DOI 10.1051/LHB/2010048

Théorie statistique du renouvellement pour la détermination des houles extrêmes. Partie 2 : illustrations sur sites

The Peaks-over-Threshold statistical theory for the estimation of extreme sea-states. Part 2 : case studies

FRANCK MAZAS
Ecole des Ponts ParisTech
présentement Sogreah Consultants
6 rue de Lorraine, 38130 Echirolles, France
Tél : +33 (0)4 76 33 40 00, e-mail : franck.mazas@ponts.org

LUC HAMM
Sogreah Consultants
6 rue de Lorraine, 38130 Echirolles, France
Tél : +33 (0)4 76 33 41 88, e-mail : luc.hamm@sogreah.fr

Dans la première partie de l'article, la théorie statistique des valeurs extrêmes nous avait permis d'établir un modèle, dit Poisson-GPD, permettant une détermination rigoureuse des hauteurs significatives des états de mer extrêmes. Cependant, le caractère asymptotique de ce modèle nous avait conduits à suggérer en sus une approche multi-distributions. Nous appliquons ici cette double approche à deux sites : Haltenbanken et la Réunion (avec séparation en deux systèmes de vagues homogènes), et les outils objectifs permettant la sélection des seuils, des paramètres GPD et des meilleures lois sont introduits. Ces tests confirment tout l'intérêt de la loi GPD pour des échantillons présentant une saturation. Celle-ci n'est cependant pas observable partout ; dans ce cas l'approche multi-lois avec critères objectifs de classement s'avère nécessaire pour conclure.

In the first part of this paper, a review of the theoretical background of the peaks-over-threshold statistical theory led us to select a Poisson- general Pareto distribution (GPD) enabling a consistent estimation of significant wave heights of extreme sea-states. However, its asymptotic nature was considered and it was recommended to enlarge the approach to several other distributions. We apply here both approaches on two datasets, namely field measurements at Haltenbanken and hindcast data off La Réunion island (with separation of wind seas and swells) and we introduce rational tools to select thresholds, parameters of the GPD and to rank the different distribution fittings. The results confirm the choice of the GPD when a trend towards a saturation limit of wave heights is observed in the dataset. In the contrary, alternative distribution shall be also tested and information criterions used to select the best fitting.

I ■ INTRODUCTION

Dans la première partie de l'article, nous avons fait le point sur la théorie des statistiques des valeurs extrêmes et son application pour la détermination des plus fortes hauteurs significatives des états de mer à long terme. Nous appliquons dans cette seconde partie la méthode proposée (modèle Poisson-GPD associé à l'estimateur du maximum de vraisemblance EMV), que nous étendons ici à l'approche multi-distributions évoquée en fin de première partie, à deux séries de données : des mesures par bouée sur le site d'Haltenbanken en Norvège et des données reconstituées et ajustées par mesures satellitaires pour l'île de la Réunion. Cette dernière technique de reconstitution des états de mer joue actuellement un rôle de plus en plus important pour pallier le manque de données mesurées directement sur site.

Nous pourrons ainsi comparer les avantages et inconvénients des deux approches et identifier les difficultés auxquelles l'analyste reste malgré tout confronté.

II ■ PREMIERS TESTS SUR LE SITE D'HALTENBANKEN

● II.1 MODÈLE POISSON-GPD ET EMV

Nous traitons d'abord les données du site d'Haltenbanken, en Atlantique Nord, au large de la Norvège, qui avaient été

utilisées en 1993 pour l'établissement de recommandations internationales sur le sujet (van Vledder *et al.* (1993) [1]). Un premier test est mené en utilisant la loi GPD : nous supposons donc avoir atteint le domaine asymptotique de la théorie des statistiques extrêmes. Nous disposons d'un échantillon de N_T = 128 pics de tempêtes supérieurs à u_1 = 7 mètres sur une période de K = 9 ans (soit λ_T = 14,22).

Nous utilisons la version 1.55 du *package* extRemes [2] du logiciel et système d'analyse statistique et graphique R. Ce *package*, qui repose sur les méthodes spécifiques développées par Coles (2001) [3], permet l'analyse de données extrêmes par le biais de la loi GPD (donc par une approche du type POT) et propose des outils objectifs pour déterminer la valeur haute du double seuil : d'une part en examinant la stabilité des paramètres de forme et d'échelle k et ψ (on observe, à partir d'un certain seuil et sous réserve que la taille de l'échantillon restant demeure suffisante, une *zone de stabilité* laissant espérer qu'on se situe dans le domaine asymptotique), d'autre part en étudiant le *mean excess plot* ou *mean residual life plot*, grâce à des propriétés théoriques de la loi GPD. Nous ne détaillons pas ici le fondement théorique de ces outils [4], qui suggèrent de fixer le second seuil à u_2 = 8,57 mètres *(figure 1)* ; nous effectuerons donc l'ajustement sur un échantillon de N = 46 valeurs (soit v = 0,36 : ce logiciel permet la prise en compte de la censure).

À l'aide du paramètre de Poisson λ, également estimé par le maximum de vraisemblance et alors assimilable à la moyenne empiriqueλ_T, extRemes renvoie les résultats suivants *(figure 2)* : ψ = 1,90, k = - 0,42 et une houle centennale à 12,7 mètres avec un intervalle de confiance à 90 % de [12,2 ; 14,7]. Nous pouvons également d'ores et déjà remarquer que d'après les graphes classiques quantile-quantile et probabilité-probabilité, l'ajustement semble de bonne qualité.

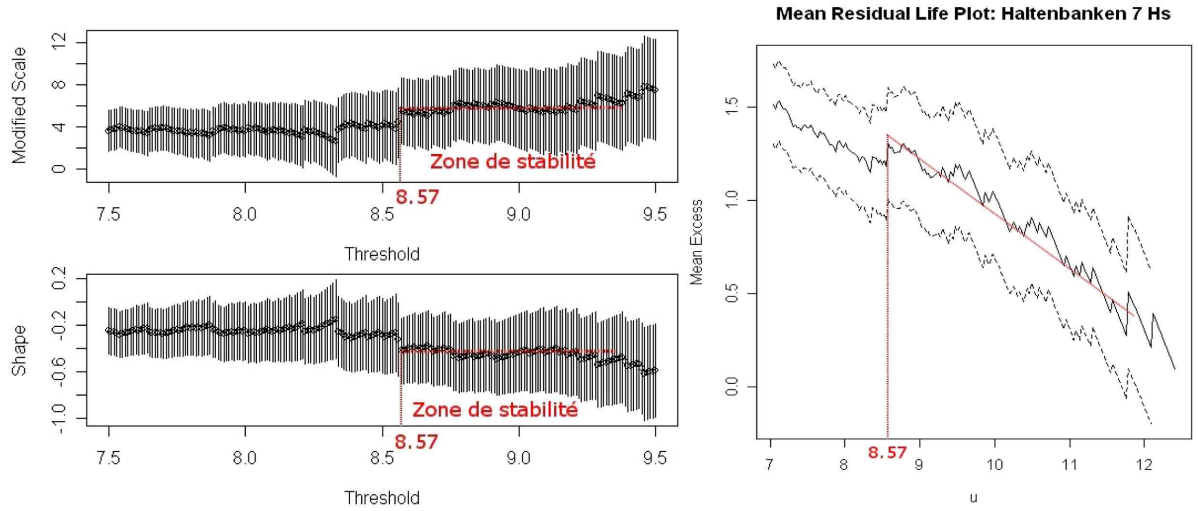

Figure 1 : Graphes extRemes pour la détermination du seuil haut de l'échantillon de Haltenbanken

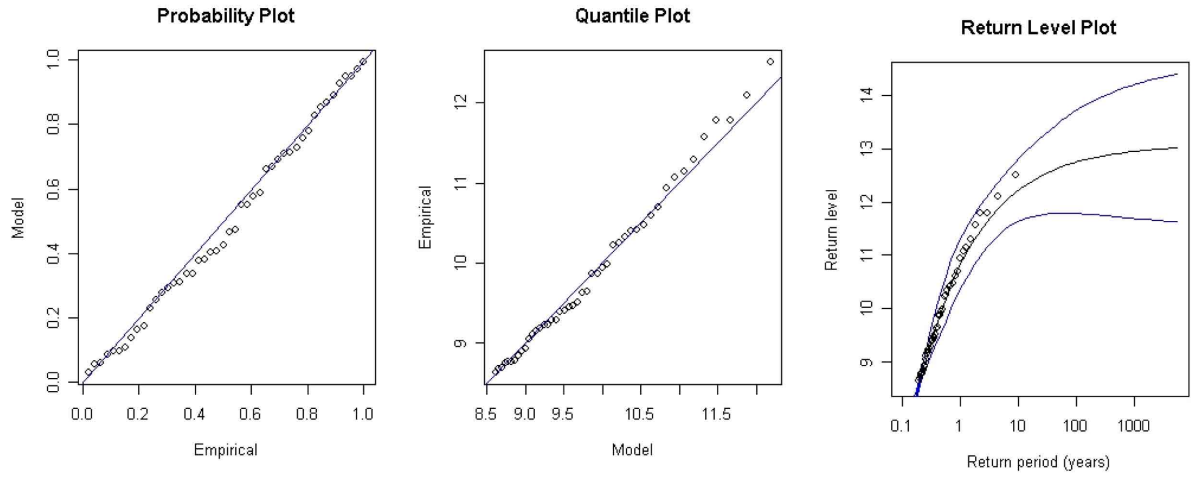

Figure 2 : Graphes extRemes pour l'ajustement GPD des données de Haltenbanken

II.2 ÉLARGISSEMENT À UN GRAND NOMBRE DE DISTRIBUTIONS

La validité de l'hypothèse précédente, à savoir que l'on se situe dans le domaine asymptotique justifiant la loi GPD, ne peut être garantie. Reprenons donc l'analyse en essayant d'ajuster à l'échantillon de nombreuses familles de distributions : GPD, Gumbel, Weibull, Gamma, exponentielle, GEV, log-Pearson III, avec le logiciel HYFRAN [5] développé par l'équipe du professeur Bobet à l'université du Québec (INRS-ETE). Notons que ce logiciel ne permet pas de prendre en compte la censure éventuelle de données.

On le voit (*fig. 3* et *4*) : difficile de privilégier une loi particulière sur la foi d'un simple examen graphique, alors même qu'on se rend compte que ces lois ont des comportements très différents au niveau des quantiles extrêmes. Il faut quantifier la qualité de l'ajustement. Pour cela, le logiciel HYFRAN inclut deux critères de comparaison : le *Bayesian Information Criterion* (BIC) [6] qui est une minimisation du biais entre le modèle ajusté et la vraie distribution inconnue, et l'*Akaike Information Criterion* (AIC) [7] qui sélectionne le modèle réalisant le meilleur compromis biais-variance. Le meilleur modèle minimise ces critères. Leurs résultats, avec les valeurs des houles centennales, les intervalles de confiance à 90 % (lorsqu'ils sont calculables) et le nombre de paramètres pour chaque loi, sont fournis par ce logiciel et sont résumés dans le *tableau 1*.

C'est bien la loi GPD qui est sélectionnée ici. En revanche, la valeur retournée est 0,9 mètre plus haute qu'avec extRemes ! L'utilisation par HYFRAN d'un estimateur différent (les moments pondérés) explique sans doute cet écart. Plusieurs points sont à relever : les lois renvoient des hou-

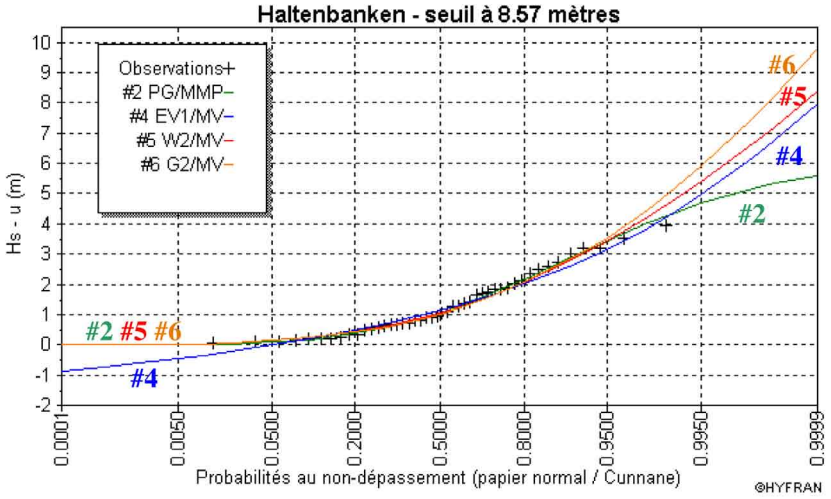

Figure 3 : Ajustement par des lois GPD (PG - #2), Gumbel (EV1 - #4), Weibull (W2 - #5) et Gamma (G2 - #6) aux données de Haltenbanken

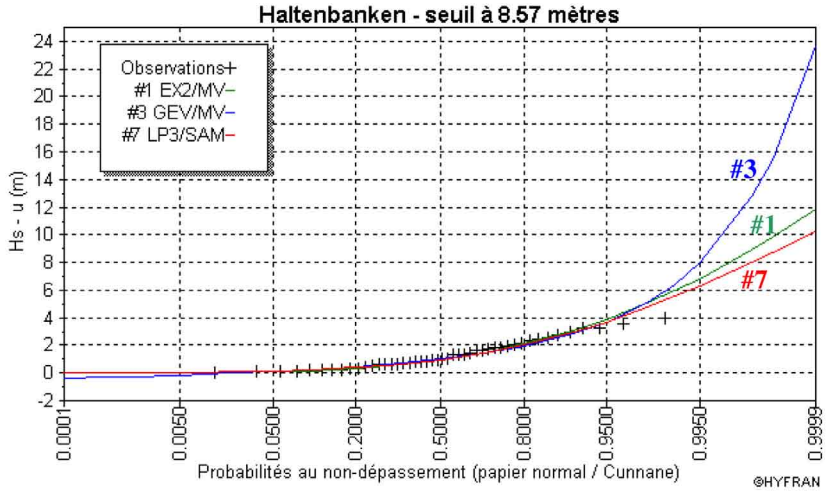

Figure 4 : Ajustement par des lois exponentielle (EX2 - #1), GEV (#3) et log-Pearson de type III (LP3 - #7) aux données de Haltenbanken

Tableau 1 : Houle centennale, intervalle de confiance, critères BIC et AIC et nombre de paramètres pour chaque loi ajustée aux données de Haltenbanken pour un seuil à 8,57 m

	GPD #2	Weibull #5	Gamma #6	Exp. #1	LP III #7	Gumbel #4	GEV #3
$H_{100\ ans}$	13,6	14,7	15,4	16,6	15,9	14,3	19,1
IC 90 %	-	12,9-16,5	13,6-17,3	14,7-18,6	-	13,3-15,3	-
BIC	120,941	121,962	122,427	122,815	126,897	130,949	133,057
AIC	117,283	118,304	118,770	119,157	121,412	127,292	127,371
K_i	2	2	2	2	3	2	3

les centennales très différentes ; les critères BIC et AIC se rejoignent pour fournir le même classement ; et les lois à trois paramètres sont plus biaisées, puisqu'un paramètre, lui-même estimé, apporte sa propre incertitude. L'exception de Gumbel semble due à son mauvais ajustement sur les plus petites valeurs.

III ■ TESTS SUR LE SITE DE LA RÉUNION

III.1 PRÉSENTATION DU SITE

L'île de la Réunion est soumise à un climat tropical complexe qui l'expose à de fortes houles de trois grands types : la houle cyclonique, la houle d'alizé et la houle australe. L'activité des alizés de sud-est est maximale lors de la saison fraîche, de mai à octobre, pendant laquelle ceux-ci soufflent régulièrement dans le secteur S-SE. La houle australe est elle générée par les tempêtes australes très violentes de l'extrême sud de l'Océan Indien. Leur intensité est maximale lors de la saison fraîche. La houle résultante se propage donc sur 3 000 kilomètres jusqu'à la Réunion qu'elle touche par le Sud. La très grande période de ces vagues (une vingtaine de secondes) caractérise leur très grande énergie. À proximité de l'île, elles vont gonfler et déferler sur le rivage, pouvant causer de très importants dégâts.

Les données utilisées sur ce site ne sont pas des données mesurées par bouée, mais reconstituées et ajustées par mesures satellitaires. En toute rigueur, une sélection précise des tempêtes nécessite un examen systématique de chaque événement sur la base d'une analyse météorologique détaillée. En pratique, une automatisation de cette sélection a permis de séparer la composante d'alizé de la composante australe. D'autres systèmes de vagues interviennent bien sûr ; on constate cependant qu'en moyenne, plus de 98 % de la hauteur significative des vagues résulte des composantes australe et d'alizé. L'intérêt de cette séparation est de mener à bien l'analyse sur deux échantillons bien plus homogènes. L'expérience montre que l'on peut ainsi espérer obtenir des estimations plus fiables et précises pour chacun des échantillons homogénéisés.

III.2 MODÈLE POISSON-GPD ET EMV

Cette fois-ci, les outils de détermination du seuil haut suggèrent de fixer u_2 = 3,0 mètres pour la houle d'alizé et u_2 = 2,0 mètres pour la houle australe *(figure 5)*.

Nous obtenons *(figure 6)* une houle d'alizé centennale à 4,8 mètres avec un intervalle de confiance à 90 % de [4,6 ; 5,3] et une houle australe centennale à 4,8 mètres également avec un intervalle de confiance à 90 % de [4,5 ; 5,3]. extRemes permet également de visualiser le profil de la courbe de log-vraisemblance dans le domaine de l'intervalle de confiance (cf. *figure 7*).

On observe sur l'échantillon de la houle australe un phénomène fréquent dans une telle analyse : deux *outliers*, c'est-à-dire des mesures ou des observations très supérieures aux

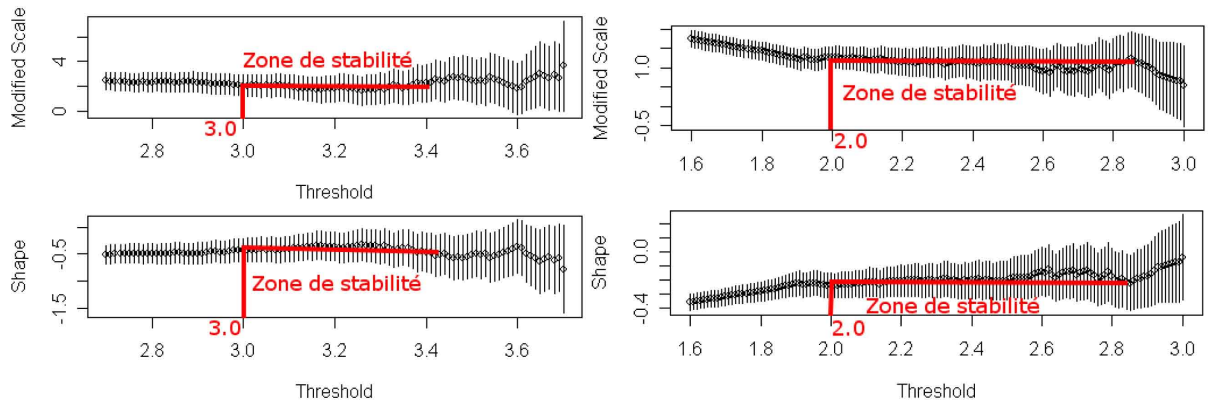

Figure 5 : Graphes extRemes pour la détermination du seuil haut des échantillons de la Réunion : houle d'alizé (gauche) et houle australe (droite)

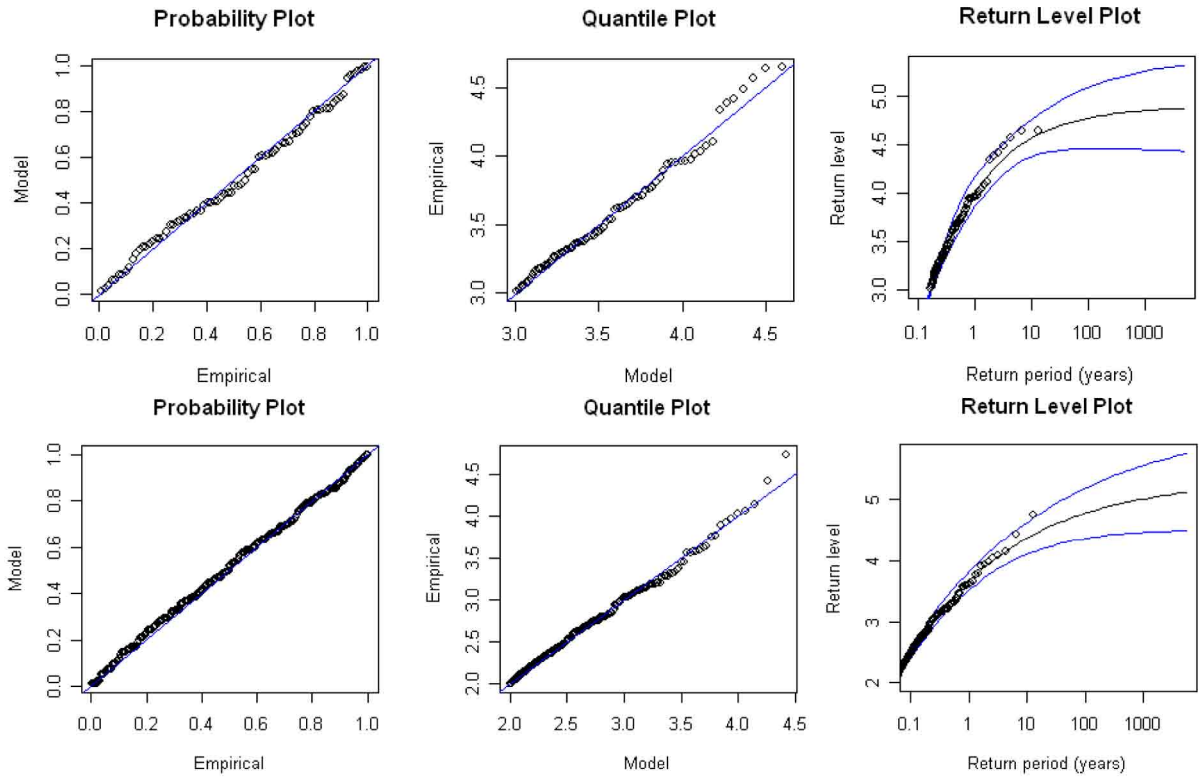

Figure 6 : Graphes extRemes pour l'ajustement GPD des données de houle d'alizé (haut) et australe (bas)

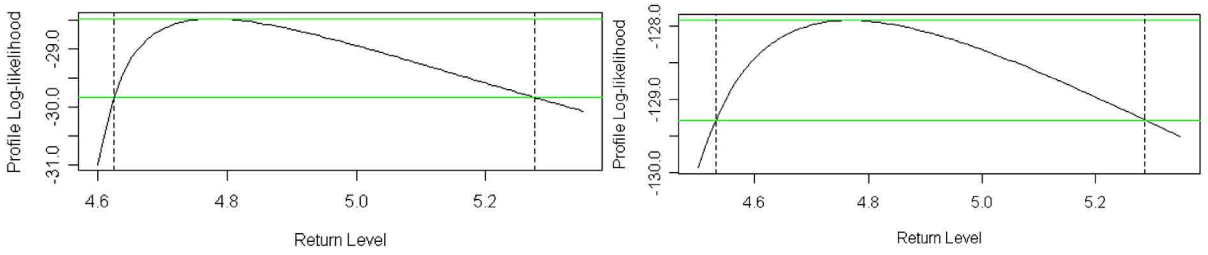

Figure 7 : Détermination des intervalles de confiance pour houle d'alizé (gauche) et australe (droite)

autres, causées soit par une erreur de mesure, soit par une tempête de période de retour grande devant la période de mesure ou d'observation, soit par un événement météorologique très différent des autres, comme une houle cyclonique. La présence d'un *outlier* doit mener à effectuer une analyse météorologique spécifique, en relation avec les limites physiques que l'on peut attendre de la houle sur le site étudié.

Le graphe de la valeur de retour en fonction de la période de retour montre la robustesse de l'estimateur LME : l'ajustement n'est guère « tiré vers le haut », contrairement à la méthode des moindres carrés.

III.3 ÉLARGISSEMENT À UN GRAND NOMBRE DE DISTRIBUTIONS

Reprenons une analyse multi-distributions, comme précédemment. On présente ici uniquement *(figure 8)* les lois s'adaptant le mieux aux échantillons, c'est-à-dire les lois GPD, Gamma, exponentielle, de Weibull et log-Pearson de type III (LP III). Là encore, les critères BIC et AIC aident à sélectionner le meilleur modèle *(tableau 2)* :

Pour la houle d'alizé, la loi GPD est celle qui s'ajuste le mieux à l'échantillon, et elle renvoie une valeur très proche de l'approche précédente. Ici, comme pour les mesures d'Haltenbanken, on observe un replat correspondant à un phénomène de saturation, qu'on peut instinctivement rapprocher du fait que la hauteur des vagues est physiquement limitée. Lorsque cette saturation en H_s est observée, la loi GPD, qui donne la plupart du temps des valeurs de retour plus faibles, est généralement celle qui convient le mieux.

En revanche, la loi de Weibull ajuste mieux l'échantillon de houle australe. Graphiquement, l'ajustement par les lois de Weibull et de Gamma semble avoir été plus perturbé par les deux valeurs maximales, sans doute des *outliers*, et la

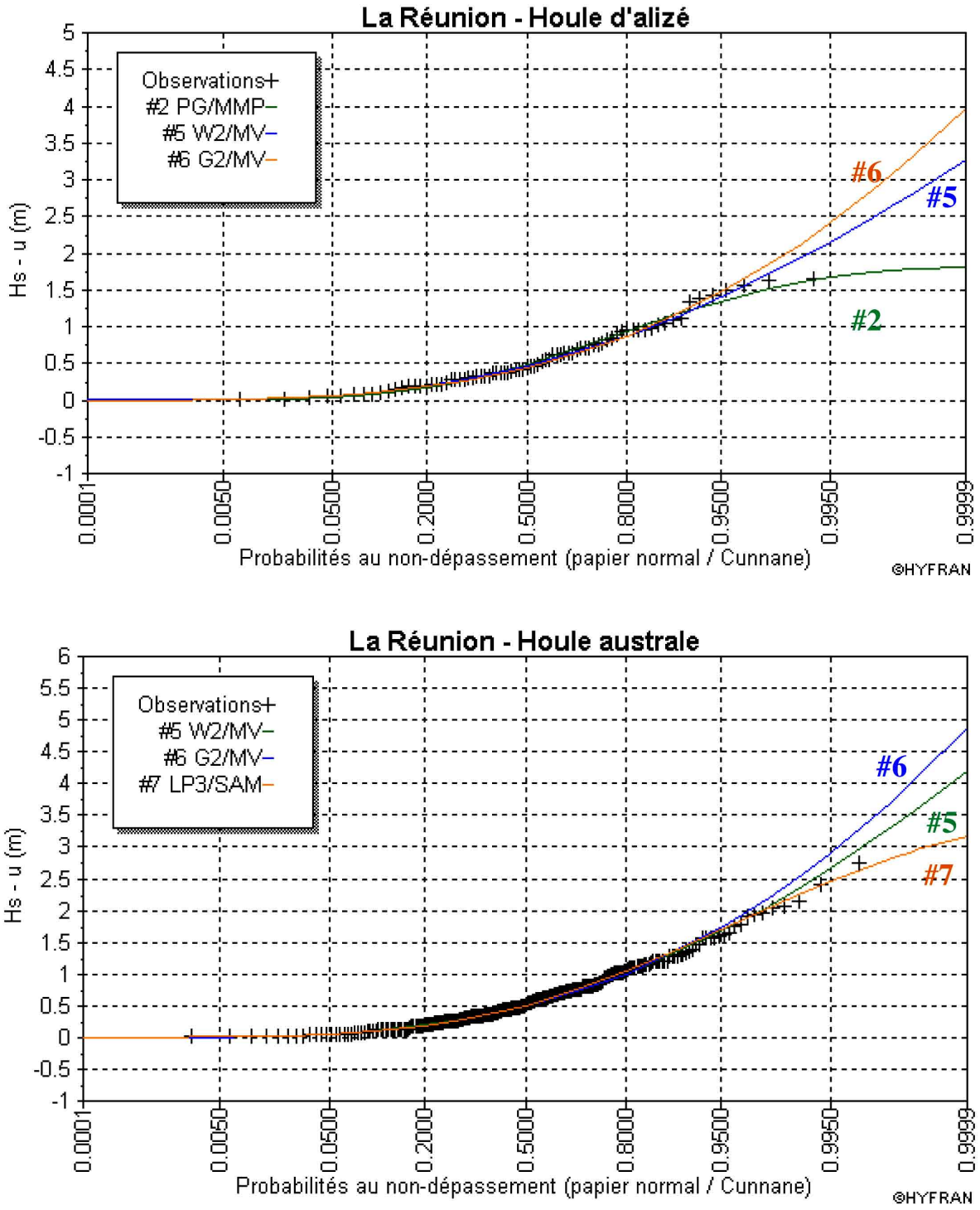

Fig. 8 : Ajustement par des lois GPD (PG - #2), de Weibull (W2 - #5), Gamma (G2 - #6) et log-Pearson de type III (LP3 - #7) aux données de la Réunion

loi LP III paraît meilleure. Elle donne d'ailleurs un résultat très proche de la loi GPD avec extRemes, mais est sans doute désavantagée par son paramètre supplémentaire. Au final, les deux approches divergent de 0,8 m ! Ici, c'est un peu le « flair » de l'analyste qui départage les deux approches ; dans ce cas précis on aurait tendance à privilégier la loi GPD donnée par extRemes. On voit d'ailleurs qu'avec l'analyse HYFRAN, les critères BIC et AIC intervertissent parfois le classement de certaines lois. Au final, nous avons ici l'exemple de deux échantillons moins « propres » qu'à

Tableau 2 : Houle centennale, intervalle de confiance, critères BIC et AIC et nombre de paramètres pour chaque loi ajustée aux données de la Réunion

	Houle d'alizé				Houle australe			
	GPD #2	Weibull #5	Gamma #6	LP III	Weibull #5	Gamma #6	LP III #7	Exp.
$H_{100\ ans}$	4,7	5,5	5,9	5,2	5,6	6,1	4,9	6,8
IC 90 %	-	5,0 – 6,0	5,3 – 6,4	-	5,1 – 6,1	5,6 – 6,6	-	6,3 – 7,3
BIC	65,794	67,213	68,813	70,702	267,076	270,257	272,385	275,242
AIC	61,055	62,474	64,074	63,594	260,073	263,255	261,882	268,240
K_i	2	2	2	3	2	2	3	2

Haltenbanken, et plus difficiles à analyser. En outre, nous avons fait ici l'hypothèse que le seuil optimal au sens de la loi GPD l'est également pour les autres lois. Cette hypothèse nécessite sans doute d'être confortée théoriquement.

III.4 RECOMBINAISON DES ÉCHANTILLONS HOMOGÉNÉISÉS

Il reste maintenant à recombiner ces différents résultats de façon à obtenir des prévisions sur les hauteurs de houle totales. Cela est loin d'être trivial : c'est une source de recherches actuellement.

Reprenons l'exemple de la Réunion. On cherche donc à déterminer si une corrélation entre la houle d'alizé et la houle australe peut être mise en évidence. On observe que la plupart du temps, les deux composantes sont fortement présentes, avec un minimum de 0,65 m de houle d'alizé et de 0,4 m de houle australe *(figure 9)*.

Appliquons ce résultat au cas particulier du phénomène qui a touché la côte sud-ouest de l'île, et particulièrement St-Pierre, les 13 et 14 mai 2007 : une forte houle australe, avec des hauteurs significatives allant jusqu'à 5 à 6 mètres (soit des hauteurs maximales de 10 à 12 mètres), a déferlé sur ces rivages, causant deux morts. La *figure 10* donne les relevés de la bouée omnidirectionnelle de St-Pierre.

Le modèle Poisson-GPD donne une houle australe centennale à 4,8 mètres, avec un intervalle de confiance à 90 % de [4,5 ; 5,3]. On est manifestement en présence d'un phénomène dont la période de retour est de quelques dizaines d'années, ce que confirment d'ailleurs les « anciens » de l'île. Mais s'il est évident que la houle australe est bien le phénomène dominant, on peut se demander quel est le poids des autres composantes.

On aurait besoin pour conclure sur la période de retour du phénomène de pouvoir séparer les composantes, ce que les relevés de la bouée ne permettent pas. On sait cependant que

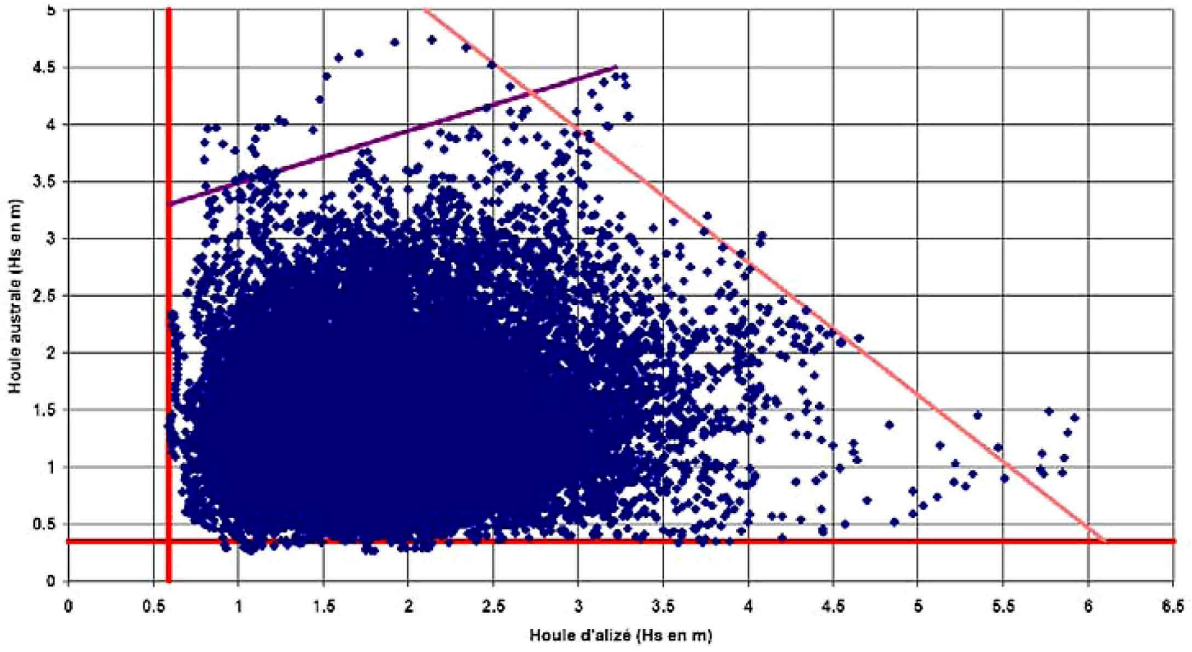

Figure 9 : Corrélation houle australe / houle d'alizé à la Réunion

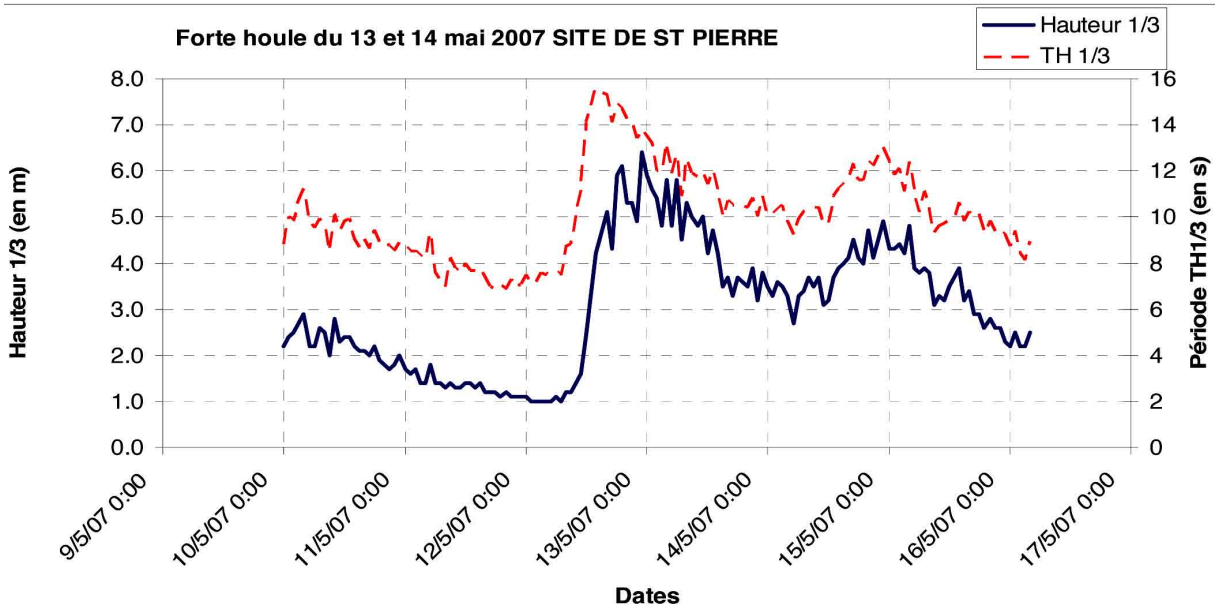

Figure 10 : forte houle à Saint-Pierre de la Réunion les 13 et 14 mai 2007

la veille, les alizés soufflaient très peu. On peut alors envisager raisonnablement un niveau bas de la composante d'alizé autour d'1 mètre, soit une composante australe autour de 5 mètres, ce qui donne des périodes de retour raisonnables, de l'ordre de quelques dizaines d'années.

La séparation des composantes d'un état de mer permet ainsi de travailler sur des échantillons beaucoup plus homogènes et de réduire ainsi les incertitudes. Cependant, la recombinaison des résultats pour obtenir les valeurs de retour globales est beaucoup plus difficile et fait encore l'objet d'études. La recherche de corrélation entre les systèmes de vagues paraît nécessaire.

IV ■ CONCLUSIONS

Les tests présentés ici sur Haltenbanken et la Réunion illustrent les méthodes statistiques rigoureuses décrites dans la première partie de l'article. Les résultats obtenus restent cependant en demi-teinte. Ainsi, sur le site d'Haltenbanken, les deux approches convergent vers la même loi d'ajustement (loi GPD) mais fournissent des hauteurs significatives centennales assez éloignées, qui peuvent refléter la sensibilité d'un échantillon de faible durée (neuf ans) au choix de l'estimateur. Sur l'île de La Réunion, il a d'abord été nécessaire de séparer les deux systèmes de vagues pour bien homogénéiser les échantillons. Pour les houles d'alizé, les deux approches convergent vers la même loi et la même estimation. Ce n'est pas le cas de la houle australe, ce qui justifie l'utilisation d'une approche multi-lois, *a priori* plus sûre. La recombinaison de ces deux résultats s'avère cependant un peu délicate et nécessite une étude détaillée de corrélation, qui reste à faire.

Plus généralement, on constate en pratique que la loi GPD est souvent très adaptée en présence d'un phénomène de saturation (replat). Sinon, l'approche multi-lois est judicieuse en complément. En l'absence d'outils permettant la détermination d'un seuil de censure spécifique à chaque loi, l'utilisation du seuil u_2 optimisé pour la loi GPD est recommandée.

V ■ RÉFÉRENCES

[1] Van Vledder G., Goda Y., Hawkes P., Mansard E., Martin M.J., Mathiesen M., Peltier E., Thompson E. (1993) — Case studies of extreme wave analysis : a comparative analysis. *Proc. Second symposium on ocean wave measurements and analysis, New-Orleans, Louisiana, ASCE*. 978-992

[2] Gilleland E., Katz R., Young G. (2004) — *The extRemes Package. http ://cran.r-project.org/doc/packages/extRemes.pdf.*.

[3] Coles S. (2001) — *An introduction to statistical modeling of extreme values. Springer-Verlag, London.*

[4] Smith R.L. (2001) — *Environmental statistics. Department of Statistics, University of North Carolina. URL : http ://www.stat.unc.edu/postscript/rs/envnotes.ps.*.

[5] Bobee B., Fortin V., Perreault L., Perron H. (1999) — *HYFRAN 1.0 (logiciel hydrologique : Chaire en hydrologie statistique CRNSG/Hydro-Québec), INRS-Eau, Terre et Environnement, Université du Québec, URL : http ://www.inrs-ete.uquebec.ca/activites/groupes/chaire_hydrol/hyfran.html.*

[6] Schwarz G. (1978) — Estimating the dimensions of a model. *Annals of statistics*. **6** 461-464

[7] Akaike H. (1973) — Information theory as an extension of the maximum likelihood principle. *Second International Symposium on Information Theory. Akademiai Kiado, Budapest*. 267-281