Réconcilier théorie et pratique dans la détermination des houles extrêmes

Luc Hamm*, Franck Mazas, Nicolas Garcia*, Benjamin Bailly***

** SOGREAH Maritime,*
6 rue de Lorraine,
38130 Échirolles, France.
luc.hamm@sogreah.fr

*** Ecole des Ponts ParisTech.*
franck.mazas@ponts.org

RÉSUMÉ. *Le but de cet article est d'améliorer les méthodes statistiques actuelles de détermination des houles extrêmes en proposant des solutions justifiées théoriquement mais aussi applicables en pratique tout en pointant la nécessité pour les praticiens de construire des bases de données solides (statistiquement parlant) en préalable. La méthode du renouvellement (ou POT) est préconisée ici et des outils objectifs de détermination du seuil sont présentés. Le choix de la loi statistique est discuté ; la loi GPD est soulignée et une approche multi-lois justifiée. L'ajustement par l'estimateur du maximum de vraisemblance est fortement recommandé. Des tests sont menés sur le site d'Haltenbanken pour illustrer les améliorations proposées. Enfin, une application pratique dans le détroit de Gibraltar est présentée en détail.*

ABSTRACT. *This article aims to improve the current statistical methods for the determination of extreme wave heights. It proposes both theoretically justified and user-friendly solutions and insists on the necessity for practitioners to build reliable wave databases. Use of the POT method is advocated and objective tools for threshold determination are presented. The choice of the statistical law is discussed; the GPD law is stressed and a multi-law approach is justified. The adjustment by the likelihood maximum estimator is strongly recommended. Tests were conducted on the site of Haltenbanken to illustrate the proposed improvements. Finally, a practical application in the Gibraltar straight is described in details.*

MOTS-CLÉS : *Valeurs extrêmes – houle – POT – EMV.*

KEYWORDS: *Extreme values - significant wave height – POT - LME.*

1. Introduction

Prévoir la hauteur significative des états de mer extrêmes pour des périodes de retour élevées (de l'ordre de cent ans) est primordial pour le dimensionnement des ouvrages portuaires sur un site maritime mais relève de la gageure. « *Prediction is very difficult, especially about the future* », disait Niels Bohr. Pourtant, les méthodes statistiques développées depuis quelques décennies ambitionnent d'offrir à l'analyste des outils objectifs.

La méthode la plus répandue actuellement a été proposée par le Professeur Goda (Goda, 1988b ; Goda & Kobune, 1990 ; Goda, 2000). Elle a été largement reprise par le Groupe de Travail sur les Statistiques des Houles Extrêmes créé au sein de l'AIRH (Association Internationale pour la Recherche Hydraulique) dans son document de synthèse (Mathiesen *et al.*, 1994) et tout récemment dans le *Guide Enrochement* du CIRIA-CUR-CETMEF (2009). Devant les difficultés à concilier théorie statistique et pratique de l'ingénieur, elle se veut globale et relativement légère à mettre en œuvre. Un rapide tour d'horizon international (Franco & Piscopia, 2004 ; Thompson, 2002) montre d'ailleurs qu'elle reste très suivie.

Nous avons choisi ici d'aller plus loin et tenté notamment d'introduire dans le monde de l'ingénierie côtière des méthodes plus fines du point de vue de la théorie des statistiques tout en cherchant à établir clairement les limites d'une telle tentative. Nous examinerons donc dans ce papier la question du choix des distributions statistiques à ajuster aux données de tempêtes, la méthode d'ajustement adéquate et les outils objectifs de détermination de seuil et de choix de la meilleure distribution. Nous validerons notre nouvelle approche par rapport à un jeu de données classique dans le domaine et présenterons en détail un exemple pratique issu de notre expérience récente dans lequel des mesures de bouées et des reconstitutions historiques d'états de mer sont utilisées pour définir des hauteurs significatives d'états de mer extrêmes dans le détroit de Gibraltar.

2. Traitement de l'échantillon

2.1 Choix du type de jeux de données

L'ingénieur analyste travaille à partir d'échantillons de données environnementales, réelles ou simulées, comme ici la hauteur significative des vagues. Il existe alors trois approches de ces jeux de données : celle de l'échantillon complet (*total sample method*) qui ajuste une distribution statistique à toutes les données collectées, la méthode des *block maxima* qui n'analyse que les valeurs maximales sur un intervalle de temps donné, souvent un an (on parle alors des maxima annuels) et enfin la méthode du renouvellement ou méthode POT (*peaks-over-threshold*). Cette méthode ne retient que les valeurs maximales des épisodes de tempêtes, grâce à la fixation d'un seuil (*threshold*).

Un échantillon statistique devant réunir des conditions d'*indépendance* et d'*homogénéité*, c'est-à-dire être identiquement distribué (on le qualifie alors d'indépendant et identiquement distribué, i.i.d.), la plupart des analystes rejettent la première méthode. La deuxième méthode a l'inconvénient d'écarter des valeurs qui apportent une information valorisante, information au contraire recueillie par la méthode POT. Aussi retiendrons-nous cette dernière méthode, conformément aux conclusions du Groupe de Travail de l'AIRH.

2.2 Censure des données et double seuil

Les tempêtes retenues par cette méthode sont d'intensités très diverses, si le seuil est assez bas. Cette constatation n'est pas anodine : les faibles tempêtes peuvent en effet distordre l'ajustement à une distribution en apportant trop de poids aux faibles valeurs de pics, donc en introduisant un biais négatif. Cependant, elles apportent une information valorisante sur les fréquences d'apparition qu'il est bon de prendre en compte. Dans ce cas, on applique alors un *processus de censure* (Goda, 1988) : un seuil bas permet de sélectionner toutes les tempêtes (de quantité N_T) alors qu'un seuil plus haut, dont la détermination est essentielle, retient les N plus hauts pics auxquels on ajustera la loi. Nous appellerons ce doublet seuil bas – seuil haut, *double seuil*. On peut ainsi définir un *paramètre de censure* ν correspondant au rapport N sur N_T. L'intérêt d'un tel processus de censure a été souligné par le Groupe de Travail, même si la théorie montre que son influence est asymptotiquement nulle. Notre expérience montre d'ailleurs que la prise en compte de ce processus ne fait que très faiblement varier les résultats. Nous verrons plus loin que l'estimateur du maximum de vraisemblance permet de le traiter correctement.

3. Analyse rigoureuse de l'échantillon

3.1 Un peu de statistique des extrêmes

Considérons un échantillon de variables aléatoires réelles, indépendantes et identiquement distribuées. On s'intéresse aux valeurs extrêmes, ici aux maxima, d'un tel échantillon. Jenkinson (1955) a généralisé les résultats de Fréchet (1927) et Fisher & Tippett (1928) en montrant que la loi du maximum de l'échantillon tend vers la *loi généralisée des valeurs extrêmes* (GEV, *Generalized Extreme Value distribution*), qui a trois paramètres x_0, ψ et k, respectivement appelés paramètres de localisation, d'échelle et de forme. Sa fonction de répartition est donnée par l'équation [1]. Le cas $k > 0$ correspond à la loi de Fréchet ; le cas $k < 0$ à la loi de Weibull ; enfin, en faisant tendre k vers 0, on obtient la loi de Gumbel par passage à la limite.

$$F_{x_0,\psi,k}(x) = \exp\left[-\left(1+k\frac{x-x_0}{\psi}\right)^{-\frac{1}{k}}\right] \quad [1]$$

On s'intéresse à présent à la loi régissant le dépassement d'un seuil u au sein d'un échantillon, soit l'approche de la méthode POT. Soit X une variable aléatoire réelle de fonction de répartition F, u le seuil fixé et posons $Y = X - u$ sous condition que $X > u$. (Y est donc la variable aléatoire représentant le dépassement du seuil u par la variable X. Dans cette étude, X est la variable aléatoire représentant la hauteur significative de la houle.) Pickands (1975) a montré que lorsque u approche le point terminal de l'échantillon (c'est-à-dire une valeur, finie ou infinie, dont la probabilité de dépassement est nulle), la loi des dépassements de u peut être approchée par la *distribution généralisée de Pareto* (GPD, *Generalized Pareto Distribution*) donnée par [2] :

$$F_{\psi,k}(y) = 1-\left(1+k\frac{y}{\psi}\right)^{-\frac{1}{k}} \quad [2]$$

Cette approximation se justifie pour une taille d'échantillon assez grande, et pour un seuil u assez élevé. De même que pour la loi GEV, ψ et k sont appelés paramètres d'échelle et de forme car ils déterminent respectivement l'échelle linéaire et la forme fonctionnelle de la distribution. Le cas $k = 0$ (par passage à la limite) correspond à la distribution exponentielle d'espérance ψ. (On rappelle que $y = x – u$).

Le nombre N_1 de dépassements du seuil u (c'est-à-dire le nombre de tempêtes considérées) dans une année pouvant être considéré comme régi par un processus poissonien, on suggère le modèle suivant, appelé *modèle Poisson-GPD* et défini ainsi : les dépassements de seuil obéissent à une loi GPD et sont i.i.d., et N_1 suit une loi de Poisson.

3.2 Choix des distributions candidates

La théorie dit ainsi que la loi correspondant à des échantillons POT est la loi GPD. Dans une analyse simple, c'est donc bien cette loi qu'il s'agit d'utiliser, et non celle de Gumbel ou de Weibull comme recommandé par le Groupe de Travail. Il est alors pertinent de mettre en place un modèle Poisson-GPD.

Mais on peut (doit ?) approfondir l'analyse. En effet, la théorie des valeurs extrêmes est certes très séduisante, mais il est primordial de garder à l'esprit son caractère *asymptotique*. Pour que son utilisation soit vraiment pertinente, il faut des échantillons de taille beaucoup plus grande que ce dont on dispose habituellement, c'est-à-dire quelques dizaines d'années. En outre, nous n'avons aucune information sur la vitesse de convergence de la loi de l'échantillon vers ces lois asymptotiques : or rien ne garantit qu'elle ne soit pas très faible.

En conséquence, les lois des valeurs extrêmes (GEV, GPD) sont bien des candidates privilégiées pour modéliser les valeurs maximales et/ou les dépassements de seuil d'un échantillon. Mais la taille de ces échantillons comme la gamme des probabilités considérées dans les applications hydrologiques et maritimes font que d'autres distributions (Weibull, Gumbel, log-normale, log-Pearson de type III, Gamma…) peuvent *a priori* fournir une meilleure modélisation. Une analyse plus approfondie devra donc utiliser en principe plusieurs distributions candidates, car bien que beaucoup plus lourde, c'est l'approche la plus justifiable. La restriction à deux lois (Weibull à 3 paramètres et Gumbel) ne nous paraît donc pas devoir être maintenue en l'état, comme cela a déjà été évoqué par Goda (2000) et Thompson (2002). Une réflexion sur les domaines d'attraction pour les maxima de ces lois aide cependant à restreindre le choix (Castillo & Sarabia, 1992). En effet, une distribution ayant un point terminal fini ne peut se trouver dans le domaine d'attraction de Fréchet. Or, la hauteur des vagues est physiquement limitée. On peut donc exclure de l'étude les lois appartenant à ce domaine d'attraction, comme les lois de Cauchy, de Pareto, de Student, du χ^2 ou les lois α-stables.

3. 3 Méthode d'ajustement

Pour réaliser un ajustement rigoureux, il faut disposer d'un estimateur *robuste* et *efficace*. Un estimateur est dit robuste s'il est peu perturbé par une valeur rare et extrême (*outlier*) ; il est d'autant plus efficace que sa variance est faible. Enfin, on cherche à ce que cet estimateur ait un biais le plus faible possible, et notamment qu'il soit asymptotiquement non biaisé, i. e. que le biais tende vers 0 lorsque la taille de l'échantillon tend vers l'infini.

La méthode des moindres carrés présente le grave défaut de donner beaucoup trop de poids aux évènements rares, ce qui conduit à des ajustements biaisés. Les corrections proposées par le Groupe de Travail AIRH nous paraissent insuffisantes pour pallier ce défaut. Pour des processus non linéaires, elle est aujourd'hui fortement déconseillée par les statisticiens. La méthode des moments consiste à utiliser les relations entre les moments de l'échantillon et les paramètres de la loi que l'on cherche à ajuster. Cette méthode construit certes des estimateurs convergents, mais ceux-ci sont souvent entachés de biais négatifs importants pour les petits échantillons. La méthode des moments pondérés (Hosking & Wallis, 1987) tente d'y remédier en pondérant les moments par leur probabilité. Dans le cas d'échantillons de taille réduite (inférieure à 500), pour l'ajustement à une loi GPD, Hosking et Wallis ont montré que cet estimateur était plus efficace que le maximum de vraisemblance pour $k < ½$. Dans la pratique, cette condition est certes souvent vérifiée, mais ce résultat ne concerne que la loi GPD.

De manière générale, l'estimateur du maximum de vraisemblance (EMV) est considéré comme le plus rigoureux par les statisticiens. L'EMV consiste à maximiser la fonction de vraisemblance en fonction des paramètres de la famille de lois choisie

pour l'ajustement. La méthode du maximum de vraisemblance repose sur des bases théoriques plus solides que celles de la méthode des moments. En particulier, on montre que, sous des conditions très générales, un estimateur MV est convergent, asymptotiquement normal et efficace. La méthode du maximum de vraisemblance est aujourd'hui la principale méthode d'estimation. En particulier, elle semble s'adapter beaucoup plus facilement à l'utilisation de données censurées que la méthode des moments.

Nous recommandons ainsi l'utilisation de l'estimateur du maximum de vraisemblance, même si une pondération judicieuse des moments peut éventuellement donner de bons résultats dans les domaines de validité *ad hoc*.

3.4 Mettre la théorie en pratique

À partir de cette base théorique, certains points restent à approfondir dans les applications pratiques.

Tout d'abord, le choix du seuil haut de censure des données demeure assez libre et la sensibilité à ce paramètre reste à explorer. Les illustrations présentées plus bas montrent qu'il existe des outils objectifs, basés sur des propriétés théoriques de la loi GPD, permettant d'identifier la valeur optimale de ce seuil haut au sens de cette distribution. Une difficulté pratique sera donc de déterminer si cette valeur est également pertinente pour les autres lois utilisées. En d'autres termes, on choisit donc de privilégier le modèle Poisson-GPD par ce choix, puis de vérifier si d'autres lois n'ajustent pas mieux l'échantillon.

Ensuite, le choix de la meilleure distribution reste aussi ouvert, les différentes lois utilisées ayant souvent des comportements très différents dans les quantiles extrêmes. Il faut donc quantifier la qualité de l'ajustement. L'utilisation de l'estimateur du maximum de vraisemblance nous conduit à utiliser deux critères de comparaison définis par Schwarz (1978) et Akaike (1973) : le *Bayesian Information Criterion* (BIC) qui est une minimisation du biais entre le modèle ajusté et la vraie distribution inconnue, et l'*Akaike Information Criterion* (AIC) qui sélectionne le modèle réalisant le meilleur compromis biais-variance. Ces critères font intervenir la vraisemblance L de l'ajustement, le nombre de données N et le nombre de paramètres k de la loi ajustée, comme le montrent les équations [3] et [4] :

$$BIC = -2\log(L) + 2k\log(N) \quad [3]$$

$$AIC = -2\log(L) + 2k \quad [4]$$

Pour maximiser la vraisemblance, la meilleure loi sera donc celle qui minimise ces critères.

4. Tests : le site d'Haltenbanken

4.1 Loi GPD et EMV

Nous prenons ici l'exemple classique d'Haltenbanken, en Atlantique Nord, au large de la Norvège, qui est le site utilisé par le Groupe de Travail. Le jeu de données est décrit par Hawkes sur le site Internet de l'AIRH. Un premier test est mené en n'utilisant que la loi GPD : nous supposons donc que nous nous trouvons dans le domaine asymptotique de la théorie des statistiques extrêmes. Nous disposons d'un échantillon de N_T = 128 pics de tempêtes supérieurs à 7 mètres (c'est le seuil bas) sur une période de K = 9 ans, fournis avec une précision à deux décimales significatives.

Nous utilisons la version 1.60 du paquet `extRemes` (Gilleland *et al.*, 2004) du logiciel d'analyse statistique `R`, qui repose sur les méthodes spécifiques développées par Coles (2001). Dans le cas d'une analyse type Poisson-GPD, ce paquet dispose d'outils objectifs pour déterminer la valeur haute du double seuil. En premier lieu, on examine (figure1) la stabilité des paramètres de forme et d'échelle k et ψ lorsque l'on balaye le seuil haut entre 7,5 et 9,5 m par pas de 0,01 m (correspondant à la précision sur les données) : on observe, à partir d'un certain seuil et sous réserve que la taille de l'échantillon restant demeure suffisante, des décrochages suivis de *zones de stabilité*. La zone de stabilité correspondant aux seuils les plus hauts peut nous laisser espérer que l'on se situe dans le domaine asymptotique. Nous choisissons alors de nous caler sur *le seuil le plus bas de la zone de stabilité la plus haute*, pour maximiser le nombre de données restantes et donc réduire les incertitudes.

Par ailleurs, on peut étudier (figure 2) le *mean excess plot* ou *mean residual life plot*, grâce à des propriétés théoriques de la loi GPD détaillées par Smith (2001). Notre analyse des graphes des figures 1 et 2 nous conduit à fixer le second seuil à 8,57 mètres, où une rupture apparaît assez nettement. Nous effectuerons donc l'ajustement sur un échantillon de N = 46 valeurs (soit un paramètre de censure ν = 0,36), ce qui correspond à une moyenne de 5,11 tempêtes par an. Sur d'autres échantillons (par exemple ceux caractérisés par une très grande stabilité de la loi GPD), ce choix peut cependant être bien plus difficile qu'ici ; une part importante de subjectivité est alors réintroduite.

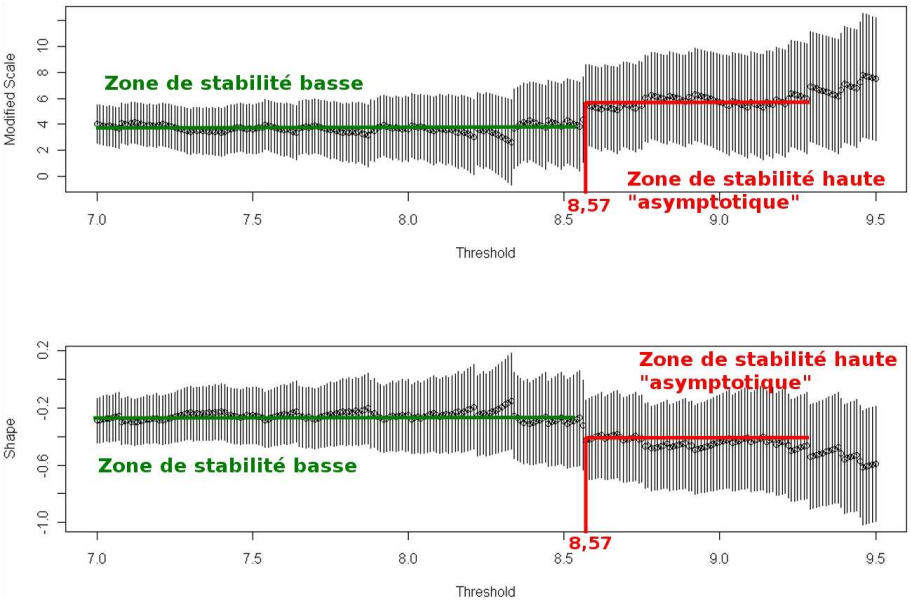

Figure 1. *Graphes extRemes de stabilité des paramètres d'échelle et de forme avec leurs intervalles de confiance associés pour la détermination du seuil haut de l'échantillon de Haltenbanken*

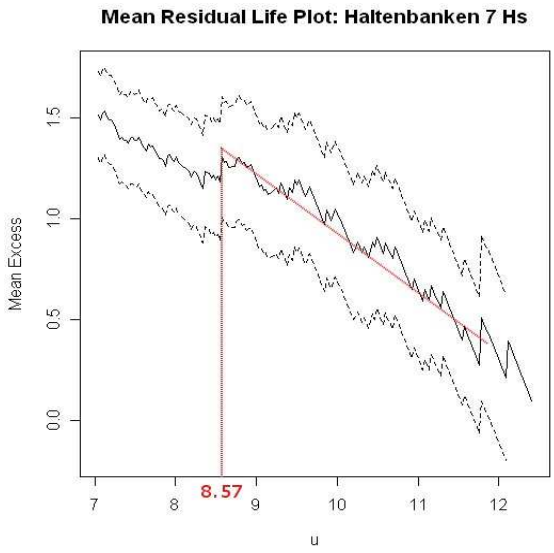

Figure 2. *Graphe extRemes du* mean residual life *pour la détermination du seuil haut de l'échantillon de Haltenbanken*

À l'aide du paramètre de Poisson λ, également estimé par le maximum de vraisemblance et alors assimilable à la moyenne empirique $\lambda_T = N_T / K$, `extRemes` renvoie les résultats suivants (figure 3) : $\psi = 1{,}90$, $k = -0{,}42$ et une houle centennale à 12,75 mètres avec un intervalle de confiance à 90 % de [12,4 ; 14,8], d'une amplitude de 2,4 mètres. Cet intervalle est cependant calculé par la méthode delta, qui n'est guère fiable lorsque la période de retour est grande devant la durée de l'échantillon, ce qui est le cas ici. Enfin, la log-vraisemblance nous permet de calculer les critères BIC et AIC.

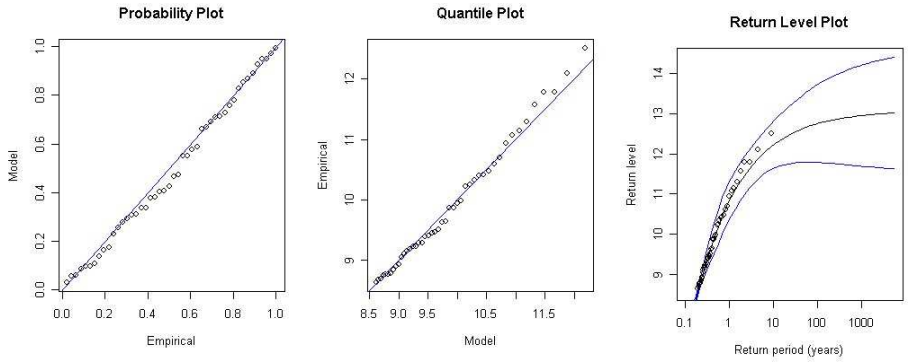

Figure 3. *Graphes extRemes pour l'ajustement GPD des données de Haltenbanken*

4.2 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'adapter l'échantillon à de nombreuses familles de distributions : GPD, Gumbel, Weibull, Gamma, Halphen de type A, log-Pearson de type III, avec le logiciel HYFRAN (Bobée *et al.*, 1999). Ce logiciel ne prend pas en compte la censure, aussi nous choisissons le seuil haut déterminé précédemment comme seuil unique. De plus, il ne propose pas l'estimateur MV pour la loi GPD, mais ceux des moments ou des moments pondérés, ce qui nous permet d'en comparer trois différents. Les intervalles de confiance sont calculés par la méthode asymptotique classique.

Le résultat d'une telle analyse est fourni sur les figures 4 et 5 et les résultats, avec les valeurs des houles centennales, les intervalles de confiance à 90 % (lorsqu'ils sont calculables), leur amplitude et le nombre de paramètres pour chaque loi, sont résumés dans le tableau 1, avec un rappel des résultats obtenus par le paquet `extRemes`.

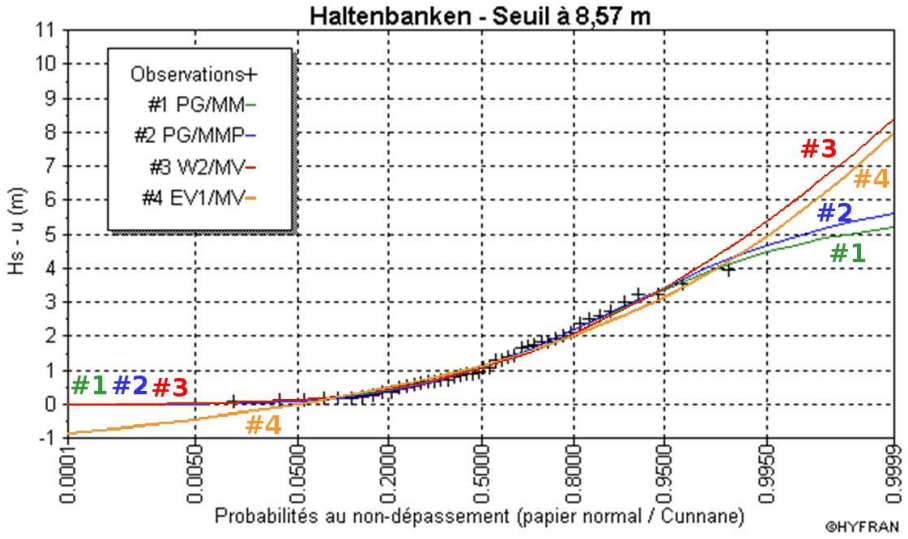

Figure 4. *Ajustement des données de Haltenbanken par les lois GPD (méthode des moments, PG/MM - #1), GPD (méthode des moments pondérés, PG/MMP - #2), de Weibull (maximum de vraisemblance, W2 - #3) et de Gumbel (maximum de vraisemblance, EV1 - #4)*

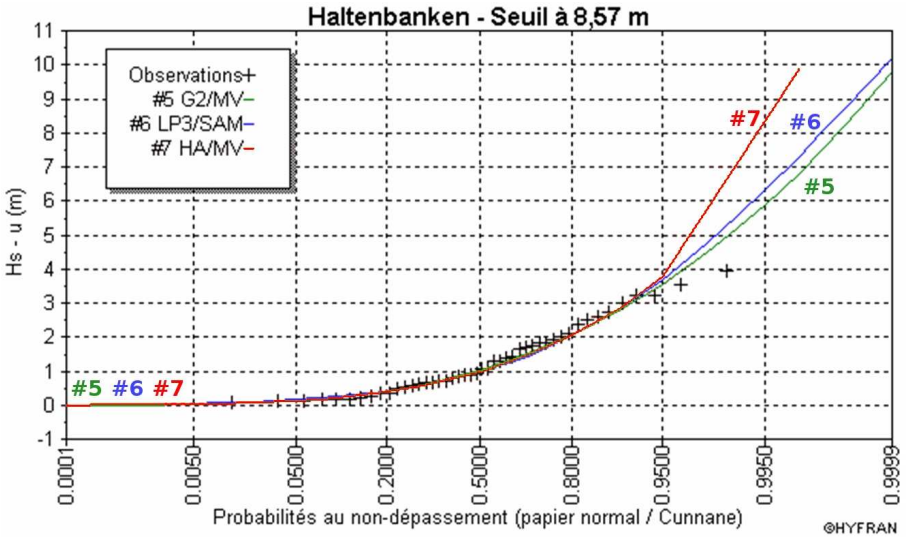

Figure 5. *Ajustement des données de Haltenbanken par les lois Gamma (maximum de vraisemblance, G2 - #5), log-Pearson de type III (LP3 - #6) et Halphen de type A (maximum de vraisemblance, HA - #7)*

Méthode	**extRemes**	**HYFRAN**						
Loi statistique	GPD	GPD	GPD	Weibull	Gamma	Halphen A	LP-III	Gumbel
Méthode d'ajustement	EMV	Moments	Mmts pondérés	EMV			SAM	EMV
Critère BIC	119,0	120,5	120,9	122,0	122,4	125,9	126,9	130,9
Critère AIC	116,3	116,9	117,3	118,3	118,8	120,5	121,4	127,3
Hs 100 ans (m)	**12,75**	**13,30**	**13,56**	**14,71**	**15,41**	**18,51**	**15,90**	**14,25**
borne inf IC 90% (m)	12,4	-	-	12,9	13,6	14,8	-	13,3
borne sup IC 90% (m)	14,8	-	-	16,5	17,3	22,2	-	15,3
amplitude IC 90% (m)	*2,4*	-	-	*3,6*	*3,7*	*7,4*	-	*2,0*
Nombre paramètres *k*	*2*	*2*				*3*		*2*

Tableau 1. *Houle centennale, intervalle de confiance à 90%, critères BIC et AIC et nombre de paramètres pour chaque loi ajustée aux données de Haltenbanken*

La meilleure loi au sens des critères BIC et AIC reste ici la loi GPD ajustée par EMV. HYFRAN privilégie également la loi GPD, mais la différence due aux choix d'estimateurs est clairement visible. Pour cette loi, l'écart entre le maximum de vraisemblance et les moments pondérés atteint en effet 0,8 mètre. Il a été vérifié sur plusieurs sites et de nombreux échantillons que les critères BIC et AIC privilégiaient systématiquement les ajustements EMV ; aussi on peut en conclure que les ajustements par les moments ou les moments pondérés sont inutiles et peuvent être abandonnés par la suite. Plusieurs points sont également à relever : les lois renvoient des valeurs de hauteurs centennales très différentes (alors même qu'elles sont toutes acceptées par le test d'adéquation du χ^2 !) ; les critères BIC et AIC se rejoignent pour fournir le même classement ; on observe que les lois à trois paramètres sont plus biaisées que celles n'en comprenant que deux, puisqu'un paramètre, lui-même estimé, apporte sa propre incertitude (cela est d'ailleurs pris en compte dans les critères BIC et AIC puisque le nombre de ceux-ci fait augmenter la valeur du critère). L'exception de la loi de Gumbel (implémentée avec deux paramètres dans ce logiciel) est due à son mauvais ajustement des basses valeurs.

5. Application pratique : détermination des houles extrêmes dans le détroit de Gibraltar

5.1 Présentation du site et des sources de données

L'exemple de la détermination pratique des houles extrêmes dans le détroit de Gibraltar est présenté dans ce qui suit. Le site est particulier, puisqu'il s'agit du seul passage maritime entre Atlantique et Méditerranée, que le détroit n'est large que de 15 km et qu'il s'inscrit d'un point de vue géomorphologique dans une vaste baie orientée vers l'Ouest délimitée au Nord par le cap Saint-Vincent (Portugal) et au Sud par el-Jadida (Maroc). La bathymétrie générale du détroit est relativement complexe, avec une série de caps et de baies, de hauts-fonds et de canyons sous-marins influençant la propagation à la côte des houles atlantiques. Le haut-fond localisé dans le prolongement du cap Trafalgar a de ce point de vue une importante

influence. Le site compte plusieurs ports importants, parmi lesquels ceux de Cadix et de Tanger qui sont plus particulièrement exposés aux conditions atlantiques.

Deux sources de données ont été considérées pour l'étude des houles extrêmes à l'entrée du détroit de Gibraltar : les mesures du houlographe au large de Cadix et la base de données SIMAR-44.

La bouée « Golfo de Cadiz » fait partie du réseau REDEXT de houlographes directionnels en eaux profondes de l'organisme public « Puertos del Estado » dépendant du Ministère de l'Equipement espagnol. Les mesures de cette bouée ont aussi été utilisées pour la validation des climats reconstitués.

La base de données SIMAR-44, dont la diffusion est assurée par Puertos del Estado (2008), est une base de données d'états de mer reconstitués numériquement, dans le cadre du projet européen HIPOCAS pour la partie méditerranéenne et étendue à l'Atlantique Nord par Puertos del Estado. Elle couvre une période de 44 années allant du 01.01.1958 au 31.12.2001 avec des résultats toutes les 3 heures de plusieurs paramètres caractéristiques des états de mer. Cette base de données couvre la Méditerranée occidentale (côtes méditerranéennes françaises, espagnoles, marocaines et algériennes) et les côtes atlantiques espagnoles, portugaises, marocaines ainsi que l'archipel des Canaries.

La figure 6 présentée ci-après montre la localisation du point de mesure de la bouée « Golfo de Cadiz » et du point SIMAR correspondant ainsi que l'emplacement de deux autres points analysés (Gibraltar Nord et Gibraltar Sud) de la base de donnée SIMAR.

Figure 6. *Localisation des points analysés et présentés dans cet article*

5.2 Validation/comparaison des données utilisées

Les figures 7, 8 et 9 présentent pour les deux types de données respectivement les séries temporelles de hauteur de houle significatives correspondant aux tempêtes de décembre 1997 et décembre 2000 et les courbes de dépassement sur une période d'analyse commune au point localisé au large de Cadix.

On observe sur les figures 7 et 8 que l'ajustement temporel entre les deux types de données est bon. On notera néanmoins sur les trois figures la divergence des données SIMAR en termes de hauteur de houle pour les valeurs supérieures à 3 m environ, celles-ci tendant à surestimer les valeurs extrêmes par rapport à la mesure.

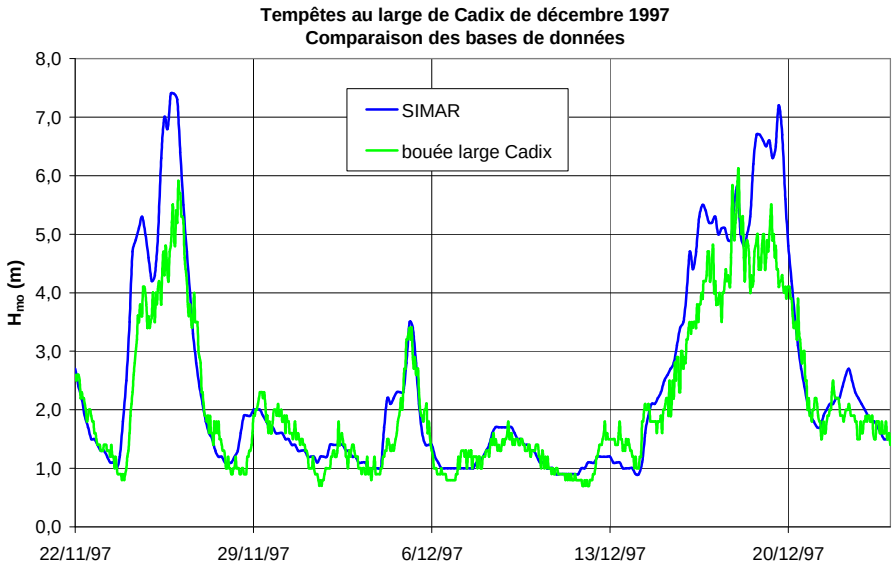

Figure 7. *Séries temporelles de hauteur significative de houle entre le 22 novembre et le 25 décembre 1997 – comparaison des données (site de Cadix)*

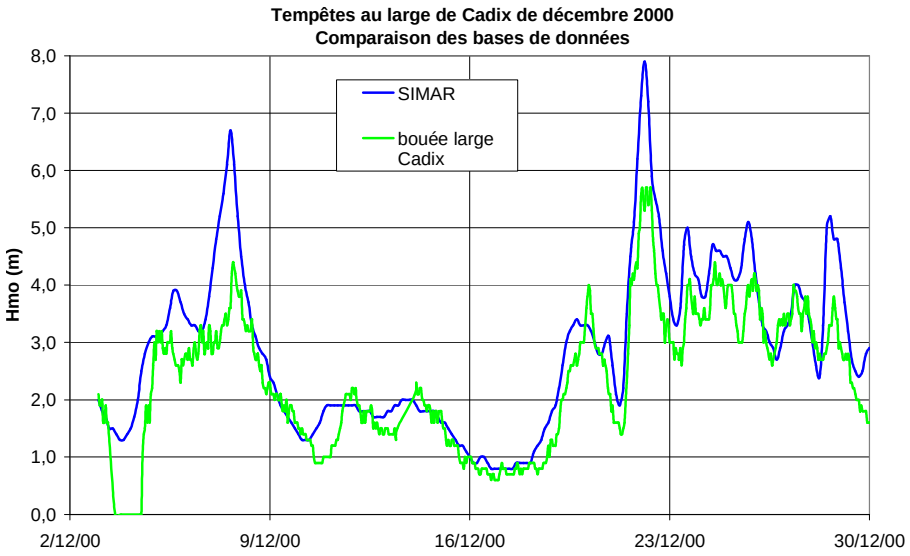

Figure 8. *Séries temporelles de hauteur significative de houle entre le 04 décembre et le 30 décembre 2000 – comparaison des données (site de Cadix)*

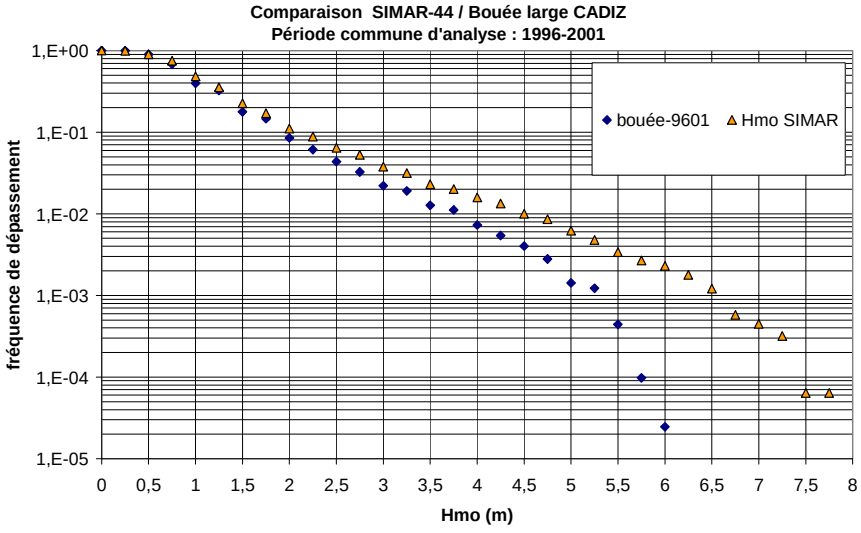

Figure 9. *Courbes de fréquence de dépassement de hauteur significative de houle – comparaison des données (site de Cadix)*

5.3 Variabilité spatiale de la houle dans le détroit

Pour déterminer les hauteurs significatives des états de mer extrêmes à l'entrée du détroit de Gibraltar, deux points alignés selon un axe Nord – Sud et distants d'une quinzaine de milles marins ont été analysés. Le point NORD de référence 1056044 dans la base de données SIMAR-44 se localise au sud du cap Trafalgar, le point SUD de référence 1056043 se localise à l'ouest du Ra's Spartel (ouest du port de Tanger Ville), comme indiqué sur la figure 6.

La comparaison du climat de houle à ces deux points montre que :

– les houles sont plus fortes au point SUD de 0,50 à 0,75 m lorsque $H_{m0} > 3{,}5$ m ;

– l'incidence des houles au point SUD est orientée de 10 à 25 ° plus au nord et il y a une forte variation en directions entre les deux points ;

– au point SUD, les périodes de pic sont majoritairement supérieures à celles du point NORD, dans un intervalle de 0 à 1 s (on observe un climat de houle longue au point Sud qui n'existe pas au point Nord, probablement en raison des différences bathymétriques et à la présence du Cap St Vincent, voir figure 10) ;

– les périodes de pic sont bornées pour chaque direction ;

– les houles de fortes périodes ont des incidences comprises entre N260° et N280° au point NORD (figure 11) ;

– les houles de fortes périodes ont des incidences comprises entre N285° et N310° au point SUD (figure 12).

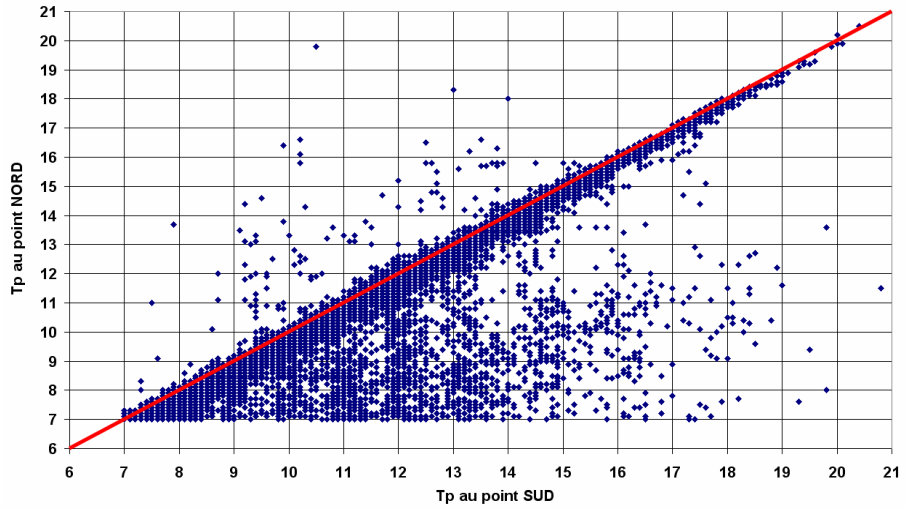

Figure 10. *Comparaison de la période (en secondes) de houle aux deux points SIMAR-44 à l'entrée du détroit de Gibraltar sur la période 1958-2001*

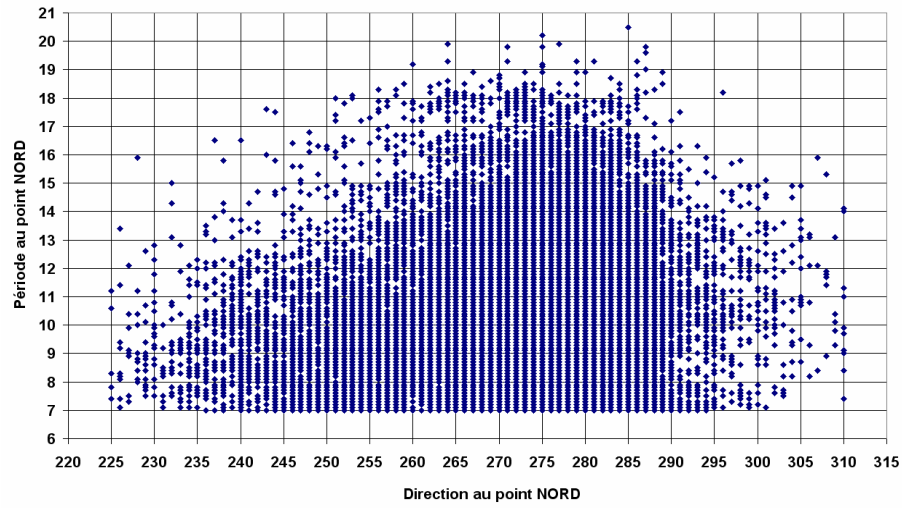

Figure 11. *Diagramme période/direction au point NORD (période 1958-2001)*

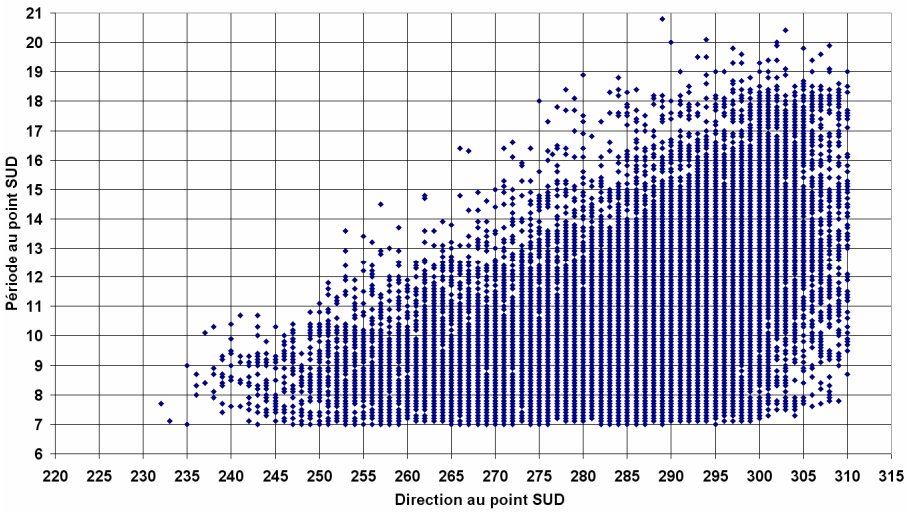

Figure 12. *Diagramme période/direction au point SUD (période 1958-2001)*

La présence de ce gradient Nord-Sud fortement marqué montre bien qu'un seul point d'analyse au large est insuffisant pour déterminer les conditions de projet d'un site côtier localisé à l'intérieur du détroit.

5.4 Analyse pluriannuelle des données SIMAR

L'analyse des données SIMAR sur l'ensemble de la période 1958-2001 permet de constater que la période 1983-2001 est sous-représentée en tempêtes majeures, comme le montre la répartition temporelle des hauteurs significatives d'états de mer pour le point Nord représentée sur la figure 13.

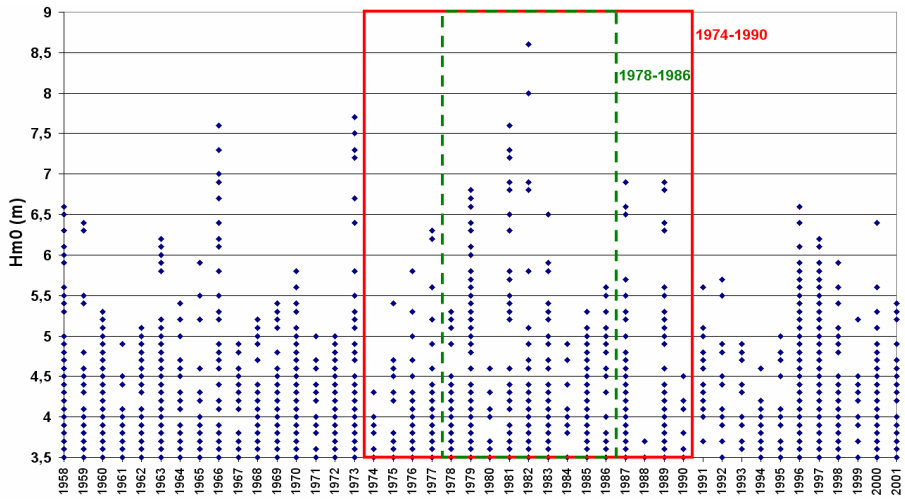

Figure 13. *Hauteurs significatives supérieures à 3,5 m de 1958 à 2001 au point SIMAR Gibraltar Nord et périodes d'étude choisies sur 9 et 17 ans*

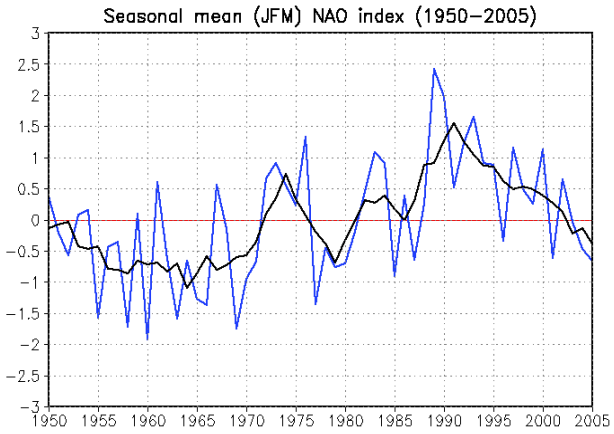

Figure 14. *Valeurs moyennes saisonnières de l'indice NAO sur la période 1950-2005 (source : NOAA)*

Cette différence d'activité pourrait être associée à l'oscillation Nord-Atlantique (NAO). L'indice NAO prend en effet des valeurs plus fréquemment et plus fortement positives sur la période 1983-2001 (voir figure 14), ce qui implique que les tempêtes dans le détroit de Gibraltar durant cette période sont moins violentes que pour la période 1950-1982. En effet, une phase positive de la NAO tend à dévier les tempêtes atlantiques sur l'Europe septentrionale, alors qu'une phase négative tend à orienter leur trajectoire sur le bassin méditerranéen. Cet exemple met en évidence la nécessité de disposer d'une base de données suffisamment longue pour être représentative des conditions extrêmes du site et en l'occurrence pour mieux appréhender les aspects météo-océanographiques de nature cyclique et à période longue qui régissent les houles à l'entrée du détroit de Gibraltar.

Sur la base des figures 13 et 14, nous avons identifié 1982 comme année charnière, date de basculement de l'indice NAO, pour mener une étude sur l'influence de la taille temporelle de l'échantillon sur les houles centennales calculées. Nous avons donc mené trois analyses sur le point Nord : une étude sur 44 ans (1958-2001), une sur 17 ans (1974-1990) et une dernière sur 9 ans (1978 – 1986), comme indiqué sur la figure 13. Le choix de ces périodes centrées sur 1982 permet d'avoir dans chaque échantillon une durée équivalente des phases négative et positive de l'indice NAO, mais également d'étudier la réponse des échantillons devant le pic maximal de tempête de 1982. Par souci de simplicité, et pour nous concentrer sur les aspects que nous venons de citer, nous nous limitons ici au modèle Poisson-GPD à l'aide du logiciel `extRemes` et aux lois de Weibull et Gamma avec HYFRAN.

		9 ans	**17 ans**	**44 ans**
	Seuil de censure (m)	4,2	4,6	4,5
	Nombre de tempêtes *N*	18	21	90
	Tempêtes/an (λ)	2	1,24	2,05
GPD EMV	**Hs 100 ans (m)**	**9,69**	**8,71**	**8,43**
	IC sup 90%	-	12,3	10,0
	BIC	49,65	56,44	174,31
Weibull EMV	**Hs 100 ans (m)**	**10,45**	**9,26**	**8,62**
	IC sup 90%	13,6	11,0	9,5
	BIC	50,49	56,96	173,79
Gamma EMV	**Hs 100 ans (m)**	**10,56**	**9,67**	**8,77**
	IC sup 90%	13,4	11,5	9,6
	BIC	50,46	57,16	173,10

Tableau 2. *Résultats du point Gibraltar Nord pour les trois durées d'étude*

Les résultats sont présentés dans le tableau 2. On peut en conclure :

- pour une loi donnée, l'analyse sur une longue période tend à diminuer la houle centennale ; cela est du au fait que la tempête de 1982 a une période de retour très grande devant, par exemple, la durée de 9 ans choisie, et de ce fait tire les résultats

vers le haut. L'estimateur MV est néanmoins reconnu pour sa robustesse face à de tels *outliers* ; un ajustement par les moindres carrés, par exemple, aurait montré une sensibilité bien plus grande ;

- la loi GPD est privilégiée par le critère BIC pour les échantillons de 9 et 17 ans. En revanche, pour l'analyse sur 44 ans, la loi Gamma s'ajuste mieux. Ce résultat justifie notre décision de ne pas se limiter au modèle Poisson-GPD ;

- l'analyse sur 44 ans fournit une estimation proche de celle portant sur 17 ans avec, comme attendu, une réduction des intervalles de confiance ;

- l'écart par rapport à la meilleure loi est au maximum de 9, 11 et 4 % sur chaque échantillon. L'expérience sur différents sites montre que cela est un bon résultat.

L'analyse sur une longue période est donc intéressante, surtout si l'on soupçonne la présence dans l'échantillon d'un pic de tempête à très grande période de retour, où si l'on veut lisser les effets d'un phénomène cyclique, ici la NAO. Une autre méthode, détaillée par Mendez *et al.* (2006), consiste à rendre les paramètres des lois dépendants du temps. Ainsi, pour rendre compte de l'oscillation nord-atlantique, les paramètres se décomposeraient en une valeur moyenne à laquelle s'ajouteraient deux termes sinusoïdaux (on suppose la période de l'oscillation connue). Une telle approche augmente cependant le nombre de paramètres à estimer et donc l'incertitude associée. De plus, on ne sait si la période et l'intensité de la NAO resteront stationnaires sur de longues périodes de temps. Cela est néanmoins une approche très innovante si l'on veut tenir compte de la non-stationnarité du climat sur de longues périodes, et en particulier du changement climatique attendu.

6. Conclusions

La prise en compte de l'environnement dans les projets d'aménagement a pris de plus en plus d'importance ces dernières décennies, conduisant dans différents domaines scientifiques à des suivis en nature réguliers de plus en plus poussés. Parallèlement, les statisticiens ont été mis à contribution pour développer des outils mathématiques d'analyse de ces données pour différents besoins. L'analyse des valeurs extrêmes est une branche des statistiques qui s'est donc considérablement développée ces dernières années pour répondre à la demande des ingénieurs.

Nous nous sommes ainsi appuyés sur les résultats théoriques obtenus pour rappeler que la distribution généralisée de Pareto (GPD), associée à la loi de dépassement de seuil de Poisson, est la distribution asymptotique vers laquelle tend un échantillon de hauteurs significatives sélectionnées par la méthode du renouvellement. Nous avons souligné d'autre part que l'ajustement à cette loi GPD devait s'effectuer par l'estimateur du maximum de vraisemblance.

Nous avons ensuite recensé les difficultés pratiques relatives à la mise en œuvre de ce modèle Poisson-GPD. En premier lieu, elles concernent l'acquisition des données. Leur nature peut être très diverse (mesures de bouées, états de mer

reconstitués et validés ou non par mesures satellitaires…) et leur fiabilité est parfois aléatoire. Pour une bonne analyse statistique, tous les efforts doivent être faits pour que ces données soient au maximum indépendantes et homogènes (identiquement distribuées). La sélection des tempêtes doit être effectuée rigoureusement, en prenant garde notamment aux fluctuations au cours d'un même évènement météorologique. Nous avons également présenté des outils aidant à une détermination la plus objective possible du seuil haut. Malgré tout, une part de subjectivité peut demeurer pour certains échantillons caractérisés par une très grande stabilité de la loi GPD.

L'autre difficulté majeure provient du caractère asymptotique du modèle Poisson-GPD. En effet, on ne peut être assuré de se trouver dans le domaine de validité de cette loi. Pour y remédier, tout en privilégiant ce modèle en choisissant le seuil qui lui est optimal, nous proposons d'élargir l'analyse à d'autres distributions. Des critères objectifs (BIC et AIC) permettent alors d'identifier le meilleur ajustement pour en tirer les houles de projet. L'analyse sur 44 ans à Gibraltar a illustré le fait qu'une autre loi peut donner des résultats légèrement meilleurs que ceux du modèle Poisson-GPD.

Cette approche à double détente permet d'améliorer la confiance dans les résultats obtenus, tout en gardant un regard critique sur ceux-ci. Un bon analyste sait décomposer judicieusement son échantillon en fonction des spécificités météorologiques et maritimes du site étudié pour le rendre le plus homogène possible, l'analyser rigoureusement et surtout prendre le recul nécessaire face aux résultats obtenus, qui ne sont jamais un but en soi mais toujours insérés dans la conception d'un projet pour lequel l'enchainement des méthodes de calcul et des choix de conception doit garder sa cohérence (choix des coefficients de sécurité et des niveaux de risques à chaque étape selon le type d'ouvrage). Dans ce contexte, une suite à ce travail pourrait ainsi s'orienter vers une meilleure appréciation de l'étalement des intervalles de confiance qui reste un peu rudimentaire actuellement.

L'application pratique présentée sur Gibraltar met également l'accent sur l'importance d'établir une base de données suffisamment longue et validée comme préalable à la détermination de houles extrêmes. Ces deux conditions sont rarement rencontrées simultanément actuellement. En effet, la reconstitution des données historiques par calcul nécessite une assimilation des données satellitaires. Cela n'est possible que sur la période récente (typiquement 1991-2008). Au-delà, vers le passé, des hypothèses doivent être faites, comme c'est le cas pour la base de données SIMAR couvrant une période de 44 ans. Cette longue période de reconstitution permet en outre d'inclure un cycle complet de l'oscillation nord-Atlantique qui joue clairement un rôle sur la climatologie pluriannuelle du site étudié. La réflexion sur la meilleure manière de prendre en compte la non-stationnarité d'un échantillon (par exemple le changement climatique) reste cependant à approfondir.

7. Références bibliographiques

Akaike H., « Information theory as an extension of the maximum likelihood principle », *Second International Symposium on Information Theory*, 1973, p. 267-281. Akademiai Kiado, Budapest.

Bobée B., Fortin V., Perreault L., Perron H., *HYFRAN 1.0*. INRS-Eau, Terre et Environnement, Université du Québec, Québec, 1999.

Castillo E., Sarabia J.M., « Engineering analysis of extreme value data: selection of models », *J. Wtrwy., Port, Coast., and Oc. Engrg.*, vol. 118, Issue 2, p. 129-146, 1992.

CIRIA, CUR, CETMEF, *Guide Enrochement - L'utilisation des enrochements dans les ouvrages hydrauliques*, 2009. ISBN 978-2-11-098518-7.

Coles S., *An introduction to statistical modelling of extreme values*, Springer-Verlag, London, 2001.

Fisher R.A., Tippett L.H.C., « Limiting forms of the frequency distributions of the largest or smallest member of a sample », *Proceedings of the Cambridge Philosophical Society*, vol. 24, p. 180-190, 1928.

Franco L., Piscopia R., *Atlante delle onde nei mari italiani – Italian wave atlas*, Agenzia per la Protezione dell'Ambiente e per i Servizi Tecnici, 2004.

Fréchet M., « Sur la loi de probabilité de l'écart maximum », *Annales de la Société polonaise de Mathématique*, vol. 6, Cracovie, 1927.

Gilleland E., Katz R., Young G., *The extRemes Package*, 2004. URL: http://cran.r-project.org/doc/packages/extRemes.pdf.

Goda Y., « On the methodology of selecting design wave height », *Proceedings of the 21st International Conference on Coastal Engineering*, Malaga, ASCE, p. 899-913, 1988.

Goda Y., Kobune K., « Distribution function fitting for storm wave data », *Proceeding of the 22nd International Conference on Coastal Engineering*, Delft, ASCE, p. 18-31, 1990.

Goda Y., « Statistical analysis of extreme waves », *Random seas and design of maritime structures*, chap. 19. Advanced Series on Ocean Engineering, vol. 15, World Scientific, p. 377-425, 2000.

Hawkes P.J., « Case 5: Extreme Wave Analysis Data ». URL : http://www.iahr.net/site/e_library/links/databases/CASE5/Case5Hawkes.doc

Hosking J.R.M., Wallis J.R., « Parameter and quantile estimation for the generalized Pareto distribution », *Technometrics*, vol. 29, p. 339-349, 1987.

Jenkinson A. F., « The frequency distribution of the annual maximum (or minimum) values of meteorological events », *Quaterly Journal of the Royal Meteorological Society*, N°81, p. 158-171, 1955.

Mathiesen M., Goda Y., Hawkes P.J., Mansard E., Martín M.J., Peltier E., Thompson E.F., Van Vledder G., « Recommended practice for extreme wave analysis »., *Journal of Hydraulic Research*, vol. 32, N°6, 1994.

Méndez F.J., Menendez M., Luceno A., Losada I.J., « Estimation of the long-term variability of extreme significant wave height using a time-dependent Peak Over Threshold (POT) model », *J. Geophys. Res., 111*, 2006, C07024, doi : 10.1029/2005JC003344.

Pickands J. III, « Statistical Inference Using Extreme Order Statistics », *The Annals of Statistics*, vol. 3, No. 1, p. 119-131, 1975.

Puertos del Estado, « Conjunto de datos SIMAR-44 (Proyecto HIPOCAS) », 2008. Fiche de présentation des données disponible sur le site web de Puertos del Estado : http://www.puertos.es/es/oceanografia_y_meteorologia/banco_de_datos/index.html.

R Development Core Team, « R: A language and environment for statistical computing », R Foundation for Statistical Computing, Vienna, Austria, 2007. ISBN 3-900051-07-0. URL: http://www.R-project.org.

Schwarz G., « Estimating the dimensions of a model », *Annals of statistics*, vol. 6, p. 461-464, 1978.

Smith R.L., *Environmental statistics.* Department of Statistics, University of North Carolina, 2001. http://www.stat.unc.edu/postscript/rs/envnotes.ps.

Thompson E.F., « Hydrodynamic Analysis and Design Conditions », *Coastal Engineering Manual*, J. Pope (ed.), Part II Chapter 8, US Army Corps of Engineers, 2002.