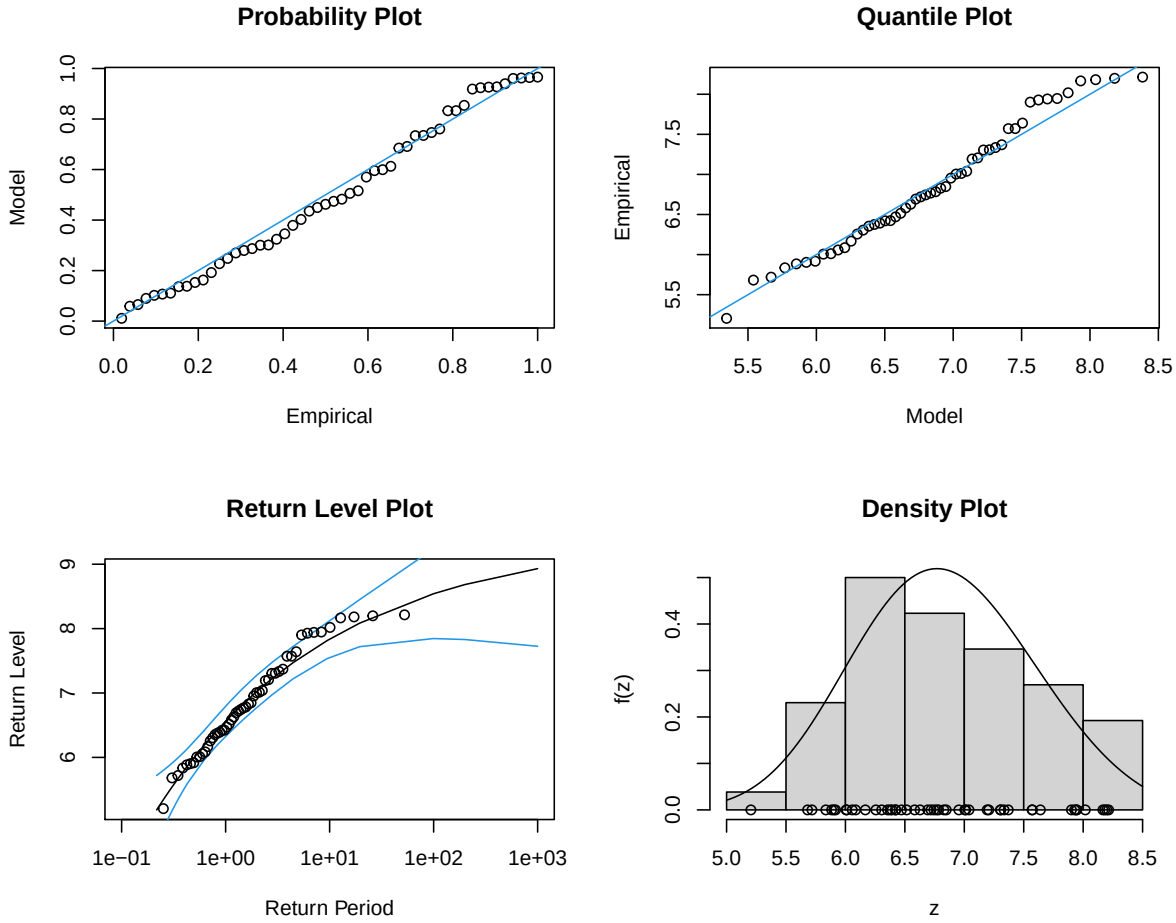


FIGURE 3 – Evaluation graphique de la validité du modèle.

Le Probability Plot compare les fonctions de répartition empirique avec celle issue de l'ajustement. L'étude des extrêmes s'interesse à la qualité du modèle pour les plus grandes observations. Cependant, ce graphique compare la f.d.r empirique à celle issue du modèle qui sont contraintes par construction à s'approcher de 1 d'autant plus que les observations sont grandes. Ainsi ce graphique fournit peu d'information sur la région dont l'interêt est le plus grand.

Le P-plot et le QQ-plot indiquent une linéarité assez correcte.
La courbe du niveau de retour semble tendre vers un niveau de retour fini en conséquence de l'estimation négative du paramètre de queue de distribution γ. Ce qui est aussi cohérent avec nos données dont les vagues les plus hautes ne dépassent pas $8,5$ mètres. La courbe fournit également une représentation satisfaisante pour une grande partie des observations. Enfin la densité issue du modèle semble consistante avec l'histogramme.

Comme l'intervalle de confiance du paramètre de queue indique la plausibilité de la présence d'une Gumbel, une alternative pour évaluer le modèle est d'effectuer un test de rapport de vraisemblance entre les deux modèles, donc $\gamma = 0$ et $\gamma < 0$.

L'ajustement sur une Gumbel fournit les estimateurs

$$(\hat{\mu}, \hat{\sigma}) = (6.465,\ 0.687),$$

avec des écarts type réspectifs de 0.101 et 0.072.
La vraisemblance maximisée est de -60.74.
La statistique de test pour la réduction au modèle de Gumbel suit une loi de $\mathcal{X}^2$ à un degrès de liberté. En considérant un seuil de $\alpha = 0.05$, elle vaut dans cet exemple

$$\mathcal{D} = 2\{-59,02 - -(60,74)\} = 3.44 < 3.84.$$

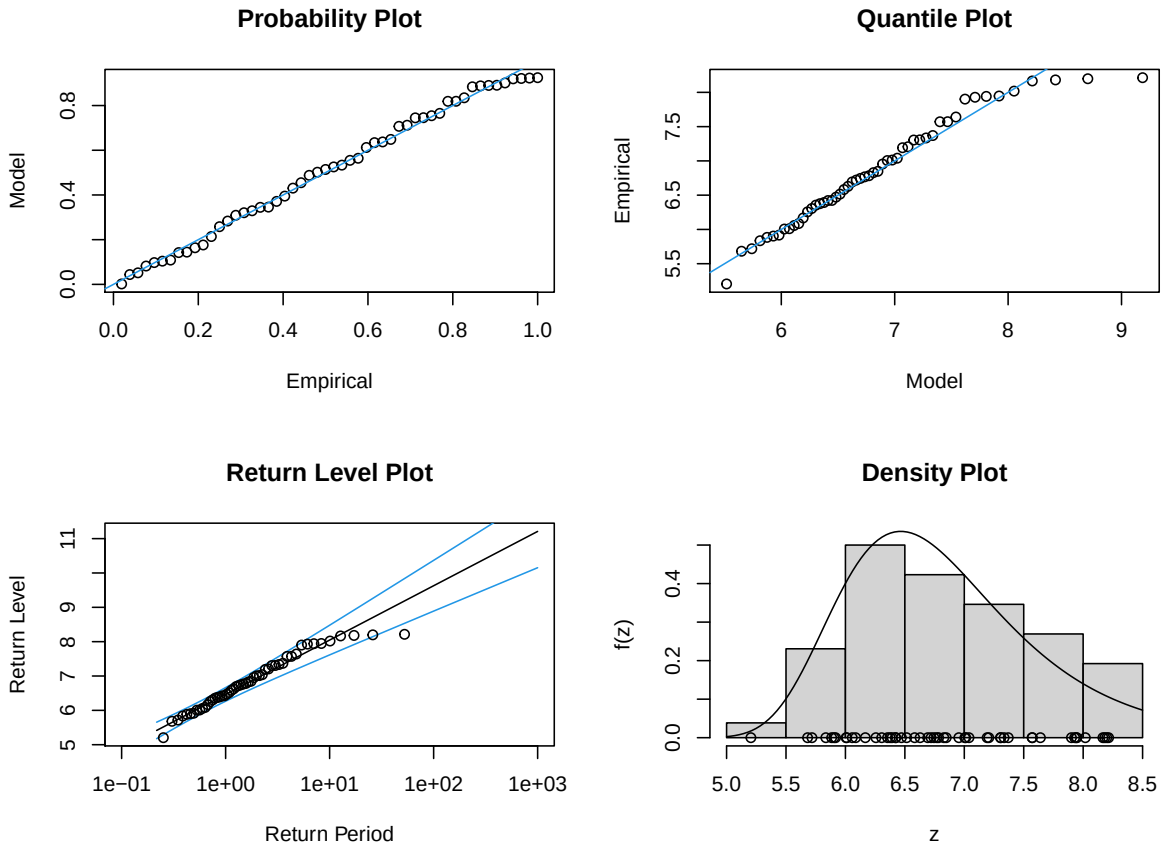


FIGURE 4 – Graphes de diagnostic de l'ajustement sur le modèle de Gumbel.

Commentaire

La statistique de test est assez proche du seuil critique, ce qui rend la validité du sous modèle discutable. De plus le diagnostic graphique de la Figure 4 ci-dessus revèle une extrapolation parlante des observations par rapport au modèle, ce qui nous pousse à préferer le modèle complet avec $\gamma < 0$ au sous modèle de Gumbel.

1.1.4 Niveaux de retour et incértitude

Les estimations des quantiles extrêmes de la distribution peuvent être retrouvées par inversion de la fonction de répartition de la GEV et en utilisant les paramètres estimées à l'aide de la relation

$$z_p = \begin{cases} \mu - \frac{\sigma}{\gamma}[1 - \{-\log(1-p)\}^{-\gamma}] & \text{si } \gamma \neq 0 \\ \mu - \sigma \log\{-\log(1-p)\} & \text{si } \gamma = 0 \end{cases} \tag{2}$$

.

Avec $G(z_p) = 1 - p$.

z_p est le niveau de reetour associé à la periode de retour $\frac{1}{p}$ et représente le niveau moyen qu'on s'attend à voir être dépassé tous les $\frac{1}{p}$ ans.
La delta-methode rend possible le calcul de la variance du quantile extrême estimé z_p à l'aide de la matrice de variance covariance des estimateurs $\hat{\mu}$, $\hat{\sigma}$, $\hat{\gamma}$. Nous implémentons alors une fonction `gev_ret()` prenant en entrée les données et la période de retour souhaitée, et qui rend en sortie l'estimation du niveau de retour ainsi que son écart type. Les résultats obtenus sont

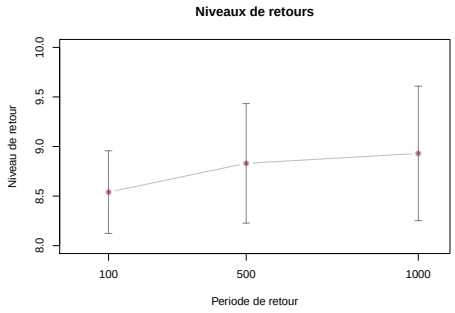


FIGURE 5 – Représentation graphique des niveaux de retours et leurs écart types respectifs pour les périodes de retour de 100, 500, et 1000 ans

	Période de retour en année		
	100	500	1000
Niveau de retour (m)	8.54	8.83	8.93
Ecart type	0.417	0.603	0.678

TABLE 2 – Niveaux de retours estimés

Nous observons encore une fois un effet de plateau concernant les niveau de retour. Ceci est cohérent avec les données qui ne dépassent pas 8 mètres. L'incertitude sur le niveau de retour est davantage grande que la période de retour est longue.
Par ailleurs, comme le paramètre de queue γ est estimé à une valeur négative, la queue de distribution des maximum est bornée. Il est donc possible d'inférer la valeur du point terminal de la distribution modélisée par

$$\hat{z_0} = \hat{\mu} - \hat{\sigma}/\hat{\gamma} = 6.56.$$

Ce qui nous laisse, disons...perplexe en vue des données dont on dispose et aussi des estimations des niveaux de retour observés.

1.1.5 Présence d'une dépendance temporelle ?

Il n'est pas déraisonnable de supposer la non-stationnarité dans le temps quand il s'agit d'étudier des processus environnementaux. Particulièrement, face à des phénomènes susceptibles de changer à long terme, l'hypothèse d'une modélisation des extrêmes avec des paramètres fixes au cours du temps peut porter défaut à la pertinence du modèle. Nous nous penchons sur cette question dans cette partie.

Linéarité

Supposons que la hauteur des vagues maximum suit une GEV où le paramètre localisation dépend linéairement du temps : $\mu(t)$ avec

$$\mu(t) = \beta_0 + \beta_1 t.$$

La dépendance temporelle du paramètre de location produit une vraisemblance de -58.154 et les estimateurs des paramètres sont :

$$(\hat{\mu}(t), \hat{\sigma}, \hat{\gamma}) = (6.3 + 0.01t,\ 0.7,\ -0.21).$$

Les écarts type sont respectivement 0.23, 0.007, 0.083, 0.12.
La statistique de déviance étant

$$\mathcal{D} = 2(59.02 - 58.15) = 1.73 < \mathcal{X}_1^2(5\%) = 3.84$$

nous suggère qu'une tendance temporelle linéaire n'est pas justifiable par les données.

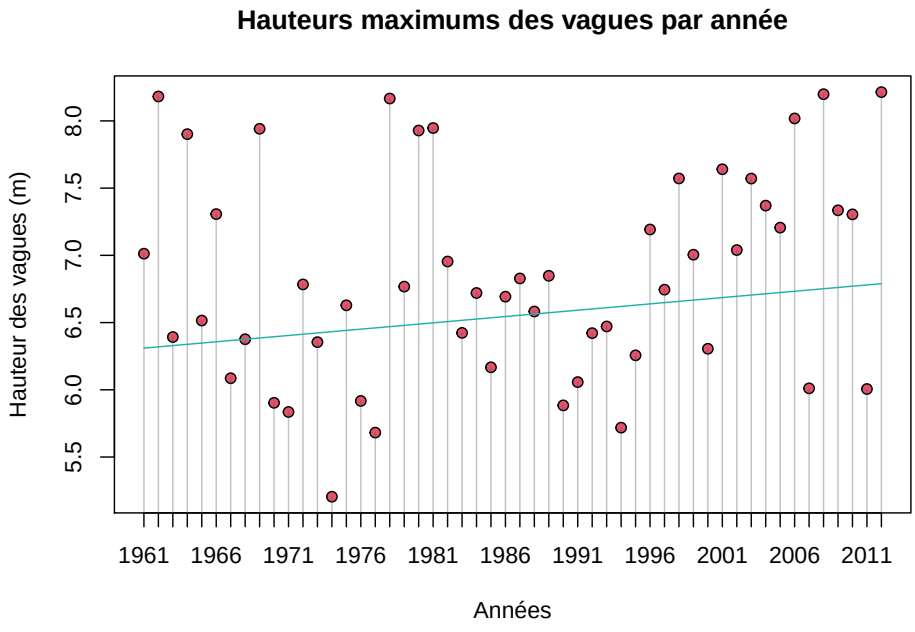


FIGURE 6 – Représentation de la courbe de tendance du paramètre de localisation de la GEV.

Comme nous pouvons observer par ailleurs sur la figure et sur la valeur du coefficient β_1, il n'y pas réellement de tendance sur toute la totalité de la periode considérée. Nous pourrions supposer à l'allure du graphique une tendance linéaire par morceaux.

Une augmentation tous les siècles ?

Nous regardons maintenant l'évolution de la hauteur des vagues à l'échelle d'un siècle.

$$\mu(t) = \beta_0 + \beta_1 \frac{t}{100}.$$

L'ajustement d'une GEV sur ce modèle renvoie pour une vraisemblance de $58,15$ les estimateurs suivants

$$(\beta_0, \beta_1, \sigma, \gamma) = (6.29,\ 0.93,\ 0.705,\ -0.2082468).$$

Avec des écart types respectifs de $0.231,\ 0.716,\ 0.083,\ 0.130$.

Ceci résulte en u ngrand intervalle de confiance pour β_1. Ce modèle est il vraiment nécessaire ? Le test d'hypothèse

$$H_0 : \beta_1 = 0 \qquad \text{v.s} \qquad H_1 : \beta_1 \neq 0$$

renvoie une statistique de test $|\frac{\beta_1}{\sigma_{\beta_1}}|$ de 1.3 indiquant le rejet de l'hypothèse H_0 au seuil de 5%avec une probabilité critique de 0.18.

Par ailleurs, le test de rapport de vraisemblance entre le modèle avec un paramètre de localisation fixe et ce modèle indique une statistique de test inférieure à une $\mathcal{X}_1^2(5\%)$ suggèrant que le modèle sans dépendance temporelle est préférable.

1.1.6 Limitations

Le modèle des maxima par bloc fait intervenir la question cruciale du choix de la taille des blocs. Comme nous avons cité dans la Section 1.1.1, un choix de bloc optimisée permetterait une étude plus pertinente du phénomène souhaité. A savoir, un bloc d'un an produit une seule observation par année. un bloc par mois produit un nombre satisfaisant d'observations, mais capture de l'information indésirable, ce qui viendrait biaiser les estimations et invalider les hypothèses du modèle. Les deux méthodes ne capturent pas le fait que les vagues extrêmes se produisent successivement en temps de tempêtes, mais aussi la variabilité saisonnière.

Nous étudions dans la suite une méthode basée sur le dépassement de seuil qui ne consiste pas en regroupement des données par bloc mais par l'étude des dépassements au dessus d'un seuil fixé permettant une meilleur exploitation de l'information que dans un modèle de maxima par bloc.

1.2 Dépassements de seuils

Dans ce modèle, une vague extrême est considérée comme une vague dont la hauteur dépasse un seuil supposé être grand. La loi d'interêt est la loi des dépassements

$$\mathbb{P}(X - u > z | X > u) \quad , \quad z > 0.$$

Le modèle de dépassements de seuils s'est developpé après les travaux fondateurs de Balkema et De Haan en 1974 [5]. Le modèle des dépassements de seuils est fondé sur le résultat suivant : si le maxima par bloc peut être approché par une GEV, alors les dépassements de seuils peuvent être approximés par une distribution de la famille Paretto généralisée.

1.2.1 Quel seuil choisir ?

Le problème du choix du seuil est analogue au problème du choix de la taille des blocs. Un seuil trop grand resulterait en un petit nombre d'observations menant à plus de variance dans les estimations, un seuil trop petit, ne vérifierait pas les fondations asymptotiques du modèle et induira donc du biais.

Espérence des dépassements de seuil. Une première méthode basée sur la moyenne de la GDP, consiste à exploiter la linéarité de l'espérence des dépassements de seuil vue comme une fonction de u (Mean excess function) sous l'hypothèse que l'ajustement d'une GDP est valide pour un certain seuil u_0

$$u \longmapsto \mathbb{E}[X - u | X > u].$$

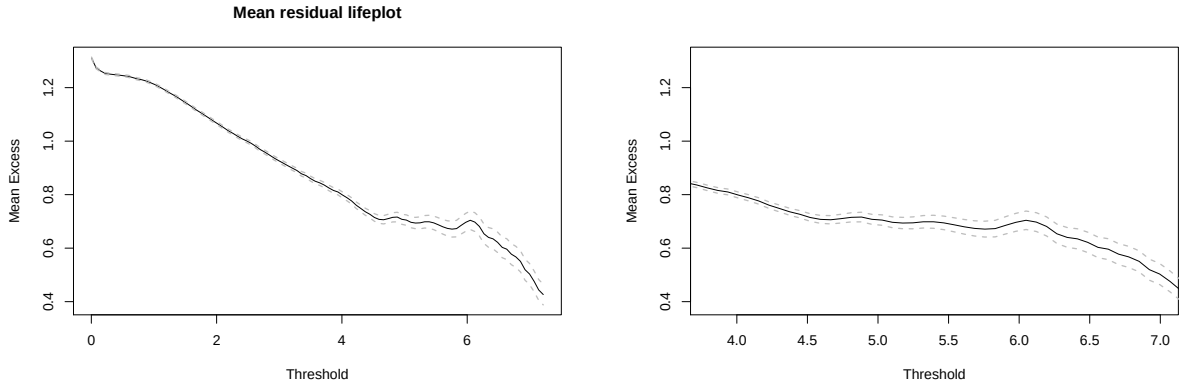


FIGURE 7 – **A gauche** Espérence des dépassements de seuils pour les hauteurs de vagues de la station 6. Les courbes en pointillées représentes les intervalles de confiance à 95%. **A droite** Agrandissement de l'échelle entre 4 et 6.

En pratique, nous cherchons un seuil u_0 à partir duquel $u \longmapsto \mathbb{E}[X - u | X > u]$ est approximativement linéaire. Au vu du grand nombre d'observations, les intervalles de confiance sont assez petits et la linéarité est visible à partir de 1, mais ce seuil est trop bas. Une investigation entre 4 et 6(A droite de la Figure 8) revèle une quasi linéarité à partir de 4.5.

Stabilité des paramètres Cette méthode est basée sur la modélisation de la GPD et suggère de choisir un seuil u pour lequel les paramètres estimés sont stable.

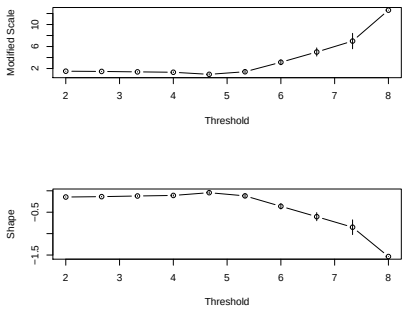


FIGURE 8 – Estimations des paramètres de la GPD en fonction du seuil u

Les deux méthodes nous mènent à choisir comme seuil $u = 4.5$ résultant en un nombre satisfaisant d'observations pour faire de l'inférence ~ 9200.

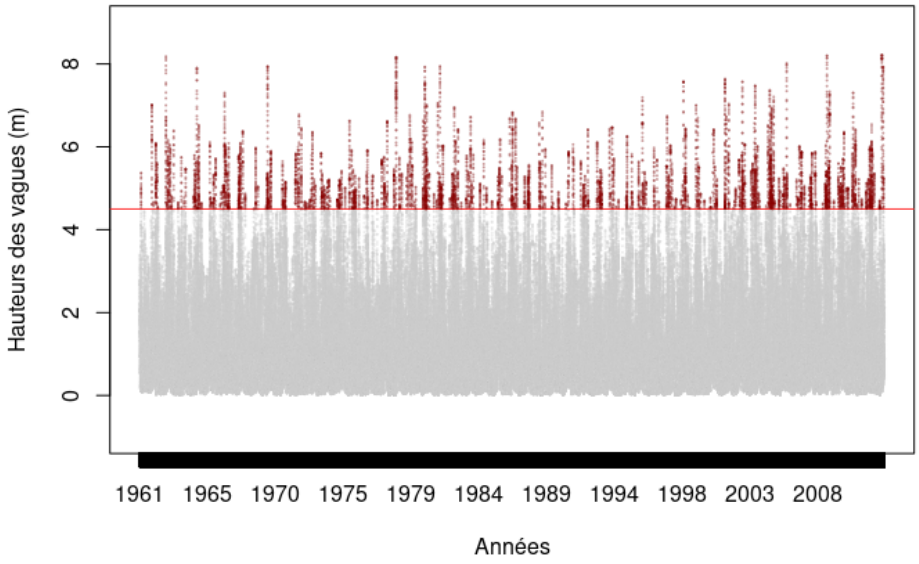


FIGURE 9 – Excès au dessus du seuil $u = 4.5$

1.2.2 Ajustement du modèle

Afin de justifier l'utilisation de la GPD, nous considérons deux observations au dessus du seuil $u > 4.5$ commes indépendantes si elles sont séparées par une période d'une semaine. Ceci repose sur l'hypothèse naive qu'une série d'évènements extrêmes comme des vents forts ou des tempêtes ne durerait pas une semaine entière. Par ailleurs, une investigation de l'auto-correlation de la série temporelles des données démontre une faible et stable auto-corrélation($\sim$0.2) à partir de 72h.

L'ajustement de la GPD en définissant le nombre de clusters à 168 observations (correspondant à une semaine), résulte en 351 clusters. Nous obtenons les estimateurs suivants.

	Borne inf.	Estimation	Borne sup.
$\hat{\sigma}$	1.033	1.20	1.383
$\hat{\gamma}$	-0.317	-0.214	-0.111

TABLE 3 – Estimateurs des paramètres de la GPD et leur intervalles de confiance à \(95%)

La vraisemblance profilée produit des intervalles de confiance similaires pour les deux paramètres.

1.2.3 Evaluation du modèle et diagnostic

Nous procédons à une évaluation graphique du modèle.

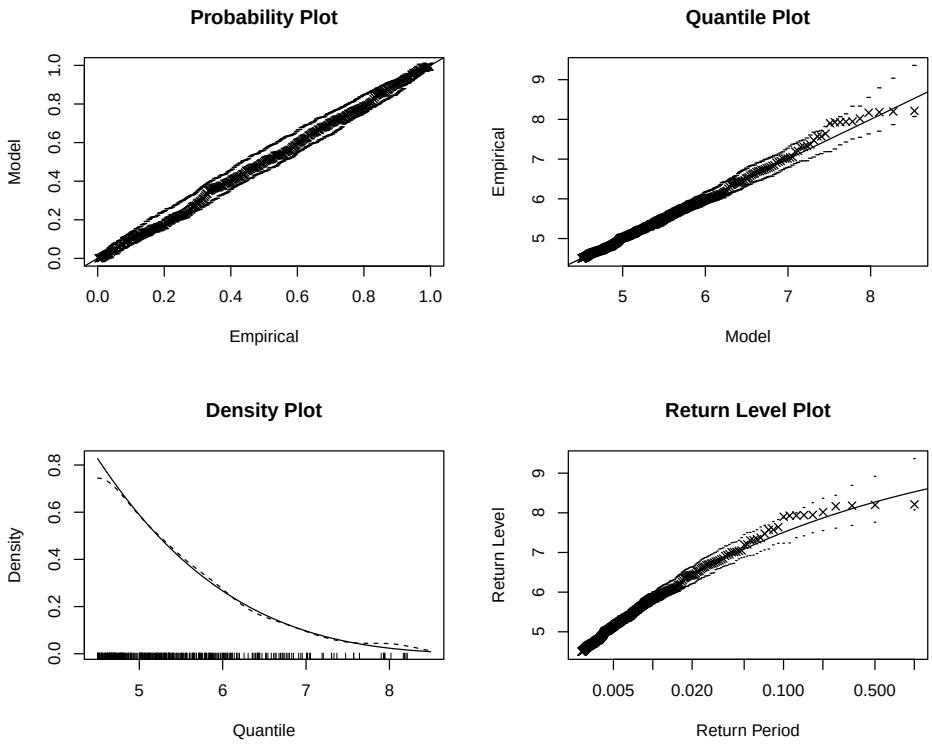


FIGURE 10 – Graphiques de diagnostic du modèle de dépassements de seuil pour la station 6.

Le modèle semble être adapté au données au regard de la Figure 10. Notons par ailleurs que l'ajustement d'une GPD en ignorant les autocorrélations de la série résulte en un paramètre de forme similaire à celui issu du modèle maxima par blocs, mais extrapole fortement les grandes hauteurs de vagues.

1.2.4 Niveaux de retour et incértitude

		Niveau de retour		
		Borne inf.	Estimation	Borne sup.
Periode de retour	100 ans	8.06	8.73	9.42
	500 ans	8.22	9.14	10.06
	1000 ans	8.26	9.3	10.3

TABLE 4 – Niveaux de retours et leurs intervalles de confiance à 95%

Les niveaux de retours suggèrent une lente augmentation et se caractérisent par des intervalles de confiance d'autant plus grand que la période de retour est longue. Notons que les bornes inférieurs collent aux données dont on dispose.

2 Etude bidimensionnelle

2.1 Etude des stations 6 et 16

2.1.1 Maxima par bloc

Dans une tentative de modéliser des blocs saisonniers, le regroupement par bloc résulte en des maximums saisonniers contenant des hauteurs de vagues communes par exemple 2 mètres pour la station 6. Ce qui correspond au quantile à 80%. Nous pouvons le remarquer visuellement au regard de la Figure 1. Cette approche, générera donc du biais dans l'estimation. Une modélisation par bloc annuel a été préferée.
En supposant que les couples d'observations d'origine sont indépendants, les maximum annuels pour chaque composante résultent en :

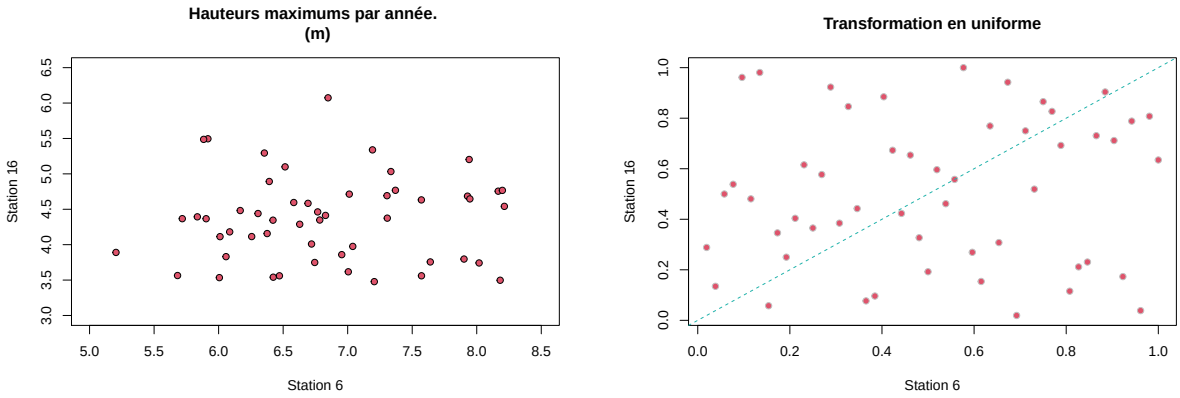


FIGURE 11 – **A gauche** Données d'origines : Hauteurs maximum annuelles jointes des vagues dans les stations 6 et 16 regroupées par **A droite** Transformations des marginales en uniforme.

Etude de la dépendance

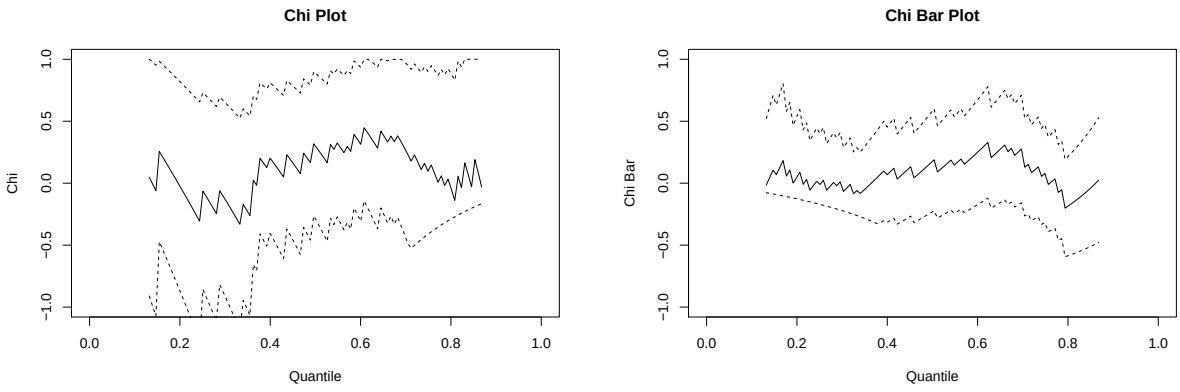


FIGURE 12 – Estimations empiriques de $\mathcal{X}(u)$ et de $\overline{\mathcal{X}(u)}$ avec leurs intervalles de confiance à 95%.

Au regard des estimations de $\overline{\mathcal{X}(u)}$ indiquent que les distributions des extrêmes pour les deux stations sont asymptotiquement indépendantes ; la valeur $\overline{\mathcal{X}} = 1$ comme limite

de $\overline{\mathcal{X}(u)}$ ne semble pas être plausible au regard de la Figure 12. Nous observons une convergence vers 0 quand $u \to 1$. D'autre par les intervalles de confiances ont une grande amplitude pour $\mathcal{X}$(qui indique par ailleurs l'indépendance), ceci est typique des modélisations du type maxima par bloc à cause du faible nombres d'observations.

Ajustement du modèle

Le choix du modèle a été fait sur la base d'un diagnostic naif de la symétrie des observations après la transformation des marginales en uniforme. En effet, au regard du graphique de droite de la Figure 11, la distribution des points a l'air symétrique. Nous optons alors pour le modèle logistique défini par :

$$G(x,y) = \exp\{-\left(x^{-1/\alpha} + y^{-1/\alpha}\right)^{\alpha}\} \quad , \quad x > 0,\ y > 0.$$

En adoptant l'approche proposée par Coles(2001) [6], les stations sont modélisées séparéement par des GEV, ensuite, le modèle bivarié est produit via EMV après la transformation des marginales en Fréchet unitaire. Les estimations des paramètres du modèle logistique sont :

	Station 6			Station 16			
	μ_x	σ_x	γ_x	μ_y	σ_y	γ_y	α
EMV	6.56	0.735	-0.257	4.11	0.522	-0.107	0.99
Ecart type	0.119	0.089	0.138	0.083	0.06	0.114	0.000002

TABLE 5 – Estimateurs du maximum de vraisemblance pour le modèle logistique ajusté.

Nous retrouvons des estimations similaires pour la stations 6 (Eq 1.1.2). Le paramètre α indique une dépendance extrêmement faible entre les occurences des extrêmes des deux stations. Une interprétation du paramètre αselon les travaux de Ledford et Tawn (1998)[7] : la probabilité d'occurence d'un extrême pour une pair d'observations converge vers $1-\alpha$ quand la taille de l'échantillon tend vers l'infini.
Notons que les stations 6 et 16 sont à 150 km de distance, l'indépendance des observations est naturelle.

D'autre part, pour le modèle logistique, le paramètre $\mathcal{X}$ peut être retrouvé par

$$\mathcal{X} = 2 - 2^{\alpha}.$$

Comme $\hat{\alpha}$ est très proche de 1, $\mathcal{X}$ est très proche de 0 indiquant une force de dépendance extremement faible.

Enfin, une alternative au choix de modèles paramétriques serait de tester plusieurs modèles et retenir celui dont le critère AIC est le plus petit.

2.1.2 Estimation du quantile extrême conditionnel

Pour la modélisation via maximum par bloc, après transformation des marginales en Fréchet, (X, Y) suit la distribution GEV logistique bivarié. Et comme la dépendance des stations est extrememement faible,

$$\mathbb{P}(X > z_p | Y > y) \sim \mathbb{P}(X > z_p).$$

Il est alors possible de retrouver le quantile z_p :

En citant Ribatet(2022) [8], si $Y \sim GEV(\mu,\ \sigma, \gamma)$, alors

$$Z = max(1 + \gamma \frac{Y - \mu}{\sigma}, 0)^{\frac{1}{\gamma}}.$$

est distribuée selon une Frechet unitaire.

Ainsi en fixant par exemple $p = 0.005$, le quantile de Frechet unitaire correspondant est

$$200.49.$$

En utilisant la transformation citée ci haut, on retrouve

$$z_p = 6.22$$

2.1.3 Un autre regard

Dans ce cas précis, nous aurions pu mener une analyse plus directe, en premier lieu l'étude de la dépendance montrerait que les observations sont indépendantes d'une station à une autre et donc qu'on pouvait analyser individuellement la station 16 dans le but d'en estimer le quantile par un modèle maxima par bloc ou dépassement de seuil.

2.2 Etude des stations 6 et 4

Nous nous intéressons maintenant à la modélisation jointe des hauteurs des vagues pour des stations proches géographiquement. Les hauteurs des vagues sont significativement différente au regard de la Figure 13.

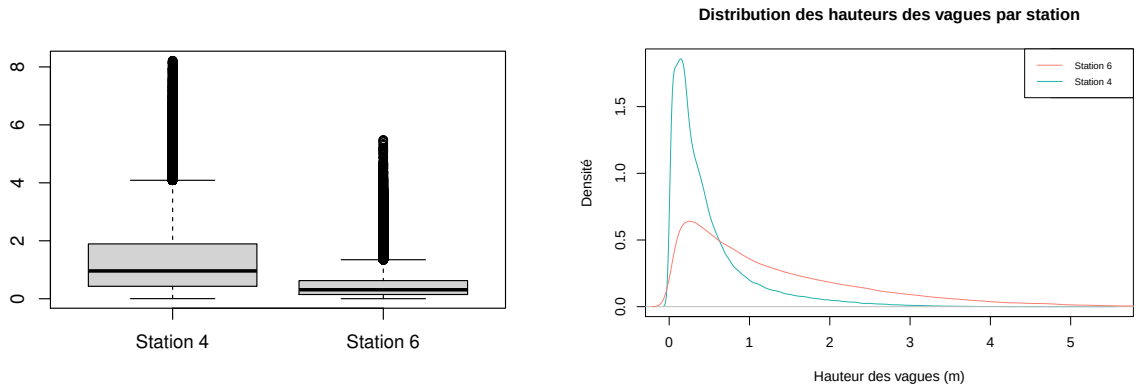


FIGURE 13 – Distribution des hauteurs de vagues enregistrées dans les stations 4 et 6. La station 4 est à $2.7km$ de la station 6.

La station 4 à 2.7 km de la station 6 peut être sujet aux mêmes vents forts et tempêtes par exemple.

2.2.1 Etude de la dépendance

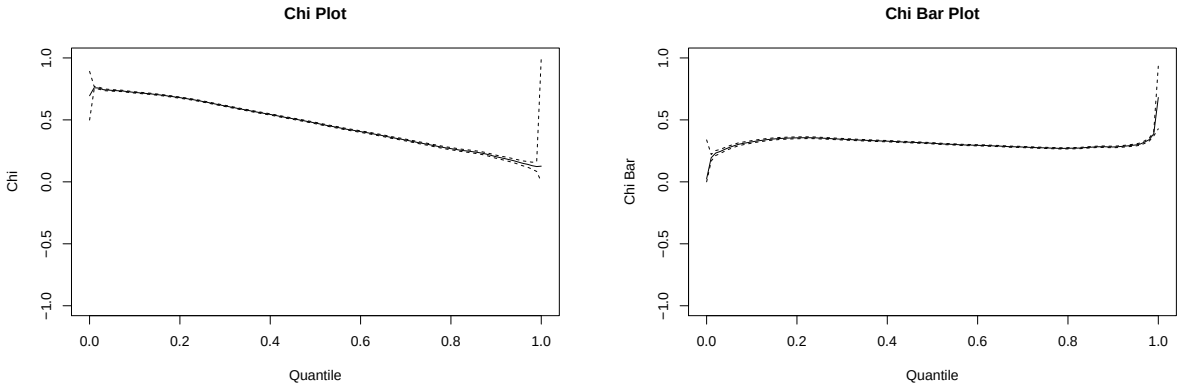


FIGURE 14 – Estimations des coefficients de dépendances $\mathcal{X}(u)$ et $\overline{\mathcal{X}}(u)$ avec leurs intervalles de confiances à 95%.

Nous observons ici que $\mathcal{X}$ étant la limite de $\mathcal{X}(u)$ quand $u \to 1$ indique par définition que les marginales sont asymptomatiquement indépendantes. Selon coles(2001) [6] Le coefficient $\overline{\mathcal{X}}(u)$ permet d'évaluer la dépendance pour les classes des distributions asymptotiquement independantes.
En particulier, pour ces deux stations, $\overline{\mathcal{X}}(u)$ est stable autour de 0.3. Cependant, autour des quantiles extrèmes, à partir de 98%, nous observons une nette augmentation indiquant la possibilité que les deux stations enregistrent des évènements extrêmes en même temps (forte tempete ...).

2.2.2 Modélisation

$\mathcal{X}(u)$ semble stable pour $u \geq 0.98$, nous choisissons alors d'utiliser comme seuils les quantiles à 98% dans notre modélisation par dépassement de seuils représentant 9111 observations pour chaque station.

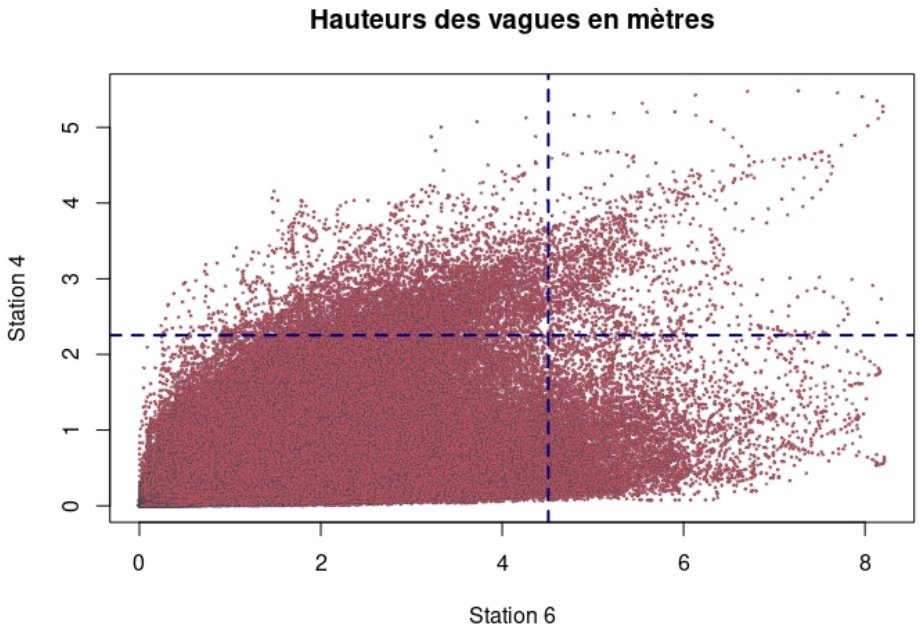


FIGURE 15 – Représentation graphique des hauteurs de vagues dans chaque station. Les courbes bleues en pointillés désignent les seuils respectifs pour chaque marginale.

Le modèle logistique et bilogistique ont été considérés avec la méthode de vraisemblance censurée. Ce sont des modèles emboités. Un test de rapport de vraisemblance résulte en une statistique de déviance largement supérieure à celle d'une $\mathcal{X}_1^2$ considérée au niveau α. Ceci suggère donc de l'asymetrie dans les données.

Nous obtenons les estimateurs suivants.

	Station 4		Station 6			
	σ_x	γ_x	σ_y	γ_y	α	β
EMV	0.751	-0.047	0.55	-0.09	0.957	0.635
Ecart type	0.011	0.010	0.007	0.008	0.0017	0.0185

TABLE 6 – Estimateurs du maximum de vraisemblance pour le modèle bilogistique ajusté.

Le nombre grand nombre d'observations résulte en de faibles intervalles de confiance.

Références

[1] R. A. Fisher et L. H. C. Tippett. « Limiting forms of the frequency distribution of the largest or smallest member of a sample ». In : *Mathematical Proceedings of the Cambridge Philosophical Society* 24.2 (1928), p. 180-190. doi : 10.1017/S0305004100015681 (page 3).

[2] Jonathan RM Hosking. « L-moments : Analysis and estimation of distributions using linear combinations of order statistics ». In : *Journal of the Royal Statistical Society : Series B (Methodological)* 52.1 (1990), p. 105-124. doi : https://doi.org/10.1111/j.2517-6161.1990.tb01775.x (page 4).

[3] Pierre Ailliot, Craig Thompson et Peter Thomson. « Mixed methods for fitting the GEV distribution ». In : *Water Resources Research* 47.5 (2011) (page 4).

[4] P Prescott et AT Walden. « Maximum likelihood estimation of the parameters of the generalized extreme-value distribution ». In : *Biometrika* 67.3 (1980), p. 723-724 (page 4).

[5] August A Balkema et Laurens De Haan. « Residual life time at great age ». In : *The Annals of probability* 2.5 (1974), p. 792-804 (page 10).

[6] Stuart Coles, Joanna Bawa, Lesley Trenner et Pat Dorazio. « An introduction to statistical modeling of extreme values ». T. 208. Springer, 2001 (pages 15, 17).

[7] Anthony W Ledford et Jonathan A Tawn. « Concomitant tail behaviour for extremes ». In : *Advances in applied Probability* 30.1 (1998), p. 197-215 (page 15).

[8] Mathieu Ribatet et Christophe Dutang. « POT : Generalized Pareto Distribution and Peaks Over Threshold ». R package version 1.1-8. 2022. url : https://CRAN.R-project.org/package=POT (page 16).

[9] Stuart G Coles et Elwyn A Powell. « Bayesian methods in extreme value modelling : a review and new developments ». In : *International Statistical Review/Revue Internationale de Statistique* (1996), p. 119-136.

[10] J. Beirlant, Y. Goegebeur, J.J.J. Segers et J. Teugels. « Statistics of Extremes : Theory and Applications ». English. Pagination : 522. Wiley, 2004. isbn : 0471976474.

[11] Janet E Heffernan et Jonathan A Tawn. « A conditional approach for multivariate extreme values (with discussion) ». In : *Journal of the Royal Statistical Society : Series B (Statistical Methodology)* 66.3 (2004), p. 497-546.

Annexe

```
library(tidyverse)
library(ismev)
library(fExtremes)
library(stargazer)
library(xtable)
library(evd)
library(extRemes)
load("donneesStations.RData")
load("donneesVagues.RData")


stations <- buoysInfos
waves <- donneesVague


# DATA PRE PROCESSING
dates <- waves[,1]
waves <- waves[!duplicated(dates),]

needed_culomns <- c("date", "station6", "station16", "station4")
df_temp <- waves[, needed_culomns]
df1 <- data.frame("month" = substr(df_temp$date, 6, 7),
                  "year" = as.numeric(substr(df_temp$date, 1, 4)),
                  "station_6" = df_temp$station6, "station_16" = df_temp$station16,
                  "station_4" = df_temp$station4)

df2 <- data.frame("date" = substr(df_temp$date, 1, 7),
                  "station_6" = df_temp$station6,
                  "station_16" = df_temp$station16)


# Density plots
dsty_s6 <- density(df1$station_6)
dsty_s16 <- density(df1$station_16)

plot(dsty_s16,
     xlab = "Hauteur des vagues (m)",
     ylab = "Densité",
     main = "Distribution des hauteurs des vagues par station",
     col = "lightseagreen")
lines(dsty_s6$x, dsty_s6$y, col = "salmon")
legend(x = "topright",
       legend = c("Station 6", "Station 16"),
       lty = c(1,1),
       col = c("salmon", "lightseagreen"), cex = 0.7)
```

```r
#########################################################
################## Univariate extremes ##################
#########################################################

#####
#######
## STATION 6
#######
#####


# Maxima per year
year_df6 = df1 %>%
  group_by(year) %>%
  summarise(station_6 = max(station_6))
year_size <- length(month_df6$date)

# Maxima per month
month_df6 = df2 %>% group_by(date) %>% summarise(station_6 = max(station_6))
month_size <- length(month_df6$date)

# Plot
g8_t <- adjustcolor("gray30", alpha.f = 0.35)
plot(year_df6$station_6, pch=21, col=1, bg=2, xaxt = "n",
     xlab = "Mois",
     ylab = "Hauteur des vagues (m)",
     main = "Hauteurs maximums par mois")
axis(1, at=1:year_size, labels=year_df6$date)
segments(1:year_size, 0, 1:year_size, year_df6$station_6, col = g8_t)




# FITTING THE GEV
# Model fit
fit_year <- gev.fit(year_df6$station_6)
fit_year$mle

# Confidence interval
round(fit_year$mle - 1.96*fit_year$se, digits = 4 )
round(fit_year$mle + 1.96*fit_year$se, digits = 4)

# Diagnosis
gev.diag(fit_year)
```

```r
gev.profxi(fit_year,xlow=-0.55,xup=0, conf = 0.90)

# Gumbel model
fit_gumbel <- gum.fit(year_df6$station_6)
gum.diag(fit_gumbel)
deviance <- 2 * (-fit_year$nllh - (-fit_gumbel$nllh))

# return level inference
gev_ret <- function(data, period){
  model_fit <- gev.fit(data)

  V <- matrix(ncol=3,nrow=3)
  V[1,1] <- model_fit$cov[1,1]
  V[2,2] <- model_fit$cov[2,2]
  V[3,3] <- model_fit$cov[3,3]
  V[1,2] <- model_fit$cov[1,2]
  V[1,3] <- model_fit$cov[1,3]
  V[2,1] <- model_fit$cov[2,1]
  V[2,3] <- model_fit$cov[2,3]
  V[3,1] <- model_fit$cov[3,1]
  V[3,2] <- model_fit$cov[3,2]

  yp <- -log(1-(1/period))

  diff1 <- 1
  diff2 <- -((model_fit$mle[3])**(-1))*(1-(yp**(-model_fit$mle[3])))
  diff3 <- ((model_fit$mle[2])*((model_fit$mle[3])**(-2))*(1-((yp)**(-model_fit$mle[3]

  del.t <- matrix(ncol=3,nrow=1)
  del.t[1,1] <- diff1
  del.t[1,2] <- diff2
  del.t[1,3] <- diff3

  del <- matrix(ncol=1,nrow=3)
  del[1,1] <- diff1
  del[2,1] <- diff2
  del[3,1] <- diff3

  A <- matrix(ncol=3,nrow=1)
  A[1,1] <- del.t[1,1]*V[1,1]
  A[1,2] <- (del.t[1,2]*V[2,2]) + (del.t[1,3]*V[3,2])
  A[1,3] <- (del.t[1,2]*V[2,3]) + (del.t[1,3]*V[3,3])

  ret.var <- (A[1,1]*diff1) + (A[1,2]*diff2) + (A[1,3]*diff3)
  ret.se <- sqrt(ret.var)
  ret.level <- (model_fit$mle[1])-((model_fit$mle[2])/(model_fit$mle[3]))*(1-(-log(1-

  result <- vector(mode ="list", length = 2)
```

```
  names(result) <- c("return_level","return_se")
  result$return_level <- ret.level
  result$return_se <- ret.se
  return(result)
}

# plot return level
g8_t <- adjustcolor("gray50", alpha.f = 0.5)
r_lvls <- c(8.54, 8.83, 8.93)
r_periods <- c(100, 500, 1000)
r_sd <- c(0.417, 0.603, 0.678)
plot(r_periods, r_lvls, type = "b", xaxt = "n",
     xlab = "Periode de retour",
     ylab = "Niveau de retour",
     main = "Niveaux de retours",pch=21, bg=2,
     xlim = c(0,1050),
     ylim = c(8, 10), col = g8_t)
axis(1, at=r_periods)
g7_t <- adjustcolor("gray30", alpha.f = 0.4)
arrows(x0=r_periods, y0=r_lvls-r_sd,
       x1=r_periods, y1=r_lvls+r_sd,
       code=3, angle=90,
       length=0.05, col =g7_t)

# inference for upper end point
mu_hat <- fit_year$mle[1]
sigma_hat <- fit_year$mle[2]
gamma_hat <- fit_year$mle[3]
end_point <- mu_hat - (sigma_hat/gamma_hat)


# LINERAR DEPENDANCE
# linear
time <- matrix(1:length(year_df6$station_6),ncol=1)
linear_gev <- gev.fit(year_df6$station_6, ydat=time, mul=1)
beta0 <- linear_gev$mle[1]
beta1 <- linear_gev$mle[2]

trend <- vector(mode = "numeric", length = length(year_df6$station_6))
trend <- beta0 + beta1*time[,1]

# Evolution over century
time2 <- matrix(1:length(year_df6$station_6),ncol=1)
time2 <- time2/100
linear_gev2 <- gev.fit(year_df6$station_6, ydat=time2, mul=1)
beta02 <- linear_gev$mle[1]
beta12 <- linear_gev$mle[2]
```

```r
# Hypothesis testing
# H_0 : Beta1 = 0 vs H_1 : Beta1 =! 0
statistic <-  abs(linear_gev2$mle[2]/linear_gev2$se[2])
critical <- 2*pnorm(statistic, lower.tail = FALSE)
critical

# rapport de vraisemblance
W <- 2*(fit_year$nllh - linear_gev2$nllh)
pchisq(W, df = 1, lower.tail = FALSE)

# plot linear trend
plot(year_df6$station_6, pch=21, col=1, bg=2, xaxt = "n",
     xlab = "Années",
     ylab = "Hauteur des vagues (m)",
     main = "Hauteurs maximums des vagues par année")
axis(1, at=1:year_size, labels=year_df6$year)
segments(1:year_size, 0, 1:year_size, year_df6$station_6, col = g8_t)
lines(1:year_size, trend, col = "lightseagreen")




# PEAKS OVER THRESHOLD

# Treshold selection
#1 - Mean residual life plot
extRemes::mrlplot(df1$station_6)
title("Mean residual lifeplot")
abline(v = 5)

#2 - Stability of parameters
ismev::gpd.fitrange(df1$station_6,2,8)

# Plot
indexes <- which(df_temp$station6 > 4.5)
excesses <- data.frame("date" = df_temp[which(df_temp$station6 > 4.5),1],
                       "station_6" = df_temp$station6[which(df_temp$station6 > 4.5)])


g5_t <- adjustcolor("gray80", alpha.f = 0.35)
plot(1:length(df1$station_6), df1$station_6, xaxt ="n",
     cex = 0.1, ylim = c(-1, 9), col = g5_t,
     xlab = "Années",
     ylab = "Hauteurs des vagues (m)")
axis(1, at=1:length(df1$year), labels=df1$year)
abline(h= 4.5, col = "firebrick1")
darkred <- adjustcolor("darkred", alpha.f = 0.2)
```

```r
points(indexes, excesses$station_6, cex = 0.1, col = darkred)

# Fitting GPD
library(evd)
library(extRemes)
GPD <- fpot(df1$station_6, 4.5, model = "gpd", cmax = T,r=168)
GPD <- fevd(df1$station_6, threshold = 4.5, type = "GP")
GPD$estimate - 1.96*GPD$std.err
round(confint(GPD, level = 0.95), digits = 3)
plot(profile(GPD,conf= 0.999))



# Return levels
# 100 years
GPD_100 = fpot(df1$station_6, 4.5,cmax = T,r=168,npp=365*24,mper=100)
GPD_100$param[1]
round(GPD_100$param[1] - 1.96*GPD_100$std.err[1], digits = 2)
round(GPD_100$param[1] + 1.96*GPD_100$std.err[1], digits = 2)

# 500 years
GPD_500 = fpot(df1$station_6, 4.5,cmax = T,r=168,npp=365*24,mper=500)
GPD_500$param
round(GPD_500$param[1] - 1.96*GPD_500$std.err[1], digits = 2)
round(GPD_500$param[1] + 1.96*GPD_500$std.err[1], digits = 2)

# 1000 years
GPD_1000 = fpot(df1$station_6, 4.5,cmax = T,r=168,npp=365*24,mper=1000)
GPD_1000$param
round(GPD_1000$param[1] - 1.96*GPD_1000$std.err[1], digits = 2)
round(GPD_1000$param[1] + 1.96*GPD_1000$std.err[1], digits = 2)



#########################################################
################## Bivariate extremes ###################
#########################################################
st_6 <- df1$station_6
st_16 <- df1$station_16

year_df16 = df1 %>% group_by(year) %>% summarise(station_16 = max(station_16))
df3 <- data.frame(year_df6, "station_16" = year_df16$station_16)
names(df3) <- c("year", "station_6", "station_16")


plot(df3$station_6, df3$station_16, pch=21, col=1, bg=2,
     xlab = "Station 6",
     ylab = "Station 16",
```

```
     main = "Hauteurs maximums par année. \n (m)",
     xlim = c(5,8.5),
     ylim = c(3,6.5))


gev_station16 <- fgev(df3$station_16)
gev_station6 <- fgev(df3$station_6)


# on transforme les marginales en Frechet :
Frechet_s16 <- qgev(pgev(df3$station_16,
                         loc = gev_station16$estimate[1],
                         scale = gev_station16$estimate[2],
                         shape = gev_station16$estimate[3]),
                    loc =1, shape =1, scale =1)

Frechet_s6 <- qgev(pgev(df3$station_6,
                         loc = gev_station6$estimate[1],
                         scale = gev_station6$estimate[2],
                         shape = gev_station6$estimate[3]),
                    loc =1, shape =1, scale =1)



plot(b1,b2 , pch=21, col=g8_t, bg=2,
     xlab = "Station 6",
     ylab = "Station 16",
     main = "Transformation en uniforme",)
abline (0 ,1, lty = 2, col = "lightseagreen")


# Parametric
fbv_log <- fbvevd(df3[,2:3], model = "log")
fbv_bilog <- fbvevd(df3[,2:3], model = "bilog",std.err = FALSE)

db <- cbind(Frechet_s6,Frechet_s16)
fit_frech <- fbvevd(db, model = "log")




library(POT)

year_df4 = df1 %>%
  group_by(year) %>%
  summarise(station_ = max(station_4))
year_size <- length(month_df6$date)
```

```r
#####
#######
## Station 4 et 6
#######
#####


# descriptive plots
plot(dsty_s4,
     xlab = "Hauteur des vagues (m)",
     ylab = "Densité",
     main = "Distribution des hauteurs des vagues par station",
     col = "lightseagreen")
lines(dsty_s6$x, dsty_s6$y, col = "salmon")
legend(x = "topright", legend = c("Station 6", "Station 4"),
       lty = c(1,1),
       col = c("salmon", "lightseagreen"), cex = 0.7)
boxplot(df1[,c(3,5)])

# tresholds
qt4 <- quantile(df1$station_4, 0.98)
qt62 <- quantile(df1$station_6, 0.98)
s64 <- data.frame(df1$station_6, df1$station_4)


# transform to uniform

b4 <- to.uniform ( s64[ ,2])
b6 <- to.uniform ( s64[ ,1])
plot(b6,b4, cex = 0.0001)
chiplot(s64)


# plot
plot(df1$station_6, df1$station_4,
     pch = 21, col = g8_t, bg = 2, cex = 0.3,
     xlab = "Station 6", ylab = "Station 4",
     main = "Hauteurs des vagues en mètres")

abline(v = qt62, col ="navy", lty = 2,lwd = 2)
abline(h = qt4, col = "navy", lty = 2, lwd = 2)

library(evd)
library(POT)
# log model
thr_64 <- c(qt62,qt4)
fit_64 <- fbvpot(x = s64, threshold = thr_64, model = "log" )
```

```r
deplph <- fit_64$estimate[5]
vev_of_z <- seq(from=2, to=5, by=0.1)
mat <- as.matrix(cbind(vev_of_z, rep(k_y, length.out = length(vev_of_z))))

probs <- pbvevd(mat, alpha =0.957, beta = 0.635,
                model = "bilog", lower.tail = FALSE )
pp <- probs/0.98
plot(pp,vev_of_z, type ="l", xlab = "p", ylab = "Quantile conditionnel z_p")

fla <- apply(-1/log(s64), 1, min)
thresh <- quantile(fla, probs = c(0.025, 0.975))
library(evd)
```

```r
############################
###### cond quant
##############################
library(texmex)
library(gridExtra)
names(s64) <- c("station_6", "station_4")
mex_64 <- mex(s64, dqu = 0.98, penalty = "none", which = "station_6")
marg <- migpd(s64, mqu=0.98, penalty="none")
mex_64 <- mexDependence(marg, which = "station_6")

# diagnostic plots
ggplot(mex_64)
mrf <- mexRangeFit(marg, "station_6", trace=11)
ggplot(mrf)
```