Modélisation du comportement extrême de processus spatio-temporels. Applications en océanographie et météorologie.

Nicolas Raillard

N° d'ordre : 4487 ANNÉE 2011

THÈSE / UNIVERSITÉ DE RENNES 1
sous le sceau de l'Université Européenne de Bretagne

pour le grade de

DOCTEUR DE L'UNIVERSITÉ DE RENNES 1

Mention : Mathématiques appliquées

Ecole doctorale MATISSE

présentée par

Nicolas Raillard

préparée à l'unité de recherche IRMAR - n°6625
Institut de Recherche Mathématique de Rennes
Université de Rennes 1

Modélisation du comportement extrême d'un processus spatio-temporel.

Applications en océanographie et en météorologie.

Thèse soutenue à Rennes le 13 Décembre 2011
devant le jury composé de :

Liliane Bel
Professeur, Université Paris-Sud / rapporteur

Pierre Ailliot
Maitre de conférence, Université de Brest / examinateur

Catherine Laredo
Professeur, Université Diderot-Paris 7 / examinateur

Valérie Monbet
Professeur, Université de Rennes 1 / Présidente

Bertrand Chapron
Chercheur, IFREMER Brest / co-directeur de thèse

Remerciements

Je tiens en premier lieu à remercier tous ceux avec qui j'ai eu le plaisir de travailler au cours de cette thèse, et notamment Bertrand Chapron et Jian-feng Yao mes deux directeurs. Je tiens particulièrement à remercier Pierre Ailliot, qui m'a constamment aidé, écouté et corrigé pendant ces trois années, son aide fût précieuse.

Mes remerciements vont également à l'ensemble des membres du jury, et aux rapporteurs Liliane Bel et Peter Challenor pour m'avoir fait l'honneur d'évaluer ces travaux.

Le soutient amical a également été important, et je pense en particulier à toutes les personnes qui ont travaillé et qui se sont succédées au laboratoire de mathématiques de Brest et ceux qui y sont encore.

Last but not least, j'ai une pensée particulière pour Anne-lise, mon amie et probablement ma plus fidèle lectrice pour son soutient et ses encouragements tout au long de ces trois années.

Table des matières

Introduction

Motivations

Les évènements extrêmes sont des facteurs déterminants dans de nombreuses activités humaines, que ce soient par exemple les tempêtes, les inondations ou encore les chocs économiques ; alors que souvent les risques faibles à modérés sont bien connus et pris en compte. Il est donc nécessaire de disposer d'outils permettant d'assurer au long terme la pérennité de structures, la continuité des activités ou encore de s'assurer de la solidité d'un acteur économique. Les années passées ont montré l'importance de la prise en compte de ces évènements extrêmes parfois catastrophiques, entre autres dans les domaines financiers ou environnementaux. L'augmentation de la population mondiale implique que de plus en plus de territoires soient habités et que par conséquent de plus en plus d'individus soient soumis aux aléas climatiques, rendant nécessaire la description des risques les plus extrêmes sur des zones de plus en plus larges. Le problème commun à toute modélisation des extrêmes est le manque d'observations : en effet, on cherche à estimer la quantité la plus importante observée, disons en 100 ans, alors que la durée typique d'une série d'observations est de quelques années. Il est donc crucial de développer des méthodes prenant en compte d'une part le maximum d'observations, et d'autre part permettant d'utiliser le plus de sources de données différentes décrivant le même phénomène.

Dans les domaines d'applications usuels de la théorie des valeurs extrêmes, la mer tient une place particulièrement importante puisque ce sont des problématiques de construction de digues qui ont motivé les premières applications, en particulier suite à la tempête de Février 1953 qui a causé de nombreux dégâts matériels et humains. Les travaux présentés dans ce document se placent également dans le domaine des applications à l'océan, puisque nous nous intéresserons à des données de hauteur significative de vagues, quantité que nous détaillerons dans la suite, mais qui est au coeur de préoccupations pratiques, comme le calibrage de structures marines ou encore la prévision de production d'électricité par des systèmes de récupération d'énergie des vagues. La mesure de cette quantité peut être réalisée par différents systèmes, qu'ils soient *in-situ* comme ceux embarqués sur les bateaux ou sur les bouées météorologiques, ou effectuée à distance à l'aide des satellites dotés d'altimètres qui observent la Terre continument depuis quelques années. La hauteur significative des vagues est également une valeur utilisée par des modèles climatologiques, permettant d'intégrer de nombreuses observations au sein d'un modèle numérique. Comme nous le préciserons dans la suite, chacune de ces sources présente des avantages et des inconvénients, mais il semble que les bouées fournissent les observations les plus précises, mais sont peu nombreuses en espace. Les données satellitaires quant à elles sont précises également, mais souffrent de difficultés de traitements que nous détaillerons dans la suite de ce document. La particularité de l'échantillonnage spatio-temporel des données satellitaires a motivé l'introduction de nouvelles méthodes, que ce soit pour la reconstruction de séries temporelles sur des grilles régulières aussi proche que possible de ce que donnerait une bouée située au même endroit, ou encore pour la modélisation de séries temporelles

dont les observations sont réalisées à pas de temps irrégulier. En particulier, il est de prime importance de pouvoir utiliser ces données pour décrire les extrêmes de hauteur de vagues à un endroit où il n'y a pas de bouée.

Plan

Ce document est organisé de la manière suivante : dans une première partie, nous faisons un rappel des théories probabilistes des valeurs extrêmes et des méthodes statistiques correspondantes, dans les cas uni-variés, bi-variés. Puis nous présentons dans ce même chapitre les résultats plus récents concernant la théorie des valeurs extrêmes de processus, ainsi que des extrêmes d'observations univariées dépendantes, telles que les séries temporelles.

Dans une seconde partie, nous présentons un article publié au cours de cette thèse, développant une méthode d'interpolation des données satellitaires permettant la reconstruction de séries temporelles en des points de l'espace où il n'y a pas d'observations in-situ. La méthode proposée utilise une estimation des déplacements sur des données de réanalyse à l'aide de méthode de filtrage particulaire. Cette méthode montre une amélioration de l'adéquation aux données de bouée par rapport aux méthodes existantes (krigeage 'simple' des données satellitaires, données de réanalyses), et permet la création de bouées virtuelles. Nous verrons que cette approche n'est pas satisfaisante pour les extrêmes, et qu'il est donc nécessaire de développer d'autres approches, ne pouvant être basée sur le kriggeage.

Les deux chapitres suivants sont consacrés respectivement au développement et à l'étude d'un nouveau modèle pour le premier, chapitre dans lequel nous établirons les propriétés de ce nouveau modèle, avant de le tester sur diverses séries temporelles usuelles pour vérifier sa capacité à décrire les extrêmes dans de nombreuses situations. Le chapitre 4 quant à lui est consacré à l'application de cette nouvelle approche sur les divers jeux de données en notre possession. Nous présenterons également à cette occasion les résultats d'une première extension de notre approche pour modéliser les dépassements de seuils dans un contexte spatio-temporel.

Une conclusion générale dresse une synthèse des résultats présentés dans les différents chapitres de cette thèse, tout en fournissant quelques perspectives envisagées pour des études ultérieures.

Chapitre I

Outils probabilistes et statistiques pour la modélisation des queues de distributions

1 Théorie des valeurs extrêmes dans le cas de variables IID unidimensionnelles

1.1 Extrema par blocs

1.1.1 Résultats théoriques

Si l'on se donne un échantillon de variables aléatoires $(X_i)_{i=1,\dots,n}$ indépendantes et identiquement distribuées (i.i.d), la statistique classique s'intéresse à la loi de $S_n = \sum_{i=1}^{n} X_i$, convenablement centrée et normalisée. On sait grâce aux travaux de Paul Lévy que S_n (centré et normalisé) converge en loi vers un élément de la classe des lois stables pour l'addition, et même le plus souvent vers son élément le plus connu, la loi normale, qui en est un cas particulier (voir, entre autres, [67]). Or, quand on s'intéresse aux queues de distributions, ce n'est plus la moyenne que l'on va chercher à décrire, mais le comportement de $M_n = \max_{i=1,\dots,n} X_i$, à nouveau correctement centré et normalisé. De même que dans le cas de S_n, la famille de lois possibles pour M_n est connue, cette fois-ci grâce aux travaux de Fisher et Tippett ([27]). Comme dans le cas du comportement asymptotique de la moyenne, la famille de lois limites sera stable, non par l'application de la somme, mais par l'application du maximum, d'où appellation de loi max-stables. Cette famille est caractérisée par une équation fonctionnelle portant sur sa fonction de répartiton, et dont les solutions sont les membres de la famille *GEV* (Generalized Extreme Value Distribution). Plus précisément, on a le théorème suivant :

Théorème 1.1 (Fisher-Typett, 1928)**.** S'il existe des suites $(a_n)_{n\in\mathbb{R}} > 0$ et $(b_n)_{n\in\mathbb{R}}$ telles que

$$\lim_{n\to\infty} \mathbb{P}\Big\{a_n^{-1}(M_n - b_n) \leq x\Big\} \to F(x) \tag{I.1}$$

alors F est un membre de la famille *GEV*, c'est-à-dire qu'il existe des paramètres μ, $\sigma > 0$ et ξ tels que

$$F(x) = \begin{cases} \exp\left[-(1-\xi\frac{x-\mu}{\sigma})^{1/\xi}\right] & \forall\ x/1+\xi(x-\mu)/\sigma) > 0, \text{ si } \xi \neq 0 \\ \exp(-e^{-x}) & , \text{ si } \xi = 0 \end{cases} \tag{I.2}$$

où $\mu \in \mathbb{R}$, $\sigma > 0$ et $\xi \in \mathbb{R}$ sont trois paramètres dépendants de la loi initiale des observations $(X_i)_{i=1,\dots}$

Les trois paramètres de F, μ, σ et ξ, sont respectivement ses paramètres de position, d'échelle, et de queue. Les comportements diffèrent largement suivant ce dernier paramètre : en effet, pour $\xi > 0$ on a une loi à queue lourde (loi de type Fréchet), pour $\xi < 0$ on a une loi à support borné (loi de type Weibull), alors que le cas $\xi = 0$ est intermédiaire, correspondant à une loi de type Gumbel. Pour plus de détails sur ce théorème, on peut consulter le chapitre 3 de [10] et pour plus de détails mathématiques, on peut se référer à la partie 3.2 de [23]. La figure I.1 montre la fonction de répartition de certains membres de cette famille, montrant bien les différents comportements de la queue de la distribution.

Il est intéressant à la vue de ce théorème de se demander s'il est possible de prévoir la valeur des paramètres de la loi limite du maximum quand on connait la loi dont sont issues les observations : c'est la question du domaine d'attraction, défini de la manière suivante :

Définition 1.1. Une variable aléatoire X appartient au domaine d'attraction de la distribution des valeurs extrêmes F s'il existe des suites déterministes $(a_n)_{n\in\mathbb{N}} > 0$ et $(b_n)_{n\in\mathbb{N}}$ telles que I.1 est vérifiée.

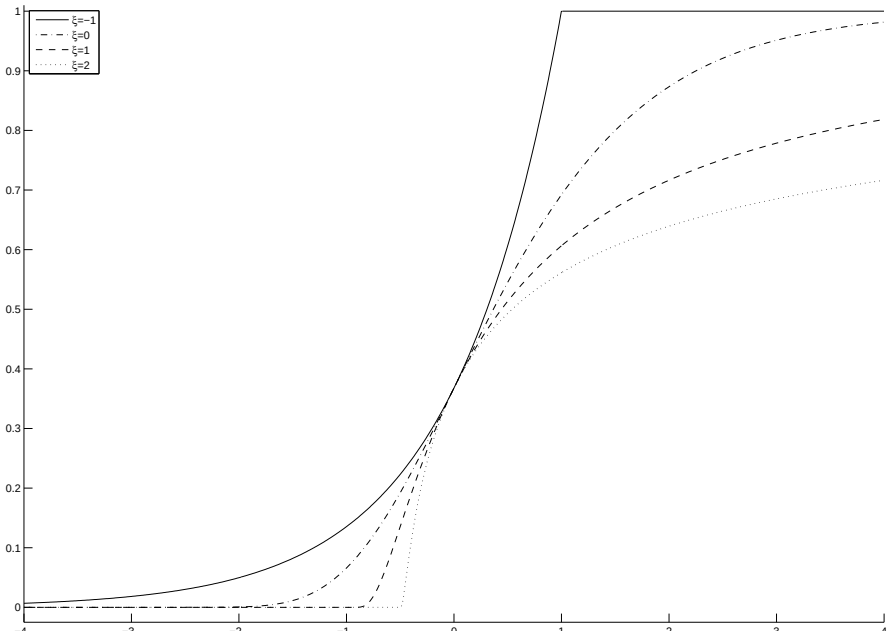


FIGURE I.1 – Fonctions de répartition de lois GEV pour différentes valeurs de ξ, avec $\sigma = 1$ et $\mu = 0$

On peut alors chercher à déterminer dans quel domaine d'attraction appartiennent les lois usuelles. Cette approche n'a pas été retenue dans notre étude car dans le cas général, la loi des observations est inconnue, et on ne peut donc pas étudier le domaine d'attraction de cette façon. Un apprentissage important néanmoins de cette approche est la classification des lois usuelles suivant leur appartenance à un domaine d'attraction de distributions aux queues plus ou moins lourdes. Une telle classification se trouve par exemple dans [23]. On pourra retenir que la loi normale, très couramment —voire abusivement— utilisée, est dans le domaine d'attraction de la loi de Gumbel, correspondant, on le rappelle, à une queue de distribution légère, quoique non bornée supérieurement.

1.1.2 Estimation

D'un point de vue statistique, l'approche couramment employée consiste à utiliser le théorème I.1 pour estimer le comportement du maximum : les données sont séparées en groupes de tailles égales (par exemple des blocs d'une année), sur lesquels on calcule les maxima, et on suppose alors que l'on dispose d'un échantillon i.i.d de loi GEV, ce qui serait réaliste pour des blocs dont la taille tend vers l'infini. De nombreux travaux se sont attachés à l'estimation des paramètres en partant de cette modélisation : maximum de vraisemblance ([44, 57]), la méthode des L-moments ([31]), la méthode des moments pondérés ([1]) pour l'estimation des trois paramètres ; estimateur de Hill ([21]), de Pickands ([42]), de Dekkers-Einmahl-De Hann ([18]) pour l'estimation du paramètre de queue ; chaque méthode citée étant valable dans un domaine particulier du paramètre de queue. Par exemple, il est connu depuis les travaux de Smith [54] que l'estimateur au maximum de vraisemblance fonctionne sur un domaine restreint en ce qui concerne les valeurs de ξ, en partie à cause du fait que le support de la loi dépend de ce paramètre, contrairement aux cas usuels, comme par exemple la famille exponentielle. On a plus particulièrement le résultat

suivant :

- Si $\xi > -0.5$, l'estimateur au maximum de vraisemblance est régulier, dans le sens où il vérifie les propriétés habituelles d'efficacité et de normalité asymptotique ;
- Si $-1 < \xi < -0.5$, l'estimateur au maximum de vraisemblance peut être généralement calculé, mais ne possède pas les propriétés habituelles, en particulier la normalité asymptotique ;
- Si $\xi < -1$, l'estimateur au maximum de vraisemblance n'est généralement pas calculable du fait de problèmes numériques.

Ces limitations sont souvent peu contraignantes en pratique, car le cas où $\xi < -0.5$ correspond à des queues bornées très courtes, qui est un cas assez peu fréquent lorsque l'on s'intéresse à la modélisation des valeurs extrêmes, en particulier lorsque que l'on s'intéresse aux extrêmes de données environnementales ou financières, pour lesquelles on constate des paramètres de queue positifs, ou nuls.

Un des problèmes de cette approche est que toute l'information collectée n'est pas retenue pour l'estimation, étant donné que l'on ne conserve que les maxima sur un bloc, par exemple une année. Il y a alors un compromis à établir entre la taille des blocs, pour justifier de l'utilisation de la loi limite pour le maximum, mais aussi sur le nombre de blocs résultants pour avoir assez d'observations afin de mener à bien l'estimation des paramètres. Lorsque l'on regroupe les observations pour n'en conserver que le maximum par bloc, on perd non seulement l'information temporelle sur les moments où ces maxima sont observés et sur leur étendue temporelle comme sur leur tendance à l'agrégation ou à la répulsion, mais on perd aussi une information sur le comportement de la distribution aux alentours du maxima, comme la forme d'un évènement extrême. De plus, les conditions d'application du théorème de Fisher-Tippett ne sont que rarement vérifiées, étant donné que les observations initiales ne sont pas systématiquement i.i.d., et il faut donc développer des modèles dans ce cas-là : ce point fera l'objet de la dernière partie de ce chapitre, les sections intermédiaires étant dédiées à l'extension de la description des lois max-stables au cas multivarié et à celui des processus stochastiques. Auparavant, nous nous intéresserons à la description d'une méthode se passant du regroupement par blocs.

1.2 Loi des dépassements de seuils

1.2.1 Résultats théoriques

Dans cette partie, nous allons présenter une approche utilisée pour résoudre le problème de la perte d'information qui a lieu lorsque que l'on ne conserve que les maxima par blocs. Un évènement extrême sera défini par le dépassement d'un seuil u élevé, et l'on s'intéressera à la loi des dépassements, c'est-à-dire à la probabilité conditionnelle

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

Ce domaine suit les travaux fondateurs de Balkema et de Hann ([2]) ainsi que Pickands ([42]). Un théorème similaire à celui sur la convergence de M_n indique que l'ensemble des lois possibles est la famille des lois de Pareto généralisée (GPD), grâce au théorème suivant (Pickands [42]) :

Théorème 1.2. Soit Xdans le domaine d'attraction d'une loi GEV de paramètres μ, σ et ξ. Alors on a le résultat suivant :

$$\sup_{z \in \mathbb{R}} |\mathbb{P}\left(X - u \leq z \,|X > u\right) - G(z; \xi, \tilde{\sigma}_u)| \underset{u \to +\infty}{\longrightarrow} 0$$

où $G(z;\xi,\sigma) = 1 - \left(1 - \frac{\xi z}{\sigma}\right)^{1/\xi}$ et $\tilde{\sigma}_u = \sigma + \xi(\mu - u)$. Les lois du même type que G constituent la famille **GPD**.

La figure I.2 montre les comportements de la famille ci-dessus pour différentes valeurs du paramètre de queue.

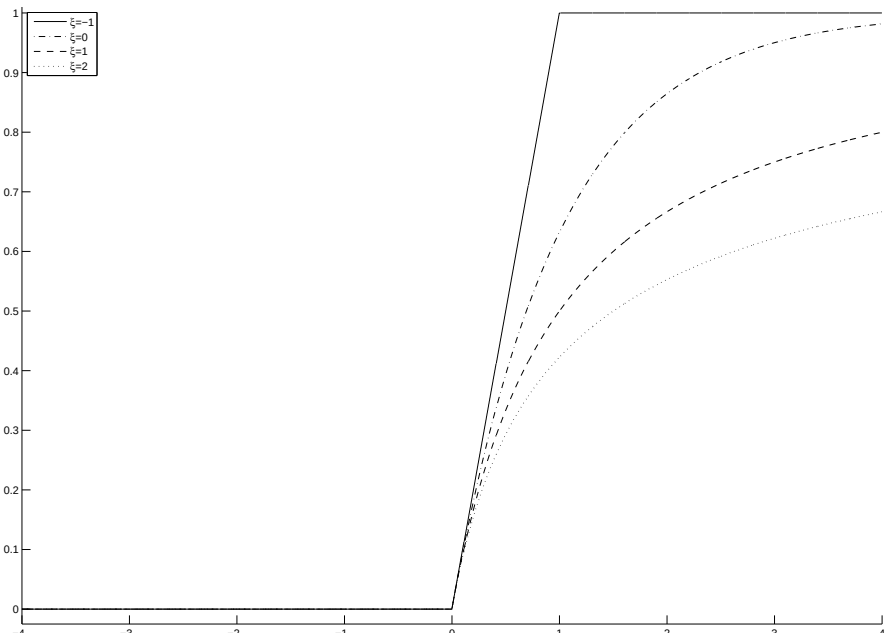


FIGURE I.2 – Fonctions de répartition de lois GPD pour différentes valeurs de ξ, avec $\sigma = 1$ et $\mu = 0$

Le chapitre 4 de [10] propose une justification de ce théorème ainsi qu'une discussion sur le choix du seuil et l'estimation des paramètres, une discussion détaillée étant disponible dans [13].

1.2.2 Estimation

Choix du seuil Le résultat théorique énoncé précédemment fait apparaitre un comportement quand on observe les excès au-delà d'un seuil qui devient arbitrairement grand. Or, en pratique, il faut le laisser fini sous peine de ne plus avoir d'observations le dépassant. On voit alors apparaitre une nécessité de compromis entre la justesse de l'approximation, engendrant un biais, et le nombre d'observations conservées, dont la diminution quand le seuil augmente génère une variance d'estimation plus importante. Deux méthodes principales sont utilisées par les praticiens, une exploratoire et une basée sur l'estimation des paramètres :

– Supposons qu'au-delà d'un seuil u_0, X suive effectivement une loi GPD de paramètres σ_{u_0} et $\xi < 1$, ce dernier paramètre étant invariant du seuil. On peut montrer le résultat suivant pour un autre seuil $u > u_0$:

$$\mathbb{E}\left(X - u | X > u\right) = \frac{\sigma_u}{1-\xi} = \frac{\sigma_{u_0} + \xi u}{1-\xi}$$

et donc l'espérance des dépassements de seuil u est une fonction linéaire de u. Puisque cette quantité est facilement estimable par sa contrepartie empirique, nous avons une

première approche permettant de choisir un seuil convenable, en cherchant le seuil au-delà duquel la fonction précitée est linéaire.
– Une approche alternative consiste à estimer les paramètres de la loi GPD (voir ci-après les méthodes d'estimation) et de chercher une valeur du seuil au-delà de laquelle ces estimations sont stables. Pour obtenir des paramètres indépendants du seuil, Coles (2001) propose de tracer le comportement de $\sigma^* = \widehat{\sigma_u} - \hat{\xi}u$ et de $\hat{\xi}$.

Estimation des paramètres Une fois le seuil fixé, plusieurs approches existent pour l'estimation des paramètres : on peut calculer la vraisemblance associée aux dépassements de seuils, car la loi des excès suit — asymptotiquement— une loi de Pareto généralisée. Il est ainsi possible de construire l'estimateur au maximum de vraisemblance pour estimer les paramètres σ et ξ, ce dernier étant le même que lorsque l'on s'intéresse au maxima par bloc.

Plus précisément, le modèle retenu est le suivant :

$$\mathbb{P}\{X > x | X > u\} = \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}.$$

Ce que l'on peut aussi écrire :

$$\mathbb{P}\{X \leq x\} = 1 - \lambda_u \left[1 + \xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi},$$

avec $\lambda_u = \mathbb{P}\{X > u\}$. On voit ainsi qu'il est nécessaire d'estimer trois paramètres, et bien que l'estimation de λ_u puisse se faire trivialement à l'aide du nombre de dépassement, l'erreur d'estimation devrait être prise en compte lors du calcul des intervalles de confiance sur les quantiles extrêmes, ce qui semble rarement fait en pratique. Nous proposerons dans la suite une méthode alternative qui permet de s'affranchir de ce problème, tout en retirant le fait que le paramètre σ estimé dépende du seuil u, grâce à une autre approche sur les dépassements de seuils.

Le lecteur peut aussi se référer aux articles mentionnés dans la partie précédente et qui s'intéressent à l'estimation de ξ, vu que les méthodes d'estimation de ce paramètre sont transposables au cas présent.

1.3 Applications

De nombreuses applications des résultats présentés ci-dessus existent dans la littérature, si bien qu'il serait fastidieux d'en dresser un portrait exhaustif. Nous pouvons cependant dégager quelques domaines principaux dans lesquels la théorie des valeurs extrêmes a été utilisée par les praticiens, ce qui lui confère une reconnaissance toute particulière : ces domaines sont l'assurance et la finance comme en témoigne [23] ou [4], la fiabilité et ses applications industrielles ([30] ou [22]), et enfin les sciences environnementales au sens large, domaine qui retiendra notre attention tout particulièrement.

Dans le domaine de l'environnement, la question des valeurs extrêmes est en effet primordiale du fait de l'impact de tels évènements sur la société : séismes, inondations, tempêtes, mais aussi en terme de pollution par exemple. Il est donc naturel que ces applications soient très souvent citées en exemple, d'autant plus que la statistique des extrêmes a été initiée par des préoccupations issues des problématiques de prévention d'inondations aux Pays-Bas. De nombreuses applications de la théorie univariée que l'on vient de présenter sont disponibles dans l'ouvrage collectif [19]. Citons également des applications de la théorie usuelle aux données de hauteurs significatives : [8] ou [63], mais les exemples

d'application de méthodes alternatives comme la formule de Rice existent également (voire par exemple [50]). Être proche des personnes utilisant la théorie des valeurs extrêmes permet également de développer de nouvelles applications, en prise directe avec les problèmes que rencontrent les praticiens pour modéliser les évènements extrêmes, et de nouvelles méthodes découlent directement de ces problématiques : citons par exemple [60], mais les exemples d'applications aux données environnementales sont trop nombreuses pour que l'on puisse prétendre à l'exhaustivité dans ce domaine.

2 Statistique des extrêmes bi- et multivariés

Dans certains contextes, il peut être intéressant d'étudier des comportements extrêmes pour des observations multivariées, telles des couples : non seulement pour modéliser des risques joints, mais aussi en vue d'améliorer la modélisation du comportement extrême d'une quantité grâce à l'information contenue dans une autre. Citons par exemple la construction de digues, pour lesquelles les hauteurs d'eau et celle des vagues pourraient être modélisées conjointement. Nous allons tenter ici de répondre aux questions suivantes : comment définir un extrême multivarié, suffit-il par exemple qu'une composante soit extrême ? Comment lier les extrêmes de différentes variables ? Comment décrire, étudier, modéliser et estimer la structure de dépendance ? Comment étendre les méthodes de dépassements de seuils en dimension supérieure à un ?

Ces réponses se trouveront au moins en partie dans les paragraphes suivants. On pourra pour plus de détail se référer à [10], [3] et [15], ainsi que les références qui s'y trouvent.

2.1 Maxima par composantes

2.1.1 Caractérisation de la famille de lois limites

Dans tout ce qui suit, nous nous intéresserons uniquement à des observations bivariées, l'extension aux dimensions (finies) supérieures ne présentant pas de difficultés théoriques majeures. Nous supposerons que nous disposons d'un échantillon $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ de vecteurs aléatoires indépendants et identiquement distribués, de même distribution que (X, Y), et de fonction de répartition bivariée $F(x, y)$. Les lois marginales seront notées $F_1(x) = \mathbb{P}(X \leq x) = F(x, \infty)$, et $F_2(y) = \mathbb{P}(Y \leq y) = F(\infty, y)$ de manière analogue.

Nous allons considérer les maxima par composantes, $M_n^X = \max_{1 \leq i \leq n} X_i$ $M_n^y = \max_{1 \leq i \leq n} Y_i$. On définit le maxima par composante

$$M_n = (M_n^X, M_n^Y),$$

qui n'est pas forcement un point faisant parti des observations.

La question qui se pose alors est de savoir quel est le comportement de $\mathbb{P}(M_n \leq z) \overset{\text{def.}}{=} \mathbb{P}(M_n^X \leq x, M_n^Y \leq y) = F(z)^n$ avec $z = (x, y)$, quand $n \to \infty$: c'est-à-dire a quelles conditions il existe des suites $a_n > 0$, $a'_n > 0$, b_n et b'_n et une fonction de répartition bivariée G telles que $F^n(a_n x + b_n, a'_n y + b'_n) \underset{n \to \infty}{\to} G(x, y)$, avec G une loi max-stable bivariée, c'est-à-dire vérifiant pour tout n,

$$G^n(a_n x + b_n, a'_n y + b'_n) = G(x, y).$$

La convergence des lois marginales X et Y est relativement facile à caractériser : en effet, la convergence de F^n vers G implique une converge des loi marginales également. Les lois marginales F_i sont donc dans le domaine d'attraction d'une certaine loi max-stable

univariée, G_i. Pour étudier la structure de dépendance d'une loi max-stable G, on va se ramener à des lois marginales identiques, a l'instar de l'étude des copules : la loi Fréchet unitaire définie par

$$\Psi_1(z) = \mathbb{P}(Z \leq z) = \exp(-1/z)$$

qui est un cas particulier de loi GEV avec $\mu = \sigma = \xi = 1$ et qui vérifie la propriété de max-stabilité avec $a_n = 1/n$ et $b_n = 0$.

La première étape est donc de transformer les marges vers la loi commune, Fréchet(1) :

Proposition 2.1. Soit G une loi max-stable bivariée, de marginale G_1 et G_2. Si $(X, Y) \sim G$, alors :

$$\left(-\frac{1}{\log G_1(X)}, -\frac{1}{\log G_2(Y)}\right) \sim G^*$$

où G^* est une loi max-stable bivariée de lois marginales G_1^* et G_2^* Fréchet unitaire (i.e $G_1^* = \Psi_1$.

On a de plus le théorème suivant :

Théorème 2.1 (Resnick, 1985)**.** Soit (X_i^*, Y_i^*) un vecteur aléatoire de marges Fréchet unitaire et $(M_n^{X^*}, M_n^{Y^*})$ défni comme précédemment. Si

$$\mathbb{P}\left(M_n^{X^*} \leq x, M_n^{Y^*} \leq y\right) \to G(x, y)$$

où G est une distribution non-dégénéré, alors elle est de la forme

$$G(x, y) = \exp(-V(x, y)),\ x, y > 0$$

où :

$$V(x, y) = 2\int_0^1 \max\left\{\frac{w}{x}, \frac{1-w}{y}\right\} dH(w)$$

et H est une distribution sur $[0, 1]$ vérifiant $\int_0^1 w dH(w) = 1/2$.

La fonction V définie précédemment vérifie la propriété suivante, dite d'homogénéité d'ordre -1 :

$$V(ax, ay) = a^{-1}V(x, y),$$

et cette propriété caractérise également la famille des lois extrêmes multivariés. Une autre fonction importante pour représenter la structure de dépendance extrême entre deux composantes d'un couple est la fonction de Pickands [43], notée A et définie sur $[0, 1]$ par :

$$A(t) = V\left(\frac{1}{1-t}, \frac{1}{t}\right)$$

Cette fonction vérifie les propriétés suivantes ([3]) :

- A est une fonction convexe de $[0, 1]$ dans $[1/2, 1]$
- $\forall t,\ \min(t, 1-t) \leq A(t) \leq 1$
- $A(0) = A(1) = 1$
- $-1 \leq A'(0) \leq 0$ et $0 \leq A'(1) \leq 1$

Un point important à noter est que les extrêmes bivariés ne sont pas décrits par une famille paramétrique mais par une équation fonctionnelle. D'un point de vue statistique, il est intéressant de développer des modèles paramétriques, sans pour autant sacrifier la flexibilité et l'importante variabilité possible pour les fonctions de dépendance extrême.

Cette approche paramétrique est d'autant plus importante qu'il est difficile d'estimer non-paramétriquement H (ou V) du fait des contraintes existant sur ces fonctions, bien que des méthodes aient été développées pour estimer ces fonctions ([43], [6], [29]).

Des exemples de fonctions de dépendance sont disponibles dans [3] ou encore la documentation du package evd [1] pour le logiciel **R** dans lequel plus de 8 modèles sont disponibles.

2.1.2 Estimation

Le choix d'un modèle paramétrique sur V permet de faire de l'extrapolation au-delà du support des observations, et afin de comparer les influences des variables, mais ce au prix d'une perte en flexibilité comparativement aux modèles non-paramétriques. Ces derniers sont préférés pour guider le choix d'un modèle paramétrique.

En ce qui concerne l'estimation des paramètres, deux approches sont possibles : estimer les marges indépendamment par une des méthodes décrites dans la partie sur les lois univariées, transformer les données en marges Fréchet grâce à la proposition 2.1, puis maximiser la vraisemblance (calculée grâce à 2.1) pour un modèle paramétrique donné. Une approche alternative consiste à faire l'estimation des lois marginales et de la structure de dépendance en même temps, en introduisant la transformation des marges dans la vraisemblance. On peut penser que la seconde approche donnera de meilleurs résultats, surtout en ce qui concerne les erreurs d'estimation que la première approche aura tendance à sous-estimer, mais ceci au prix d'une fonction de vraisemblance plus compliquée à estimer (le nombre de paramètres est de 6 sans compter ceux décrivant la dépendance), avec le risque que l'optimisation mène à un maxima local de la fonction de vraisemblance. Une comparaison de ces deux approches dans le cas de l'estimation de copule est établie par Joe [33].

2.1.3 Autres fonctions de dépendance

Il n'est en pratique pas évident de choisir une famille paramétrique, car peu de familles sont assez souples pour permettre à la fois de la dépendance asymptotique et de l'indépendance, suivant la valeur d'un ou des paramètres. C'est la raison pour laquelle des outils ont été développés pour quantifier ce niveau de dépendance, c'est ce qui fera l'objet de ce paragraphe. Dans ce qui suit, (X, Y) est un couple de variables aléatoires de fonctions de répartitions marginales F_X et F_Y respectivement.

La fonction χ

Définition 2.1. Soit $u \in]0, 1[$, et soit $\chi(u) = \mathbb{P}\left(F_2(Y) > u | F_1(X) > u\right)$.

On appelle coefficient de dépendance de queue la quantité $\chi \underset{u\to 1}{=} \chi(u)$. On dit que X et Y sont asymptotiquement indépendantes si $\chi = 0$. Au contraire, si X et Y sont parfaitement dépendants, alors $\chi = 1$.

On peut résumer les propriétés de χ (voir [9, 10]) :

1. $0 \leq \chi \leq 1$
2. $\chi = 0$ pour des variables asymptotiquement indépendantes, sans pour autant que les deux variables soient indépendantes.
3. Si $\chi > 0$, alors il mesure le degré de dépendance entre les deux composantes aux niveaux élevés.

1. http ://cran.r-project.org/web/packages/evd/evd.pdf

4. Si (X, Y) appartient à la famille des lois extrêmes bivariées de fonction de dépendance V, alors $\chi = 2 - V(1, 1)$.

Même si cette quantité est intéressante pour quantifier le degré de dépendance entre deux variables, elle est en pratique difficile à estimer en partant d'un graphique de $\chi(u)$: en effet, il peut arriver que la fonction $\chi(u)$ converge très lentement vers 0, c'est le cas par exemple d'un couple de variables gaussiennes, et dans ce cas on aura tendance à conclure à tort que les observations sont dépendantes. C'est la raison de l'introduction d'une quantité complémentaire.

La fonction $\bar{\chi}$

Définition 2.2. Soit $u \in]0, 1[$, et soit $\bar{\chi}(u) = \frac{2\mathbb{P}(F_X(X)>u)}{\mathbb{P}(F_X(X)>u, F_Y(Y)>u)} - 1$.
On note alors $\bar{\chi} \underset{u\to 1}{=} \bar{\chi}(u)$.

A la différence de χ, ce coefficient vaut 1 quand les variables sont asymptotiquement dépendantes, et vaut toujours 0 dans le cas indépendant (asymptotiquement). Cette fois ci, un graphique de $\bar{\chi}(u)$ nous apporte de l'information quand les variables sont asymptotiquement indépendantes, puisque la limite en 1 donne une indication du degré de dépendance entre X et Y.

Ces deux objets sont ainsi complémentaires, puisqu'ils décrivent chacun le degré de dépendance dans les deux classes limites possibles de comportement limite :

- Si $\chi = 0$, les variables sont asymptotiquement indépendantes et $\bar{\chi}$ donne le degré de dépendance.
- Si $\bar{\chi} = 1$, les variables sont asymptotiquement dépendantes et χ donne le degré de dépendance.

2.2 Dépassements de seuils bivariés

Comme dans le cas univarié, la modélisation par les maxima par bloc présente l'inconvénient de laisser de côté un grand nombre d'observations, qui peuvent contenir de l'information, par exemple dans le cas où deux observations extrêmes sont observées dans le même bloc. Pour palier ce problème, nous allons présenter une extension de l'approche par dépassement de seuil que nous avons déjà étudiée dans le cas univarié au cas où les observations sont issues d'un couple de variables aléatoires.

Nous présenterons l'approche retenue ici par Coles [10] dont nous reprenons les notations. Nous présenterons dans la suite de ce document une approche alternative. Pour le moment, nous allons considérer un couple de variables aléatoires (X, Y), et des observations $(x_1, y_1), \ldots, (x_n, y_n)$ i.i.d issues de ce couple de variables aléatoires. Pour deux seuils u_x et u_y qui peuvent être choisis indépendamment en suivant la procédure indiquée dans la partie 1.2, chacune des composantes peut être approchée par une loi GPD de paramètres respectifs $\theta_X = (\lambda_X, \sigma_X, \xi_X)$ et $\theta_Y = (\lambda_Y, \sigma_Y, \xi_Y)$, de telle sorte que l'on peut définir un nouveau couple de variables aléatoires $(\widetilde{X}, \widetilde{Y})$ dont chacune des lois marginales est distribuée approximativement suivant une loi Fréchet unitaire :

$$\begin{aligned}
\widetilde{X} &= -\left(\log\left\{1 - \lambda_X\left[1 + \xi_X\left(\frac{X - u_x}{\sigma_X}\right)\right]^{-1/\xi_X}\right\}\right)^{-1} \\
\widetilde{Y} &= -\left(\log\left\{1 - \lambda_Y\left[1 + \xi_Y\left(\frac{Y - u_Y}{\sigma_Y}\right)\right]^{-1/\xi_Y}\right\}\right)^{-1},
\end{aligned}$$

et la fonction de répartition de ce couple sera notée $\widetilde{F}(.,.)$ et en utilisant (2.1), on a :

$$\begin{aligned}\widetilde{F}(\widetilde{x},\widetilde{y}) &= \left[\widetilde{F}^n(\widetilde{x},\widetilde{y})\right]^{1/n} \\ &\approx \left[\exp\left\{-V(\widetilde{x}/n,\widetilde{y}/n)\right\}\right]^{1/n} \\ &= \exp\left\{-V(\widetilde{x}/n,\widetilde{y}/n)\right\}\end{aligned}$$

car la fonction V est homogène d'ordre -1. On en déduit donc une approximation de la fonction de répartition bivariée au-delà d'un certain seuil pour $\tilde{F}$ et donc pour F. Cette procédure est plus complexe et moins satisfaisante que dans le cas univarié, car il n'existe pas de description précise de la famille des lois acceptables pour les dépassements de seuils. Il est cependant possible de calculer la densité associée à une telle modélisation, en distinguant suivant les cas :

$$\widetilde{f}(\widetilde{x},\widetilde{y}) = \begin{cases} \widetilde{F}(\lambda_x,\lambda_y) & \text{si } x \le u_x,\ y \le u_y \\ \frac{\partial \widetilde{F}}{\partial \widetilde{x}}(\widetilde{x},\lambda_y) & \text{si } x \le u_x,\ y > u_y \\ \frac{\partial F}{\partial \widetilde{y}}(\lambda_x,\widetilde{y}) & \text{si } x > u_x,\ y > u_y \\ \frac{\partial^2 \widetilde{F}}{\partial \widetilde{x}\partial \widetilde{y}}(\widetilde{x},\widetilde{y}) & \text{si } x > u_x,\ y \le u_y \end{cases} \tag{I.3}$$

$$f(x,y) = \begin{cases} \widetilde{f}(\widetilde{x},\widetilde{y}) & \text{si } x \le u_x,\ y \le u_y \\ \frac{\partial \widetilde{x}}{x}(x).\widetilde{f}(\widetilde{x},\widetilde{y}) & \text{si } x > u_x,\ y \le u_y \\ \frac{\partial \widetilde{y}}{y}(y).\widetilde{f}(\widetilde{x},\widetilde{y}) & \text{si } x \le u_x,\ y > u_y \\ \frac{\partial \widetilde{x}}{x}(x).\frac{\partial \widetilde{y}}{y}(y).\widetilde{f}(\widetilde{x},\widetilde{y}) & \text{si } x > u_x,\ y \le u_y \end{cases} \tag{I.4}$$

Il est ensuite possible, par un choix adapté de la fonction de répartition bivariée, à choisir par les modèles présentés précédemment, d'utiliser cette expression pour calculer la vraisemblance associée à un échantillon et donc d'estimer les paramètres du modèle associé. Il n'existe pas à notre connaissance de références concernant l'estimation des paramètres, mais cette approche a été adaptée pour traiter différentes données, en particulier issues du domaine de la finance. Pour plus de détails, cf. [10].

3 Modélisation des extrêmes spatiaux

Nous venons d'exposer les résultats théoriques et pratiques développés pour modéliser les queues de distribution, aussi bien dans le cas univarié classique que dans le cas, un peu moins traité en pratique, des extrêmes multivariés. Nous allons dans la partie qui suit présenter les méthodes disponibles pour traiter des maxima observés sur une zone, car cette question est d'intérêt dans de nombreux domaines où l'on peut mesurer une quantité sur tout l'espace, et donc en particulier en météorologie. Nous ne traiterons pas le cas des dépassements de seuils dans ce cadre, car cette problématique, quoique très importante en pratique, ne dispose pas encore de réponse. Les observations sont alors constituées de cartes de maxima, disponibles par exemple pour différentes années, ces différentes réalisations étant supposées indépendantes. Il existe deux façons principales de traiter le cas des données spatialement réparties : la première est de supposer l'existence d'un processus spatial de paramètres cachés, la seconde étant de chercher une extension des lois max-stables, coeur de la modélisation dans les cas uni- et bi-variés, pour le processus des maxima.

3.1 Processus spatial latent de paramètres

Etant donné un processus de maxima $Z(s)$, où s est un point de $\mathbb{R}^d$, il peut sembler légitime d'utiliser des processus latents de paramètres, de sorte que le modèle peut s'écrire :

$$\forall s,\ Z(s) \sim \text{GEV}(\mu(s), \sigma(s), \xi(s)), \tag{I.5}$$

où les processus $\mu(s), \sigma(s)$ et $\xi(s)$ sont des processus cachés. L'hypothèse importante de ces modèles est alors que les observations en deux sites s_i et s_j soient indépendantes conditionnellement à ces processus. Une telle démarche est relativement usuelle dans le domaine de la géostatistique. Dans ce domaine d'application en effet, il est fréquent que les lois marginales soient supposées gaussiennes, de moyenne $\mu(r)$, un processus gaussien lisse. Un travail classique dans ce domaine est celui de Diggle [20].

L'application à des modèles hiérarchiques pour la modélisation des valeurs extrêmes est plus récente. Les premières incursions dans le domaine datent de 1999, par Casson et Coles [7] dans le cas d'un modèle unidimensionnel. Les travaux plus récents sont plus nombreux : citons Cooley (2007) [11], Sang et Gelfand (2007) [51] ou Mendes et al. (2008) [38]. Ce dernier inclut une prise en compte de la variation temporelle des paramètres GEV en plus d'introduire une dépendance spatiale. Dans le même esprit, un autre travail, réalisé par Gaetan et al. [28], introduit une variation continue de la dépendance en espace des paramètres, en plus d'une variation temporelle, permettant d'analyser l'importance du changement climatique.

L'approche d'un processus latent pour modéliser les variations spatiales a l'avantage d'être relativement flexible et semble extensible, notamment en ce qui concerne l'ajout de covariables. Cependant, il semble peu probable qu'un tel modèle soit à même de modéliser une dépendance spatiale forte du fait du modèle spatial adopté, et de manière générale, l'hypothèse d'indépendance des observations conditionnellement aux valeurs des paramètres est une hypothèse forte et difficilement vérifiable. Cependant, de par sa flexibilité, cette approche peut permettre de modéliser des dépassements de seuils dans un contexte spatio-temporel, en remplaçant la loi de $Z(s)$ par la loi conditionnelle au-delà d'un seuil $u(s)$ et en remplaçant la loi GEV par la loi GPD.

3.2 Processus Max-Stable

L'idée de cette caractérisation est d'étendre les idées des cas 'classiques' uni- et multidimensionnels au cas où les observations proviennent d'un champ aléatoire : description des processus admissibles, caractérisation de la convergence, établissement du domaine d'attraction...

Définition 3.1. Un processus $X(r), r \in \mathbb{R}^d$ de loi marginale une loi de Fréchet unitaire est dit *max-stable* s'il a même loi que $Z_n(r) = \max_{i=1,\dots,n} \left\{ \frac{X_i(r)}{n} \right\}$, avec $X_1(r), \dots, X_n(r)$ des réplications indépendantes de $X(r)$. De même que dans le cas unidimensionnel, on a que si $Z_n \xrightarrow{d} Z$ où Z est un processus bien défini, alors Z est un processus max-stable.

Le point important de cette théorie est la représentation spectrale, due à de Haan [14], qui caractérise complètement les processus max-stables, analogue spatial du théorème de Fisher-Tippett :

Théorème 3.1. Soient $(u_1, s_1), (u_2, s_2), \dots$ les points d'un processus de Poisson sur $(0, \infty) \times S$, où S est un espace mesurable arbitraire. L'intensité du processus est

$$\lambda(du, ds) = \frac{du}{u^2} \times \nu(ds),$$

avec ν une fonction positive d'intensité sur S. Soit encore une fonction f positive de S dans $\mathbb{R}$ telle que $\int_S f(s)\nu(ds) = 1$.

Alors le processus $Z(r) = \max_i \{u_i f(s_i - r)\}$ est un processus max-stable de loi marginale Fréchet unitaire. De plus, tout processus max-stable peut être représenté de cette façon, pour un certain choix de S, ν et f.

On peut montrer la propriété suivante ([15]) :

Propriété 3.1. Soit $k \in \mathcal{N}$ et un ensemble fini d'indices $r_1, \ldots, r_K \in \mathbb{R}^d$ et des valeurs positives $z_1, \ldots, z_K \in \mathbb{R}$. Alors la distribution du vecteur aléatoire $(Z(r_1), \ldots, Z(r_K))'$ est donnée par sa fonction de répartition :

$$\mathbb{P}(Z(r_k) \leq z_k, k = 1, \ldots, K) = \exp\left[-\int_S \max_{k=1,\ldots,K}\left\{\frac{f(s-r_k}{z_k}\right\}\nu(ds)\right]. \tag{I.6}$$

On en déduit en particulier que le processus est de loi marginale Fréchet unitaire :

$$\mathbb{P}(Z(r) \leq z) = \exp\left[-z^{-1}\int_S f(s-r)\nu(ds)\right] = \exp(-1/z). \tag{I.7}$$

Smith ([55]) donne l'interprétation suivante du théorème précédent : l'ensemble S est un ensemble de "centre de tempêtes" ; la mesure ν décrit la répartition spatiale des tempêtes sur l'espace S ; λ_i représente l'intensité de la i$^{\text{ème}}$ tempête ; s_i sa position et $\lambda_i f(s_i, r)$ représente l'impact de cette tempête au point r, par exemple la quantité de pluie provenant de cette tempête en ce point. Ainsi, $f(s, .)$ est le profil de la tempête. La résultante $Z(r)$ est donc le maxima sur toutes les réalisations possibles de tempêtes. Comme on peut le voir, la famille des processus max-stables est décrite par plusieurs fonctions et non par un paramètre d'un espace de dimension finie. Cela rend l'estimation délicate en pratique, et c'est la raison pour laquelle des modèles spécifiques sur S et f ont été développés pour palier ce problème.

Ainsi, Smith ([55]) montre certaines propriétés de la loi du processus Z dans le cas particulier où $S = R = \mathbb{R}^d$, $f(s,t) = f_0(s-t)$ avec f_0 la densité d'une loi normale de dimension d, centrée, de matrice de variance-covariance Λ. Il donne notamment la distribution jointe en deux points, la loi marginale étant une loi de Fréchet unitaire. Cependant, il n'a pas encore été trouvé d'expression analytique exploitable pour plus de deux points, et en tout état de cause, le calcul de la densité pour plus de points aboutit rapidement à une explosion du nombre de termes à calculer. La figure I.3 montre des réalisations d'un tel processus sur $\mathbb{R}$, dont les marges ont été transformées en Gumbel pour améliorer la représentation : on voit notamment l'influence de la forme prise pour la forme des tempêtes, ainsi que l'influence de la valeur du paramètre.

Des avancées récentes [53] étendent ce modèle en intégrant le cas où la covariance varie suivant les points. Un autre modèle, dû à Schlather ([52]) permet d'introduire un processus gaussien sous-jacent décrit par une fonction de corrélation qui intervient dans la fonction de répartition jointe en deux points. Le problème de cette approche est de ne pas obtenir l'indépendance asymptotique. Citons également Kabluchko [34] qui permet une généralisation de la représentation de Smith ou le processus de Brown-Resnik ([5]) qui est la limite du maxima de processus d'Orstein-Uhlenbeck. Un des travaux importants dans ce domaine est réalisé par Stoev ([59]) qui montre sous certaines conditions, l'ergodicité et le mélange de certaines familles de processus max-stables, et en particulier pour le processus de Smith.

Peu d'articles traitent de la question de l'estimation des paramètres d'un processus max-stable. On peut citer à ce sujet les travaux de [16] et plus récemment [40] qui traitent

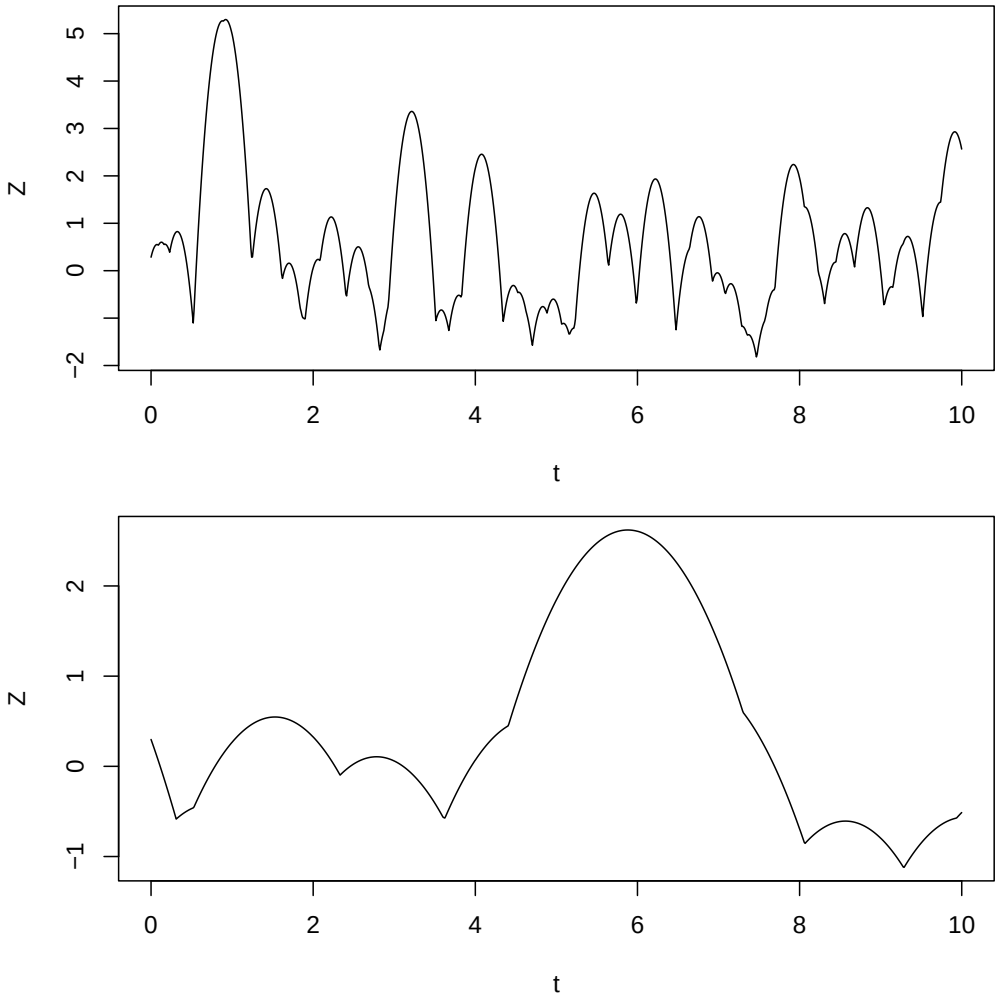


FIGURE I.3 – Réalisations de processus max-stables de marges Gumbel sur $\mathbb{R}$ pour $\sigma^2 = 0.01$ (haut) et $\sigma^2 = 1/2$ (bas)

respectivement d'estimation non-paramétrique et d'estimation par maximum de vraisemblance composite. Pour plus de détails sur cette méthode, on peut se référer à l'article fondateur de Linsday ([36]), [12] pour l'application à des données spatiales, et enfin [62] pour une discussion plus générale sur l'application de cette méthode.

Un des défauts majeurs de cette approche semble être la difficulté d'interprétation des résultats, au delà de l'interprétation de la forme de la fonction f : bien que convenable d'un point de vue statistique, cette théorie est relativement complexe et ne permet de disposer que de la structure bivariée, et seulement pour quelques modèles. Il existe un package **R** dédié à ces modèles [2]. On notera cependant que les travaux se plaçant dans ce contexte sont nombreux et que les évolutions sont rapides. Il a récemment été proposé ([64]) une méthode pour la simulation conditionnelle d'un processus max-stable, ouvrant encore plus la voie vers les applications, en permettant la prédiction quand des observations sont données. Une des faiblesses de cette approche est que la modélisation des dépassements de seuils n'a pas été traitée dans la littérature, et seule la méthode des maxima par bloc est disponible.

2. `http://spatialextremes.r-forge.r-project.org/`

Nous avons vu différentes approches qui permettent de modéliser les extrêmes du cas univarié jusqu'au cas spatial, voire même spatio-temporel. Cependant, tous ces modèles font l'hypothèse que l'on dispose d'observations IID, même si les observations sont des cartes de maxima. Une des solutions couramment évoquée est de grouper les observations par blocs, typiquement de taille commune d'un an pour les applications en météorologie. Mais se pose alors la question du choix de la taille des blocs (an, mois, semaine,...) permettant de considérer les observations comme indépendantes, tout en conservant suffisamment de données pour l'estimation. Nous allons donc dans la partie suivante nous intéresser à la modélisation d'extrêmes dans le cas où les observations sont dépendantes.

4 Modélisation d'extrêmes de séries temporelles

Dans bien des cas, il est dommage de sacrifier de l'information sur le comportement au cours du temps contenue dans les données, en ne considérant par exemple que des données de maxima annuels. En effet, les dommages les plus importants sont souvent causés par une succession rapide d'événements extrêmes (vents, vagues ou pluies en sont de très bons exemples). Si l'on considère donc toutes les données élevées par rapport au comportement moyen des observations, il est peu probable que l'hypothèse d'indépendance soit réaliste : il faut donc des extensions des modèles vus précédemment à ce cas-là. Il ne semble pas exister de littérature sur le sujet couplé dépendance spatiale-dépendance temporelle, nous allons donc faire une étude de la littérature disponible dans le cas de données ponctuelles, en dégageant les deux méthodes principales existant pour prendre en compte la dépendance des extrêmes d'une série stationnaire :

Définition 4.1. Une série $(X_1, \ldots)$ est dite stationnaire, si pour tous entiers i, j, k, le vecteur $(X_i, \ldots, X_{i+k})$ a même loi que $(X_j, \ldots, X_{j+k})$.

Cette hypothèse ne doit pas être confondue avec la stationnarité d'ordre deux, qui correspond à l'invariance au cours du temps de la moyenne, la variance, l'auto-corrélation, etc. Cette dernière définition est moins forte que celle que nous avons donné, mais est difficilement applicable vu que les lois adaptées aux extrêmes n'ont pas nécessairement de moments finis, en particulier ceux d'ordre deux.

4.1 Indice extrémal

Sous une contrainte de non-dépendance à long terme, Leadbetter ([35]) énonce le théorème suivant, qui permet de traiter les maxima par blocs de façon similaire au cas initial lorsque nous sommes en présence d'observations dépendantes :

Définition 4.2 (Condition $\mathcal{D}(u_n)$)**.** Une série stationnaire $(X_1, \ldots, X_n)$ satisfait la condition $\mathcal{D}(u_n)$ si quels que soient les entiers

$$i_1 < \ldots < i_p < j_1 < \ldots < j_q \text{ avec } j_1 - i_p > l_n,$$

$$\begin{aligned} |\mathbb{P}(X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n, X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n) \\ - \mathbb{P}(X_{i_1} \leq u_n, \ldots, X_{i_p} \leq u_n)\mathbb{P}(X_{j_1} \leq u_n, \ldots, X_{j_q} \leq u_n)| \leq \alpha(n, l_n), \end{aligned} \quad \text{(I.8)}$$

où $\alpha(n, l_n) \to 0$ pour une suite l_n telle que $l_n/n \to 0$ quand $n \to \infty$.

Théorème 4.1 (Leadbetter, 1983)**.** Soient $(X_1, \ldots, X_n)$ une série stationnaire vérifiant la condition $\mathcal{D}(u_n)$ ci-dessus et $(X_1^*, \ldots, X_n^*)$ une série **indépendante** de même distribution marginale que $(X_1, \ldots, X_n)$. Soient $M_n = \max_{i=1,\ldots,n} X_n$ et $M_n^* = \max_{i=1,\ldots,n} X_n^*$. S'il existe des suites $(a_n)_{n\in\mathbb{R}} > 0$ et $(b_n)_{n\in\mathbb{R}}$ telles que

$$\lim_{n\to\infty} \mathbb{P}\Big\{a_n^{-1}(M_n^* - b_n) \leq x\Big\} \to H_1(x) \tag{I.9}$$

et

$$\lim_{n\to\infty} \mathbb{P}\Big\{a_n^{-1}(M_n - b_n) \leq x\Big\} \to H_2(x), \tag{I.10}$$

alors il existe un paramètre appelé *extremal index* ou indice extrémal, noté θ, à valeur dans $(0, 1]$ tel que $H_2 = H_1^\theta$.

On déduit du théorème précédent qu'il existe un lien entre les paramètres de positions, d'échelle et de forme entre les distributions limites H_1 et H_2, en fonction de la valeur du paramètre que nous venons d'introduire. Ce paramètre θ est souvent interprété comme l'inverse de la taille moyenne des blocs d'extrêmes et apporte par là même une caractérisation importante du processus des dépassements, cette interprétation étant justifiée par les travaux de Hsing, 1988 [32]. Il donne une mesure de la dépendance à court terme entre les extrêmes, mais il est à interpréter avec prudence, car s'il est égal à un pour des variables indépendantes, la réciproque n'est pas vérifié.

Nous pouvons citer plusieurs estimateurs de ce paramètre : [58] propose un estimateur de cet indice basé sur le nombre de dépassements consécutifs, [26] propose un estimateur sans paramètre supplémentaire à fixer, alors que [48] propose un estimateur mobile. D'autres méthodes pour l'estimer sont décrites dans [3].

Les conséquences pratiques du théorème précédent sont une extension de la méthode **POT** aux données dépendantes : une première méthode est de considérer que les extrêmes proches en temps sont issus d'une même observation, et de définir ainsi de nouvelles observations, supposées indépendantes, et dont la loi est une loi de Pareto. Par rapport à la méthode décrite dans le cas d'observation indépendante, seule l'estimation de la probabilité de dépasser le seuil, λ_u, se trouve modifiée. Cette méthode est appelée 'dégroupement' (*declustering*), et l'on voit qu'il est nécessaire de définir précisément un cluster de valeurs extrêmes, et surtout d'être capable de les identifier précisément. On peut noter un travail de Fawcett [26] qui a montré que l'estimation de paramètres d'une loi GEV est sensible au choix du critère de dégroupement.

4.2 Modèles sur les dépassements

Une approche alternative consiste à proposer un modèle sur la dépendance des dépassements de seuils, typiquement un modèle markovien ([39, 56, 41] ou plus récemment [24, 25, 46]), des modèles type 'Moyenne Mobile' ([49]) ou des modèles ARCH ([17]). Ces modèles ont l'avantage de conserver toutes les données, alors que les modèles de *declustering* occasionnent une perte d'information. Cependant, il peut être difficile de trouver un modèle sur les observations dépassant le seuil pour lequel des informations sur son comportement extrême sont disponibles. Le grand avantage de ces modèles est de prendre en compte toutes les valeurs qui dépassent un seuil, tout en proposant une prise en compte de la dynamique des extrêmes, qui peut elle-aussi être d'intérêt dans de nombreuses applications.

Une autre modélisation possible est de s'intéresser au temps d'occurrence des dépassements de seuil, principalement pour décrire la fréquence des événements extrêmes. Dans ce

cadre, les temps d'occurrence sont modélisés par un processus de Poisson non-homogène dont l'intensité est modifiée lors de l'apparition d'un évènement extrême afin de prendre en compte la succession des dépassements de seuils (voir [37]). L'application de Luceno à des données de vagues montre une forte amélioration de l'adéquation aux données par rapport à la méthode POT classique. Le problème de cette approche est de ne permettre de modéliser que les temps d'occurrence et non pas les niveaux de dépassement de seuil. Mendes ([38]) introduit une vraisemblance prenant en compte les deux aspects, dans le cadre où l'espace est discrétisé, c'est-à-dire découpé en régions.

5 Conclusions du chapitre

Comme nous avons pu le voir dans le bref rappel bibliographique précédent, les approches pour modéliser le comportement extrême d'un jeu de données sont nombreuses et complémentaires, mais ont pour but commun de proposer des approximations, si possible paramétriques, des queues de distributions afin de permettre l'extrapolation au-delà du support des observations. Le choix de tel ou tel modèle est laissé au praticien, suivant l'objectif recherché, mais surtout suivant les données disponibles. Ces dernières en effet apportent leur lot de contraintes, et il est souvent nécessaire de chercher l'approche la plus adaptée pour répondre au problème de modélisation des valeurs extrêmes. Les conditions d'applications des différents théorèmes énoncés précédemment peuvent êtres ardues à vérifier, mais cette étape est cependant indispensable à effectuer afin de pouvoir accorder un crédit aux extrapolations faites.

Le chapitre qui va suivre va s'attacher à décrire les données utilisées dans cette étude, les avantages, inconvénients et contraintes apportées par chacune d'entre elles en ce qui concerne l'estimation de quantités extrêmes.

Chapitre II

Données de hauteur significative des vagues

1 Données

1.1 Hauteur significative des vagues

Pour un état de mer donné sur une zone où l'on considère qu'il est stationnaire en espace, on peut donner différents paramètres qui permettent de caractériser cet état, et en particulier de décrire la houle qui s'y trouve. C'est le rôle de la hauteur significative, notée H_s qui est définie comme quatre fois l'écart-type de la hauteur des vagues. Cette définition permet de correspondre avec la valeur que donnerait un navigateur expérimenté au sujet de l'état de mer étudié. Sous certaines hypothèses quant à la distribution des hauteurs de vagues, elle correspond également à la moyenne du tiers des vagues les plus hautes. Il est utile de noter que la hauteur des vagues n'est pas la seule quantité d'intérêt dans un état de mer, car les vagues sont réparties suivant leur direction, mais aussi suivant leur période, c'est la raison pour laquelle de nombreux appareils servent à mesurer le spectre directionnel, c'est-à-dire la répartition de l'énergie suivant les directions et les périodes. La hauteur significative, qui peut aussi s'interpréter comme une mesure de l'énergie, est une valeur intégrée sur ce spectre, et, par ce calcul même, induit une perte d'information quant à la description de l'état de mer.

La hauteur significative est néanmoins une quantité très utilisée en pratique, que ce soit en ce qui concerne les mesures, mais aussi les modèles numériques, car il existe des algorithmes permettant de reconstituer la distribution des hauteurs à partir du H_s. De plus, étant donné qu'elle s'interprète couramment comme une mesure de l'énergie véhiculée par les vagues, elle est utilisée pour calibrer des structures marines, ou encore pour utiliser l'énergie des vagues, ce qui est couramment appelé SREV (Système de Récupération de l'Energie des Vagues).

Nous allons par la suite nous attacher à décrire les sources de données que nous avons utilisées pendant cette thèse.

1.2 ERA-Interim

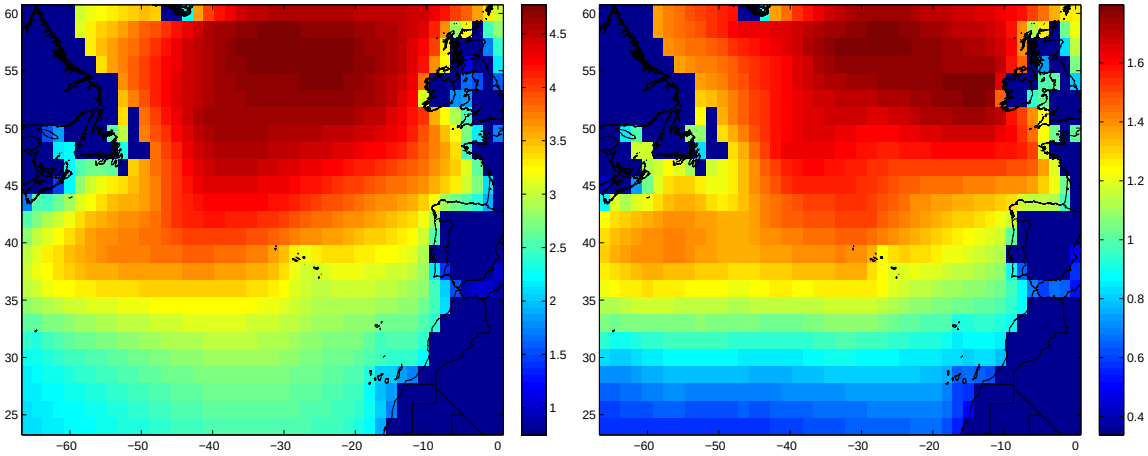
1.2.1 Description

La première source de données est un modèle de réanalyse dit de *hindcast*. Elle est issue d'un projet d'assimilation de données météorologiques coordonné par l'ECMWF [1] qui vise à intégrer toutes les données historiques disponibles, grâce à un modèle numérique complexe, l'atmosphère tout comme l'océan, ainsi que les diverses interactions aux frontières. Les données utilisées pour calibrer ce modèle sont nombreuses, et contiennent en particulier les données issues des satellites et des bouées que nous décrirons et utiliserons par la suite. Ce jeu de données est, de part sa nature, couteux à produire, en temps de calcul comme en espace de stockage mais permet d'obtenir les valeurs de nombreuses quantités physiques, comme la hauteur significative des vagues, la température, la vitesse du vent, etc., et ce sur une grille régulière en temps et en espace. Nous allons nous intéresser dans un premier temps à la description du comportement extrême des vagues grâce aux modèles classiques de la littérature sans prendre en compte la structure spatiale des données C'est-à-dire que chaque point de la grille est vu comme une série temporelle de données, indépendante des séries des points adjacents. Ces données sont disponibles sur l'ensemble des océans, avec une résolution spatiale de 1.5°, une résolution temporelle de 6 heures, et est disponible de Janvier 1989 à Avril 2011, au moment où ce document est rédigé. Ces données sont

1. European Centre for Medium-Range Weather Forecasts

librement accessibles pour les organismes de recherche et sont régulièrement mises à jour [2].

1.2.2 Description du comportement extrême



(a) (b)

FIGURE II.1 – (a) : hauteur significative moyenne sur ERA-Interim ; (b) : écart-type de la hauteur significative sur ERA-Interim.

Max annuels Une première étude descriptive du jeu de données consiste donc à étudier les queues de distribution en chacun des points de la grille, avant de comparer aux autres jeux de données décrits pas la suite. Nous avons donc appliqué la méthode des maxima annuels, comme décrite dans le chapitre précédent, en chaque point de la grille, les données retenues étant celles du mois de Décembre. Ce choix a été fait afin de garantir une certaine stationnarité, tout en conservant suffisamment de données pour le calcul du maxima. La méthode d'estimation retenue est le maximum de vraisemblance, car les méthodes exposées ultérieurement se basent sur cet estimateur, ici 22 ans de données.

On peut faire les remarques suivantes sur les cartes de la figure II.1 : les paramètres μ et σ ont une tendance spatiale claire, avec une zone au nord où ces deux paramètres sont plus élevés, alors que le paramètre de queue a un comportement plus compliqué en espace, avec une zone centrale où il est plus faible (approximativement autour du parallèle de 40 ° Nord), bordée de zones où il est plus élevé. On peut distinguer quatre zones au comportement diffèrent sur l'Atlantique Nord. La première, située au nord-est (30 ° à 10 ° Ouest et 50 ° à 60 ° Nord) est une zone où la moyenne est élevée, de même que la variance : les tempêtes y sont fortes, mais aussi très variables. La seconde, située à l'ouest de la précédente (30 ° à 50 ° Ouest, 50 ° à 60 ° Nord), présente une moyenne assez élevée, mais une variance bien plus faible, ce qui correspond à des situations tempétueuses assez fortes comparativement au reste de la carte, mais moins variables que dans la zone décrite précédemment. La troisième zone est la zone intermédiaire, centrée sur le 40^{e} parallèle Nord, qui est caractérisée par une moyenne peu élevée et une variance forte, voire très forte : les évènements dépressionnaires y sont peu forts mais très variables : il existe des situations où de forts extrêmes sont observés, même s'ils doivent être rares. La quatrième et dernière zone se trouve au sud, entre 25 ° et 35 ° Nord et 25 ° à 40 ° Ouest : elle se caractérise par une moyenne et une variance faibles, correspondant à des extrêmes faibles

2. `http://data-portal.ecmwf.int/data/d/interim_daily/`

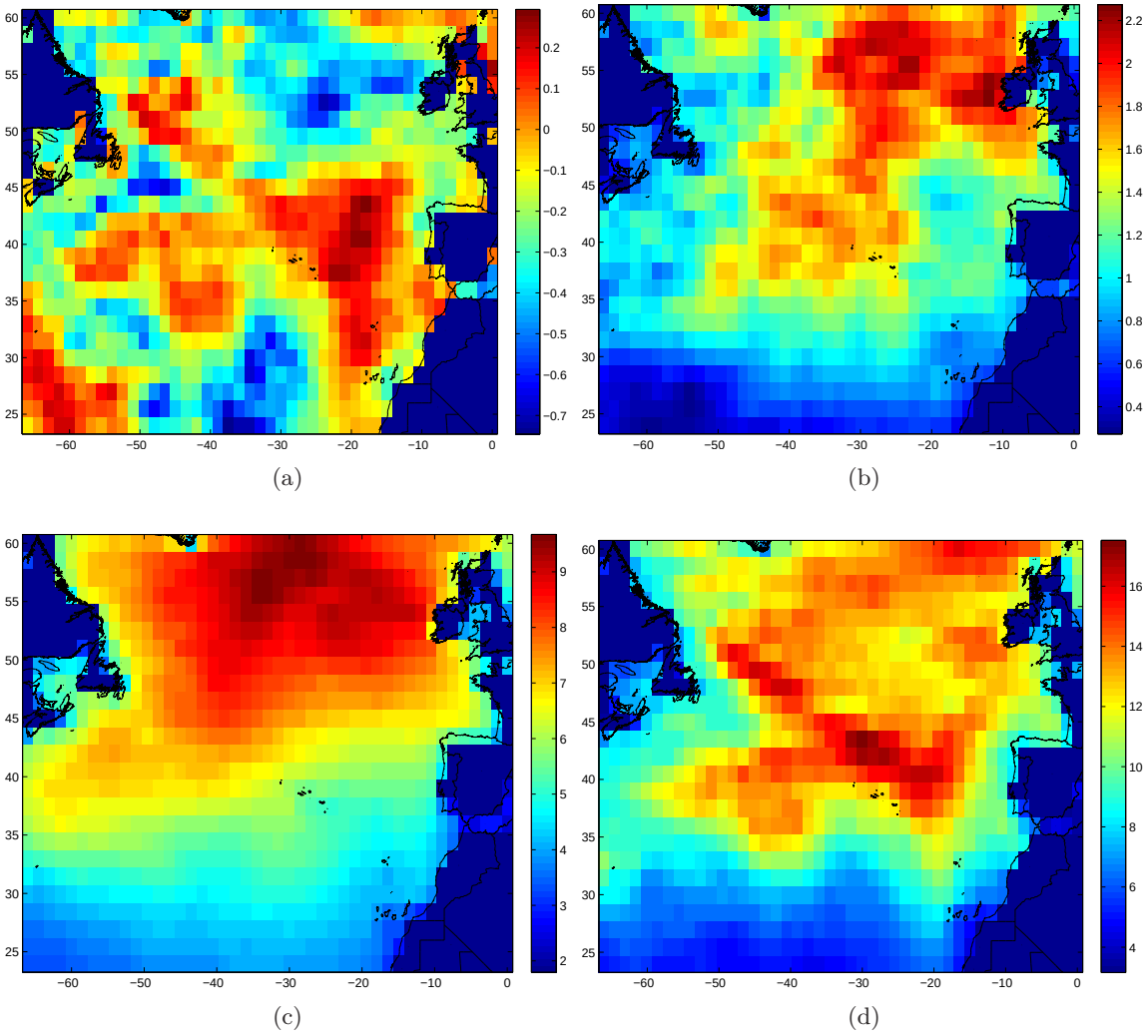


FIGURE II.2 – (a) : carte des paramètres de forme, ξ ; (b) : carte des paramètres d'échelle, σ ; (c) : carte des paramètres de position, μ et (d) : carte des niveaux de retour à 100 ans.

et relativement similaires d'années en années. La carte en bas à droite de cette figure représente le niveau de retour à 100 ans : cette valeur est le niveau dépassé en moyenne une fois en 100 ans, niveau z_p donné par la formule II.1, où $y_p = -\log(1 - 1/N)$ avec N la période de retour (100 ans ici) :

$$z_p = \begin{cases} \mu + \frac{\sigma}{\xi}\left[1 - y_p^{\xi}\right] & \text{si} \quad \xi \neq 0 \\ \mu - \sigma \log y_p & \text{sinon} \end{cases} \tag{II.1}$$

La carte des niveaux de retour apporte des éléments complémentaires pour l'interprétation des paramètres estimés : on observe sur ces cartes que les niveaux de retour les plus élevés se situent en effet là où la moyenne de la loi est la plus élevée (carte des niveaux de retour à 50 ans) , mais aussi là où la variance est la plus élevée (carte des niveaux à 100 ans). L'autre point intéressant est que l'on observe sur ces cartes le caractère borné des distributions dont le paramètre ξ est élevé : en effet, dans la zone — spécifiée précédemment — où ξ est plus faible, on observe que le niveau de retour continue de croître quand on augmente la période de retour, alors que dans la zone où ξ est élevé, le niveau

de retour semble ne plus croître au-delà de 12 à 14 mètres de hauteur significative.

Cette variabilité spatiale est intéressante à plus d'un titre : d'une part, on observe des comportements radicalement différents entre l'est et l'ouest de l'Atlantique Nord, les zones aux queues les plus lourdes étant situées plus près des côtes européennes, ce qui correspond relativement bien à la dynamique des tempêtes, créées —en hiver— le long des côtes canadiennes, et traversant l'Atlantique avant d'atteindre l'Europe, période pendant laquelle les vagues sont formées. Une autre observation importante est la diagonale de niveau de retours élevés par rapport au reste, observée sur la carte des niveaux de retours, allant du Labrador à la pointe sud du Portugal. Cette diagonale est absente des cartes de moyennes et d'écart-types de la figure II.1, signifiant que le comportement extrême n'est d'une part pas stationnaire en espace, et d'autre part montre qu'il existe une zone de vagues plus fortes, zone très probablement liée aux lignes de propagation des tempêtes.

Dépassement de seuils Nous avons indiqué dans le chapitre précédent que la méthode des maxima annuels induisait une forte perte d'informations, c'est pourquoi il est intéressant de regarder également le comportement des données au-dessus d'un certain seuil. Nous avons, en chaque point de la grille, ajusté le modèle **POT** par la méthode du maximum de vraisemblance pour les données dépassant le quantile de 93%. Ce seuil a été choisi en accord avec des études précédentes ([63] et [65]).

Les résultats concernent donc les estimations des paramètres ξ et σ tels que définis dans le chapitre précédent, mais aussi l'estimation du paramètre de dépendance des extrêmes, l'*extremal index*, estimé par l'estimateur de Fawcett-Walshaw ([26]). Les résultats de ces estimations sont représentés sur la figure II.3, avec une estimation du niveau de retour à 100 ans.

Comme lors de l'étude des maxima annuels, on remarque dans un premier temps une certaine cohérence spatiale, c'est-à-dire que chacun des indicateurs précités varie de manière plus ou moins régulière en espace, même si la variabilité est plus importante dans le cas des dépassements de seuils. Une autre remarque concerne la non-stationnarité, qui est également marquée : la zone aux niveaux de retour plus élevés identifiée précédemment au nord-est des Acores est toujours présente, et la dissymétrie est-ouest est encore plus marquée, en particulier en ce qui concerne l'*extremal index*. Rappelons que ce dernier peut s'interpréter comme l'inverse du nombre moyen de points consécutifs au-dessus du seuil, indiquant en quelque sorte une durée des évènements extrêmes. On en déduit que la durée des tempêtes croît lors de leur traversée de l'Atlantique Nord, ce qui correspond bien à l'a priori physique expliqué précédemment. Par contre la carte des indices extrêmes ne montre pas de chemin prédominant emprunté par les tempêtes, contrairement à la carte des niveaux de retours.

Le gain entre le maxima annuel et les dépassements de seuils est appréciable : on remarque que la quantité d'information présente est comparable dans les deux cas, puisque les mêmes remarques s'appliquent aux résultats sur les dépassements de seuil, mais le nombre d'observations utilisées est bien plus important, puisque nous conservons alors plus de 100 observations, à comparer à la vingtaine de la méthode précédente. On dispose de plus d'une information intéressante et complémentaire : le durée moyenne des clusters, qui d'une part influe sur le niveau de retour, et d'autre part permet de préciser le comportement des évènements extrêmes.

Variabilité annuelle Une question que nous n'avons pas encore évoquée est celle de la stationnarité temporelle des données : nous avons en effet supposé dans les deux applications précédentes que les données ne présentaient pas de variation au cours de temps, ce

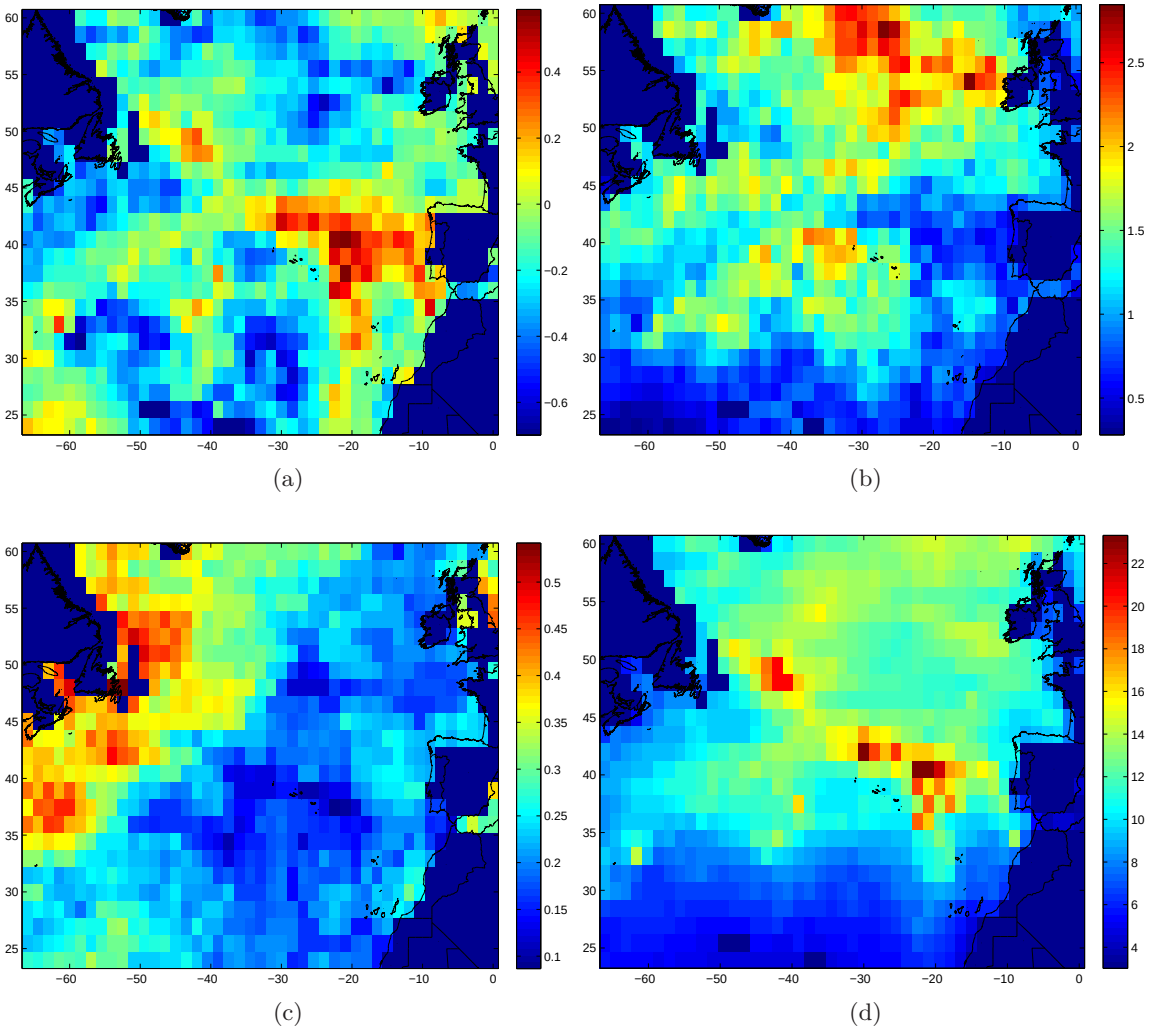


FIGURE II.3 – (a) : carte des paramètres de forme, ξ ; (b) : carte des paramètres d'échelle, σ ; (c) : carte de l'extremal index et (d) : carte des niveaux de retour à 100 ans.

qui permet par exemple de supposer que les maxima annuels sont i.i.d. Or cette hypothèse implique qu'il n'y a ni cycle ni tendance au sein des données. Pour vérifier ce point, nous avons en chaque point ajusté une tendance, linaire en fonction du temps. Vu la période de temps considérée, nous n'avons pas jugé nécessaire de chercher la présence de cycles.

Quatre quantités calculées sur les mois de décembre ont été étudiées afin de discerner les changements du comportement moyen, et ceux du comportement extrême : la hauteur moyenne, l'écart-type de la hauteur, la hauteur maximale et enfin le nombre de dépassements anuels d'un seuil élevé. Les valeurs du paramètre de tendance sont représentées sur la figure II.4, et la p-valeur du test de significativité de ce paramètre se trouvent sur la figure II.5.

Le point commun entre ces différents graphiques est l'absence de tendance décelable, aussi bien en terme de moyenne qu'en terme d'évènements extrêmes, dans la majeure partie de l'Atlantique Nord, tout du moins sur les données ERA dont nous disposons. On observe néanmoins une zone, située au sud-est de l'Atlantique Nord dans laquelle il existe une tendance significativement négative, sur l'ensemble des indicateurs considérés. Comme nous le préciserons par la suite, nous nous sommes focalisés dans la suite de

cette étude sur des données disponibles au large de la pointe bretonne, zone où nous ne décelons pas de problèmes majeurs de non-stationnarité, aussi bien sur la moyenne, que sur le comportement extrême, raison pour laquelle nous maintiendrons cette hypothèse de stationnarité tout au long de notre étude.

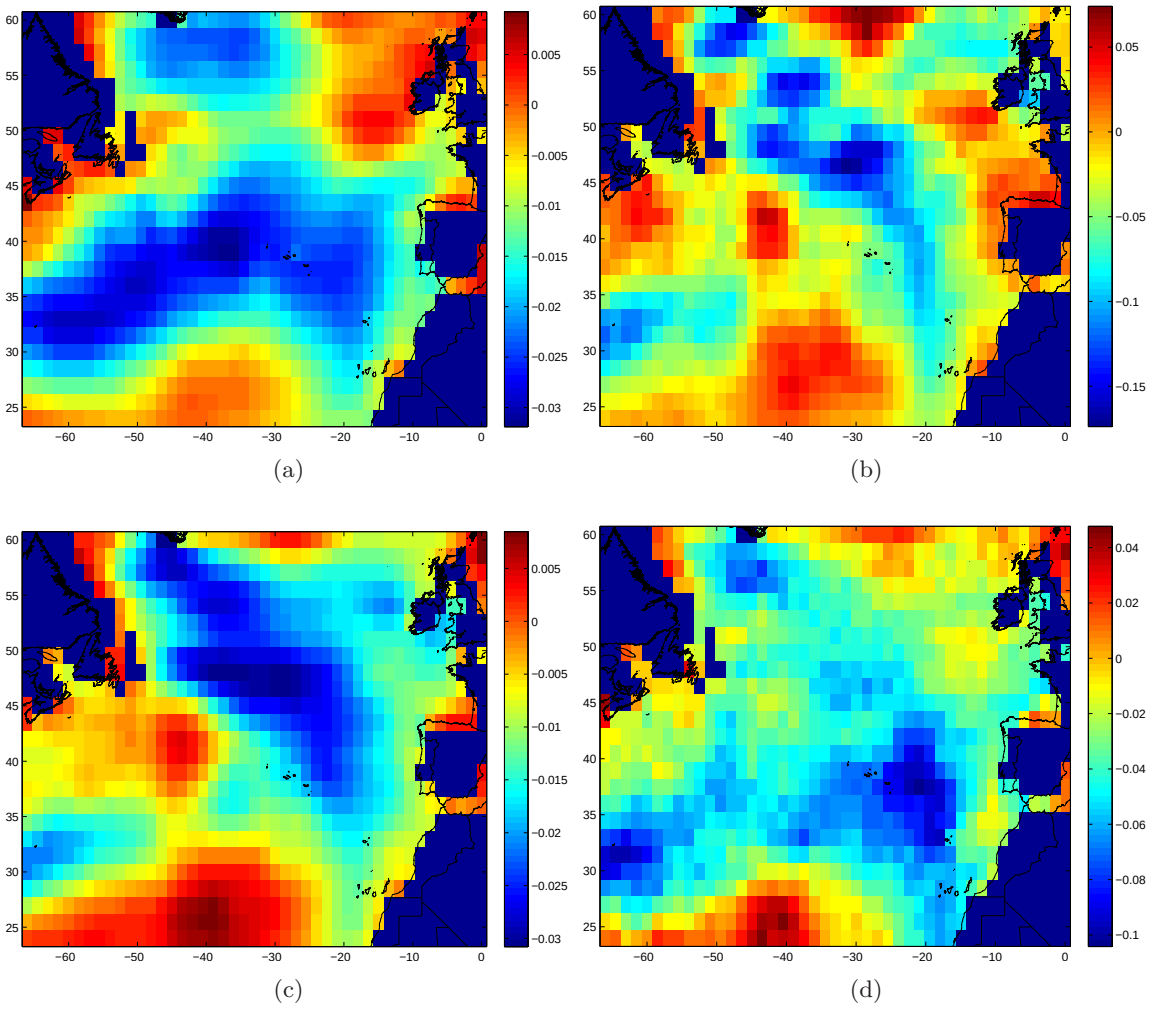


FIGURE II.4 – valeur du paramètre de tendance pour : (a), la moyenne ; (b), le maxima annuel ; (c), l'écart-type et (d), le nombre de dépassement.

1.3 Données de bouées

1.3.1 Description

Une autre source de donnée utilisée dans cette thèse est issue des bouées météorologiques qui sont des bouées équipées de systèmes permettant la mesure de paramètres météorologiques et océanographiques telles que : la pression atmosphérique, la vitesse et la direction du vent, la température de l'air et de la mer, mais aussi la période, la hauteur et la direction des vagues. Nous n'utiliserons que des bouées fixes, c'est-à-dire ancrées au sol marin, qui délivrent donc des mesures en un point donné uniquement, mais ce de manière précise. Les bouées utilisées délivrent des mesures de la hauteur significative toutes les trois heures, avec une précision qui leur fait jouer un rôle de référence pour la calibration

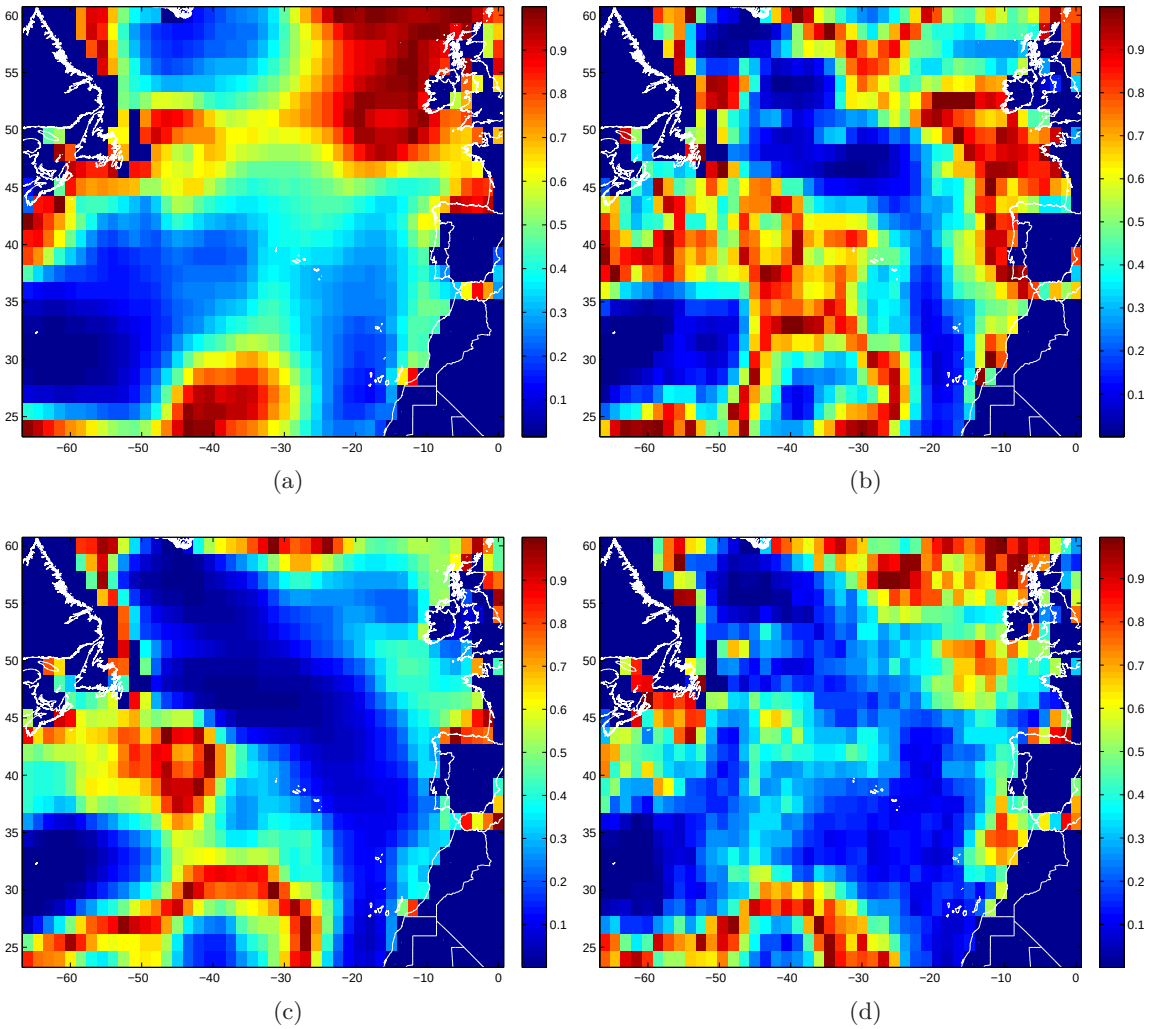


FIGURE II.5 – Significativité du paramètre de tendance pour : (a), la moyenne ; (b), le maxima annuel ; (c), l'écart-type et (d), le nombre de dépassement.

des modèles comme ERA-Interim, où pour les méthodes de mesure à distance (*remote sensing* en anglais) comme les satellites dont nous parlerons par la suite.

L'inconvénient majeur de ces bouées est leur faible couverture spatiale. En effet, leur implantation est une tâche relativement lourde et complexe, et elles nécessitent un entretient régulier car ces structures sont exposées à de fortes contraintes, en particulier lors des tempêtes. Ces raisons sont autant de freins à la multiplication des bouées, ce qui limite fortement la couverture spatiale, comme en témoigne la carte II.6 où l'on constate par exemple qu'il n'y a aucune bouée au centre de l'Atlantique, et donc il n'y a pas de mesure in-situ dans cette zone, pourtant cruciale pour prévoir l'arrivée des tempêtes sur nos côtes.

Nous allons dans un premier temps étudier une série temporelle observée sur la station 62163, la bouée Brittany, située au point 47.5 ° Nord, 8.5 ° Ouest, soit au large des côtes bretonnes. Cette station est détenue et maintenue par le UK Met Office, en coopération avec Météo France. Nous avons obtenu des données sur la période 1995–2005, et pour des raisons de stationnarité, nous n'avons retenu que les mois de décembre de cette période.

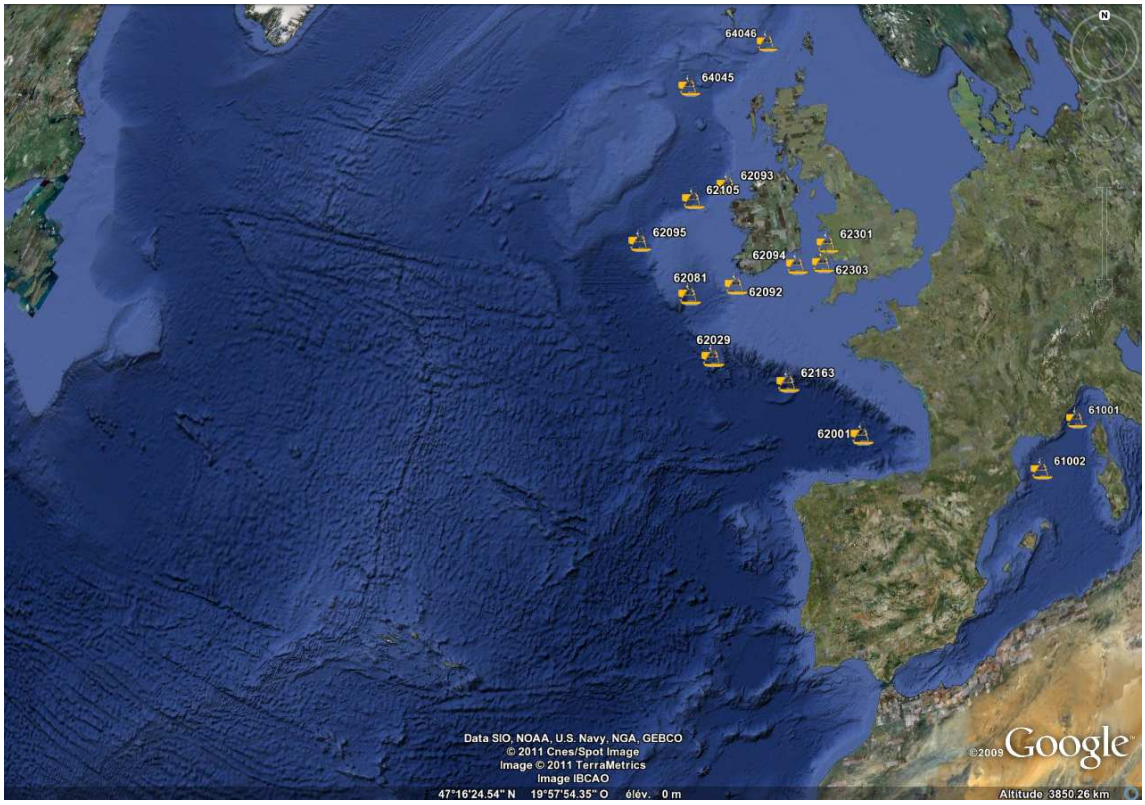


FIGURE II.6 – Répartition des bouées dans l'Atlantique Nord-Est.

1.3.2 Statistiques descriptives

Accord avec ERA-Interim Comme décrit précédemment, ces données sont assimilées dans ERA-Interim, et on devrait donc trouver une grande cohérence entre ces deux jeux de données, mais du fait de la différence de résolution temporelle, il peut y avoir des écarts, d'autant plus que l'on sait que le modèle numérique a tendance à présenter un caractère assez lisse. Ce point est représenté sur la figure II.7, sur laquelle on peut observer d'une part les fréquences d'échantillonnage différentes, mais aussi que ERA-Interim aura tendance à manquer des évènements extrêmes, à cause des deux points précités : fréquence temporelle relativement faible, et caractère lisse. On peut pour illustrer ce point s'en référer à la tempête observée par la bouée le 2 Décembre 2005, qui n'apparait pas complètement sur ERA, probablement à cause de l'échantillonnage temporel. Un peu plus tard, au cours du 9 Décembre 2005, on peut observer que ERA-Interim a des difficultés à coller aux observations, à cause de leur variation rapide en ce point, et on observe également plus tard dans le mois des décalages temporels plus ou moins importants entre ces deux sources, même si l'adéquation globale est très bonne. Les points soulevés précédemment encouragent plutôt à douter de la capacité à ERA-Interim à modéliser correctement les extrêmes, et on peut penser que les évènements les plus élevés sont en-deçà de ce que la bouée donnerait sur la même période.

Cette remarque est confirmée par le graphique II.8, représentant un graphique quantile-quantile (QQ-plot), représentant les quantiles du H_s observé sur la bouée en abscisses contre les données du modèle numérique en ordonnées. La ligne de '+' rouges est la droite passant par les premier et troisième quartiles de chaque distribution. On observe sur ce

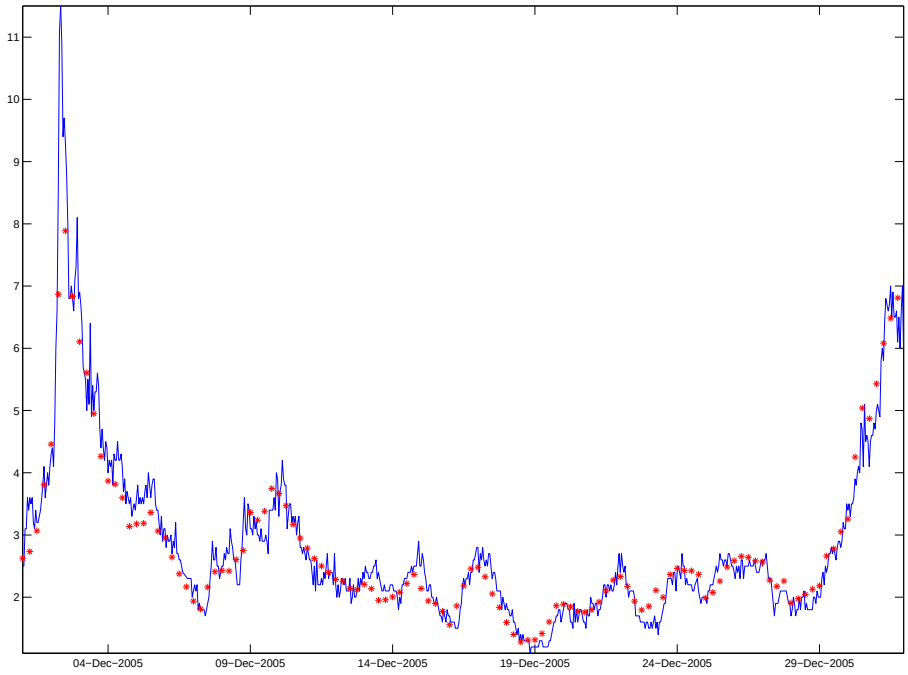


FIGURE II.7 – Observation de la hauteur significative des vagues sur la bouée Britanny (bleu) avec les données issues du modèle ERA (points rouges).

graphique que dans le coeur de la distribution, les deux jeux de données sont très proches, le qq-plot étant quasiment linéaire ce qui est un signe d'adéquation des deux distributions. En revanche, en ce qui concerne les extrêmes, la situation est plus problématique. En effet, le graphique montre que les quantiles d'ERA restent plus bas que ceux de la bouée, montrant que les niveaux les plus élevés de hauteur significative des vagues ne seront pas restitués par le modèle numérique. Il ressort de cette remarque une contrainte à prendre en compte lors des applications : bien que le modèle numérique présente des avantages certains en ce qui concerne la répartition spatio-temporelle des données, il a une certaine tendance à sous-estimer les extrêmes, et à ce titre, il ne doit pas constituer une référence pour représenter les valeurs les plus élevées. Cette intuition est confirmée par le tableau II.1, représentant les résultats de la méthode *POT* usuelle sur les deux séries temporelles. Le seuil retenu pour l'estimation est le quantile de 97.5%. Sur ce tableau, les données satellitaires ne semblent pas meilleures que les données ERA, nous illustrerons ce point par la suite.

Données	ξ	σ	μ	Extremal Index	Niveau de retour
ERA	-0.1229	0.3986	7.3061	0.2497	8.6476
Bouée	0.3167	0.4059	7.5849	0.0643	11.8997
Satellite	0.1990	1.7167	2.3818	0.5182	7.0682

TABLE II.1 – Résultats de la méthode POT usuelle sur les différents jeux de données

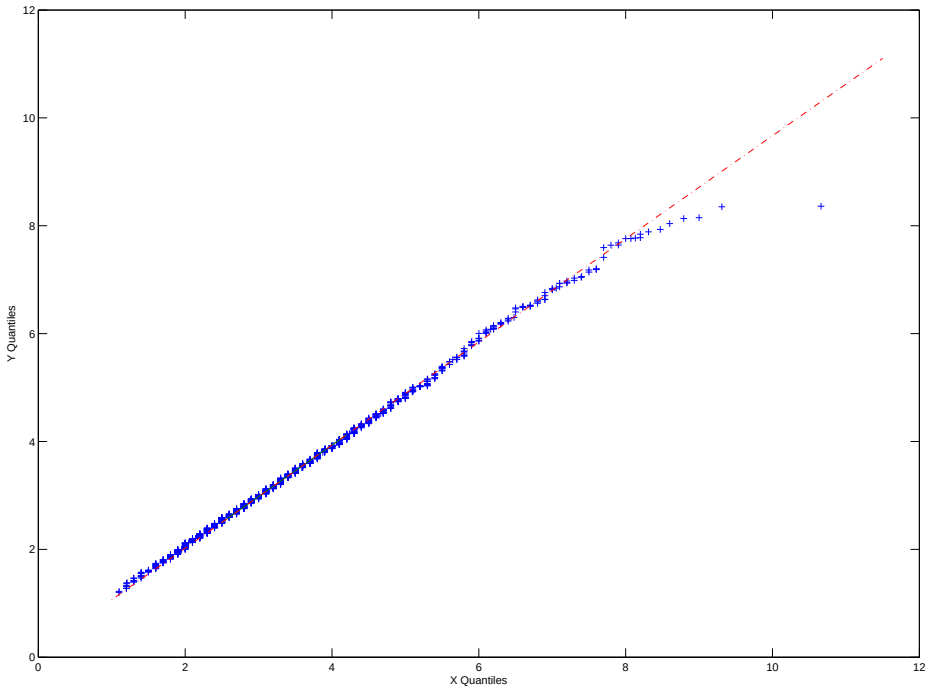


FIGURE II.8 – QQ-Plot de la hauteur significative des vagues sur la bouée Britanny (abscisses) contre les données issues du modèle ERA (ordonnées).

Comportement extrême Nous illustrerons dans ce paragraphe la dynamique des extrêmes des données issues de la bouée. Pour ce faire, nous disposons pour le moment de peu d'outil : en effet, la seule quantité présentée jusqu'à présent est l'*extremal index*, qui peut s'interpréter comme l'inverse de la taille des groupes consécutifs d'extrêmes, groupes que nous appellerons par la suite **clusters**, les extrêmes considérés étant les dépassements d'un seuil élevé. Nous nous intéresserons donc bien entendu à la valeur de ce paramètre, mais également à d'autres quantités qui permettent de compléter cette information sur la dynamique des extrêmes.

Trois autres statistiques seront donc calculées sur les données, chacune étant définie au-delà d'un certain seuil :

- Le nombre de clusters : aux effets de bords près, cette quantité est le nombre de dépassements du seuil (ou *up-crossing*). Il représente la fréquence avec laquelle les évènements extrêmes apparaissent ;
- La longueur des clusters : temps écoulé entre un *up-crossings* et un *down-crossings*. Il représente la persistance des extrêmes. Dans le cas d'observations indépendantes, cette quantité est proche de 1 ;
- Le temps entre les clusters : cette quantité complète la précédente dans le sens où elle est définie comme le temps entre *down-crossings* et un *up-crossing*.

Ces quantités, facilement estimables, seront estimées sur chaque mois de décembre à notre disposition, et nous regarderons la moyenne sur ces années, afin de pouvoir dresser des conclusions. Vu que chacune des définitions fait intervenir la notion de dépassement de seuil, nous regarderons comment elles se comportent quand le seuil augmente : c'est le

contenu des figures II.9 et II.10. Elle contiennent respectivement les moyennes sur toute l'étendue possible pour le seuil, et un agrandissement sur une zone où les seuils sont suffisamment élevés pour pouvoir parler de comportement réellement extrême, mais s'arrêtant avant le maximum des observations, afin d'éviter les effets de bords.

Ces graphiques nous apportent différentes informations : tout d'abord, précisons que les diverses quantités sont d'intérêt quand le seuil est élevé (par rapport à la distribution des observations), et en particulier les propriétés de l'estimateur de l'extremal index utilisé ont été démontrées quand le seuil tend vers la borne supérieure du support (cf [26]) . On remarque que l'extremal index est toujours inférieur à 1, ce qui correspond à un clustering des extrêmes, c'est-à-dire que l'on a tendance à observer des groupes d'extrêmes consécutifs. Cette caractéristique est confirmée par le troisième graphique, sur lequel on observe un palier autour d'une durée de deux heures, pour des niveaux compris entre 8 et 9. Cela signifie qu'une fois ce seuil passé, il faudra du temps pour redescendre, pour revenir à une situation plus calme, et en moyenne ce retour prend deux heures, à un facteur d'incertitude près car cette durée est le temps qui sépare deux observations, il se peut donc que la durée réelle soit un peu plus courte, ou un peu plus élevée.

On peut faire les commentaires suivants à la vue de ces graphiques : les courbes semblent très bruitées pour les seuils élevés, principalement celle des temps inter-clusters. Ce phénomène normal s'explique par le fait que ces estimations sont très difficiles, vu le peu d'observations dont on dispose à un tel niveau. Ainsi, il faudra prêter une attention particulière quand l'on cherchera à comparer ces estimations avec celles basées sur le modèle, du fait que leur manque de fiabilité et de représentativité. Un phénomène qui peut paraître surprenant est la chute brutale du temps inter-cluster entre 8 et 10 mètres de H_s : cela s'explique par le fait que certaines années ne comportent aucun dépassement de 8m. Le second graphique quant à lui nous permet de définir une quantité très utilisée en pratique : le niveau de retour. Cette quantité est définie comme le niveau dépassé en moyenne une fois sur la période considérée. Ainsi, le niveau de retour à 100 ans, ou la vague centennale dans notre cas, est l'évènement qui apparait en moyenne une fois tous les 100 ans. Dans le cas d'observations indépendantes, cette quantité est facile à définir, mais quand elles sont dépendantes, il y a un choix à opérer sur la définition "d'évènement". Nous avons fait le choix de retenir un évènement comme l'apparition d'un cluster de valeurs extrêmes. Ce choix parait assez naturel en termes de gestion des risques, d'autant plus qu'il est alors possible d'attribuer une durée moyenne à l'évènement centennal en plus d'un niveau, ces deux quantités étant des caractéristiques très utiles pour les applications. Avec cette définition, le niveau de retour à p ans se trouve à l'intersection entre la courbe représentée sur le deuxième graphique et la droite d'ordonnée $1/p$. Il est ensuite possible de regarder quelle est la durée moyenne des évènements extrêmes pour le niveau qui a été trouvé dans l'étape précédente, ce qui donne une description plus complète de l'évènement. Pour pouvoir obtenir des valeurs précises de ces quantités, il est donc nécessaire de pouvoir extrapoler cette courbe au-delà du support des observations, ce qui sera l'objectif du présent travail.

Les figures II.11 et II.12 représentent quant à elles les même quantités que celles présentées ci-dessus, mais calculées sur les données ERA-Interim, déjà présentées. On peut noter que l'estimation de l'extremal index aboutit à la même conclusion que précédemment : les extrêmes ont une tendance à arriver par blocs et non par valeurs isolées, et une stabilisation de l'estimateur autour de la valeur 0.4 est encore plus flagrante. On constate cependant de grandes différences entre les données de bouées et ERA-Interim pour les trois autres courbes représentées. Sur le second graphique en effet, ce que l'on pourrait nommer 'niveau de retour à deux ans', qui n'est d'autre que l'intersection entre la droite $y = 0.5$

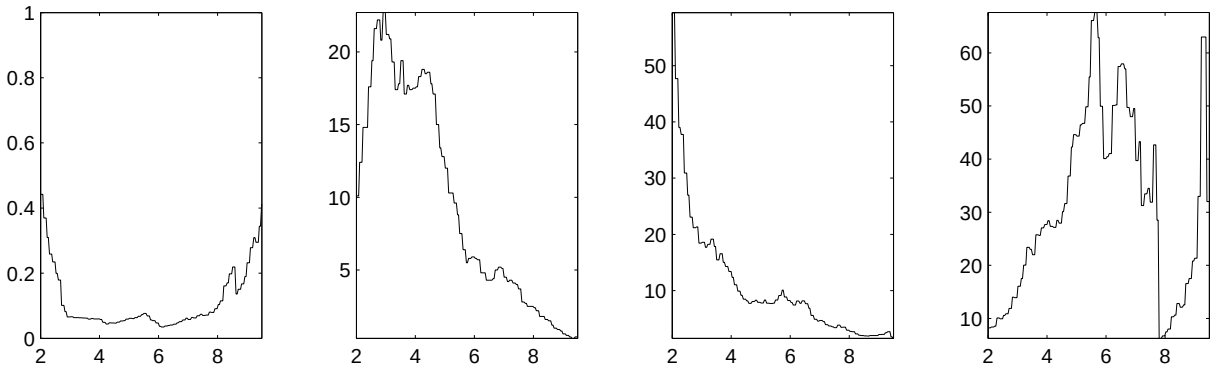


FIGURE II.9 – Statistiques extrêmes calculées sur la bouée, en fonction du seuil. De gauche à droite : extremal index, nombre moyen de dépassements du seuil par mois de décembre, longueur moyenne des clusters (en heures), temps moyen entre les clusters (en heures).

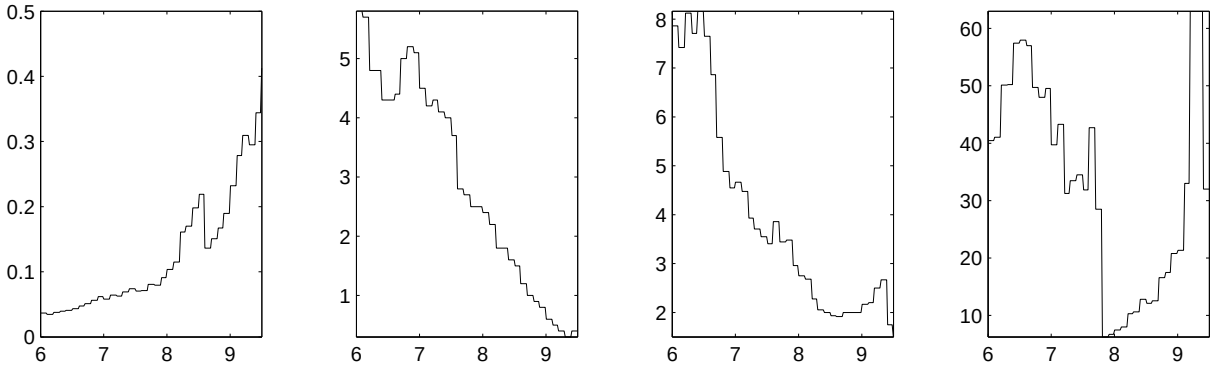


FIGURE II.10 – Statistiques extrêmes calculées sur la bouée, en fonction du seuil, pour des valeurs de seuil élevées. De gauche à droite : extremal index, nombre moyen de dépassements du seuil par mois de décembre, longueur moyenne des clusters (en heures), temps moyen entre les clusters (en heures).

et la courbe, est à près de 9 sur la bouée, alors qu'elle vaut environ 8 sur ERA-Interim, ce qui constitue une différence certaine. Au-delà de cet exemple, les niveaux extrêmes de ERA-Interim sont plus faibles que ceux observés sur la bouée : ce comportement était prévisible aux vues des comparaisons faites précédemment. Mais la dynamique des extrêmes est également différente entre les deux sources de données : on constate en effet sur le troisième graphique que la durée moyenne des extrêmes se situe autour de 12 heures, ce qui correspond comme pour les bouées à deux observations, mais cette fois-ci la fréquence d'échantillonnage est plus faible. Le même commentaire s'applique au dernier graphique : une fois que ERA-Interim a donné un extrême, il met plus de temps que la bouée avant d'en produire un nouveau. Ces remarques contribuent à confirmer ce qui apparaissait déjà : les extrêmes de ERA-Interim sont plus faibles, mais ce jeu de données est également plus lisse, dans le sens où il peut difficilement mettre en évidence des variations rapides, tout du moins pour les extrêmes.

Toutes ces observations contribuent au fait de préférer ne pas utiliser les données ERA pour modéliser des comportements extrêmes. Mais comme dit précédemment, le réseau des bouées est très peu dense en espace, et il y a peu de chances donc d'en avoir une proche

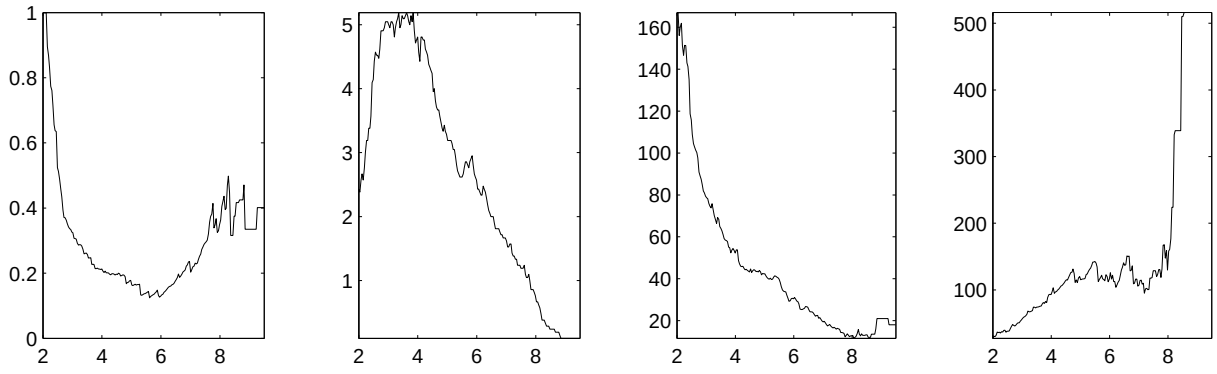


FIGURE II.11 – Statistiques extrêmes calculées sur ERA-Interim, en fonction du seuil. De gauche à droite : extremal index, nombre moyen de dépassements du seuil par mois de décembre, longueur moyenne des clusters (en heures), temps moyen entre les clusters (en heures).

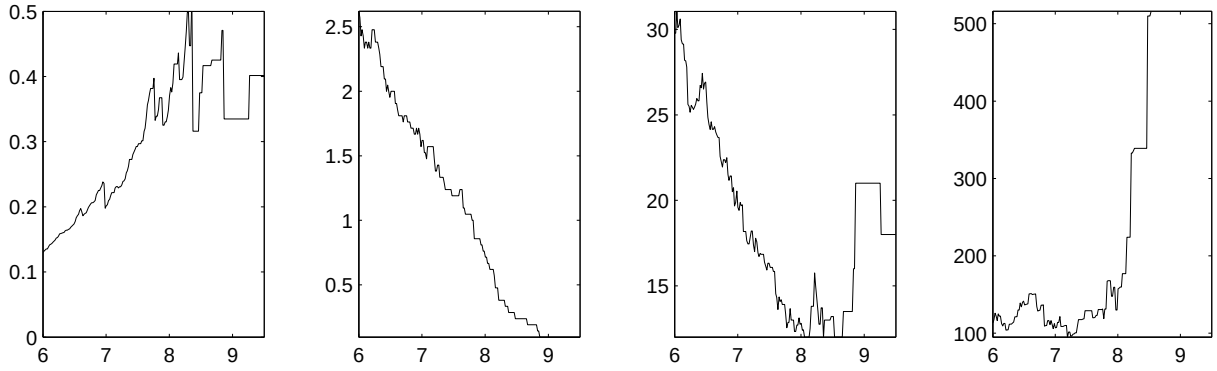


FIGURE II.12 – Statistiques extrêmes calculées sur ERA-Interim, en fonction du seuil, pour des valeurs de seuil élevées. De gauche à droite : extremal index, nombre moyen de dépassements du seuil par mois de décembre, longueur moyenne des clusters (en heures), temps moyen entre les clusters (en heures).

de l'endroit où l'on souhaite faire une étude des extrêmes. C'est la raison pour laquelle les données satellitaires sont intéressantes pour palier ce problème, point que nous allons préciser dans le paragraphe suivant.

1.4 Données satellitaires

1.4.1 Description

Les données *hindcast* présentent l'avantage d'être disponibles sur une grille régulière en temps et en espace, régularité qui simplifie le traitement et la modélisation. Cependant, un important défaut de ces données est sa tendance à lisser les extrêmes, et de fait à sous-estimer les niveaux de retour associés, comme nous avons pu le voir dans la partie qui précède. C'est la raison pour laquelle des données issues de l'observation directe sont souvent préférées, mais ce au prix d'une nécessitée d'élaborer des modèles plus complexes. Les données qui nous intéressent ici sont issues de mesures satellitaires de la hauteur significative des vagues, données qui nous ont été fournies par le CERSAT, laboratoire de

l'IFREMER.

On dispose donc de traces des satellites, qui prennent une photo d'environ 10km de coté sur laquelle sont mesurés les différents paramètres d'état de mer (hauteur des vagues, direction moyenne, vitesse et direction du vent...). Un exemple de telles mesures de la hauteur significative des vagues est représenté sur la figure II.13, où l'on observe toutes les traces satellitaires sur une journée, carte qui permet d'apprécier la couverture spatiale sur une journée.

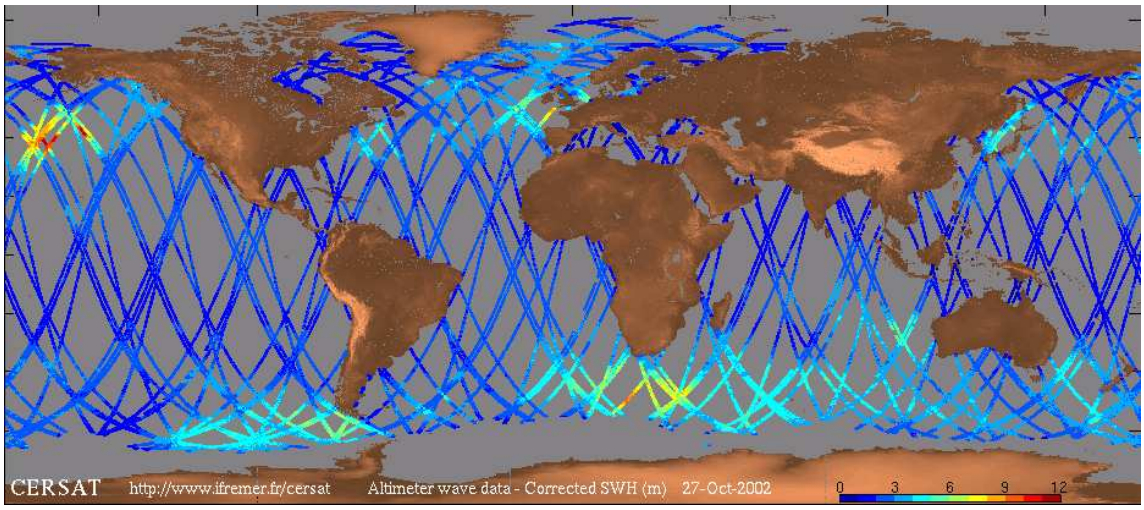


FIGURE II.13 – Exemple de mesure de H_s par satellite, le 27 Octobre 2002

Les données sont disponibles d'avril 1992 à maintenant [3]. Sept satellites différents ont collaboré à l'établissement de cette base de données : ERS-1, ERS-2, ENVISAT, TOPEX, Poséidon, Jason1 et GEOSAT FollowOn, mais tous ne sont pas disponibles en même temps sur tous les endroits du globe, en fonction de leurs dates de mises en service ou a cause de défaillances internes. La figure II.14 montre les couvertures temporelles de chacun des satellites pris en compte dans cette étude (source : [66]). Ces données ont montré une adéquation globalement satisfaisante aux données de bouées lors d'un passage à proximité de ces dernières, tout du moins en ce qui concerne le comportement moyen ([45]), le comportement extrême étant plus complexe à étudier du fait de la nature des données.

Ces données présentent en effet certaines difficultés pour leur traitement : d'une part tous les satellites peuvent ne pas avoir la même précision, ce qui donnerait deux mesures différentes au même endroit et au même moment. Mais l'autre difficulté, plus importante, est qu'ils ne passent pas aux mêmes endroits au cours du temps, même pour un satellite donné. En effet, les points de mesure changent au cours du temps, et la densité spatiale est relativement faible. S'ajoute également une couverture temporelle assez faible et surtout irrégulière, bien que non représentée sur ces graphiques. Un problème non visible ici est celui des données manquantes. Il arrive en effet qu'un satellite présente une défaillance (la cause principale est une avarie de la mémoire interne du satellite, faisant qu'il ne peut communiquer que lorsqu'il est à vue d'une station de réception) et la trace comporte donc des données manquantes le long du passage du satellite.

Cet échantillonnage irrégulier nous fait perdre l'échantillonnage spatio-temporel régulier des précédentes données, ce qui exclut d'apporter un traitement similaire, en particulier

3. Avec un léger délai de quelques mois

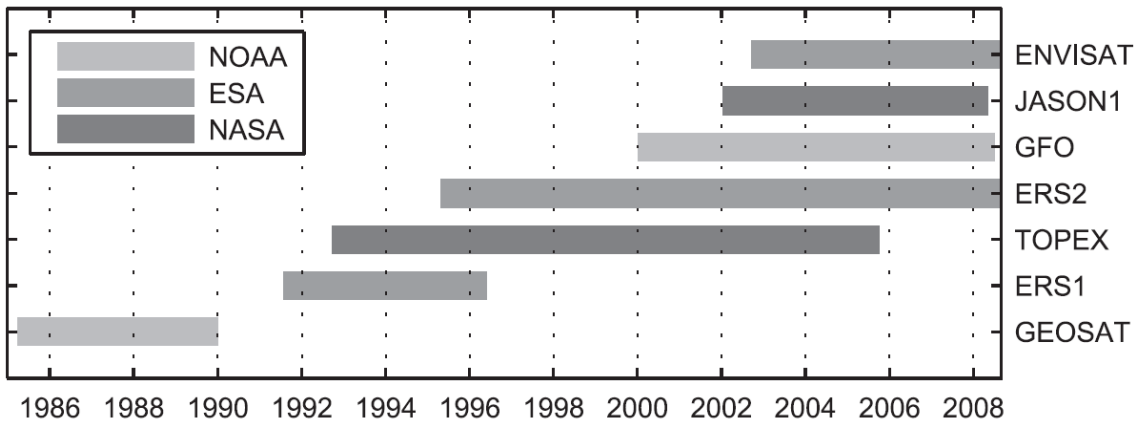


FIGURE II.14 – Couverture temporelle des différents satellites

les maxima par blocs : même si nous définissions des blocs spatiaux, tous ne comporteraient pas le même nombre d'observations, et qui plus est, deux années différentes sur un même bloc spatial ne comporteraient pas non plus le même nombre d'observations. Il est donc nécessaire, afin de proposer une description du comportement extrême, de développer des modèles spatio-temporels basés sur de telles particularités. Une étude des valeurs extrêmes des altimètres au sein de boites de taille constante a été réalisée récemment dans [63] par une approche POT classique, montrant qu'une telle analyse est néanmoins possible et montre une bonne adéquation aux données de bouées disponibles.

1.4.2 Statistiques descriptives

Le but de cette partie sera d'effectuer une comparaison entre les sorties du modèle et les observations. Pour cela, nous avons calculé la moyenne en chaque site : nous avons discrétisé en espace notre zone d'étude de façon à obtenir des boites de 1.5°, comme pour ERA-Interim.

Dans un premier temps, nous avons représenté le nombre de passages dans chaque site, sur le graphique II.15. On remarque que le nombre d'observations par site est à peu près uniforme, même s'il semble que la zone la plus au sud soit moins bien échantillonnée : cela vient du fait que dans les zones plus équatoriales, les traces satellitaires sont plus verticales, donc restent moins longtemps dans la zone. Ensuite vient la carte de la moyenne en chaque site, ainsi que l'écart-type sur la figure II.16. Nous ne reprendrons pas ici la description de cette répartition, mais plutôt les différences avec la figure II.1, 25 : en ce qui concerne la moyenne, on observe que l'estimation donnée par les données satellitaires est moins lisse spatialement, mais surtout que les niveaux sont légèrement différents, le satellite donnant une estimation un peu supérieure à celle donnée par ERA. Quant à l'écart-type, l'estimation donnée par le satellite est beaucoup moins lisse que celle donnée par ERA, mais surtout les zones ayant la plus forte variabilité ne sont pas les mêmes, et une fois encore les niveaux de variabilité sont plus élevés pour les données satellitaires que pour les données de réanalyse. Ce point n'est pas surprenant du fait que le modèle numérique est réputé pour son caractère lisse, mais cela confirme que ce dernier risque de ne pas être adapté à l'étude des valeurs extrêmes, une fois encore de par son caractère lisse.

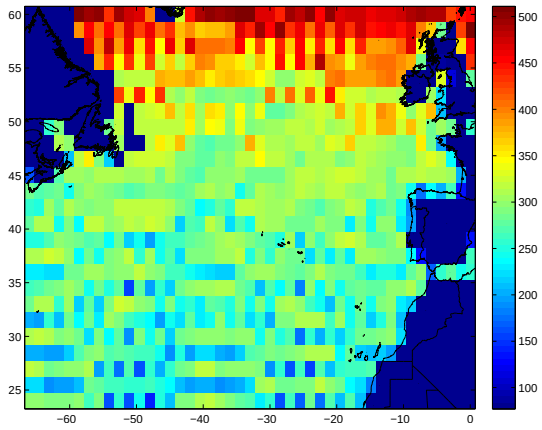
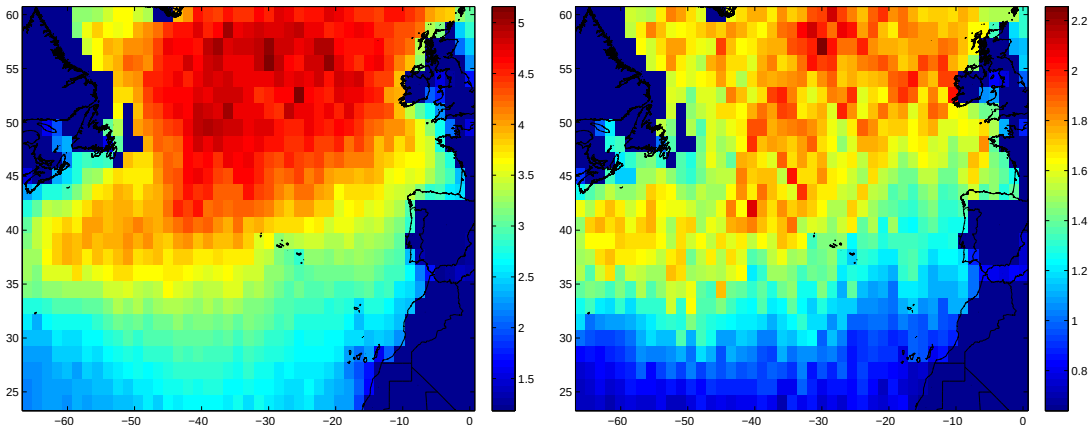


FIGURE II.15 – Nombre de traces de satellite par site sur la durée observée



(a) (b)

FIGURE II.16 – (a) : hauteur significative moyenne observée par satellite ; (b) : écart-type des observations.

1.5 Interpolation des traces satellitaires

On constate un problème majeur lors du traitement de ces données : la discrétisation de l'espace n'est pas un moyen de traitement adapté à ces données. En effet, une tempête peut être observée en un point de l'espace et du temps par le satellite, et en se déplaçant va contribuer à créer un état de mer perturbé plus tard, et dans une autre zone ou à un instant où l'on ne dispose pas d'observation. La discrétisation de l'espace nous fait perdre cette influence spatiale des tempêtes. Une approche que nous avons adoptée et qui sera détaillée par la suite (III) est d'utiliser l'information sur les déplacements des tempêtes, contenue dans les données ERA-Interim, afin de déplacer les traces satellitaires en un point donné de l'espace et du temps.

Plus concrètement, les données ERA-Interim permettent d'estimer entre chaque pas de temps un déplacement des structures, en recherchant les points les plus ressemblants. Cette estimation a été réalisée à l'aide de filtrage particulaire, sur un ensemble de caractéristiques plus large que celui traité jusqu'à présent : nous avons pris en compte la hauteur significative, la direction moyenne de propagation et la période moyenne.

Cette approche a permis la création de bouées virtuelles en améliorant l'adéquation avec les observations comparativement aux données ERA-Interim, on se référera au chapitre III pour plus de détails à ce sujet.

1.5.1 Description du comportement extrême

Afin de dresser une comparaison avec les données ERA-Interim, nous avons calculé les maxima annuels en chacun des points de la grille obtenue précédemment. Cette démarche n'est pas tout-à-fait légitime, car le nombre d'observations sur lequel est calculé le maxima annuel change chaque année, en fonction du nombre de traces satellitaires, et change de site en site, en fonction de la trajectoire des satellites qui ne couvrent pas uniformément l'espace. Cependant, cette approche nous permet d'avoir une idée sur le comportement extrême des données satellitaires, et en particulier si l'on retrouve les caractéristiques observées dans les données ERA-Interim.

Comme on peut l'observer sur les graphiques de la figure II.17, il existe des différences majeures entre les deux sources de données, qui ne peuvent s'expliquer que par l'inadéquation de l'approche pour les données satellitaires : en effet, ces deux sources sont connues pour être proches (voir les travaux de [45]), et même si certains évènements extrêmes peuvent être manqués par les satellites de par leur faible couverture spatiale, de même que ERA-Interim a tendance à lisser les évènements les plus forts. Pour ces raisons, il semblerait pertinent de retrouver au moins la même information spatiale entre les deux sources, même si les données satellitaires en fournissent une image plus bruitée pour les raisons évoquées dans le point précédent. Or ici il ne semble pas se dessiner de zone géographique au sein de laquelle le comportement extrême serait similaire, en particulier si l'on compare la carte des niveaux de retour à 100 ans : le satellite donne des niveaux de retours plus élevés, ce qui est cohérent avec le fait qu'il ne lisse pas les extrêmes qu'il observe, mais l'information spatiale semble perdue par l'approche que nous avons adoptée.

Au-delà de l'information spatiale, ces données permettent-elles d'estimer des niveaux de retours plus proches de ceux données par une bouée, qui, rappelons-le, peut être considérée comme une référence ? Cette question n'est pas évidente à traiter de part la nature des données traitées : en effet, comme nous l'avons expliqué dans ce qui précède, l'échantillonnage temporel des satellites en un point est fortement irrégulier, comme on peut le voir sur la figure II.18, qui représente conjointement chacun des jeux de données sur un mois donné. On observe que deux observations issues des satellites peuvent être très

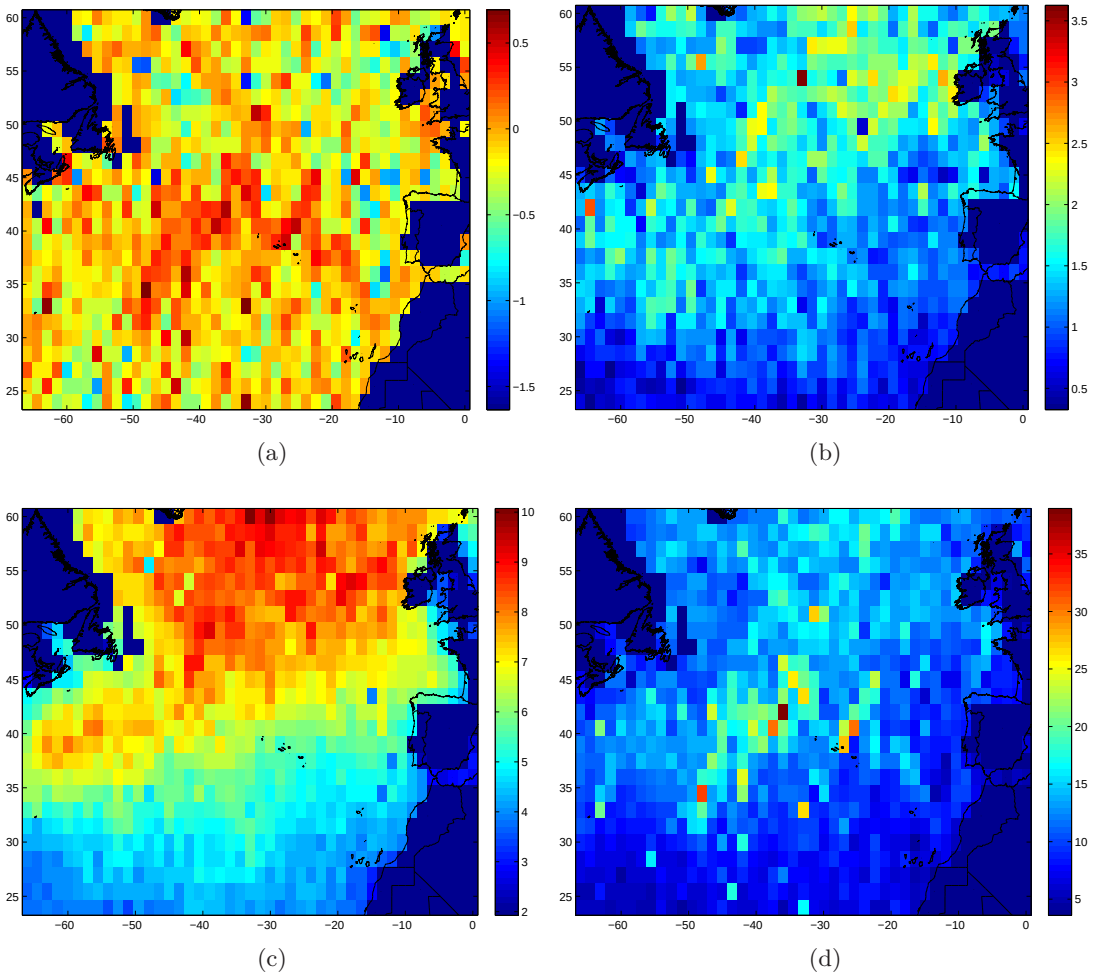


FIGURE II.17 – Résultats de l'ajustement des maxima annuels sur les données satellitaires : (a) : ξ, (b) : σ, (c) : μ, (d) : niveau de retour à 100 ans.

proches, ce qui arrive par exemple quand deux satellites se suivent en vue d'un étalonnage des outils de mesure, ou être très éloignées, ce qui arrive quand aucun satellite ne passe suffisamment proche de la bouée. Nous avons choisi pour cette représentation une boite de 1.5 deg autour de la bouée. Une information complémentaire est apportée par les graphiques de la figure II.19, représentant les qq-plots des données satellitaires contre les données de bouée.

Deux informations importantes sont à tirer de ces graphiques : d'une part, quand un satellite donne une mesure, on peut constater qu'elle est proche de celle donnée par la bouée, ce qui est visible sur le graphique II.18 et confirmé en partie par les qq-plots, sur lesquels les distributions paraissent proches, même si dans ce cas l'information temporelle ne rentre pas en compte. Du point de vue de la modélisation des valeurs extrêmes, on rencontre une importante difficulté, car les modèles classiques de maxima par blocs ou de **POT** ne convienent pas. Pour le premier en effet, il est difficile d'exhiber des blocs ayant une signification et ayant tous le même nombre d'observations , alors que pour le second, on aimerait appliquer les travaux réalisés dans le cas de séries temporelles, mais l'échantillonnage fortement irrégulier rend difficile d'une part la définition d'un cluster si l'on souhaite appliquer la technique du *declustering*, et d'autre part la modélisation par une chaine de Markov des dépassements successifs.

Nous avons cependant appliqué la méthode **POT** usuelle sur les diverses sources de données, afin de comparer les résultats. C'est l'objet du tableau II.1 page 32, dans lequel sont consignées les valeurs des paramètres marginaux ainsi que de l'*extremal index* pour chacun des jeux de données, ainsi qu'une estimation du niveau de retour à 100 ans, obtenu à partir de ces paramètres. Pour chacun, une étape de declustering a été appliquée pour les dépassements du quantile à 95%. On remarque que les données satellitaires donnent des résultats quelque peu différents. Nous voyons deux raisons principales à cela : d'une part le modèle ajusté est mal choisi du fait des particularités d'échantillonnage, et d'autre part, il y a une perte d'information due au fait que nous ayons moyenné dans une zone restreinte autour de la bouée.

2 Objectif du présent travail

Ce chapitre a cherché à présenter les différentes données en notre possession pour caractériser les évènements extrêmes de hauteur significative des vagues. Il a été montré que les données ERA-Interim, bien qu'elles présentent un atout évident qui est celui d'être disponible sur une assez longue période de temps sur une grille régulière en temps comme en espace, ne permettent pas de bien modéliser les comportements extrêmes. On observe en effet de grandes différences entre ces données et les données de bouées sur les plus hautes valeurs, mais aussi sur la dynamique des extrêmes, qui est lissée par ERA-Interim. Cette remarque motive à l'utilisation de données satellitaires, plus précises comme nous l'avons vu, mais aussi plus difficiles à traiter. Les données de bouées montrent en effet que les observations extrêmes semblent arriver par paquets, ou *clusters*, ce qui invalide la méthode POT la plus rudimentaire, et pose de réelles difficultés à son extension au cas qui nous intéresse, puisqu'il est difficile de déterminer des clusters de valeurs extrêmes sur le satellite. Toutes ces remarques incitent donc aux développements que nous allons effectuer par la suite. Dans un premier temps, nous allons utiliser les données satellitaires conjointement avec les données ERA pour tirer meilleur partie de chacune : la précision des niveaux pour la première, et la couverture spatiale permettant d'estimer le déplacement des structures pour la seconde. Ces travaux feront l'objet du chapitre suivant. Un second objectif sera de se donner les moyens d'obtenir un modèle valide sur chacun des jeux de

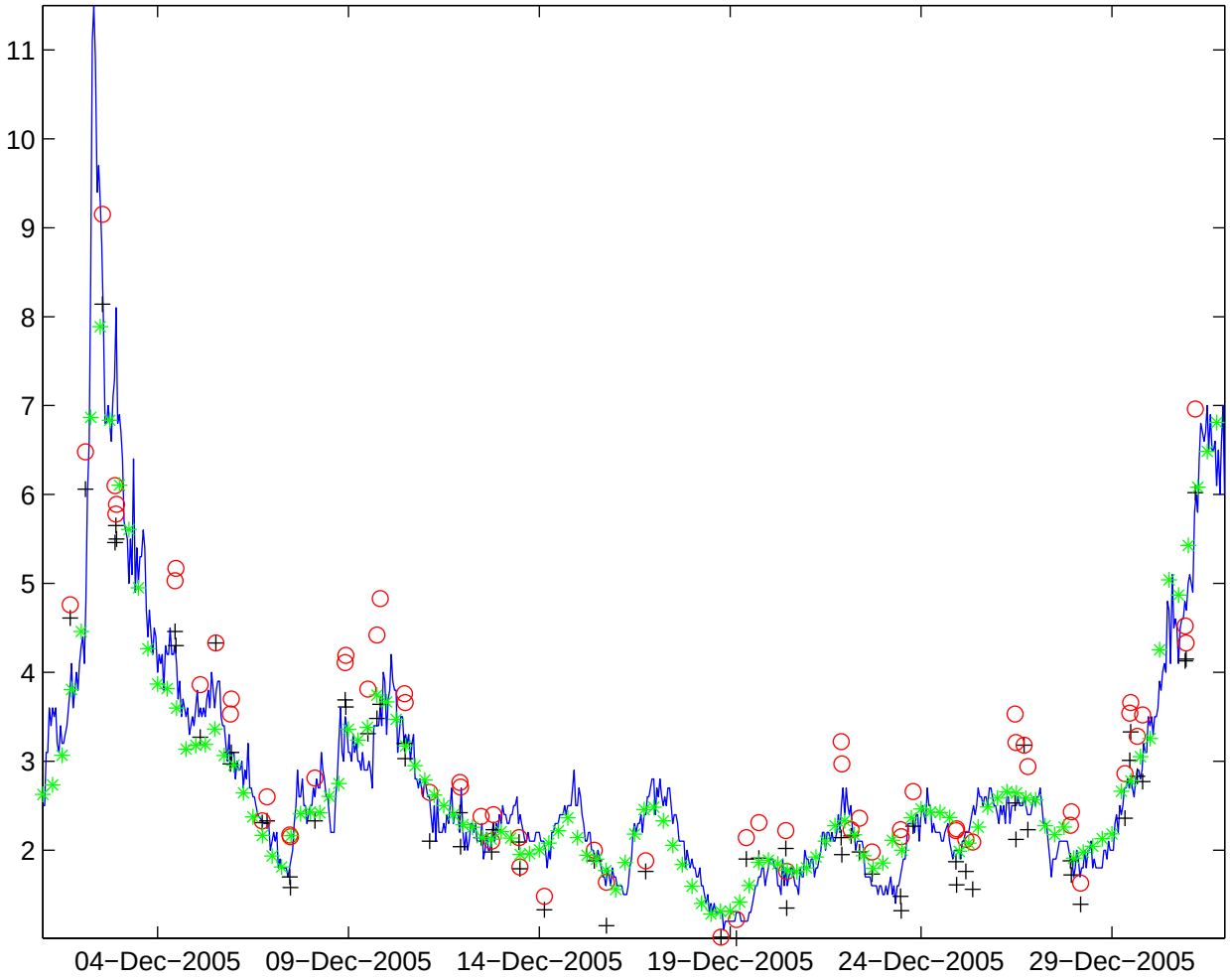


FIGURE II.18 – Exemple d'un mois d'observation des différents jeux de données en un point donné. Ligne bleue : bouée ; étoiles vertes : ERA-Interim ; Plus noirs : donnée de la trace la plus proche de la bouée ; Rond rouge : maximum sur la trace.

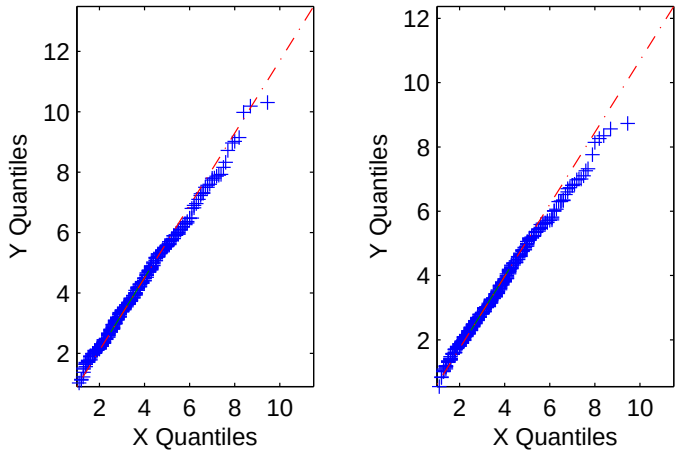


FIGURE II.19 – QQ-plot des observations issues de la bouée (abcisse) contre les valeurs observées par satellite (ordonnées) en conservant : (a) la maximum par trace, (b) la valeur du point le plus proche de la bouée dans une boite de 2.5° de coté.

données présenté dans le présent chapitre, afin de pouvoir mener des comparaisons sur les informations contenues dans chacun au sujet des valeurs extrêmes. Ayant remarqué que peu d'outils de diagnostics existent pour s'assurer du bon ajustement du modèle aux données, nous utiliserons certains des graphiques présentés dans cette partie, qui ont l'avantage d'être rapides à calculer et faciles à interpréter, comme le nombre de clusters au-delà d'un certain seuil par exemple, afin de s'assurer que le modèle choisi s'ajuste bien aux données. En particulier, nous nous intéresserons à la dynamique des extrêmes, en plus de l'ajustement des lois marginales.

Chapitre III

Interpolation des données satellitaires

1 Introduction

Dans cette section, nous présentons un article publié dans la revue Environmetrics. Ce travail concerne une méthode d'interpolation de données satellitaires. Comme expliqué dans la partie qui précède, les données satellitaires sont difficiles à traiter du fait de leur répartition spatiale, et il est donc intéressant de proposer une méthode d'interpolation spatio-temporelle. L'originalité de ce papier est de proposer une estimation des déplacements des états de mer, à l'aide des données ERA-Interim, afin de réaliser une interpolation en temps et en espace des données satellitaires, en prenant en compte le déplacement des états de mer au cours du temps. Les vitesses de déplacement ont été estimées grâce à un modèle à espace d'état. Afin de conserver les références se trouvant à l'intérieur de l'article intact, nous avons intégré cet article directement à l'intérieur de ce document.

2 Article

Research Article

Received: 19 May 2010, Revised: 19 May 2010, Accepted: 20 May 2010, Published online in Wiley Online Library: 29 December 2010

(wileyonlinelibrary.com) DOI: 10.1002/env.1061

Space–time models for moving fields with an application to significant wave height fields

Pierre Ailliot[a*,†], Anastassia Baxevani[b], Anne Cuzol[c], Valerie Monbet[c] and Nicolas Raillard[a,c,d]

The surface of the ocean, and so such quantities as the significant wave height, H_s, can be thought of as a random surface that develops over time. In this paper, we explore certain types of random fields in space and time, with and without dynamics that may or may not be driven by a physical law, as models for the significant wave height. Reanalysis data is used to estimate the sea-state motion which is modeled as a hidden Markov chain in a state space framework by means of an AR(1) process or in the presence of the dispersion relation. Parametric covariance models with and without dynamics are fitted to reanalysis and satellite data and compared to the empirical covariance functions. The derived models have been validated against satellite and buoy data. Copyright © 2010 John Wiley & Sons, Ltd.

Keywords: space–time model; significant wave height; state-space models

1. INTRODUCTION

Spatio-temporal modeling is an important area in statistics that is one of rapid growth at the moment, with various applications in environmental science, geophysical science, biology, epidemiology and others. See for instance Christakos and Hristopoulos (1998), Cressie and Huang (1999), and Gneiting *et al*. (2007).

Especially after all the recent technological advances such as satellite scanning that resulted in increasingly complex environmental data sets, estimating and modeling the covariance structure of a space–time process have been of great interest although not as well developed as methods for the analysis of purely spatial or purely temporal data. Because it is often difficult to think about spatial and temporal variations simultaneously, it is tempting to focus on the analysis of how the covariances at a single point vary across time or how the covariances at a single time vary across space. If these were the only characteristics that mattered, then separable models would suffice. Allowing though the merely spatial and merely temporal covariances to define the space–time dependence is a severe restriction, that is not satisfied by many geophysical processes, such as meteorological systems, rainfall cells, air pollution, etc., that exhibit motion. Hence the need for non-separable covariance models which include interactions between the spatial and the temporal variability, see Christakos (2000), Ma (2002), and Gneiting *et al*. (2007) among others for recent contributions.

In this paper, we explore methods for constructing models for the significant wave height, a parameter related to the energy of the sea-state, based on fitting random field models to data collected from different sources. A full description of the data, some aspects of its limitations and some assumptions that are reasonable in modeling it are given in the next subsection. The models are then described and interpreted in terms of these assumptions.

An important feature of the significant wave height fields, non-compatible with the assumption of separability, is their motion. The apparent motion of the significant wave height fields is actually the composition of various motions, that of the wind fields (see Ailliot *et al*., 2006) that generate the waves and those of the various wave systems that compose each sea-state. In order to simplify the analysis though, we consider that each sea-state moves with a single velocity that is the composition of the ones mentioned above.

In the area of interest, a part of the North Atlantic Ocean, westerly winds are prevailing and low-pressure systems are generally moving to the East. As a consequence, the significant wave height fields are also moving to the East, although the important variability in the

* *Correspondence to: P. Ailliot, Laboratoire de Mathématiques, UMR 6205, Université Européenne de Bretagne, Brest, France. E-mail: ailliot@univ-brest.fr*

a *Laboratoire de Mathématiques, UMR 6205, Université Européenne de Bretagne, Brest, France*

b *Department of Mathematical Sciences, Chalmers University of Technology, University of Gothenburg, Sweden*

c *Lab-STICC, UMR 3192, Université Européenne de Bretagne, Vannes, France*

d *Laboratoire d'Océanographie Spatiale, IFREMER, France*

meteorological conditions in this area imply also an important variability in the speed and direction at which the sea-states are traveling. In this study, we propose the use of the output of numerical weather forecast systems to estimate the resulting sea-state motions. Although these numerical models are sometimes inaccurate, we show that they provide enough information on the state of the atmosphere and ocean in order to get at least a rough estimate of the prevailing motion.

Using a Lagrangian reference frame, instead of a Eulerian (fixed) reference frame, appears to be natural for modeling moving processes. Such an idea has already been used in the literature (see e.g., Gneiting *et al*. (2007) and references therein), but it is generally assumed that the motion is constant in time (*frozen velocity*) and eventually in space. The Lagrangian reference frame moves with the sea-states, and as a consequence we expect longer range dependence in the temporal domain than in the Eulerian reference frame. Indeed, we show that the main difference between the covariance structures for the two reference frames is a slower decrease to zero with time for the Lagrangian one, and that changing the reference frame to the Lagrangian actually leads to more accurate space–time interpolation.

Another originality of this work is that we combine different data sources that provide information at different scales: reanalysis (or *hindcast*) data, satellite data, and buoy data, a short description of which is given in Section 1.1. In Section 2, a covariance model with constant velocity is introduced. The method used to estimate the motion of the sea-states and the field covariance model in the associated Lagrangian reference frame are discussed in Section 3. A comparison of the ability of the different models to produce accurate space–time interpolations is presented in Section 4, where is also shown that the model with changing velocity outperforms the one with constant velocity. Finally, some conclusions are presented in Section 5.

1.1. Data

The data used in this work come from three different sources that are briefly described next.

- *Satellite data*: The observations consist of the significant wave height taken at discrete locations along one-dimensional tracks from seven different satellite altimeters that have been deployed progressively since 1991 and whose operation times can be seen in Figure 1. These tracks have two orientations and those with the same orientation are roughly parallel. For convenience, we shall use the term passage for the observations from one pass along a single track. It should be noted that, for each track, each passage may have observations at slightly different locations (in terms of longitude and latitude), which are consequently neither equidistant nor, from passage to passage, identical. Each observation collected is a summary statistic from a sampling window of about 7 km × 7 km. These windows are about 5.5–7 km apart and so the windows can overlap marginally. This smooths the observations for the passage and introduces some short-term dependence. Another characteristic of the data is that the observations have been discretized. This, unfortunately, has removed the very small-scale variation and, in many places, especially where the observations for a passage are low, has resulted in flat sections in the sample path. Moreover, satellite altimeter data have been calibrated using buoy data and the adequacy is generally satisfying (see Queffeulou, 2004). More information on the data sets and their particulars can be found in the URL: `ftp://ftp.ifremer.fr/ifremer/cersat/products/ swath/altimeters/waves`
- *Hindcast data*: Recently, a wave reanalysis 6-hourly data set on a global 1.5 × 1.5 latitude/longitude grid (see Figure 2) covering the period of 1957 to 2002 was made available, the ERA-40 data set. This reanalysis was carried out by the European Centre for Medium-Range Weather Forecasts (ECMWF) using its Integrated Forecasting System, a coupled atmosphere-wave model with variational data assimilation. Shortly after, new progress was made by producing the ERA-Interim data set which has progressed beyond the end of ERA-40 data set and now covers the period up to 2007. A distinguished feature of ECMWF's model is its coupling through the wave height dependent Charnock parameter (see Janssen *et al*., 2002), to a third generation wave model, the well-known WAM (Komen *et al*., 1994), which makes wave data a natural output of both the ERA-40 and the ERA-Interim system, the latter having a variational bias correction using satellite data. A large subset of the complete ERA-Interim data set, including significant wave height estimates, can be freely downloaded and used for scientific purposes at the URL `http://data.ecmwf.int/data/`. In this work, we use the ERA-Interim data set from 1992 until 2007.
- *Buoy data*: Buoy data are often considered as a reference to provide ground-truth for hindcast and satellite data. In this study, we use K1 buoy data (station 62029), which is part of the UK Met Office monitoring network. It is located at position (48.701 N, 12.401 W) (see

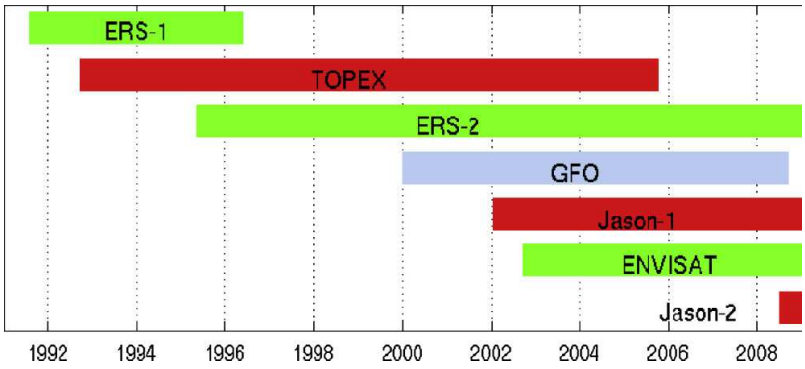


Figure 1. Time periods when altimeter data are available for each satellite. This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

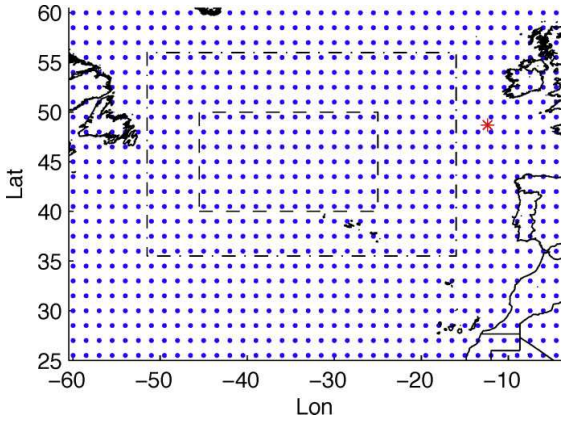


Figure 2. Grid of ERA-Interim data and domain of interest D_0 (dashed box). The dashed-dotted box corresponds to the domain D where the velocity fields are computed in Section 3.2 and the * indicates the location of buoy 62029. This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

Figure 2) and provides hourly significant wave height estimates. In this work, we only consider data for the period from April 2002 until December 2007.

A typical example of the data coverage over a 24 h time window can be seen in Figure 3. Hindcast data, like the ERA-Interim data, are sampled over a regular 1.5 × 1.5 degrees grid at synoptic times every 6 h starting at midnight, in contrast to the irregular, both in space and time, sampling provided by the satellite altimeter. However, the ERA-Interim data set tends to underestimate the variability of the significant wave height (see next section) and it only provides information at a synoptic scale whereas satellite data also give smaller scale information. For the two data sets to become more compatible, the satellite data have been smoothed in order to eliminate small-scale components. In practice, a moving average filter with a window of size 10 observations, which covers a distance of about 50 km, has been applied on each track and, to reduce computational cost, the data have been under-sampled so that only one observation every 10 is used.

Different analyses have been presented studying the correlation structure of the sea surface energy as measured by the significant wave height. For example Cotton *et al.* (2001) carried out a comparison of model (ERA-WAM, the 15-year first version of the ERA-40) and satellite climatologies (provided by the Southampton Oceanography Centre from altimeter data) to find that the model data showed similar tendencies when compared to altimeter data. Moreover, the ERA-40 data have been extensively validated against observations and other reanalysis data sets, (Caires *et al.*, 2004; Caires and Sterl, 2005), and turns out that performs well when compared with measurements from *in situ* buoy and global altimeter data. It is reasonable therefore to assume that the dynamics and the shape of spatio-temporal covariance structure of the

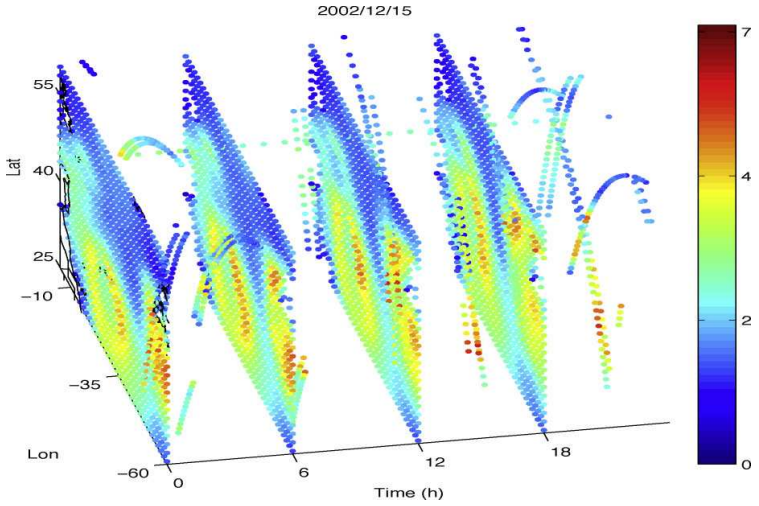


Figure 3. 3D representation of the available data for 15 December 2002. The 2D fields at times 0, 6, 12, 18 correspond to ERA-Interim data, the 24 observations at location (48.701 N, 12.401 W) correspond to buoy data and the other data to the various satellite tracks. This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

field is the same as those computed using the significant wave height estimates from the ERA-Interim data set. However, we keep in mind that the satellite altimeter provides with significant wave height estimates that are more accurate, although not as regular, and use these data sets for the final estimation of any parameters entering the covariance structure of the underlying field.

This principle is illustrated in the next section, where an empirical estimate of the spatio-temporal covariance structure is obtained using the ERA-Interim data set and is used to check for the presence of stationarity and isotropy of the underlying field. The affirmative answer simplifies the model considerably making it possible to compute again the empirical covariance of the field using satellite data and then fit an appropriate parametric model.

Buoy data will be used only in Section 4.2 as a reference to compare the accuracy of the significant wave height obtained by interpolating satellite data using the different space–time models. There are some questionable extremely high values (above 20 m!) in the time series considered, so in order to filter out these values we have applied a moving median filter, with a time-window of width 6 h. This is also expected to bring the temporal scale closer to the one of ERA-Interim data set.

2. MODEL WITH CONSTANT VELOCITY

The observations consist of the significant wave heights taken at discrete locations either along the one-dimensional satellite tracks or at the ERA-Interim grid points. In order to simplify the initial analysis, we only consider a central region of the North Atlantic, with latitudes ranging between 25W and 45.5W and longitudes ranging between 40N and 50N (see Figure 2), which from now on will be denoted by D_0. It would seem unlikely that the field is stationary in time, its characteristics are expected to change seasonally at least, although probably inter-annual components (see e.g., Athanassoulis and Stephanakos, 1995) are also present. The annual cycle generally dominates the within-year variability of the significant wave height field, especially in the northern hemisphere, see Baxevani *et al*. (2005). One way of dealing with seasonal components is to fit an annual cycle to the data. An alternative way, which is employed here, is to focus on 1 month at a time and consider the data from that month over the different years as independent realizations of the same time-stationary random field. This is a usual assumption for meteorological processes, despite the fact it does not take into account the low frequency variation due to inter-annual variability (trend, NAO, etc). In this paper we present results only for the month of December.

Figure 4 shows the empirical mean and standard deviation at each grid point in the area D_0, computed using the ERA-Interim significant wave height data set from the 16 different December months (period 1992–2007). It is apparent that the data are non-stationary in space, the mean values and variability in the north of D_0 appear to be higher than in the south. Notice the correlation between the magnitude of the mean values of significant wave height and variability. High mean values correspond to high variability and areas with calmer conditions like in the south of D_0 have smaller variance. The data have been standardized locally, that is at each grid point, by removing the mean and scaling by the standard deviation.

In order to make the two estimates, from the ERA-Interim and the satellite altimeter, compatible, we have considered 1.5×1.5 degree boxes centered at the ERA-Interim grid points and used all the satellite data that fall inside each one to obtain estimates of the mean and standard deviation, for the 16 December months. These estimates, which can be seen in Figure 5, although they exhibit an important spatial variability, which may be due to the low space–time sampling of the satellite data (the sample size for each box can be seen in Figure 5, (right panel)), present the same overall spatial trend with the ERA-Interim data, compare to Figure 4, although the actual values of the mean and variance are quite different as can be seen in Table 1.

This sampling variability could probably be reduced by either increasing the box size or by using some type of spatial smoothing. Here we decided to scale the data by removing the mean-field estimates obtained using the ERA-Interim data to which the difference of the two overall means (i.e., the difference between the two readings in the first row of Table 1) was added, and then scale by dividing by the ERA-Interim standard deviation corrected by the ratio of the two overall standard deviations given in Table 1 (the two readings on the second row).

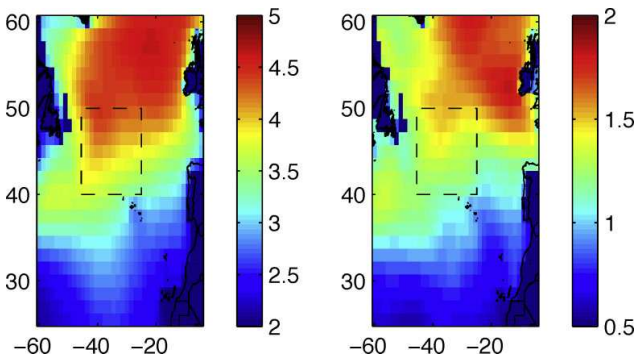


Figure 4. Mean (left panel) and standard deviation (right panel) of H_s computed using ERA-Interim data (December 1992–2007). The dashed box indicates the domain D_0. This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

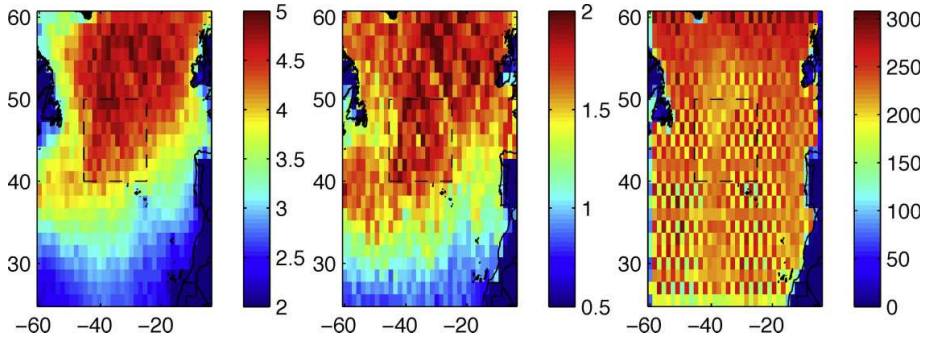


Figure 5. Mean (left panel) and standard deviation (middle panel) of H_s computed from satellite data (December 1992–2007). The right panel is the sample size for each 1.5×1.5 degree box. The dashed boxes indicate the domain D_0. This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

In the following, we shall consider that the standardized versions of the significant wave heights are partial observations of a random field $Y(\mathbf{p}, t)$ where $\mathbf{p} \in D_0 \subset \mathbb{R}^2$ and $t \in T$, measured in days. Hence,

$$E(Y(\mathbf{p}, t)) = 0, \quad \mathrm{Var}(Y(\mathbf{p}, t)) = 1.$$

We further assume that the field is stationary both in space and time, i.e., for all $\mathbf{p}, \mathbf{p}' \in D_0$ and $t, t' \in T$

$$\mathrm{Cov}(Y(\mathbf{p}, t), Y(\mathbf{p}', t')) = C_Y(\mathbf{p}' - \mathbf{p}, t' - t). \tag{1}$$

To check this simplifying assumption, we have computed the covariance $\mathrm{Cov}(Y(\mathbf{p}, t), Y(\mathbf{p}', t'))$ for various values of $\mathbf{p}$ and t as functions of $\mathbf{p}'$ and t'. The resulting estimates seem relatively independent of the choice of $\mathbf{p}$ and t.

An empirical estimate of C_Y for different time lags computed using the ERA-Interim data, can be seen in Figure 6. Notice that the maximum of the spatial covariance $C_Y(\cdot, t)$ drifts to the West as the time lag increases. This is due to the prevailing sea-state motion to the East. Such behavior strengthens the arguments for using a covariance structure that includes some sort of dynamics, like the covariance

$$C_Y(\mathbf{p}' - \mathbf{p}, t' - t) = C(\mathbf{p}' - \mathbf{p} - \mathbf{V}_0(t' - t), t' - t) \tag{2}$$

where $\mathbf{V}_0$ denotes the mean velocity of the sea-state, and which can be thought as the covariance of a static field subordinated by the constant velocity $\mathbf{V}_0$. That is, let $X(\mathbf{p}, t)$ be a field with spatio-temporal covariance function $C(\mathbf{p} - \mathbf{p}', t - t')$, then the field $X(\mathbf{p} - \mathbf{V}_0 t, t)$ has covariance function

$$\mathrm{Cov}(X(\mathbf{p} - \mathbf{V}_0 t, t), X(\mathbf{p}' - \mathbf{V}_0 t', t')) = C(\mathbf{p}' - \mathbf{p} - \mathbf{V}_0(t' - t), t' - t).$$

Various parametric covariance models have been considered for C, including some usual separable models as well as the non-separable model proposed in Gneiting *et al.* (2007). The best fit to the empirical covariance function using standard weighted least-square method (see e.g., Cressie, 1993), has been obtained for the simple non-separable rational quadratic model

$$C(\mathbf{p}' - \mathbf{p}, t' - t) = (1 - C_0)\mathbf{1}_{\{0,0\}}(\mathbf{p}' - \mathbf{p}, t' - t) + \frac{C_0}{1 + \frac{d(\mathbf{p},\mathbf{p}')^2}{\theta_S^2} + \frac{|t'-t|^2}{\theta_T^2}}. \tag{3}$$

The parameters θ_S and θ_T describe the spatial and temporal range respectively, $1 - C_0$ is the space–time nugget effect and $d(\mathbf{p}, \mathbf{p}')$ denotes the geodesic distance on the sphere between locations $\mathbf{p}$ and $\mathbf{p}'$. The nugget effect may model small-scale structures, those usually not observed in the ERA-Interim data because of the space–time sampling resolution, and the measurement error which may be present in the satellite data. Eventually, different nugget effects could be used for the different satellites.

Table 1. Sample mean and sample variance of H_s in D_0 over the months of December for the period 1992–2007

	ERA-Interim	Satellite
Mean	4.25	4.39
Variance	2.29	3.07

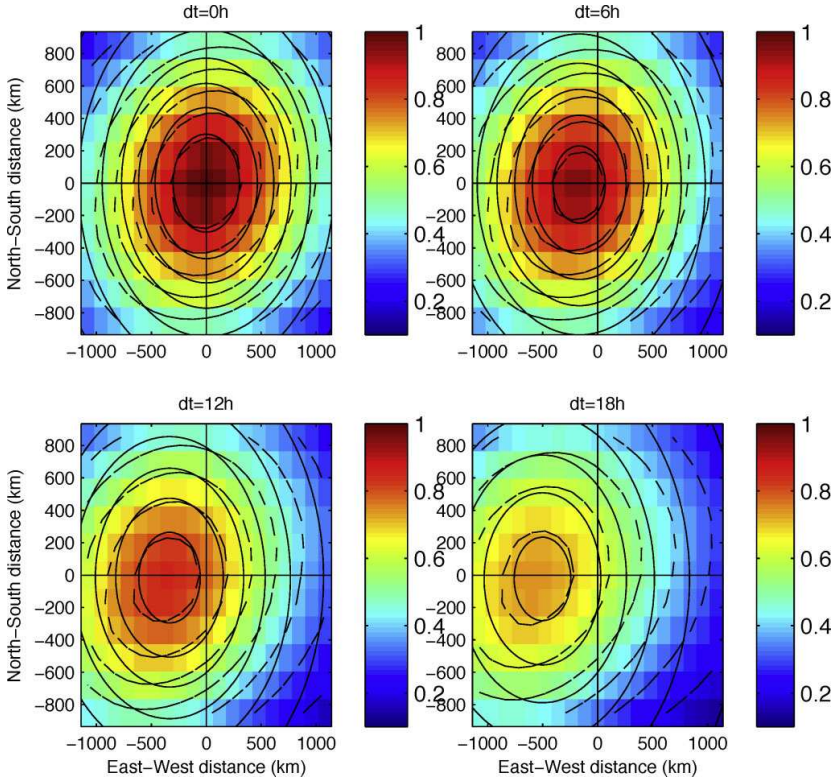


Figure 6. Empirical covariance function for ERA-Interim data for time lags $|t - t'| = 0\,\text{h}$ (top left panel), $|t - t'| = 6\,\text{h}$ (top right panel), $|t - t'| = 12\,\text{h}$ (bottom left panel) and $|t - t'| = 18\,\text{h}$ (bottom right panel). The dashed lines represent the levels of the empirical covariance function and the full lines the levels of the fitted parametric model. This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

Model (3) has been fitted to the empirical covariance function, see Figure 6. The elliptic level curves of the parametric model seem to fit the overall shape of the empirical covariance function both in space and time and in particular the model seems to be able to describe the mean motion of the sea-state. The estimated mean velocity $\mathbf{V}_0$ corresponds to a mean motion of $27.8\,\text{km}\,\text{h}^{-1}$ to the East and $1.3\,\text{km}\,\text{h}^{-1}$ to the North, which seems to be a physically realistic velocity for that area.

Notice that model (3) becomes isotropic after removing the mean motion, whereas Figure 6 suggests the presence of a small anisotropy. And indeed a slightly better fit was achieved by using an anisotropic model. However, because of the geometry of the satellite tracks that provide information only along two main directions, the assumption of isotropy simplifies considerably the estimation of the covariance function using satellite data (see also Baxevani *et al*., 2008).

3. MODEL WITH DYNAMIC VELOCITY

In the previous section, we incorporated the motion of the sea-states into the model by considering a static field subordinated by a constant velocity. Such a field is still not optimal since the assumption of constant velocity over such large region and for a time span greater than 10 h, is not realistic, see Baxevani *et al*. (2009). In this section, we propose a new approach by subordinating a static field with a dynamically changing velocity. The section is organized as follows: first we define velocity through a flow of diffeomophisms that are the solution to the transport equation and discuss ways to incorporate these velocity fields into the covariance model. Then, we describe a method for estimating the sea-state velocity fields within a state-space model framework, with the aim towards a spatio-temporal consistency.

3.1. Dynamic velocity

In this section we introduce sea-state motion using a flow of diffeomorphisms. Denote by $\boldsymbol{\phi}(\mathbf{p}, t, s)$ the motion of the point $\mathbf{p}$ between times t and s in the interval $[t_0 - \Delta T, t_0 + \Delta T]$ for some arbitrary reference point t_0. Assume also that the flow satisfies $\boldsymbol{\phi}(\mathbf{p}, t, t) = \mathbf{p}$ and the flow property $\boldsymbol{\phi}(\cdot, t, s) = \boldsymbol{\phi}(\cdot, u, s) \circ \boldsymbol{\phi}(\cdot, t, u)$ and that the inverse of $\boldsymbol{\phi}(\mathbf{p}, t, s)$ exists and is denoted by $\boldsymbol{\phi}^{-1}(\mathbf{p}, t, s)$. This construction is sensible both from the physical and mathematical point of view. The physical analogy is that the sea-state develops over time. Mathematically for

every diffeomorphism $\boldsymbol{\phi}$ there exist a velocity field $\mathbf{V}(\mathbf{p}, t) = (u(\mathbf{p}, t), v(\mathbf{p}, t))$ such that (see Markussen, 2007):

$$\boldsymbol{\phi}(\mathbf{p}, t, s) = \mathbf{p} + \int_t^s \mathbf{V}(\boldsymbol{\phi}(\mathbf{p}, t, \tau), \tau)\, \mathrm{d}\tau. \tag{4}$$

This differential equation has been employed to describe sea-state dynamics in Baxevani *et al*. (2009) and in Joshi and Miller (2000) to solve the landmark matching problem.

For mathematical convenience we will make assumptions on the flow of diffeomorphisms which may be unrealistic from a physical point of view. For example, the sea-state motions should be differentiable and define one to one transformations: every sea-state at time t should be uniquely matched to a sea-state at time t' and thus different sea-states cannot "cross" as they do in the real world. There may also be some boundary problems since sea-states appear or disappear from the domain D_0 as they travel. In order to avoid losing too much data when the sea-states move, in the next section the velocity field is computed on a bigger domain D which includes D_0. In practice, D has been chosen in order to ensure that the flow of differomorphisms $\boldsymbol{\phi}(\mathbf{p}, t_0, t)$ is generally defined for all $\mathbf{p} \in D_0$ and $t \in [t_0 - \Delta T, t_0 + \Delta T]$ for ΔT less than one day.

Using the flow of diffeomorphisms defined in Equation (4), a natural generalization of the frozen covariance model discussed in Section 2 is obtained by assuming that

$$\mathrm{Cov}(Y(\mathbf{p}, t), Y(\mathbf{p}', t')) = C(\boldsymbol{\phi}^{-1}(\mathbf{p}', t_0, t') - \boldsymbol{\phi}^{-1}(\mathbf{p}, t_0, t), t' - t) \tag{5}$$

for the covariance function C given in Equation (3). The physical interpretation of this model is that the covariance function depends on the locations where points $\mathbf{p}'$ and $\mathbf{p}$ were at time t_0 before arriving at times t' and t respectively to their current locations. For velocity fields constant in space, i.e., $\mathbf{V}(\mathbf{p}, t) = \mathbf{V}_0$ for all $\mathbf{p}, t$, the covariance function in Equation (5) simplifies to the covariance function in Equation (2). Relation (5) is equivalent to assuming C is the covariance function of the field in the Lagrangian reference frame, i.e., the covariance function of the field Z defined by

$$Z(\mathbf{p}, t) = Y(\boldsymbol{\phi}^{-1}(\mathbf{p}, t_0, t), t). \tag{6}$$

Here we should notice that the covariance given in Equation (5) is generally a non-stationary one, except in the particular case the flow of diffeomorphisms is such that $\boldsymbol{\phi}^{-1}(\mathbf{p}', t_0, t') - \boldsymbol{\phi}^{-1}(\mathbf{p}, t_0, t)$ is only a function of $t' - t$ and $\mathbf{p}' - \mathbf{p}$. This is the case for example when the velocity field is constant in space (i.e., $\mathbf{V}(\mathbf{p}, t) = \mathbf{V}(t)$ for all $\mathbf{p}, t$), case which includes the constant velocity model introduced in Section 2.

This apparent contradiction with the stationarity assumptions which have been made in Section 2 may be better understood if we consider the velocity field as a random field, in which case the covariance function in Equation (5) should be understood as a conditional expectation given the velocity field, and a more correct way of writing the model is

$$\mathrm{Cov}(Y(\mathbf{p}, t), Y(\mathbf{p}', t')) = E[\mathrm{Cov}(Z(\boldsymbol{\phi}^{-1}(\mathbf{p}, t_0, t), t), Z(\boldsymbol{\phi}^{-1}(\mathbf{p}', t_0, t'), t'))], \tag{7}$$

where the expectation is taken with respect to the random field $\boldsymbol{\phi}$. Writing a proper stochastic model for the velocity field $\mathbf{V}$ for which Equation (7) leads to a second-order stationary space–time covariance function for Y will be the subject of future work.

3.2. Motion estimation

After having introduced the motion of sea-states through a flow of diffeomorphisms that are the solution to the transport equation, in this section we present a method for estimating the associated velocity field.

3.2.1. Framework

To estimate the velocity of the sea-state motion we use the ERA-Interim data presented in Section 1.1, since the coverage of the satellite data is generally poor and does not provide enough information to track correctly the motion of the sea-state systems. The regular coverage of the ERA-Interim data does not only provide us with more information but also simplifies the modeling since it allows the use of existing models for processes on regular space–time lattices. The objective is nevertheless to use the estimated velocity fields to model the satellite data.

The most widely used technique for motion estimation is based on maximizing the "local" correlation between two rectangular regions in successive images, and this method has been used in particular for meteorological and oceanographic applications , see e.g., Schmetz *et al*. (1993) and Marcello *et al*. (2008). In parallel, in the computer vision community, the problem of motion estimation has been addressed using differential methods. It consists in solving a partial differential equation (PDE) system which is built assuming conservation of the intensity of the displaced object and a certain spatial regularity of the flow, suggestions for which in the particular case of fluid motion can be found in Corpetti *et al*. (2002).

One drawback of the above-mentioned methods is that the velocity fields are estimated independently of each other, using only the information contained in pairs of successive images. As a consequence the temporal consistency of the estimated velocity fields is not guaranteed although is expected since the wave systems are traveling with almost constant velocity. Indeed the local correlation method when applied to the ERA-Interim data failed to reproduce the temporal consistency that is usual in geophysical flows. Different classes of methods allow to deal with this lack of coherence by introducing dynamical information on the velocity fields using *a priori* information that may

come from physical knowledge. The variational methods, for instance, are related to optimal control theory (Le Dimet and Talagrand, 1986). In that framework, a sequence of motion fields is estimated knowing an initial state, a dynamical model, and noisy and possibly incomplete observations. Dynamically consistent motion estimates are then obtained on a fixed time interval, given observations over the whole period (Korotaev *et al*., 2008; Papadakis *et al*., 2007).

In this work, the motion estimation problem is formulated and solved sequentially within a state-space model framework. The hidden state is the velocity field, which is supposed to be a Markovian process with a transition kernel that is parameterized using a simple physical model to warranty that the velocity fields evolve slowly in space and time. The hidden state (velocity field) is related to the observations (ERA-Interim data) through a conservation of the characteristics of the moving sea-states between successive times. Then, the velocity fields are estimated using a particle filter which permits to compute approximations of the distribution of the hidden state given the observations. State-space models have already been used for motion estimation in Ailliot *et al*. (2006) and Cuzol and Mémin (2009). In Ailliot *et al*. (2006), the method is applied to a similar hindcast data set, although the velocity was supposed to be constant in space and the hidden state was discretized for computational reasons. In Cuzol and Mémin (2009), the hidden velocity fields are guided by an *a priori* dynamical law constructed from fluid flow equations.

3.2.2. State-space model

In this section, we describe the state-space model which has been used to estimate the sea-state motion.

We commence by introducing some notation. Let $D = (\mathbf{p}_1, \ldots, \mathbf{p}_N)$ be the set of the $N = 336$ grid points of the ERA-Interim data set which are located between longitude 16.5W and 51W and latitude 36N and 55.5N (see Figure 2). This domain has been chosen empirically so that the motion of the sea states in D_0 can be followed on a ± 24 h time window. Moreover, let $\mathbf{V}(D, t) = (u(\mathbf{p}_1, t), \cdots, u(\mathbf{p}_N, t), v(\mathbf{p}_1, t), \cdots, v(\mathbf{p}_N, t))^{\mathrm{T}}$ be the vector corresponding to the velocity field at time t, with u and v denoting the zonal and meridional components respectively.

Usually for the description of the sea-state we have available only some statistics related to the spectrum governing the sea surface process. Most often these are the significant wave height (H_s), the mean direction of propagation (Θ_m), and the mean period (T_m) which are included among others in the ERA-Interim data set. Let us then denote by

$$\begin{aligned} S(D, t) = \; & (H_s(\mathbf{p}_1, t)\cos(\Theta_m(\mathbf{p}_1, t)), \cdots, H_s(\mathbf{p}_N, t)\cos(\Theta_m(\mathbf{p}_N, t)), \\ & H_s(\mathbf{p}_1, t)\sin(\Theta_m(\mathbf{p}_1, t)), \cdots, H_s(\mathbf{p}_N, t)\sin(\Theta_m(\mathbf{p}_N, t)), \\ & T_m(\mathbf{p}_1, t), \cdots, T_m(\mathbf{p}_N, t))^{\mathrm{T}}, \end{aligned}$$

the vector that describes the sea-state conditions at time t over the region D (see Krogstad and Barstow, 1999). In practice, this information is available at discrete times with a regular time step of 6 h.

Dynamics of the hidden velocity field. The time evolution of the hidden state $\mathbf{V}$ is modeled using a mixture of two models. The first model is a physical approximation of the velocity of a wave group which is valid for sea-states with narrow band spectrum (see e.g., Whitham, 1974). Hereafter, $\mathbf{V}_{\mathrm{disp}}(\mathbf{p}_i, t) = (u_{\mathrm{disp}}(\mathbf{p}_i, t), v_{\mathrm{disp}}(\mathbf{p}_i, t))^{\mathrm{T}}$ denotes the group velocity at time t and location $\mathbf{p}_i$ and $\mathbf{V}_{\mathrm{disp}}(D, t)$ the associated velocity field obtained by concatenating the velocity at the different locations. The group velocity is related to the mean period and direction of the sea-state through the dispersion relation:

$$\mathbf{V}_{\mathrm{disp}}(\mathbf{p}_i, t) = \frac{g}{4\pi} T_m(\mathbf{p}_i, t) \begin{pmatrix} \cos(\Theta_m(\mathbf{p}_i, t)) \\ \sin(\Theta_m(\mathbf{p}_i, t)) \end{pmatrix}, \tag{8}$$

where g is the gravitational constant. Equation (8) should provide a good approximation of the velocity of the sea-state when a unique swell system and calm wind conditions are dominating. However, such weather conditions are not always prevailing in the North Atlantic in which case a simple AR(1) model may provide a better approximation of the sea-state dynamics.

To be specific, let C_t denote a Bernoulli variable which governs the choice of the model at time t, then

$$\mathbf{V}(D, t) = C_t \mathbf{V}_{\mathrm{disp}}(D, t) + (1 - C_t)[\mu + A(\mathbf{V}(D, t - \Delta t) - \mu)] + \epsilon_t \tag{9}$$

with $A \in \mathbb{R}^{2N \times 2N}$, $\mu \in \mathbb{R}^{2N}$, and $\{\epsilon_t\}$ a Gaussian white noise sequence with zero mean and covariance matrix Σ^{dyn}. We further assume that $\{C_t\}$ is an i.i.d. sequence of Bernoulli variables with parameter π_c which is independent from both the velocity and the sea-state fields. Note that particle-based multiple model filters may also introduce a Markov structure on $\{C_t\}$ (see McGinnity and Irwin, 2001 for instance).

Observation equation. In order to relate the hidden velocity fields to the observed sea-state conditions, we assume that the characteristics of a sea-state evolve slowly between two time instants if we follow correctly its motion, i.e., that for all $i \in \{1, \ldots, N\}$,

$$S(\mathbf{p}_i, t) = S(\mathbf{p}_i - \mathbf{V}(\mathbf{p}_i, t)\Delta t, t - \Delta t) + \eta(\mathbf{p}_i, t), \tag{10}$$

where $\eta(\mathbf{p}_i, t)$ denote the evolution of the sea-state fields between two time steps, then it is further assumed that $\{\eta\}$ is a Gaussian white noise sequence with covariance matrix Σ^{obs}.

Parametrization. Let us now give more details on the parametrization of the various matrices of parameters which appear in the state-space model:

- *Autoregressive matrix A*: We assume a block structure

$$A = \begin{pmatrix} A_{1,1} & 0 \\ 0 & A_{2,2} \end{pmatrix}$$

with $A_{1,1} = A_{2,2}$ and $A_{1,1}(i, j) \propto \exp(-\frac{d(\mathbf{p}_i, \mathbf{p}_j)}{\lambda_A})$ and the normalizing constraint $\sum_{j=1}^{N} A_{1,1}(i, j) = \theta$ for some constant θ. The velocity field at time t is thus obtained by smoothing the velocity field at time $t - \Delta t$, using weights which decrease with distance, and satisfy a certain constrain $\theta < 1$ in order to warranty the stability of the AR model. Due to the difficulties presented in setting up an automatic procedure for estimating the unknown parameters, those have been chosen using empirical knowledge hence, $\theta = 0.9$ and $\lambda_A = 200$ km, these values agree with the correlation length found in (Baxevani *et al*., 2005, 2008).
- *Mean vector* μ: μ denotes the mean of the stationary distribution of the AR(1) model. We assume that this mean is the same at all locations and we used the parameter values obtained when fitting the model with constant velocity (see Section 2).
- *Covariance matrix* Σ^{dyn}: We assume that the velocity innovations on the zonal and the meridional component are independent, i.e., that Σ^{dyn} has also a block structure

$$\Sigma^{\text{dyn}} = \begin{pmatrix} \Sigma^{\text{dyn}}_{1,1} & 0 \\ 0 & \Sigma^{\text{dyn}}_{2,2} \end{pmatrix}$$

where $\Sigma^{\text{dyn}}_{1,1} = \Sigma^{\text{dyn}}_{2,2}$ are $N \times N$ spatial covariance matrices with standard exponential shape

$$\Sigma^{\text{dyn}}_{1,1}(i, j) = \sigma^2_{\text{dyn}} \exp\left(-\frac{d(\mathbf{p}_i, \mathbf{p}_j)}{\lambda_{\text{dyn}}}\right) \tag{11}$$

where σ^2_{dyn} represents the variance of the innovation at each location and $\lambda_{\text{dyn}} > 0$ its spatial range. In practice, we have used the parameter values $\sigma_{\text{dyn}} = 1$ and $\lambda_{\text{dyn}} = 500$ km.
- *Covariance matrix* Σ^{obs}: We also use a block structure

$$\Sigma^{\text{obs}} = \begin{pmatrix} \Sigma^{\text{obs}}_{1,1} & 0 & 0 \\ 0 & \Sigma^{\text{obs}}_{2,2} & 0 \\ 0 & 0 & \Sigma^{\text{obs}}_{3,3} \end{pmatrix}$$

where $\Sigma^{\text{obs}}_{1,1} = \Sigma^{\text{obs}}_{2,2} = \Sigma^{\text{obs}}_{3,3}$ are $N \times N$ spatial covariance matrices with coefficients

$$\Sigma^{\text{obs}}_{1,1}(i, j) = \sigma^2_{\text{obs}} \exp\left(-\frac{d(\mathbf{p}_i, \mathbf{p}_j)}{\lambda_{\text{obs}}}\right)$$

In practice, we have used the parameter values $\sigma_{\text{obs}} = 0.5$ and $\lambda_{\text{obs}} = 500$ km.

3.2.3. *Particle filter*

The model described in the previous section is used to estimate hidden velocity fields that are consistent with the sea-state fields given by the ERA-Interim data set. For that, we compute the conditional expectation

$$\hat{\mathbf{V}}(D, t) = E[\mathbf{V}(D, t)|S(D, t_0), \cdots, S(D, t - \Delta t), S(D, t)]$$

of the velocity field at time t given the history of the sea-state fields up until time t. Notice that a possible refinement would consist in computing the smoothing probabilities and then estimate the hidden velocity by the conditional expectation $E[\mathbf{V}(D, t)|S(D, t_0), \cdots, S(D, t - \Delta t), S(D, t), S(D, t + \Delta t), \ldots, S(D, T)]$ of the velocity field given both past and future sea-state conditions.

The observation Equation (10) is nonlinear, and in such situation it is usual to compute sequential Monte Carlo approximations of the conditional expectation using a particle filter algorithm. At each time step t, the conditional filtering distribution $p(\mathbf{V}(D, t)|S(D, t_0), \cdots, S(D, t - \Delta t), S(D, t))$ is approximated by a set of M weighted particles $\{\mathbf{V}^{(i)}(D, t), \omega_t^{(i)}\}_{i=1:M}$ where $\mathbf{V}^{(i)}(D, t) \in \mathbb{R}^{2N}$ are velocity fields and $\omega_t^{(i)} \in [0, 1]$ the associated weights. The conditional expectation is then approximated by

$$\hat{\mathbf{V}}(D, t) = \sum_{i=1}^{M} \omega_t^{(i)} \mathbf{V}^{(i)}(D, t)$$

and the weighted set of particles is updated at each time according to a Sequential Importance Sampling and Resampling scheme (see e.g., Cappé *et al*., 2005). The sampling step at time t consists in simulating the velocity field using the dynamical model (9). We first generate

independently $\{C_t^{(i)}\}_{i=1:M}$ according to a Bernoulli distribution with probability π_c, and then $\mathbf{V}^{(i)}(D,t)$ is simulated using either the group velocity or the AR process depending on the value of $C_t^{(i)}$. The weights $\{\omega_t^{(i)}\}_{i=1:M}$ are then computed as follows:

$$\omega_t^{(i)} \propto \exp(-(\|S(D,t) - S(D - \mathbf{V}^{(i)}(D,t)\Delta t, t - \Delta t)\|_{\Sigma_{\text{obs}}})^{\nu}) \tag{12}$$

with $\|d\|^2_{\Sigma_{\text{obs}}} = d^{\mathrm{T}}\Sigma_{\text{obs}}^{-1}d$ and the normalizing constraint $\sum_{i=1}^{M}\omega_t^{(i)} = 1$. The parameter ν controls the decay of the weights with distance between the observed sea-state field at time t and the previous field after moving. The Gaussian model (10) corresponds to the case $\nu = 2$. In practice, due to the high dimension of the observation space, this choice leads to a quick degeneracy of the particles: after a few iterations, one weight is almost equal to one whereas all the others close to zeros (see also Berliner and Wikle, 2007). In order to avoid this problem, after several trials ν was chosen equal to $2/3$.

Notice that the computation of the weights using Equation (12) requires the sea-state fields $S(D - \mathbf{V}^{(i)}(D,t)\Delta t, t - \Delta t)$ at other locations than those given by ERA-Interim data set. In practice, we use a simple linear interpolation of the ERA-Interim field at time $t - 1$.

The particle filter was run with $M = 10^4$ particles. It seems to be a good compromise since it leads to reasonable CPU time and robust estimations of the velocity fields.

3.2.4. Visual validation

Before including the velocity fields in the covariance model, we have first performed visual check of their realism. An example of results is shown in Figure 7. It displays a 24 h sequence of ERA-Interim fields where two low pressure systems can easily be identified: one in the North-West which is moving to the East and one in the South-East which is almost fixed. The velocity fields obtained with a mixing proportion $\pi_c = 1/3$ seem to be able to track these two systems, whereas use of a pure autoregressive model ($\pi_c = 0$) leads to a bad estimation of the velocity of the static system. In general, introducing the dispersion relation in the dynamics leads only to minor improvements except in some specific situations. We do believe however that the benefits would be substantial in areas where swell conditions are more dominating.

It is also important to check that the estimated velocities are realistic for the significant wave height as measured by the satellites. The left panel in Figure 8 shows all the satellite data $H_s^{\text{sat}}(\mathbf{p},t)$ as a function of $\mathbf{p} \in D$ available at time $t \in I$ where I is a 48 h time window centered at some arbitrary time t_0. For comparison purposes, we also show the significant wave height field given by the ERA-Interim data set at time t_0. Due to the motion of sea-states, there are important differences between some data $H_s^{\text{sat}}(\mathbf{p},t)$ and $H_s^{\text{sat}}(\mathbf{p}',t')$ which are close to each other on the Figure 8 (left panel) ($\mathbf{p} \approx \mathbf{p}'$) but which have been moving differently between time t (resp. t') and t_0. The right panel in Figure 8 shows the same satellite data in the Lagrangian reference frame induced by the estimated velocity field $\hat{\mathbf{V}}$. More precisely, if $\boldsymbol{\phi}$ denotes the flow of diffeomorphisms associated to $\hat{\mathbf{V}}$, we plot $H_s^{\text{sat}}(\boldsymbol{\phi}^{-1}(\mathbf{p},t_0,t),t)$ as a function of $\mathbf{p}$. Since the flow of diffeomorphisms permits to follow the motion of the sea-states we expected an improved spatial coherence in the new reference frame. And indeed, on the right panel of Figure 8, the high significant wave height values are all located in the same area which corresponds to a storm location at time t_0. The storm can also be seen on the ERA-Interim data.

Figure 9 shows the estimated empirical joint probability density function of the zonal and the meridional components of the velocity at a central location $\mathbf{p}_0$ =(46.5N,34.5W). The distribution shows again that sea-state systems are generally traveling to the east, but also that there

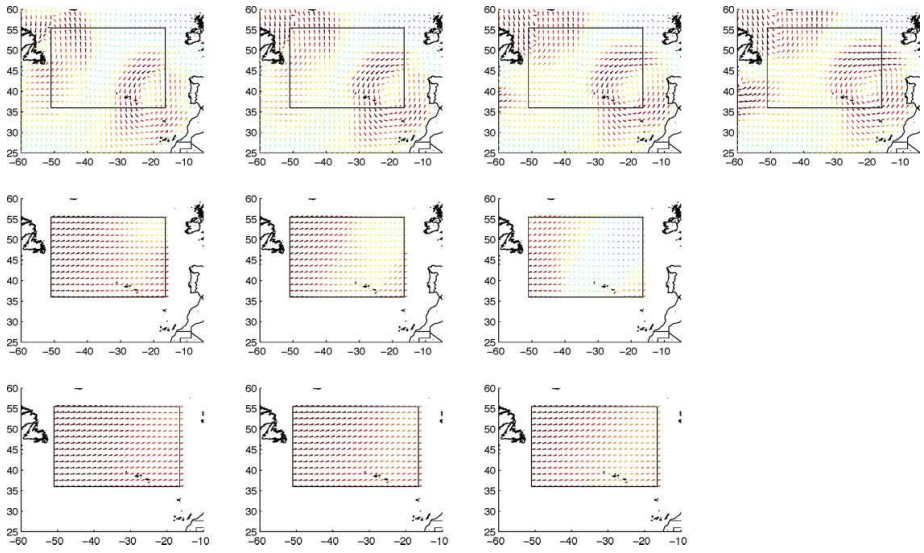


Figure 7. ERA-Interim fields at consecutive times at 6 h intervals between 16 December 2002 at 18:00 and 17 December 2002 at 12:00 (top panels) and associated motions with the dispersion relation (middle panels), and without the dispersion relation (bottom panels). This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

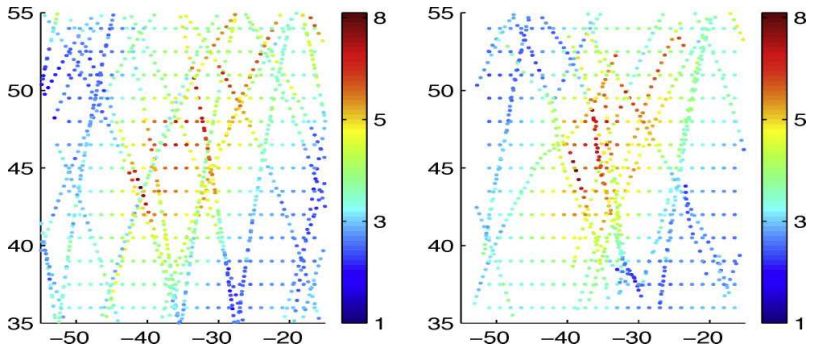


Figure 8. All available satellite tracks for a period of time of 48 h centered at time t_0 with ERA-Interim H_s field at time t_0: 17 December 2002; left: without use of the displacements (Eulerian fields) and right: using the displacements (Lagrangian fields). This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

is an important variability which cannot be accommodated by the constant velocity model. The mean velocity over the months of December 1992–2007 in D_0 is 19.5 $\mathrm{km h}^{-1}$ to the East and 0.5 $\mathrm{km h}^{-1}$ to the North, and this is slightly smaller than the mean velocity computed in Section 2 by fitting the constant velocity model. Figure 9 also shows that there is a positive dependence between the estimated velocities at location $\mathbf{p}_0$ and both the wind speed and the mean period of the sea state which are given by the ERA-Interim data set. The analogous plots for the direction show also a clear relation between the direction the sea-states are traveling and both the wind direction and the mean direction of the sea-state Θ_m. This confirms that the estimated velocities are linked not only to the physical propagation of the waves but also to the motion of the wind fields.

3.3. Covariance estimation

Figure 10 shows empirical estimates of the covariance function computed using ERA-Interim and satellite data in the Lagrangian reference frame, i.e., as a function of $d(\boldsymbol{\phi}^{-1}(\mathbf{p}', t_0, t'), \boldsymbol{\phi}^{-1}(\mathbf{p}, t_0, t))$ and $|t' - t|$ and the fitted parametric model (3). This has been done for three different configurations. First for the static case, that is no dynamics are present, secondly, for constant velocity $\mathbf{V}_0$ and finally for the case where velocity is modeled in the presence of the dispersion relation. Inclusion of velocity into the covariance functions results to longer temporal range dependence, even longer when velocity is modeled using dispersion relation, (see also Table 2). This effect was to be expected since the Lagrangian reference system moves along with the sea-states. The rate of the spatial dependence seems to be unaffected by the introduction of dynamics in the model.

Overall, the agreement between the empirical covariance function computed using the ERA-Interim and the satellite altimeter data and the fitted parametric covariance models, seem to be satisfactory as manifested in Figure 10. However, comparing the entries in Table 2, we can draw some further conclusions. There seems to be some inconsistency in the sizes of spatial dependence between the ERA-Interim and the satellite data sets, which is of the order of 100 km (relative difference of about 10%) in the absence of dynamics. This difference however drops to half when dynamics (non-constant velocity) are included in the model.

Summarizing, modeling sea-state motion using the dispersion relation leads to a covariance model with longer temporal dependence which should allow better tracking of the sea-state motion.

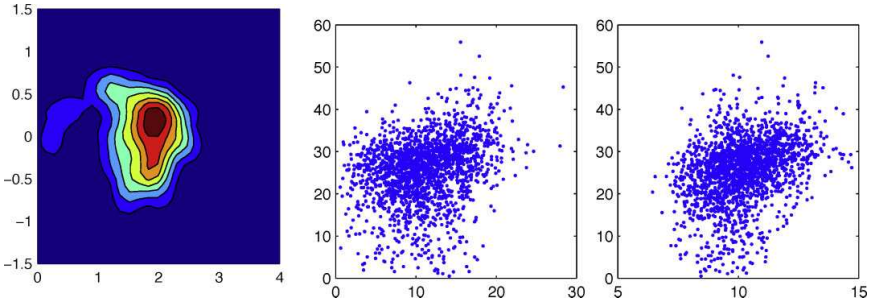


Figure 9. Left panel: empirical bivariate pdf of $(u(\mathbf{p}_0), v(\mathbf{p}_0))$ at location $\mathbf{p}_0$ with coordinates (46.5N,34.5W). Middle panel: relation between the wind speed in ms^{-1} (x-axis) and the velocity of the sea-state motion in $\mathrm{km\,h}^{-1}$ (y-axis) at location (46.5N,34.5W). Right panel: relation between the mean period of the sea-state T_m in s (x-axis) and the velocity of the sea-state motion in $\mathrm{km\,h}^{-1}$ (y-axis) at location (46.5N,34.5W). This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

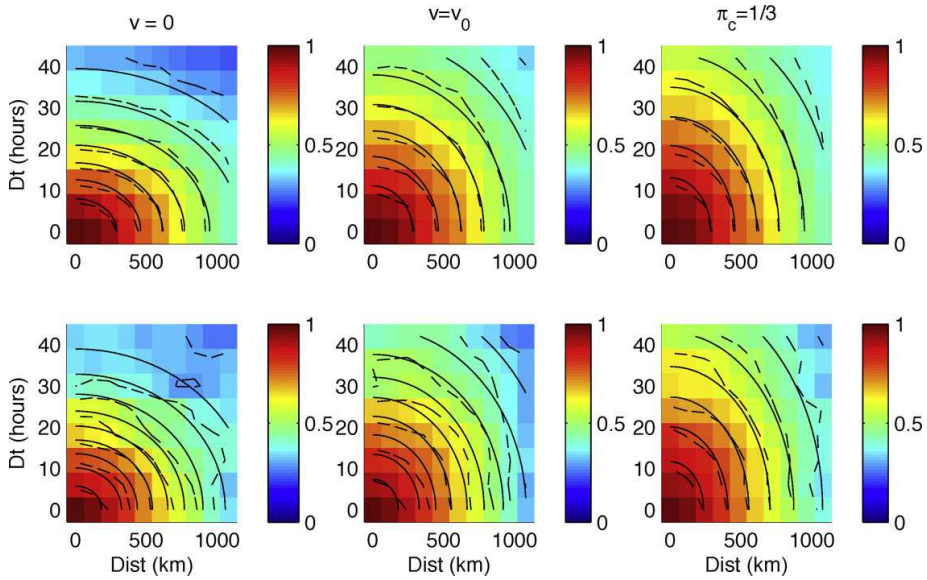


Figure 10. Empirical correlation function of the Lagrangian field computed using ERA-Interim data (top panels) and satellite data (bottom panels). Left panels: no motion ($\mathbf{V}(\mathbf{p}, t) = 0$), middle panels: constant velocity ($\mathbf{V}(\mathbf{p}, t) = \mathbf{V}_0$), right panels: dynamic velocity with dispersion relation ($\pi_c = 1/3$). The dashed lines represent the levels of the empirical covariance function and the full lines the levels of the fitted parametric model. This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

4. NUMERICAL RESULTS

In this section, we check the accuracy of the fields obtained by interpolating satellite data using the model presented in this paper and ordinary kriging (see e.g., Cressie, 1993). This could be useful, in order for example, to produce historical data for metocean studies. We first perform cross-validation on satellite data and then make a comparison with a buoy.

4.1. Cross-validation of satellite data

In this section, the whole methodology is validated using cross-validation on satellite data. For that, each satellite passage which intersects D_0 is predicted using all other satellite passages that are available in a 2 day time window. The global root mean square errors (RMSE) of the difference between the true satellite data and the interpolated values are given in Table 3. The model with dynamic velocity performs best and the introduction of the velocity (even constant) clearly improves the static model. Introducing the dispersion relation in the dynamics does not seem to improve things a lot, but again we expect more benefits in an area dominated by swell conditions.

Table 2. Parameter values of the fitted parametric covariance model (3) for different velocity fields

Velocity	Static model		Constant velocity		Dynamic ($\pi_c = 0$)		Dynamic ($\pi_c = \frac{1}{3}$)	
data	ERA	Satellite	ERA	Satellite	ERA	Satellite	ERA	Satellite
θ_S (km)	966	873	997	896	966	913	967	909
θ_T (h)	26.1	27.1	39.1	36.6	41.2	43.3	43.9	44.7
$1 - C_0$	0.018	0.047	0.029	0.051	0.019	0.047	0.021	0.051

Table 3. RMSE for different velocity fields computed using cross-validation

Velocity	Null	Frozen	$\pi_c = 0$	$\pi_c = \frac{1}{3}$
RMSE (m)	0.5250	0.5018	0.4810	0.4801

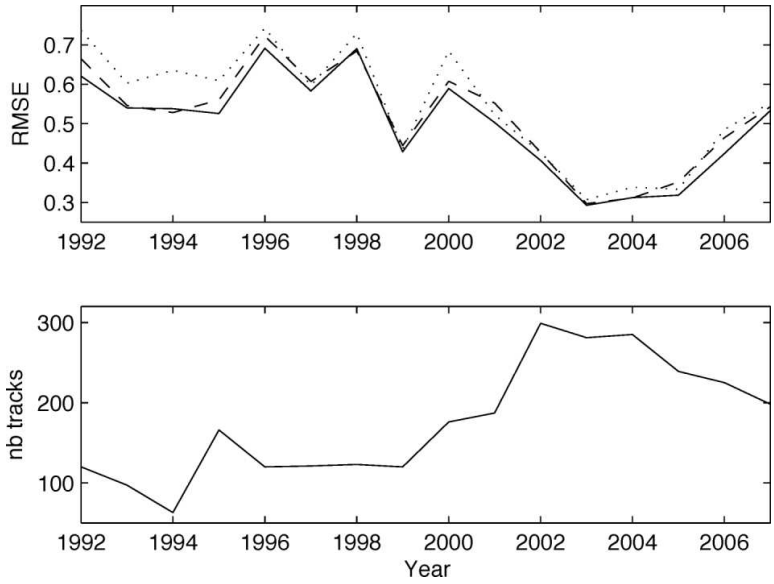


Figure 11. Top panel: time evolution of the root mean square error computed using cross-validation on the period 1992–2007. Solid line: changing velocity ($\pi_c = 1/3$), dashed line: frozen velocity, and dotted line: no velocity. Bottom panel: evolution of the number of satellite tracks which cross the domain D

Figure 11 gives the RMSE for each year: it is obviously correlated to the amount of available satellite data and the RMSE decreases when the number of operational satellites increases (see Figure 1). The inter-annual variability though, could also be due to other factors like meteorological conditions (years with stormy conditions and high significant wave height are more difficult to forecast) and the geometry of the tracks (for example, new satellites are calibrated by working simultaneously with another satellite, and this may reduce the benefit of introducing the motion in the model).

4.2. Virtual buoy

For many offshore applications, it is beneficial to have time series describing significant wave height conditions on a long time period at some specific locations. When there is no *in situ* information available at the location of interest, hindcast data are generally used in operational applications. However these data have well known limitations such as the under-estimation of the severe sea-states.

Using the model proposed in this paper, we can produce significant wave height time series at any location by interpolating the satellite data. Here, we focus on the location (48.7N, 12.4W) for which there is also buoy data for the time period 2002–2007. As the buoy is not inside the domain D_0 considered previously, we have modified the domains D and D_0 as shown in Figure 12. Then, the model has been fitted in the new areas using the method described in the previous sections. Again, we used a time window $\Delta T = \pm 24h$.

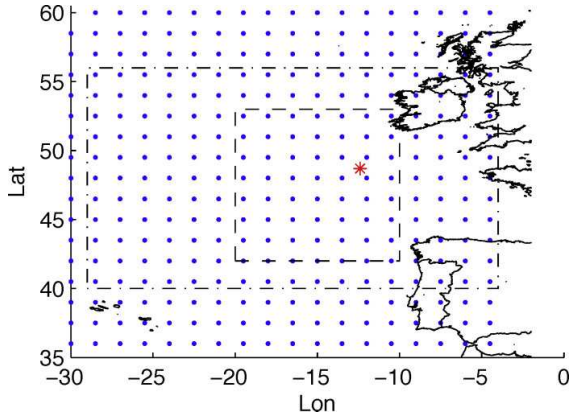


Figure 12. Grid of ERA-Interim data, K1 buoy (*), domains D_0 (dashed box), and D (dash-dotted box). This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

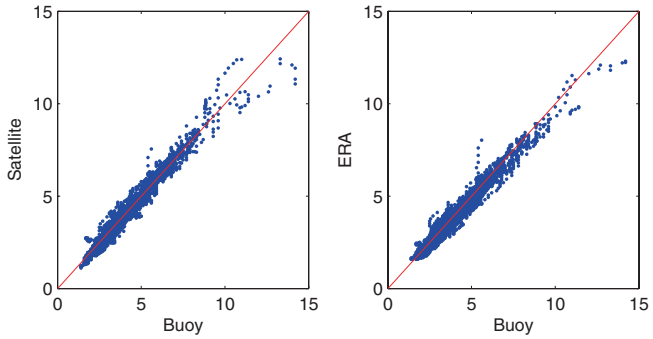


Figure 13. Scatter plot of H_s measured at the buoy (x-axis) against interpolated value (y-axis) obtained from satellite (left) and ERA-Interim data (right). This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

Table 4. RMSE of the time series of the buoy for different velocity fields and data sets

Velocity data	Null Sat.	Constant Sat.	Changing Sat.	Changing ERA
RMSE	0.45	0.36	0.33	0.35

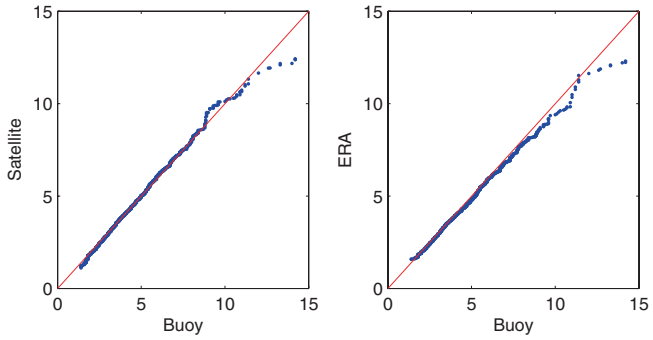


Figure 14. Quantile–quantile plot of H_s measured at the buoy (x-axis) against interpolated value (y-axis) obtained from satellite (left) and ERA-Interim data (right). This figure is available in color online at wileyonlinelibrary.com/journal/environmetrics

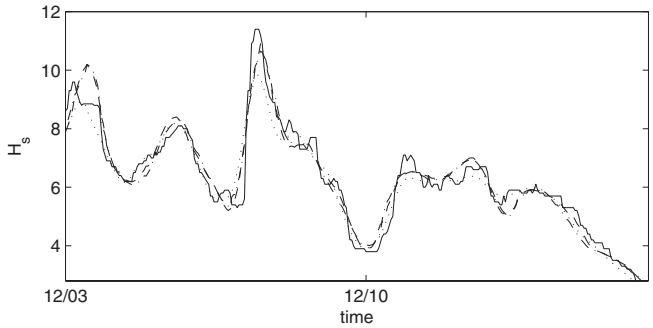


Figure 15. Example of H_s time series measured at K1 buoy (solid line), compared with interpolated using satellite data and changing velocity [resp. constant velocity] (dashed line [resp. dash-dotted line]), interpolated using ERA-Interim data and changing velocity (dotted line)

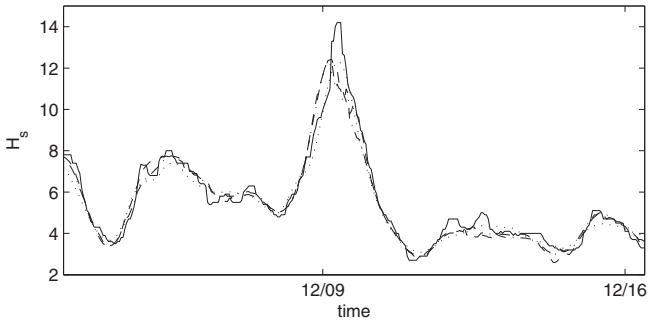


Figure 16. Example of H_s time series measured at K1 buoy (solid line), compared with interpolated using satellite data and changing velocity [resp. constant velocity] (dashed line [resp. dash-dotted line]), interpolated using ERA-Interim data and changing velocity (dotted line)

Table 4 shows that the method that gives a virtual buoy that best matches the *in situ* observations is obtained by interpolating satellite data using the velocity fields computed using particle filter. It slightly improves the results obtained with ERA-Interim data (we use the same interpolation method with changing velocity than for satellite data) and the model with satellite data and the constant velocity. If hindcast data are not available to estimate the motion, this last interpolation method should be favored since it clearly improves the results obtained with a static covariance model.

Figure 13 shows that ERA-Interim data systematically under-estimates high significant wave height values, and this is problematic for extreme value analysis. This is also evident on the quantile–quantile plots in Figure 14. The two examples of time-series shown in Figures 15 and 16 show also that ERA-Interim data tend to smooth the temporal variability observed at the buoy and under-estimate the significant wave height in the storms. But, storms are not always well reconstructed using satellite data neither. For example, the biggest storm observed in buoy data occurred around the 12 September 2007 (see Figure 16) and the time-series obtained using satellite data fail to reproduce both the intensity of the event and the date of arrival of the storm. The main reason is probably the poor sampling of this storm by satellites: only one satellite passage crosses the track of the storm and at locations far from the position of the buoy (about 200 km).

5. CONCLUSIONS

The contributions of this paper are mainly for applications. First of all, in order to model the evolution of the significant wave height over space and time, three different sources of data, satellite altimeter, reanalysis and buoy data, have been combined. These sources provide information at different scales.

Since the ocean, and so quantities such as the significant wave height, can be thought of as a random surface which develops over time, a static model for the covariance structure of the field does not suffice. Instead, we propose to study the problem using the Lagrangian reference frame; that is, we include dynamics in the model by considering a static field subordinated by a velocity field. Assuming the velocity is constant over space and time, although improving the results obtained using a static model, is still unrealistic and hence a new velocity field introduced through a flow of diffeomorphisms is considered.

The important problem of motion estimation is then formulated and solved sequentially within a state-space model framework. The motion is described by a hidden Markov chain, including a physical *a priori* law, the dispersion relation, to describe the evolution of the sea-states. The data used are the ERA-Interim data set. On the one hand, the satellite coverage is generally poor for tracking the motion of the sea-states. On the other hand, the sampling regularity of the ERA-Interim data, both in time and space, allows for the use of existing models for processes on regular grids.

The fitted models have been cross-validated against satellite and buoy data. In both cases, the best results are obtained by interpolating satellite data using velocity fields computed by means of the state-space model. If hindcast data are not available for estimating the velocity fields, the model with constant velocity should be favored since it clearly improves the results obtained with the static covariance model. For operational applications, interpolating satellite data using the dynamic covariance model proposed in this paper should lead to a better description of extreme events compared to hindcast data.

ACKNOWLEDGMENTS

The authors thank the Gothenburg Mathematical Modelling Center for financial support.

REFERENCES

Ailliot P, Monbet V, Prevosto M. 2006. An autoregressive model with time-varying coefficients for wind fields. *Environmetrics* **17**(2): 107–117.
Athanassoulis GA, Stephanakos CN. 1995. A nonstationary stochastic model for long-term time series of significant wave height. *Journal of Geophysical Research* **100**(C8): 16149–16162.

Baxevani A, Borgel C, Rychlik I. 2008. Spatial models for the variability of the significant wave height on the world oceans. *International Journal of Offshore and Polar Engineering* **18**(1): 1–7.
Baxevani A, Caires S, Rychlik I. 2009. Spatio-temporal statistical modelling of significant wave height. *Environmetrics* **20**: 14–31.
Baxevani A, Rychlik I, Wilson RJ. 2005. A new method for modelling the space variablilty of significant wave height. *Extremes* **8**(4): 267–294.
Berliner LM, Wikle CK. 2007. Approximate importance sampling monte carlo for data assimilation. *Physica D* **230**: 37–49.
Caires S, Sterl A. 2005. A new non-parametric method to correct model data: application to significant wave height from the era-40 reanalysis. *Journal of Atmospheric and Oceanic Technology* **22**(4): 443–459.
Caires S, Sterl A, Bidlot JR, Graham N, Swail V. 2004. Intercomparison of different wind wave reanalyses. *Journal of Climate* **17**(10): 1893–1913.
Cappé O, Moulines E, Rydén T. 2005. *Inference in Hidden Markov Models*. Springer-Verlag: New York.
Christakos G. 2000. *Modern Spatiotemporal Geostatistics*. Oxford University Press: New York.
Christakos G, Hristopoulos DT. 1998. *Spatiotemporal Environmental Health Modeling: A Tractatus Stochastic*. Kluwer Academic Publishers: USA.
Corpetti T, Mémin E, Pérez P. 2002. Dense estimation of fluid flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(3): 365–380.
Cotton PD, Challenor PG, Redbourn-Marsh L, Gulev S, Sterl A, Bprtkovskii RS. 2001. An intercomparison of voluntary observing satellite data and modelling wave climatologies. In *Satellite Microwave Remote Sensing*, Swail VR (ed.). WMO: Geneva, Switzerland; 451–460.
Cressie N. 1993. *Statistics for Spatial Data*, Rev Sub edition. Wiley-Interscience.
Cressie N, Huang HC. 1999. Classes of nonseparable, spatiotemporal stationary covariance functions. *Journal of the American Statistical Association* **94**: 1330–1340.
Cuzol A, Mémin E. 2009. A stochastic filtering technique for fluid flow velocity fields tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **31**(7): 1278–1293.
Gneiting T, Genton MG, Guttorp P. 2007. Geostatistical space-time models, stationarity, separability and full symmetry. In Finkenstadt, B., Held, L. and Isham, V. (eds.), *Statistical Methods for Spatio-Temporal Systems*. Chapman Hall/CRC Boca Raton, 151–175.
Janssen PAEM, Doyle JD, Bidlot J, Hansen B, Isaksen L, Viterbo P. 2002. *Impact and feedback of ocean waves on the atmosphere*. In *Advances in Fluids Mechanics*. Atmosphere-ocean Interactions. Vol. 1. W. Perrie (ED.); WIT press; **1**: 155–197.
Joshi SC, Miller MI. 2000. Landmark matching via large deformation diffeomorphisms. *IEEE Transactions on Image Processing* **9**: 1357–1370.
Komen GJ, Cavaleri L, Donelan M, Hasselmann K, Hasselmann S, Janssen PAEM. 1994. *Dynamics and Modelling of Ocean Waves*. Cambridge University Press.
Korotaev GK, Huot E, Le Dimet FX, Herlin I, Stanichny SV, Solovyev DM, Wu L. 2008. Retrieving ocean surface current by 4-d variational assimilation of sea surface temperature images. *Remote Sensing of Environment* **112**: 1464–1475.
Krogstad H, Barstow S. 1999. Directional distributions in ocean wave spectra. In *Proceedings of 9th International Offshore and Polar Engineers Conference, ISOPE*, Vol. III; 76–89.
Le Dimet FX, Talagrand O. 1986. Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A* **38**: 97–110.

Ma C. 2003. Nonstationary covariance functions that model space-time interactions. *Statistics and Probability Letter* **61**(4): 411–419.
Marcello J, Eugenio F, Marqués F, Hernàndez-Guerra A, Gasull A. 2008. Motion estimation techniques to automatically track oceanographic thermal structures in multisensor image sequences. *IEEE transactions on Geosciences and Remote sensing* **46**(9): 2743–2762.
Markussen B. 2007. Large deformation diffeomorphisms with application to optic flow. *Computer Vision and Image Understanding* **106**(1): 97–105.
McGinnity S, Irwin GW. 2001. Maneuvering target tracking using a multiple-model bootstrap filter. *Sequential Monte Carlo Methods in Practice*, Doucet A, de Fretitas N, Gordon N (eds). Springer: USA, 247–271.
Papadakis N, Corpetti T, Mémin E. 2007. Dynamically consistent optical flow estimation. In *Proceedings of the International Conference on Computer Vision (ICCV'07)*, Rio de Janeiro, Brazil.
Queffeulou P. 2004. Long term validation of wave height measurements from altimeters. *Marine Geodesy* **27**: 495–510.
Schmetz J, Holmlund K, Hoffman J, Strauss B, Mason B, Gaertner V, Koch A, Van De Berg L. 1993. Operational cloud-motion winds from Meteosat infrared images. *Journal of Applied Meteorology* **32**: 1206–1225.
Whitham GB. 1974. *Linear and Nonlinear Waves*. John Wiley Sons: New York.

3 Conclusion du chapitre

Nous reprendrons ici principalement les commentaires apportés dans la conclusion de l'article, article qui a permis de combiner l'information contenue dans deux sources de données différentes, les satellites et les données de réanalyse ERA-Interim, afin de reconstituer aussi fidèlement que possible les données de bouées, qui sont, comme nous l'avons vu dans le chapitre précédent, une référence en ce qui concerne la mesure de la hauteur significative des vagues.

La surface de l'océan, et donc en particulier la hauteur significative, peut être considérée comme une surface aléatoire, qui évolue au cours du temps, raison pour laquelle la description de cette surface par la seule covariance spatiale est insuffisante. Les outils mis en place dans cet article permettent d'intégrer une dynamique à l'aide d'un référentiel Lagrangien, dynamique qui peut être soit statique, soit dynamique. L'estimation dans cette dynamique a été réalisée dans le cadre des modèles à espace d'état, à l'aide d'un filtrage particulaire et d'un a priori issu de la physique, en se basant sur les données ERA-Interim.

Cette approche est intéressante, car elle permet d'améliorer la modélisation de la hauteur significative des vagues en un point du globe où il n'y a pas de bouée, et ce même si les données ERA-Interim ne sont pas disponibles, à l'aide du modèle avec une vitesse constante, qui constitue déjà une amélioration du simple krigeage des données satellitaires.

Chapitre IV

Modélisation des dépassements de seuil

1 Introduction

Nous avons expliqué dans le chapitre 2 la nécessité de développer de nouveaux modèles pour la modélisation de processus observés sur une grille irrégulière. Ce point a été illustré par les séries temporelles dont nous disposons, mais il est évident que de telles remarques s'appliquent encore plus facilement dans le cas de processus spatiaux, pour lesquels un échantillonnage régulier fait rarement sens. Ce contexte justifie également que l'on ne retienne pas l'approche markovienne sur les dépassements de seuils successifs, car une telle structure est généralement inadaptée au contexte spatial. L'objectif de ce chapitre est donc de proposer une méthode alternative pour les dépassements de seuils, basée sur les procédures existantes, mais permettant d'introduire une structure de dépendance plus souple qui doit être à même de modéliser des processus observés à pas de temps irréguliers. Cette partie est constituée d'un article soumis à la revue *Annals Of Applied Statistics* le 16/09/2011, et contient la description d'une méthode alternative à celles existantes. Nous avons établi une preuve de la convergence des estimateurs correspondants et effectué une analyse du comportement de ces estimateurs sur des données simulées afin de vérifier le comportement à distance finie. Une partie importante et originale, est de proposer une vérification du comportement du modèle proposé pour la description du comportement extrême de séries temporelles classiques, selon des critères nouveaux, qui ont déjà été présentés dans le chapitre 3. Le dernière partie de cet article concerne une application à des données réelles, décrites auparavant. Cette partie étant succincte, nous détaillerons les applications aux diverses données en notre possession dans le dernier chapitre de ce document.

2 Article

Modelling extreme values of processes observed at irregular time step. Application to significant wave height.

Nicolas Raillard[1,2,3], Pierre Ailliot[1], Jian-feng Yao[4]

Septembre 16, 2011

[1] Laboratoire de Mathématiques, Université de Bretagne Occidentale, Brest, France.
[2] Laboratoire d'Océanographie Spatiale, IFREMER, France.
[3] Laboratoire de Mathématiques de Rennes, Université de Rennes 1, Rennes, France.
[4] Department of Statistics & Actuarial Sciences, The University of Hong Kong, Pokfulam, Hong-Kong.

Abstract

The distribution of extremes such as flood peaks, maximum wave height or minimum daily returns over annual or other time intervals is of common interest to many disciplines including the natural and social sciences. This work is motivated by the analysis of extreme values from times series of significant wave heights observed in North Atlantic. One of these time series exhibits missing data (buoy data) and another one irregular time sampling (satellite data). This situation is frequent when considering environmental data sets and new statistical methods are needed to analyze the extremal behavior of such time series. The method proposed in this work consists in assuming that the behavior of the process above a high threshold is well approximated by a max-stable process which parameters are estimated by maximizing a composite likelihood function. The consistency of these estimates is established. Then, using an extensive set of simulated time series, we assess the finite-sample behavior of our estimates for small to medium sample sizes and compare them to other available estimation methods proposed in the literature for analyzing the extremal behavior of stochastic processes on the basis of standard validation statistics. Finally, a detailed study of significant wave height data is performed. It is shown that the proposed methodology may be used to estimate characteristics of extreme significant wave height at any location in the ocean from altimeter satellite data.

Keywords: Extreme values, time series, max-stable process, composite likelihood, consistency, irregular time sampling, significant wave height, altimeter

1 Introduction

Extreme events have become a major concern in risk management or engineering and appropriate statistical methods are needed to derive estimations of the extremal properties of various phenomena from complex data sets. For example, the design of marine structure depends on the extreme waves that they may face and mainly three sources of data can be used to estimate the extreme quantiles of the significant wave height (Hs) distribution :

- *Reanalysis data* which provide long time series (typically a few decades) at regular time step and without missing values but which tend to smooth out extreme values.

- *Buoy data* which generally give more accurate observations but on shorter time period (typically a few years with missing values) and have a poor spatial distribution.
- *Satellite data* which also provide accurate observations of the wave height over the last 20 years. However, the time series obtained by selecting all the satellite data available at a given location exhibits a complex irregular time sampling depending on the number and the tracks of the operating satellites.

The motivation of this work is to develop statistical methods for analyzing the extremal properties of Hs based on such data sets. In particular, the method that we propose can be used for estimating various characteristics of the extremal behavior of processes (high quantiles, return period, storm durations,...) observed at regular or irregular time steps whereas the other existing methods are not appropriate in the last case.

Two methods are commonly used in the literature for the statistical analysis of extreme events (see e.g. [9], [4], [3], [8], and references therein). The first one, generally referred as the block maxima method, relies on probabilistic results which suggest using the generalized extreme-value (GEV) distribution for modeling the maximum of a large number of identically distributed random variables. The main drawback of this approach is the waste of data induced by taking the maximum over a large block, typically one year for meteorological applications, before fitting the GEV distribution. Hence another approach, generally referred as Peaks Over Threshold (POT), consists in keeping all the observations above a threshold which is chosen high enough in order to ensure that the distribution of the excesses above this threshold is well approximated by a generalized Pareto distribution (GPD). A problem when using POT approach for time series of dependent data is that clusters of consecutive dependent exceedances are generally observed, especially when the time-lag between successive observations is smaller than the characteristic duration of extreme events and thus some changes are needed in practice when using POT method in this context. Usually, the first step consists in identifying clusters ("declustering" step) before fitting a GPD distribution to the sample of clusters maxima. This methodology also leads to waste data since only maxima within each cluster are used to fit the GPD and relies on arbitrary rules for declustering the data which are even more difficult to chose in presence of missing values or irregular time sampling. Hence another approach, initially proposed in [20], consists in keeping all exceedances and model the dependence structure between neighboring excesses as a first order Markov chain which transition kernel is derived from bivariate extreme value theory. This method has been applied successfully to various meteorological time series (see e.g. [16]).

In Section 2 of this paper we propose an alternative approach for modeling the dependence structure above a high threshold in which the time series of exceedances is assumed to be a realization of a censored max-stable process. Although the methodology is not restricted to a specific model of max-stable processes, we focus on the continuous-time model proposed in [19] which has a simple meteorological interpretation and can easily handle observations available at irregular time steps.

Parameter estimation is discussed in Section 3. Since the likelihood function is not tractable, we propose to estimate the unknown parameters by maximizing a composite likelihood function and we prove theoretical results which indicate that these estimates are consistent. We also discuss the properties of the estimates for small to medium size samples, comparable to those typically available in usual applications, using simulations.

In Section 4, we validate the methodology on time series simulated using different classical time series models. The fitted censored max-stable process can be easily simulated and this allows to estimate various quantities of interest for the applications, such as quantiles, return periods or characteristics of the sojourns above high threshold using Monte-Carlo simulations.

The results are compared to the ones obtained using the most classical approaches mentioned above.

One advantage of our approach is that it can easily handle time series with missing values or irregular sampling. This is illustrated in Section 5 on Hs data in North Atlantic where the methodology is used to compare the extremal properties of reanalysis, buoy and satellite data.

Conclusions and key findings are given in Section 7.

2 Censored max-stable process

In this section we introduce an original model for describing the extremal behavior of a time series of dependent observations. In order to motivate this model, we focus in Section 2.1 on the simpler case of independent and identically distributed (iid) random variables and discuss the interpretation of the POT method in terms of censoring. The generalization to time sequences of dependent observations is discussed in Section 2.2.

2.1 Threshold models and censoring in the independent case

Probably the most classical approach for modelling the extremal properties of an iid sample $(X_1, ..., X_n)$ consists in using the "block maxima" approach. It relies on probabilistic results originating in [10] which suggest approximating the distribution of $M_n = \max_{i=1,...,n} X_i$ by a GEV distribution. The GEV distribution has the following cumulative distribution function (cdf)

$$F(x;\mu,\sigma,\xi) = \begin{cases} \exp\{-[1+\xi\frac{x-\mu}{\sigma}]^{-\frac{1}{\xi}}\} & (\xi \neq 0) \\ \exp\{-\exp[-\frac{x-\mu}{\sigma}]\} & (\xi = 0) \end{cases} \quad (1)$$

defined for x such that $1+\xi\frac{x-\mu}{\sigma} > 0$ with parameters $\mu \in \mathbb{R}$, $\sigma > 0$ and $\xi \in \mathbb{R}$. For practical applications, the data are blocked into blocks of equal length and a GEV distribution is fitted to the sample obtained by keeping the maxima within in each block. The choice of the size of the blocks is critical in practice. As concerns environmental time series, the GEV distribution is generally fitted to the time series of annual maxima in order to remove seasonal effects. This leads to an important waste of data and the three parameters of the GEV distribution have to be estimated based on small samples which size corresponds to the number of years of observation with no or few missing values (a few decades in the best cases). Although many methods have been proposed in the literature to provide estimates which have a good behavior on small samples (see [2] and references therein), it remains an important issue for practical applications.

The POT approach is the classical alternative to the block maxima approach. It is less wasteful of data since it keeps all the data above a high threshold u which is chosen such that the conditional distribution $\mathbb{P}[X_i \leq x | X_i > u]$ is well approximated by a GPD with cdf

$$G(x;\mu,\sigma,\xi) = \begin{cases} 1-\left(1+\xi\frac{x-\mu}{\sigma}\right)^{-1/\xi} & (\xi \neq 0) \\ 1-\exp[-\frac{x-\mu}{\sigma}] & (\xi = 0) \end{cases}$$

defined for $x \geq \mu$ such that $\left(1+\xi\frac{x-\mu}{\sigma}\right) \geq 0$ with $\mu = u$ and parameters $\sigma > 0$ and $\xi \in \mathbb{R}$. Again, the use of the GPD is motivated by probabilistic results and various methods have been proposed in the literature for estimating the two parameters of the GPD based on the sample of exceedances as well as to choose u, although the latter is a more difficult problem to handle (see [6]). Once u is chosen, the most usual method for estimating the unknown parameters consists

in maximizing the likelihood function given by

$$\begin{aligned} L(\lambda,\sigma,\xi;X_1,...,X_n) &= \lambda^{N_u}(1-\lambda)^{n-N_u}\prod_{i=1|X_i>u} g(x_i;u,\sigma,\xi) \\ &= \prod_{i=1|X_i\leq u}\lambda\prod_{i=1|X_i>u}(1-\lambda)g(x_i;u,\sigma,\xi) \end{aligned} \quad (2)$$

where $\lambda=\mathbb{P}(X_i\leq u)$, N_u is the number of observations below the threshold u and $g(x;\mu,\sigma,\xi)$ is the probability density function (pdf) of the GPD. It is well known that the conditional distribution of the exceedances of a GPD above an arbitrary threshold is also a GPD and this allows to interpret (2) as the likelihood of an iid sample of a GPD censored at the threshold u. More precisely, let $(\tilde{X}_1,\tilde{X}_2,...,\tilde{X}_n)$ be an iid sample of a GPD with parameter $(\mu,\tilde{\sigma},\xi)$ and consider the censored random variable

$$Y_i=u\mathbb{1}_{[\tilde{X}_i\leq u]}+\tilde{X}_i\mathbb{1}_{[\tilde{X}_i>u]}=\begin{cases} u & \text{if } \tilde{X}_i\leq u \\ \tilde{X}_i & \text{if } \tilde{X}_i>u \end{cases}$$

where the threeshold u belongs to the support of the GPD distribution. We have $P(Y_i=u)=\lambda$ with $\lambda=G(u;\mu,\tilde{\sigma},\xi)$ and, for $x>u$,

$$\begin{aligned} \mathbb{P}(Y_i\geq x) &= \mathbb{P}(\tilde{X}_i\geq u)\mathbb{P}(\tilde{X}_i\geq x|\tilde{X}_i\geq u) \\ &= (1-G(u;\mu,\tilde{\sigma},\xi))\frac{1-G(x;\mu,\tilde{\sigma},\xi)}{1-G(u;\mu,\tilde{\sigma},\xi)} \\ &= (1-\lambda)(1-G(x;u,\sigma,\xi)) \end{aligned}$$

with $\sigma=\tilde{\sigma}\left(1+\xi\frac{u-\mu}{\tilde{\sigma}}\right)$ and thus (2) is the likelihood of $(Y_1,...,Y_n)$. Finally, the assumptions made when using the POT approach are equivalent to assuming that the original sample $(X_1,...,X_n)$ satisfies

$$u\mathbb{1}_{[X_i\leq u]}+X_i\mathbb{1}_{[X_i>u]}=u\mathbb{1}_{[\tilde{X}_i\leq u]}+\tilde{X}_i\mathbb{1}_{[\tilde{X}_i>u]} \quad (3)$$

for all $i\in\{1,...,n\}$ where $\left(\tilde{X}_1,...,\tilde{X}_n\right)$ is an iid sample of a GPD.

We will see below that this interpretation of the POT approach in terms of censoring may have advantages for modeling purpose. From a numerical point of view, it can be viewed as a reparametrization of the likelihood function. Maximizing (2) over (λ,σ,ξ) leads to the following estimate for λ, $\widehat{\lambda}=\frac{N_u}{n}$ and this estimate has the desirable property to be easy to interpret and be independent of the estimates of the two other parameters. On the other hand, maximizing (2) over $(\mu,\tilde{\sigma},\xi)$ leads to a more complicated 3-dimensional optimization problem and estimates which are correlated and this can be problematic for some applications (see [17]).

Although the GPD distribution is the most usual approximation for the tail of the distribution when modeling the exceedances over a high threshold , other approximations have been proposed in the literature. In particular, we have

$$G(x;\mu,\sigma,\xi)\approx F(x;\mu,\sigma,\xi)$$

for "high" values of x and this suggests that similar results will be obtained if we model the distribution of $\tilde{X}_i$ by a GEV distribution instead of a GPD. We performed various tests on simulated samples which confirms that both approximations lead to similar results in practice.

It has been shown that the tail approximations discussed above are still valid for dependent sequences under mild conditions (see [12]) and this justify the use of both the block maxima and POT approaches for analyzing the extremal behavior of a time series. One difficulty when

using POT approach in this context is that clusters of consecutive dependent exceedances are generally observed whereas the likelihood function (2) is the true likelihood of the sample only if the exceedances are independent. The most usual method is then to apply a declustering step and keep only the maxima within each cluster in order to obtain a sample of approximatively independent exceedances. It leads to waste data and thus degrade the quality of the estimates but also to lose information on the dynamics of the process inside the clusters. This may be problematic for applications sensitive to the within-cluster behavior. A method which generalizes the principle of censoring for dependent sequences and permits to keeps all exceedances above the selected threshold is proposed in the next section.

2.2 Censored max-stable process

We now consider a sample $(X_{t_1}, ..., X_{t_n})$ of a stochastic process $\{X_t\}$ observed at time $(t_1, ..., t_n)$. It is generally assumed in the literature that the observations are available at regular time step (i.e. $t_{i+1} - t_i = t_{j+1} - t_j$ for all $(i, j) \in \{1, ..., n-1\}$) but we would like to have a method which is flexible enough to deal with irregular time sampling (see Section 5 for practical motivations). We propose to analyse the extremal behaviour of such data set by extending the POT approach discussed above and model the distribution of the process $\{X_t\}$ over a high threshold u by a censored max-stable process. The theory of max-stable processes (see [7], [8]) is a natural generalization of the traditional univariate max-stable theory which was used above to motivate the choice of the GEV distribution in the iid case. Several families of max-stable process have been proposed in the literature (see e.g. [19, 18]), but hereafter we will focus on the specific *Gaussian extreme value process* introduced in [19] although the methodology is flexible enough to deal with other models.

More precisely, we assume that

$$u\mathbb{1}_{[X_t \leq u]} + X_t \mathbb{1}_{[X_t > u]} = u\mathbb{1}_{[\tilde{X}_t \leq u]} + \tilde{X}_t \mathbb{1}_{[\tilde{X}_t > u]} \tag{4}$$

holds for all t where u is a fixed threshold and $\{\tilde{X}_t\}$ is a stationary Gaussian extreme value process with parameter $\theta = (\mu, \sigma, \xi, \nu) \in (-\infty, +\infty) \times (0, +\infty) \times (-\infty, +\infty) \times (0, +\infty)$ defined below.

- The marginal distribution of $\{\tilde{X}_t\}$ is a GEV distribution with parameter (μ, σ, ξ). With this assumption, the process $\{Z_t\}$ obtained by applying the following marginal transformation

$$Z_t = -\frac{1}{\log(F(\tilde{X}_t; \mu, \sigma, \xi))} \tag{5}$$

  is a stationary process with unit Fréchet (i.e. GEV distribution with parameter $(1, 1, 1)$) marginal distribution

- We further assume, following [19], that

$$Z_t = \max\left\{\frac{\zeta_i}{\nu\sqrt{2\pi}} \exp\left(-\frac{(s_i - t)^2}{2\nu^2}\right)\right\}$$

  where $\{(\zeta_i, s_i), i \geq 1\}$ denote the points of a Poisson process on $(0, \infty) \times \mathbb{R}$ with intensity measure $\zeta^{-2} d\zeta \times ds$.

We focus on Gaussian extreme value processes because they have a nice meteorological interpretation (see [19]) and provide a flexible class of models. The parameters (μ, σ, ξ) are related

to the marginal distribution and can be interpreted respectively as location, scale and shape parameters, whereas the parameter ν is related to the temporal structure of the process and may be interpreted as the typical duration of the storms. More precisely, we have (see [19]):

$$\mathbb{P}(Z_{t_1} \leq z_{t_1}, Z_{t_2} \leq z_{t_2}) = F_Z(z_{t_1}, z_{t_2}; \nu) = \exp\left[-V(z_{t_1}, z_{t_2}; \nu)\right], \tag{6}$$

where

$$V(z_{t_1}, z_{t_2}; \nu) = \frac{1}{z_{t_1}}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_2}}{z_{t_1}}\right) + \frac{1}{z_{t_2}}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_1}}{z_{t_2}}\right) \tag{7}$$

with $a = \frac{|t_1 - t_2|}{\nu}$ and Φ the cdf of the standard normal distribution. The limit cases $\nu \to 0$ and $\nu \to +\infty$ corresponds respectively to independence and perfect dependence.

Applying the inverse marginal transformation leads to the following bivariate cdf for the Gaussian extreme value process $\{\tilde{X}_t\}$:

$$\begin{aligned} F_{\tilde{X}}(\tilde{x}_{t_1}, \tilde{x}_{t_2}; \theta) &= \mathbb{P}(\tilde{X}_{t_1} \leq \tilde{x}_{t_1}, \tilde{X}_{t_1} \leq \tilde{x}_{t_2}) \\ &= \exp\left[-\frac{1}{z_{t_1}}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_2}}{z_{t_1}}\right) - \frac{1}{z_{t_2}}\Phi\left(\frac{a}{2} + \frac{1}{a}\log\frac{z_{t_1}}{z_{t_2}}\right)\right] \end{aligned} \tag{8}$$

with $a = \frac{|t_1 - t_2|}{\nu}$ and $z_{t_i} = \frac{-1}{\log F(\tilde{x}_{t_i}; \mu, \sigma, \xi)}$.

The distribution of (Y_{t_1}, Y_{t_2}), where $Y_t = u\mathbb{1}_{[\tilde{X}_t \leq u]} + \tilde{X}_t\mathbb{1}_{[\tilde{X}_t > u]}$ is the censored Gaussian extreme value process, has the following bivariate pdf

$$p_Y(y_{t_1}, y_{t_2}; \theta) = \begin{cases} F_{\tilde{X}}(u, u; \theta) & \text{if } y_{t_1} = u \text{ and } y_{t_2} = u, \\ \frac{\partial F_{\tilde{X}}}{\partial \tilde{x}_{t_1}}(y_{t_1}, u; \theta) & \text{if } y_{t_1} > u \text{ and } y_{t_2} = u, \\ \frac{\partial F_{\tilde{X}}}{\partial \tilde{x}_{t_2}}(u, y_{t_2}; \theta) & \text{if } y_{t_1} = u \text{ and } y_{t_2} > u, \\ \frac{\partial^2 F_{\tilde{X}}}{\partial \tilde{x}_{t_1} \partial \tilde{x}_{t_2}}(y_{t_1}, y_{t_2}; \theta) & \text{if } y_{t_1} > u \text{ and } y_{t_2} > u, \end{cases} \tag{9}$$

with respect to the product measure $m \times m$, where $m(dx) = \delta_u(dx) + dx$ is the measure obtained by mixing the Dirac measure at u with the Lebesgue measure.

Similar approximations, motivated using probabilistic results from the bivariate extreme value theory, have already been proposed in the literature for modeling the bivariate distribution of neighboring exceedances (see [20], [16] and references therein). In these papers, it is further assumed that the censored process is a Markov chain observed at regular time step. With these assumptions the likelihood function can be derived from the bivariate distribution of successive observations and optimized to compute the maximum likelihood estimates. In our case, the likelihood function is not tractable and an alternative estimation strategy is needed. This is discussed in the next section.

3 Parameter estimation

In this section, we address the problem of estimating the parameters of the censored Gaussian extreme value process introduced in the previous section. In Section 3.1, we introduce the composite likelihood function which will be maximized to define the maximum composite likelihood estimates. Then, in Section 3.2, we prove various results related to the asymptotic properties of these estimates before illustrating these results using simulations in Section 3.3.

3.1 Composite likelihood functions

In this section $(y_{t_1}, ..., y_{t_n}) \in (u, +\infty)^n$ denotes a realization of a Gaussian extreme value process $\{Y_t\}$ with unknown parameter $\theta^* = (\mu^*, \sigma^*, \xi^*, \nu^*)$ censored at the threshold $u \geq 0$ and observed at times $(t_1, ..., t_n)$. There is no known tractable expression for the joint distribution of such sample and, as a consequence, the maximum likelihood estimates can not be computed. We have seen however in the previous section that the marginal and bivariate distributions have tractable expressions and this suggests replacing the likelihood by the composite likelihood functions introduced below (see e.g. [13, 23, 22, 5]).

- The *independent likelihood* function defined as

$$IL(\theta; y_{t_1}, ..., y_{t_n}) = \prod_{i=1}^{n} p_Y(y_{t_i}; \theta). \tag{10}$$

  where $p_Y(y_t; \theta)$ is the pdf of the marginal distribution of Y_{t_i}, with respect to the measure m. It is given by

$$p_Y(y_t; \theta) = \begin{cases} F(u; \mu, \sigma, \xi) & \text{if } y_t = u \\ f(y_t; \mu, \sigma, \xi) & \text{if } y_t > u \end{cases}$$

  where F and f denotes respectively the cdf and pdf of the GEV distribution. It corresponds to the likelihood function of an iid sample of a censored GEV distribution (see Section 2.1) and does not depend on the parameter ν which describes the dependence structure of the process. We denote MILE the estimates of (μ, σ, ξ) obtained by maximising this function.

- The *pairwise likelihood* function

$$PL(\theta; y_{t_1}, ..., y_{t_n}) = \prod_{i=1}^{n-1} \prod_{j>i} p_Y(y_{t_i}, y_{t_j}; \theta)^{\omega_{t_i,t_j}}. \tag{11}$$

  with $p_Y(y_{t_i}, y_{t_j}; \theta)$ given by (9) and $\omega_{t_i,t_j} \in \{0, 1\}$ indicates if the pair of observation (y_{t_i}, y_{t_j}) contributes to the pairwise likelihood function. This approach has already been considered for time series with regular time sampling (see [22] and references therein). It is generally assumed that

$$\omega_{t_i,t_j} = \mathbb{1}_{[|i-j| \leq K]} \tag{12}$$

  such that only the pairs of observations which are less than K time unit apart are kept to build the pairwise likelihood function. Hereafter we will denote PL_K the corresponding pairwise likelihood function estimates and MPL_KE the estimates obtained by maximizing this function. Keeping only the neighboring observations (i.e. using $K = 1$) has clear computational benefits, since it permits to significantly reduce the number of terms in the product (11), and may also lead to more efficient estimates in practice (see [22] and Section 3.3.). Another strategy would consist in keeping the pairs of observations separated by a time lag smaller than T and take

$$\omega_{t_i,t_j} = \mathbb{1}_{[|t_i - t_j| \leq T]} \tag{13}$$

  This second strategy is similar to the first one when the process is observed at regular time sampling but not in the irregular case. This will be further discussed using simulations in Section 3.3 .

- The *Markovian likelihood* function is defined as:

$$\begin{aligned} ML(\theta; y_{t_1}, ..., y_{t_n}) &= p_Y(y_{t_1}; \theta) \prod_{i=2}^{n} p_Y(y_{t_i} | y_{t_{i-1}}; \theta) \\ &= \frac{\prod_{i=1}^{n-1} p_Y(y_{t_i}, y_{t_{i-1}}; \theta)}{\prod_{i=2}^{n-1} p_Y(y_{t_i}; \theta)} \\ &= \frac{\prod_{i=1}^{n-1} p_Y(y_{t_i}, y_{t_{i-1}}; \theta)}{\prod_{i=2}^{n-1} p_Y(y_{t_i}; \theta)} \\ &= \frac{PL_1(\theta; y_{t_1}, ..., y_{t_n})}{IL(\theta; y_{t_2}, ..., y_{t_n})} \end{aligned} \tag{14}$$

  We denote MMLE the values of θ maximizing this function. When the process is observed at regular time step, we retrieve the Markovian model considered in [20, 16] for the specific bivariate max-stable distribution associated to the Gaussian extreme value process.

From a numerical point of view, we found useful to use a two-stage procedure, where the parameters (μ, σ, ξ) of the marginal distribution are first estimated by maximizing the independent likelihood function, before estimating the dependence parameter ν by maximizing the pairwise likelihood function over ν with the parameters of the marginal distribution being fixed to the values obtained in the first step. It permits to reduce computational time (the independent likelihood function can be evaluated quickly compared to the pairwise likelihood function) and avoid divergence problems which may occur when optimizing the pairwise likelihood function simultaneously over all the parameters with an inappropriate starting point. Eventually, a global optimization of the pairwise likelihood function may be performed to refine the estimates obtained with the two-stage procedure.

3.2 Consistency of MPL$_1$E with known marginal distribution

In this section, we prove results related to the consistency of the MPL$_1$E introduced in the previous section in an idealized situation. More precisely, we assume that there is no censoring ($u = -\infty$), that the process is observed at regular time step and that the marginal parameters (μ^*, σ^*, ξ^*) are known. This latter assumption is not realistic in practice but mimics the second step of the two-stage procedure introduced in the previous section and permits to avoid supplementary difficulties which appear when the support of the distribution depends on the parameters. With these assumptions, we can apply the marginal transformation (5) and assume, without loss of generality, that the process has a unit Fréchet marginal distribution. Finally, we denote $(Z_1, ..., Z_n)$ a Gaussian extreme value process observed at time $t_i = i$ for $i \in \{1, ...n\}$ with parameter $(1, 1, 1, \nu^*)$ and $\hat{\nu}$ the estimate of $\nu^* \in (0, +\infty)$ obtained my maximizing the PL_1 function over ν.

The proof of the following theorem is postponed to Section 6.

Theorem 3.1. *Assume that parameter space is* $\Theta = [\nu_-,\ \nu_+]$ *with* $0 < \nu_- < \nu_+$ *for* ν. *Then* $\hat{\nu}$ *is a consistent estimate of* ν^*.

3.3 Simulation Study

A simulation study was undertaken to assess the accuracy of the estimates introduced in Section 3.1 for practical applications. Random samples were generated from a Gaussian extreme value process with parameters $\mu = 0$, $\sigma = 1$ and $\xi = 0.3$ which corresponds to realistic values for

environmental applications. We performed various experiments to study the impact of the sample size n (see Figure 1), of the dependence parameter ν (see Figure 2), of the threshold u (see Figure 3) and finally of the strategy considered to select the pair of observations which contribute to the pairwise likelihood function (see Figure 4) on the accuracy of the estimates. For each experiment, 200 independent random samples were generated from a Gaussian extreme value process and the various estimates of (μ, σ, ξ, ν) were determined to derive an empirical root-mean-squared error (RMSE). The MILE only provides estimates of (μ, σ, ξ) but we have also computed the estimate of ν considered in Section 3.2 which is obtained by maximizing the PL_1 function over ν with (μ, σ, ξ) fixed to their true values (see the right panels of Figures 1, 2 and 3).

Let us first focus on the case when the time step between successive observations is constant. According to Figure 1 all the estimates seem to be convergent when the sample size increases and we checked empirically that, when multiplied by $\sqrt{n}$, the errors are almost constant and thus that we retrieve the usual speed of convergence. The MILE are clearly the less efficient estimates whereas MPL_1E, MPL_5E and MMLE give similar results. A closer look reveals that MPL_5E is the less accurate of these three estimates. MMLE slightly outperforms MPL_1E in estimating the scale parameter σ and the shape parameter ξ whereas MPL_1E provides the best estimate of the dependence parameter ν. Figure 2 shows that the RMSE of all the estimates decreases with the dependence parameter ν and that it is more difficult to get reliable estimates when there is a stronger dependence between successive observations. The efficiency of MMLE generally deteriorates quicker when ν increases, and it provides the worst estimates of μ, σ and ν when $\nu \geq 1$ but provides the best estimate of ξ for all the values of ν considered in this experiment. Those results indicate that the Markovian approximation may not be appropriate when the dependence in strong. Figure 3 depicts the behavior of the different estimators when censoring occurs. As expected, all of them worsen when the threshold increases and the number of observations decreases. We can notice that the estimator which most suffers from censoring is the MILE, that MMLE outperforms both MPL_1E and MPL_5E in estimating the parameters of the marginal distribution, but MPL_1E again provides the best estimate of the dependence parameter ν.

Figure 4 shows the influence of the windows considered when defining the neighborhood which are taken into account in the pairwise likelihood functions. The Gaussian extreme value process was simulated using an irregular time sampling (the time lags between successive observations were drawn from a uniform distribution on $[0, 2]$) in order to allow a comparison between the two strategies discussed in Section 3.1: the first one consist in using the K closest observations (see Equation (12)) whereas the second one consists in taking into account all the observations falling within less that K time steps (see Equation (13)). The first strategy is found to always be the best. The evolution of RMSE with K differs according to the strategy, since it is increasing for the first strategy, meaning that the best results will be obtained with $K = 1$, but is generally decreasing for the second strategy and the difference between both strategies decreases when K increases. The comparison with MILE and MMLE indicates again that MMLE slightly outperforms MPL_1E for estimating μ, σ and ξ but MPL_1E provides the best estimate of ν.

The results given in this section suggest using MMLE or MPL_1E for practical applications since their RMSE is generally lower than the ones of the other estimates. Both estimates leads to similar computational cost, the first one may provide slightly better estimates when the dependence between successive observations is small, but is clearly less efficient if the dependence is strong. Finally, it seems reasonable to use the MPL_1E for practical applications and we will mainly focus on this estimate in the next sections.

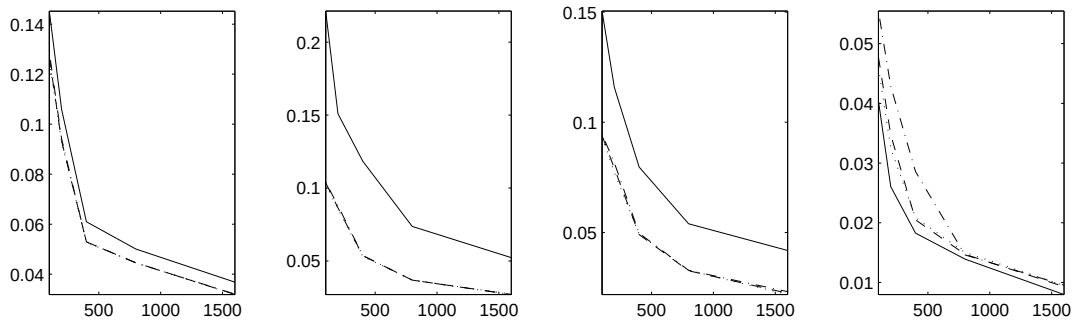


Figure 1: RMSE of the different estimates (y-axis) for various sample size n (x-axis). Results obtained using 200 simulations of a Gaussian extreme value process with parameters values $\mu = 0$, $\sigma = 1$, $\xi = 0.3$ and $\nu = 0.5$, no censoring ($u = -\infty$) and regular time sampling. From left to right panels: estimates of μ; estimates of σ; estimates of ξ; estimates of ν. Solid line: MILE ; dotted line: MMLE ; dashed line: MPL_1E ; dashed-dotted line: MPL_5E.

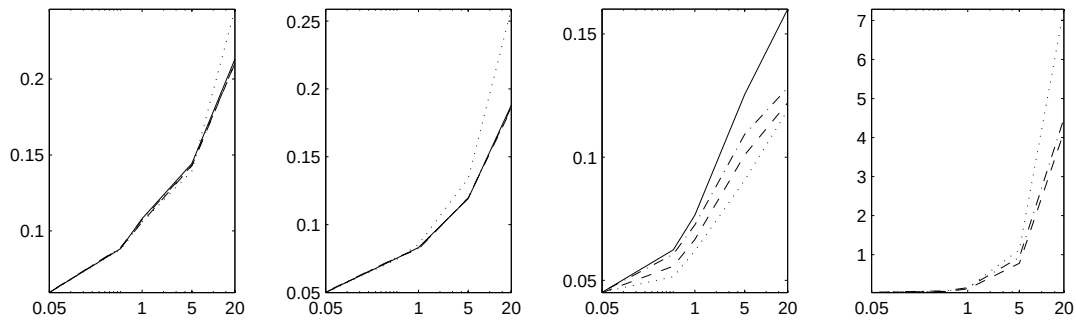


Figure 2: RMSE of the different estimates (y-axis) for various values of ν (x-axis, logscale). Results obtained using 200 simulations of a Gaussian extreme value process with parameters values $\mu = 0$, $\sigma = 1$, $\xi = 0.3$, sample size $n = 300$, no censoring ($u = -\infty$) and regular time sampling. From left to right panels: estimates of μ; estimates of σ; estimates of ξ; estimates of ν. Solid line: MILE ; dotted line: MMLE ; dashed line: MPL_1E ; dashed-dotted line: MPL_5E.

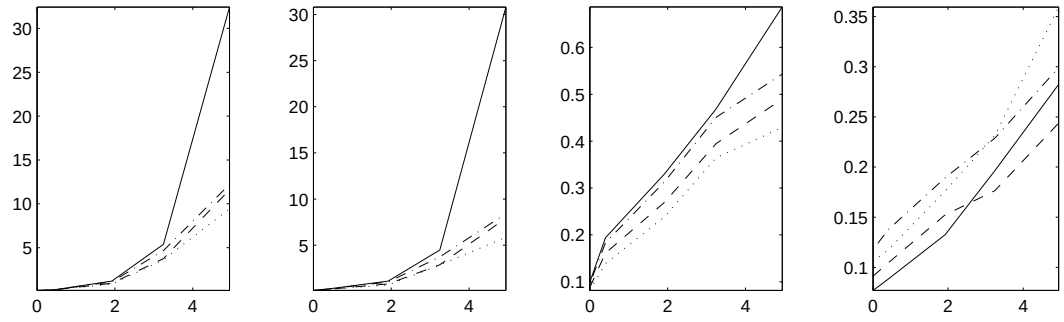


Figure 3: RMSE of the different estimates (y-axis) for various values of the threshold u (x-axis, in terms of quantiles). Results obtained using 200 simulations of a Gaussian extreme value process with parameters values $\mu = 0$, $\sigma = 1$, $\xi = 0.3$, $\nu = 0.5$, sample size $n = 300$ and regular time sampling. From left to right panels: estimates of μ; estimates of σ; estimates of ξ; estimates of ν. Solid line: MILE; dotted line: MMLE ; dashed line: MPL_1E ; dashed-dotted line: MPL_5E.

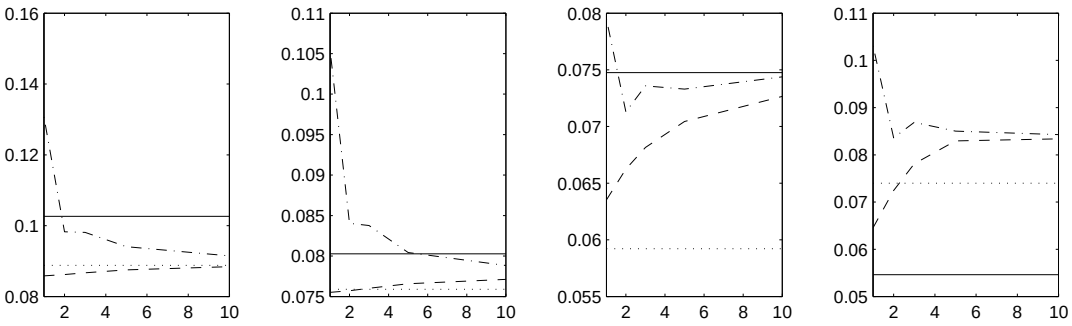


Figure 4: RMSE of the different estimates (y-axis) for various values of K (x-axis). Results obtained using 200 simulations of a Gaussian extreme value process with parameters values $\mu = 0$, $\sigma = 1$, $\xi = 0.3$, $\nu = 0.5$, sample size $n = 300$ and no censoring ($u = -\infty$). The time step between successive observations is drawn from a uniform distribution on the interval $(0, 2)$. From left to right panels: estimates of μ; estimates of σ; estimates of ξ; estimates of ν. Solid line: MILE; dotted line: MMLE ; dashed line: MPL_KE with the first weighting strategy (see (12)) ; dashed-dotted line: MPL_KE with the second weighting strategy (see (13)).

4 Performance on classical time series models

The lack of data makes it generally difficult to validate models for extreme values when facing real data. In this Section we thus perform a simulation study to check if the proposed methodology is able to catch the extremal properties of some widely used time series models. In Section 4.1, we simulate large samples in order to get parameters estimates with low variance and check if the Gaussian extreme value process provides an appropriate approximation of the extremal behavior of the time series models under consideration. Then, in Section 4.2 we simulate samples with smaller sizes in order to validate the whole methodology in a more realistic context for practical applications.

4.1 Model validation

We have chosen to focus on the following time series models in the sequel:

- **IID**: $\{X_t\}$ is an iid sequence of standard normal distribution.
- **AR(1)**: $\{X_t\}$ is a discrete time stationary process which satisfies

  $$X_t = \alpha X_{t-1} + \sqrt{1-\alpha^2}\epsilon_t$$

  for all t, where $\alpha \in (-1, 1)$ describes the dependence between successive observations and $\{\epsilon_t\}$ is an iid sequence of standard normal distribution. The marginal distribution of $\{X_t\}$ is a standard normal distribution and the extremal index (see [4]) is equal to one (no clustering of extremes). We use the value $\alpha = 0.2$ in the sequel.
- **logARMAX(1)**: $\{X_t\}$ is a discrete time stationary process which satisfies $X_t = \log(U_t)$ where $\{U_t\}$ is a usual ARMAX(1) process defined as

  $$U_t = \max\{(1-\alpha)U_{t-1}, \alpha\epsilon_t\}$$

  for all t, where $\alpha \in (0, 1)$ describes the dependence between successive observations and $\{\epsilon_t\}$ is an iid sequence of unit Fréchet distribution. The log transformation is used to avoid numerical problems which occur when estimating quantities related to heavy tail distributions by simulation: the marginal distribution of $\{U_t\}$ is unit Fréchet whereas the one of $\{X_t\}$ is Gumbel. The extremal index is α. We use the value $\alpha = 0.2$ in the sequel.
- **OU** : $\{X_t\}$ is a continuous time stationary process which satisfies

  $$dX_t = -\alpha X_t dt + \sqrt{2\alpha} dW_t$$

  where $\alpha > 0$ describes the dependence structure and $\{W_t\}$ is a standard Brownian motion. The marginal distribution of $\{X_t\}$ is a standard normal distribution and we retrieve the AR(1) model when the process is sampled at regular time step. We use the value $\alpha = 0.05$ in the sequel and the time lags between successive observations are drawn from a uniform distribution on $[0, 2]$.

For each of these models, we first generate a long realization (equivalent to 100 years with one observation per day) and fit a Gaussian extreme value process to the simulated sequence censored at the 95% quantile by computing the $\mathrm{MPL}_1\mathrm{E}$. Then we compare various characteristics of the original model and the fitted censored Gaussian extreme value process model, namely

- mean number of up-crossings during a given time period (1 year) as a function of the threshold

- mean length of the sojourns above a varying threshold
- mean length of the sojourns below a varying threshold

These quantities have been selected because they summarize important properties of the extremal behavior of the processes and may be important for practitioners. All these quantities were computed using a large number of simulations of both the original time series model (IID, AR(1), logARMAX(1) or OU generated using standard algorithms) and the fitted Gaussian extreme value process (see [18] for more details on simulation algorithms). This is illustrated on Figure 5 which shows realizations of both the AR(1) model and the fitted Gaussian extreme value process whereas Figure 7 permits a more systematic comparison of the behavior of both processes. According to this last figure, the fitted model seems to be able to reproduce both the frequency of up-crossings and time between up-crossings, even for high thresholds, but slightly overestimates the mean length of the clusters above high thresholds. Indeed (see middle panel of Figure 7) for the AR(1) model the mean length of the clusters tends to one when the threshold increases as expected from the theory (no clustering of extremes) whereas the fitted Gaussian extreme value process exhibits some small extremal dependence and thus clustering of extremes. Using a higher threshold for censoring before fitting the Gaussian extreme value process may help improving these results and retrieving extremal independence. According to Figure 6, the results are indeed better for the IID model which is a particular case of the AR(1) model with no dependence between successive between observations and the fitted model seems to be able to catch the extremal properties of an iid sequence of a standard normal distribution. It is not surprising to get similar results for the OU (see Figure 9) and AR(1) models since they are equivalent when the observation step is regular.

Contrary to the other models, the extreme values of logARMAX(1) process tend to cluster. According to Figure 8, the fitted model seems to be able to reproduce both the frequency of up-crossings and the mean length of the cluster, which tends to a limit greater than one as expected from the theory, but seems to overestimate the mean length of the sojourn below high thresholds although the erratic behavior of the curves suggests that the observed differences may be due to the sampling error.

These results indicate that the Gaussian extreme value process is able to catch important properties of the extremal behavior of some usual time series model. Similar results have been obtained on the other models that we have tried.

4.2 Simulation results in a realistic context

In the previous Section, the Gaussian extreme value process was fitted on long realizations to test its ability to describe the extremal behavior of usual time series models. In this Section, we validate the proposed methodology on shorter synthetic time series which length correspond to the amount of data typically available in environmental applications (a few years of data). For each time series model introduced in the previous Section, we have repeated the following numerical experiment 200 times:

- Generate a 5-years sequence (one observation per day) of the reference time series model.
- Fit the censored Gaussian extreme value process on this sequence after censoring at the 95% quantile.
- Compute the 100 years return level for the fitted model. It is defined as the level such that the mean number of **clusters** above this level on a 100 years time period is equal to 1 and is computed by simulating a long realization (1000 years) of the fitted model. This

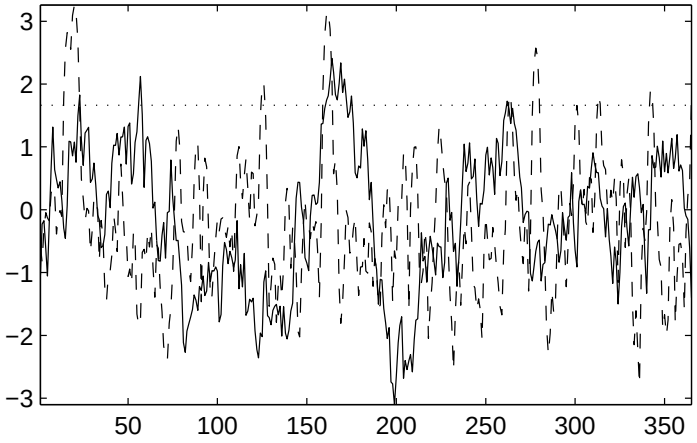


Figure 5: Short samples of the AR(1) model (solid line) and the fitted Gaussian extreme value process (dashed line). The horizontal dotted line is the threshold used for censoring.

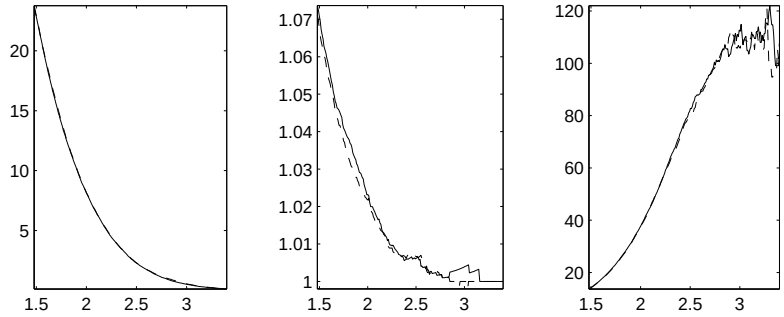


Figure 6: Comparison of the extremal behavior of the IID model (full line) against the fitted Gaussian extreme value process (dashed line). From left to right panels: mean number of up-crossings per year, mean length of the clusters and mean time between consecutive up-crossings as a function of the threshold (x-axis). Results obtained by simulating 1000 years of each model (one observation per day).

quantity was chosen because it is probably the most usual quantity of interest for practical applications. It depends on both the marginal distribution and the dependence structure of the process.

The results are given in Table 1. In every case, the results obtained with the fitted censored Gaussian extreme value process clearly outperform the results obtained with the usual POT method. Both MMLE and CPL_1E give similar results for the three models with no extremal dependence (IID, AR(1) and OU) but the results obtained with the CPL_1E are clearly superior to the ones obtained with MMLE in terms of accuracy and bias for the logARMAX(1) model. The IID and AR(1) models have almost the same 100 years return period, which is expected from the

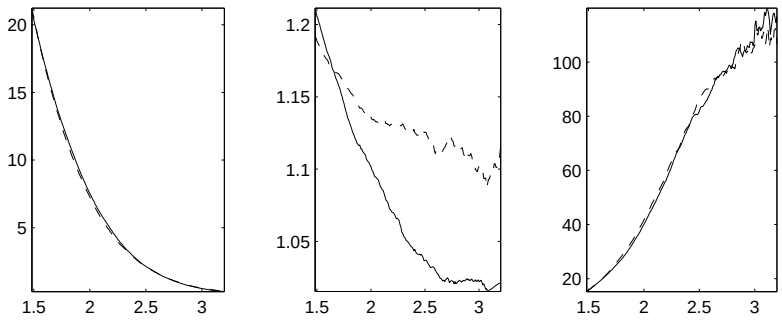


Figure 7: Comparison of the extremal behavior of the AR(1) model (full line) against the fitted Gaussian extreme value process (dashed line). From left to right panels: mean number of up-crossings per year, mean length of the clusters and mean time between consecutive up-crossings as a function of the threshold (x-axis). Results obtained by simulating 1000 years of each model (one observation per day).

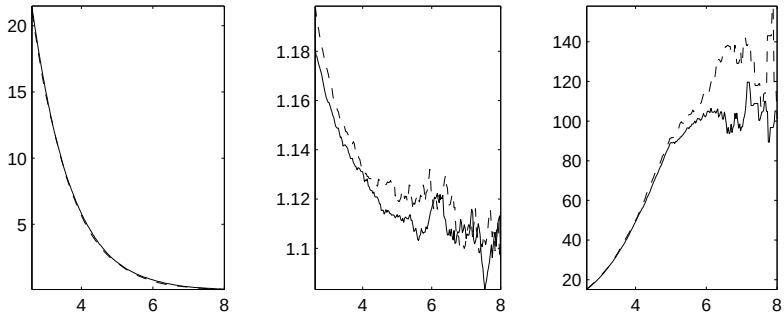


Figure 8: Comparison of the extremal behavior of the logARMAX(1) model (full line) against the fitted Gaussian extreme value process (dashed line). From left to right panels: mean number of up-crossings per year, mean length of the clusters and mean time between consecutive up-crossings as a function of the threshold (x-axis). Results obtained by simulating 1000 years of each model (one observation per day).

theory since they have the same marginal distribution and no extremal dependence. The lower return period of the OU process, which has also the same marginal distribution and no extremal dependence when observed at regular time step, may be due to the irregular time sampling. The extremal dependence of the logARMAX(1) model leads to a higher return value, in comparison with the other models, but also to more important bias and variance in the estimates. This is in adequacy with the simulation results given in Section 3.3 (see Figure 2).

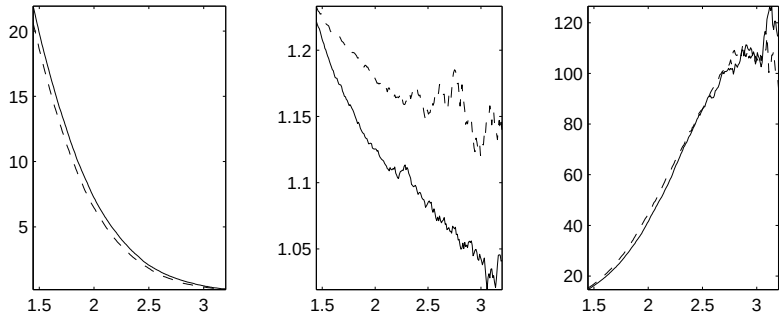


Figure 9: Comparison of the extremal behavior of the OU model (full line) against the fitted Gaussian extreme value process (dashed line). From left to right panels: mean number of up-crossings per year, mean length of the clusters and mean time between consecutive up-crossings as a function of the threshold (x-axis). Results obtained by simulating 1000 years of each model (one observation per day).

Method	IID	AR(1)	logARMAX(1)	OU
True value	4.03	4.03	8.90	3.79
POT	4.18 (3.12—6.02)	4.16 (3.14—5.68)	12.19 (5.31 — 29.96)	3.21 (2.31—5.04)
MMLE	3.89 (3.12—5.15)	3.90 (3.09—4.95)	18.77 (5.88—36.96)	3.62 (2.61—5.05)
CPL_1E	3.84 (3.11—5.17)	3.76 (3.16—4.72)	9.52 (5.54—17.09)	3.62 (2.61—4.98)

Table 1: Mean value of the estimated 100 years return level with 90% fluctuation intervals in parenthesis. Simulation results based on 200 independent 5-years synthetic sequences of each model. A declustering step was applied in the POT method (see [4]).

5 Application to significant wave height

In this section, the proposed methodology is used to analyze the extremal behavior of the three time series of significant wave height (Hs) briefly introduced below.

- **Buoy data**. We focus on data from the buoy Brittany (station 62163), which is part of the UK Met Office monitoring network. It is located at position (47.5 N, 8.5 W) and provides hourly significant wave height data. Buoys are often considered as a reference to provide ground-truth for the other datasets introduced below. In this work, we consider 10 years of data, from 1995 until 2005 (no data available for 2000) and focus on the months of December in order to get rid of the seasonal components. Blocking the data by month leads to waste data and probably to lose important information on extreme events. The development of non-stationary models which include seasonal and interannual components and can be fitted on the whole time series will be the topic of future works. Missing values represent about 7.7% of the data and are generally associated to extreme events (breakdowns generally occur during storms) and this is an important issue when implementing block maxima or POT approaches. Buoys provide accurate information on sea-state conditions but are sparsely distributed over the ocean and there is generally no buoy at the location of interest for a particular application. In such situation, it is important to be able to derive estimates

of the extremal behavior of Hs from the other sources of data introduced below which are available all over the oceans.

- **Reanalysis data**. The ERA-interim dataset provides from a global reanalysis carried out by the European Centre for Medium-Range Weather Forecasts (ECMWF). It can be freely downloaded and used for scientific purposes[1] For this study, we have considered the Hs data available at the location (48 N, 9 W) which is the closest grid point to the buoy Brittany and also extracted the months of December. It leads to a long time series (21 years from 1989 until 2009) which can be analyzed using standard statistical methods since the time sampling is regular (6 hours between successive observations) and there is no missing data. Figure 5 shows both the buoy and ERA-interim time series for the month of December 2005. The agreement is generally good although reanalysis data tends to be smoother than boy data and the quantile-quantile plot (see Figure 11) exhibits some differences between the empirical marginal distributions.

- **Satellite data**. The observations consist of the significant wave height measured at discrete locations along one-dimensional tracks from seven different satellite altimeters which have been deployed progressively since 1991.The dataset and information on it can be freely downloaded[2]. Satellite data exhibits a rather complicated space-time sampling and generally do not provide observations exactly at the location of interest. In order to avoid using interpolation methods, which may smooth the data (see [1]), we have decided to consider the time series obtained by keeping all the closest observations in the tracks which intersect a $1.5^o \times 1.5^o$ box centered on the location of interest (see e.g [24] and [25]). We have retrieved Hs data for 18 months of December, from 1992 until 2009. Figure 5 shows the resulting series for the month of December 2005 at the location of the buoy Britanny. There is fewer data compared to the other datasets and the time sampling is irregular (durations between successive observations ranges from a few minutes to 3 days with a mean value of 19 hours) which prevents from using usual statistical methods (block maxima or POT) for analyzing the extremal behavior of the dataset. Satellite altimeter data have been calibrated using buoys data and Figure 5 shows indeed a good overall agreement between the two time series (see also e.g [15]). Figure 11 suggests however that the empirical distribution of satellite data have an heavier tail than the one of buoy data.

We first focus on reanalysis data in order to avoid estimation problems related to irregular sampling or missing values and detail the practical implementation of the proposed methodology on this particular dataset. We first need to chose a censoring threshold u. It is a crucial choice since u must be chosen high enough to justify the approximation by probabilistic models derived from extreme value theory but not too high in order to keep enough observations to fit the model. A common tool for selecting an appropriate threshold is to fit the model for various thresholds and chose the lowest one such that the estimates are almost stable for any higher threshold value. Indeed, from a theoretical point of view, if the fitted censored Gaussian extreme value process is an appropriate model to describe the behavior of the observed process above a threshold u then it should also be appropriate above higher threshold. However, in practice it is known that is generally difficult to come up to a decision using such diagnostic plots. Figure 12 shows indeed that the values of the parameters do not seem to stabilize when the threshold increases. Comparable results were obtained using similar diagnostic plots when implementing the POT method (see [4] for more details). It may be due to model misspecification or to the estimation error which increases with the threshold since the number of exceedances decreases.

[1] http://data.ecmwf.int/data/
[2] ftp://ftp.ifremer.fr/ifremer/cersat/products/swath/altimeters/waves/

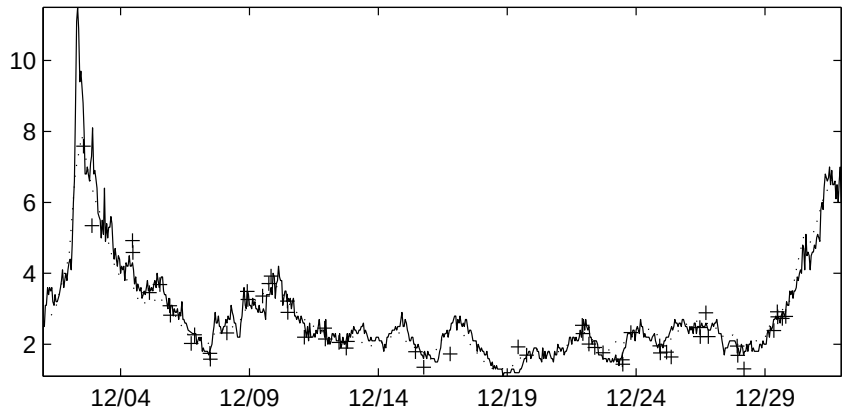


Figure 10: Comparison of the three time series for the month of December 2005. Solid line: buoy data (location (47.5 N, 8.5 W)); dotted line: reanalysis data (location 48 N, 9 W)); plus points: closest satellite observation to the buoy from each satellite track within a 1.5° box.

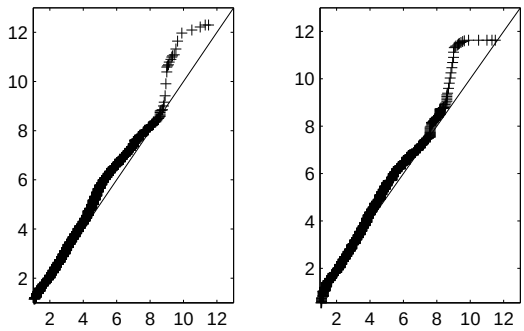


Figure 11: QQ-plot of the sample distribution of the buoy (x-axis) against the empirical distribution of reanalysis data (y-axis on the left panel) and satellite data (y-axis on the right panel)

In this context, adding confidence intervals on Figure 12 would provide helpful information but unfortunately our method does not allow the computation of such intervals. Similar figures were done for buoy and satellite data and led us to select a threshold $u = 6$ meters. The same threshold was used for all the datasets in order to facilitate the comparison.

According to Figure 13, the fitted censored Gaussian extreme value process seems to provide a realistic description of the extremal behavior of the reanalysis data. Indeed, the statistics computed from the data generally lie in the 95% prediction intervals for the fitted model. This indicates that the difference between the statistics computed from the observations and the fitted model may be due to the sampling error and gives confidence in the results obtained when

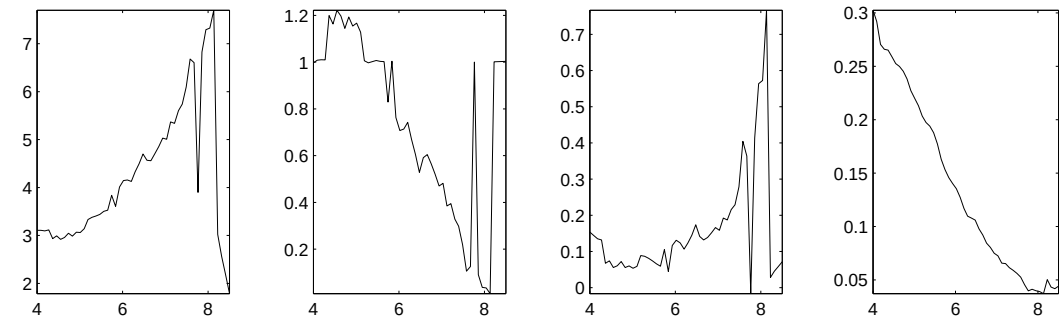


Figure 12: Values of $\mathrm{MPL_1E}$ as a function of the threshold u (x-axis). From left to right panels: estimate of μ, σ, ξ and ν.

extrapolating the extremal behavior of the data using the model to compute, for example, the 100 years return levels discussed below.

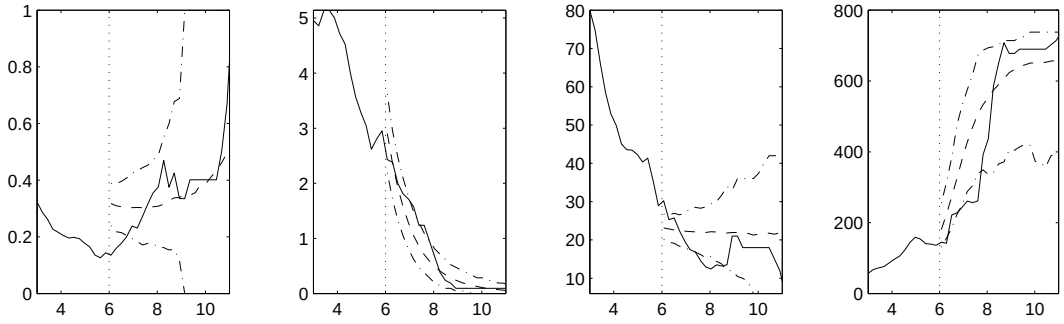


Figure 13: Comparison of extremal behavior of the reanalysis data (full line) and the fitted model (dashed line) as a function of the threshol (x-axis). The dotted lines represents 95% prediction intervals computed by simulating 1000 sequences of the same length than the original data from the fitted model. From left to right panels: estimate of the extremal index (see [12]), number of up-crossings, mean length of the clusters and mean time between consecutive up-crossings.

The censored Gaussian extreme value processes was also fitted to the buoy and satellite time series introduced above by computing the $\mathrm{MPL_1E}$. For comparison purpose, we have also applied the usual POT method with a declustering scheme to the reanalysis dataset (this method is not readily available for other datasets due to the irregular time sampling). Parameter values are given in Table 2 and Figure 14 provides a graphical comparison of the properties of the censored Gaussian extreme value processes fitted on the various datasets. The parameters values of the models fitted on buoy and satellite time series are relatively close but there are important differences with the ones obtained on reanalysis data. In particular, the estimated value of the extremal index is higher for reanalysis data and this additive extremal dependence is probably related to the smoothness of this dataset which tends to smooth out the short term temporal variability. The values of the shape parameter ξ obtained with the $\mathrm{MPL_1E}$ indicate

Data source	μ	σ	ξ	ν	Extremal index	$q_{100\text{ years}}$
ERA-Interim (POT)	3.6328	1.1236	0.0394	NA	0.129	12.47
ERA-Interim (MPL_1E)	4.2002	0.6853	0.1380	0.1341	0.056	14.42
Buoy (MPL_1E)	2.9514	0.9923	0.0533	0.0070	0.213	14.60
Satellite(MPL_1E)	3.6229	1.0913	0.0783	0.0113	0.185	17.96

Table 2: Parameter values for the different datasets. The last column gives the 100 return level.

that the model fitted on reanalysis data also exhibits a marginal distribution with heavier tail, which seems consistent with Figure 11. This seems however contradictory with the results obtained with the POT method on reanalysis data which leads to a smaller value of the shape parameter. However the observed differences in the values of the shape parameter are probably not statistically significant. This is suggested by the following 95% asymptotic confidence interval for ξ, $[-0.22,\ 0.30]$, which has been derived from the information matrix when using the POT method. Figure 14 confirms that the models identified on buoy and satellite exhibits similar characteristics but noticeably differ from the model fitted on the reanalysis dataset. For example, the storm have a typical duration of about 5 hours for the first two datasets whereas the reanalysis dataset identifies storms which mean durations of about 22 hours. Table 2 also gives the 100 years return level, and the results may seem surprising according to the discussion above. Indeed reanalysis and buoy data lead to similar return values (about 14 meters) whereas satellite data leads to a higher value (17 meters). Actually, the return level is a complex function of the four parameters of the Gaussian extreme value process and the differences on the parameter values may compensate to provide similar return levels. Again, the observed differences may not be statistically significant.

Similar results were obtained at other locations where buoys data are available. Buoys and satellite generally lead to identifying model which are close to each other whereas reanalysis data identifies more extremal dependence and longer storms. If we believe that buoys data are a good reference, these results suggest that satellites may provide more accurate information on the extremal behavior of Hs than reanalysis data. In this context, the proposed methodology can be an efficient tool to derive estimates of any quantities of interest, related to the extremal properties of H_s, at any location of the ocean where satellite data are available.

6 Proof of Therorem 3.1

A Gaussian extremal process $\{Z_t\}$ is a moving maxima process as defined in [21]. Using the results given in this paper, we deduce that $\{Z_t\}$ is a stationary unit Fréchet process, continuous in probability, mixing, and hence ergodic. It allows us to use the following theorem which is a straightforward generalisation of Theorem 1.12 in [14]:

Theorem 6.1. *Let $\{Z_i\}_{i=1,\dots,n}$ be a stationary and ergodic process which distribution depends on a parameter $\nu^* \in \Theta$, where Θ is a compact subset of $\mathbb{R}$ and let Q_n be a contrast function defined as*

$$Q_n(\nu) = \frac{1}{n}\sum_{i=1}^{n-1} f(Z_i, Z_{i+1}; \nu),$$

where f is a measurable function with real values and continuous in ν. Suppose that

1. $\mathbb{E}\inf_{\nu\in\Theta} f(Z_1, Z_2; \nu) > -\infty$;
2. $\nu \mapsto \mathbb{E}f(Z_1, Z_2; \nu)$ *has a unique finite minimum at ν^*.*

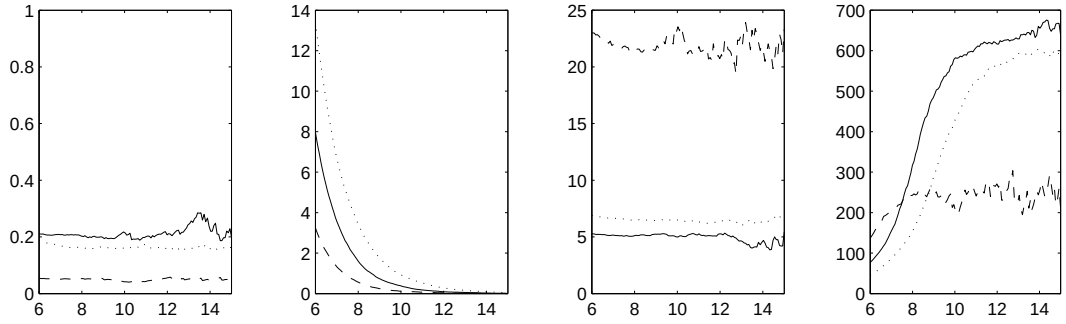


Figure 14: Comparison of the extremal behavior of the models fitted on buoy data (solid line), reanalysis data (dashed line) and satellite data (dotted line). From left to right panels: extremal index, mean number of up-crossings per year, mean length of the clusters and mean time between consecutive up-crossings as a function of the threshold (x-axis). The results were obtained by simulating 1000 years of hourly data with each model.

Then the minimum contrast estimator $\hat{\nu}_n = argmin_{\nu\in\Theta} Q_n(\nu)$ *is strongly consistent:*

$$\lim_{n\to\infty} \hat{\nu}_n = \nu^* \ a.s.$$

We will use this theorem with

$$f(Z_1, Z_2; \nu) = -\log p(Z_1, Z_2; \nu)$$

where $p(z_1, z_2; \nu)$ denotes the joint pdf of (Z_1, Z_2), $\theta = \nu$, $\Theta = (\nu_-, \nu_+)$ with $0 < \nu_- < \nu_+$. An explicit expression for f is given in Section 6.1. It will be used to prove the properties (1) and (2) of Theorem 3.1 in Sections 6.2 and 6.3 respectively.

6.1 Expression of the constrast

We have

$$p(Z_1, Z_2; \nu) = \frac{\partial^2}{\partial z_1 \partial z_2} F_Z(z_1, z_2; \nu)$$

with $F_Z(z_1, z_2; \nu) = \exp((-V(z_1, z_2; \nu))$ and V defined in (7) and thus

$$f(Z_1, Z_2; \nu) = V(z_1, z_2; \nu) - \log\left(\frac{\partial V}{\partial z_1}(z_1, z_2)\frac{\partial V}{\partial z_2}(z_1, z_2) + \frac{\partial^2 V}{\partial z_1 \partial z_2}(z_1, z_2)\right)$$

V and its derivatives satisfy

$$\begin{aligned}
V(z_1, z_2; \nu) &= \frac{\Phi(a/2 + 1/a \log\frac{z_2}{z_1})}{z_1} + \frac{\Phi(a/2 + 1/a \log\frac{z_1}{z_2})}{z_2} = \frac{\Phi(w)}{z_1} + \frac{\Phi(v)}{z_2} \\
\frac{\partial V}{\partial z_1}(z_1, z_2; \nu) &= -\frac{\Phi(w)}{z_1^2} - \frac{\varphi(w)}{a z_1^2} + \frac{\varphi(v)}{a z_1 z_2} \\
\frac{\partial V}{\partial z_2}(z_1, z_2; \nu) &= -\frac{\Phi(v)}{z_2^2} - \frac{\varphi(v)}{a z_2^2} + \frac{\varphi(w)}{a z_1 z_2} \\
\frac{\partial^2 V}{\partial z_1 \partial z_2}(z_1, z_2; \nu) &= -\frac{v\varphi(w)}{a^2 z_1^2 z_2} - \frac{w\varphi(v)}{a^2 z_1 z_2^2}
\end{aligned}$$

with

$$\begin{aligned} w &= \frac{a}{2} + \frac{1}{a}\log\frac{z_2}{z_1} \\ v &= a - w = \frac{a}{2} + \frac{1}{a}\log\frac{z_1}{z_2} \\ \varphi(w) &= \frac{1}{\sqrt{2\pi}} e^{-a^2/8} e^{-\frac{\log^2(z_1/z_2)}{2a^2}} \sqrt{\frac{z_1}{z_2}} \\ \varphi(v) &= \frac{1}{\sqrt{2\pi}} e^{-a^2/8} e^{-\frac{\log^2(z_1/z_2)}{2a^2}} \sqrt{\frac{z_2}{z_1}} \end{aligned}$$

Taking into account that $\frac{\varphi(w)}{az_1^2} - \frac{\varphi(v)}{az_1z_2} = 0$ leads to the following simplified expressions for the derivatives of V

$$\frac{\partial V}{\partial z_1}(z_1, z_2) = -\frac{\Phi(w)}{z_1^2},$$
$$\frac{\partial V}{\partial z_2}(z_1, z_2) = -\frac{\Phi(v)}{z_2^2},$$

$$\frac{\partial^2 V}{\partial z_1 \partial z_2}(z_1, z_2) = -\frac{\varphi(w)}{az_1^2 z_2} = -\frac{\varphi(v)}{az_2^2 z_1} = \frac{e^{-a^2/8}\exp\left(\frac{\log^2 \frac{z_2}{z_1}}{2a^2}\right)}{(z_1 z_2)^{3/2}}.$$

Finally, we deduce that

$$f(z_1, z_2; \nu) = V(z_1, z_2; \nu) - \log\left[\frac{\Phi(w)\Phi(v)}{a^2 z_1 z_2} + \frac{\varphi(w)}{az_1^2 z_2}\right] \tag{15}$$

6.2 Minoration (1)

We have

$$\inf_\nu\{f(z_1, z_2; \nu)\} \geq \inf_\nu\{V(z_1, z_2; \nu)\} + \inf_\nu\left\{-\log\left[\frac{\Phi(w)\Phi(v)}{a^2 z_1 z_2} + \frac{\varphi(w)}{az_1^2 z_2}\right]\right\}$$

and each term of the right hand term of this expression are treated separately below.

- **Term** $V(z_1, z_2; \nu)$.
  V can be bounded using the Fréchet Bound ([11])

  $$P(Z_1 \leq z_1) + P(Z_2 \leq z_2) - 1 \leq \mathbb{P}(Z_1 \leq z_1, Z_2 \leq z_2) \leq \min\{P(Z_1 \leq z_1), P(Z_2 \leq z_2))$$

  which implies that

  $$\min\left(-\frac{1}{z_1}, -\frac{1}{z_2}\right) \leq V(z_1, z_2; \nu) \leq \exp\left(-\frac{1}{z_1}\right) + \exp\left(\frac{1}{z_2}\right)$$

  and thus

  $$\inf_\nu\{V(z_1, z_2; \nu)\} \geq \min\left(-\frac{1}{z_1}, -\frac{1}{z_2}\right)$$

  The right hand term of the last expression has finite expectation since $\frac{1}{Z_1}$ and $\frac{1}{Z_2}$ have unit exponential distributions.

- **Term** $-\log\left[\frac{\Phi(w)\Phi(v)}{a^2 z_1 z_2}+\frac{\varphi(w)}{a z_1^2 z_2}\right]$.

  We have

$$\begin{aligned}-\log\left[\frac{\Phi(w)\Phi(v)}{z_1 z_2}\nu^2+\frac{\varphi(w)}{z_1^2 z_2}\nu\right] &\geq 1-\frac{\Phi(w)\Phi(v)}{z_1 z_2}\nu^2+\frac{\varphi(w)}{z_1^2 z_2}\nu\\ &\geq 1-\frac{\nu^2}{z_1 z_2}\\ &\geq 1-\frac{\nu_+^2}{z_1 z_2}\end{aligned}$$

  and the Cauchy's inequality implies that the right hand term of the last expression has finite expectation since

$$\mathbb{E}\left[\frac{1}{Z_1 Z_2}\right]\leq\sqrt{\mathbb{E}[1/Z_1^2]\mathbb{E}[1/Z_2^2]}=1$$

6.3 Identifiability

We have to prove that $\nu\mapsto\mathbb{E}f(Z_1,Z_2;\nu)$ has a unique finite minimum at ν^* on $\Theta=[\nu_-\nu_+]$. To this end, denote by P_ν the distribution with density function $p(z_1,z_2;\nu)$. Then,

$$\mathbb{E}_{\nu^*}\left[-\log\frac{p(Z_1,Z_2;\nu)}{p(Z_1,Z_2;\nu^*)}\right]=K(P_{\nu^*},P_\nu),$$

is the Kullback-Leibler divergence between P_{ν^*} and P_ν. We have then $K\geq 0$ and $K=0$ iff $P_\nu=P_{\nu^*}$. As the density functions are positive and continuous in (z_1,z_2), this is equivalent to $p(z_1,z_2;\nu^*)=p(z_1,z_2;\nu)$ for all (z_1,z_2). In particular, $K=0$ implies that for all $z_1=z_2=z>0$, we have:

$$\exp\left[-\frac{2}{z}\Phi\left(\frac{1}{2\nu}\right)\right]\left[\frac{\Phi\left(\frac{1}{2\nu}\right)^2}{z^2}\nu^2+\frac{\varphi\left(\frac{1}{2\nu}\right)}{z^3}\nu\right]=\exp\left[-\frac{2}{z}\Phi\left(\frac{1}{2\nu^*}\right)\right]\left[\frac{\Phi\left(\frac{1}{2\nu^*}\right)^2}{z^2}\nu^{*2}+\frac{\varphi\left(\frac{1}{2\nu^*}\right)}{z^3}\nu^*\right].$$

Letting $z\to 0$ while $z>0$ we see that the exponents in the exponential function must be equal, i.e.

$$\Phi\left(\frac{1}{\nu}\right)=\Phi\left(\frac{1}{\nu^*}\right).$$

Hence $\nu=\nu^*$. The proof is complete.

7 Conclusion

In this paper we propose an original method to analyze the extremal behavior of univariate time series. It was motivated by the need to analyze environmental time series with missing values or irregular sampling but the tests performed on classical time series model indicate that the method also performs well on time series with regular sampling compared to the other methods which have been proposed in the literature. The parameters are estimated by using a composite likelihood method and both theoretical and simulation results indicate that it leads to consistent estimates. Results obtained on Hs data indicate that the proposed methodology could be used

to estimate the extremal behavior of Hs from satellite data and produce climatology of extreme Hs all over the ocean which are more accurate than the ones obtained from reanalysis data.

We believe that our methodology is flexible enough to build extensions which may be useful for practical applications. For example, the methodology could deal with other max-stable processes, such as the Brown-Resnick process which could give more flexibility to the model, include non-stationary components or be extended to a space-time model. This will be the subject of future research.

Various aspects of the proposed methodology need also to be improved. In particular, the theoretical results are incomplete. Consistency is only proven in an idealized situation where some of the parameters are known and asymptotic normality has not been established yet. Solving this last problem would have important practical implications since it would give tools to estimate uncertainties.

Acknowledgment

The authors are indebted to the Laboratoire d'Océanographie Spatiale, IFREMER and to the ECMWF for having provided the data used in this study.

References

[1] P. Ailliot, A. Baxevani, A. Cuzol, V. Monbet, and N. Raillard. Space-time models for moving fields with an application to significant wave height fields. *Environmetrics*, 22(3):354–369, 2011.

[2] P. Ailliot, C. Thompson, and P. Thomson. Mixed methods for fitting the gev distribution. *Water Resources Research*, 47, 2011.

[3] J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. *Statistics of extremes*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 2004. Theory and applications, With contributions from Daniel De Waal and Chris Ferro.

[4] S. G. Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag London Ltd., London, 2001.

[5] D. R. Cox. A note on pseudolikelihood constructed from marginal densities. *Biometrika*, 92:729–737, 2004.

[6] A. C. Davison and R. L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B*, 52(3):393–442, 1990.

[7] L. de Haan. A spectral representation for max-stable processes. *Annals of Statistics*, 12(4):1194–1204, 1984.

[8] L. de Haan and A. Ferreira. *Extreme value theory*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006. An introduction.

[9] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997.

[10] R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(02):180–190, 1928.

[11] M. Fréchet. Sur les tableaux de corrélation dont les marges sont données. *Annales de l'Université de Lyon*, 14:53–77, 1951.

[12] M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes*. Springer Series in Statistics. Springer-Verlag, New York, 1983.

[13] B. G. Lindsay. Composite likelihood methods. In *Statistical inference from stochastic processes*. American Mathematical Society, 1988.

[14] J. Pfanzagl. A characterization of sufficiency by power functions. *Metrika*, 21:197–199, 1974.

[15] P. Queffeulou. Long-term validation of wave height measurements from altimeters. *Marine Geodesy*, 27:495–510, 2004.

[16] M. Ribatet, T. B. M. J. Ouarda, E. Sauquet, and J. M. Gresillon. Modeling all exceedances above a threshold using an extremal dependence structure: Inferences on several flood characteristics. *Water Resources Research*, 45(3), March 2009.

[17] P. Ribereau, P. Naveau, and A. Guillou. A note of caution when interpreting parameters of the distribution of excesses. *Advances in Water Resources*, 34:1215–1221, 2011.

[18] M. Schlather. Models for stationary max-stable random fields. *Extremes*, 5(1):33–44, 2002.

[19] R. L. Smith. Max-stable processes and spatial extremes. Unpublished, 1990.

[20] R. L. Smith, J. A. Tawn, and S. G. Coles. Markov chain models for threshold exceedances. *Biometrika*, 84(2):249–268, 1997.

[21] S. A. Stoev. On the ergodicity and mixing of max-stable processes. *Stochastic Processes and their Applications*, 118(9):1679–1705, 2008.

[22] C. Varin. On composite marginal likelihoods. *Advances in Statistical Analysis*, 92(1):1–28, 2008.

[23] C. Varin and P. Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528, 2005.

[24] J. Vinoth and I. R. Young. Global estimates of extreme wind speed and wave height. *Journal of Climate*, 24(6):1647–1665, 2011.

[25] W. Wimmer, P. Challenor, and C. Retzler. Extreme wave heights in the north atlantic from altimeter data. *Renewable Energy*, 31(2):241–248, 2006.

3 Addendum

Afin de compléter l'article ci-dessus, nous fournissons au lecteur quelques informations complémentaires sur les différents outils présentés dans cet article. En premier lieu, nous étudierons le modèle alternatif proposé pour la modélisation des queues de distributions d'observations iid, à savoir le modèle GEV censuré. Dans un second temps, nous donnons quelques résultats complémentaires concernant l'ajustement du modèle proposé à des séries temporelles usuelles. Enfin, nous présentons des estimations de la variance des estimateurs dans différents contextes puisque nous ne disposons pas d'expression analytique, ainsi que la corrélation éventuelle existant entre les estimations des paramètres.

3.1 Comparaison des approximations de queues : loi GEV censurée contre loi de Pareto

Nous allons dès ce paragraphe nous attarder à comparer les deux approches décrites dans l'article qui précède. Nous y avons en effet présenté une approche alternative en ce qui concerne la modélisation de la queue d'une distribution. Comme présenté dans le chapitre précédent, il est usuel en statistique, et dans les applications, d'approcher la loi des dépassements de seuils par une loi de type Pareto (loi GPD), qui s'obtient asymptotiquement à partir de la loi GEV. Comme expliquée dans l'article ci-dessus, une alternative possible consiste à utiliser directement la loi GEV, censurée au-delà d'un seuil. Cette approche est intéressante car elle permet alors d'appliquer la même procédure lorsque les observations sont issues d'un processus, temporel ou spatial : utiliser les lois ou processus max-stables correspondant au contexte étudié. Il est cependant nécessaire de s'assurer de la validité de cette approche, ce qui est le but de ce paragraphe. Pour comparer ces méthodes, nous nous sommes intéressés à l'estimation d'un quantile élevé d'une loi Normale centrée-réduite, afin de ne privilégier aucune méthode. La figure (IV.1) (resp. (IV.2)) montre le biais et l'erreur quadratique moyenne estimés sur 400 réalisations indépendantes d'un échantillon i.i.d pour une taille variable et un seuil fixé au quantile à 90% (resp. une taille fixée à 5000 et un seuil qui varie).

On remarque sur la figure (IV.1) que le biais est toujours négatif en ce qui concerne l'estimation du quantile, ce qui correspond à une sous-estimation en moyenne du quantile cherché. On observe également que ce biais, ainsi que l'erreur quadratique moyenne, décroit quand la taille de l'échantillon augmente, ce qui correspond bien au comportement attendu. Il est intéressant de remarquer également que la méthode basée sur la loi GEV censurée est préférable à la méthode POT classique, et ce que ce soit en terme de biais (figure de gauche) ou d'erreur quadratique moyenne. Ce résultat nous conforte dans le choix de cette méthode pour baser notre extension présentée dans l'article ci-dessus, bien que les deux méthodes soient comparables en termes d'efficacité.

La figure (IV.2) permet quant à elle d'illustrer un propos peu explicité jusqu'alors : le compromis à trouver lors du choix du seuil. On sait en effet que plus le seuil choisi pour ajuster le modèle est élevé, plus le biais dû à l'erreur du modèle sera faible, ce que l'on observe bien sur le graphique de gauche dans la figure précitée. En revanche, le nombre d'observations utilisées alors pour ajuster le modèle est faible, ce qui entraine de plus importantes erreurs d'estimations, ce qui est confirmé par le graphique de droite de cette figure. On observe alors une valeur pour laquelle l'erreur quadratique moyenne, qui est — rappelons-le— la somme du carré du biais et de la variance, est minimale, traduisant ainsi le choix optimal du seuil. En ce qui concerne les méthodes étudiées ici, ce choix optimal semble se situer autour du quantile à 90%. Dans ce cas, on observe également que notre nouvelle approche dépasse la méthode POT usuelle, que ce soit en terme de biais ou de

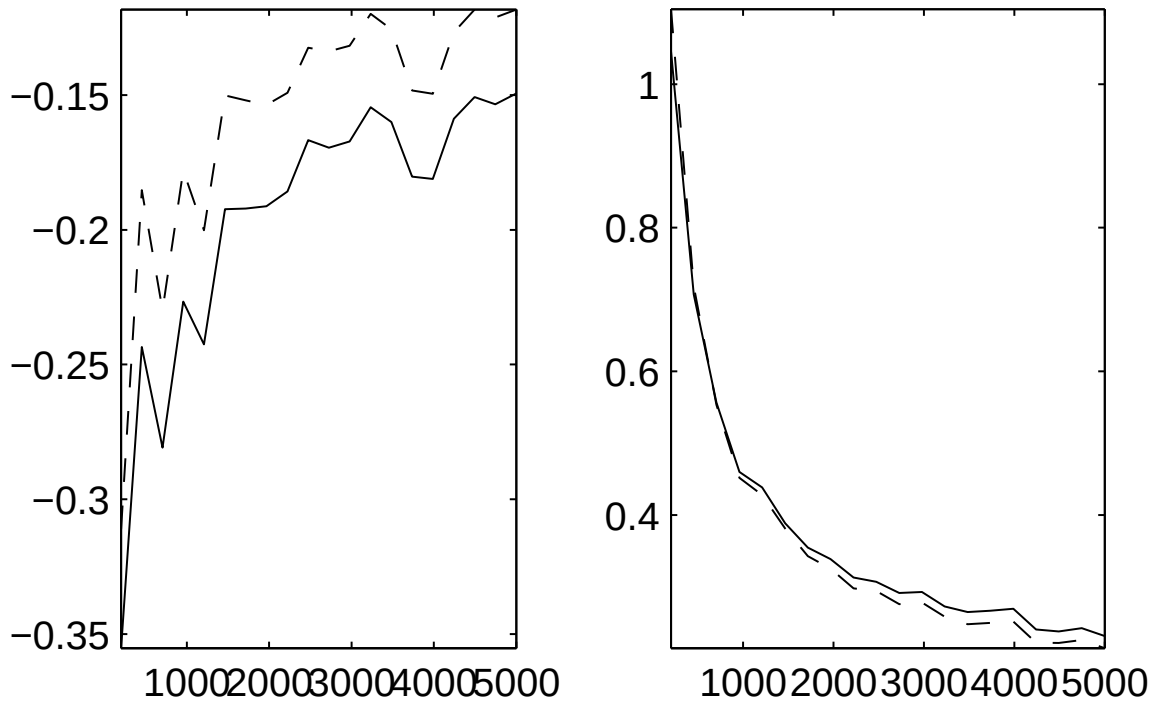


FIGURE IV.1 – Résultats de l'estimation du quantile à 99.99% d'une loi normale standard, par la méthode POT classique (ligne continue) contre la méthode GEV censurée (ligne discontinue). Les deux modèles sont ajustés au-delà du quantile à 90%, la taille de l'échantillon est indiquée en abscisse. En ordonnées : biais (gauche), racine carrée de erreur quadratique moyenne (droite).

variance, tout en étant d'ordre comparable.

3.2 Ajustement sur des séries temporelles classiques

Ci-dessous se trouvent des résultats complémentaires sur la modélisation des extrêmes de séries temporelles classiques, que nous avons retirés de l'article ci-dessus pour des raisons de concision, mais nous pensons qu'ils apportent plus d'informations sur la flexibilité du processus extrémal gaussien ajusté. On peut remarquer en particulier sur les graphiques IV.3 à IV.9 que le processus ajusté permet de bien capturer le comportement extrême des séries, mais a tendance à imposer une structure trop rigide, et en particulier quand la dépendance décroit faiblement, ce qui est le cas pour les processus autorégressifs et d'Ornstein-Uhlenbeck, où il impose une dépendance là où il n'y en a pas. On peut remarquer cependant sur le tableau IV.1 que notre méthodologie permet d'améliorer les résultats de la méthode **POT** declusterisé dans un grand nombre de situations, et en particulier quand les observations sont indépendantes, quoique ce dernier cas est un peu biaisé, car on sait que la méthode **POT** classique serait alors préférable au **POT** declusterisé, mais il est difficile de le savoir lors de l'étude de données réelles. Pour des commentaires plus précis on peut se référer à l'article ci-dessus.

Le tableau (IV.1) nous permet de constater que la méthode proposée permet d'améliorer, dans tous les cas, la procédure POT telle que décrite dans [10]. Dans la majorité

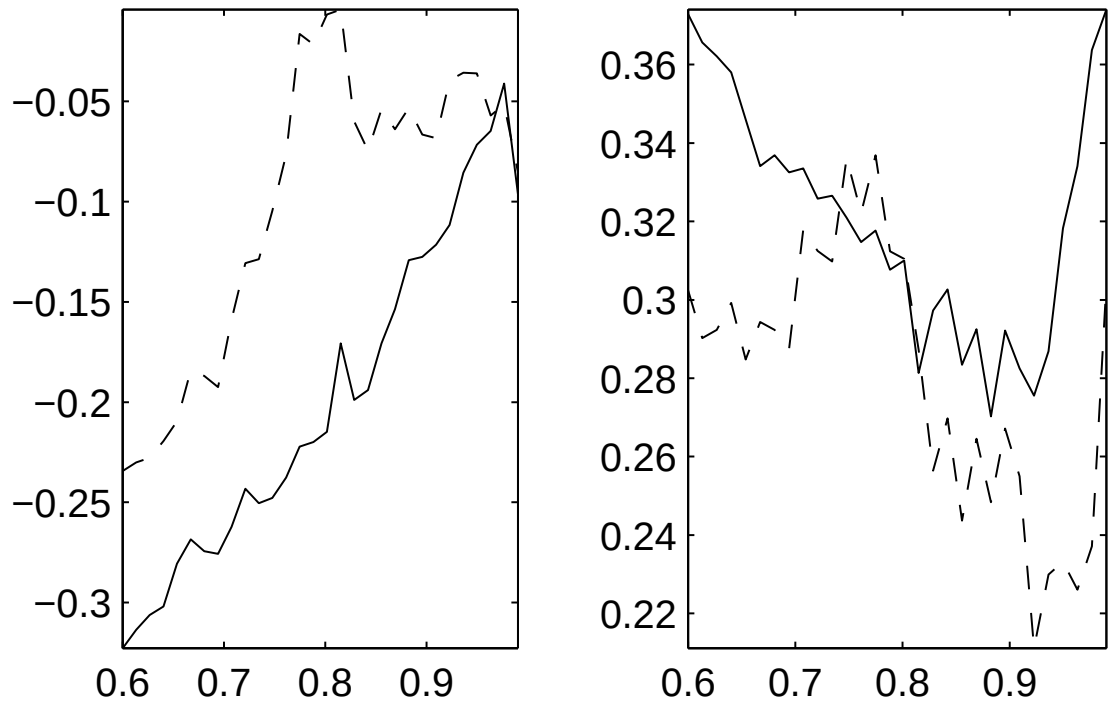


FIGURE IV.2 – Résultats de l'estimation du quantile à 99.99% d'une loi normale standard, par la méthode POT classique (ligne continue) contre la méthode GEV censurée (ligne discontinue). Les deux modèles sont estimés sur un échantillon de taille 5000, le quantile utilisé pour ajuster le modèle est indiqué en abscisse. En ordonnées : biais (gauche), racine carrée de erreur quadratique moyenne (droite).

Model	True	POT	MMLE	CPLE
$\mathcal{N}(0,1)$ 4.03	4.18 (3.12–6.02)	3.89 (3.12–5.15)	3.84 (3.11 – 5.17)	
AR(1) $\phi = 0.2$	3.99	3.84 (3.14–4.79)	3.93 (3.31–4.71)	3.92 (3.32–4.71)
AR(1) $\phi = 0.9$	3.97	3.43(2.63–4.82)	3.87 (2.90– 4.93)	3.89 (2.88–5.01)
logARMAX $\theta = 0.2$	8.80	12.19 (5.31–29.96)	18.77(5.88–36.96)	9.52(5.54–17.09)
logARMAX $\theta = 0.9$	10.42	10.29(6.08–15.17)	9.96 (6.97–13.38)	10.00 (6.95–13.71)
OU $\lambda = 0.05$	3.79	3.21(2.31–5.04)	3.62 (2.61–5.05)	3.62 (2.61–4.98)
OU $\lambda = 2$	4.01	3.69 (3.10–4.55)	3.83 (3.13–4.83)	3.83 (3.14–4.86)

TABLE IV.1 – Comparaison des estimations des niveaux de retour en fonction des modèles de séries temporelles et des méthodes d'estimations. Le modèle a été estimé sur une longue série (1000 ans). Les intervalles de confiance ont été calculés sur 200 réalisations indépendantes des séries temporelles.

des cas, et en particulier quand la dépendance extrêmes est forte (logARMAX(1) avec $\theta = 0.1$) le gain et encore plus important. L'autre avantage majeur de la méthode est de proposer une description plus fine du comportement extrêmes, là où la procédure POT ne donne que la durée moyenne des clusters, par l'extrémal index. Notons de plus que plus la dépendance est forte, plus le niveau de retour estimé sera faible, ce qui confirme l'importance de prendre en compte cette dépendance lors des applications.

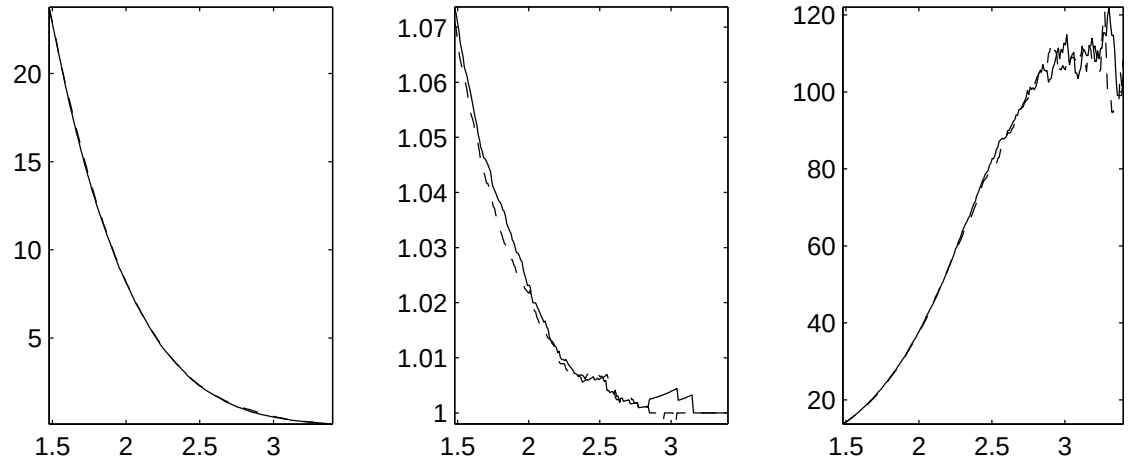


FIGURE IV.3 – Comparaison du comportement extrême de d'observations i.i.d issues d'une loi normale centrée-réduite (ligne continue) contre le modèle Extremal Gaussien ajusté par vraisemblance composite (ligne pointillée). De gauche à droite : nombre moyen de dépassements du seuil par an, longueur moyenne de cluster et temps moyen entre deux dépassements du seuil en tant que fonctions du seuil (abscisse). Résultats obtenus sur 1000 simulations de longueur 1 an pour chaque modèle (une observation par jour).

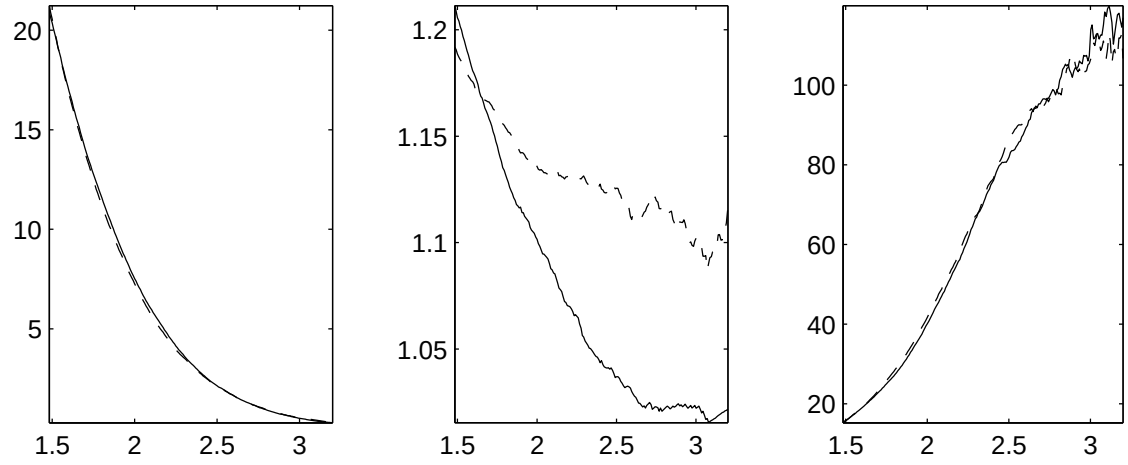


FIGURE IV.4 – Comparaison du comportement extrême de d'observations AR(1) de paramètre 0.2 (ligne continue) contre le modèle Extremal Gaussien ajusté par vraisemblance composite (ligne pointillée). De gauche à droite : nombre moyen de dépassements du seuil par an, longueur moyenne de cluster et temps moyen entre deux dépassements du seuil en tant que fonctions du seuil (abscisse). Résultats obtenus sur 1000 simulations de longueur 1 an pour chaque modèle (une observation par jour).

3.3 Correlation entre les estimateurs

Des travaux récents ([47]) ont rappelé que les estimateurs des paramètres dans le cadre d'estimation des valeurs extrêmes étaient dépendants, ce qui peut poser des problèmes en pratique, en particulier pour calculer des intervalles de confiances, mais aussi en ce

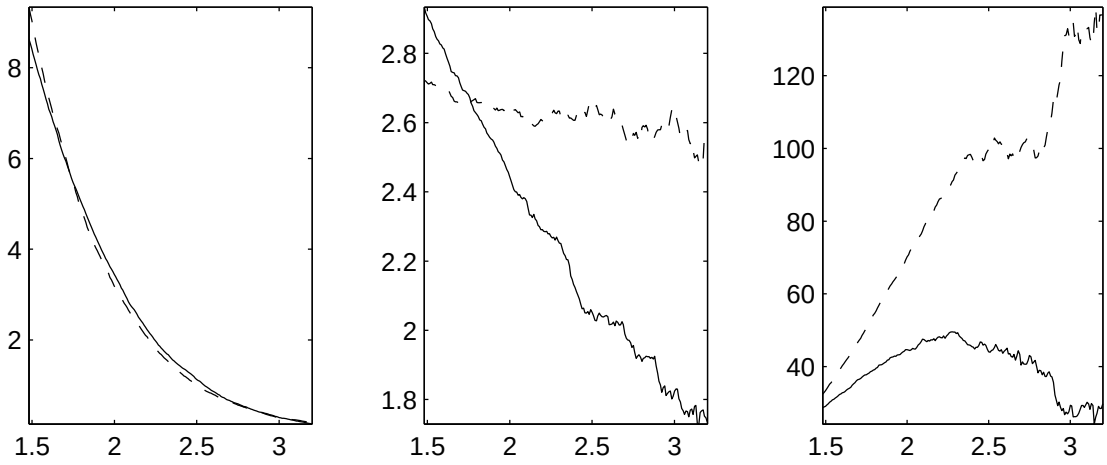


FIGURE IV.5 – Comparaison du comportement extrême de d'observations AR(1) de paramètre 0.9 (ligne continue) contre le modèle Extremal Gaussien ajusté par vraisemblance composite (ligne pointillée). De gauche à droite : nombre moyen de dépassements du seuil par an, longueur moyenne de cluster et temps moyen entre deux dépassements du seuil en tant que fonctions du seuil (abscisse). Résultats obtenus sur 1000 simulations de longueur 1 an pour chaque modèle (une observation par jour).

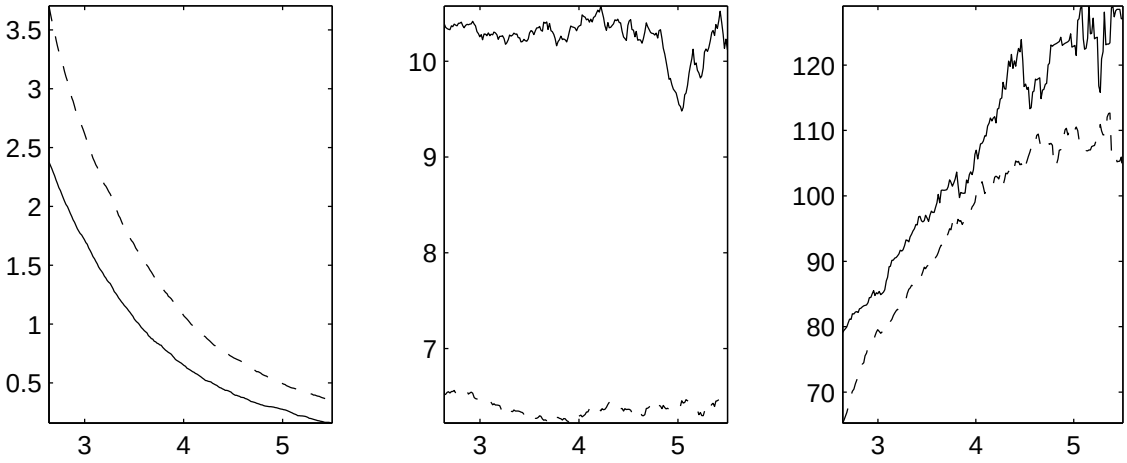


FIGURE IV.6 – Comparaison du comportement extrême de d'observations issues du modèle logARMAX(1) de paramètre 0.1 (ligne continue) contre le modèle Extremal Gaussien ajusté par vraisemblance composite (ligne pointillée). De gauche à droite : nombre moyen de dépassements du seuil par an, longueur moyenne de cluster et temps moyen entre deux dépassements du seuil en tant que fonctions du seuil (abscisse). Résultats obtenus sur 1000 simulations de longueur 1 an pour chaque modèle (une observation par jour).

qui concerne les valeurs des paramètres obtenus par maximisation de la vraisemblance : une corrélation entre les estimateurs induit une corrélation entre les erreurs d'estimations. Nous avons donc cherché à observer ce phénomène, pour juger en particulier si l'estimation du paramètre de dépendance est très liée à l'estimation des paramètres marginaux. Le tableau IV.3 (resp. IV.4) contient la matrice de covariance des estimateurs calculés par

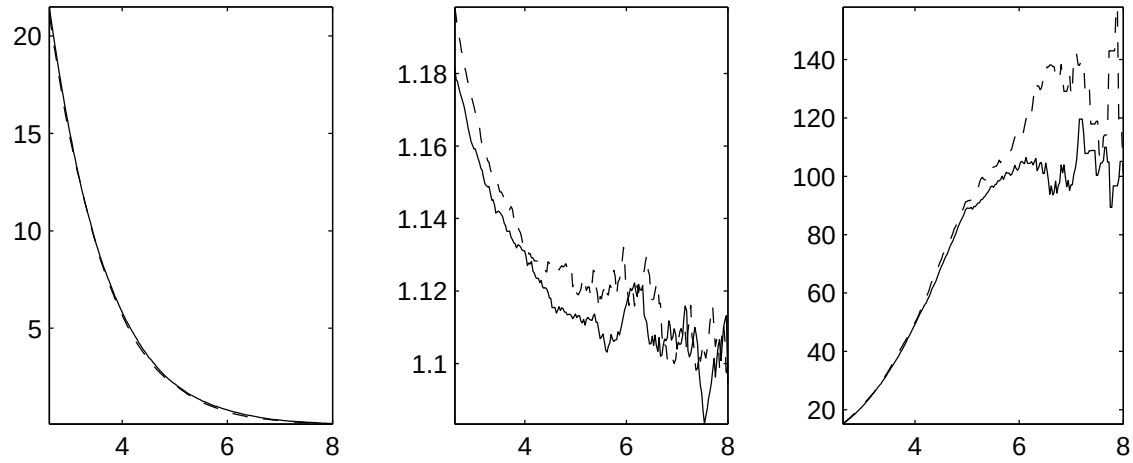


FIGURE IV.7 – Comparaison du comportement extrême de d'observations issues du modèle logARMAX(1) de paramètre 0.9 (ligne continue) contre le modèle Extremal Gaussien ajusté par vraisemblance composite (ligne pointillée). De gauche à droite : nombre moyen de dépassements du seuil par an, longueur moyenne de cluster et temps moyen entre deux dépassements du seuil en tant que fonctions du seuil (abscisse). Résultats obtenus sur 1000 simulations de longueur 1 an pour chaque modèle (une observation par jour).

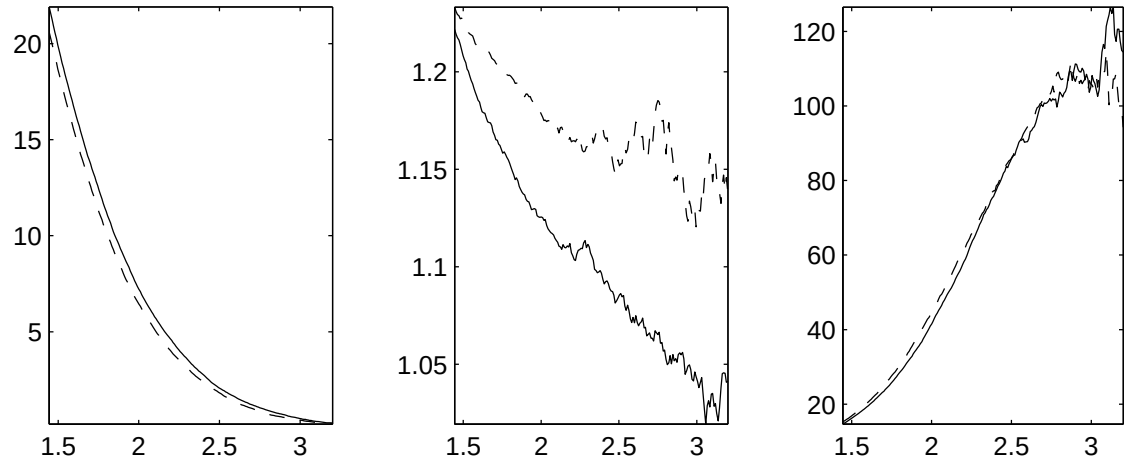


FIGURE IV.8 – Comparaison du comportement extrêmes de d'observations issues du modèle Ornstein-Uhlenbeck de paramètre 0.05 (ligne continue) contre le modèle Extremal Gaussien ajusté par vraisemblance composite (ligne pointillée). De gauche à droite : nombre moyen de dépassements du seuils par an, longueur moyenne de cluster et temps moyen entre deux dépassements du seuil en tant que fonctions du seuil (abscisse). Résultats obtenus sur 1000 simulations de longueur 1 an pour chaque modèle (une observation par jour).

la maximisation de la fonction de vraisemblance dans le cas indépendant (MILE) (resp. la fonction de vraisemblance composite MPL_1E), le tableau IV.2 contenant les écart-types des estimateurs obtenus par les deux méthodes. Ces quantités ont été estimées en effectuant 200 estimations indépendantes des paramètres, sans censure, sur un processus

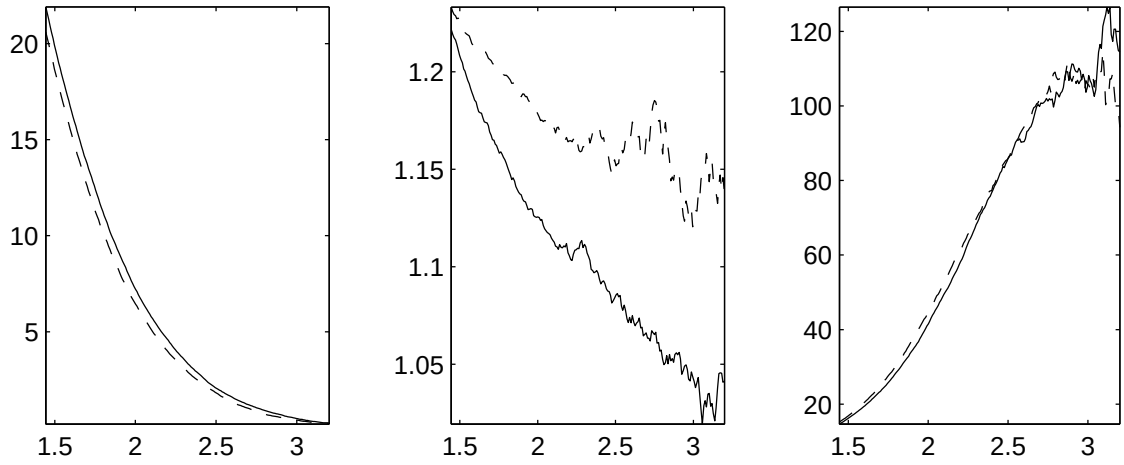


FIGURE IV.9 – Comparaison du comportement extrêmes de d'observations issues du modèle Ornstein-Uhlenbeck de paramètre 2 (ligne continue) contre le modèle Extremal Gaussien ajusté par vraisemblance composite (ligne pointillée). De gauche à droite : nombre moyen de dépassements du seuils par an, longueur moyenne de cluster et temps moyen entre deux dépassements du seuil en tant que fonctions du seuil (abscisse). Résultats obtenus sur 1000 simulations de longueur 1 an pour chaque modèle (une observation par jour).

de longueur 400, dont les paramètres sont fixés.

Méthode	μ	σ	ξ	ν
$MILE$	0.0966	0.0992	0.0817	NA
MPL_1E	0.0802	0.1000	0.0812	0.0618

TABLE IV.2 – Ecart-types des estimateurs des paramètres marginaux pour un processus Extremal Gaussien de paramètres $\mu = 0,\ \sigma = 1,\ \xi = 0.3,\ \nu = 0.5$ de longueur 500.

1.0000	-0.3915	0.1435
-0.3915	1.0000	0.1805
0.1435	0.1805	1.0000

TABLE IV.3 – Matrice de covariance de l'estimateur $MILE$ des paramètres marginaux pour un processus Extremal Gaussien de paramètres $\mu = 0,\ \sigma = 1,\ \xi = 0.3,\ \nu = 0.5$ de longueur 500.

1.0000	-0.3359	0.1468	0.5642
-0.3359	1.0000	0.1753	0.2750
0.1468	0.1753	1.0000	0.0116
0.5642	0.2750	0.0116	1.0000

TABLE IV.4 – Matrice de covariance de l'estimateur MPL_1E des paramètres marginaux pour un processus Extremal Gaussien de paramètres $\mu = 0,\ \sigma = 1,\ \xi = 0.3,\ \nu = 0.5$ de longueur 500.

Si l'on considère les écarts types des estimateurs, on constate qu'elles sont comparables suivant les deux méthodes considérées, ce qui est intéressant en pratique, étant donné que l'on ne dispose pas de formule donnant des intervalles de confiance sur les paramètres : on peut supposer qu'ils seront du même ordre de grandeur que ceux obtenus habituellement par le maximum de vraisemblance dans le cas d'ajustement d'une loi GEV. Cela est d'autant plus intéressant que la fonction de vraisemblance maximisée est plus complexe, et pourrait entrainer des complications numériques. Comme remarqué dans le papier précité, ce sont les estimateurs de ξ et σ qui sont principalement concernés par ce problème, et d'une intensité comparable à ce qui a déjà été remarqué. On notera en particulier que cette corrélation est négative, ce qui s'explique facilement intuitivement : une loi à queue lourde (ξ élevé) produit des observations dont la dispersion est élevée, et la procédure d'estimation aboutit à un maxima local de la vraisemblance où la variance de la loi ajustée est élevée (σ élevé, même si le lien entre la variance de la loi et ce paramètre est un peu plus complexe) et la queue plus faible que ce qui est en réalité. On remarque sur ce même graphique que l'estimateur de μ est très peu corrélé avec les deux autres estimateurs. En ce qui concerne l'estimateur MPL_1E, le comportement est relativement similaire à ce qui vient d'être décrit : les estimations de ξ et σ sont corrélées de manières similaires, de même que l'estimateur de μ est peu corrélé avec les deux autres estimateurs. Si l'on s'intéresse à la dernière colonne de cette matrice, on observe que l'estimateur du paramètre de dépendance ν a une corrélation importante avec les estimateurs de ξ et σ : cela semble s'expliquer par le fait qu'une forte dépendance à paramètres marginaux égaux tend à créer des extrêmes moins forts, mais plus regroupés afin de conserver la loi marginale. En apparence, on a donc une surestimation de la dépendance qui induit une surestimation de la queue des distributions, phénomène qui s'explique, comme nous l'avons vu précédemment par une surestimation de ξ ou σ, et dans notre cas, l'erreur d'estimation se répercute sur les deux paramètres. Les tableaux suivants, IV.5, IV.6 et IV.7 contiennent les mêmes informations que celles détaillées précédemment, à ceci près que la valeur de ξ est différente : nous sommes dans un contexte de queue bornée, avec une valeur de -0.3. Pour la vraisemblance indépendante, le comportement est similaire, à ceci près que l'estimateur de μ est bien plus corrélé avec celui de σ que dans le cas précédent : cette fois-ci, cela s'explique par le fait que le support de loi est influencé par ces deux valeurs, et qu'une erreur sur l'un se propage dans la valeur de l'autre. Pour la vraisemblance composite PL_1 en revanche, le comportement est différent dans ce cas : en effet, l'estimateur de ν est très peu corrélé avec celui de ξ contrairement au cas précédent, alors qu'il est toujours corrélé avec celui de σ : le même phénomène de compensation que celui expliqué dans le cas précédent semble s'appliquer ici uniquement au paramètre σ. Cette caractéristique peut provenir du fait qu'il y a dans ce cas-là moins de liberté pour faire varier ξ, étant donné qu'une valeur trop faible conduit à une distribution dont la borne supérieure est trop faible pour être adaptée aux observations.

Méthode	μ	σ	ξ	ν
$MILE$	0.0569	0.0839	0.0853	NA
MPL_1E	0.0432	0.0859	0.0829	0.0582

TABLE IV.5 – Ecart-types de l'estimateur $MILE$ des paramètres marginaux pour un processus Extremal Gaussien de paramètres $\mu = 0,\ \sigma = 1,\ \xi = -0.3,\ \nu = 0.5$ de longueur 500.

1.0000	-0.6697	0.0529
-0.6697	1.0000	-0.3310
0.0529	-0.3310	1.0000

TABLE IV.6 – Matrice de covariance de l'estimateur $MILE$ des paramètres marginaux pour un processus Extremal Gaussien de paramètres $\mu = 0,\ \sigma = 1,\ \xi = -0.3,\ \nu = 0.5$ de longueur 500.

1.0000	-0.7410	0.0165	-0.0713
-0.7410	1.0000	-0.3409	0.5367
0.0165	-0.3409	1.0000	-0.1316
-0.0713	0.5367	-0.1316	1.0000

TABLE IV.7 – Matrice de covariance de l'estimateur MPL_1E des paramètres marginaux pour un processus Extremal Gaussien de paramètres $\mu = 0,\ \sigma = 1,\ \xi = -0.3,\ \nu = 0.5$ de longueur 500.

4 Conclusions du chapitre

Nous avons dans ce chapitre mis en place un nouvel outil, le processus extrémal gaussien censuré, basé sur des idées proches de celles déjà mises en oeuvre dans de nombreuses applications, à savoir l'approximation des queues de distribution par une loi max-stable, ou par une extension de ces lois dans le cas de la méthode POT. Nous avons de plus proposé des estimateurs des paramètres de ce modèle basés sur des méthodes de vraisemblance composite, et nous avons montré leur consistance dans un cadre théorique simplifié. Nous avons également étudié les propriétés à distance finie de ces estimateurs, mais nous avons aussi cherché à étudier la robustesse du modèle, c'est-à-dire sa faculté à capturer la structure de dépendance extrême de séries temporelles usuelles.

Cette étude nécessite également de nombreuses améliorations que nous avons déjà pointé du doigt : d'une part, il serait sans nul doute intéressant de la mettre en oeuvre avec un processus max-stable plus flexible que celui utilisé, de même que pour les applications il serait très intéressant de développer un modèle permettant de prendre en compte les non-stationnarités.

Concernant les applications, nous avons montré jusqu'à présent une application succincte de cette méthode aux données réelles, une application plus étendue sur les différents jeux de donnés en notre possession fera donc l'objet du chapitre suivant.

Chapitre V

Application : modélisation des extrêmes de la hauteur significative des vagues

L'objectif de ce chapitre sera de présenter les résultats de l'application du modèle présenté dans les chapitres 1 et 3 aux données décrites dans le chapitre 2, et de manière plus approfondie que dans l'application présentée dans l'article du chapitre précédent. On rappelle que le but est de proposer une modélisation du comportement extrême, dans le cas où les observations sont échantillonnées de façon irrégulière, comme rappelé dans le chapitre 2. Cette partie présente trois volets, contenant les points suivants : dans un premier temps, nous nous intéresserons au cas des données de bouées, et plus précisément des bouées K3 et Brittany, sur lesquelles nous ajusterons notre modèle après avoir rappelé le descriptif plus précis de leurs extrêmes, réalisé dans le chapitre 2. Dans une seconde partie, nous présentons les résultats de l'ajustement de notre modèle aux données ERA-Interim aux mêmes points que les bouées, données qui présentent l'avantage d'offrir une discrétisation temporelle constante, sans valeur manquante. La dernière partie contient les résultats de l'ajustement du modèle développé dans le chapitre 3 sur les données satellitaires, et nous présenterons les résultats de l'extension de la méthode proposée jusqu'à présent au cas spatial.

1 Ajustement sur les données de Bouées

L'objectif de ce point est de présenter à la fois les données, issues d'observations *in-situ* sur des bouées situées dans la partie Est de l'Atlantique Nord, au large de l'Irlande et de la Bretagne respectivement pour les bouées K3 et Brittany. Nous passerons en revue chacune de ces bouées, en adoptant la démarche suivante : nous nous intéresserons à la description des caractéristiques extrêmes de ces deux processus temporels, puisque la description univariée plus classique a déjà été faite dans le chapitre 2 ; puis, dans un second temps, nous étudierons et validerons le modèle du chapitre 3.

1.1 Résultats sur la bouée Britanny

Pour plus de détails sur les données de bouées, nous renvoyons le lecteur au chapitre 2, nous ne décrirons ici que le comportement extrême des données de H_s. Lorsque nous nous intéressons au comportement extrême, plusieurs quantités sont d'intérêt : bien souvent dans la littérature sur le sujet, seul le comportement marginal est modélisé, c'est-à-dire quelle valeur maximale peut prendre la variable étudiée. C'est l'objet de tous les résultats présentés précédemment, à savoir que les valeurs extrêmes suivent une loi GEV ou GPD suivant les cas et la modélisation retenue. Or, dans bien des situations, à commencer par la fatigue de structures marines par exemple, ce ne sont pas uniquement les valeurs extrêmes isolées qui posent problème, mais la succession plus ou moins rapide d'extrêmes, raison pour laquelle nous allons nous intéresser aussi à décrire le comportement temporel des extrêmes. Plus précisément, nous allons reprendre les outils présentés dans la fin du chapitre 3, lorsque que nous nous étions intéressés à la robustesse du modèle proposé ; c'est-à-dire que les quantités calculées ci-après sont, pour un seuil donné que l'on fera varier librement :

- Nombre de points dépassant le seuil : ce graphique donne une indication sur le comportement marginal des données ;
- Longueur des clusters, c'est-à-dire le nombre de points consécutifs se trouvant au-dessus du seuil et qui définissent un groupe, ou cluster. Cette quantité peut poser des problèmes à estimer de fait de la présence de données manquantes, car il faut se demander si une telle donnée signe la fin d'un cluster ou non. Nous avons choisi ici que c'était le cas. Ce nombre indique donc une durée des évènements extrêmes ;

– Nombre de clusters observés sur une année, quantité qui est reliée au nombre d'évènements 'indépendants' que l'on va observer au cours d'une période d'observation. Ce graphique permet d'accéder au niveau de retour à p ans, avec p donné : en effet, ce dernier est défini comme le niveau au-delà duquel on observe qu'un seul évènement en moyenne toutes les p années ;
– Nombre d'observations entre les clusters : cette statistique concerne la récurrence des évènements, c'est-à-dire que l'on va chercher quel est le temps moyen qu'il faut attendre avant d'observer un dépassement du seuil à nouveau, un fois l'évènement terminé.

Les quatre statistiques pré-citées ont déjà étés représentées sur le graphique V.2, page 35, on s'y réfèrera pour plus de détails. Nous allons maintenant étudier l'ajustement du modèle présenté dans le chapitre précédent à ces données.

1.1.1 Ajustement du modèle CGEVP

Choix du seuil Comme dans toutes les méthodes de dépassement de seuil telles que présentées dans le chapitre 1, il est nécessaire de se doter d'outils de décision pour choisir au mieux le seuil qui conditionnera grandement les résultats d'ajustement. La méthode du *Mean Residual Life Plot* couramment utilisée n'est pas valable ici, car l'approximation par une loi **GEV** que l'on fait ne dispose pas de la même propriété que lors que la modélisation par une loi *GPD*, à savoir la linéarité de la fonction $u \mapsto \mathbb{E}\left(X - u | X > u\right)$. L'étude de la stabilité des paramètres au-delà d'un certain seuil est par contre toujours valable, et nous détaillerons de plus une autre méthode basée sur les statistiques utilisées pour valider l'ajustement dans la suite de ce paragraphe, ces deux critères pouvant être utilisés conjointement pour choisir le seuil. Dans un premier temps, nous allons nous intéresser à la stabilité des paramètres du modèle ajusté pour choisir un seuil adéquat : pour chaque seuil u, le modèle est ajusté au-delà de u : si le modèle est valide pour un certain seuil u_0, les paramètres doivent se stabiliser pour tout seuil supérieur, même s'il est parfois difficile de juger étant donné que l'intervalle de confiance de chaque estimation s'élargit quand le seuil augmente. Les résultats de cette procédure sont représentés sur le graphique V.1. Nous différencions ici deux comportements districts : pour les paramètres marginaux, on observe effectivement une certaine stabilisation pour des seuils situés entre 5 et 7 mètres. En revanche, pour le paramètre de dépendance, ν, ce palier n'existe pas ou tout du moins n'est pas visible ici. Deux explications sont alors possibles : soit le modèle utilisé est inadapté car les données sont asymptotiquement dépendantes dans les extrêmes, auquel cas le vrai paramètre est effectivement nul. Ce comportement semble pourtant invalidé par les graphiques de la figure (V.2), sur laquelle on peut constater que même pour les niveaux les plus élevés du seuil, les extrêmes semblent arriver par groupe, validant ainsi l'hypothèse de dépendance extrême. On peut alors supposer que la non-stabilisation est due à une erreur d'estimation, causée par exemple par la difficulté à estimer une dépendance quand peu d'observations consécutives sont non-censurées.

La seconde approche consiste à considérer les graphiques de la figure V.2, tout du moins le graphique représentant la durée moyenne des clusters. En effet, quand on est suffisamment dans la queue de la distribution, la durée des clusters converge vers l'inverse de l'indice extrémal, comme nous l'avons rappelé dans le premier chapitre. Ainsi, les valeurs de ces deux quantités devraient se stabiliser, aux fluctuations numériques près. Ce critère nous permet d'avoir un outil complémentaire afin de se guider dans le choix du seuil.

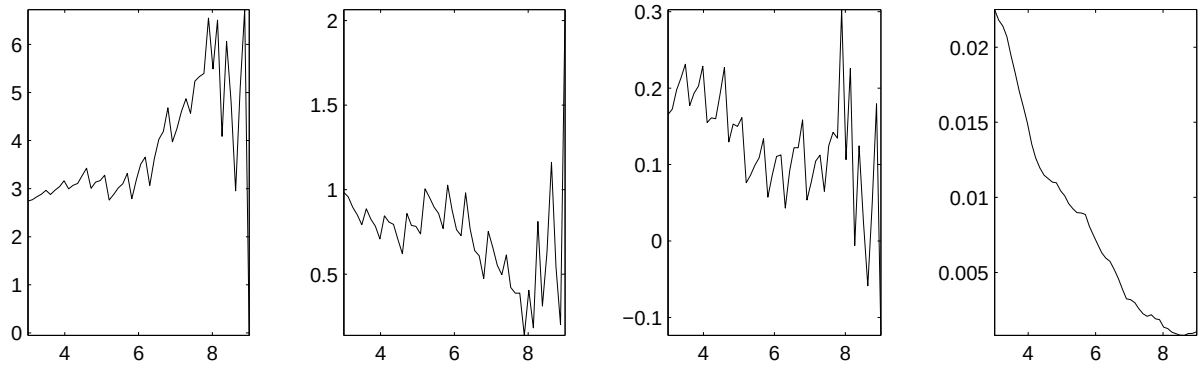


FIGURE V.1 – Valeurs des paramètres ajustés pour différentes valeurs de seuils. De gauche à droite : μ, σ, ξ, ν. Abscisse : valeur du seuil utilisé pour l'ajustement. Ordonnée : valeurs des paramètres.

Résultat d'ajustement Le tableau V.1 contient les valeurs des paramètres du modèle CGEV estimés sur la bouée Brittany par le maximum de la vraisemblance composite à 1 pas de temps (MPL_1E). Ces paramètres sont proches de ceux observés habituellement sur les données de vagues, et en particulier cohérentes avec les valeurs trouvées lors de l'ajustement de la méthode POT usuelle. La figure V.2 permet de comparer les structures extrêmes des observations et du modèle ajusté. On remarque que l'indice extrémal du modèle ajusté est proche pour des valeurs élevées du seuil, de même que pour les nombres d'*up-crossings*. On constate en revanche une certaine différence pour les deux graphiques qui suivent. Au vu des résultats sur les données ERA, qui ne présentent pas de valeurs manquantes, on peut raisonnablement supposer que ce sont ces dernières qui posent problème lors de l'estimation des deux dernières quantités, la longueur des clusters et le temps inter-clusters, plutôt qu'un mauvais ajustement. Il est certes évident que le modèle ajusté est rigide comparativement à la diversité des situations extrêmes, mais comme nous l'avons sur les résultats de simulations, les niveaux sont éloignés de la réalité lorsque le modèle n'est pas adapté. Nous avons ici un écart moins important que celui observé sur les simulations de processus asymptotiquement indépendant, tel que le processus AR, ce qui laisse présager une mauvaise estimation de la durée des clusters et du temps inter-cluster lors de la présence de valeurs manquantes. En effet, les estimateurs utilisés ne sont pas robustes aux valeurs manquantes, et une observation manquante marque la fin d'un cluster d'extrêmes, amenant à une sous-estimation de la longueur des clusters, problème que ne rencontrent pas ces estimateurs sur le modèle ajusté, qui est simulé sur un pas de temps régulier, sans valeurs manquantes. En ce qui concerne l'ajustement, plusieurs remarques s'imposent. En premier lieu, on constate une décroissance très faible de la longueur moyenne des clusters lorsque le seuil augmente, ce qui peut paraitre peu intuitif. Cette caractéristique est due au fait que le processus extrémal gaussien est asymptotiquement dépendant, ce qui signifie que même au-delà d'un seuil élevé, les observations apparaissent par groupe. Cette remarque est intéressante, car ce comportement est important lors de la conception de structures marines, la succession d'évènements extrêmes causant potentiellement d'importants dégâts. Leur durée également est intéressante, puisque l'on observe en moyenne des clusters de 3.5 heures (donc moins de 4 observations sur la bouée). Cette durée est inférieure à celle de l'échantillonnage de ERA-Interim, ce qui laisse supposer que ces données auront quelques difficultés à présenter le même comportement. Pour les satellites également on peut s'attendre à un problème : en moyenne, on a une observation tous les 9 jours à moins

de 1.5° de la bouée Britanny, et une observation tous les 7 jours à moins de 2°.

Données	μ	σ	ξ	ν	θ	q_{100}
Brittany	2.951	0.992	0.005	0.007	0.213	14.60
K3	4.5502	1.030	0.1403	0.0047	0.2468	26.80

TABLE V.1 – Valeur des paramètres du modèle estimé par le MPL_1E sur les données de bouées, ainsi que l'indice extremal et le niveau de retour à 100 ans associés.

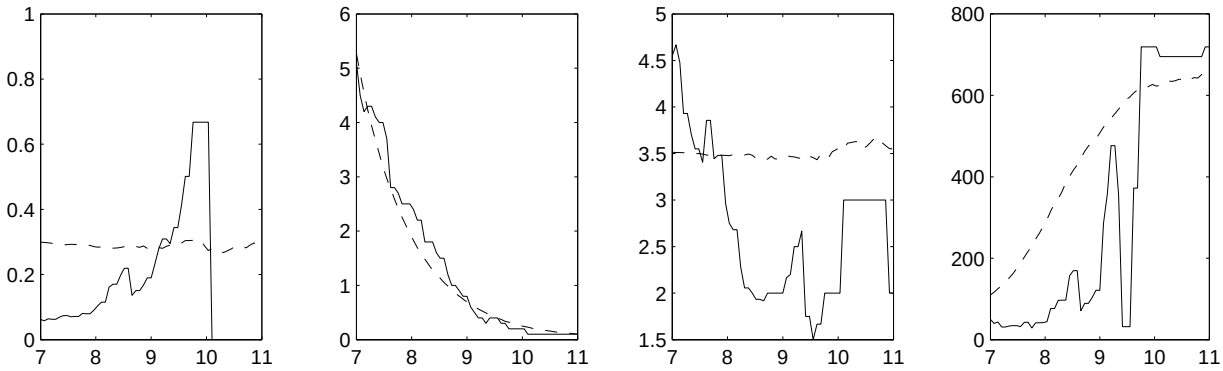


FIGURE V.2 – Statistiques extrêmes calculées sur la bouée Brittany, en fonction du seuil, pour des valeurs de seuil élevées. De gauche à droite : extremal index, nombre moyen de dépassements du seuil par mois de décembre, longueur moyenne des clusters (en heures), temps moyen entre les clusters (en heures). Ligne continue : estimations sur la bouée ; Ligne pointillée : estimations sur le modèle ajusté.

1.2 Résultats sur la bouée K3

Nous avons également ajusté notre modèle à une autre bouée, la bouée K3, située à 53.5° de latitude nord et 19.5° de longitude ouest. Comme précédemment, nous nous somme restreints aux mois de décembre, pour des raisons de stationnarité, et les données sont disponibles de 1997 à 2007, avec un échantillonnage horaire. On a cependant une source avec plus de valeurs manquantes : 14.6%, contre 6.7% sur la bouée Brittany. Les valeurs des paramètres sont disponibles dans le tableau V.1 et les statistiques de validation sur la figure V.3. On peut remarquer tout d'abord l'écart entre les paramètres obtenus sur deux bouées relativement proches (1000km). Cette variation des comportements avait déjà été observé dans le chapitre 2 : les deux bouées présentent des comportements distincts dû au fait que les tempêtes peuvent en toucher une sans être observée par l'autre. En ce qui concerne l'ajustement du modèle, on observe des différences du même ordre que précédemment entre le modèle ajusté et les observations. On remarque également que l'on a une durée des clusters qui est à nouveau supérieure à une heure, ce qui traduit une réelle dépendance au niveau des extrêmes, et que cette information importante pour les applications, est bien restituée par le modèle.

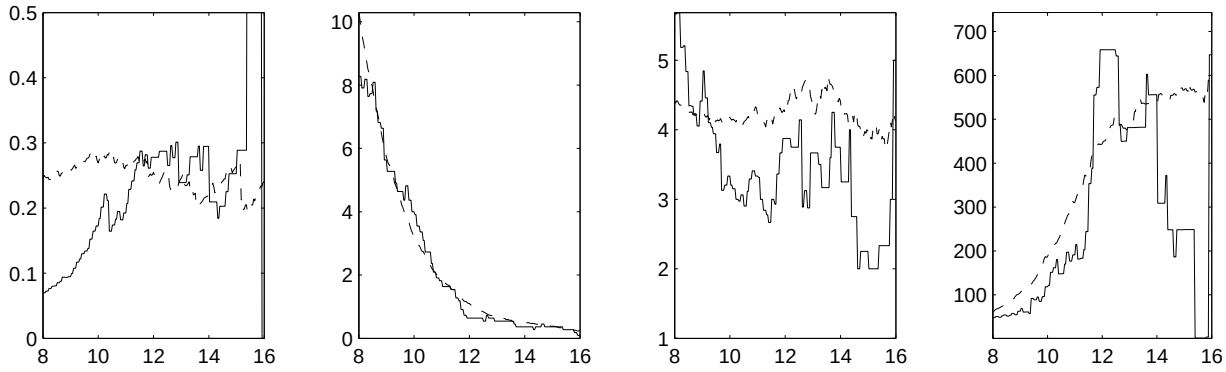


FIGURE V.3 – Statistiques extrêmes calculées sur la bouée K3, en fonction du seuil, pour des valeurs de seuil élevées. De gauche à droite : extremal index, nombre moyen de dépassements du seuil par mois de décembre, longueur moyenne des clusters (en heures), temps moyen entre les clusters (en heures). Ligne continue : estimations sur la bouée ; Ligne pointillée : estimations sur le modèle ajusté.

2 Ajustement sur les données ERA-Interim

Le tableau V.2 contient les résultats de l'ajustement du modèle aux données ERA-Interim. Pour rappel, ces données sont issues de modèles numériques, et ont donc tendance à lisser les évènements extrêmes. Elles sont disponibles sur une grille régulière en espace, et délivrent en chaque point une valeur toutes les 6 heures. Elles sont disponibles de 1989 à 2009.

Emplacement	μ	σ	ξ	ν	θ	$q_1 00$
Brittany	4.200	0.685	0.138	0.134	0.056	14.42
K3	3.749	1.346	0.059	0.140	0.060	18.41

TABLE V.2 – Valeur des paramètres du modèle estimé par le MPL_1E sur les données ERA-Interim aux emplacements des bouées, ainsi que l'indice extremal et le niveau de retour à 100 ans associés.

Le tableau V.2 montre les résultats de l'ajustement du modèle sur les données ERA : comme sur les données de bouée, il existe une certaine différence entre les bouées, qui s'explique par leur situation géographique différente : l'une, plus au Nord, est plus exposée au tempêtes venant du Nord-Ouest, ce qui explique que le niveau de retour calculé soit plus élevé. On observe une différence plus faible entre les paramètres obtenus sur la bouée Brittany et ceux obtenus par les données ERA au même endroit qu'entre les paramètres obtenus sur la bouée K3 et ceux obtenus par les données ERA au même point : cette remarque confirme que les données ERA sont inappropriées au traitement des extrêmes puisqu'aux endroits où les extrêmes sont les plus forts, ERA est plus éloigné de la bouée qu'aux endroits où ces évènements sont d'intensité moindre. Les figures (V.4) et (V.5) montrent quant à elles que le modèle est en relative adéquation avec les observations, ce qui permet d'utiliser les ajustements obtenus pour comparer les sources de données.

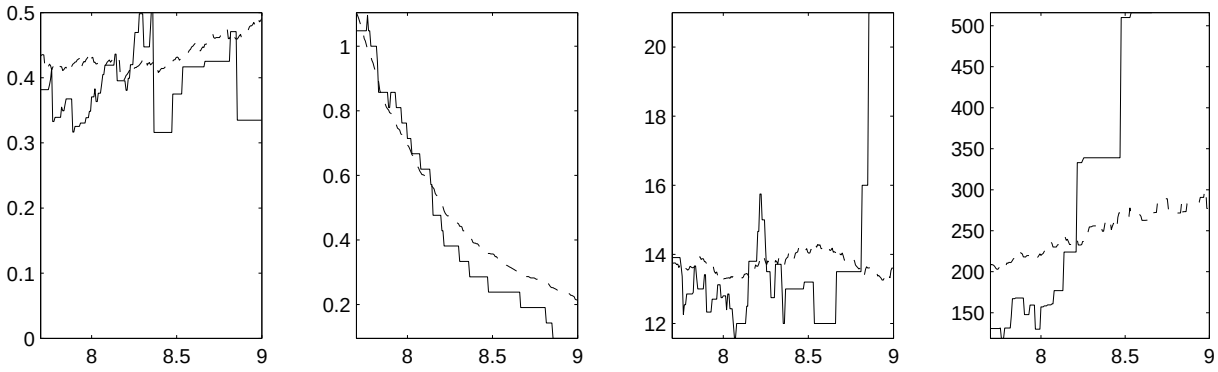


FIGURE V.4 – Statistiques extrêmes calculées sur ERA-Interim, à l'emplacement de la bouée Brittany, en fonction du seuil, pour des valeurs de seuil élevées. De gauche à droite : extremal index, nombre moyen de dépassements du seuil par mois de décembre, longueur moyenne des clusters (en heures), temps moyen entre les clusters (en heures). Ligne continue : estimations sur la bouée ; Ligne pointillée : estimations sur le modèle ajusté.

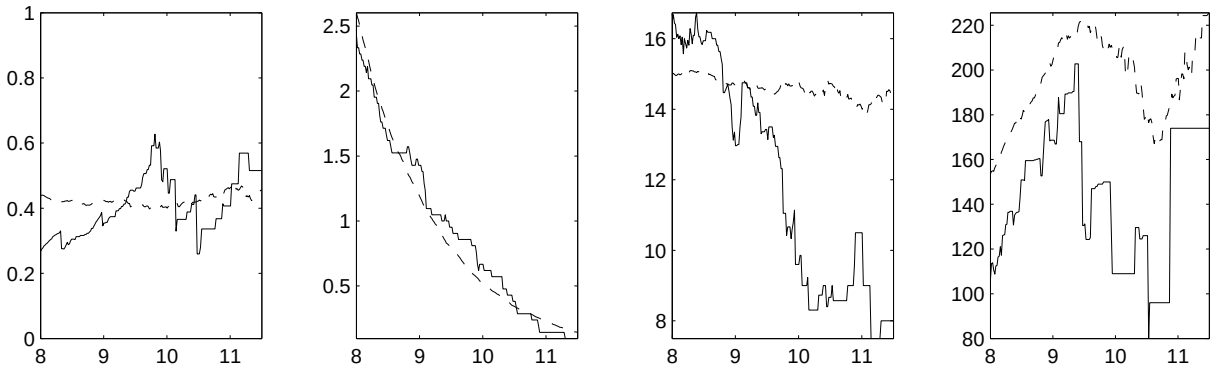


FIGURE V.5 – Statistiques extrêmes calculées sur ERA-Interim, à l'emplacement de la bouée K3, en fonction du seuil, pour des valeurs de seuil élevées. De gauche à droite : extremal index, nombre moyen de dépassements du seuil par mois de décembre, longueur moyenne des clusters (en heures), temps moyen entre les clusters (en heures). Ligne continue : estimations sur la bouée ; Ligne pointillée : estimations sur le modèle ajusté.

3 Ajustement sur les données satellitaires

Nous allons désormais nous intéresser à l'ajustement du modèle sur les données satellitaires, données qui sont au coeur de nos préoccupations. En effet, ces données ont l'avantage d'être disponibles sur tout le globe, avec une couverture temporelle assez importante désormais puisqu'elles sont disponibles depuis 1992. Cependant, comme nous avons pu le décrire précédemment, ces données ne sont pas disponibles à pas de temps régulier, pas plus qu'elles ne sont réparties sur une grille régulière en espace, caractéristiques qui nous ont demandées de proposer une modélisation particulière. Nous allons nous intéresser tout à l'ajustement du modèle présenté jusqu'à présent, avant de décrire et étudier une extension au cas spatial.

3.1 Résultats à l'emplacement de la bouée Britanny

Nous nous intéresserons dans cette partie à l'ajustement du modèle tel que défini dans le chapitre précédent, c'est-à-dire décrivant le comportement extrême d'une série temporelle observée à pas de temps irréguliers. Etant donné que les satellites ne passent pas exactement sur la bouée Brittany, nous avons décidé de conserver les observations tombant à l'intérieur d'une zone autour de cette bouée. Deux tailles seront étudiées ici : des boites de 1.5^o de coté, en accord avec les données ERA-Interim, et des boites de 2^o de coté, conformément aux conclusions de [8], [61] et également [63]. Plusieurs approches sont possibles encore, nous en avons choisi deux : conserver le maxima de chaque trace, ou conserver l'observation la plus proche de la bouée. Il serait également possible de conserver n'importe quel quantile, ou encore la hauteur moyenne par trace, mais il n'y a en général pas de grandes différences avec les deux statistiques étudiées ici (voir par exemple [63]). Pour réaliser l'estimation, nous avons retenu le quantile à 93% pour être consistant avec les études précédentes ([63]), ce qui explique les différences avec les résultats du chapitre 3.

Paramètre	ξ	σ	ν	μ	θ	q_{100}
Valeur (max, boite de 1.5^o)	0.0765	0.9531	0.0053	3.9542	0.2476	25.6708
Valeur (max, boite de 2^o)	0.0038	1.2642	0.0180	3.7466	0.1411	15.6675
Valeur (plus proche obs., boite de $1,5^o$)	-0.1608	1.6304	0.0026	2.8084	0.3450	10.8532
Valeur (plus proche obs., boite de 2^o)	-0.0584	1.4391	0.0029	2.4790	0.3808	13.4917

TABLE V.3 – Valeur des paramètres du modèle estimé par le MPL_1E sur les données satellitaires, ainsi que l'indice extremal et le niveau de retour à 100 ans associés. Le seuil utilisé pour l'estimation est le quantile à 93%.

3.2 Comparaison des ajustements

L'obtention des paramètres pour chaque jeu de données nous permet de construire des comparaisons entre les sources de données : nous avons déjà observé que les valeurs des paramètres étaient différentes, mais nous allons voir quelles conséquences cela implique. Afin de comparer les modèles ajustés, nous avons simulé des réalisations indépendantes, de longueur un mois, avec un écart temporel égal à une heure, ce qui correspond à ce que l'on observe sur la bouée quand il n'y a pas de valeurs manquantes. Ensuite, nous avons calculé les statistiques qui nous sont désormais familières, à savoir l'extremal index, le nombre moyen de dépassements du seuil par mois, la longueur moyenne des clusters (en heures) et enfin le temps moyen entre les clusters (en heures) pour chaque modèle. Les résultats sont sur la figure (V.6). On constate sur cette figure que les données satellitaires sont plus à même de donner des paramètres proches de ceux constatés sur la bouée que sur les données ERA-Interim. Cette figure semble montrer que ces données sont trop lisses, et on tendance à sur-estimer la durée des tempêtes. Ceci peut aussi être du à la fréquence d'échantillonnage : en effet, on constate une durée moyenne des tempêtes les plus importantes d'environ 5 heures, ce qui est inférieur à la résolution spatiale des données de réanalyse. Il se peut qu'une tempêtes se trouve entre deux observations de ERA, ce qui conduit à deux observations moyennes là où une seule plus intense aurait été préférable.

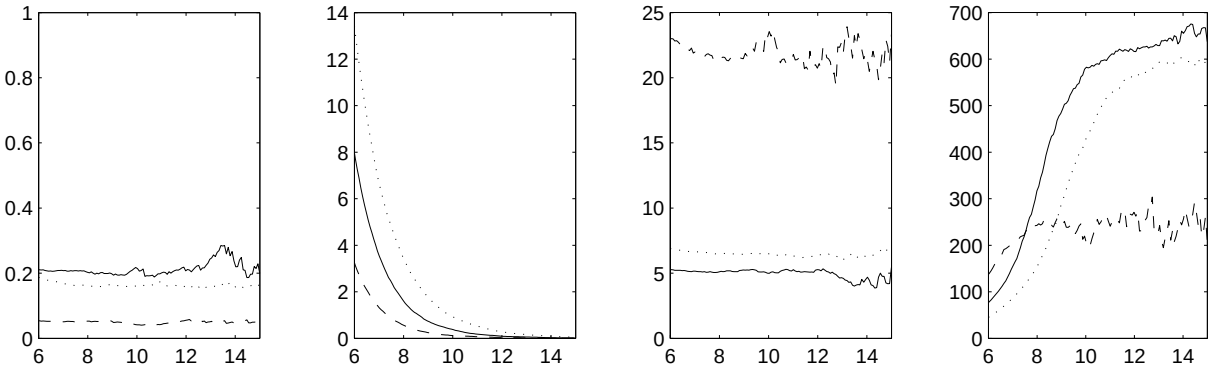


FIGURE V.6 – Statistiques extrêmes calculées sur les différents modèles ajustés. De gauche à droite : extremal index, nombre moyen de dépassements du seuil par mois de décembre, longueur moyenne des clusters (en heures), temps moyen entre les clusters (en heures). Ligne continue : modèle ajusté sur la bouée ; Ligne pointillée : modèle ajusté sur le satellite ; Ligne point-tiret : modèle ajusté sur ERA-Interim.

3.3 Ajustement spatial

Comme précisé tout au long de ce document, l'idée initiale de cette approche fût de proposer une méthode d'estimation adaptée aux données spatiales. Pour des raisons de temps, mais aussi pour motiver l'utilisation de cette modélisation, nous nous sommes restreints jusqu'à présent à des données uniquement temporelles. Dans ce paragraphe, nous allons étendre la procédure développée jusqu'à présent afin de l'utiliser sur les données spatiales dont nous disposons. Le principe d'ajustement sera le même que précédemment, à cela près que la distance entre deux observations est changée. En effet, nous sommes dans un contexte spatio-temporel, et le processus extremal gaussien sur $\mathbb{R}^3$, $Z(x, y, t)$ est caractérisé par une matrice de variance-covariance, donc par une matrice $\Sigma \in \mathcal{S}^3(\mathbb{R})$, ce qui correspond à 6 paramètres. Pour des raisons de temps de calcul, mais aussi pour des raisons d'interprétabilité des paramètres, nous nous sommes restreints aux deux cas suivants :

1. Deux paramètres : $\Sigma = \begin{pmatrix} \sigma_x & 0 & 0 \\ 0 & \sigma_x & 0 \\ 0 & 0 & \sigma_t \end{pmatrix}$
2. Quatre paramètres : $\Sigma = \begin{pmatrix} \sigma_x & \sigma_{xy} & 0 \\ \sigma_{xy} & \sigma_y & 0 \\ 0 & 0 & \sigma_t \end{pmatrix}$

Dans les deux cas, les tempêtes sont modélisées par des formes statiques, correspondant respectivement à des cercles et des ellipses se déplaçant au cours du temps. Une fois encore, ces représentations sont un peu trop simplistes pour modéliser convenablement toute la diversité et la complexité des situations créant de fortes vagues, mais d'une part le modèle est extensible à des processus plus flexibles, et d'autre part nous avons déjà observé que les résultats en 1D étaient en accord avec les observations, ce qui accrédite en partie la méthode. S'il n'a pas été retenues de structures évoluant au cours du temps, c'est que dans le cas présent, il est facile de retrouver la fonction de répartition bivariée du processus en un point, i.e. du processus $Z_t^{(x_0,y_0)} = Z(x_0, y_0, t)$ où le point (x_0, y_0) est fixé, par exemple sur l'emplacement d'une bouée. En particulier, cette loi ne dépend que de la valeur de

σ_t, les autres paramètres n'intervenant pas. Etant donné que la simulation d'un processus en dimension 3 n'a pas encore été implémentée, cette remarque permet tout de même de calculer des niveaux de retour en n'importe quel point de l'espace, bien que cela ne permette pas de calculer le niveau de retour sur une zone géographique donnée. Une autre hypothèse forte a été faite sur les données, puisque nous avons supposé que les paramètres marginaux des dépassements des seuils sont constants en espace, ce qui nous contraint à n'utiliser les observations que sur une petite zone de l'espace.

L'avantage principal de cette approche appliquée aux données satellitaires, est qu'elle permet de prendre en compte toutes les observations d'une trace et non plus uniquement la valeur la plus proche du point étudié. Cela permet donc de conserver plus d'observations lors de l'estimation, et obtenir de ce fait des estimations que l'on espère plus précises. Comme il est habituel de le constater en statistique, il y a un compromis biais-variance à effectuer : la volonté d'avoir plus d'observations pour faire baisser la variance d'estimation, et donc d'augmenter la taille de la zone, menant à une augmentation possible du biais due à la probable non stationnarité spatiale des données.

Modèle	ξ	σ	μ	σ_t	σ_x	σ_y	σ_{xy}
Un paramètre	0.231	0.739	2.924	0.3076	NA	NA	NA
Deux paramètres	0.009	0.873	3.773	0.235	0.003	NA	NA
Quatre paramètres	0.033	1.011	3.559	0.106	3.242	0.006	0.007

TABLE V.4 – Ajustement des modèles spatiaux sur les données satellitaires dans le voisinage de la bouée Brittany

Le tableau V.4 contient les résultats de l'ajustement des deux modèles spatiaux ci-dessus, de même que le résultat de l'ajustement du modèle utilisé dans l'article de la section 3, ne prenant donc pas en compte la dépendance spatiale des données. Pour rester cohérent avec l'analyse effectuée précédemment, nous avons conservé une zone de 1.5^o autour de la bouée Brittany pour ajuster les modèles, mais dans ce cadre, toutes les observations d'une trace sont conservées, ce qui permet de réaliser l'estimation sur un nombre d'observations plus conséquent, et ce au prix d'une vraisemblance plus complexe. Précisons également que la performance des estimateurs étudiés dans le chapitre précédent n'a pas été étudiée dans ce cas spatial. En particulier, les méthodes de vérification de l'ajustement du modèle ne sont pas disponibles directement, et il faudrait développer de nouveaux outils.

On peut cependant continuer à étudier la composante purement temporelle du processus ajusté, puisque comme indiqué précédemment, la fonction de répartition bivariée pour deux points espacé uniquement en temps ne fera intervenir que les paramètres marginaux, ainsi que le paramètre de dépendance temporelle, notée ici σ_t.

Le tableau V.4 nécessite plusieurs commentaires : on remarque déjà que les paramètres marginaux n'évoluent que peu lors de l'extension : en effet, la première ligne contient pour rappel les estimations déjà données, alors que les lignes suivantes contiennent les résultats des ajustements du modèle spatial. Il apparait que seul le paramètre de queue est diminué, bien que nous ne disposions d'outils pour tester si ce changement est significatif : ce dernier peut néanmoins s'expliquer par la présence de plus d'observations moins extrêmes, ce qui se traduit par une diminution du paramètre de queue. Concernant les paramètres décrivant la dépendance du processus, notons les différences qui existent au sujet des paramètres de dépendance spatiale, celui de dépendance temporelle changeant peu en comparaison : en effet, lorsque qu'une liberté est introduite sur la forme des tempêtes, c'est-à-dire entre les modèles à deux et quatre paramètres, on obtient des tempêtes très fortement dissymétriques, avec un étalement Est-Ouest, caractérisé par la valeur de σ_x,

bien plus important que l'étalement Nord-Sud, caractérisé par la valeur de σ_y. On observe également une valeur de σ_{xy} faible, ce qui correspond à des tempêtes dont la forme est proche d'une ellipse dont le grand axe est orienté du Nord au Sud. Cette forte dissymétrie peut être surprenante, d'autant plus qu'un tel comportement n'est pas celui attendu, tout du moins pas dans cette ampleur. Ce résultat peut donc provenir de la répartition spatiale des données, puisque les satellites échantillonnent le long de traces d'orientation quasiment constantes, orientées soit du Nord-Ouest au Sud-Ouest, soit du Sud-Ouest au Nord-Est de la zone étudiée, comme on peut le voir sur la figure II.14.

En tout état de cause, cette application très succincte est principalement destinée à montrer la faisabilité de l'utilisation du modèle proposé pour la modélisation de dépassements de seuils spatio-temporels, même si en l'état des investigations plus poussées sont nécessaires avant d'obtenir une procédure acceptable.

4 Conclusions du chapitre

Cette partie a permis de montrer la flexibilité de notre procédure et sa capacité à se plier à diverses données réelles. Nous avons vu qu'il est possible grâce au modèle ajusté de comparer le comportement extrême de séries ne présentant pas les mêmes caractéristiques. En effet, nous savons que la méthode POT est difficile à appliquer aux données de bouées du fait des valeurs manquantes, et qu'elle n'est même pas applicable aux données satellitaires. Notre procédure d'estimation permet d'obtenir des valeurs de paramètres qui nous permettent de ramener les différentes sources de données à un socle commun, référence que nous avons chois être la bouée en accord avec son rôle de référence qui lui est généralement attribué.

Conclusion générale

Nous nous sommes intéressés dans cette thèse à la modélisation de données relativement particulières que sont les données satellitaires. Ces données, en effet, ont une structure complexe, du fait de l'absence de grille régulière sur laquelle seraient faites les observations. Nous avons à notre disposition un jeu de données important, de qualité comme en attestent les études comparatives, mais de traitement complexe. Ces données ont été utilisées de deux manières différentes, mais à chaque fois dans le but de disposer d'information fiable à des endroits où il y a peu d'information.

Nous avons tout d'abord dressé une analyse descriptive des différentes données en notre possession, à l'aide des procédures usuelles disponibles dans la littérature. Cette étape nous a permis de constater que la structure des données satellitaires était inappropriée à un tel traitement, et qu'en particulier la discrétisation en espace de ces données induisait une perte conséquente d'information. Nous avons alors mis en oeuvre une méthode avancée d'interpolation, qui, tout en prenant en compte une certaine dynamique estimée à l'aide des données de réanalyse, permet de déplacer les traces satellitaires de manière à conserver les structures. L'estimation de cette dynamique a été effectuée à l'aide de modèles à espace d'états, dont l'estimation a pu être réalisée à l'aide de filtrage particulaire. Cette procédure à été appliquée avec succès pour la création de bouées virtuelles, là où nous disposions de données de bouées n'ayant pas servi à ajuster notre modèle, et montre une amélioration de l'adéquation avec la bouée par rapport aux méthodes d'interpolation usuelles. Les possibilités d'évolution de cette approche sont nombreuses, et passent tant par une amélioration de l'estimation des vitesses que par l'amélioration de l'estimation de la structure de covariance des données.

Dans le chapitre 3, nous introduisons une nouvelle procédure de modélisation du comportement extrême d'un processus observé à pas de temps irrégulier. A l'aide de processus max-stable, nous construisons une fonction de vraisemblance faisant intervenir la distance entre deux observations, qu'elle soit spatiale ou temporelle. Cette fonction de vraisemblance nous permet d'obtenir des estimateurs des paramètres décrivant le comportement extrême, estimateurs dont nous avons établi la consistance. Nous avons aussi étudié les propriétés de ces estimateurs à distance finie, dans diverses situations. Nous avons également testé la flexibilité du modèle ajusté, c'est-à-dire sa faculté à capturer la structure extrême du processus initial. Cette vérification a été effectuée sur de nombreux modèles usuels, et de nouveaux outils de diagnostics nous ont permis d'apprécier la qualité de l'ajustement. Nous avons également réalisé une étude empirique sur la variance des estimateurs, puisque nous ne disposons pas de formule analytique, et avons trouvé un comportement similaire à celui observé dans des cas usuels pour lesquels les formules existent. Les pistes pour des recherches futures sont nombreuses, tant d'un point de vue théorique de pratique. Il reste en effet à montrer une convergence en loi des estimateurs afin de proposer une variance d'estimation, mais il est également intéressant d'étendre les résultats au cas où les paramètres marginaux ne sont pas connus. D'un point de vue applicatif, il serait intéressant de relaxer l'hypothèse de stationnarité, en intégrant par exemple des paramètres variant au cours du temps ou des saisons. On peut également envisager d'introduire un processus max-stable sous-jacent plus flexible que celui utilisé, tel que le processus de Brown-Resnick, nos procédures restant similaires.

Dans un dernier chapitre, nous avons mis en oeuvre la procédure décrite précédemment sur les divers jeux de données en notre possession, et avons montré que les données ERA-Interim sous-estimaient d'autant plus les extrêmes que ceux-ci sont importants, en se comparant aux résultats de notre procédure obtenus sur la bouée et montrant une bonne adéquation avec elle. Un autre résultat important est celui obtenu sur les données satellitaires, car on obtient des résultats proches de ceux obtenus à l'aide de la bouée, alors

que les données sont bien moins nombreuses. On note de plus que les résultats sont alors meilleurs que ceux obtenus sur ERA-Interim. Ces résultats montrent qu'il est réaliste d'utiliser les données satellitaires pour estimer des quantités extrêmes, ce qui est intéressant car les bouées sont une source très coûteuse de donnée, quoique fiable. Il serait intéressant de pousser l'investigation plus loin en regardant à d'autre points si ce comportement est confirmé. Nous avons également présenté les premiers résultats de l'extension de cette méthode à des dépassements de seuils spatio-temporels, afin de prendre en compte la structure spatiale des données. Cette brève application est la porte d'entrée vers de nombreuses extensions, que ce soit en terme de modèle que d'estimation. Il est alors également important de permettre une non-stationnarité pour les paramètres marginaux car on sait qu'ils varient fortement en espace, comme rappelé dans le deuxième chapitre.

Ces travaux de recherche ont été l'occasion de présentations orales dans plusieurs conférences internationales, comme Extreme Environmental Events à Cambridge en Décembre 2010 ou à l'EVA [1] à Lyon en Juin 2011. Deux articles découlent de ces travaux de thèse, et ont été intégrés dans le présent document. Le premier d'entre eux a déjà été accepté et publié dans la revue *Environmetrics*, le second étant en cours de relecture par la revue *Annals Of Applied Statistics* au moment de finalisation du présent document.

1. Extreme Value Analysis, Probabilistic and Statistical Models and their Applications

Bibliographie

[1] P. Ailliot, C. Thompson, and P. Thomson. Mixed methods for fitting the gev distribution. *Water Resources Research*, 47, 2011.

[2] A. A. Balkema and L. de Haan. Residual life time at great age. *Ann. Probability*, 2 :792–804, 1974.

[3] J. Beirlant, Y. Goegebeur, J. Teugels, and J. Segers. *Statistics of extremes*. John Wiley & Sons Ltd., 2004.

[4] E. Brodin and C. Klüppelberg. *Extreme Value Theory in Finance*. John Wiley & Sons, Ltd, 2008.

[5] B. M. Brown and S. L. Resnick. Extreme values of independent stochastic processes. *Journal of Applied Probability*, 14(4) :732–739, 1977.

[6] P. Caperaa, A.-L. Fougères, and Genest C. A nonparametric estimation procedure for bivariate extreme value copulas. *Biometrika*, 84(3) :567–577, 1997.

[7] E. Casson and S. Coles. Spatial regression models for extremes. *Extremes*, 1(4) :449–468, 1999.

[8] P. G. Challenor, S. Foale, and D. J. Webb. Seasonal changes in the global wave climate measured by the geosat altimeter. *International Journal of Remote Sensing*, 11(12) :2205–2213, 1990.

[9] S. Coles, J. Heffernan, and J. Tawn. Dependence measures for extreme value analyses. *Extremes*, 2 :339–365, 1999.

[10] S. G. Coles. *An introduction to statistical modeling of extreme values*. Springer Series in Statistics. Springer-Verlag London Ltd., London, 2001.

[11] D. Cooley, D. Nychka, and P. Naveau. Bayesian spatial modeling of extreme precipitation return levels. *Journal of the American Statistical Association*, 102(479) :824–840, 2007.

[12] F. C. Curriero and S. Lele. A composite likelihood approach to semivariogram estimation. *Journal of Agricultural, Biological, and Environmental Statistics*, 4(1) :9–28, 1999.

[13] A. C. Davison and R. L. Smith. Models for exceedances over high thresholds. *Journal of the Royal Statistical Society Series B*, 52(3) :393–442, 1990.

[14] L. de Haan. A spectral representation for max-stable processes. *Annals of Statistics*, 12(4) :1194–1204, 1984.

[15] L. de Haan and A. Ferreira. *Extreme value theory*. Springer, 2006.

[16] L. de Haan and T. T. Pereira. Spatial extremes : models for the stationary case. *Annals of Statistics*, 34(1) :146–168, 2006.

[17] L. de Haan, S. I. Resnick, H. Rootzén, and C. G. de Vries. Extremal behaviour of solutions to a stochastic difference equation with applications to ARCH processes. *Stochastic Processes and their Applications*, 32(2) :213–224, 1989.

[18] A. L. M. Dekkers, J. H. J. Einmahl, and L. de Haan. A moment estimator for the index of an extreme-value distribution. *Annals of Statistics*, 17(4) :1833–1855, 1989.

[19] H. F. Diaz and R. J. Murnane. *Climate Extremes and Society*. Cambridge University Press, 2008.

[20] P. J. Diggle, R. A. Moyeed, and J. A. Tawn. Model-based geostatistics. *Applied Statistics*, 47 :299–350, 1998.

[21] A. Diop and G. Samb Lo. Generalized hill's estimator. *Far East Journal of Theoretical Statistics*, 20(2) :119–131, 2006.

[22] J. Ekengren and J. Bergstrüm. Extreme value distributions of inclusions in six steels. *Extremes*, pages 1–9, 2011.

[23] P. Embrechts, C. Klüppelberg, and T. Mikosch. *Modelling extremal events*, volume 33 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1997.

[24] L. Fawcett and D. Walshaw. A hierarchical model for extreme wind speeds. *Journal of the Royal Statistical Society Series C*, 55(5) :631–646, 2006.

[25] L. Fawcett and D. Walshaw. Markov chain models for extreme wind speeds. *Environmetrics*, 17(8) :795–809, 2006.

[26] L. Fawcett and D. Walshaw. Improved estimation for temporally clustered extremes. *Environmetrics*, 18(2) :173–188, 2007.

[27] R. A. Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 24(02) :180–190, 1928.

[28] C. Gaetan and M. Grigoletto. A hierarchical model for the analysis of spatial rainfall extremes. *Journal of Agricultural, Biological, and Environmental Statistics*, 12(4) :434–449, 2007.

[29] P. Hall and N. Tajvidi. Distribution and dependence-function estimation for bivariate extreme-value distributions. *Bernoulli Journal*, 6(5) :835–844, 2000.

[30] R. Harris. An application of extreme value theory to reliability theory. *The Annals of Mathematical Statistics*, 41(5) :1456–1465, 1970.

[31] J. R. M. Hosking. *L*-moments : analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society Series B*, 52(1) :105–124, 1990.

[32] T. Hsing, J. Hüsler, and M. R. Leadbetter. On the exceedance point process for a stationary sequence. *Probability Theory and Related Fields*, 78(1) :97–112, 1988.

[33] H. Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94 :401–419, June 2005.

[34] Zakhar Kabluchko, Martin Schlather, and Laurens de Haan. Stationary max-stable fields associated to negative definite functions, 2008.

[35] M. R. Leadbetter, G. Lindgren, and H. Rootzén. *Extremes and related properties of random sequences and processes*. Springer-Verlag, 1983.

[36] B. G. Lindsay. Composite likelihood methods. In *Statistical inference from stochastic processes*. American Mathematical Society, 1988.

[37] A. Luceño, M. Menéndez, and F. J. Méndez. The effect of temporal dependence on the estimation of the frequency of extreme ocean climate events. *Proceedings of the Royal Society of London. Series A. Mathematical, Physical and Engineering Sciences*, 462(2070) :1683–1697, 2006.

[38] J. Mendes, P. C. de Zea Bermudez, José Pereira, K. Turkman, and M. Vasconcelos. Spatial extremes of wildfire sizes : Bayesian hierarchical models for extremes. *Environmental and Ecological Statistics*, 2008.

[39] G. L. O'Brien. Extreme values for stationary and markov sequences. *Annals of Probability*, 15(1) :281–291, 1987.

[40] S. A. Padoan, M. Ribatet, and S. A. Sisson. Likelihood-based inference for max-stable processes. *Journal of the American Statistical Association*, 105(489) :263–277, 2010.

[41] R. Perfekt. Extremal behaviour of stationary markov chains with applications. *Annals of Applied Probability*, 4(2) :529–548, 1994.

[42] J. Pickands. Statistical inference using extreme order statistics. *Annals of Statistics*, 3 :119–131, 1975.

[43] J. Pickands. Multivariate extreme value distributions. In *Proceedings of the 43rd session of the International Statistical Institute, Vol. 2 (Buenos Aires, 1981)*, 1981.

[44] P. Prescott and A. T. Walden. Maximum likelihood estimation of the parameters of the generalized extreme-value distribution. *Biometrika*, 67(3) :723–724, 1980.

[45] P. Queffeulou. Long-term validation of wave height measurements from altimeters. *Marine Geodesy*, 27 :495–510, 2004.

[46] M. Ribatet, T. B. M. J. Ouarda, E. Sauquet, and J. M. Gresillon. Modeling all exceedances above a threshold using an extremal dependence structure : Inferences on several flood characteristics. *Water Resources Research*, 45(3), March 2009.

[47] P. Ribereau, P. Naveau, and A. Guillou. A note of caution when interpreting parameters of the distribution of excesses. *Advances in Water Resources*, 34 :1215–1221, 2011.

[48] C. Y. Robert, J. Segers, and C. A. T. Ferro. A sliding blocks estimator for the extremal index. *Electronic Journal of Statistics*, 2008.

[49] H. Rootzen. Extreme value theory for moving average processes. *Annals of Probability*, 14(2) :612–652, 1986.

[50] I. Rychlik, J. RydÃ©n, and C. Anderson. Estimation of return values for significant wave height from satellite data. *Extremes*, 14 :167–186, 2011.

[51] H Sang and A.E. Gelfand. Hierarchical modeling for extremes values. Technical report, Duke University, 2007.

[52] M. Schlather. Models for stationary max-stable random fields. *Extremes*, 5(1) :33–44, 2002.

[53] E. L. Smith and A. G. Stephenson. An extended gaussian max-stable process model for spatial extremes. *Journal of Statistical Planning and Inference*, 139(4) :1266–1275, 2009.

[54] R. L. Smith. Maximum likelihood estimation in a class of nonregular cases. *Biometrika*, 72(1) :67–90, 1985.

[55] R. L. Smith. Max-stable processes and spatial extremes. Unpublished, 1990.

[56] R. L. Smith. The extremal index for a markov chain. *Journal of Applied Probability*, 29(1) :37–45, 1992.

[57] R. L. Smith. Likelihood and modified likelihood estimation for distributions with unknown endpoints. In *Recent advances in life-testing and reliability*, pages 455–474. CRC, 1995.

[58] R. L. Smith and I. Weissman. Estimating the extremal index. *Journal of the Royal Statistical Society. Series B*, 56(3) :515–528, 1994.

[59] S. A. Stoev. On the ergodicity and mixing of max-stable processes. *Stochastic Processes and their Applications*, 118(9) :1679–1705, 2008.

[60] G. Toulemonde, A. Guillou, P. Naveau, M. Vrac, and F. Chevallier. Autoregressive models for maxima and their applications to ch4 and n2o. *Environmetrics*, 21 Issue 2 :113–220, 2010.

[61] J. Tournadre and R. Ezraty. Local climatology of wind and sea state by means of satellite radar altimeter measurements. *Journal of Geophysical Research*, 95 :18255–18268, 1990.

[62] C. Varin. On composite marginal likelihoods. *Advances in Statistical Analysis*, 92(1) :1–28, 2008.

[63] J. Vinoth and I. R. Young. Global estimates of extreme wind speed and wave height. *Journal of Climate*, 24(6) :1647–1665, 2011.

[64] Y. Wang and S. A. Stoev. Conditional sampling for spectrally discrete max-stable random fields. *ArXiv e-prints*, 2010.

[65] W. Wimmer, P. Challenor, and C. Retzler. Extreme wave heights in the north atlantic from altimeter data. *Renewable Energy*, 31(2) :241–248, 2006.

[66] S. Zieger, J. Vinoth, and I. R. Young. Joint calibration of multiplatform altimeter measurements of wind speed and wave height over the past 20 years. *Journal of Atmospheric and Oceanic Technology*, 26(12) :2549–2564, 2009.

[67] V. M. Zolotarev. *One-dimensional stable distributions*, volume 65. American Mathematical Society, 1986.