Réconcilier théorie et pratique dans la détermination des houles extrêmes

Franck MAZAS[1], Luc HAMM[2]

[1] Ecole Nationale des Ponts et Chaussées, ParisTech.
franck.mazas@eleves.enpc.fr

[2]Directeur technique, SOGREAH Maritime, 6 rue de Lorraine, 38130 Échirolles, France.
luc.hamm@sogreah.fr

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. La méthode POT est préconisée 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. Enfin, des tests sont menés sur le site d'Haltenbanken pour illustrer les améliorations proposées.

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. 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. Finally, tests were conducted on the site of Haltenbanken to illustrate the proposed improvements.

Mots clés :
Valeurs extrêmes – houle – POT – EMV.

DOI: 10.5150/jngcgc.2008.037-M (disponible en ligne - http://www.paralia.fr - available online)

1. Introduction

Sur un site maritime, prévoir les hauteurs de vagues extrêmes sur de longues périodes de retour (de l'ordre de quelques dizaines d'années) est primordial pour le dimensionnement des ouvrages portuaires 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 a été proposée par le Professeur Goda (GODA, 1988b ; GODA & KOBUNE, 1990). Elle a été largement reprise par le Groupe de Travail sur les Statistiques des Houles Extrêmes dans son document de synthèse (MATHIESEN *et al.*, 1994) et tout récemment dans le *Rock Manual* du CIRIA (2007). Devant les difficultés à concilier théorie statistique et pratique de l'ingénieur, elle se veut globale et relativement légère à mettre en œuvre.

Nous examinerons ici la question du choix des distributions statistiques à ajuster aux données de tempêtes, à la méthode d'ajustement adéquate et à des outils objectifs de détermination de seuil et de choix de la meilleure distribution.

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, 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.

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* : un seuil bas permet de sélectionner toutes les tempêtes alors qu'un seuil plus haut, dont la détermination est essentielle, retient les plus hauts pics auxquels on ajustera la loi. Nous appellerons ce doublet seuil bas – seuil haut, *double seuil*. L'intérêt du processus de censure a été souligné par le Groupe de Travail. 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 (voir 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 $X > u$. Lorsque u approche le point terminal (fini ou infini), la loi des dépassements de u peut être approchée par la *distribution généralisée de Pareto* (GPD, *Generalized Pareto Distribution*) donnée par :

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

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

Le nombre N_1 de dépassements du seuil u dans une année pouvant être considéré comme régi par un processus poissonien, on suggère le modèle suivant, appelé *modèle Poisson-GPD*, où les dépassements de seuil obéissent à une loi GPD et sont i.i.d., et où 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 l'on dispose habituellement, c'est-à-dire quelques dizaines de valeurs. 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 (log-normale, log-Pearson de type III, Gamma, χ^2…) peuvent *a priori* fournir une meilleure modélisation. Une analyse plus approfondie utilisera donc avec bonheur un grand nombre de distributions candidates : bien que beaucoup plus lourde, c'est l'approche la plus justifiable.

3. 3 Ajustement

Pour réaliser un ajustement rigoureux, il faut disposer d'un estimateur *robuste* et *efficace*. Un estimateur est dit robuste s'il est très 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. Pour des processus non linéaires, elle est aujourd'hui fortement déconseillée par les statisticiens. La méthode des moments, la plus intuitive, consiste à utiliser les relations entre les moments de l'échantillon et les paramètres de la loi que l'on

cherche à ajuster. La méthode des moments 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 souvent vérifiée… mais ce résultat ne concerne que la 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, ce qui nous intéresse particulièrement.
L'estimateur du maximum de vraisemblance est donc recommandé, même si une pondération judicieuse des moments peut donner de meilleurs résultats dans les domaines de validité *ad hoc*. Enfin, la détermination des intervalles de confiance sera une étape cruciale de l'analyse.

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. 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 128 pics de tempêtes supérieurs à 7 mètres sur une période de 9 ans.
Nous utilisons le *package* `extRemes` (GILLELAND *et al.*, 2004) du logiciel système d'analyse statistique `R`. Dans le cas d'une analyse type Poisson-GPD, ce *package* dispose d'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, d'autre part en étudiant le *mean excess plot* ou *mean residual life plot*, grâce à des propriétés théoriques de la loi GPD. SMITH (2001) détaille ces techniques. Ici, ils suggèrent de fixer le second seuil à 8.57 mètres ; nous effectuerons donc

l'ajustement sur un échantillon de 46 valeurs (soit un paramètre de censure ν = 0.36).

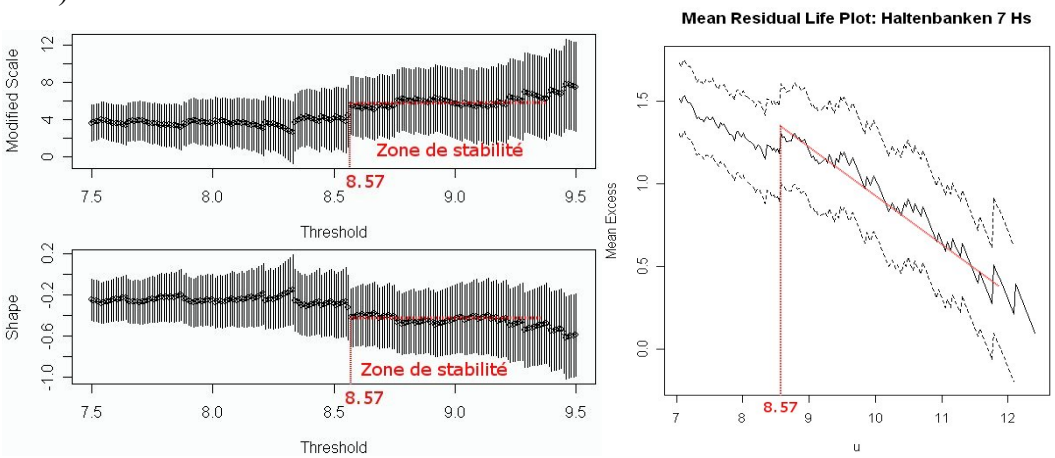

Figure 1. Graphes `extRemes` *pour la détermination du seuil haut de l'échantillon de Haltenbanken.*

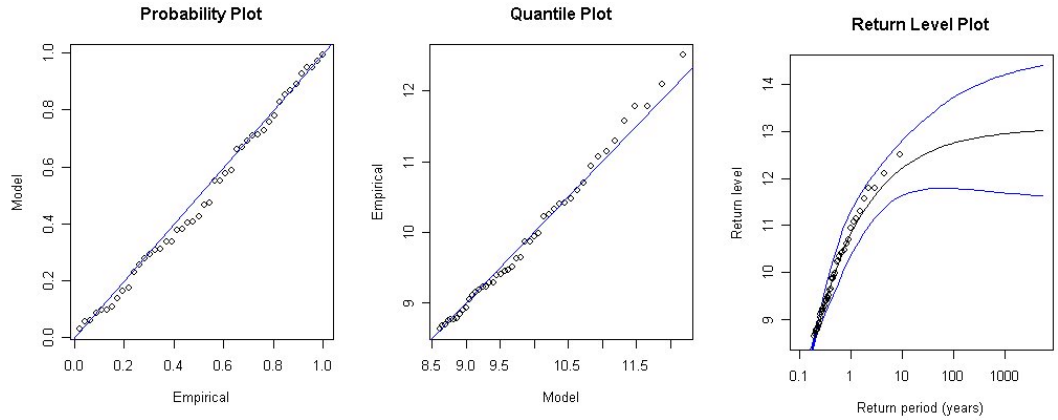

Figure 2. Graphes `extRemes` *pour l'ajustement GPD des données de Haltenbanken.*

Nous obtenons les résultats suivants : 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].

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, exponentielle, GEV, log-Pearson de type III, avec le logiciel HYFRAN (BOBÉE *et al.*, 1999).

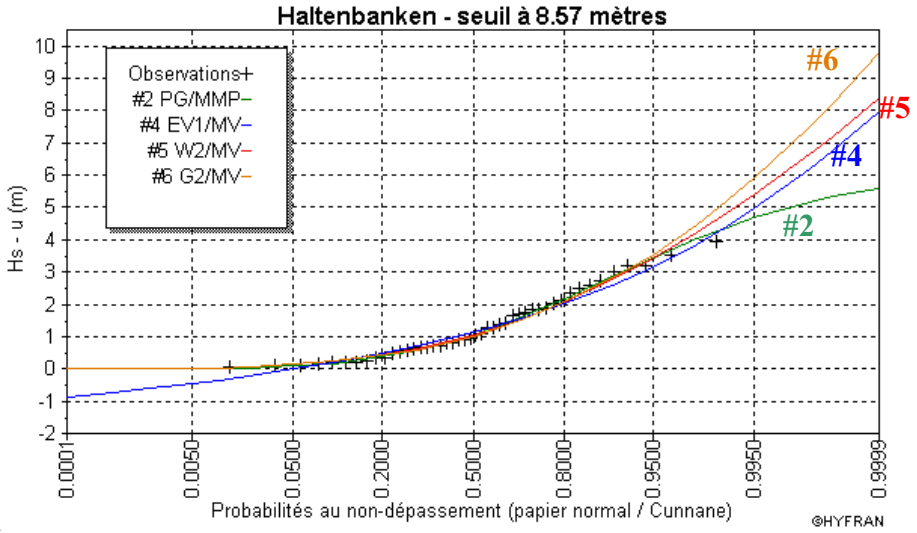

Figure 3. Ajustement par des lois GPD (PG - #2), Gumbel (EV1 - #4), Weibull (W2 - #5) et Gamma (G2 - #6) aux données de Haltenbanken.

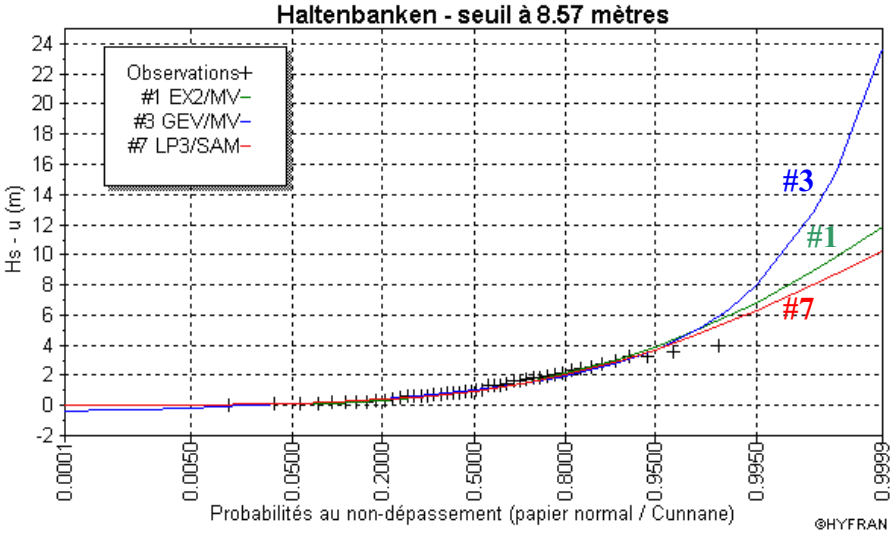

Figure 4. Ajustement par des lois exponentielle (EX2 - #1), GEV (#3) et log-Pearson de type III (LP3 - #7) aux données de Haltenbanken.

On voit qu'il est difficile de privilégier une loi particulière sur la foi d'un simple examen graphique, alors même que ces lois ont des comportements très différents au niveau des quantiles extrêmes. Il faut quantifier la qualité de l'ajustement ; pour cela on peut utiliser, entre autres, deux critères de comparaison : 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. La meilleure loi minimise ces critères. Les 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 résumés dans le tableau 1.

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 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

Plusieurs remarques sont ici à soulever. La première est que c'est bien la loi GPD qui est sélectionnée ici. De plus, les critères BIC et AIC se rejoignent pour fournir le même classement. On remarque qu'à l'exception de la loi de Gumbel, dont la relégation paraît *a priori* étrange, les lois retournant de très fortes valeurs sont rejetées en fin de classement. Il faut d'ailleurs noter la très forte disparité des houles centennales, alors même que toutes ces lois ont été acceptées par le test d'adéquation du χ^2 ! Enfin, les lois à trois paramètres sont plus biaisées que les lois à deux paramètres, car rajouter des paramètres accroît l'incertitude sur ces mêmes paramètres. Cela est d'ailleurs pris en compte dans les critères BIC et AIC puisque le nombre de paramètres fait augmenter la valeur du critère.

5. Conclusions

De telles analyses sur des échantillons de données environnementales recueillies sur une grande période de temps sont très délicates. De nombreux tests sont nécessaires pour appréhender les difficultés de toute sorte qui interviennent. Celles-ci sont généralement de deux types : des difficultés intrinsèques à l'échantillon et des difficultés purement statistiques. Les premières sont dues au caractère non indépendant mais surtout non homogène et non stationnaire de l'échantillon. C'est la plus grande source d'imprécision et, partant, la plus grande source potentielle d'améliorations.

Parallèlement, les outils numériques nous permettent aujourd'hui d'utiliser des méthodes statistiques plus performantes et plus justifiables théoriquement. Nous insistons ici sur la détermination du seuil haut par des outils objectifs, et nous recommandons deux méthodes : une, légère, basée sur un modèle Poisson-GPD,

et une autre plus rigoureuse mais plus lourde fondée sur une approche multi-distributions. L'ajustement par l'EMV est fortement recommandée.
Enfin, l'accent doit être mis sur l'importance des intervalles de confiance, dont la largeur peut fortement varier. 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 de notre travail doit clairement s'orienter vers une meilleure appréciation de l'étalement des intervalles de confiance qui reste un peu rudimentaire actuellement.

6. Références bibliographiques

BOBÉE B., FORTIN V., PERREAULT L., PERRON H. (1999). *HYFRAN 1.0*. INRS-Eau, Terre et Environnement, Université du Québec, Québec.

CIRIA (2007). *Manual on the use of rock in coastal and shoreline engineering*.

FISHER R.A., TIPPETT L.H.C. (1928). *Limiting forms of the frequency distributions of the largest or smallest member of a sample.* Proceedings of the Cambridge Philosophical Society, 24:180-190.

FRÉCHET M. (1927). *Sur la loi de probabilité de l'écart maximum*. Annales de la Société polonaise de Mathématique, vol. 6, Cracovie.

GILLELAND E., KATZ R., YOUNG G. (2004). *The extRemes Package*. URL : http://cran.r-project.org/doc/packages/extRemes.pdf.

GODA Y. (1988). *On the methodology of selecting design wave height*. Proc. 21st Int. Conf. Coastal Engrg. Malaga, pp. 899-913.

GODA Y., KOBUNE K. (1990). *Distribution function fitting for storm wave data.* Proc. 22nd Int. Conf. Coastal Engrg, Delft, pp. 18-31.

HOSKING J.R.M., WALLIS J.R. (1987). "*Parameter and quantile estimation for the generalized Pareto distribution*", *Technometrics*, 29:339-349.

JENKINSON A. F. (1955). *The frequency distribution of the annual maximum (or minimum) values of meteorological events*. Quaterly journal of the Royal meteorological society, 81, pp. 158-171.

MATHIESEN M., GODA Y., HAWKES P.J., MANSARD E., MARTÍN M.J., PELTIER E., THOMPSON E.F., VAN VLEDDER G. (1994). *Recommended practice for extreme wave analysis.* Journal of Hydraulic Research, Vol. 32, N°6.

R DEVELOPMENT CORE TEAM (2007). *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.
SMITH R.L. (2001). *Environmental statistics*. Department of Statistics, University of North Carolina. http://www.stat.unc.edu/postscript/rs/envnotes.ps.