

Récapitulatif 2025 - Caractérisation des états de mer dans le Raz Blanchard

Bilan de la première année de thèse - Paul-Adrien Alves

22 décembre 2025

Ce document présente brièvement l'ensemble des travaux effectués dans le cadre de ma thèse au cours de l'année 2025.

1 Modélisations statistiques et extrêmes

- Intégration du courant à la relation de dispersion pour le calcul de la vitesse orbitale, prenant explicitement en compte l'influence du courant sur la dynamique des vagues (*janvier 2025*).
- Simulations de séries temporelles complexes couplées à l'estimation de paramètres de distributions de Pareto généralisées (GPD) avec seuil non stationnaire ; comparaison des approches avec et sans non-stationnarité. Les travaux ont donné des résultats peu conclusifs, à creuser ultérieurement si nécessaire (*été 2025*).
- Suite du développement de modèles additifs généralisés (GAM) pour estimation des paramètres de la GPD, en versions stationnaires et instationnaires. Tentatives d'adaptation du niveau de retour avec covariables, mais limites importantes identifiées : absence d'expression analytique stable car le paramètre de fréquence d'occurrence est instable en espace discrétisé, rendant impossible la sélection cohérente d'un quantile. En conséquence, les valeurs de retour ainsi calculées ne sont pas interprétables (*tout au long de l'année, particulièrement au printemps et en été, hors novembre-décembre 2025*).
- Revue bibliographique approfondie sur les méthodes de contours environnementaux : IFORM, DIFORM (en particulier), direct sampling, highest density contour, SPAR, et approches basées sur copules. Ces techniques géométriques offrent un intérêt opérationnel pour explorer des "conditions limites" possibles en termes de variables environnementales, mais présentent des défis pour quantifier l'incertitude et assurer une rigueur statistique d'estimation comparable aux autres méthodes... (*activité récurrente, particulièrement novembre 2025*).
- Modèles d'apprentissage automatique avec XGBoost pour estimer la réponse (vitesse orbitale), surpassant les forêts aléatoires et LightGBM mais pas parfait pour les extrêmes (pour Ciaran, sa prédiction est 0,64 contre 0,71 la valeur "observée" de Resourcecode). Exploration en cours des forêts aléatoires généralisées (pour estimation de quantiles), régression quantile extrême et sa variante avec gradient boosting. Objectif : meilleure estimation des quantiles extrêmes de la hauteur significative (H_s) ou vitesse orbitale, identification des variables les plus influentes (*décembre 2024 et novembre-décembre 2025*).
- Bibliographie variée sur sujets connexes : analyse de sensibilité, océanographie, extrêmes multivariés (copules et mesures de dépendance, processus max stables...)

2 Analyses spécifiques et outils développés

- Analyse par site de courants forts via méthodologies d'extrêmes (contours et GAM), révélant des différences inter-sites significatives (*octobre-novembre 2025*).
- Comparaisons des extrêmes de Resourcecode et HYWAT. H_s de Resourcecode plus élevés, notamment dans les extrêmes (e.g. Ciaran 7.59m (RSCD) versus 6.2m (HYWAT)). Resourcecode surestime légèrement les extrêmes. (*avril-mai 2025*)
- Analyses spectrales : ajustement JONSWAP aux données (souvent insuffisant, surtout extrêmes avec $\gamma < 1$ théoriquement impossible ; surestimation énergie au pic si $\gamma \geq 1$). Comparaisons inter-sites (Bréhat, Fromveur, Barfleur, Raz Blanchard) et profils distincts. Début d'analyse fonctionnelle pour typologie spectrale au Raz Blanchard et stratégie de modélisation (*récurrent toute l'année*).
- Soutien mineur à une alternante : comparaison modèle numérique Resourcecode vs données in-situ Copernicus (réflexions sur le nettoyage de données, calcul d'indicateurs) (*mai-septembre 2025*).

- Travail mineur pour stage M2 : graphiques descriptifs cartographiés (Leaflet) de paramètres d'état de mer en de nombreux points de la Manche, conditions normales et extrêmes ($H_s > \text{quantile } 95\%$), afin d'y identifier des conditions extrêmes réalistes (*mai-juin 2025*).
- Application Shiny pour analyse de points individuels Resourcecode : graphiques descriptifs (histogrammes, séries temporelles, roses des vents/vagues, scatterplots bivariés), quantiles, moyennes par catégorie, régressions linéaires/quantile (B-splines), modèles GPD/GEV basiques, corrélations linéaires, analyse fréquentielle. Outil potentiellement utile et partageable (*janvier-mars 2025*).
- Application Shiny d'analyse spatiale : sélection possible d'une zone de Resourcecode, avec importation des données d'un nombre de points au choix, puis calculs du min/moyenne/max/niveaux de retour GEV avec interpolation spatiale. Intégration de la représentation de données extrêmes travaillée pour le stage M2 sur conditions réalistes en Manche (avec filtrage $H_s > \text{quantile souhaité}$) ; importation des données lente dû à multiples appels API pour paramètres et points multiples (*décembre 2025*).
- Courte note sur l'effet du changement climatique sur les vagues et le vent réalisée pour HydroQuest. Travail s'appuyant sur les conclusions du projet 2CNOW de France Energies Marines (*février-avril 2025*).

3 Points d'intérêt à creuser

- Identification de deux modèles dont le courant n'est pas exclusivement limité au courant de marée (IBI-MFC, HYWAT), en revanche il faudrait les courants de surface pour comparer avec Resourcecode (pas encore vérifié s'ils étaient disponibles). Regarder lien entre vent et vitesse de courant dans un autre rejeu (e.g. HYWAT) Communication avec Martin Goix pour obtenir données ADCP en un point au Raz Blanchard : estimer le courant résiduel en ce point en pourcentage du courant total. Voir si ce point est comparable au point 158348 de Resourcecode utilisé pour mes analyses dans le Raz Blanchard.
- Observer l'impact que peut avoir la pression atmosphérique en conditions extrêmes (pression descendue à 958 hPa au Cap de la Hague lors de Ciaran), sorties ERA5 à récupérer => après discussion, regarder la pression apporterait *a priori* peu d'informations, et il faudrait plutôt regarder ce qu'on pourrait manquer dans Resourcecode du fait qu'on ne prenne pas en compte les surcotes (regarder le rejeu HYWAT).
- Identification de tempêtes historiques majeures avant date du début du rejeu Resourcecode : 1987 et 1990 ("ouragan de 1987" (rafales Jobourg 242 km/h et Cap de la Hague 216 km/h) et tempête Herta (rafales Granville 180 km/h)) vs Ciaran (rafales cap de la Hague 131 km/h). Estimation du H_s et des vitesses orbitales à proposer, malgré les limitations inhérentes à un tel calcul.
- Modélisations statistiques : voies d'amélioration. (i) contours => comparer avec d'autres méthodologies, travailler sur la quantification d'incertitude. (ii) apprentissage automatique => mieux prédire les extrêmes. (iii) modélisation des spectres extrêmes (non encore véritablement débutée).
- Pour le calcul des vitesses orbitales, ne plus se limiter à Airy et passer à du Stokes du 3e ordre.

4 Valorisation envisagée

- **Prioritaire sur début 2026** : Papier en cours de rédaction. Parties descriptive et sur la méthode univariée bien avancées.

Une sorte de boîte à outils pour Resourcecode, par exemple sous la forme d'un package R, qui s'articule autour de cinq axes :

- Des représentations descriptives => applis Shiny ;
- Une modélisation explicative : comment la valeur d'une variable change en fonction d'une ou plusieurs autres ? => GAM, travaux de stage. Méthode spécialisée dans les extrêmes. Méthodologie qui pourra être élargie au corps de distribution (e.g. loi EGPD, un peu travaillée pendant le stage).
- Une modélisation "opérationnelle" avec les contours environnementaux : quelles sont les pires combinaisons simultanées de conditions climatiques susceptibles de se produire une fois toutes les N années ? => contours ;
- Une approche prédictive qui peut venir compléter/se comparer à l'approche explicative. Dans cette approche, on peut intégrer plus de variables qu'en GAM et mesurer l'importance de chacune dans la prédiction => apprentissage automatique. Méthodologie qui pourra être améliorée pour mieux estimer les extrêmes (e.g. extremal random forest) ;
- Modélisation des spectres de vagues lors de conditions extrêmes. Pour l'instant, peu de travaux effectués en dehors de modélisations JONSWAP, qui montrent des limites (trop d'énergie dans le pic au détriment d'autres fréquences).

Ces approches ne répondent pas aux mêmes questions mais se complètent, de sorte à caractériser au mieux l'état de mer.

5 Communications scientifiques

Passées :

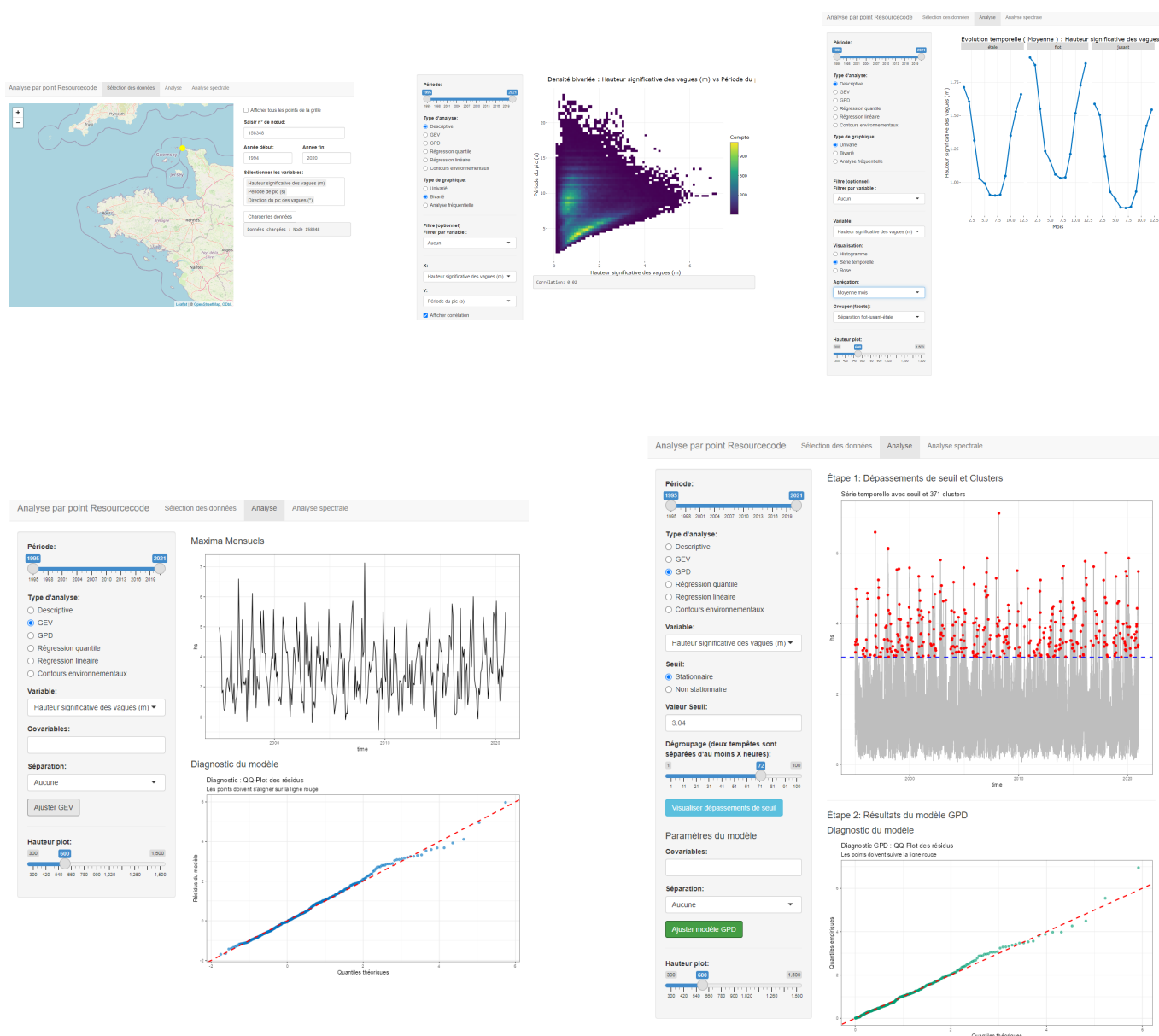
- Colloque “Data Science pour les risques Hydro-climatiques et côtiers”, Roscoff (*fin mars-début avril 2025*).
- 56e Journées de la Société Française de Statistique (SFdS), Marseille (*juin 2025*).
- Présentation au séminaire des doctorants en statistique, Rennes (*visio, mai 2025*).

À venir :

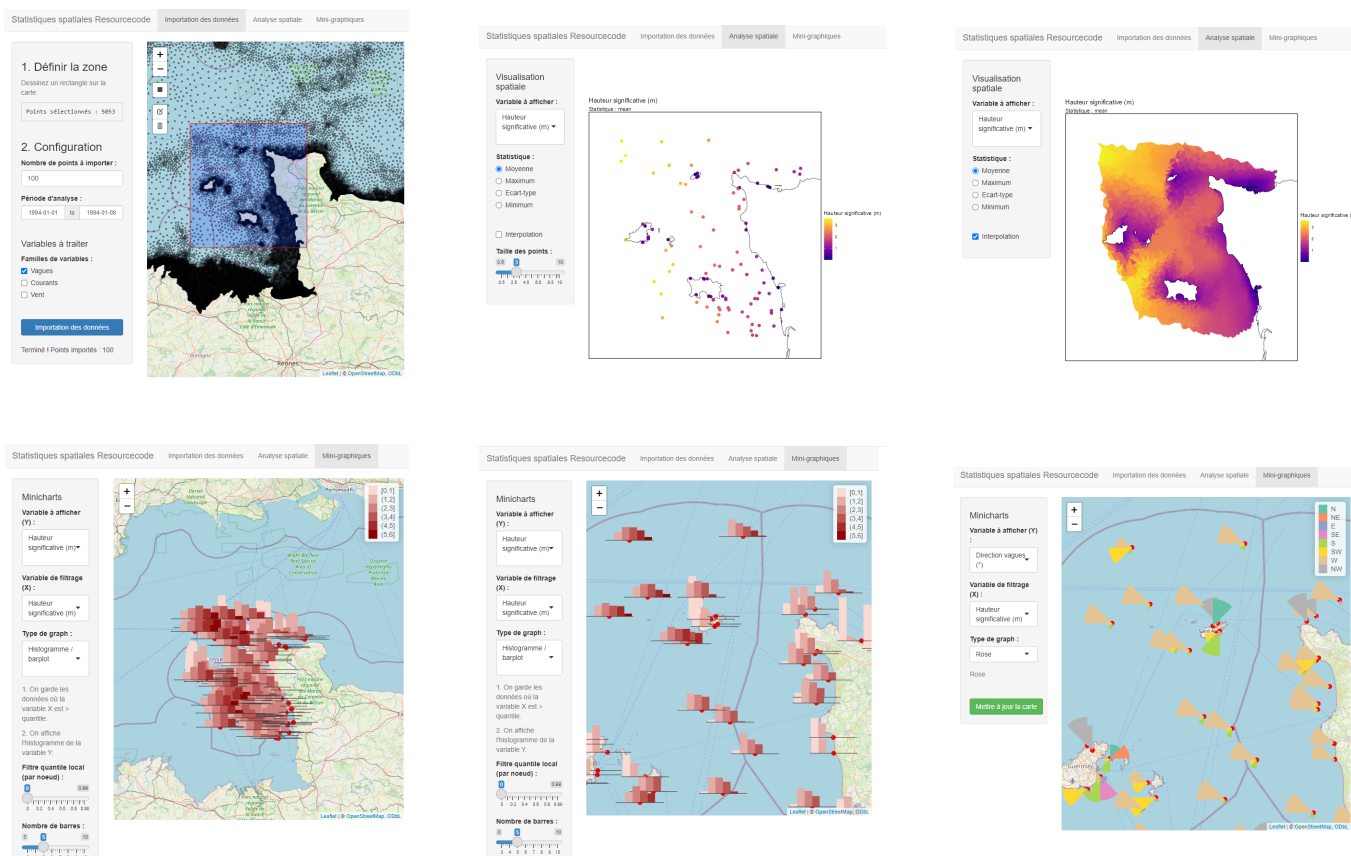
- Présentation au séminaire DeepDive d'Erwan Le Roux, IMT Atlantique (*début 2026*).
- Abstract à envoyer aux 57e Journées SFdS, Clermont-Ferrand (*juin 2026*).
- Abstract à envoyer à EVAN2026 (Extreme Value Analysis and application to Natural hazards), Delft, Pays-Bas, (*juillet 2026*).

6 Applis Shiny, captures d'écran

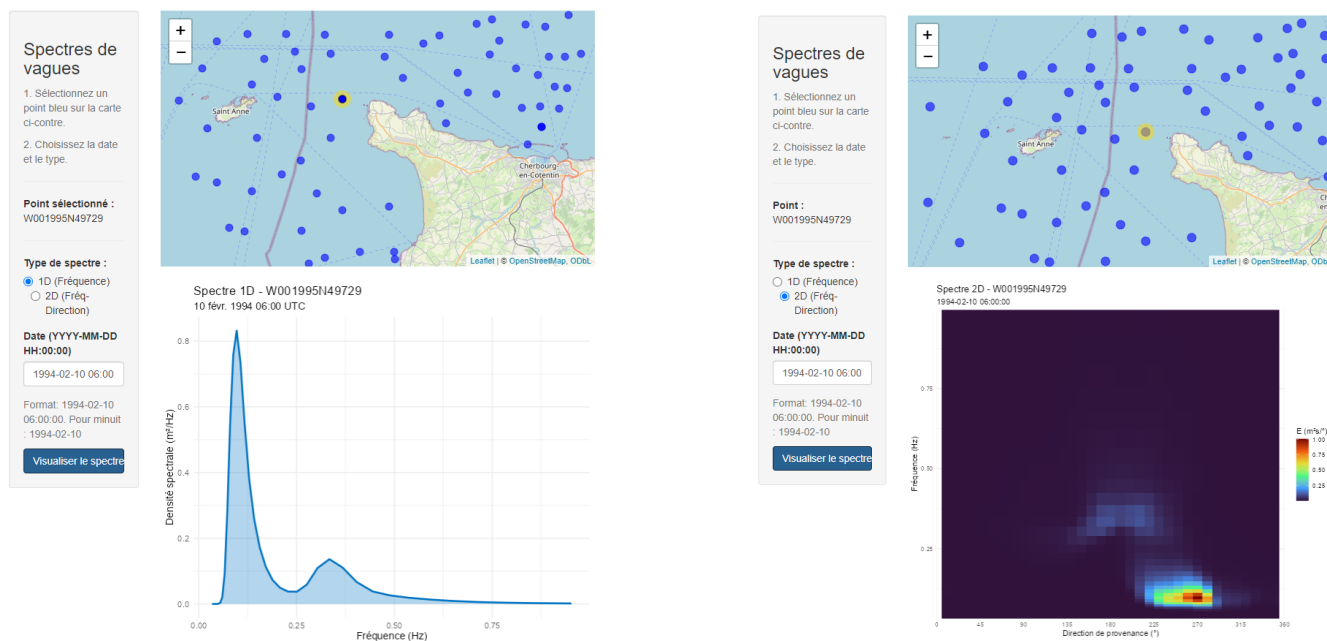
6.1 Appli descriptive en un point



6.2 Appli de visualisation spatiale



6.3 Importation de spectres



7 Méthodologies employées (stage + 1ere année de thèse) et en cours d'exploration

1. Approches explicatives des valeurs extrêmes (méthodes explorées)			
Méthodologie	Principe	Avantages	Limites
Régression quantile (Koenker 1978)	Estimation de quantiles conditionnels (ex : $q_{50}, q_{75}, q_{90}, \dots$) en minimisant une fonction de perte asymétrique (au lieu de la perte quadratique de la régression linéaire).	<ul style="list-style-type: none"> — Simple. — Utilise toutes les données. — Ne dépend pas d'un seuil. 	<ul style="list-style-type: none"> — Incapable d'extrapoler au-delà du support des données. — N'est pas adaptée à l'estimation de quantiles très élevés, tels que $> 95\%$.
Maxima par blocs (loi GEV)	Division de la série temporelle en blocs (ex : mois ou années) et ajustement de la loi GEV (μ, σ, ξ) sur le maximum unique de chaque bloc.	<ul style="list-style-type: none"> — Fondement théorique robuste (Théorème de Fisher-Tippett). — Indépendance des blocs généralement assurée. 	<ul style="list-style-type: none"> — Perte d'information massive (ne garde qu'une donnée par an ou mois) \Rightarrow problème si plusieurs événements extrêmes surviennent dans le même bloc. — Ne donne pas les conditions d'état de mer dans lesquelles se produisent les événements extrêmes.
Dépassements de seuil (loi GP)	Ajustement d'une loi GP sur les pics indépendants au-dessus d'un seuil u (POT). Les paramètres (σ, ξ) sont constants.	<ul style="list-style-type: none"> — Simplicité et rapidité. — Utilise plus de données que la méthode des Maxima par Blocs (GEV). — Niveau de retour unique. 	<ul style="list-style-type: none"> — Ne donne pas les conditions d'état de mer dans lesquelles se produisent les événements extrêmes. — Sensibilité au choix du seuil.
Dépassements de seuil (loi GP non stationnaire)	Les paramètres de la loi GP varient en fonction de covariables via des splines de lissage. On utilise pour cela un modèle additif généralisé (GAM) (Hastie et Tibshirani 1987, Wood 2015).	<ul style="list-style-type: none"> — Capture l'effet des covariables (ex : interaction vagues-courant). — Possibilité d'ajuster un seuil non stationnaire. 	<ul style="list-style-type: none"> — Sensibilité élevée à la base et au nombre de splines. — Niveau de retour non calculable en raison de la nécessaire discrétisation de l'espace des covariables.
Extended GPD (EGPD) (Naveau et al., 2016)	Modélisation conjointe des valeurs faibles, modérées et élevées/extrêmes via une famille paramétrique étendue.	<ul style="list-style-type: none"> — Supprime la sélection du seuil. — Utilise toute les données. 	<ul style="list-style-type: none"> — Inférence plus complexe (estimateur des moments pondérés) et longue. — Sensibilité à la forme paramétrique choisie. — En version non stationnaire, niveau de retour non calculable.

2. Approches explicatives des valeurs extrêmes (méthodes à explorer / en cours d'exploration)			
Méthodologie	Principe	Avantages	Limites
Régression quantile extrême	Extension de la régression quantile classique aux quantiles très élevés ($\tau \rightarrow 1$) en combinant la flexibilité de la régression avec l'extrapolation de la théorie des valeurs extrêmes (queue lourde).	<ul style="list-style-type: none"> — Permet d'extrapoler au-delà du support des données observées (contrairement à la régression quantile classique). — Gère des relations non-linéaires complexes entre covariables et extrêmes. 	<ul style="list-style-type: none"> — Complexité d'optimisation (fonctions de perte non différentiables ou complexes). — Nécessite beaucoup de données pour converger sur les quantiles très élevés (> 0.99).

TABLE 1 – Comparaison des méthodes univariées et de régression : méthodes explorées (haut) vs en cours d'exploration (bas).

3. Contours environnementaux (méthodes explorées)			
Méthodologie	Principe	Avantages	Limites
Contours IFORM (inverse first order reliability method, Winterstein et al., 1993)	Transformation de la loi jointe vers l'espace normal standard (Rosenblatt), définition d'une sphère de rayon β , puis transformation inverse. Approximation linéaire de la surface de défaillance (inconnue) au point de design.	<ul style="list-style-type: none"> — Standard industriel (DNV). — Interprétation géométrique simple dans l'espace standard. 	<ul style="list-style-type: none"> — Biais introduit par la transformation de Rosenblatt non-linéaire. — Forte sensibilité au modèle de loi jointe choisi.
Contours D-IFORM (Direct IFORM, Derbanne et de Hauteclercq, 2019)	Projection des données sur des droites selon u , ajustement GPD sur les maximums projetés pour trouver z_u . Contour formé par intersection d'hyperplans. Approximation linéaire de la surface de défaillance (inconnue) au point de design.	<ul style="list-style-type: none"> — Passage en dimension $d > 2$ plus aisé. — Évite la transformation vers l'espace normal standard. 	<ul style="list-style-type: none"> — Pas de quantification native de l'incertitude. — Sensibilité au nombre de directions et au seuil GPD. — Sensibilité au lissage des contours.

4. Contours environnementaux (méthodes à explorer / en cours d'exploration)			
Méthodologie	Principe	Avantages	Limites
Highest density contour (HDC) (Haselsteiner et al., 2017)	Définit le contour comme la frontière de la région de plus haute densité de probabilité f_m englobant une masse $1 - \alpha$. Occupe le volume minimal pour une probabilité donnée.	<ul style="list-style-type: none"> — Définition probabiliste stricte : plus conservateur et sûr pour le design que l'IFORM. — Ne dépend pas de la convexité de la surface de défaillance. 	<ul style="list-style-type: none"> — Nécessite l'estimation de la densité jointe sur une grille.
Direct sampling contour (DSC) (Huseby et al., 2015)	Construction du contour directement dans l'espace physique par simulation Monte Carlo utilisant le modèle de loi jointe, sans passer par l'espace normal standard.	<ul style="list-style-type: none"> — Élimine le biais lié à la transformation de Rosenblatt. — Interprétation probabiliste directe dans l'espace des variables. 	<ul style="list-style-type: none"> — Coût calculatoire : nécessite beaucoup de simulations pour stabiliser le contour extrême. — Sensibilité au lissage. — Sensibilité au modèle de loi jointe.
Modèle SPAR (semi-parametric angular radial) (Mackay et al., 2024)	Transformation en coordonnées polaires (R, Θ) . Modélisation non-paramétrique de la densité angulaire Θ et modélisation GPD conditionnelle du rayon $R \Theta$.	<ul style="list-style-type: none"> — Cadre rigoureux pour l'extrapolation multivariée. — Aucune hypothèse a priori sur la structure de dépendance. — Capture bien la dépendance asymptotique. 	<ul style="list-style-type: none"> — Sensibilité à de nombreux hyper paramètres (largeur de bande noyau, seuil radial, norme L1/L2...). — Inférence complexe et longue.

TABLE 2 – Comparaison des méthodes de contours : méthodes explorées (haut) vs en cours d'exploration (bas).

5. Approches prédictives : méthodes d'apprentissage automatique (méthodes explorées)			
Méthodologie	Principe	Avantages	Limites
Random forest (Breiman, 2001) & XGBoost (Chen et al., 2015)	Méthodes d'ensemble basées sur des arbres de décision (Breiman et al., 1984). Optimisent généralement l'erreur quadratique moyenne (MSE) ou l'erreur de classification pour prédire l'espérance conditionnelle $E[Y X]$. La forêt aléatoire est une méthode de <i>bagging</i> qui permet une estimation stable à partir d'estimateurs instables (les arbres de décision). XGBoost est une méthode de <i>boosting</i> qui corrige l'erreur séquentiellement de chaque estimateur instable.	<ul style="list-style-type: none"> — Très performants pour capturer des interactions non-linéaires complexes. — Robustes et faciles à implémenter. — Gestion simple des variables catégorielles et directionnelles. 	<ul style="list-style-type: none"> — Inadaptés aux extrêmes : Les prédictions se concentrent vers la moyenne et échouent à extrapoler vers les quantiles élevés.
6. Méthodes d'apprentissage automatique pour les extrêmes (méthodes à explorer / en cours d'exploration)			
Méthodologie	Principe	Avantages	Limites
Extremal random forest (ERF) (Gnecco et al., 2024)	Adaptation des forêts aléatoires pour estimer les paramètres (σ, ξ) d'une GPD conditionnelle locale. Utilise des poids de similarité dérivés de la forêt aléatoire pour maximiser une vraisemblance GPD locale pondérée.	<ul style="list-style-type: none"> — Combine la flexibilité des forêts aléatoires avec la rigueur d'extrapolation de la GPD. — Permet l'estimation de quantiles extrêmes en grande dimension de covariables. — Consistant théoriquement (sous conditions de domaine d'attraction). — Performance supérieure aux forêts quantiles classiques pour $\tau \approx 1$. 	<ul style="list-style-type: none"> — Coût computationnel plus élevé que les RF classiques (optimisation locale). — Nécessite un réglage fin des hyperparamètres (taille minimale des feuilles pour avoir assez d'excès locaux). — Mauvaises estimations si les données extrêmes sont trop éparées localement (malédiction de la dimensionnalité).

TABLE 3 – Comparaison des méthodes de Machine Learning : méthodes explorées (haut) vs en cours d'exploration (bas).

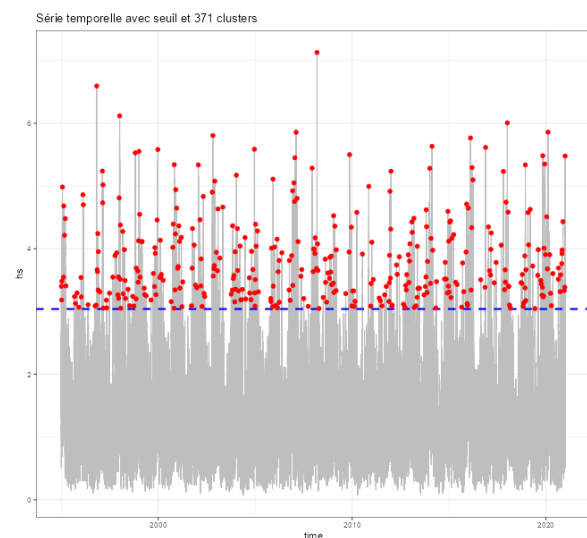
7. Modélisation des spectres (méthodes explorées)			
Méthodologie	Principe	Avantages	Limites
Spectre JONSWAP (Hasselmann et al., 1973)	Modélisation paramétrique empirique prolongeant le spectre de Pierson-Moskowitz (1964). Introduit un facteur de rehaussement du pic ($\gamma > 1$).	<ul style="list-style-type: none"> — Standard industriel (DNV). — Formule analytique simple dépendant de peu de paramètres (H_s, T_p, γ). — Bonne représentation des mers de vent en développement. 	<ul style="list-style-type: none"> — Incapable de modéliser les états de mer multimodaux (ex : houle lointaine + mer du vent locale), pourtant fréquents. — Surestimation de l'énergie au pic observée lors de tempêtes intenses (bien qu'ayant un spectre unimodal), même avec un γ minimal.

TABLE 4 – Méthode de modélisation spectrale explorée.

7.1 Exemple de visualisation de méthodes employées

7.1.1 Approches explicatives des extrêmes

Étape 1: Dépassements de seuil et clusters



Étape 2: Résultats du modèle GPD

Diagnostic du modèle

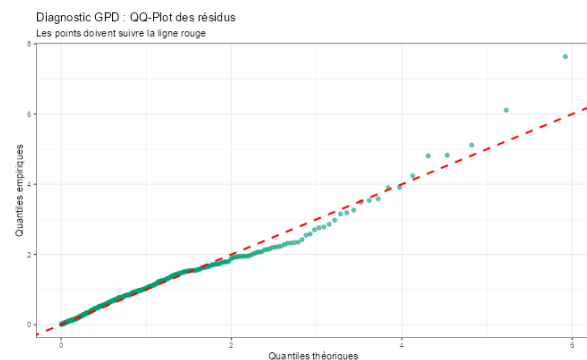


FIGURE 1 – Modèle GPD non stationnaire. Maximums de tempête (H_s) et diagnostic du modèle (quantile-quantile plot).

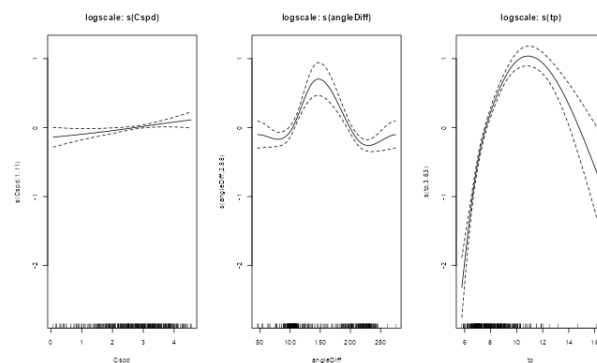


FIGURE 2 – Modèle GPD non stationnaire. Estimation de la valeur des paramètres de la GPD en fonction de de la vitesse du courant, de l'écart angulaire et de T_p .

7.1.2 Contours environnementaux

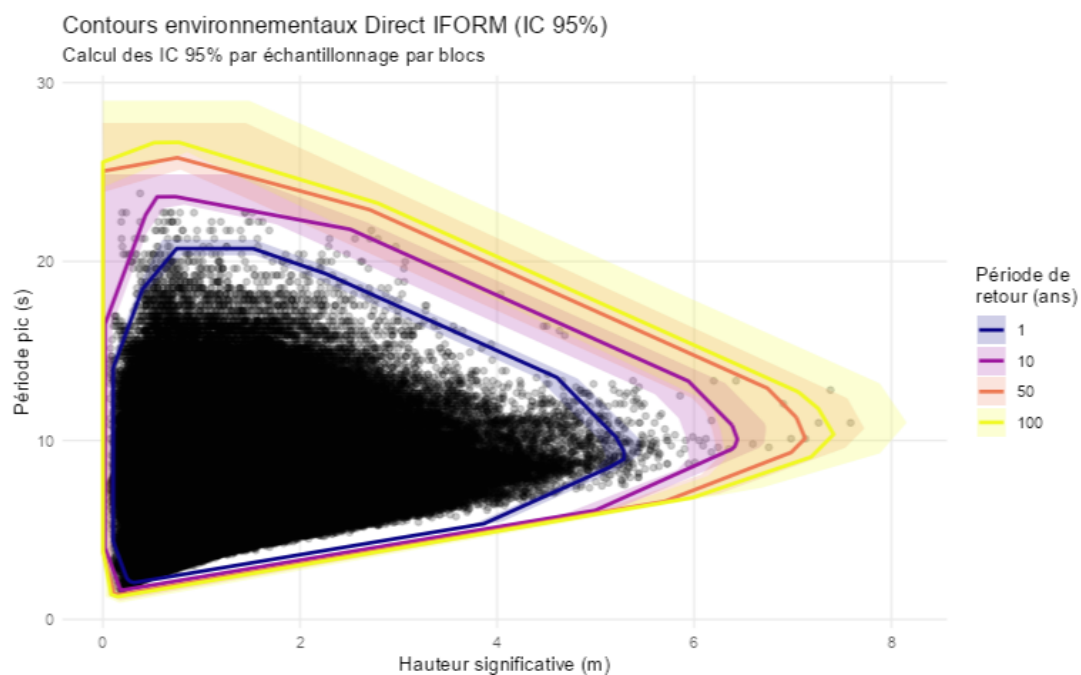


FIGURE 3 – Contours environnementaux H_s-T_p au Raz Blanchard (1994-2024) pour plusieurs périodes de retour, avec intervalles de confiance à 95% prenant en compte l'incertitude d'échantillonnage.

7.1.3 Approches prédictives : méthode d'apprentissage automatique

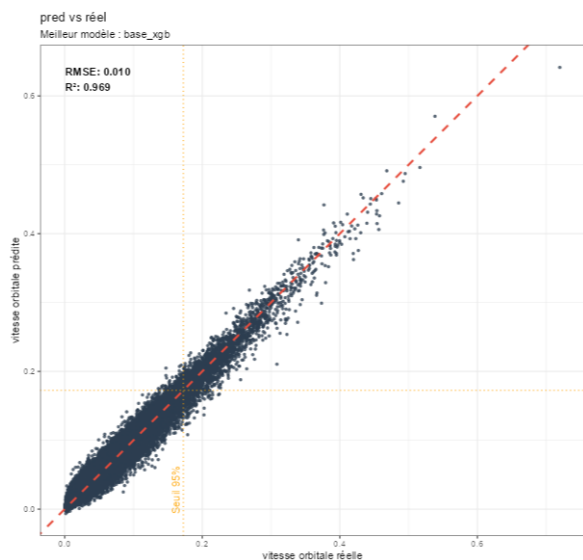


FIGURE 4 – Modèle XGBoost. Estimation contre prédictions de la vitesse orbitale.

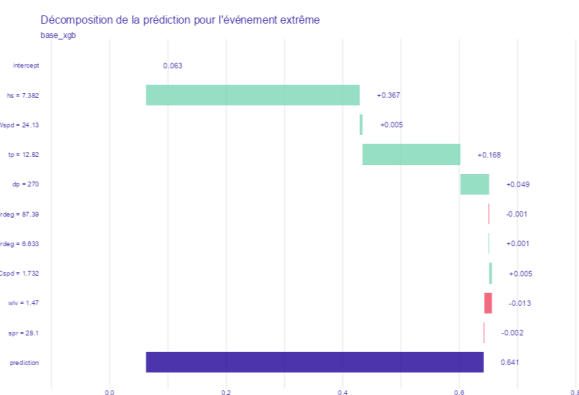


FIGURE 5 – Modèle XGBoost. Contribution des variables du modèle à la prédiction (0.64 m/s) pour la vitesse orbitale max du jeu de données (Ciaran, 0.72 m/s).