IRM | institut du risk management BY Institut.

Mémoire présenté pour la validation de la Formation
« Certificat d'Expertise Actuarielle »
de l'Institut du Risk Management
et l'admission à l'Institut des actuaires
le / /2022

Par : Benjamin Van Rij

Titre : Modélisation de l'influence du réchauffement climatique sur la fréquence des ouragans aux Etats-Unis

Confidentialité : ☒ NON ☐ OUI (Durée : ☐ 1an ☐ 2 ans)
Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus

Membres présents du jury de l'Institut des actuaires :

Membres présents du jury de l'Institut du Risk Management :

Secrétariat :

Bibliothèque :

Entreprise :

Nom : SCOR

Signature et Cachet :

SCOR SE
5, avenue Kléber
F - 75795 PARIS CEDEX 16
Tél. : + 33 (0)1 58 44 70 00

Directeur de mémoire en entreprise :

Nom : Sidney Rostan

Signature :

Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels
(après expiration de l'éventuel délai de confidentialité)

Signature du responsable entreprise

Signature(s) du candidat(s)

Remerciements

J’exprime tout d’abord ma profonde reconnaissance à mon épouse Mathilde, qui m’a apporté un soutien constant et infaillible tout au long de cette épreuve.

Ma gratitude va également à ma famille, ma belle-famille et mes amis pour leur soutien inconditionnel, leur patience et leur compréhension.

Enfin, je souhaite remercier mes collègues qui se sont intéressés au sujet et, pour le suivi académique, Olivier Lopez et Christophe Dutang pour les échanges constructifs que nous avons pu avoir à des points cruciaux d’avancement de mon mémoire, mais aussi Dr Jérôme G., Anas E., Maxime L. et Julien S. pour leur intérêt et leurs riches conseils en *machine learning* ainsi que Pierre P. pour son aide précieuse sur des points délicats du pré-traitement des données sous Python.

Résumé

La température globale à la surface de la terre a augmenté en moyenne de 1°C par rapport à l'ère préindustrielle (1850-1900) et les sept dernières années ont été les plus chaudes jamais enregistrées. La température moyenne globale en 2020 par exemple était de 1,2°C au-dessus du niveau de l'ère préindustrielle, alors que l'accord de Paris[1] prévoit de contenir cette augmentation « bien en dessous de 2°C » tout en poursuivant les efforts pour la contenir à 1,5°C.

Parallèlement, l'activité cyclonique semble s'intensifier, en particulier dans l'Atlantique Nord, menaçant les Etats-Unis, actuellement le plus gros marché d'assurance.

Peut-on pour autant en déduire que le réchauffement climatique augmente la fréquence des ouragans, ou celle des ouragans les plus intenses ? Nous avons proposé dans ce mémoire une modélisation de la relation entre températures de l'eau et nombre annuel d'ouragans et d'ouragans majeurs (catégorie 3 à 5) aux Etats-Unis, à partir des données historiques depuis 1850 et en utilisant à la fois des méthodes de régression classiques et des algorithmes d'apprentissage supervisé. Appliqués aux projections de températures du GIEC[2], ces modèles nous ont donné une prédiction du nombre annuel moyen d'ouragans et d'ouragans majeurs à horizon milieu et fin du XXIème siècle. Enfin, par une application actuarielle simple, nous en avons déduit une estimation de la sinistralité moyenne au niveau marché à ces deux horizons de projection.

Mots clés : réchauffement climatique, catastrophes naturelles, ouragans, Atlantique Nord, Etats-Unis, apprentissage supervisé

[1] COP *(Conference of the Parties)* 21, Paris, 2015

[2] GIEC (Groupe d'experts Intergouvernemental sur l'Evolution du Climat), sixième rapport d'évaluation (AR6), Groupe de travail I, 2021

Abstract

Global surface temperature has increased by an average of 1°C compared to the pre-industrial era (1850-1900) and the last seven years have been the warmest on record. The global average temperature in 2020, for example, was 1.2°C above the pre-industrial level, while the Paris Agreement[3] calls for containing this increase "well below 2°C" while continuing efforts to contain it within 1.5°C.

At the same time, hurricane activity seems to be intensifying, particularly in the North Atlantic, threatening the United States, currently the largest insurance market.

Can we conclude that climate change increases the frequency of hurricanes, or that of the most intense hurricanes? In this thesis, we proposed a model of the relationship between sea surface temperatures and the annual number of hurricanes and major hurricanes (category 3 to 5) in the United States, based on historical data since 1850 and using both classical regression methods and supervised learning algorithms. Applied to IPCC[4] temperature projections, these models gave us a prediction of the average annual number of hurricanes and major hurricanes in the mid to late 21st century. Finally, through a simple actuarial application, we provided an estimate of the average insured industry loss at these two projection horizons.

Keywords: climate change, natural catastrophes, hurricanes, North Atlantic, United States of America, supervised machine learning

[3] COP (Conference of the Parties) 21, Paris, 2015

[4] IPCC (Intergovernmental Panel on Climate Change), sixth Assessment Report (AR6), Working Group I, 2021

Sommaire

Abstract 4
Introduction 7
1. Ouragans et changement climatique 9
1.1. Formation des ouragans 9
1.2. Réchauffement climatique 12
1.3. Données et approche 14
1.4. Prédiction du modèle sur des projections 23
1.5. Limites 25
2. Modélisation du nombre annuel d'ouragans 27
2.1. Etude préliminaire exploratoire 27
2.2. Modélisation 35
2.3. Prédiction future 55
3. Application actuarielle simplifiée à l'impact sur les pertes assurées du marché 68
3.1. Approche méthodologique 68
3.2. Estimation dans le scénario le plus émetteur SSP5-8.5 69
3.3. Estimation dans le scénario intermédiaire SSP2-4.5 70
3.4. Incertitude et erreur stochastique 71
3.5. Sensibilité et élasticité 72
3.6. Résultats et intervalles de confiance 73
Conclusion 75
Bibliographie 78
Annexes 80
Annexe 1 : coefficients des modèles GLM testés pour la modélisation du nombre annuel d'ouragans 80

Introduction

La température globale à la surface de la terre a augmenté en moyenne de 1°C par rapport à l'ère préindustrielle (1850-1900) et les sept années de 2015 à 2021 inclus ont été les plus chaudes jamais enregistrées. La température moyenne globale en 2020 par exemple était de 14,9°C soit 1,2 (± 0,1) °C au-dessus du niveau de l'ère préindustrielle alors que l'accord de Paris (COP21, 2015) prévoit de contenir cette augmentation « bien en dessous de 2°C » tout en poursuivant les efforts pour la contenir à 1,5°C. Les caractéristiques remarquables de l'année 2020 ont été une chaleur persistante et des feux de forêt en Sibérie, la faible étendue de la banquise arctique et une saison cyclonique record dans l'Atlantique avec 31 tempêtes tropicales au total - contre 11 en moyenne depuis 1950 - dont 13 ont atteint le stade d'ouragan (vitesse de vent supérieure à 119 km/h) – contre 6 en moyenne (depuis 1950) - et 6 d'ouragan majeur (vitesse de vent supérieure à 252 km/h) - contre 3 en moyenne (depuis 1950). Pour la seconde fois dans l'Histoire seulement, il a été nécessaire d'utiliser l'alphabet grec pour nommer les derniers ouragans, notre alphabet s'arrêtant à 26 lettres. Cette année a aussi été la cinquième année consécutive avec une activité cyclonique au-dessus de la moyenne. Malheureusement, une année de bien tristes records donc. Selon Météo France (Méteo France, 2022), l'année 2021 a certes globalement été moins chaude que 2020, mais serait malgré tout la sixième année la plus chaude au niveau mondial et elle marque un record notable, avec l'été européen le plus chaud jamais enregistré.

Le réchauffement climatique est un des enjeux majeurs de notre génération, et de la prochaine, avec des effets déjà visibles à différents niveaux (air, océan, biodiversité…) et pour certains irréversibles, par exemple la fonte des glaces et en conséquence l'augmentation du niveau de la mer. S'il est établi qu'il aggrave la fréquence et l'intensité de certaines catastrophes naturelles comme les épisodes de sécheresse, d'inondation ou encore les feux de forêt, son influence sur l'activité cyclonique fait moins consensus. Les cyclones tropicaux dans l'Atlantique Nord constituent pour la plupart des acteurs internationaux du marché de l'assurance et de la réassurance l'un des risques majeurs, si ce n'est le premier, dans leur distribution de perte, à la fois en moyenne et dans les scénarios extrêmes. A lui seul, ce risque peut représenter jusqu'à 70% des pertes assurantielles mondiales, comme cela a été le cas lors de l'année record de 2017 avec 135 milliards de dollars de pertes assurantielles dont 92 milliards attribuables aux ouragans dévastateurs Harvey, Irma et Maria d'après le service *NatCatSERVICE* du réassureur Munich Re, dédié notamment au suivi des événements marché (MunichRe, 2018). Sur l'année 2021, d'après ce même service, les ouragans américains ont causé environ 40 milliards de dollars de pertes assurées, comptant pour environ un tiers des pertes assurées mondiales.

La forte contribution des cyclones tropicaux de l'Atlantique Nord dans le total tient au fait qu'il s'agit d'un péril à la fois fréquent (plusieurs occurrences par an), relativement dévastateur et qui touche les Etats-Unis, de loin le plus gros marché d'assurance au monde avec près de 2000 milliards de dollars de primes émises en assurance non-vie d'après l'institut de recherche *Sigma* du réassureur mondial Swiss Re (SwissRe, 2021). Il s'agit du plus gros marché d'assurance du monde car c'est à la fois une région très peuplée, avec plus de 331 millions d'habitants recensés en 2020, et avec la pénétration d'assurance la plus forte au monde en non-vie.

Pour la première fois dans son dernier rapport, le *Groupe d'experts intergouvernemental sur l'évolution du climat* (GIEC, Sixième rapport d'évaluation (AR6), 2021) mentionne le fait que l'activité cyclonique est accrue par des températures plus chaudes, ce qui semble assez logique au vu du phénomène naturel que nous décrivons en détail plus loin. L'augmentation des précipitations et du risque de submersion marine (*storm surge*) accompagnant les tempêtes tropicales, notamment du fait de la montée du niveau des océans, est quant à elle établie avec un niveau de confiance élevé. Toutefois, il ne ressort pas d'estimation claire des effets sur la fréquence et/ou la sévérité des cyclones, et encore moins sur l'impact en termes de pertes économiques et assurantielles. Leur évaluation est en effet relativement complexe car elle dépend, en plus des caractéristiques physiques des cyclones (trajectoire, vitesse de déplacement, précipitations qui l'accompagnent, *storm surge*…) de leur probabilité de toucher la terre – on utilisera dans la suite principalement le terme anglais *landfall* très utilisé dans le domaine – et le cas échéant de la localisation de ce *landfall*. Aussi, l'évolution des pertes assurées est-elle fonction de

l'évolution du montant et/ou de la répartition géographique des valeurs assurées ou encore de la présence éventuelle de moyens de protection, tels que des digues.

C'est dans ce contexte, que nous nous proposons de mettre en évidence et quantifier l'impact du réchauffement climatique sur la fréquence des ouragans dans le bassin Atlantique Nord affectant notamment les Etats Unis, à l'aide de méthodes statistiques et d'apprentissage supervisé (*supervised machine learning*). Notre objectif est d'obtenir une estimation du nombre annuel d'ouragans dans un monde plus chaud, en milieu et fin du XXIème siècle, sous un scénario défavorable. Nous proposerons également une application actuarielle simplifiée à l'estimation des pertes assurées qui en découlent.

Le bassin cyclonique de l'Atlantique Nord s'étend en longitude du Golfe du Mexique à l'Afrique de l'Ouest et inclut les Caraïbes, les Bahamas et la côte sud et est des Etats-Unis du Texas au Maine.

Cette problématique s'inscrit donc dans un contexte plus général d'une prise de conscience accrue et d'une volonté se faisant de plus en plus pressante de tous les acteurs – individus, Etats et entreprises – de mieux appréhender le réchauffement climatique et en évaluer ses impacts sur leur activité, à court et long terme, afin notamment de le réduire au maximum et de s'adapter à ses conséquences.

Pour les assureurs et réassureurs, il s'agit en particulier de comprendre et quantifier l'influence du réchauffement climatique sur la survenance des catastrophes naturelles, afin de pouvoir mieux anticiper et intégrer ces effets dans leurs modèles et ainsi avoir une tarification, un provisionnement et un calcul de capital économique adéquats. De plus, ces acteurs ont également un rôle à jouer en termes de prévention et d'action, par la sensibilisation de leurs assurés, le soutien au monde de la recherche sur la compréhension des risques climatiques ou encore la mise en place d'une politique durable et responsable en termes de souscription et d'investissement. Un assureur peut par exemple décider d'exclure les secteurs d'activité les plus émetteurs de gaz à effet de serre de sa souscription et/ou de ces investissements.

Nous choisissons ici de nous concentrer sur les ouragans de l'Atlantique Nord, qui comme nous l'avons indiqué, constituent l'un des risques de catastrophes naturelles principaux du marché mondial de l'assurance et de la réassurance. L'originalité de cette étude réside en le fait de modéliser exclusivement la relation entre les températures à la surface de l'eau et la fréquence annuelle d'ouragans, en distinguant les ouragans majeurs, avec notamment une approche *machine learning*. Elle n'a pas vocation principale à prédire de façon fiable l'activité cyclonique année par année mais plutôt de fournir une estimation adaptée de la tendance moyenne sur une période longue d'au moins dix ans, afin de pouvoir disposer d'une projection du nombre annuel moyen d'ouragans à horizon 2041-2060 et 2081-2100, qui sont les deux périodes de projection principales retenues par le GIEC.

1. Ouragans et changement climatique

1.1. Formation des ouragans

Les cyclones, appelés ouragans dans l'Atlantique Nord et typhons en Asie, sont des systèmes de vents rotatifs se formant dans différentes zones tropicales du globe, à certaines périodes de l'année (phénomène saisonnier) où des conditions thermodynamiques précises sont vérifiées. D'après Météo France (Météo France, 2020) et le chapitre *Hurricane Dynamics* de l'article (Asthana, Krim, Sun, Roheda, & Xie, 2021), ces conditions sont les suivantes.

Il doit tout d'abord préexister une perturbation atmosphérique basse-pression.

Ensuite, la force de Coriolis doit être suffisante pour mettre en rotation les vents, ce qui implique en général d'être à au moins 10° de l'équateur, la force de Coriolis étant nulle à l'équateur.

En outre, la température à la surface de la mer doit être relativement élevée, généralement supérieure à 80°F (26,5°C) sur au moins 50 mètres de profondeur. C'est véritablement la chaleur de l'eau qui donne son énergie au cyclone.

Dans le même temps, l'humidité absolue doit être suffisamment élevée dans la basse et moyenne troposphère. La troposphère est la couche la plus basse de l'atmosphère, dont la limite varie avec la latitude et la saison entre 8 et 15 km d'altitude.

Enfin, les ouragans se forment uniquement en l'absence de, ou avec un faible, cisaillement vertical des vents (*Vertical Wind Shear*). Le changement des vents avec l'altitude, tant en direction qu'en vitesse, doit être de moins de 10 m/s (soit 36 km/h) entre la surface et la tropopause. La tropopause est la zone de l'atmosphère terrestre qui fait la transition entre la troposphère (au-dessous) et la stratosphère (au-dessus). Son altitude diminue avec la latitude, depuis environ 17 km à l'équateur jusqu'à environ 9 km aux pôles, et varie également selon les saisons.

Ces phénomènes tourbillonnaires, de pression centrale très basse, tournent dans le sens des aiguilles d'une montre dans l'hémisphère sud et dans le sens contraire dans l'hémisphère nord, sous l'action de la force de Coriolis. Ils s'étendent sur des diamètres de 500 km à plus de 1000 km et leur centre, appelé œil du cyclone, est bien visible sur les images satellitaires. D'un diamètre généralement de 30 à 60 km, parfois jusqu'à 150 km, cet œil est une zone d'accalmie (pas de pluie, vent faible).

Les cyclones provoquent également deux phénomènes maritimes très dangereux. Le premier est une houle longue générée par le vent et qui se déplace plus rapidement que le cyclone ; elle est parfois observée jusqu'à 1 000 km à l'avant du cyclone.

Le second, plus connu chez les assureurs, consiste en une surélévation anormale du niveau de la mer, connue sous le nom d'*onde de tempête*, *onde de marée* ou *marée de tempête* ou *storm surge* en anglais. Il s'agit parfois du phénomène le plus meurtrier associé au passage d'un cyclone.

Si la vitesse des vents en rotation autour du centre très élevée, le cyclone ne se déplace quant à lui qu'à une vitesse de 20 ou 30 km/h en moyenne, comme illustré notamment dans le schéma ci-dessous qui récapitule la structure d'un cyclone.

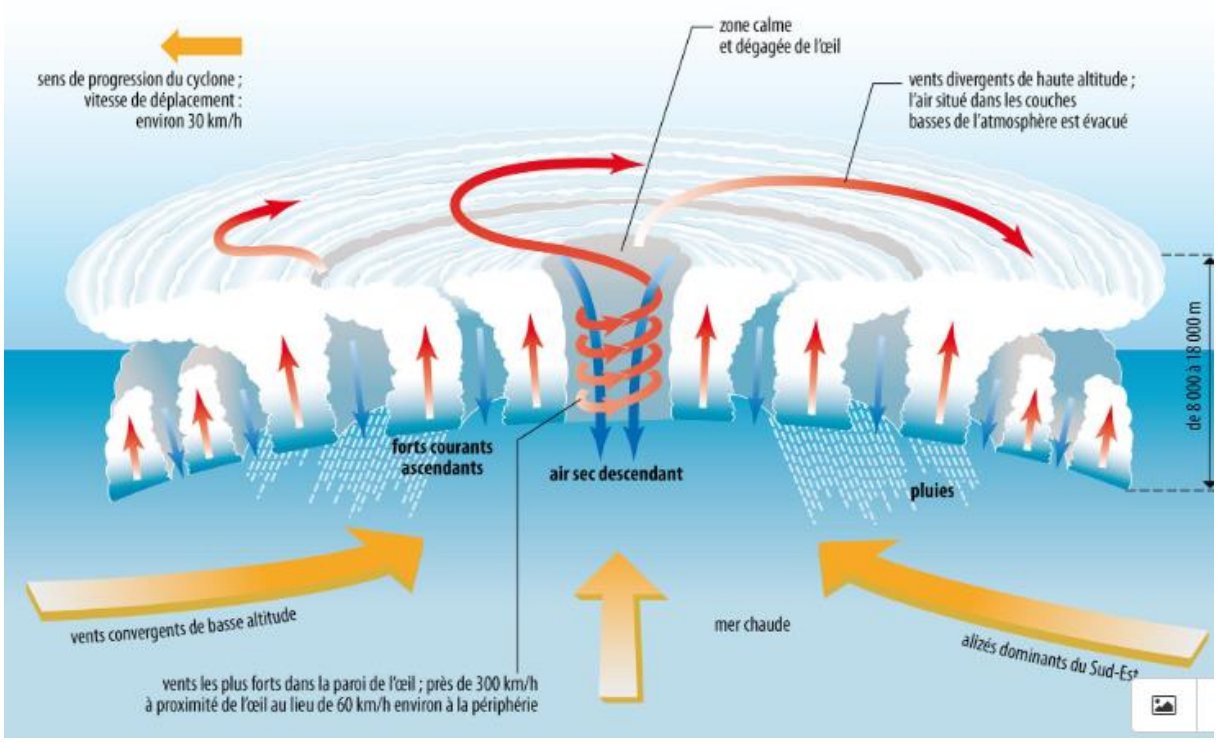

FIGURE 1 : STRUCTURE D'UN CYCLONE (SOURCE : ENCYCLOPEDIA UNIVERSALIS)

On note chaque année entre 80 et 90 cyclones tropicaux en moyenne à travers le monde, dont près de la moitié atteignent le niveau d'ouragans (ou typhons) et environ un quart celui d'ouragans majeurs, c'est-à-dire de catégorie 3 ou plus sur l'échelle de Saffir-Simpson décrite ci-dessous.

Saffir–Simpson scale

Category	Wind speeds			
	m/s	knots (kn)	mph	km/h
Five	≥ 70 m/s	≥ 137 kn	≥ 157 mph	≥ 252 km/h
Four	58–70 m/s	113–136 kn	130–156 mph	209–251 km/h
Three	50–58 m/s	96–112 kn	111–129 mph	178–208 km/h
Two	43–49 m/s	83–95 kn	96–110 mph	154–177 km/h
One	33–42 m/s	64–82 kn	74–95 mph	119–153 km/h

Related classifications

Tropical storm	18–32 m/s	34–63 kn	39–73 mph	63–118 km/h
Tropical depression	≤ 17 m/s	≤ 33 kn	≤ 38 mph	≤ 62 km/h

FIGURE 2 : ECHELLE DE SAFFIR-SIMPSON

A noter que cette échelle est fondée sur la vitesse de vents maximale mesurée sur une minute (*1-minute maximum sustained winds*).

Dans la suite et par simplification, nous appellerons simplement *tempête* toute système tropical dont les vents atteignent au moins 63km/h, qu'elle atteigne ou nom le stade d'ouragan (vitesse de vents supérieure ou égale à 119km/h). Il s'agira donc de l'ensemble le plus large considéré dans cette étude puisque nous ne prenons pas en compte les systèmes plus petits restant au stade de dépressions tropicales.

Dans l'Atlantique Nord, la saison s'étend officiellement du 1er juin au 30 novembre bien que les tempêtes se forment majoritairement entre fin juillet et mi-octobre, quand les températures de l'eau atteignent leur pic. La principale zone de genèse se situe dans le rectangle 10°N–20°N, 15°W–75°W entre le Cap Vert et les Caraïbes. Elle est appelée *Atlantic main development region (AMDR)*.

Néanmoins, les tempêtes peuvent aussi se former, ou s'intensifier plus à l'ouest dans la mer des Caraïbes et le Golfe du Mexique.

A titre d'illustration, nous proposons une visualisation de l'AMDR sur la carte ci-dessous qui décrit les anomalies de températures à la surface de l'eau, i.e. les écarts à la climatologie 1981-2010, dans l'Atlantique Nord le 12 août 2019.

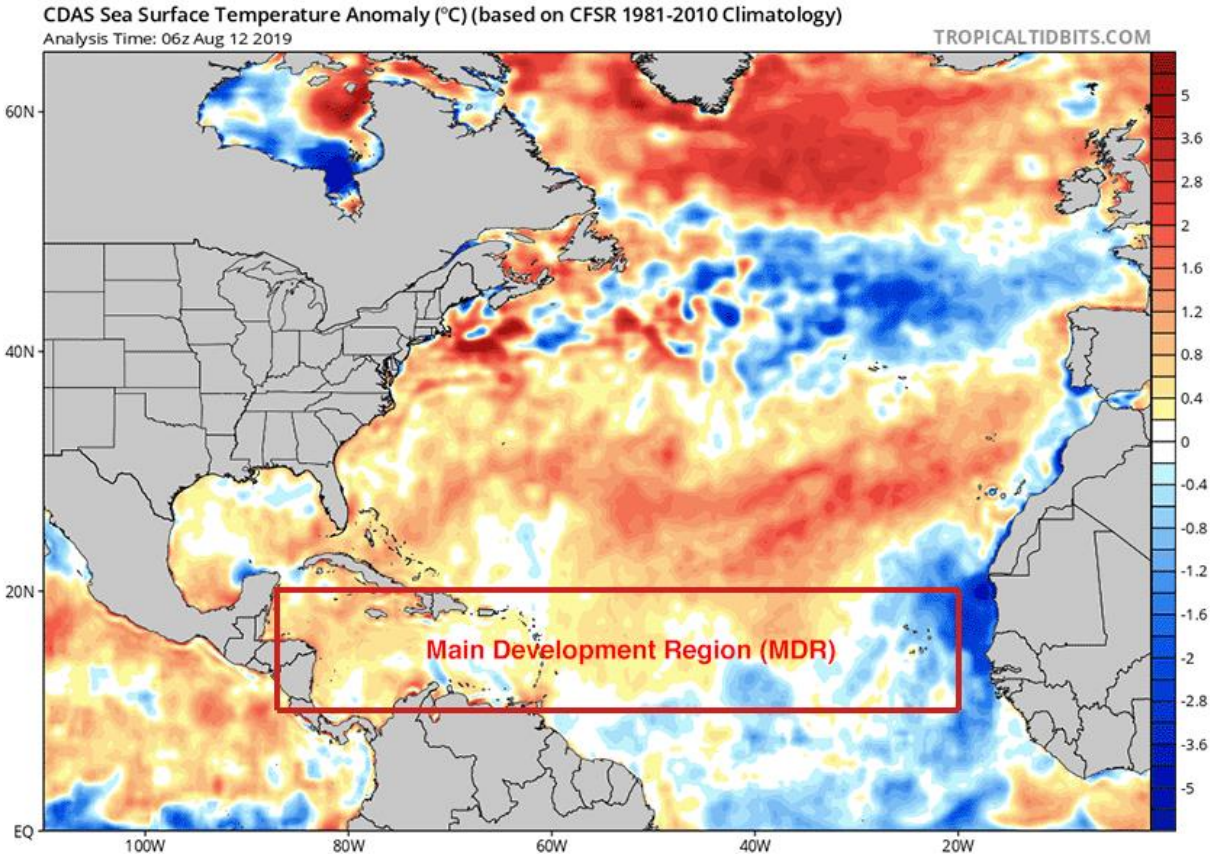

FIGURE 3 : ANOMALIES DES TEMPERATURES A LA SURFACE DE L'EAU DANS L'ATLANTIQUE NORD LE 12 AOUT 2019

1.2. Réchauffement climatique

Comme le rappelle dernier rapport du GIEC (GIEC, Sixième rapport d'évaluation (AR6), 2021), l'influence des êtres humains sur le réchauffement de l'atmosphère, des océans et des continents est sans équivoque. En émettant des gaz à effet de serre (GES), l'humanité a provoqué des changements rapides et étendus au niveau de l'atmosphère, de la cryosphère (glaces terrestres et marines), de la biosphère (les êtres vivants) et des océans.

Les principaux GES émis par les êtres humains sont le dioxyde de carbone (CO2), le méthane (CH4), et le protoxyde d'azote (N2O). Une partie des émissions humaines sont captées par l'océan et à la surface des continents (56% pour le CO2). Le reste est stocké dans l'atmosphère, augmentant leur concentration. Cette concentration s'exprime en partie par million (ppm) ou en partie par milliard (ppb pour billion qui est la traduction de milliard en anglais).

De 1750 à 2019, les concentrations atmosphériques sont passées d'environ 280 à 410 (+46%) ppm pour le CO_2, d'environ 800 à 1866 (+133%) ppb pour le CH_4, et d'environ 270 à 332 (+23%) ppb pour le N_2O. Bien que la concentration atmosphérique en CH_4 soit environ 220 fois plus faible que celle du CO_2, le CH_4 est responsable de plus d'un quart du réchauffement en raison de son pouvoir réchauffant plus fort. L'amplitude de ces variations pour le CO_2 et le CH_4 dépasse de loin celle des variations naturelles entre périodes glaciaires et périodes interglaciaires au cours des 800 000 dernières années. La concentration actuelle en CO_2 est la plus haute depuis au moins 2 millions d'années.

Comme l'explique le résumé pour les décideurs, cette augmentation de la concentration en GES est à l'origine d'une accumulation importante d'énergie sous forme de chaleur à la surface du globe. Cette énergie se retrouve à 91% dans les océans, 5% sur les surfaces continentales, 3% dans les glaces (causant leur fonte), et 1% seulement dans l'atmosphère. Pourtant, cette infime proportion (1%) de chaleur piégée dans l'atmosphère est la principale cause du réchauffement observé depuis 1850 : la température mondiale de surface sur la période 2011-2020 était 1,09°C plus chaude que celle sur la période 1850-1900, avec un réchauffement plus important sur les continents (+1,59°C) qu'au-dessus des océans (+0,88°C). Le réchauffement de la température de surface mondiale directement imputable aux activités humaines est estimé à 1,07°C, soit la quasi-intégralité du réchauffement observé. Le réchauffement s'est également accéléré, avec +0,9°C entre la période 2003-2012 et la période 2011-2020. Le rythme du réchauffement sur les 50 dernières années est sans précédent depuis au moins 2000 ans. Les températures actuelles dépassent le maximum de la dernière période chaude datant d'il y a 65 00 ans, et il est probable qu'elles dépassent également le précédent maximum datant d'il y a 125 000 ans.

Le graphique suivant donne la température observée par la NASA par rapport à la moyenne de 1850-1900 comme référence préindustrielle.

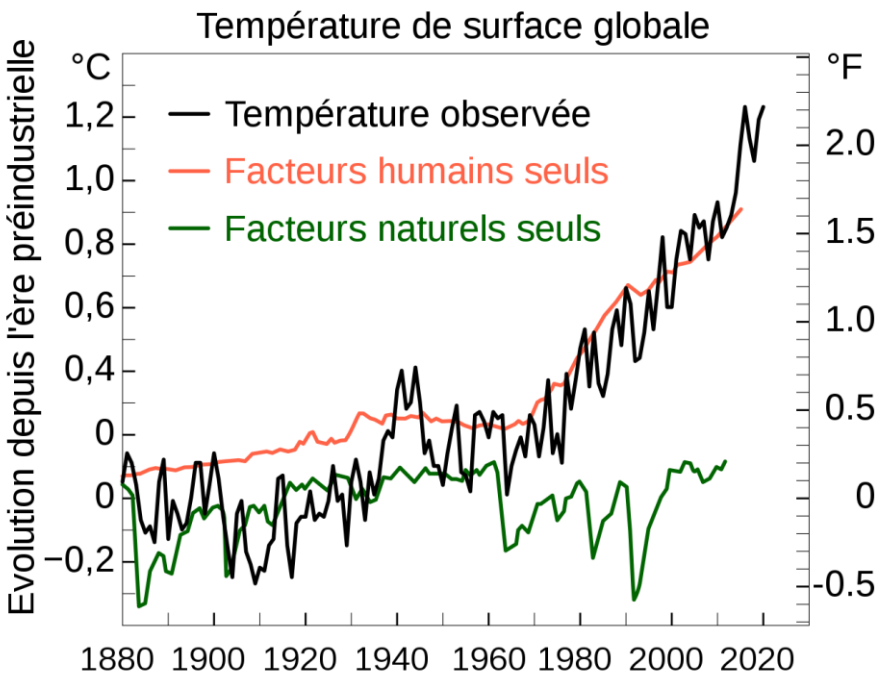

FIGURE 4 : EVOLUTION DE LA TEMPERATURE GLOBALE A LA SURFACE DE LA TERRE PAR RAPPORT A LA PERIODE PREINDUSTRIELLE 1850-1900 (SOURCE : NASA)

Il permet de mettre en évidence que le principal facteur d'augmentation des températures mondiales à l'ère industrielle est l'activité humaine, les forces naturelles ajoutant de la variabilité, et de plus on y voit nettement l'accélération du réchauffement dans les 50 dernières années.

Le réchauffement climatique anthropogène est donc avéré avec des conséquences déjà visibles et pour certaines « irréversibles pour des siècles et des millénaires » d'après le GIEC.

En perturbant le climat, l'humanité a également provoqué des changements dans la distribution de fréquence des événements météorologiques et climatiques extrêmes. En particulier, il est établi avec une probabilité très forte que le réchauffement climatique augmente la fréquence et l'intensité des épisodes de sécheresse, de vagues de chaleur et de fortes précipitations. Il n'y avait jusque-là pas suffisamment de consensus international de la communauté scientifique à propos d'une potentielle influence sur l'activité cyclonique. Mais, pour la première fois, dans son sixième rapport (GIEC, Sixième rapport d'évaluation (AR6), 2021), le GIEC mentionne qu'il est probable que la proportion mondiale de cyclones tropicaux majeurs (catégorie 3 à 5) ait augmenté au cours des quatre dernières décennies. C'est la première fois que le GIEC décrit une augmentation observée des événements extrêmes pour les ouragans. Plus généralement, c'est cette relation entre augmentation des températures et survenance des ouragans que nous allons étudier et tenter de modéliser dans cette étude. Nous nous limitons ici au bassin Atlantique Nord par souci de simplification et de matérialité pour le marché de l'assurance, mais la méthodologie suivie est réplicable et une étude similaire pourrait être conduite sur un autre bassin, par exemple le Pacifique Est concernant les pays d'Asie Pacifique ou le Pacifique Ouest concernant la côte ouest des Etats-Unis et du Mexique.

1.3. Données et approche

1.3.1. Variable à prédire : le nombre d'ouragans

La variable que nous nous proposons de prédire est le nombre annuel d'ouragans se formant dans le bassin Atlantique Nord. Nous disposons de la base de données *HURDAT2* du *National Oceanic and Atmospheric Administration* (NOAA) *National Hurricane Center* (NHC), c'est-à-dire le centre d'étude et de prévisions des ouragans qui est hébergé au sein de l'agence météorologique américaine, le *National Weather Service* qui est lui-même rattaché à l'Organisation Météorologique Mondiale (OMM), en anglais *World Meteorological Organization* (WMO). Nous utilisons ici exclusivement la version *Atlantic* de cette base HURDAT2, mais il existe aussi une version *Pacific* pour les tempêtes touchant la côte Pacifique des Etats-Unis et du Mexique i.e. le bassin Pacifique Nord Est.

La base *HURDAT2 Atlantic* est la base officielle répertoriant l'ensemble des systèmes tropicaux dans l'Atlantique (océan Atlantique, mer des Caraïbes et Golfe du Mexique) depuis 1851, avec leurs positions toutes les six heures accompagnées d'un relevé de vitesse de vents et de pression centrale. Elle est disponible librement au format texte (*.txt*) sur le site de la NOAA-NHC (Best Track Data (HURDAT2)). Pour notre étude, nous avons considéré l'extrait de la base au 10 juin 2021 contenant les données les plus à jour pour des événements jusqu'à fin 2020.

Au sein de la NHC, les équipes de la *Hurricane Research Division* (HRD) effectuent un travail continu de « réanalyse » des événements passés afin de corriger certaines données, par exemple de trajectoire ou d'intensité d'ouragan, voire d'ajouter ou de supprimer certaines observations.

Il pourrait être intéressant d'expliquer et prédire, plutôt que le nombre annuel d'ouragans, l'énergie cumulative des cyclones tropicaux, connue sous son acronyme anglais *ACE (Accumulated Cyclone Energy)* sur une saison qui correspond à la quantité d'énergie globale des différents cyclones estimée à partir de la vitesse maximale (au carré) des vents pour chaque période de six heures. Comme les vents varient dans le temps selon l'évolution de l'intensité du cyclone, l'ACE est donc la somme sur chaque période de six heures des vitesses de vent maximales v_{max} (exprimées en kn) élevées au carré, ce qui s'écrit :

$$ACE = 10^{-4} \sum {v_{max}}^2.$$

Elle est donc d'autant plus grande que les vents sont puissants et que la durée de vie de l'ouragan est grande.

Cette grandeur peut avoir l'avantage d'être continue, à l'inverse du nombre d'ouragans qui est discret, et permet de quantifier de façon assez complète l'activité cyclonique d'un bassin. Toutefois, elle est plus difficile à relier directement à un dommage ou une perte assurantielle, notamment car elle ne tient pas compte du fait que l'ouragan ait touché terre ou non, ni à quel endroit (sur une zone densément peuplée ou non). En outre, nous noterons que, d'après sa définition, elle n'est calculée qu'à partir du moment où le système atteint le niveau de tempête tropicale et ne tient donc pas compte des dépressions tropicales plus faibles et souvent de courte durée de vie.

Il existe une limite importante inhérente aux données de la base *HURDAT2*. Il est très probable que les données de tempêtes ne soient pas aussi fiables et exhaustives avant 1900, surtout pour les « petites » tempêtes et celles qui n'ont pas touché terre, comme le mentionne le département HRD au sein de la NHC sur sa page dédiée à la réanalyse des données historiques (HRD, HURDAT, 2021). Cela est dû à la faible densité de population aux Etats Unis avant 1900 notamment dans certaines régions costales. En effet, d'après (Wikipedia, 2021), le pays ne comptait que 23 millions d'habitants en 1850 et 76 millions en 1900 contre 151 millions en 1950 et 331 millions en 2020.

Même au-delà, jusque dans les années 1960 environ, la faible densité des bateaux dans l'Atlantique n'aurait pas permis un relevé exhaustif des tempêtes et de leurs positions. La HRD dans sa note (HRD, HURDAT2 format, 2019) détaille la qualité de chaque donnée physique collectée sur l'historique et précise que l'exhaustivité du comptage et la fiabilité des mesures ne sont réellement

satisfaisantes qu'à partir du début de la reconnaissance par avion (1944 pour la partie ouest du bassin) et par satellite (1966).

Néanmoins, on réduit considérablement ce biais d'exhaustivité si l'on s'intéresse uniquement au nombre d'ouragans (catégorie 1 minimum) ayant touché terre aux Etats-Unis, qui ont une grande probabilité d'avoir été tous rapportés par les populations et/ou les bateaux. En revanche, on imagine aisément qu'une tempête tropicale ou un ouragan (même potentiellement un ouragan majeur) qui serait resté dans l'océan, loin des côtes, n'ait pas été relevé dans la deuxième moitié du XIXème siècle et même potentiellement jusqu'en 1944, et le début de la reconnaissance par avion, voire 1966, début de l'observation par satellite.

L'agence américaine de la protection de l'environnement (*Environmental Protection Agency* (*EPA*)) a réalisé une analyse de l'évolution de la fréquence, de l'intensité et la durée des ouragans dans l'Atlantique Nord entre 1878 et 2020 qui semble corroborer cette hypothèse (EPA, 2021). En effet, on y constate qu'un ajustement correctif à la hausse a été appliqué sur les nombres annuels totaux d'ouragans avant 1966 mais qu'aucun ajustement n'a été réalisé sur le nombre d'ouragans ayant touché terre aux Etats-Unis.

Ce biais de qualité de données et d'autant plus réduit qu'on ne s'intéresse qu'au comptage annuel et non en détail à tous les relevés de position et données de mesure physiques (e.g. pression centrale) à intervalle de six heures. Ainsi, nous pouvons raisonnablement considérer que la qualité de données de notre variable à prédire est homogène sur l'historique.

Nous choisissons donc de travailler sur le nombre d'ouragans. La base HURDAT2 ne donne pas directement la catégorie sur l'échelle de Saffir-Simpson. Par conséquent, nous définissions pour chaque tempête sa catégorie à partir de la mesure maximale de vitesse des vents sur une minute atteinte par la tempête au cours de sa trajectoire. Dans la mesure où nous souhaitons *in fine* quantifier l'effet du réchauffement climatique en termes de pertes assurantielles et en particulier aux Etats-Unis, nous devons nous intéresser aux ouragans qui ont fait *landfall* et plus précisément qui ont fait *landfall* sur les côtes des Etats-Unis. Ainsi, nous capterons également, de façon implicite, un éventuel effet du réchauffement climatique sur les trajectoires d'ouragans. En effet, si l'augmentation des températures a tendance par exemple à augmenter le nombre moyen d'ouragans mais dans le même temps si ces ouragans se forment en moyenne plus à l'est dans l'océan Atlantique, ou s'incurvent plus rapidement vers l'est au cours de leur trajectoire dans l'Atlantique Nord, nous pourrions constater *in fine* une réduction du nombre d'ouragans atteignant les côtes américaines, i.e. une baisse du nombre de *landfalls* aux Etats-Unis.

Deux voies s'offrent à nous pour capturer correctement l'information de *landfall* dans notre modélisation.

La première consiste à filtrer la base de données HURDAT2 sur les ouragans qui ont touché terre aux Etats-Unis et d'entraîner et calibrer nos modèles sur cette base filtrée.

La seconde option consiste à considérer l'ensemble des ouragans présents dans la base HURDAT2 et de modéliser par ailleurs la probabilité moyenne (puisque nous ne nous intéressons qu'à l'effet sur la fréquence annuelle moyenne) pour un ouragan de toucher terre aux Etats-Unis. Il suffit alors d'appliquer cette probabilité à la prédiction de nombre d'ouragans faite par notre modèle calibré sur la base de données totale. Cette modélisation du taux de *landfall* peut se faire de façon déterministe en modélisant la probabilité moyenne de *landfall* observée sur l'historique, en la distinguant éventuellement par catégorie d'ouragans, ou bien de façon stochastique en calibrant une loi de probabilité (e.g. Poisson, Binomiale Négative) à partir de la moyenne et de la variance empiriques. Dans tous les cas, la limite de cette seconde méthode, que ce soit dans sa version déterministe ou stochastique, est qu'elle ne permettrait pas de capter un éventuel effet du réchauffement climatique sur la probabilité de *landfall* aux Etats-Unis. Nous retiendrons donc la première option.

La base *HURDAT2* ne donne pas directement le pays mais les coordonnées du ou des *landfall*(s), le cas échéant. Nous avons donc procédé à un géocodage inverse (*reverse geocoding*) afin d'associer aux différents couples de coordonnées (longitude, latitude) un pays. Pour ce faire, nous avons utilisé la librairie Python (Van Rossum, 2009) *geopy* (geopy, s.d.) et en particulier la fonction *Nominatim* qui permet de faire du géocodage, ou du géocodage inverse, à l'aide d'une *API* (*Application Programming*

Interface) communicant avec le site *OpenStreetMap*. Sur les 1645 tempêtes de la base, 648 ont touché terre (au moins une fois) dont 477 aux Etats-Unis. Sur les 648 tempêtes ayant touché terre, 204 ont touché terre plusieurs fois au cours de leur trajectoire, potentiellement dans des pays différents. On parle de *multiple landfall*. Nous ne devons conserver qu'une seule catégorie et qu'une seule localisation (pays) afin de ne les compter qu'une fois dans notre base d'apprentissage. Nous avons fait le choix de retenir la catégorie et la localisation du *landfall* de catégorie Saffir-Simpson la plus élevée.

Enfin, pour les besoins de l'application de notre modèle à l'évaluation des pertes assurantielles, nous distinguerons pour la modélisation, et donc pour la prédiction, deux ensembles d'ouragans : un premier constitué de la totalité des ouragans ayant touché terre aux Etats-Unis et un second constitué uniquement des ouragans majeurs (catégorie 3 à 5) ayant touché terre aux Etats-Unis. En effet, cette distinction est nécessaire pour notre application finale à l'estimation des pertes assurantielles pour le marché, où nous aurons besoin d'associer à chaque ouragan un coût pour le marché d'après le modèle RMS (Risk Management Models, Analytics, Software & Services, s.d.). Or, le coût individuel d'un ouragan est très différent selon son intensité, et donc sa catégorie sur l'échelle Saffir-Simpson.
En outre, établir une prédiction distinguant le nombre d'ouragans majeurs permet également de capter, au moins en partie, l'effet du réchauffement climatique sur l'intensité de l'*alea* naturel ou en tout cas sur la proportion d'ouragans majeurs.

1.3.2. Variables explicatives : les températures à la surface de l'eau

Nous utiliserons comme seules données explicatives les températures historiques à la surface de l'eau, en anglais *Sea Surface Temperatures (SST)*, fournies dans la base *HadSST4* du *Met Office*, le service national britannique de météorologie. Cette base contient les températures mensuelles depuis janvier 1850 dans le monde entier avec une résolution géographique de 5° en longitude et 5° en latitude. Elle est téléchargeable librement, au format *NetCDF*, sur le site du *MetOffice* (HadSST4, 2021), fichier *HadSST.4.0.1.0_actuals_median.nc*, sous une licence gouvernementale v3 ©British Crown Copyright, Met Office. Cette base est associée à l'article de recherche (Kennedy, 2019).

La version utilisée pour cette étude est l'extrait arrêté à fin juillet 2021, mais afin de ne considérer que des saisons complètes et d'être cohérents avec l'historique disponible dans la base HURDAT2, nous arrêterons les données à fin 2020. Le *Met Office* précise dans son guide d'utilisation de la base HadSST4 (J.J. Kennedy, 2021) qu'il s'agit des températures médianes au sein de chaque carré (de 5° par 5°) mesurées à une profondeur de 20 cm, à l'aide de bouées ou de bateaux, et que les valeurs proviennent du *International Comprehensive Ocean-Atmosphere Data Set (ICOADS)* versions 3.0.0 (1850-2014) et 3.0.1 (2015 à aujourd'hui) ainsi que de relevés de mesures de bouées à la dérive (*drifting buoys*) obtenues au sein du projet *Copernicus* à partir de janvier 2016 et qui viennent remplacer à partir de cette date les relevés des bouées à la dérive du jeu *ICOADS 3.0.1*. Le *Met Office* précise que les données ont été ajustées afin de minimiser les effets liés aux changements de méthodes de mesures au cours de l'historique.

A titre illustratif, en voici un extrait affichant le premier et le dernier mois disponible, au format *DataFrame* disponible sous Python (Van Rossum, 2009) au sein de la librairie *pandas* (The_Pandas_Development_team, 2020) :

			longitude	-177.5	-172.5	-167.5	-162.5	-157.5	-152.5	-147.5	-142.5	-137.5	-132.5	...	132.5	137.5	142.5	14
time_bnds	year	month	latitude															
1850-01-01	1850	1	-87.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	N
			-82.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	N
			-77.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	N
			-72.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	N
			-67.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	N
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
2021-07-01	2021	7	67.5	NaN	NaN	0.822386	0.566307	NaN	NaN	NaN	NaN	4.987915	NaN	...	NaN	NaN	NaN	N
			72.5	NaN	NaN	-1.744227	NaN	0.900149	1.280156	1.160757	NaN	-0.189553	2.19	...	NaN	NaN	NaN	N
			77.5	-0.49982	NaN	NaN	NaN	NaN	NaN	NaN	0.057486	-0.471404	NaN	...	NaN	NaN	NaN	N
			82.5	NaN	NaN	NaN	NaN	NaN	NaN	-0.737685	NaN	NaN	NaN	...	NaN	NaN	NaN	N
			87.5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	N

FIGURE 5 : EXTRAIT DE LA BASE DES TEMPERATURES DE L'EAU FOURNIE PAR LE METOFFICE

Elle se structure par couples (mois, année) qui apparaissent chacun comme une section au sein du fichier. Chacune de ces sections est composée de 36 lignes correspondant à toutes les latitudes de -87.5° à +87.5° avec un pas de 5° (i.e. balayant la totalité de l'horizon de latitudes [-90 ; 90] avec un recentrage) et 72 colonnes correspondant à toutes les longitudes de -177.5° à +177.5° avec un pas de 5° (i.e. balayant la totalité de l'horizon de longitudes [-180 ; 180] avec un recentrage).

On remarque que de nombreuses données ont la valeur par défaut *NaN*. Le *Met Office* explique, dans les instructions jointes avec cette base, qu'il peut s'agir de points géographiques qui ne sont pas situés dans l'eau mais sur terre ou de points pour lesquels la donnée de mesure de température est manquante ou incohérente donc non retenue. Nous allons tout d'abord tenter de distinguer les points géographiques sur terre, qui devraient avoir la valeur *NaN* sur tous les couples (mois, année) et que nous pourrons exclure du jeu de données, des éventuelles « vraies » données manquantes qui *a priori* devraient ne prendre cette valeur par défaut que sur certaines dates d'observation où le relevé de températures n'est pas disponible. Pour identifier les points sur terre, nous avons utilisé une fonction attribuant à un point spatial un booléen indiquant s'il est sur terre ou non. Il s'agit de la fonction *is_land* disponible au sein du package Python *globe* de la librairie *global-land-mask* (project/global-land-mask, s.d.), que nous avons appliquée à tous les carrés, de résolution spatiale 5° de notre grille. Plus précisément, pour chacun des carrés, nous avons appliqué cette fonction successivement aux quatre coins du carré ainsi qu'à son centre. Les résultats, dont nous fournissons ci-dessous une visualisation, avec en abscisses les longitudes et en ordonnées les latitudes, semblent cohérents. Les points en clair sont les points indiqués sur terre et en foncé ceux indiqués en mer.

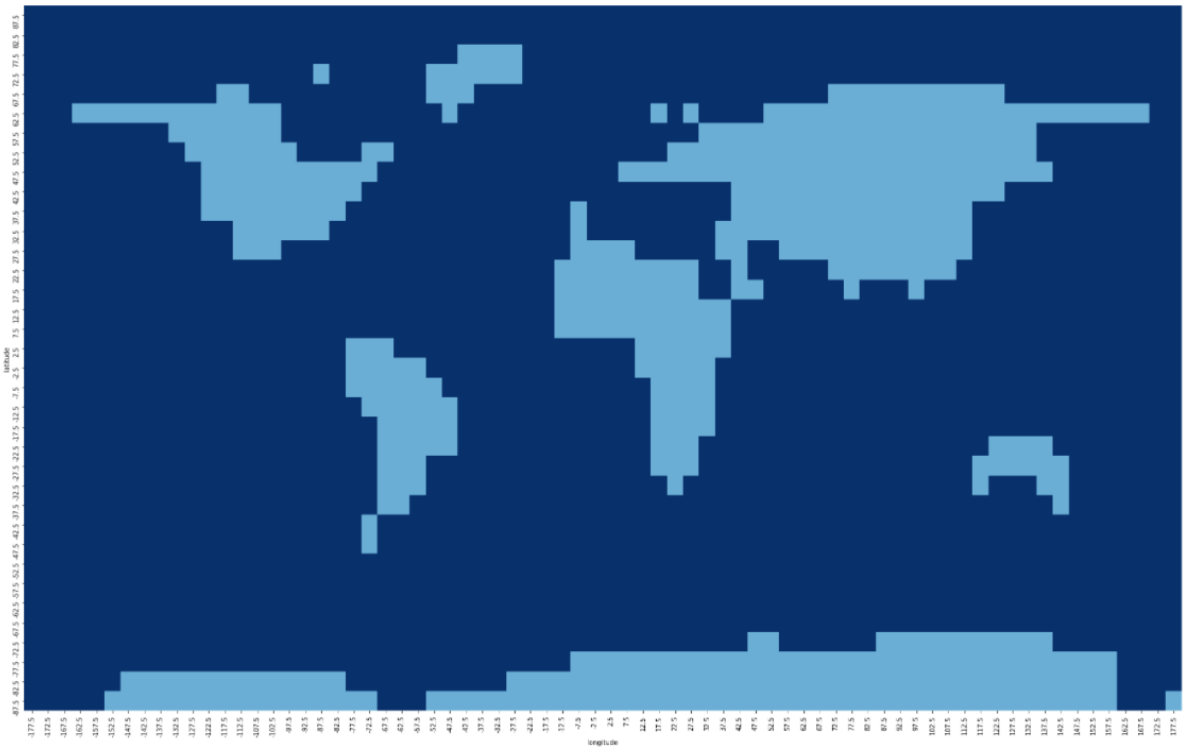

FIGURE 6 : ILLUSTRATION DES POINTS SUR TERRE ET EN MER DANS LA BASE HADSST4

Nous appliquons donc le filtre ainsi défini à notre base de données *HadSST* afin de ne conserver que les points en mer. Ce filtre permet de passer d'une base de 2052 lignes (12 mois x 171 années de 1850 à 2020 inclus) et 2592 colonnes (72 longitudes x 36 latitudes soit 2592 couples de coordonnées géographiques) à une base de même nombre de lignes mais seulement 1886 colonnes.

Dans un second temps, nous appliquons une sélection géographique pour ne conserver que les températures de l'eau dans le bassin Atlantique Nord que nous définissons comme un rectangle de latitude comprise entre 7 et 40 et de longitude entre -100 et -15 afin d'englober à la fois la zone de genèse au large du Cap Vert et la côte nord-est des Etats-Unis jusqu'à New York environ. Ceci permet de réduire la base à 111 colonnes.

Enfin, il est indispensable d'identifier et de traiter les valeurs manquantes en amont de l'apprentissage car les modèles ne sont pas en mesure de les gérer. Nous décidons tout d'abord de retirer points géographiques ayant une qualité de données insuffisante sur la totalité d'historique, c'est-à-dire n'ayant pas un taux satisfaisant de valeurs de températures renseignées. Nous devons ici choisir arbitrairement un seuil que nous fixons à 90%. Ainsi, nous retirons de la base de données les points géographiques ayant moins de 90% de valeurs remplies sur l'ensemble de l'historique, ce qui réduit la base à 79 colonnes. A ce stade, notre base de températures mensuelles de l'eau a donc la forme suivante.

new_indx	-47.5&7.5	-37.5&7.5	-32.5&7.5	-27.5&7.5	-22.5&7.5	-17.5&7.5	-57.5&12.5	-52.5&12.5	-47.5&12.5	-42.5&12.5	...	-62.5&37.5	-57.5&37.5
1-1850	26.781715	25.435608	25.858027	25.630377	27.270887	27.272804	25.983257	26.201920	26.042149	26.054138	...	19.813381	19.571449
2-1850	26.781715	26.186222	25.936396	25.883965	27.270887	27.272804	25.983257	26.201920	26.042149	26.054138	...	19.813381	19.571449
3-1850	26.526661	26.186222	25.936396	26.447533	25.600985	25.740252	26.294416	26.016861	25.230663	25.424202	...	19.401154	18.726414
4-1850	26.556194	26.209064	26.410339	26.447533	26.144981	26.363825	28.239960	25.929684	25.657295	24.157011	...	18.432301	17.475698
5-1850	27.100395	28.286087	28.688396	27.542887	26.144981	26.363825	28.239960	26.439262	26.990047	24.157011	...	21.029530	18.830715
...	...	...	...	...	...	...	...	...	...	...	...	...	...
8-2020	29.348953	28.372303	29.561155	28.326046	28.095425	27.749947	28.916494	28.792959	28.677128	28.144512	...	27.846752	27.753304
9-2020	29.348953	28.533646	28.526394	28.466019	27.817879	27.914577	29.527634	29.198223	28.677128	28.343788	...	27.846752	27.753304
10-2020	29.317755	28.807552	28.994968	29.128489	28.916748	28.450001	29.527634	29.198223	28.627655	28.343788	...	27.136837	26.756809
11-2020	28.958513	28.807552	28.994968	29.128489	29.116535	29.039318	29.060539	29.046959	28.500767	28.298763	...	25.502483	25.337404
12-2020	28.566071	28.211239	28.510708	28.596184	29.116535	29.039318	28.491879	27.939026	27.736719	27.498587	...	23.905964	23.682108

2052 rows × 79 columns

FIGURE 7 : FORMAT DE LA BASE DE DONNEES DE TEMPERATURES MENSUELLES

Sur cette base filtrée, de qualité de données satisfaisante, nous avons encore toutefois un nombre important de valeurs manquantes sur certains points géographiques à certaines dates. Il faut donc désormais proposer une stratégie d'imputation de ces données manquantes.

On peut d'abord penser à une interpolation linéaire dans le temps, par exemple lorsque, pour un point géographique donné, nous identifions une valeur mensuelle manquante entre deux valeurs mensuelles renseignées, i.e. le relevé de température du mois *n* est manquant mais nous disposons de la valeur du mois *n-1* et du moins *n+1*. Toutefois, cette hypothèse nous semble contestable du fait de la saisonnalité, et donc de la non-monotonie, des températures de l'eau au sein d'une année. Pour illustrer cette saisonnalité au sein d'une année, nous avons tracé dans le graphique suivant l'évolution des températures à la surface de l'eau en degré Celsius (°C) sur l'année complète la plus récente de notre historique soit 2020, à deux points géographiques dans l'Atlantique Nord.

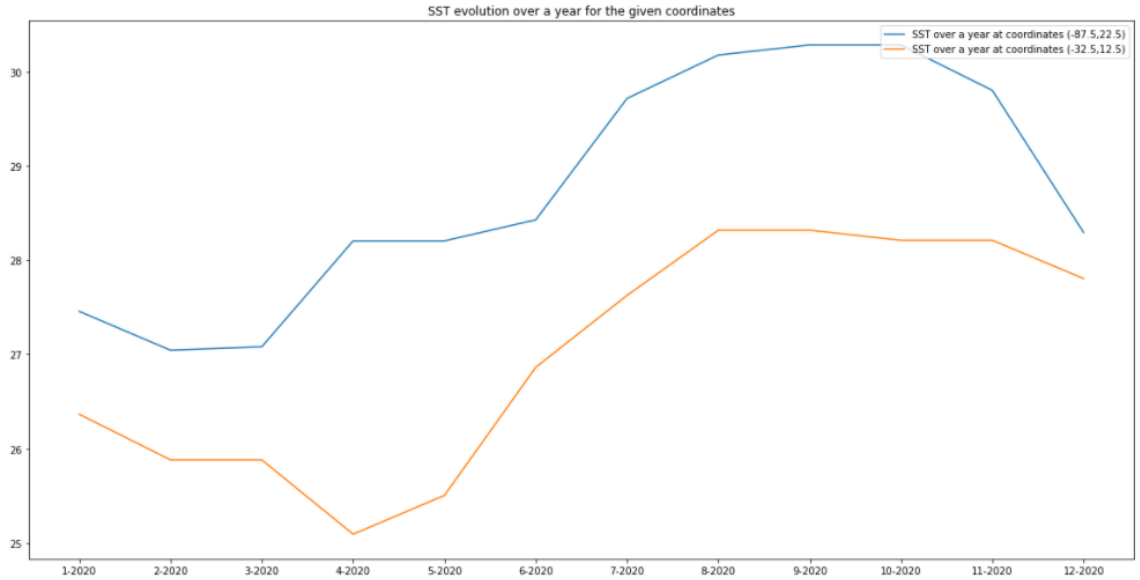

FIGURE 8 : EVOLUTION MENSUELLE DES TEMPERATURES DE L'EAU (°C) DANS L'ATLANTIQUE NORD EN 2020

Le premier point (longitude -87,5, latitude 22,5), dont la courbe est en bleu, est situé dans le golfe du Mexique dans une des zones privilégiées où les ouragans de l'Atlantique Nord s'intensifient et finissent leur course avant potentiellement de toucher terre sur les côtes du Texas, de la Louisiane ou de la Floride, et le second point (longitude -32.5, latitude 12.5) dans l'AMDR au large du Cap Vert dans une des principales zones où se forment ces mêmes ouragans.

On constate un dôme de températures entre juillet et octobre, qui correspond au centre de la saison des ouragans, suivie d'une chute à partir de novembre et jusqu'à mars ou avril selon la zone

géographique considérée. Ceci illustre la saisonnalité typique des températures à la surface de l'eau dans l'Atlantique Nord et ne permet pas de soutenir une hypothèse d'une interpolation linéaire sur une période de plusieurs mois consécutifs.

En revanche, une interpolation linéaire spatiale, entre carrés adjacents pour une date donnée, nous semble plus juste compte-tenu de la continuité de la température à la surface de l'eau sur la résolution spatiale relativement fine de notre base. Néanmoins, pour limiter le risque d'erreur due à une interpolation sur une trop grande zone, nous limitons cette interpolation au cas d'une valeur manquante située entre deux valeurs renseignées.

A ce stade, il reste donc dans notre base de données des valeurs manquantes consécutives, qui ne sont pas interpolées. Pour traiter ce dernier cas, nous procèderons à une imputation des données manquantes à l'aide de l'algorithme des k plus proches voisins que nous détaillerons plus loin dans la partie modélisation.

Une première hypothèse centrale est la granularité temporelle des données de température que nous souhaitons utiliser pour notre modélisation. Nous disposons de données mensuelles, à la fois pour les variables explicatives et les variables à prédire. Cependant, nous ne cherchons pas à prédire un nombre d'ouragans précis mois par mois, mais plutôt d'avoir une prédiction vraisemblable du nombre annuel moyen de tempêtes tropicales sur une période future d'une durée de vingt ans. Nous n'avons donc pas vocation à capter une variabilité mensuelle ni même annuelle mais plutôt une tendance sur les moyen et long termes. Aussi, nous choisirons de traiter les données au niveau annuel, et non mensuel. Ce choix permet également d'éviter les problèmes de dépendances temporelles, qui ne sont pas négligeables : les températures de l'eau de deux mois consécutifs ne sont pas indépendantes entre elles et l'activité cyclonique de fait non plus, alors qu'on peut raisonnablement considérer que l'activité cyclonique d'une année N est indépendante de celle de l'année N-1 ou N+1. La principale source de dépendance entre deux années consécutives vient du mouvement oscillatoire ENSO, responsable d'une certaine variabilité pluriannuelle, mais sa périodicité d'ailleurs est très variable - entre 2 et 7 ans – de sorte qu'on ne puisse modéliser ou prédire une dépendance explicite entre deux années consécutives.

Nous avons donc opéré une transformation de notre base de données afin de n'avoir que les années en ligne et de basculer la dimension mois en colonnes. Notre base de données finales, dont voici un extrait, est composée de 170 lignes et 948 colonnes, correspondant à 12 mois x 79 couples de coordonnées géographiques.

annual_indxs	-47.5&7.5&1	-37.5&7.5&1	-32.5&7.5&1	-27.5&7.5&1	-22.5&7.5&1	-17.5&7.5&1	-57.5&12.5&1	-52.5&12.5&1	-47.5&12.5&1	-42.5&12.5&1	...
1851	27.380344	26.899158	26.884552	26.666904	25.761581	25.913775	27.414604	27.563869	26.913494	26.321573	...
1852	27.032854	27.675678	26.919144	26.711590	25.953577	27.326984	26.676488	26.274038	27.361929	27.810509	...
1853	27.123661	27.657450	26.122650	27.034134	26.801079	27.966961	26.885332	27.290024	27.068567	26.738881	...
1854	27.173225	27.001999	27.048304	26.407215	27.350695	27.288296	27.453522	27.213091	26.027245	25.734989	...
1855	28.152588	27.911125	27.508312	27.348267	27.216808	27.609310	28.188644	27.223082	27.800503	26.975554	...
...	...	...	...	...	...	...	...	...	...	...	...
2016	27.322098	27.366449	27.624840	27.440292	27.847757	29.217842	27.696880	27.314936	27.090279	26.531651	...
2017	28.104511	28.118179	27.706638	27.642302	28.196751	28.970200	28.035748	27.865948	27.600464	26.195286	...
2018	27.520439	27.748363	27.544441	27.862085	28.094810	29.048485	27.838068	27.672203	27.697721	27.031281	...
2019	27.209993	27.317774	27.097666	27.276840	27.606979	28.795624	27.135208	26.776299	26.529194	26.166241	...
2020	27.543053	27.275047	28.133013	27.502609	28.258631	29.414728	27.729298	27.691732	27.160576	26.668575	...

170 rows × 948 columns

FIGURE 9 : EXTRAIT DE LA BASE DE DONNEES DE TEMPERATURES DE L'EAU AU FORMAT FINAL UTILISE (ANNUEL)

Sur cette base réduite et propre, nous serons alors en mesure de calibrer différents modèles prédictifs en apprentissage supervisé tels que les régressions linéaires ou linéaires généralisées (*Generalized Linear Model (GLM)*), la forêt aléatoire (*random forest)* et le *gradient boosting* (algorithme *Extreme Gradient Boosting* (*XGBoost*)). Nous utiliserons Python (Van Rossum, 2009) comme langage de programmation, et en particulier la librairie *Scikitlearn* (Pedregosa, 2011) qui contient un grand nombre de fonctions très utiles pour la préparation des données (e.g. imputation des

données manquantes, séparateur en données d'apprentissage et de test, validation croisée) et la modélisation (grande diversité d'algorithmes de *machine learning* et de métriques de *scoring*). Nous considérons ici le problème comme un problème de régression, c'est-à-dire que nous considérons nos variables à prédire, le nombre annuel d'ouragans et d'ouragans majeurs touchant terre aux Etats-Unis, comme des variables continues. En effet, considérer le problème comme de la classification mettrait en jeu un nombre de classes assez important (autant que de valeurs distinctes de nombre annuel d'ouragans observées). De plus, nous risquerions de moins facilement prédire des valeurs qui sont en dehors de ces valeurs observées, alors que notre objectif est de prédire au mieux le nombre annuel d'ouragans attendus dans un monde plus chaud, avec potentiellement des années qui pourraient avoir un nombre annuel supérieur à tous ceux observés. Ainsi, nous choisissons la régression. La prédiction fournie ne sera pas un nombre entier, comme cela l'aurait été dans un problème de classification, mais nous ferons a priori moins d'erreur en moyenne et nous ne cherchons pas ici à prédire précisément le nombre d'ouragans sur une année en particulier mais bien sur une période future, de 20 ans, sur laquelle nous disposons de projections de températures.

En termes de granularité, comme précisé plus haut, nous modéliserons et prédirons séparément le nombre annuel d'ouragans et le nombre annuel d'ouragans majeurs, notamment pour le besoin de notre application actuarielle à l'estimation des pertes assurées. Ceci permettra d'être plus fin dans nos prédictions mais également de capter dans une certaine mesure une éventuelle influence des températures de l'eau sur la distribution de l'intensité des ouragans.

Nous sommes a priori dans un problème dit de « grande dimension » avec une faible profondeur des données d'apprentissage (*scarcity*), puisque nous disposons seulement 170 ans d'historique donc 170 observations, et près de 1000 variables explicatives. Ceci peut limiter le pouvoir prédictif de nos modèles, comme le souligne l'article de recherche (Asthana, Krim, Sun, Roheda, & Xie, 2021) qui propose une modélisation de l'activité cyclonique par des méthodes d'apprentissage supervisé. De plus, cela écarte a priori tout recours aux méthodes de *deep learning*, notamment aux réseaux de neurones. C'est pourquoi, nous nous concentrerons sur les algorithmes d'apprentissage supervisé classique tels que cités plus haut ainsi qu'aux méthodes de régression linéaire et linéaire généralisée.

Dans le chapitre suivant dédié à la modélisation, nous reviendrons plus en détail sur le sujet de de la grande dimension, ses limites et la façon dont l'avons géré dans cette étude.

Dans un second temps, nous pourrions inclure davantage de variables explicatives susceptibles d'exercer une influence sur le nombre d'ouragans, par exemple l'indice ENSO (*El Niño Southern Oscillation*). Cette fluctuation à l'échelle de quelques années (2 à 7 ans) dans le système océan-atmosphère implique de grands changements dans la circulation atmosphérique dans la région tropicale de l'océan Pacifique. L'état ENSO peut être, entre autres, caractérisé par une anomalie de la température de surface de la mer (SST) dans le Pacifique équatorial oriental et central : les réchauffements dans l'est sont appelés événements *El Niño* et les refroidissements sont des événements *La Niña*.

Les épisodes *El Niño* entraînent une augmentation de la formation de cyclones tropicaux dans le Pacifique Sud et le Pacifique Nord Est central, tandis que d'autres bassins enregistrent une diminution comme dans l'Atlantique Nord, le nord-ouest du Pacifique et la région australienne.

Les épisodes *La Niña* apportent généralement des conditions opposées. Toutefois, cet indice est, au moins en partie capté par les températures de l'eau historiques utilisées, de la base *HadSST*. De plus, nous ne cherchons pas vraiment à capter des cycles de variabilité pluriannuelle et prédire précisément un nombre d'ouragans dans une année donnée mais davantage de prédire une tendance de l'évolution du nombre d'ouragans en fonction des températures à la surface de l'eau. Enfin, cet indice est disponible sur un historique moins long, depuis 1870, que celui des températures de l'eau et nous ne disposons pas de prévision ou de projection de cet indice à horizon 2050 ou 2100, dans un monde plus chaud.

Plus généralement, nous faisons le choix de ne pas ajouter d'autres variables physiques que les températures de l'eau dans la mesure où nous ne nous intéressons qu'à l'influence des températures de l'eau, selon une double dimension spatio-temporelle, sur le nombre d'ouragans et où nous ne disposons pas de projection des autres paramètres physiques pouvant influencer l'activité cyclonique comme les vents de cisaillement ou la pression atmosphérique à la surface de l'eau. De plus, l'ajout de variables explicatives ferait encore accroitre la « grande dimension » de notre jeu de données, ce qui pourrait pénaliser le pouvoir de prédiction et de généralisation de nos modèles plutôt que de l'améliorer.

1.4. Prédiction du modèle sur des projections

Nous appliquerons le modèle ainsi calibré à deux jeux de températures de l'eau projetées, respectivement à horizon 2041–2060 et 2081-2100 à partir des prévisions de températures du GIEC (GIEC, Sixième rapport d'évaluation (AR6), 2021) sur ces deux périodes.

Nous obtiendrons alors une estimation du nombre annuel d'ouragans faisant *landfall* aux Etats-Unis, en distinguant les ouragans majeurs, à ces deux horizons de projection.

Il restera alors à en déduire une estimation des pertes assurées à ces deux horizons.
Pour ce faire, nous utiliserons le logiciel de modélisation développé par l'éditeur RMS (Risk Management Models, Analytics, Software & Services, s.d.) qui est un des logiciels de référence sur le marché en matière de modélisation des risques de catastrophes naturelles. Ce système permet de modéliser les principaux pays/périls auxquels sont exposés le marché de l'assurance, notamment les risques de tremblement de terre et de tempêtes en en Amérique du Nord, en Europe, en Asie et en Australie.

Le schéma ci-dessous illustre le fonctionnement général d'un tel modèle de catastrophe naturelles, par exemple chez RMS.

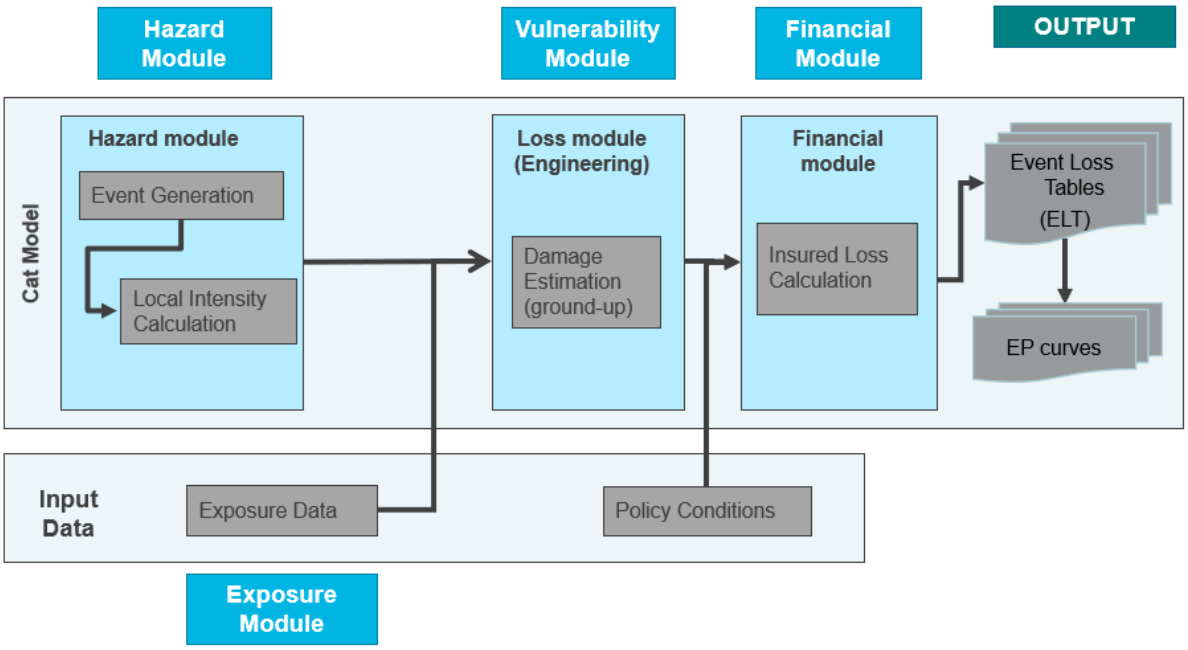

FIGURE 10 : PRINCIPES ET STRUCTURE DU MODELE DE CATASTROPHES NATURELLES RMS

Chaque modèle est constitué d'un premier module dit d'*alea*, ou *hazard* en anglais, qui a vocation à modéliser et générer de façon stochastique l'*alea* naturel, en fréquence et sévérité, à partir d'un catalogue d'événements élaboré par des spécialistes du risque en question (météorologues, géologues, ingénieurs, etc). Ce catalogue contient plusieurs milliers d'événements vraisemblables pour chaque pays/péril. Il s'agit, pour une petite partie, d'événements passés reconstitués, et pour le reste d'événements synthétiques crédibles au vu du phénomène physique en question.

Un deuxième module dit de vulnérabilité permet de modéliser les dommages associés à chacun des événements tirés, en croisant les événements physiques du catalogue avec l'exposition (valeurs assurées) et en associant des taux de destruction aux grandeurs physiques générées, par exemple la vitesse de vent pour les tempêtes et l'accélération du sol (*Peak Ground Acceleration* (PGA)) pour les séismes.

Pour chaque événement du catalogue, la localisation des valeurs assurées mais aussi leurs caractéristiques (type de bâtiment, matériau et année de construction, caractéristiques du toit, nombre

d'étages…) sont autant de paramètres qui entrent dans le calcul du taux de destruction, qui est déterminé bâtiment par bâtiment.

Enfin, un dernier module permet d'appliquer les conditions financières des contrats d'assurance ou de réassurance aux pertes brutes, dites *From Ground Up* (FGU), i.e. avant toute application de conditions d'assurance, afin de les « convertir » en pertes assurées pour le contrat modélisé. Pour un contrat d'assurance avec une franchise F et une limite L, si la perte FGU sur un site assuré est de x, alors la perte assurée y sera égale à :

$$y = \min(\max(x - F\ ;\ 0)\ ;\ L).$$

Le résultat final est une table appelée *Event Loss Table* (ELT) contenant pour tous les événements du catalogue, leur fréquence annuelle ainsi que la moyenne et l'écart-type de la perte FGU pour le portefeuille (exposition) étudié.

En pratique, lorsque l'on effectue une simulation sur un portefeuille de risques pour calculer par exemple la perte moyenne, i.e. la prime pure, ou plus généralement l'ensemble de la distribution de pertes, RMS fait appel à un pré-tirage du catalogue d'événements du région-péril en question. Ce pré-tirage est appelé *Year Event Quantile Table* (YEQT) et contient donc un tirage d'événements pour les *n* années de simulations avec pour chaque événement une valeur de quantile, entre 0 et 1. La fréquence est, pour la majorité des régions-périls modélisés, et en particulier pour les tempêtes de l'Atlantique Nord, supposée suivre une loi de Poisson et la sévérité une loi Beta.

En général, le YEQT est un tirage de n=100 000, ou 1 000 000, années. Pour chaque année, un tirage de la loi de Poisson telle que définie dans le catalogue, est réalisé en tenant compte des fréquences respectives de chaque événement dans le catalogue. Ainsi, un événement du catalogue aura d'autant plus de chance d'être tiré que sa fréquence annuelle indiquée dans le catalogue est élevée. Ensuite, pour chaque événement tiré, au sein de chacune des n années, une valeur est tirée selon une loi uniforme entre 0 et 1. Ce tirage donnera cette fois le quantile à utiliser au sein de la distribution de coût Beta qui est définie par l'ELT de façon unique grâce à la moyenne et l'écart-type.

Ainsi, en croisant le YEQT et l'ELT, RMS est en mesure de calculer pour la sélection d'événement pré-tirés, le coût de chacun d'eux pour le portefeuille de l'ELT et donc la distribution de perte globale, qui peut être décrite par les courbes d'*Occurrence Exceedance Probability* (OEP) et *Annual Exceedance Proability* (AEP) souvent rassemblées sous le terme générique d'*EP (Exceedance Probability) curves*. L'OEP est la distribution de la perte maximale par événement et l'AEP est la distribution de la perte maximale annuelle.

Dans notre cas, nous cherchons à quantifier l'impact en termes de pertes financières en général, pour l'ensemble du marché et non pour un contrat ou un portefeuille en particulier. Nous n'utiliserons donc pas le module financier et utiliserons la base de données d'exposition marché fournie par RMS, appelée *RMS Industry Exposure Database (IED)* censée couvrir l'ensemble des valeurs assurées du marché.

Pour estimer un niveau de pertes assurantielles aux deux horizons de projection à partir de la prédiction de notre modèle, deux options semblent envisageables.

La première consiste à modifier le YEQT, i.e. le pré-tirage des événements au sein du catalogue d'événements stochastiques du risque ouragan aux Etats-Unis. Plus exactement, il s'agit de faire un nouveau tirage afin d'obtenir des nombres annuels moyens d'ouragans et d'ouragans majeurs conformes aux prédictions moyennes obtenues, sur chacun des horizons de projection. Le deuxième module nous permettra d'obtenir une estimation de ce changement de fréquence sur les dommages au marché de l'assurance, en utilisant comme données d'exposition la base RMS IED. Cette méthode est a priori la plus précise mais elle est plus longue et complexe à mettre en place et, de plus, en faisant un nouveau tirage d'événements, nous sommes susceptibles d'altérer la structure de dépendance des événements, notamment des corrélations, implémentée dans le modèle RMS. En effet, le tirage d'événements effectué

par RMS prend en compte non seulement la fréquence cible mais aussi des considérations de dépendances, notamment spatiales, entre les événements.

La seconde option consiste à calculer les coûts moyens pour le marché d'un ouragan et d'un ouragan majeur aux Etats-Unis donné par RMS et de multiplier les prédictions de nombres moyens annuels respectives que nous obtenons par ces coûts moyens afin d'obtenir une estimation de la perte marché moyenne. Nous opterons pour cette méthode pour sa simplicité et sa rapidité d'application et compte-tenu de notre objectif qui est davantage d'estimer l'impact sur la sinistralité moyenne du péril ouragans aux Etats-Unis, notamment dans une optique de tarification (prime pure).

1.5. Limites

Dans cette étude, nous nous intéressons uniquement à l'influence de l'évolution des températures de l'eau sur la fréquence annuelle d'ouragans. En réalité, le réchauffement climatique peut également exercer une influence sur la sévérité des ouragans, que nous captons partiellement en distinguant les ouragans majeurs, mais aussi leurs emplacements de genèse, leurs trajectoires (incluant probabilité et emplacement de *landfall*) ou encore leur répartition dans l'année. En particulier, la saison des ouragans pourrait se prolonger si les températures de l'eau restent chaudes plus longtemps par exemple jusqu'à décembre voire janvier. Le début de saison pourrait également être avancé dès avril ou mai si les températures de l'eau atteignent des niveaux suffisants, autour de 26,5°C.

En outre, dans cette étude, nous n'apprécions le réchauffement climatique qu'à travers les températures à la surface de l'eau, par simplification et parce ce que ce sont les variables physiques reflétant le réchauffement climatique qui sont le plus directement liées à la formation de cyclones. Or, comme le rappelle l'OMM, la température n'est que l'un des indicateurs du changement climatique. Les autres indicateurs sont les concentrations de gaz à effet de serre, le contenu thermique de l'océan, le pH de l'océan - dû à l'absorption par les océans d'une partie du dioxyde de carbone présent dans l'atmosphère -, le niveau moyen des mers et océans à l'échelle du globe, la masse glaciaire, l'étendue de la banquise et les phénomènes extrêmes. Nous faisons l'hypothèse que les autres paramètres physiques influençant la formation des ouragans, comme les vents de cisaillement, ne sont pas modifiés de façon matérielle par le réchauffement climatique. Ainsi, pour simplifier, nous ne modélisons volontairement que la dépendance en température. Néanmoins, cette hypothèse nous semble raisonnable notamment au regard de l'objectif de prédiction recherché, à savoir prédire une tendance du nombre annuel moyen d'ouragans sur deux périodes de projection utilisées par le GIEC (GIEC, Sixième rapport d'évaluation (AR6), 2021) et longues de vingt ans chacune (2041-2060 et 2081-2100).

De plus, nous ne modélisons ici que l'évolution de la fréquence du nombre d'ouragans. Toutefois, comme nous distinguons les ouragans majeurs, nous captons implicitement l'évolution de leur intensité (puisque l'on prédira par exemple plus d'ouragans majeurs ou plus d'ouragans majeurs en proportion du nombre total de tempêtes) i.e. l'évolution des vitesses de vent. Or, un cyclone peut causer des dégâts non seulement par les vents violents mais aussi par les pluies torrentielles qu'il déverse ainsi que par la submersion marine déjà évoquée qu'il peut engendrer. Pour certains ouragans comme Harvey (2017), qui est l'exemple récent le plus frappant, la majorité des dégâts matériels et donc des pertes financières est même causée par ces pluies abondantes et les inondations qui en découlent. Il est établi, notamment par le GIEC, que le réchauffement climatique va augmenter le niveau général des mers et océans, accentuant ainsi mécaniquement le risque de *storm surge*. En outre, des températures plus élevées accentuent l'évaporation et donc l'humidité de l'air ce qui va très probablement accroître l'intensité et/ou la durée des précipitations, et donc le risque d'inondations liées aux pluies déversées par les tempêtes tropicales et ouragans, comme expliqué notamment par le regroupement de chercheurs universitaires sur la recherche atmosphérique *University Corporation for Atmospheric Research* (UCAR) sur leur page dédiée au lien entre réchauffement climatique et ouragans (UCAR, 2018). Certaines études comme (Kossin, 2014) montrent également que le changement climatique pourrait entraîner une migration vers les pôles (*poleward migration*) des emplacements de formation ou

d'intensité maximale observée des ouragans, effet que nous ne captons pas non plus dans notre étude puisque nous ne nous modélisons que le nombre annuel de *landfalls* et pas la trajectoire ou la répartition spatiale des ouragans.

S'ajoute à cela que statistiquement plus la durée de vie d'un ouragan est longue, plus il a de chance de causer des dégâts importants, qu'ils soient liés aux vents ou aux pluies. Or, d'après une étude du *Geophysical Fluid Dynamics Laboratory* (GFDL), laboratoire de l'université de Princeton, New Jersey, rattaché à la NOAA, le réchauffement climatique pourrait également accroitre la durée de vie des ouragans d'en moyenne 5% au cours du XXIème siècle. Tous ces effets sont inventoriés et décrits précisément sur le site du GFDL, sur la page dédiée au lien entre réchauffement climatique et ouragans (Knuston, 2021).

Par ailleurs, il existe d'autres effets anthropogéniques que l'émission de gaz à effet de serre et le réchauffement climatique susceptibles exercer une influence sur l'activité cyclonique. Un des plus couramment cités dans la littérature est l'émission d'aérosols, qui ont un effet refroidissant à court terme dans l'atmosphère. Une émission particulièrement élevée de ces gaz pendant les années 1970 permettrait d'expliquer, au moins en partie l'activité cyclonique exceptionnellement faible sur cette période.

Enfin, notre estimation des pertes assurées futures est uniquement fondée sur une modification de la distribution de fréquence des ouragans et ouragans majeurs. En effet, nous faisons l'hypothèse que le coût moyen d'un ouragan d'une catégorie donnée est inchangé, et égal au coût moyen dans le modèle RMS version 18 livré en 2019.

Comme indiqué plus haut, elle n'intègre donc pas une éventuelle augmentation de la sévérité des pertes assurantielles dues à l'augmentation de l'intensité et/ou de la durée des précipitations ou encore du niveau de la mer et donc du risque de submersion marine.

Ainsi, notre estimation ne capture pas non plus une éventuelle évolution future de l'exposition assurée aux Etats-Unis, que ce soit en termes de montant total assuré ou de répartition géographique ou encore de nature des bien assurés. Nous faisons donc ici l'hypothèse que l'exposition assurée totale (marché) est stable au cours du XXIème siècle. En effet, nous ne disposons pas de projection fiable des valeurs assurées, et encore moins de l'évolution de leur distribution géographique. A titre d'exemple, une croissance du taux de pénétration de l'assurance ou une augmentation du prix des matériaux de construction pourrait mener, pour un événement donné, à un coût moyen plus élevé pour le marché de l'assurance. Inversement, la mise en place de moyens de prévention ou de protection ou encore l'amélioration de la résistance des structures des bâtiments aux ouragans pourrait réduire le coût moyen pour le marché d'un événement donné. Suite à l'ouragan Katrina (2005) qui a été particulièrement dévastateur du fait de la rupture des digues protégeant la ville de la Nouvelle-Orléans, ces défenses ont été renforcées. Ainsi, un événement similaire ferait désormais beaucoup moins de dégâts, notamment en termes d'inondations. Cette hypothèse a été, en partie au moins, vérifiée à travers l'expérience de l'ouragan Ida (2021), relativement comparable à Katrina en termes de trajectoire et d'intensité (catégorie 4 pour Ida, 5 pour Katrina) mais qui a été bien moins destructeur, avec des dommages totaux estimés autour de 75 milliards de dollars dont environ la moitié assurée, contre 125 pour Katrina dont environ la moitié assurée là aussi. Toutefois, au-delà de cet exemple connu, il est très compliqué d'évaluer si des moyens de protection seront mis en place ou renforcés et quels seront leurs effets sur les dommages en cas d'ouragan.

2. Modélisation du nombre annuel d'ouragans

Nous allons donc modéliser le nombre annuel d'ouragans, en calibrant notre modèle sur l'ensemble des données historiques disponibles, i.e. de 1851 à 2020 qui est la période temporelle la plus longue commune à nos deux bases de données *HadSST4* et *HURDAT2*. Ensuite, nous nous intéresserons à la prédiction de ce modèle sur deux jeux de projections de températures de l'eau fondées sur différentes trajectoires du GIEC à deux horizons distincts, 2041–2060 et 2081-2100.

2.1. Etude préliminaire exploratoire

2.1.1. Analyse temporelle

Commençons par examiner séparément l'évolution historique des variables explicatives et des variables à prédire.

Le graphique ci-dessous donne l'évolution de la moyenne (en orange), de la médiane (en vert) et du maximum (en bleu) de la température globale de l'Atlantique Nord entre 1850 et 2020.

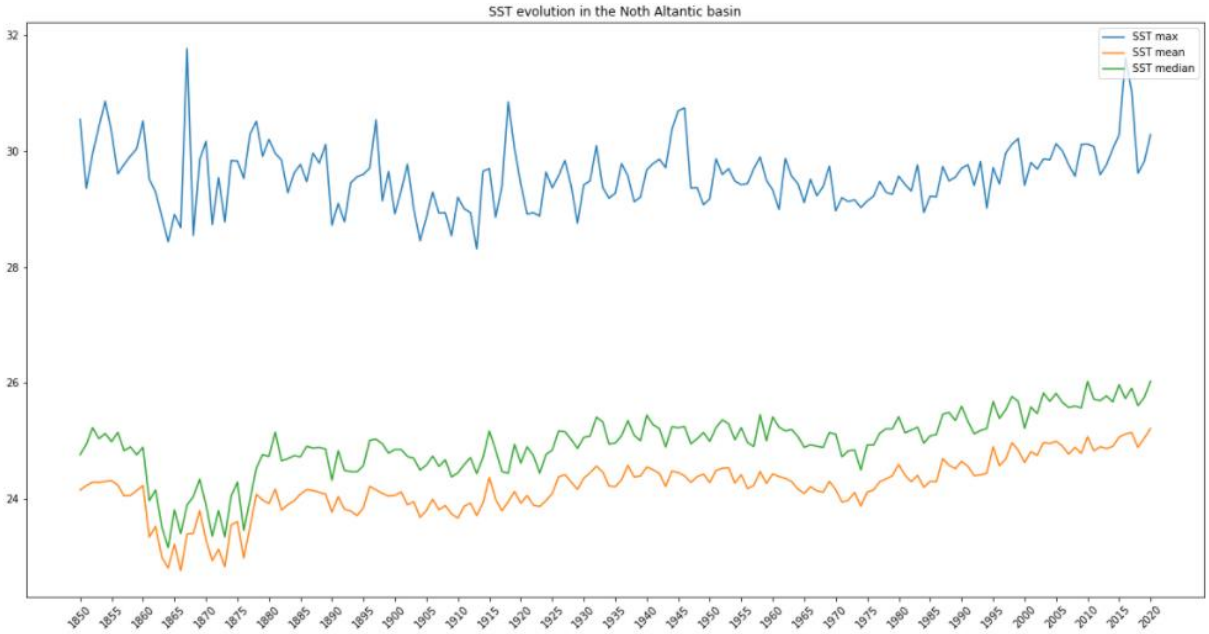

FIGURE 11 : EVOLUTION GLOBALE DES TEMPERATURES DE L'EAU DANS L'ATLANTIQUE NORD ENTRE 1850 ET 2020

On y détecte une augmentation assez claire des températures moyennes et médianes, ce qui est en ligne avec le consensus scientifique, notamment relayé par le GIEC. La tendance est moins évidente, car non monotone, sur la température maximale.

Nous nous sommes intéressés aux variables explicatives, i.e aux emplacements et mois, correspondant au maximum et au minimum de température observés sur l'historique, dans la zone sélectionnée. Le maximum est très souvent situé au sud ou à l'ouest de Cuba, aux points entre les longitudes -82,5 et -77,5 et à la latitude 22,5, aux mois d'août ou de septembre, ce qui semble tout à fait cohérent au vu de la nature et la saisonnalité du phénomène. Pour le minimum, il s'agit majoritairement du point de coordonnées (-77,5 ; 37,5) situé au large à l'Est de Washington, donc près de la limite septentrionale de notre zone sélectionnée, au mois de février, ce qui là encore semble cohérent.

Analysons désormais l'évolution historique comparée des nombres annuels de tempêtes, d'ouragans, et respectivement de tempêtes et d'ouragans ayant touché terre aux Etats-Unis sur l'ensemble de la période d'observation disponible, soit de 1851 à 2020.

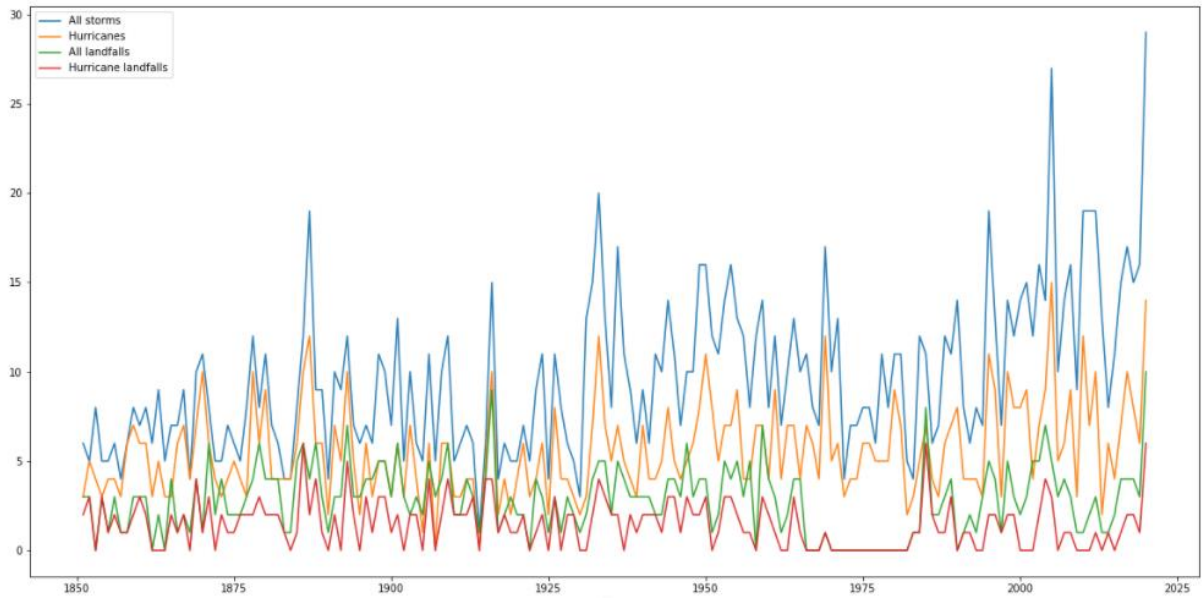

FIGURE 12 : EVOLUTION DES NOMBRES ANNUELS DE TEMPETES, D'OURAGANS ET D'OURAGANS MAJEURS DANS L'ATLANTIQUE NORD ENTRE 1851 ET 2020

Ces grandeurs paraissent assez (positivement) corrélées dans le temps, avec toutefois une période notable de quasi 20 ans entre 1966 et 1984 durant laquelle l'activité cyclonique semble anormalement faible. Seulement 3 *landfalls* y sont rapportés, et même aucun pendant une période de 13 ans entre 1970 et 1982. A partir de 1985, il semble se dessiner une tendance à la hausse avec un pic historique en 2020 avec un record historique absolu de nombre annuel total de tempêtes, 29, et de nombre annuel total de *landfalls* aux Etats-Unis, 10, ainsi qu'un record de 10 *landfalls* d'ouragans à égalité avec l'année 1985.

Commençons par analyser la période de basse activité entre 1966 et 1984. S'agissant d'une période relativement récente et, en particulier, bénéficiant de la reconnaissance par avion et par satellite, nous pouvons écarter l'hypothèse d'un problème majeur de qualité de données.

L'explication de cette longue période d'activité cyclonique notablement basse n'est pas triviale et des hypothèses variées sont avancées dans la littérature. Si certains papiers avancent la variabilité multi-annuelle due à l'oscillation atlantique muti-décennale (*Atlantic Multi-decadal Oscillation* (AMO), on trouve régulièrement citée l'hypothèse d'une émission particulièrement forte d'aérosols sulfatés issus de la combustion croissante et massive de fioul et de charbon au lendemain de la seconde guerre mondiale (Emanuel, 2021) notamment aux Etats-Unis et en Europe. Ces gaz aérosols rendent en effet l'atmosphère moins transparente et provoquent ainsi un refroidissement à court terme de la surface de la terre (effet « parasol »), et des océans, qui est susceptible d'avoir réduit voire supprimé temporairement la formation d'ouragans. Le niveau d'émissions d'aérosols due à l'activité humaine aurait par la suite diminué du fait de la mise en place de politique de réglementation pour limiter la pollution de l'air et de l'utilisation moindre d'énergies fossiles dans les pays développés.

Quelques papiers comme (Evan, 2012) évoquent le rôle des éruptions volcaniques, notamment celle majeure du Chichon (Mexique) en 1982, qui dégagent également des aérosols dans l'atmosphère.

Regardons désormais de plus près l'évolution historique de la fréquence annuelle de *landfall* avec des indicateurs tels que la moyenne glissante sur 10 ans des nombres annuels de tempêtes, ouragans et ouragans majeurs ayant touché terre.

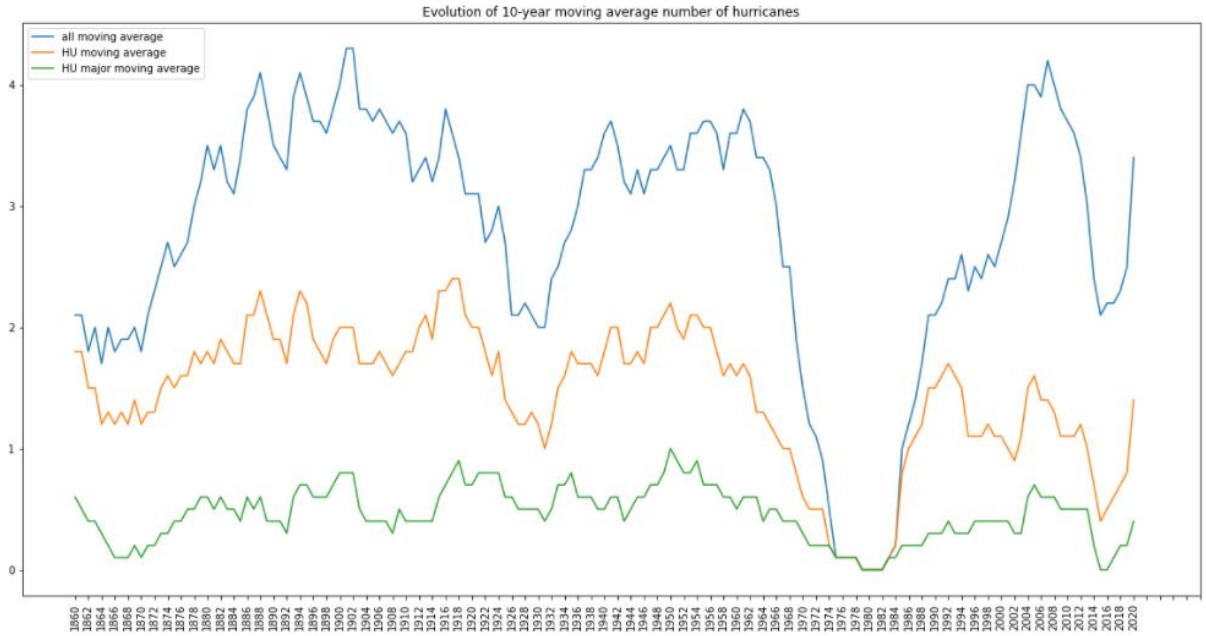

FIGURE 13 : EVOLUTION DE LA MOYENNE GLISSANTE SUR 10 ANS DES NOMBRES ANNUELS DE TEMPETES, OURAGANS ET OURAGANS MAJEURS AYANT TOUCHE TERRE

Ce graphique montre pour chaque ensemble (toutes les tempêtes, sous-ensemble des ouragans, sous-ensemble des ouragans majeurs) le creux d'activité dans les années 1970 et une tendance à la hausse depuis avec toutefois un minimum local en 2016 donc sur la moyenne entre 2007 et 2016. L'activité semble également assez réduite sur les deux premières décennies, donc jusqu'à 1870 environ, ce qui peut être dû, au moins, en partie à la qualité de données plus faible sur le début de l'historique comme évoqué plus haut. En revanche, sur la totalité de la période d'observation, il ne semble pas se dégager une tendance claire à la hausse, en tout cas pas sur les nombres annuels de *landfalls*.

En dehors de la période d'accalmie des années 1970, les nombres annuels de tempêtes, d'ouragans, de tempêtes ayant touché terre aux Etats-Unis et d'ouragans ayant touché terre aux Etats-Unis semblent corrélés positivement.

Vérifions cela en analysant les corrélations entre ces variables, en ajoutant le sous-ensemble des ouragans majeurs.

2.1.2. Analyse des corrélations

2.1.2.1. Corrélations entre variables à prédire

Nous avons tout d'abord calculé les coefficients de Pearson, qui décrivent les corrélations linéaires.

	Nb all	Nb HU	Nb Major HU	Nb all landfalls	Nb HU landfalls	Nb Major HU landfalls
Nb all	1.00	0.83	0.68	0.45	0.31	0.25
Nb HU	0.83	1.00	0.73	0.40	0.42	0.32
Nb Major HU	0.68	0.73	1.00	0.39	0.41	0.55
Nb all landfalls	0.45	0.40	0.39	1.00	0.78	0.56
Nb HU landfalls	0.31	0.42	0.41	0.78	1.00	0.71
Nb Major HU landfalls	0.25	0.32	0.55	0.56	0.71	1.00

FIGURE 14 : CORRELATIONS LINEAIRES (PEARSON) ENTRE LES NOMBRES ANNUELS

Puis, nous avons calculé les coefficients de Spearman, qui ne sont autres que les coefficients de Pearson appliqués aux rangs des deux variables et qui mesurent donc les corrélations de rang.

	Nb all	Nb HU	Nb Major HU	Nb all landfalls	Nb HU landfalls	Nb Major HU landfalls
Nb all	1.00	0.80	0.62	0.42	0.25	0.14
Nb HU	0.80	1.00	0.65	0.38	0.38	0.22
Nb Major HU	0.62	0.65	1.00	0.31	0.33	0.45
Nb all landfalls	0.42	0.38	0.31	1.00	0.75	0.50
Nb HU landfalls	0.25	0.38	0.33	0.75	1.00	0.67
Nb Major HU landfalls	0.14	0.22	0.45	0.50	0.67	1.00

FIGURE 15 : CORRELATIONS DE RANG (SPEARMAN) ENTRE LES NOMBRES ANNUELS

On remarque tout d'abord que les coefficients de Spearman sont du même ordre mais toujours légèrement inférieurs aux coefficients de Pearson. Les observations que l'on va formuler ici sur les corrélations linéaires, par exemple, sont donc également valables pour les corrélations de rang.

Il apparaît que le nombre d'ouragans majeurs est fortement corrélé, à 73%, au nombre d'ouragans et également au nombre total de tempêtes, à 68%. Il semble logique que le nombre d'ouragans majeurs soit fortement corrélé à ces deux variables avec une valeur plus forte pour la corrélation avec le nombre d'ouragans, puisque les ouragans majeurs forment un sous-ensemble du nombre d'ouragans qui constituent eux-mêmes un sous-ensemble du nombre total de tempêtes.

Nous pouvons formuler la même observation, avec la même explication sous-jacente, en ce qui concerne le nombre de *landfalls*, ce qui là encore est logique et attendu.

En revanche, le nombre annuel de *landfalls* d'ouragans n'est corrélé qu'à 38% avec le nombre annuel d'ouragans. Quand on s'intéresse aux ouragans majeurs uniquement, cette valeur est un peu plus élevée, à 45% mais reste aussi bien inférieure aux valeurs de corrélations décrites plus haut entre les sous-ensembles. Cela est dû au fait qu'il y a un *alea* supplémentaire, le fait pour une tempête de toucher terre ou non (binaire), qui est par nature non linéaire et donc réduit la corrélation linéaire entre le nombre d'un ensemble et le nombre de *landfalls* de ce même ensemble.

2.1.2.2. Corrélations entre variables explicatives et à prédire

Il est également intéressant d'analyser les corrélations, linéaires et de rang, entre les variables explicatives ou *features*, ici les températures de l'eau, et la cible, i.e. les nombres annuels d'ouragans et d'ouragans majeurs. Cela peut permettre d'identifier d'éventuelles variables explicatives qui auraient une corrélation ou anti-corrélation particulièrement forte avec la cible. En outre, cela peut également aider à effectuer une ou des sélections de *features* que nous pourrons utiliser pour réduire la dimension. La réduction de dimension peut s'avérer très utile pour améliorer le pouvoir de prédiction de notre modèle, sa capacité à généraliser, mais également avoir un modèle plus lisible et interprétable ainsi que plus rapide, même si le temps de calcul n'est pas une donnée critique dans notre étude.

Nous remarquons tout d'abord, sans grande surprise, que les températures sont corrélées entre carrés voisins, ce qui est attendu. En termes de corrélation avec le nombre d'ouragans, il n'y a toutefois pas de valeur de corrélation ou d'anti-corrélation particulièrement significative qui se démarque, les valeurs étant toutes comprises entre -30% et +20% comme illustré dans le nuage de points ci-dessous tracé à partir de la matrice de corrélations de Spearman obtenue.

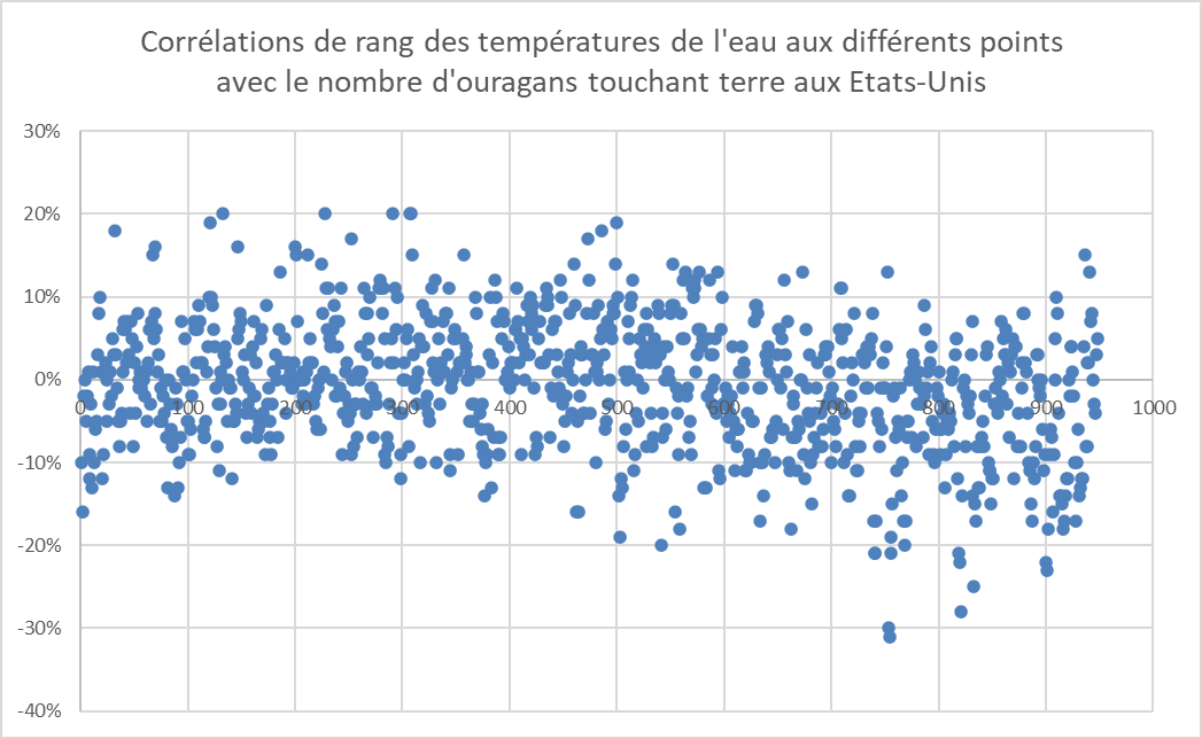

FIGURE 16 : NUAGE DE POINTS DES CORRELATIONS SPEARMAN DES TEMPERATURES DE L'EAU AVEC LE NOMBRE D'OURAGANS TOUCHANT TERRE AUX ETATS-UNIS

A noter qu'un point de ce nuage représente l'association d'un point géographique situé dans l'Atlantique Nord et d'une date (mois x année). Nous captons donc bien la corrélation à la fois spatiale et temporelle des températures de l'eau avec la fréquence annuelle d'ouragans.

La carte ci-dessous donne une visualisation en bleu plus foncé des 20 points (carrés de 5° de latitude par 5° de longitude) les plus corrélés, au sens de Spearman, avec le nombre d'ouragans ayant touché terre aux Etats-Unis.

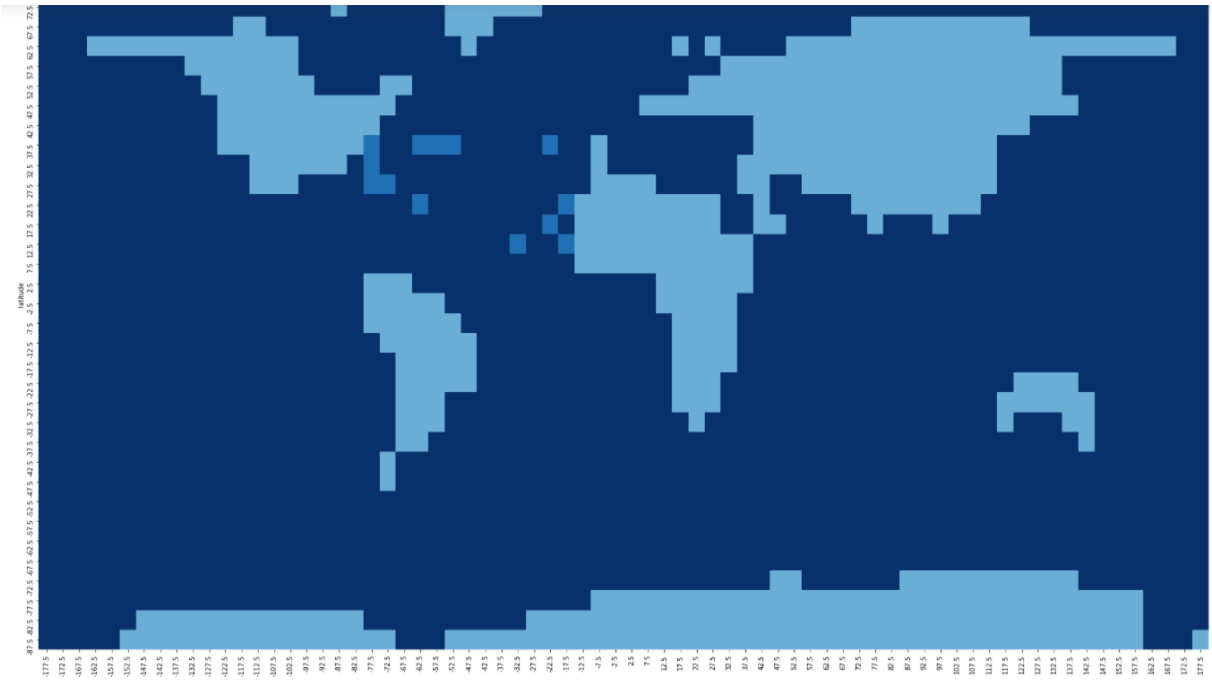

FIGURE 17 : AFFICHAGE DES 20 POINTS LES PLUS CORRELES (SPEARMAN) AU NOMBRE ANNUEL D'OURAGANS TOUCHANT TERRE AUX ETATS-UNIS

Ces valeurs de corrélations les plus élevées nous semblent assez cohérentes compte tenu des trajectoires typiques des ouragans de l'Atlantique Nord atteignant les côtes américaines. On note à droite des points situés dans la zone de genèse principale des systèmes tropicaux de l'Atlantique Nord au large du Cap Vert et à gauche des points par lesquels peuvent typiquement passer ces mêmes systèmes avant d'atteindre les Etats-Unis. A titre illustratif, nous présentons ci-dessous une carte élaborée par le *NHC* des trajectoires des tempêtes tropicales (incluant ouragans) sur la saison record de 2020.

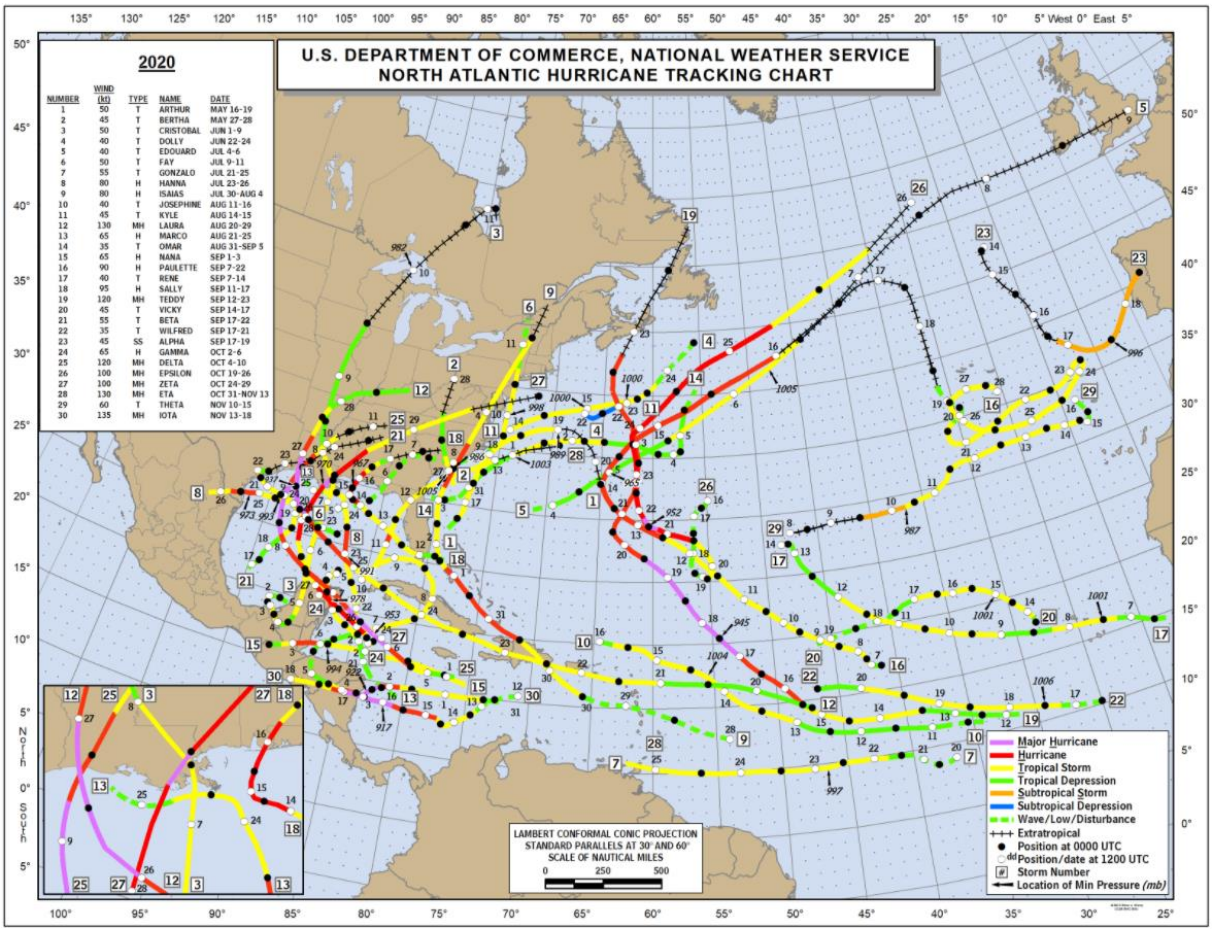

FIGURE 18 : TRAJECTOIRES DES TEMPETES DANS L'ATLANTIQUE NORD EN 2020

Pour respectivement le nombre annuel total de tempêtes, d'ouragans et d'ouragans majeurs ayant touché terre, nous avons défini plusieurs sélections de *features* constituées chacune soit des 10 soit des 20 *features* les plus corrélés à la cible et des 5 *features* les plus anti-corrélés, au sens de Pearson puis au sens de Spearman. A ce stade, cela nous donne 4 sélections.

Nous avons également défini une sélection ne conservant que les *features* corrélés positivement à plus de 30% et ceux corrélés négativement au moins à -10%, ce qui en pratique ne fait conserver que des valeurs de corrélations négatives inférieures à -10% étant donné les valeurs de corrélations observées telles que décrites plus haut et illustrées dans le nuage de points.

Enfin, nous constituons une dernière sélection de variables explicatives en retenant les *features* communs aux 3 ensembles composés des 20 *features* les plus corrélés respectivement au nombre annuel de tempêtes, ouragans et ouragans majeurs ayant touché terre et symétriquement avec les 3 ensembles des 20 les plus anti-corrélés. Nous obtenons alors la sélection suivante :

	Nb all landfalls	Nb HU landfalls	Nb Major HU landfalls
-62.5&17.5&1	0.219606	0.150346	0.183638
-22.5&27.5&6	0.216050	0.163391	0.204831
-57.5&32.5&10	-0.103375	-0.196488	-0.176186
-67.5&22.5&12	-0.106004	-0.209552	-0.215569
-62.5&27.5&10	-0.109624	-0.240355	-0.142984
-52.5&32.5&11	-0.121775	-0.211825	-0.186932
-62.5&32.5&10	-0.128807	-0.242423	-0.219852
-62.5&27.5&12	-0.142132	-0.189631	-0.146544
-57.5&27.5&12	-0.143077	-0.225806	-0.210250
-52.5&32.5&12	-0.165570	-0.235960	-0.189460
-67.5&22.5&11	-0.168357	-0.252233	-0.146352

FIGURE 19 : SELECTION DES VARIABLES EXPLICATIVES LES PLUS CORRELEES ET ANTI-CORRELEES EN COMMUN A LA FOIS AUX NOMBRES DE TEMPETES, D'OURAGANS ET D'OURAGANS MAJEURS AYANT TOUCHE TERRE

Cette sélection contient donc 2 variables corrélées positivement et 9 anti-corrélées. En se limitant aux 3 composantes communes les plus anti-corrélées, nous obtenons ainsi une sélection supplémentaire de 5 variables explicatives.

Analysons désormais les distributions des nombres annuels au travers des principaux indicateurs statistiques, en particulier la moyenne (*mean*), l'écart-type (*std*), la médiane (quantile 50%), les quantiles 25 et 75% et les valeurs extrêmes (*min* et *max*).

	Nb all	Nb HU	Nb Major HU	Nb all landfalls	Nb HU landfalls	Nb Major HU landfalls
count	170.000000	170.000000	170.000000	170.000000	170.000000	170.000000
mean	9.676471	5.605882	1.911765	2.805882	1.470588	0.470588
std	4.413002	2.627342	1.649236	1.934725	1.350768	0.698066
min	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	6.000000	4.000000	1.000000	1.000000	0.000000	0.000000
50%	9.000000	5.000000	2.000000	3.000000	1.000000	0.000000
75%	12.000000	7.000000	3.000000	4.000000	2.000000	1.000000
max	29.000000	15.000000	7.000000	10.000000	6.000000	3.000000

FIGURE 20 : RESUME STATISTIQUE DES DISTRIBUTIONS DE NOMBRE ANNUEL

Nous constatons que les distributions du nombre annuel total de tempêtes, du nombre annuel d'ouragans et du nombre annuel d'ouragans majeurs sont assez dispersées, avec une variance supérieure à la moyenne, et ce d'autant plus que l'intensité est élevée, ce qui semble assez logique s'agissant d'événements extrêmes. Il en est de même pour les distributions de nombres annuels de *landfalls* qui sont même significativement plus dispersées, ce qui semble là aussi cohérent compte tenu de l'*alea* supplémentaire binaire du *landfall*.

2.2. Modélisation

2.2.1. Préparation des données pour l'entraînement

A ce stade, les données dont nous disposons sont filtrées - sur la zone géographique d'intérêt - et nettoyées des composantes dont le taux de remplissage était jugé trop faible. L'interpolation spatiale a permis de compléter certaines valeurs manquantes isolées, i.e. encadrées de deux valeurs renseignées, mais il reste encore des valeurs manquantes consécutives, non gérées par l'interpolation spatiale que nous avons volontairement limitée à un pas de 1.

En outre, il est nécessaire pour certains modèles paramétriques, notamment ceux faisant appel à une descente de gradient comme les régressions linéaire ou logistique, que toutes les données explicatives soient normalisées, i.e. toutes comprises dans le même intervalle [-1 ; 1]. Actuellement, nos variables explicatives sont des températures de l'eau exprimées en degrés Celsius (°C) comme illustré dans le résumé statistique ci-dessous.

	-47.5&7.5&1	-37.5&7.5&1	-32.5&7.5&1	-27.5&7.5&1	-22.5&7.5&1	-17.5&7.5&1	-57.5&12.5&1	-52.5&12.5&1	-47.5&12.5&1	-42.5&12.5&1	...	-62.5&37.5&12
count	154.000000	156.000000	168.000000	170.000000	170.000000	170.000000	157.000000	158.000000	158.000000	156.000000	...	161.000000
mean	27.282016	27.009752	26.941896	26.864432	27.200999	27.825523	27.267125	27.000959	26.710989	26.310382	...	22.285502
std	0.496492	0.664713	0.588895	0.505554	0.571732	0.697898	0.559366	0.598037	0.636104	0.792017	...	0.762076
min	26.319300	25.453426	25.316072	25.749657	25.457006	25.780811	25.340189	25.219467	24.744980	23.478739	...	18.178996
25%	26.909619	26.558489	26.592416	26.533700	26.829038	27.396656	26.879494	26.571320	26.312062	25.864971	...	21.870411
50%	27.272011	27.008084	26.923979	26.792752	27.213572	27.806366	27.255630	27.069139	26.715569	26.311338	...	22.278793
75%	27.645668	27.468235	27.323322	27.132728	27.594486	28.284176	27.662500	27.433451	27.158329	26.760948	...	22.818958
max	28.537790	28.954521	28.641186	29.078863	28.317429	30.199284	28.604471	28.448406	28.994427	28.708065	...	24.050755

FIGURE 21 : RESUME STATISTIQUE DE LA BASE HISTORIQUE DE TEMPERATURES DE L'EAU

Pour l'entraînement de nos modèles, nous devons séparer notre base en données d'apprentissage et données de test. Nos modèles seront calibrés sur les premières et nous évaluerons leur capacité de prédiction, et de généralisation, sur les secondes. Nous utilisons la fonction ad-hoc *train_test_split* disponible dans la librairie Python *Scitkitlearn* (Pedregosa, 2011) en définissant une proportion de 20% de données de test, et donc 80% de données d'apprentissage. Les données de test sont supposées inconnues et nous devons donc veiller à ne pas utiliser leur information pour l'imputation des données manquantes ni pour la normalisation de nos données d'apprentissage. Autrement, nous aurons une « fuite » (*leak*) des données d'apprentissage dans l'échantillon de test. Ainsi, nous séparons notre jeu de données en amont de ces étapes de préparation de données (*pre-processing*). Le jeu d'apprentissage contient donc 136 lignes et le jeu de test 34.

Nous avons d'abord effectué un centrage et une réduction (standardisation) de chacune des variables explicatives en colonne. Pour chaque colonne, nous avons retranché la moyenne (*mean* dans le tableau ci-dessus) de la colonne à l'ensemble des observations puis divisé par l'écart-type (*std* pour *standard deviation* dans le tableau) de cette même colonne. Après cette opération, tous nos *features* sont bien des variables numériques prenant des valeurs comprises entre 0 et 1.

Nous avons alors procédé à l'imputation des valeurs manquantes consécutives, en utilisant un modèle des k plus proches voisins (*k nearest neighbours (kNN))* calibré uniquement sur les données d'apprentissage et appliqué séparément sur le jeu d'apprentissage et sur le jeu de test. Concrètement, chaque valeur manquante – située dans une année *a* – sera remplie à partir des valeurs observées dans les *k* années qui ressemblent le plus à l'année *a* dans notre base d'apprentissage. Nous choisissons k=10 voisins sans pondération, pour éviter le surapprentissage.

2.2.2. Entraînement, sensibilité et calibration des hyperparamètres

Séparément et indépendamment pour le nombre annuel de *landfalls* d'ouragans et le nombre annuel de *landfalls* d'ouragans majeurs, nous avons entrainé nos différents modèles sur le jeu d'apprentissage

complet ainsi que sur chacun des jeux d'apprentissage de dimension réduite correspondant aux différentes sélections de *features* décrites plus haut.

On est dans un cas dit de « grande dimension » puisque le nombre de *features*, noté *p*, est relativement grand par rapport au nombre d'observations noté *n* : p >> n. Il est ainsi pertinent de procéder à une réduction de dimensions et de voir si cela peut améliorer la capacité de prédiction de nos modèles. On parle couramment dans les problèmes d'apprentissage machine de gérer le « fléau de la dimension » (*curse of dimensionality*). Globalement, plus le nombre de variables explicatives augmente, plus le modèle va tenir compte de spécificités observées dans les données, notamment d'éventuelles redondances ou du bruit, ce qui augmente la complexité du modèle, favorise le surapprentissage, et peut dégrader sa performance de prédiction.

Il existe d'autres méthodes de réduction de dimension, comme l'Analyse en Composantes Principales (ACP) ou le Lasso. Nous avons également testé ces deux approches.

Toutefois, l'ACP n'a pas permis d'améliorer sensiblement la performance des modèles, ou en tout cas pas autant que la réduction faite à partir des corrélations, alors qu'elle est plus lourde à mettre en place et moins facilement interprétable que cette dernière puisqu'elle crée des nouvelles variables explicatives qui n'ont plus de sens physique.

Quant à la régression Lasso, comme nous le verrons plus loin, elle fait partie à part entière des modèles testés et nous n'avons volontairement pas souhaité l'utiliser à des fins de réduction de dimension afin de ne pas utiliser des résultats d'un modèle pour calibrer un autre modèle. Nous avons choisi de calibrer et tester les modèles indépendamment afin notamment de pouvoir les comparer de façon cohérente.

Nous avons analysé et comparé les coefficients de détermination, notés R^2, des jeux d'apprentissage et de test mais également l'erreur quadratique moyenne (*mean squarred error (MSE)*) qui est la métrique d'évaluation (ou *scoring*) que nous avons retenue, simple et adaptée à notre problème. Cette dernière est définie par la somme des différences au carré entre la valeur observée yi et sa prédiction $\hat{y}i$, sur le jeu de test, divisée par le nombre *n* de valeurs :

$$MSE(y, \hat{y}) = \frac{1}{n}\sum_{n}(yi - \hat{y}i)^2 .$$

Cette métrique quantifie donc l'écart moyen entre la prédiction et la valeur réellement observée sans distinction relative au signe de chaque écart, ce qui nous semble pertinent ici car nous ne souhaitons pas accorder une importance différente à la sur ou sous-estimation de la fréquence annuelle.

Le coefficient de détermination, quant à lui, est défini par la formule suivante, où $\bar{y}$ est la moyenne empirique des observations yi, pour i allant de 1 à n :

$$R^{2(y,\hat{y})} = 1 - \frac{\sum_n (yi - \hat{y}i)^2}{\sum_n (yi - \bar{y})^2}.$$

Il peut être vu comme le rapport de la variance expliquée par la régression sur la variance totale. C'est un indicateur de l'adéquation du modèle au problème, au travers d'une valeur entre -1 et 1. La valeur maximale 1 correspond à un modèle généralisant parfaitement sur des données inconnues. Un modèle constant qui prédirait tout le temps la moyenne empirique aurait un coefficient de détermination nul. La valeur du coefficient peut également être négative, par exemple si la valeur prédite est à chaque fois à plus éloignée de la cible que la moyenne empirique.

La valeur du coefficient de détermination sur le jeu d'apprentissage permet notamment de contrôler s'il y a un phénomène de surapprentissage (*overfitting*) du modèle et d'éliminer ces cas. Dans de tels cas, le coefficient de détermination calculé sur le jeu d'apprentissage est égal ou quasi égal à 1 et en général très faible voire négatif sur le jeu de test, traduisant l'incapacité du modèle à généraliser correctement sur des données inconnues.

L'erreur quadratique moyenne calculée sur le jeu de test nous permet de comparer la capacité de prédiction des différents modèles sur les données de test. Afin de rendre cette mesure plus robuste et complète, nous ne calculons pas seulement cette métrique sur le jeu de test mais aussi à partir d'une validation croisée, à 5 plis (*5-fold cross validation*). Cette validation croisée permet d'évaluer la métrique sur 5 découpages différents, et indépendants les uns des autres, de nos données en jeu d'apprentissage et jeu de test.

Nous avons entraîné et comparé les modèles suivants, par ordre de complexité : régression linéaire, modèle linéaire généralisé (GLM), régression *kNN*, arbre de décision, forêt aléatoire ou *Random Forest* (RF) et *Gradient Boosting* (GB).

Pour chacun de ces modèles, nous avons exploré et testé différentes hypothèses de calibration, par exemple le choix des pénalités pour la régression linéaire, le choix des lois *a priori* pour le GLM et les valeurs d'hyperparamètres pour les algorithmes de *machine learning*. Nous allons décrire plus en détail les spécificités et les tests de sensibilité réalisés pour chacun des modèles.

2.2.2.1. Régression linéaire

La régression linéaire (*linear regression* (lr)) est l'un des modèles les plus simples pour décrire le comportement d'une variable continue en fonction d'une ou plusieurs autres. Dans certains cas, où la relation entre la variable à prédire et au moins une partie des variables explicatives, est linéaire ou quasi linéaire, elle se révèle très efficace et suffit à décrire la relation avec un niveau de précision satisfaisant. Elle a l'avantage d'être particulièrement simple à utiliser et interpréter, et parcimonieuse car elle ne fait appel à aucun hyperparamètre. Le modèle linéaire de variable à prédire $\boldsymbol{Y}$ (vecteur) et de variables explicatives $\boldsymbol{X}$ (matrice) s'écrit comme suit, avec β le vecteur des coefficients, ou paramètres, à estimer :

$$\boldsymbol{Y} = \boldsymbol{X}\,\beta\ .$$

La régression consiste alors à trouver le vecteur β qui minimise l'erreur des moindres carrés (en norme L2) i.e. la quantité :

$$\|\boldsymbol{Y} - \boldsymbol{X}\,\beta\,\|_{\mathbf{2}}^{\mathbf{2}}\ .$$

Ce modèle offre également la possibilité d'avoir recours à des méthodes de régularisation, également appelées pénalités, qui permettent de « forcer » un certain nombre de coefficients à être nuls ou proches de zéro. Mathématiquement, cela se traduit par l'ajout d'une contrainte supplémentaire dans la minimisation des moindres carrés de la régression portant sur la valeur des coefficients.
On distingue principalement 3 pénalités : *Lasso (Least Absolute Shrinkage and Selection Operator)*, *Ridge* et *Elastic Net*.

Dans le cas de la pénalité Lasso, la somme des valeurs absolues des coefficients doit être inférieure à une certaine valeur choisie, i.e. qu'on va chercher le vecteur de paramètres β minimisant la quantité $\|\boldsymbol{Y} - \boldsymbol{X}\,\beta\,\|_{\mathbf{2}}^{\mathbf{2}}$ sous la contrainte $\sum_n |\beta_i|$ minimum. Sous forme synthétique, vectorielle et lagrangienne, la quantité à minimiser s'écrit :

$$\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\,\beta\,\|_{\mathbf{2}}^{\mathbf{2}} + \alpha\,\|\beta\,\|_{\mathbf{1}}\ .$$

On peut noter que cette quantité fait intervenir un hyperparamètre α en facteur de la contrainte sur les coefficients. C'est le cas pour toutes les régressions à pénalités.

Dans la régularisation Ridge, c'est la même contrainte supplémentaire mais exprimée en norme L2 (quadratique), et non L1, des coefficients, c'est-à-dire :

$$\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\,\beta\,\|_{\mathbf{2}}^{\mathbf{2}} + \alpha\,\|\beta\,\|_{\mathbf{2}}\ .$$

Enfin, la pénalité Elastic Net combine les deux précédentes en instaurant une contrainte de minimisation supplémentaire sur les coefficients à la fois en norme L1 et en norme L2, ce qui s'écrit :

$$\frac{\mathbf{1}}{\mathbf{2}}\|\boldsymbol{Y} - \boldsymbol{X}\,\beta\,\|_{\mathbf{2}}^{\mathbf{2}} + \alpha\,L1_ratio\,\|\beta\,\|_{\mathbf{2}}\ + \alpha\,(1 - L1_{ratio})\,\|\beta\,\|_{\mathbf{1}}\ .$$

La régression Elastic Net fait donc appel à un hyperparamètre supplémentaire noté $L1_ratio$ correspondant au poids à affecter à la contrainte de pénalité L1 (Lasso), le poids alloué à la pénalité L2 (Ridge) n'étant autre que son complémentaire, i.e. $1 - L1_ratio$.

Pour chacune des deux variables à modéliser, nous avons testé la régression linéaire sans pénalité, la Ridge, la Lasso et l'Elastic Net.

Comme nous l'avons vu, les modèles de régression à pénalité font appel à un hyperparamètre de régularisation α, venant en facteur de la contrainte supplémentaire sur les coefficients dans le Lagrangien, et dont la valeur est comprise entre 0 et 1. Nous avons donc recherché, dans chacun des cas, la valeur optimale, i.e. celle minimisant la MSE du jeu de test, en faisant varier la valeur de α entre 0 et 1 avec un pas de 0,1 et en calculant pour chacune de ces valeurs la MSE de test.

Pour la régression Elastic Net, nous avons fait varier la valeur de l'hyperparamètre sur la totalité de son intervalle, entre 0 et 1 par pas de 0,1, et avons calculé la MSE de test en chaque point.

Concernant la régression linéaire classique, sans pénalité, les résultats de MSE de validation croisée s'échelonnent entre 1,59 et 15 selon les sélections de *features* pour la régression linéaire du nombre annuel d'ouragans, le minimum étant obtenu avec la sélection des 10 *features* les plus corrélés et 5 les plus anti-corrélés, au sens de Pearson. Sur le jeu complet de variables explicatives, la MSE de validation croisée est égale à 2,6 et on note clairement du surapprentissage, le coefficient de détermination calculé sur le jeu d'apprentissage étant égal à 1 comme d'ailleurs pour la sélection de variables donnant la valeur maximale de MSE.

Nous conserverons donc pour la suite le modèle de régression linéaire calibré sur la sélection des 10 *features* les plus corrélés et 5 les plus anti-corrélés. A noter que la régression linéaire classique ne fait appel à aucun hyperparamètre et donc il n'y a pas eu ici de recherche d'hyperparamètres optimaux.

2.2.2.2. Régression Ridge

Comme nous l'avons mentionné plus haut, la régression Ridge fait intervenir un hyperparamètre de régularisation α sur lequel nous avons joué afin d'obtenir le modèle optimal.

Nous avons réalisé une analyse de sensibilité à partir du modèle calibré sur la sélection de *features* donnant le MSE de validation croisée la plus faible qui est ici aussi la sélection des 10 *features* les plus corrélés et 5 les plus anti-corrélés. Avec cette sélection et la valeur par défaut choisie $\alpha = 1$, la MSE de validation croisée vaut 1,57. Par comparaison, la MSE de validation croisée la plus élevée est de 2,6 ce qui est bien inférieur à la valeur maximale observée avec la régression linéaire classique et montre l'intérêt et l'efficacité de l'usage d'une pénalité.

Le graphique ci-dessous montre les valeurs de la MSE sur le jeu de test pour la régression Ridge du nombre annuel d'ouragans avec calibration sur cette sélection de *features*.

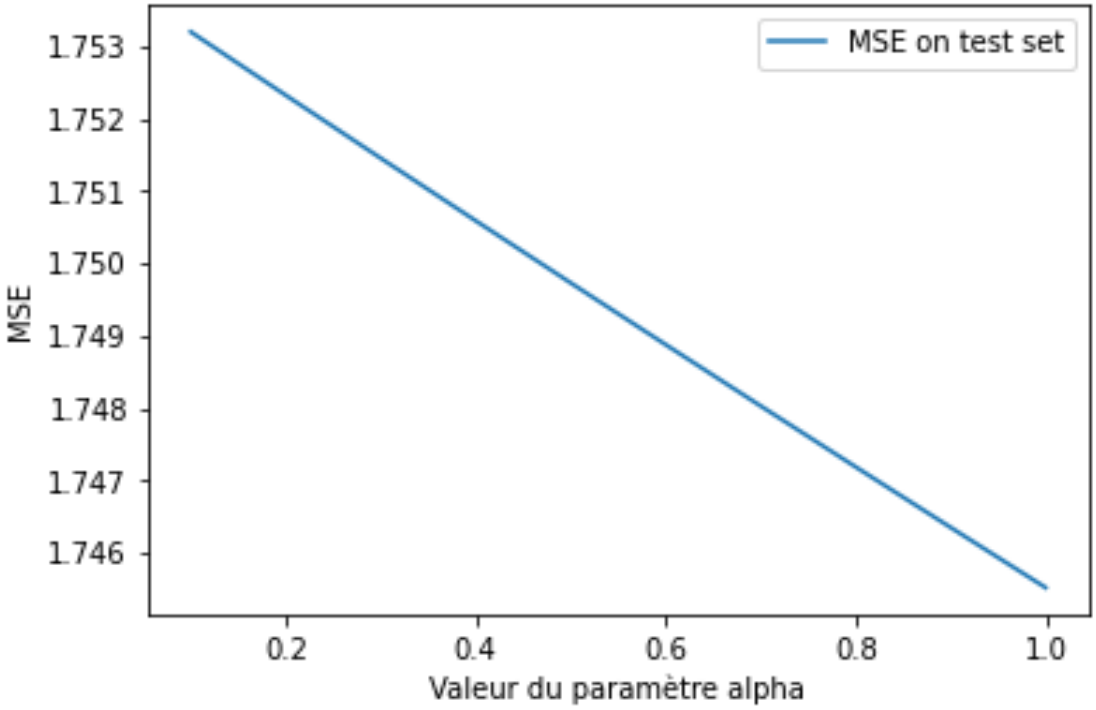

FIGURE 22 : ERREUR QUADRATIQUE MOYENNE SUR LE JEU DE TEST EN FONCTION DE LA VALEUR DE L'HYPERPARAMETRE DE REGULARISATION DANS LA REGRESSION RIDGE DU NOMBRE D'OURAGANS

Nous observons une décroissance linéaire même si de pente très faible (-0,7% environ) de la MSE de test en fonction de la valeur de α, ce qui nous conduit à choisir la valeur maximale $\alpha = 1$.

Nous n'affichons volontairement pas sur ce graphique l'évolution de la MSE calculée sur le jeu d'entraînement car elle est constante et égale à 1,2 sur tout l'intervalle [0 ; 1] et son affichage sur la même échelle occulterait la très légère variation du MSE de test sur ce même intervalle.

Quant à la régression sur le nombre d'ouragans majeurs, le graphe de sensibilité ci-dessous a une forme très similaire, avec une pente même plus faible en absolu (-0,2% environ) ce qui est cohérent avec le fait que le nombre d'ouragans majeurs est plus faible que le nombre d'ouragans.

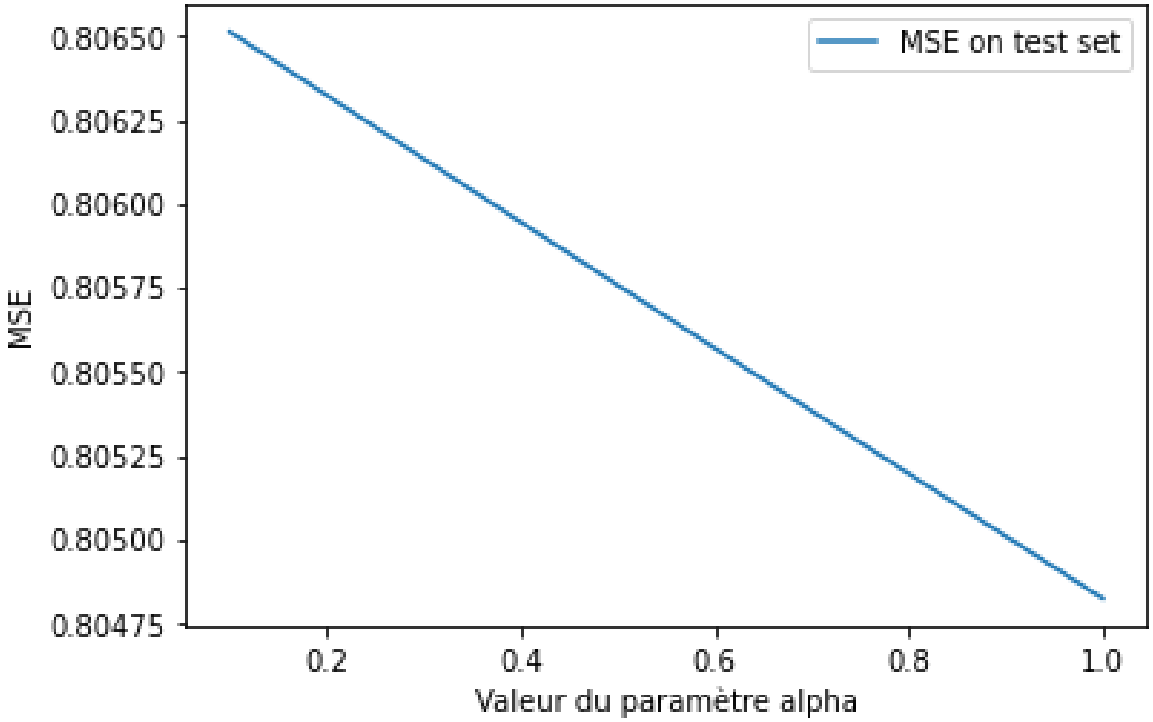

FIGURE 23 : ERREUR QUADRATIQUE MOYENNE SUR LE JEU DE TEST EN FONCTION DE LA VALEUR DE L'HYPERPARAMETRE DE REGULARISATION DANS LA REGRESSION RIDGE DU NOMBRE D'OURAGANS MAJEURS

Nous retenons donc également le choix $\alpha = 1$ pour la régression Ridge du nombre annuel d'ouragans majeurs.

2.2.2.3. Régression Lasso

Pour la régression Lasso, nous avons également commencé par définir la sélection de *features* minimisant la MSE de validation croisée, avec la valeur par défaut choisie $\alpha = 1$. Il s'agit de la même sélection que celle utilisée pour la régression Ridge, qui donne une MSE de validation croisée égale à 1,52, la MSE maximale étant égale à 2 et observée pour le jeu de *features* complet.

De façon similaire, sur ce modèle ainsi calibré, nous avons réalisé une sensibilité en faisant varier la valeur de l'hyperparamètre α et en calculant le MSE sur le jeu de test. Le graphique ci-dessous illustre les résultats obtenus.

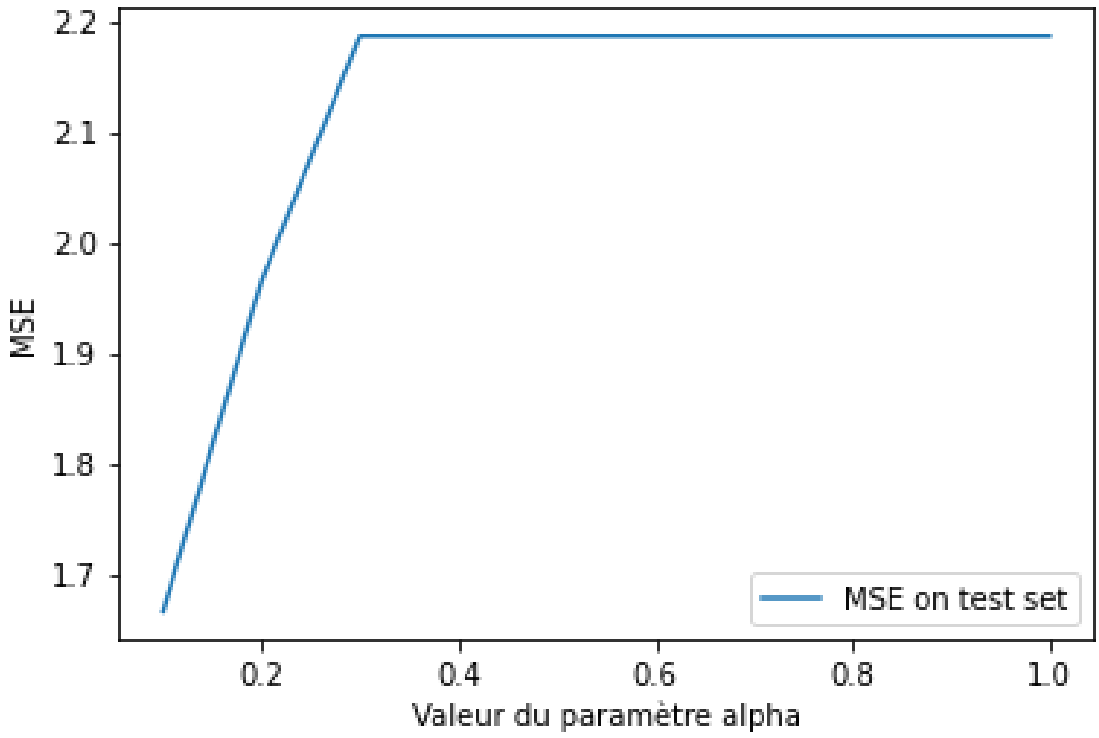

FIGURE 24 : ERREUR QUADRATIQUE MOYENNE SUR LE JEU DE TEST EN FONCTION DE LA VALEUR DE L'HYPERPARAMETRE DE REGULARISATION DANS LA REGRESSION LASSO DU NOMBRE ANNUEL D'OURAGANS

Dans le cas du Lasso, la MSE de test semble être plus sensible à la valeur de l'hyperparamètre α comme nous le montre le graphique ci-dessus, avec une pente autour de 50%, bien plus élevée que dans le cas Ridge.

La MSE calculée sur le jeu de test est minimale pour $\alpha = 0{,}1$ et augmente ensuite quasi linéairement jusqu'à atteindre un plafond à partir de $\alpha = 0{,}3$.

Nous avons observé des résultats de sensibilité très similaires pour la régression Lasso du nombre d'ouragans majeurs comme le montre le graphique ci-dessous.

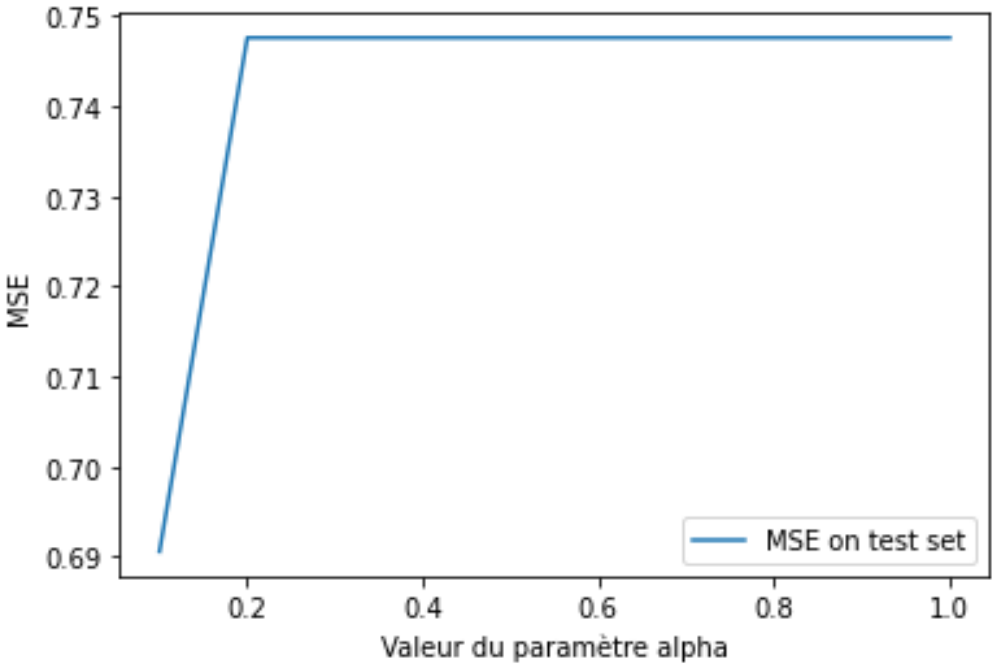

FIGURE 25 : ERREUR QUADRATIQUE MOYENNE SUR LE JEU DE TEST EN FONCTION DE LA VALEUR DE L'HYPERPARAMETRE DE REGULARISATION DANS LA REGRESSION LASSO DU NOMBRE ANNUEL D'OURAGANS MAJEURS

In fine, nous retenons donc la valeur $\alpha = 0{,}1$ à la fois pour la régression du nombre annuel d'ouragans et du nombre annuel d'ouragans majeurs.

Rétroactivement, nous avons vérifié qu'avec cette valeur de $\alpha = 0{,}1$ la sélection de *features* retenue est toujours celle donnant la MSE de validation croisée la plus faible.

2.2.2.4. Régression Elastic Net

Dans le cas de la régression Elastic Net, nous faisons varier la valeur de l'hyperparamètre $L1_ratio$, également entre 0 et 1 par pas de 0,1, comme illustré dans le graphique ci-dessous, élaboré sur la régression du nombre d'ouragans avec calibrations sur la sélection de *features* minimisant la MSE de validation croisée dans le cas Ridge et Lasso.

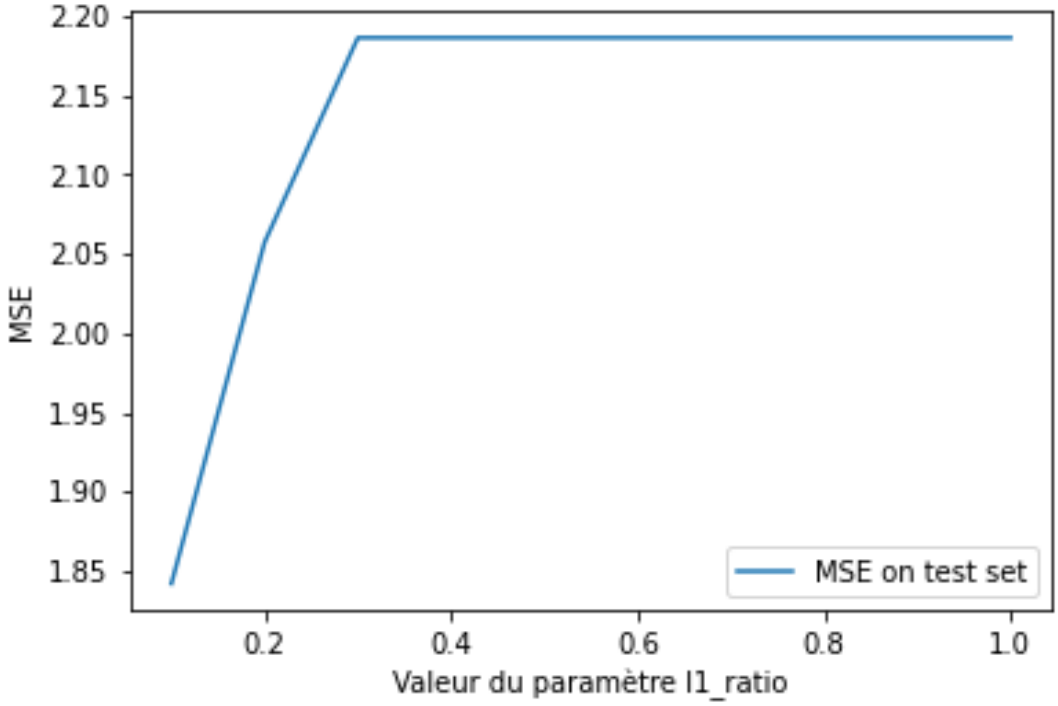

FIGURE 26 : ERREUR QUADRATIQUE MOYENNE SUR LE JEU DE TEST EN FONCTION DE LA VALEUR DE L'HYPERPARAMETRE *L1_RATIO* DANS LA REGRESSION ELASTIC NET DU NOMBRE D'OURAGANS

On retrouve une courbe similaire à celle de l'hyperparamètre α sous Lasso, avec une MSE minimale pour la valeur $L1_ratio = 0{,}1$ et stable à partir de 0,3.

On observe une courbe très similaire pour le nombre d'ouragans majeurs, où le plafond est même atteint dès la valeur 0,2 comme on le voit dans le graphique suivant.

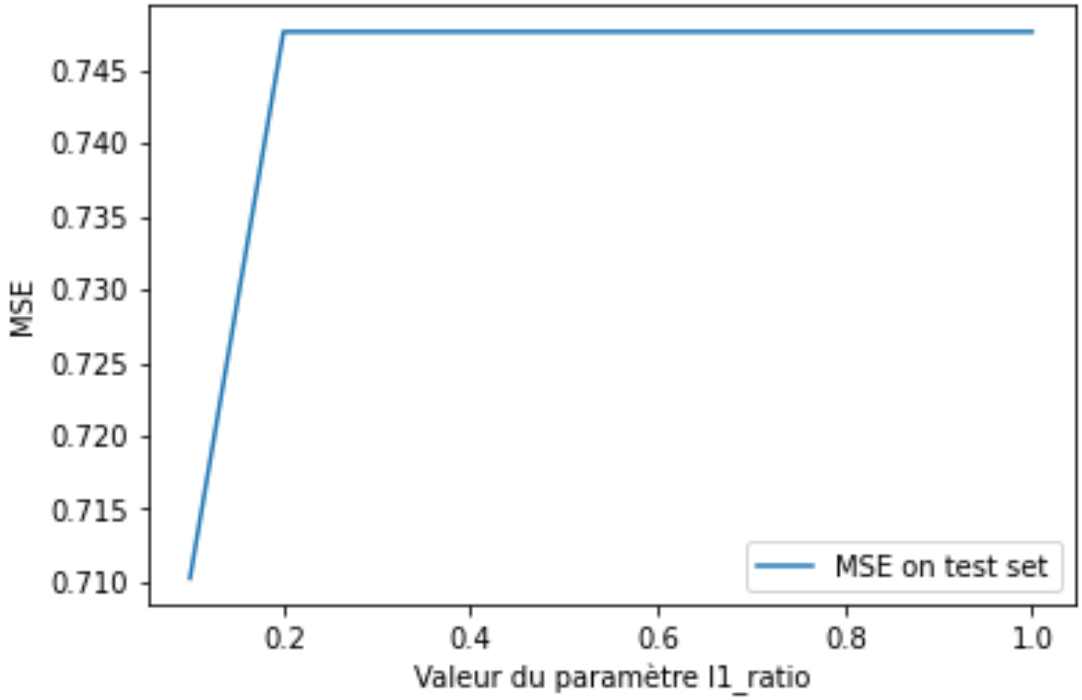

FIGURE 27 : ERREUR QUADRATIQUE MOYENNE SUR LE JEU DE TEST EN FONCTION DE LA VALEUR DE L'HYPERPARAMETRE *L1_RATIO* DANS LA REGRESSION ELASTIC NET DU NOMBRE D'OURAGANS MAJEURS

Nous choisissons donc la valeur $L1_ratio = 0{,}1$ pour la régression Elastic Net des nombres annuels d'ouragans et d'ouragans majeurs.

2.2.2.5. Régression linéaire zéro-modifiée (*hurdle regression*)

Le nombre annuel de *landfalls* d'ouragans passe régulièrement par zéro, et c'est d'autant plus vrai sur le sous-ensemble des ouragans majeurs puisque ces événements sont, par nature, moins fréquents et plus volatiles. Or, les modèles linéaires ne permettent pas toujours de prédire correctement des passages aussi fréquents par zéro.

De plus, comme nous l'avons vu dans la description du phénomène physique, un ouragan ne peut se former qu'à partir d'un certain seuil de température à la surface de l'eau, situé autour de 26,5°C.

Par conséquent, il peut être opportun d'intégrer une composante non linéaire dans la modélisation afin de capter notamment cet effet « seuil ».

Nous avons ainsi testé un modèle de régression de loi zéro-modifiée ou *hurdle regression* (« hurdle » signifie « haie » ou « barrière » en anglais) telle que décrite par (Mullahy, 1986), en choisissant une valeur de *hurdle* à 0. Il s'agit d'un modèle en deux étapes semblant bien répondre à notre besoin.

Dans la première étape, on modélise par une régression logistique la probabilité d'avoir au moins un *landfall* dans l'année.

Dans la seconde, on calibre une régression linéaire uniquement sur les données d'entraînement ayant au moins un *landfall* (annuel). Les probabilités issues de la régression logistique sont mises à 0 si la probabilité de *landfall* est inférieure à 50% et 1 sinon (binarité) et viennent en facteur multiplicatif des prédictions de notre modèle linéaire.

Cette association d'une régression logistique et d'un modèle linéaire peut également se faire avec modèle linéaire généralisé (GLM).

Aucune librairie ne proposant ces modèles de régression de loi zéro-modifiée sous Python, nous les avons développés, pour chacun des modèles de régression linéaire décrits ci-dessus ainsi que pour les GLM. Par conséquent, nous n'avons pas été en mesure de calculer de façon systématique, comme pour les estimateurs codés dans *Scikitlearn*, la métrique MSE en validation croisée. Nous n'avons donc calculé et comparé que les MSE sur le jeu de test.

Nous avons suivi la même méthodologie que pour les régressions classiques étudiées plus haut, à savoir que nous avons d'abord recherché la sélection de variables explicatives donnant la MSE de test la plus faible. Nous n'avons en revanche pas refait de recherche d'hyperparamètres optimaux, mais nous nous sommes placés directement aux valeurs optimales trouvées précédemment pour chacun des hyperparamètres.

Nous avons ensuite comparé les résultats de prédictions, en particulier les MSE, sur le jeu de test entre les versions avec et sans *hurdle*, calibrées sur la même sélection de variables explicatives et les valeurs d'hyperparamètres optimales. Le graphique ci-dessous illustre cette comparaison dans le cas de la régression linéaire (sans pénalité) et du Lasso. Le graphique contient également les observations originales en bleu, pour référence.

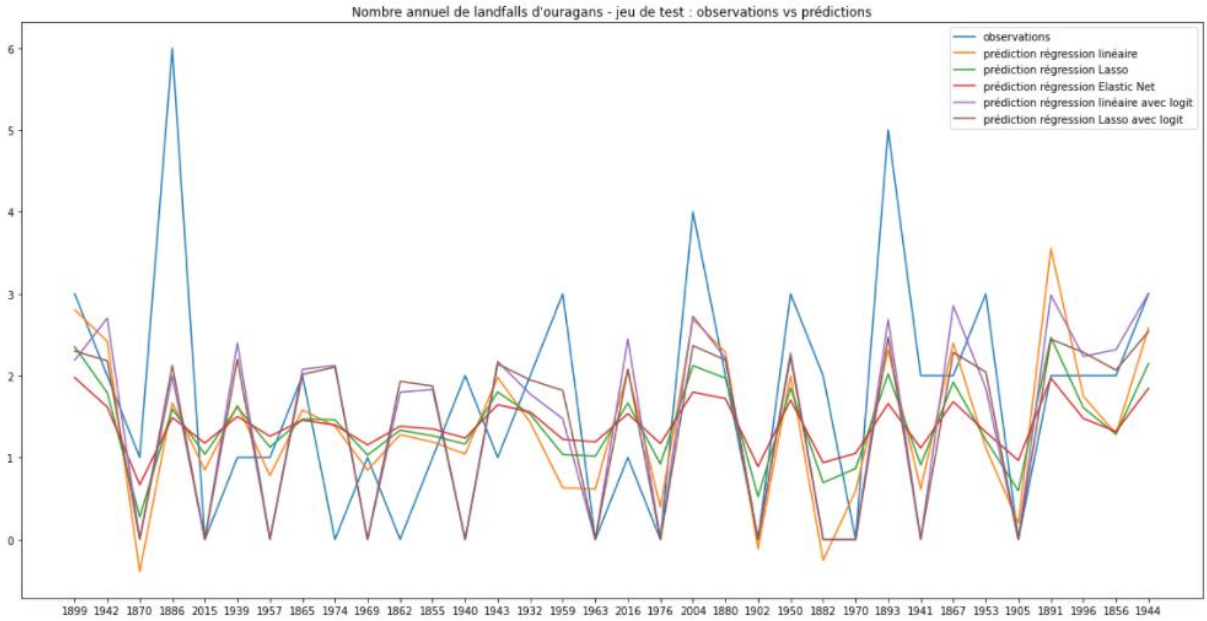

FIGURE 28 : COMPARAISON DES RESULTATS DE PREDICTIONS SUR LE JEU DE TEST ENTRE LES REGRESSIONS LINEAIRES AVEC ET SANS *HURDLE*

On constate bien que globalement les modèles linéaires zéro-modifiés prédisent plus de passages par zéro, dont certains sont corrects et donc souhaitables, par exemple pour les années 2015 et 1963, et d'autres sont erronés comme l'année 1969 ou 1940. En dehors de ces quelques différences, les prédictions semblent relativement proches, ce que nous vérifions également dans les valeurs des MSE de test.

2.2.2.6. K Plus Proches Voisins (kNN)

Pour la modélisation par les k plus proches voisins, nous avons effectué des tests de sensibilité sur le nombre de voisins k, en le faisant varier de 1 à 25 (avec un pas de 1), et sur la pondération ou non par la distance. Nous avons utilisé uniquement la distance euclidienne, la plus couramment utilisée, qui correspond à la distance de Minkowski avec une valeur du paramètre (puissance) égal à 2. C'est également la distance proposée par défaut par *Scikitlearn*.

Dans le cas d'une pondération par la distance, le poids utilisé pour une certaine observation correspond à l'inverse de la distance. Ainsi, les poids sont d'autant plus grands que les voisins sont proches de la cible.

Le graphique suivant donne les MSE calculées sur les jeux d'apprentissage et de test aux différentes valeurs de k testées, avec un modèle kNN avec pondération par la distance euclidienne.

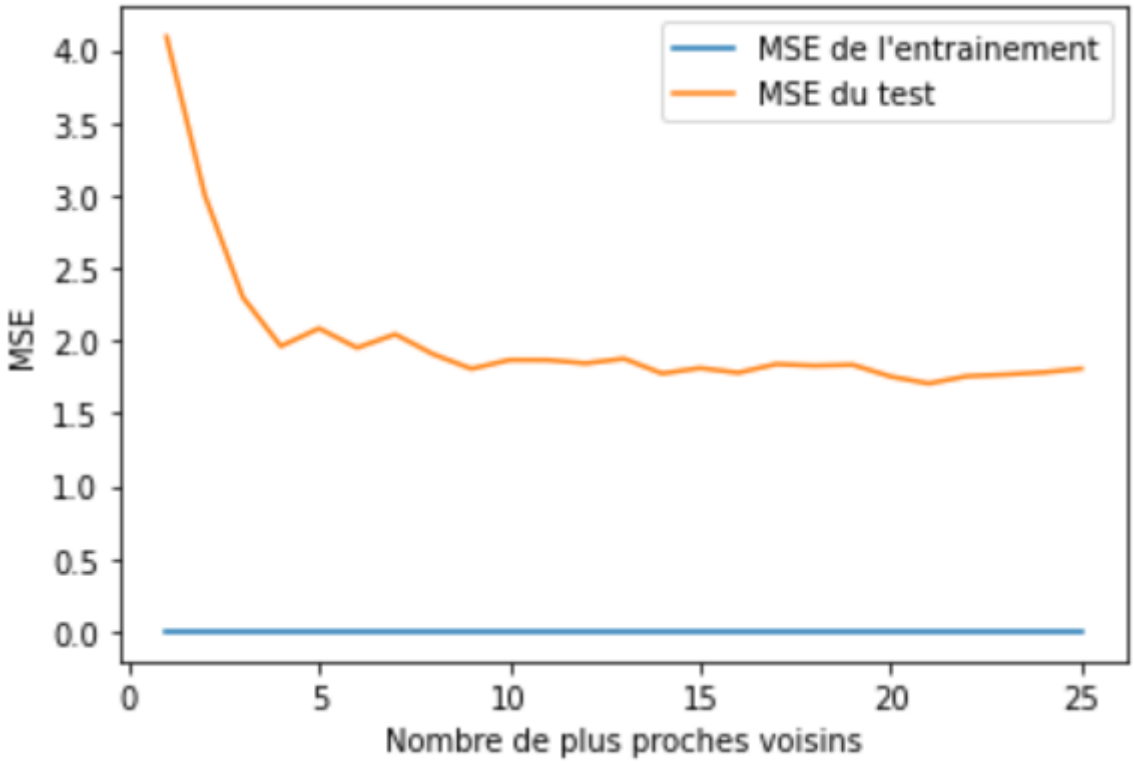

FIGURE 29 : ERREUR QUADRATIQUE MOYENNE EN FONCTION DU NOMBRE DE VOISINS DANS LE MODELE KNN DU NOMBRE D'OURAGANS AVEC PONDERATION PAR LA DISTANCE

Nous avons observé très clairement un surapprentissage dans les cas où une telle pondération est utilisée comme illustré ci-dessous pour le modèle kNN du nombre annuel de *landfalls* d'ouragans calibré sur la sélection de *features* donnant l'erreur quadratique la plus faible.
En effet, la MSE calculée sur le jeu d'apprentissage est strictement nulle, indépendamment du nombre de voisins k.

Par conséquent, nous avons utilisé une régression kNN sans pondération (i.e. avec des poids uniformes) que ce soit pour le nombre annuel de *landfalls* d'ouragans ou celui d'ouragans majeurs. Le graphique suivant donne les MSE calculées sur les jeux d'apprentissage et de test aux différentes valeurs de k testées.

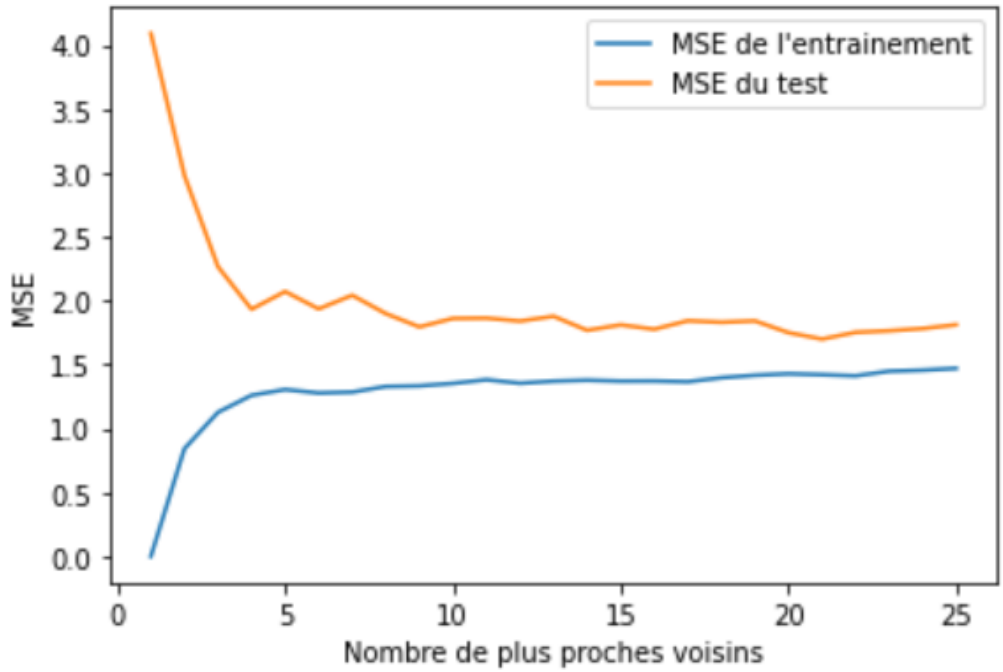

FIGURE 30 : ERREUR QUADRATIQUE MOYENNE EN FONCTION DU NOMBRE DE VOISINS DANS LE MODELE KNN DU NOMBRE D'OURAGANS SANS PONDERATION

On constate qu'il n'y a plus de phénomène de surapprentissage : la MSE sur le jeu d'entraînement est d'abord nulle pour $k = 1$, ce qui est logique, puis augmente avec le nombre de voisins avec un effet asymptote à partir de $k = 10$ environ où la MSE semble ne plus varier de façon significative. Parallèlement, la MSE sur le jeu de test décroît avec le nombre de voisins, avec une forme similaire de courbe mais inversée, et semble également stable à partir de $k = 10$ environ.

Nous avons donc choisi $k = 10$ voisins, qui semble le choix optimal, dans le modèle kNN sans pondération, à la fois pour modéliser le nombre de *landfalls* d'ouragans. Pour celui d'ouragans majeurs, nous avons choisi $k = 15$ car la convergence asymptotique observée est un peu plus lente, comme le montre le graphique ci-dessous.

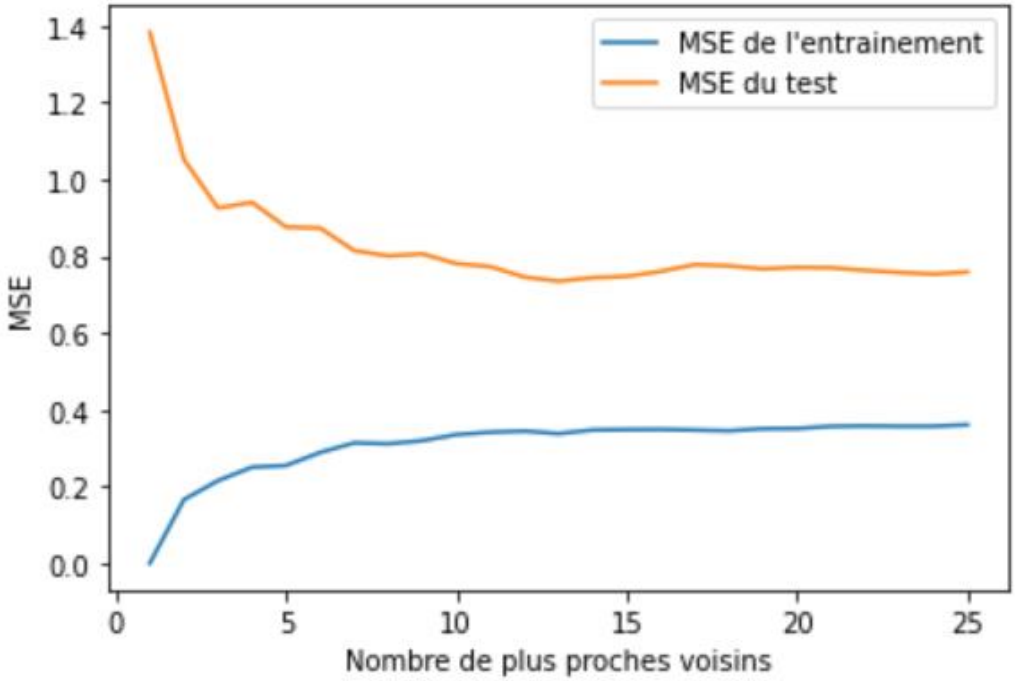

FIGURE 31 : ERREUR QUADRATIQUE MOYENNE EN FONCTION DU NOMBRE DE VOISINS DANS LE MODELE KNN DU NOMBRE D'OURAGANS MAJEURS SANS PONDERATION

2.2.2.7. Modèle linéaire généralisé (GLM)

Le GLM, en faisant appel à une fonction de lien, permet de capter une relation qui n'est pas strictement linéaire entre les variables explicatives et la variable à prédire, tout en gardant une simplicité d'exécution et d'interprétation.

On suppose que conditionnellement aux variables explicatives **X** (matrice), les variables à expliquer **Y** (vecteur) sont indépendantes et identiquement distribuées. Le modèle s'exprime de la façon suivante, avec β le vecteur des paramètres (à estimer) et g la fonction de lien, qui doit nécessairement être bijective :

$$g(E[\boldsymbol{Y}|\boldsymbol{X}]) = g(\mu) = \boldsymbol{X}\,\beta.$$

On fait également l'hypothèse que la distribution de $\boldsymbol{Y}|\mathbf{X}$ suit une loi de la famille exponentielle. A chacune des lois de cette famille est associée une ou plusieurs fonctions de lien, dont une fonction de lien canonique qui permet de relier directement l'espérance $g(\mu)$ au paramètre naturel de la loi.

Compte-tenu de notre problème où nous cherchons à modéliser un nombre (annuel), les choix pertinents sont la loi Binomiale Négative et la loi de Poisson. La Binomiale Négative est *a priori* particulièrement adaptée à la modélisation d'une variable de comptage à forte dispersion, i.e. dont la variance est supérieure à la moyenne, alors que la loi de Poisson est plus appropriée dans le cas plus courant où la variance est inférieure ou égale à la moyenne.

Empiriquement, dans l'analyse historique des données, nous avons observé que les distributions des nombres annuels avaient toute une variance supérieure à la moyenne, ce qui nous conforte *a priori* plutôt sur l'hypothèse de la loi Binomiale Négative.

Nous avons toutefois testé les deux lois, avec la fonction de lien canonique d'abord, qui est la fonction Logarithme Népérien dans les deux cas, puis avec une fonction de lien alternative. Pour la Binomiale Négative, nous avons testé la fonction identité et pour la Poisson la fonction racine carrée (*sqrt*), la fonction identité étant théoriquement utilisable mais donnant une erreur en pratique ici.

Dans le cas de la Binomiale Négative, nous avons aussi réalisé une sensibilité sur le paramètre α, second paramètre de la loi, qui intervient dans le calcul de la variance en facteur multiplicatif de la moyenne au carré. La variance est en effet égale à : $\mu + \alpha\,\mu^2$. Ainsi, plus ce paramètre est élevé, plus la Binomiale Négative modélisée aura une variance élevée. Par défaut, la valeur du paramètre est fixée à 1. Nous avons également testé les valeurs 0,1 et 2.

Nous avons fait intervenir un coefficient constant (souvent appelé *intercept*) dans tous nos modèles GLM.

Enfin, de façon analogue aux modèles linéaires zéro-modifiés nous avons testé la combinaison du GLM avec une régression logistique, respectivement sous Binomiale Négative et Poisson.

Les coefficients des différents modèles GLM testés pour la modélisation du nombre d'ouragans sont fournis en annexe [*Annexe 1 : coefficients des modèles GLM testés pour la modélisation du nombre annuel d'ouragans*].

En pratique, nous avons implémenté les GLM sous Python grâce à la librairie *Statsmodels* (Seabold, 2010), ces modèles n'étant pas disponibles dans *Scikitlearn*. Ainsi, nous ne pouvons pas calculer de façon systématique les métriques d'évaluation, comme la MSE, en validation croisée. Nous ne disposerons donc que des MSE calculées sur le jeu de test.

Nous avons relevé que le GLM renvoie une erreur lorsqu'on tente de le calibrer sur le jeu complet de variables explicatives. C'est le seul type de modèle que nous n'avons pas pu calibrer sur l'ensemble complet des *features*, en plus des différentes sélections. Pour les modèles GLM, la réduction de dimension s'est donc même révélée indispensable.

2.2.2.8. Forêt aléatoire et Gradient Boosting

Comme son nom l'indique, l'algorithme de la forêt aléatoire ou *Random Forest* met en jeu un grand nombre d'arbres de décision. Il faut donc d'abord définir la notion d'arbre.

Le principe général d'un arbre de décision, qu'il soit de classification ou de régression (*Classification And Regression Tree* (CART)), est de partitionner récursivement l'espace d'entrée de façon binaire avec des règles simples. A chaque nœud de l'arbre, deux décisions sont prises par l'algorithme : le choix de la variable explicative et la règle de partitionnement correspondante, par exemple une valeur seuil ou une tranche de valeurs dans le cas d'une variable explicative quantitative. Un arbre de décision est donc construit pour que chaque nœud corresponde à une règle simple. Bâtir un arbre CART se fait en deux étapes. Une première phase est la construction d'un arbre maximal, qui permet de définir la famille de modèles à l'intérieur de laquelle on cherchera à sélectionner le meilleur, et une seconde phase, dite d'élagage (*pruning* en anglais), qui construit une suite de sous-arbres optimaux élagués de l'arbre maximal.

Pour construire un arbre, le choix des facteurs discriminants, utilisés à chaque nœud, est primordial. L'idée est de maximiser le nombre d'éléments à éliminer après chaque question. On crée les questions de haut en bas. Pour choisir les questions optimales, l'algorithme calcule l'entropie, qui est une mesure de dispersion des données, et va retenir les questions ayant la plus faible entropie. L'entropie est d'autant plus faible qu'on peut séparer les données en deux classes bien éloignées.

Pour stopper l'algorithme de partition d'un arbre, il existe plusieurs critères d'arrêt : la profondeur maximale de l'arbre i.e. la distance entre la racine et la feuille la plus éloignée (paramètre *max_depth* dans *Scikitlearn*), le nombre minimal d'observations requis pour séparer un nœud interne (paramètre *min_samples_split*), le nombre minimal d'observations requis sur une feuille (paramètre *min_samples_leaf*), parfois appelée aussi nœud externe du fait qu'une feuille n'a pas d'enfants, ou encore le nombre maximal de feuilles (paramètre *max_leaf_nodes*).
Nous choisissons de prendre un critère d'arrêt basé sur un effectif minimal dans chacune des feuilles : 2% de la base d'apprentissage au moins. Ensuite, nous ajustons avec la profondeur de l'arbre en fonction du niveau de précision.

Dans le cadre d'une forêt, une fois un arbre construit, l'algorithme réitère le procédé pour d'autres *features*. Le principe de la forêt aléatoire est qu'on construit un grand nombre d'arbres et fonctionne par vote majoritaire pour trouver la classe de chaque instance.

Le *Random Forest* comme le *Gradient Boosting* repose sur la technique de *bagging*. Il s'agit d'une technique en intelligence artificielle qui consiste à assembler, en général de façon additive, un grand nombre d'algorithmes avec de faibles performances individuelles pour en créer un beaucoup plus efficace. Les algorithmes de faible performance sont appelés les *weak learners* et le résultat obtenu *strong learner*. L'idée derrière cet algorithme est que plusieurs petits algorithmes peuvent être plus performants qu'un seul grand algorithme. Les *weak learners* peuvent être de différentes natures et avoir des performances variées. L'assemblage des *weak learners* (arbustes) en *strong learner* (forêt) se fait par votation. C'est-à-dire que chaque *weak learner* va émettre une réponse (un vote), et la prédiction du *strong learner* sera la moyenne de toutes les réponses émises.

En pratique, le jeu de données d'apprentissage est découpé en plusieurs sous-ensembles aléatoirement constitués d'échantillons - d'où le caractère *aléatoire* de la forêt.

Sur chaque sous-ensemble, un modèle est entraîné : il y a autant de modèles que de sous-ensembles.

Puis, tous les résultats des modèles sont combinés (avec un système de vote par exemple) ce qui nous donne un résultat final.

Dans le cas du *Gradient Boosting*, les individus les plus difficiles à classifier à l'issue de la première itération se voient affectés un poids plus important, et ceux bien classifiés un poids plus faible. L'arbre calibré à la seconde itération se concentrera donc sur les « insuffisances », i.e. résidus, de l'arbre précédent. L'idée est donc, à chaque étape, d'améliorer les résultats de l'arbre précédent. Les différents

estimateurs ne sont donc pas indépendants, contrairement à la forêt aléatoire. C'est pourquoi on dit parfois que le *Gradient Boosting* « apprend de ses erreurs », à chaque itération. Il peut ainsi souvent donner de meilleurs résultats que le *Random Forest*. En revanche, il est moins rapide car le calcul ne peut pas être parallélisé, chaque estimation dépendant du résultat de la précédente.

Pour les algorithmes *Random forest* et *XGBoost*, nous avons sur effectué, sur le prédicteur le plus performant au sein des différentes sélections de *features* une recherche d'hyperparamètres optimaux (appelée aussi *tuning*) afin de trouver la combinaison d'hyperparamètres minimisant la MSE de validation croisée. Les hyperparamètres que nous avons testés sont la fonction de perte (*loss*) à optimiser (erreur quadratique avec la fonction 'ls' ou erreur absolue avec la fonction 'lad'), le nombre d'estimateurs (nombre d'itérations du *Boosting*), la profondeur maximale de chaque arbre individuel, le nombre minimal d'observations nécessaire pour séparer en deux branches et le nombre minimal d'observations sur une feuille ainsi que le taux d'apprentissage η (*learning rate* ou *shrinkage*). Ce dernier vient pondérer la contribution de chaque estimateur calibré sur les résidus, soit tous les estimateurs sauf le premier.

L'estimateur final, ou *strong learner*, dans l'algorithme *Gradient Boosting* s'écrit donc de la façon suivante :

$$\hat{y} = \hat{y}_1 + \eta \sum_{1<i\leq n} r_i \text{ avec } r_1 = y_1 - \hat{y}_1 \text{ et pour } i \geq 2, r_i = r_{i-1} - \hat{r}_{i-1}.$$

Il y a donc un compromis à trouver entre le nombre d'estimateurs et le taux d'apprentissage η.

Nous ne présentons que les résultats du *Gradient Boosting (GB)* car il donne des résultats légèrement meilleurs que ceux obtenus avec les forêts aléatoires et, comme nous l'avons vu, il limite également fortement le risque de surapprentissage.

Pour le GB, nous avons comme pour les autres régressions d'abord identifié la sélection de *features* minimisant la MSE de validation croisée.

Pour la régression du nombre d'ouragans, il s'agit des 10 variables explicatives les plus corrélées et des 5 les plus anti-corrélées, au sens de Spearman, avec le nombre annuel de *landfalls* d'ouragans.

Pour la régression du nombre d'ouragans majeurs, il s'agit des 20 variables explicatives les plus corrélées et des 5 les plus anti-corrélées, au sens de Spearman, avec le nombre annuel de *landfalls* d'ouragans majeurs.

Dans un second temps, nous avons effectué une recherche d'hyperparamètres optimaux. Le tableau ci-dessous récapitule les différentes valeurs d'hyperparamètres testées, utilisées à la fois pour la régression *GB* du nombre annuel de *landfalls* d'ouragans (abrégé « HU » pour « Hurricane » dans le tableau) et celle du nombre d'ouragans majeurs (« HU major » dans le tableau), ainsi que la valeur optimale trouvée dans chacun des deux cas.

Hyperparamètre	Plage de valeurs testées	Valeur optimale HU	Valeur optimale HU major
Loss	['ls'; 'lad']	'lad'	'ls'
learning_rate	[0,01; 0,1; 0,2]	0,2	0,01
n_estimators	[10; 31; 52; 73; 94; 115; 136; 157; 178; 200]	157	200
max_depth	[3; 8; 13; 18; 23; 29; 34; 39; 44; 50]	18	3
min_samples_split	[1; 2; 3; 4]	3	2
min_samples_leaf	[1; 2; 3; 4]	1	2

TABLE 1 : GRILLE DE VALEURS D'HYPERPARAMETRES TESTEES DANS L'ALGORITHME GRADIENT BOOSTING ET VALEUR OPTIMALES TROUVEES

En pratique, nous avons utilisé la fonction *GridSearchCV*, disponible au sein de la librairie *Scikitlearn*, qui permet de parcourir de façon exhaustive toute la grille d'hyperparamètres définie. Il existe aussi une fonction *RandomizedSearchCV* qui ne teste qu'une sélection aléatoire de valeurs parmi

toutes celles de la grille, ce qui a l'avantage d'être bien plus rapide mais ne donne pas nécessairement la meilleure solution dans l'absolu sur la grille.

Nous avons constaté que l'utilisation des hyperparamètres optimaux permet d'améliorer les résultats de façon visible mais pas très matérielle. Par rapport au paramétrage par défaut, la MSE de validation croisée baisse de 1,74 à 1,49 pour le nombre annuel d'ouragans et de 0,42 à 0,35 pour les ouragans majeurs. La fonction de perte semble être l'un des paramètres le plus sensible pour la régression du nombre d'ouragans.

A noter que nous avons aussi testé l'algorithme *Gradient Boosting* sous forme de classifieur, et non de régresseur, mais celui-ci performe moins bien et donne notamment des MSE de validation croisée plus élevées.

Pour la suite, nous ne conservons pour chacune des variables à prédire, que le modèle GB « optimal », i.e. celui avec le paramétrage, et la sélection de variables explicatives, donnant la MSE de validation croisée la plus faible.

2.2.3. Comparaison des résultats sur le jeu de test

Pour chacun des modèles décrits précédemment, nous avons retenu le meilleur prédicteur au sein de tous ceux calibrés et testés parmi les différentes sélections de *features* et jeux d'hyperparamètres, et comparer leurs résultats. Comme nous ne disposons pas de la MSE de validation croisée pour tous les estimateurs, à savoir tous ceux faisant intervenir une régression logistique (modèles de loi zéro-modifiés) et les GLM, nous faisons cette comparaison sur la MSE du jeu de test, qui est malgré tout un indicateur certes moins robuste mais pertinent.

Les deux tableaux qui suivent récapitulent les résultats respectivement pour la modélisation du nombre annuel de *landfalls* d'ouragans et d'ouragans majeurs. La MSE de test, ainsi que sa racine carrée (*Root MSE* (RMSE)) divisée par la moyenne empirique du nombre annuel, sont affichées pour chacun des modèles retenus.

Model	MSE on test data	RMSE on test data (divided by average)
lr_HU	1,754	0,901
lr_hurdle_HU	1,740	0,897
lr_lasso_HU	1,667	0,878
lr_lasso_hurdle_HU	1,646	0,872
lr_en_HU	1,842	0,923
lr_en_hurdle_HU	1,704	0,888
knn_HU	1,860	0,927
glm_negbin_wCst_HU	1,566	0,851
glm_negbin_sensialpha_wCst_HU	1,461	0,822
glm_negbin_notcano_wCst_HU	1,548	0,846
logit_glm_negbin_wCst_HU	1,592	0,858
glm_poisson_wCst_HU	1,462	0,822
glm_poisson_sqrt_wCst_HU	1,389	0,801
logit_glm_poisson_wCst_HU	1,617	0,865
gb_best_HU	1,514	0,837

TABLE 2 : COMPARAISON RECAPITULATIVE DES RESULTATS DES MEILLEURS ESTIMATEURS DU NOMBRE ANNUEL D'OURAGANS AYANT TOUCHE TERRE

On peut tout d'abord constater que, dans l'absolu, les MSE paraissent toutes assez élevées. Au mieux, notre modèle se trompe en moyenne de 1,39 ouragans par an sur la base de test, soit, rapporté à une moyenne historique de 1,47 ouragans par an touchant terre aux Etats-Unis, une erreur moyenne de 80%. Toutefois, il faut garder en tête que nous proposons ici une modélisation fondée uniquement sur les températures de l'eau comme variables explicatives, ce qui constitue une simplification importante du problème.

En effet, l'activité cyclonique d'une année dépend de nombreux autres paramètres physiques et comporte de plus un caractère aléatoire, qui est d'autant plus fort que l'on s'intéresse aux *landfalls* uniquement. Cette simplification est toutefois assumée, comme expliqué dans le premier chapitre sur les données, et se justifie aussi au regard de l'objectif de prédiction de cette étude.

Si l'on compare les modèles entre eux, on observe que les modèles GLM et GB donnent les meilleurs résultats. Au sein des modèles GLM, la version Binomiale Négative avec $\alpha = 0{,}1$ et la version Poisson, en particulier avec la fonction de lien non canonique racine carrée (*sqrt*), semblent mieux performer. Toutefois, nous privilégierons les GLM avec fonction canonique qui ont notamment l'avantage de permettre le calcul de la moyenne de façon fiable, en tout point du domaine. Côté régression linéaire, c'est le Lasso qui se démarque.

En outre, si l'ajout de la composante logistique (*hurdle*) semble améliorer la performance des modèles linéaires, notamment le Lasso, ce n'est pas le cas pour les GLM.

Ainsi, les meilleurs prédicteurs du nombre d'ouragans sont *a priori* le Lasso, avec et sans *hurdle*, le GLM, en version Binomiale Négative ou Poisson, et le GB.

Les résultats pour le nombre d'ouragans majeurs sont présentés dans le tableau suivant.

Model	MSE on test data	RMSE on test data (divided by average)
lr_HUmajor	0,675	1,746
lr_hurdle_HUmajor	0,911	2,029
lr_lasso_HUmajor	0,691	1,766
lr_lasso_hurdle_HUmajor	0,671	1,741
lr_en_HUmajor	0,710	1,791
lr_en_hurdle_HUmajor	0,669	1,738
knn_HUmajor	0,747	1,836
glm_negbin_wCst_HUmajor	0,750	1,841
glm_negbin_sensialpha_wCst_HUmajor	0,689	1,764
glm_negbin_notcano_wCst_HUmajor	0,646	1,708
logit_glm_negbin_wCst_HUmajor	0,857	1,967
glm_poisson_wCst_HUmajor	0,680	1,753
glm_poisson_sqrt_wCst_HUmajor	0,606	1,655
logit_glm_poisson_wCst_HUmajor	0,863	1,974
gb_best_HUmajor	0,623	1,678

TABLE 3 : COMPARAISON RECAPITULATIVE DES RESULTATS DES MEILLEURS ESTIMATEURS DU NOMBRE ANNUEL D'OURAGANS MAJEURS AYANT TOUCHE TERRE

On peut globalement faire les mêmes observations sur les GLM et le GB, qui se démarque d'ailleurs avec la MSE de test la plus faible. Les différentes régressions linéaires classiques donnent des résultats assez comparables. En revanche, il est notable que les versions zéro-modifiées des modèles Lasso et Elastic Net donnent des meilleures prédictions que leurs homologues sans *hurdle*, avec un petit avantage pour le Lasso là encore.

Concernant les GLM, les versions classiques performent mieux que celles zéro-modifiées.

Ici aussi, les meilleurs prédicteurs semblent être la régression Lasso, dans sa version zéro-modifiée, le GLM, en Binomiale Négative avec $\alpha = 0{,}1$ ou en Poisson, et le GB.

Nous pouvons également apprécier graphiquement la comparaison des prédictions de ces différents modèles sur le jeu de test, par rapport aux données de test originales présentées en bleu.

Nous proposons une représentation graphique pour le nombre d'ouragans et une pour le nombre d'ouragans majeurs. Par souci de lisibilité, nous n'avons conservé qu'une sélection de modèles, incluant les modèles retenus. Pour le GLM, nous avons inclus la version Binomiale Négative par défaut ($\alpha = 1$), celle avec $\alpha = 0{,}1$ et celle zéro-modifiée. Nous n'avons volontairement pas inclus la variante avec loi de Poisson car elle est très proche de celle avec Binomiale Négative.

Il faut bien noter qu'un tel graphique n'est pas chronologique. Il fournit simplement pour chaque année du jeu de test, qui est constitué d'un tirage aléatoire d'années (non ordonnées) du jeu de données intial, la donnée originale d'une part et les prédictions des différents modèles d'autre part.

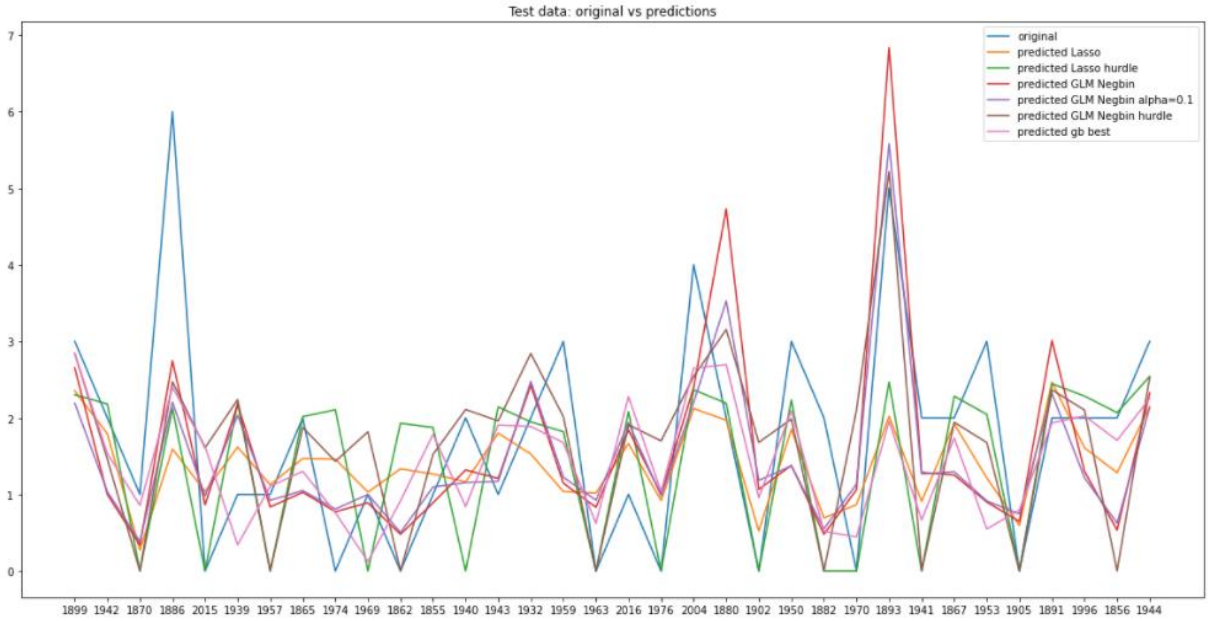

FIGURE 32 : COMPARAISON DES DONNEES DE TEST AUX PREDICTIONS DES DIFFERENTS MODELES POUR LE NOMBRE ANNUEL D'OURAGANS

Il est difficile de décrirer de façon pertinente et exhaustive cette représentative graphique des résultats ou encore d'évaluer graphiquement la performance des différents modèles sur le jeu de test, les métriques de MSE sur le jeu de test analysées plus haut livrant l'information la plus complète et synthétique.
Toutefois, la visualisation graphique permet d'illustrer plus de détail et de comparer les comportements des modèles en des points particuliers.

On peut par exemple remarquer qu'aucun des modèles ne parvient à prédire, à partir des variables explicatives disponibles, le premier pic de 6 ouragans en 1886. La prédiction la plus proche est celle des GLM, qui donnent tous trois une prédicition quasi identique entre 2,5 et 3 ouragans dans l'année. Le GLM avec Binomiale Négative et paramétrage par défaut semble se démarquer toutefois avec une prédiction plus élevée que ces deux variantes, ce qui peut sembler logique car il possède une volatilité plus forte du fait d'une valeur du paramètre α plus élevée.

Le même constat peut être formulé sur le second pic de 5 ouragans en 1893 où la prédiction du GLM par défaut dépasse même ce pic observé, alors que version avec $\alpha = 0{,}1$ et celle zéro-modifiée prédisent des valeurs plus basses et très proches de la valeur cible.
Sur ces pics, les GLM semblent donc être les modèles qui performent le mieux, les régressions linéaires (Lasso) et GB paraissant plus en difficulté à prédire des valeurs aussi élevées.

En revanche, sur la plupart des autres années, qu'il s'agisse de pics moins marqués ou de creux, les régression Lasso et GB donnent des prédictions plus proches des valeurs cibles.
C'est particulièrement visible par exemple sur l'année 1902 où le nombre d'ouragans est nul ; le GB prédit un nombre autour de 1, Lasso autour de 0,5 et le Lasso zéro-modifié la valeur 0.

Sur l'année 1856 (avant dernière sur le graphique), cette fois le nombre observé est égal à 2 et le GB et les Lasso avec et sans *hurdle* prédisent des valeurs entre 1,5 et 2 alors que les GLM prédisent une valeur trop basse autour de 0,5 et le GLM zéro-modifié 0.
On peut d'ailleurs, pour finir, remarquer que les modèles zéro-modifiés, Lasso ou GLM, sont ceux qui passent le plus par zéro, ce qui est logique et attendu. Toutefois, leurs passages par zéro semblent trop fréquents par rapport aux valeurs observées, en particulier sur les années 1940, 1882 et 1856.

L'analyse graphique du nombre d'ouragans sur le jeu de test semble donc plutôt favoriser le Lasso, pour sa régularité, ou le GLM qui suit mieux les pics mais peut se révéler moins performant sur les autres points.

Enfin, analysons graphiquement les résultats de prédiction sur le jeu de test pour le nombre annuel d'ouragans majeurs.

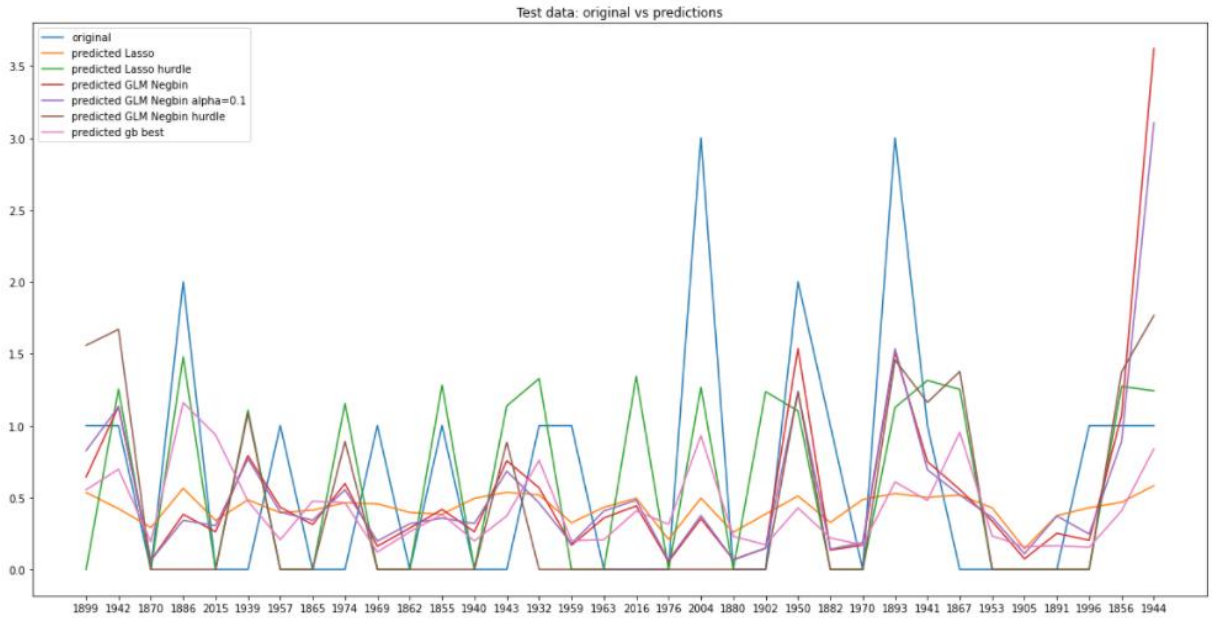

FIGURE 33 : COMPARAISON DES DONNEES DE TEST AUX PREDICTIONS DES DIFFERENTS MODELES POUR LE NOMBRE ANNUEL D'OURAGANS MAJEURS

On peut ici aussi remarquer que les deux pics de trois ouragans majeurs en 2004 et 1893 ne sont bien prédits par aucun des modèles. Pour l'année 1893, les GLM - qui prédisent d'ailleurs tous trois le même nombre - et, dans une moindre mesure, le Lasso zéro-modifié, donnent les prédictions les plus élevées et donc les plus proches de la cible.

On constate aussi le nombre d'ouragans majeurs passe régulièrement par zéro, significativement plus que le nombre d'ouragans, ce qui est logique. Les modèles zéro-modifiés devraient donc a priori mieux performer. C'est bien visible pour le GLM zéro-modifié, qui semble suivre assez fidèlement la majorité des passages par zéro. C'est moins visible en revanche sur le modèle Lasso zéro-modifié car ce dernier passe moins par zéro, ce qui est dû à la sélection de *features* différentes sur laquelle il est calibrée, et qui induit moins de zéros prédits par sa composante de régression logistique.

Le GLM zéro-modifié semble également notablement mieux se comporter que les deux autres modèles GLM sur l'année 1944 (dernière année du graphique) où les deux GLM sans *hurdle* fournissent une prédiction anormalement élevée autour de 3,5 au lieu de 1.

L'analyse graphique semble donc plutôt en faveur de ce dernier en ce qui concerne la prédiction du nombre d'ouragans majeurs.

2.3. Prédiction future

2.3.1. Construction du jeu de test final : projection des températures de l'eau

Nous devons commencer par construire le jeu de test final sur lequel nous souhaitons obtenir une prédiction. Nous allons en réalité construire deux jeux de températures de l'eau projetées, à deux horizons de temps différents, à partir des projections du GIEC les plus récentes disponibles.

Le GIEC est un organisme intergouvernemental ouvert à tous les pays membres de l'Organisation des Nations Unies (ONU) regroupant actuellement 195 Etats. Fondé en 1988, il dépend de l'Organisation Météorologique Mondiale (OMM) et du *Programme des Nations Unies pour l'environnement*. Composé de scientifiques et de représentants d'états, il a pour mission d'évaluer les informations scientifiques, techniques et socio-économiques disponibles, de façon neutre et objective en rapport avec la question du changement du climat. Il ne conduit donc pas de recherche à proprement parler mais synthétise l'ensemble des points de vue de la communauté scientifique et les incertitudes associées. Le budget annuel du GIEC varie entre 5 et 8 millions d'euros. Il est financé par les 195 Etats membres qui y contribuent de manière "indépendante et volontaire". La France par exemple, contribue à hauteur d'un million d'euros jusqu'en 2022.

Un point notable est que le GIEC ne formule pas de recommandations concrètes : il rassemble et synthétise des milliers de travaux scientifiques, afin d'en tirer des conclusions globales et des projections. Il rend accessible et publique cette information au travers de rapports réguliers, appelés rapports d'évaluation, en anglais *Assessment Reports* (AR), ainsi que de rapports *ad-hoc* sur des sujets spécifiques ou en amont de certains sommets. Le GIEC produit un rapport tous les 5 à 7 ans environ, le premier rapport d'évaluation (*First Assessment Report* (FAR)) étant paru en 1990.

Un rapport d'évaluation se décompose en trois documents de contenus scientifiques bien différents. Le rapport, en tant que tel, contient trois volumes de 1500 pages chacun. Sont également produits un rapport de synthèse de 30 à 50 pages ainsi qu'un résumé pour les décideurs de 5 à 10 pages. En pratique, c'est donc souvent ce dernier qui est le plus largement consulté, que ce soit par les populations ou les décideurs politiques.

En réalité, un rapport d'évaluation compile les rendus de trois groupes de travail distincts.
Le premier groupe de travail étudie les aspects scientifiques du changement climatique et analyse les climats passés, présents et futurs.
Le groupe II travaille sur les conséquences du changement climatique ainsi que la vulnérabilité des sociétés humaines, des écosystèmes et des systèmes économiques, et les options d'adaptation.
Enfin, le troisième groupe étudie l'atténuation du changement climatique, l'évaluation des méthodes de réduction des émissions de gaz à effet de serre ainsi que des méthodes d'élimination de gaz à effet de serre dans l'atmosphère.

Les projections les plus récentes disponibles fin 2021 sont issues du sixième rapport d'évaluation publié le 9 août 2021 (GIEC, Sixième rapport d'évaluation (AR6), 2021). Elles sont déclinées selon cinq scénarios, appelés *Trajectoires Socio-économiques Partagées* (TSP) ou *Shared Socioeconomic Pathways* (SSP) en anglais. Ils présentent diverses caractéristiques de développements socio-économiques et stratégies d'adaptation et d'atténuation jusqu'à 2100. Les trajectoires SSP sont décrites dans le tableau ci-dessous issu de l'*Institute for Climate Economics* (I4CE, 2021), de la moins émettrice à la plus émettrice.

Scenario	Narratif
SSP1 Développement durable	Forte coopération internationale, priorité donnée au développement durable, amélioration des conditions de vie et préférences des consommateurs pour des biens et services respectueux de l'environnement, peu intensifs en ressources et en énergie.
SSP2 Poursuite des tendances	Les tendances sociales, économiques et technologiques actuelles se poursuivent, le développement et la croissance progressent de manière inégale selon les pays et les régions. Les institutions nationales et internationales œuvrent à la réalisation des objectifs de développement durable qui progresse lentement. L'environnement se dégrade malgré un développement moins intensif en ressources et en énergie.
SSP3 Rivalités régionales	Résurgence des nationalismes, développement économique lent, persistance des inégalités et des conflits régionaux. Les pays sont guidés par des préoccupations en matière de sécurité et de compétitivité. Ils se concentrent sur les problèmes nationaux voire régionaux et sur les enjeux de sécurité alimentaire et énergétique. Faible priorité internationale pour la protection de l'environnement, qui se dégrade fortement dans certaines régions.
SSP4 Inégalités	Développement marqué par de fortes inégalités entre et à l'intérieur des pays. Dégradation de la cohésion sociale et multiplication des conflits. Fossé croissant entre une élite connectée et mondialisée, responsable de la majorité des émissions de GES, et un ensemble fragmenté de populations à faible revenu, peu éduqué et vulnérable au changement climatique. Le secteur énergétique se diversifie entre sources d'énergie fortement carbonées et décarbonées. Les politiques environnementales se concentrent sur les enjeux locaux.
SSP5 Développement conventionnel	Développement adossé à l'exploitation forte des énergies fossiles et marqué par des investissements élevés dans la santé, l'éducation et les nouvelles technologies. Adoption de modes de vie intensifs en ressources et en énergie à travers le monde. La croissance économique et le progrès technologique sont élevés. Les problèmes de pollution locale sont bien gérés et l'adaptation est facilitée notamment grâce au recul de la pauvreté.

TABLE 4 : DESCRIPTION DES CINQ TRAJECTOIRES SSP UTILISEES PAR LE GIEC (I4CE, 2019)

Les SSP couvrent un large spectre de futurs plausibles et leur objectif n'est pas de prédire l'avenir, aucune probabilité ne leur est d'ailleurs affecté, mais de servir de base de réflexion aux Etats et aux populations, et potentiellement d'aide à la décision.

A chacune de ces trajectoires socio-économiques de numéro *x* est associé un niveau de forçage radiatif *y* en 2100, exprimé en W/m². Chaque trajectoire porte un nom de la forme SSPx-y. Le graphique ci-dessous, extrait de l'I4CE, illustre l'évolution du niveau d'émission de dioxyde de carbone (CO_2) en giga tonnes (Gt) par an pour les différentes trajectoires SSP.

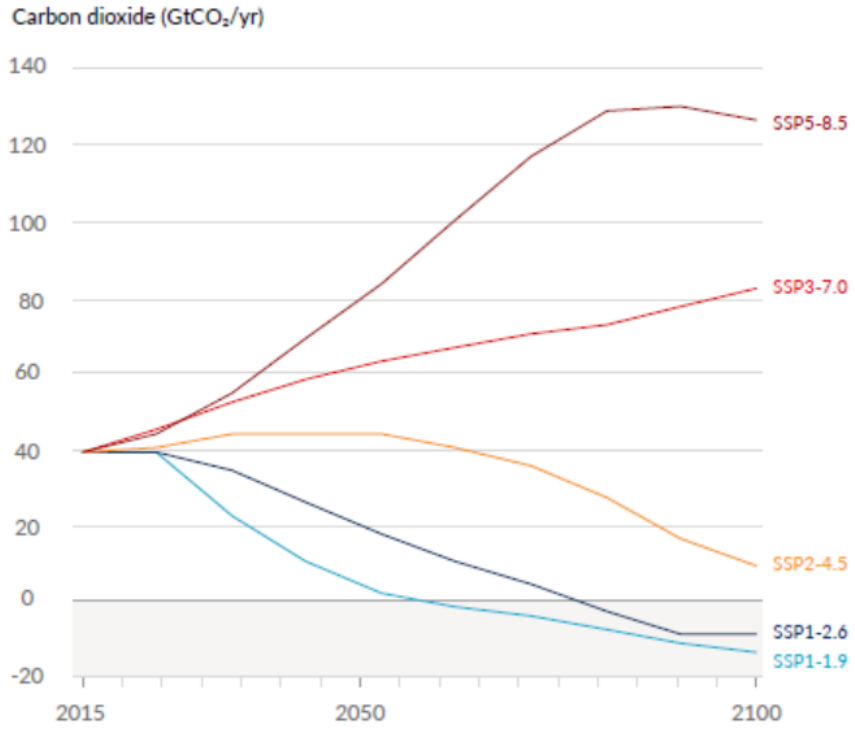

FIGURE 34 : EVOLUTION DU NIVEAU D'EMISSION DE CO2 JUSQU'EN 2100 EN FONCTION DE LA TRAJECTOIRE SSP (I4CE, 2021)

La trajectoire SSP1-1.9 correspond aux engagements de l'accord de Paris (2015, COP21), entré en vigueur le 4 novembre 2016, de contenir l'augmentation de température à 1,5°C. Elle est très ambitieuse et projette une réduction significative et rapide des émissions de CO_2 allant jusqu'à l'atteinte d'une neutralité carbone peu après 2050.

A l'autre extrême, la SSP5-8.5 projette une hausse forte et continue des émissions jusqu'en 2090 environ où semble s'amorcer ensuite une légère baisse.

Entre les deux, la SSP2-4.5 projette d'abord une légère augmentation des émissions jusqu'en 2050 avant d'amorcer une baisse continue jusqu'en 2100.
Il faut toutefois noter que, si les niveaux d'émissions baissent, il n'en est pas de même – en tout cas pas immédiatement – pour les projections de températures comme illustré dans le graphique ci-dessous construit par l'I4CE à partir de du sixième rapport du GIEC. En effet, il existe une certaine inertie du système et la durée de vie du CO_2 dans l'atmosphère est très longue, supérieure à 500 ans. Même pour le méthane (CH_4), plus puissant mais plus éphémère, la durée de vie est estimée à une dizaine d'années minimum.

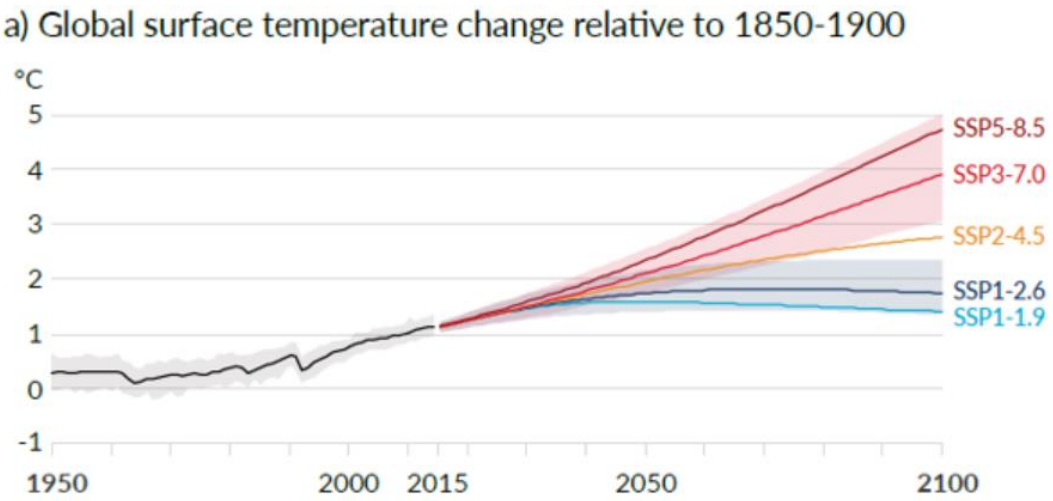

FIGURE 35 : PROJECTIONS D'AUGMENTATION GLOBALE DE TEMPERATURE A LA SURFACE DE LA TERRE JUSQU'EN 2100, PAR RAPPORT A LA PERIODE 1850-1900 EN FONCTION DE LA TRAJECTOIRE SSP (I4CE, 2021 D'APRES GIEC, AR6, 2021)

Par simplification ce sont les augmentations moyennes de températures (°C) qui sont tracées dans ce graphique, avec pour les scénarios SSSP1-2.6 et SSP3-7.0 un faisceau autour de la moyenne délimité par les quantiles 5% et 95% donnés dans les prévisions du GIEC.

Dans tous les scénarios d'émissions (à l'exception du plus bas, le SSP1-1.9), nous dépasserons le seuil de réchauffement mondial de +1,5°C dans un avenir proche, entre 2022 et 2040.

Le précédent rapport du GIEC se fondait sur d'autres scénarios dits *Representative Concentration Pathways (*RCP), également au nombre de cinq et nommés selon le niveau de forçage radiatif (en W/m²). Les nouveaux scénarios SSP sont a priori plus complets et plus précis car ils permettent de combiner un niveau d'émission radiatif et une stratégie d'atténuation, avec des effets éventuellement différenciés dans le temps. Comme le signale l'I4CE, il faut être prudent lorsque l'on compare un scénario RCP et un scénario SSP, même à niveau égal de forçage radiatif en 2100. En effet, ils ne sont pas directement comparables car la répartition des émissions de gaz à effet de serre dans le temps ainsi que la proportion des émissions par type de gaz peuvent différer nettement. C'est ainsi que « par exemple, le SSP5-8.5 a des concentrations en CO2 plus importantes que le RCP 8.5, mais des concentrations en méthane plus faibles. »
Pour la suite et notre objectif de prédiction de la fréquence d'ouragans dans un monde plus chaud, nous choisissions de ne retenir que les scénarios SSP2-4.5 et SSP5-8.5 afin de couvrir un éventail assez large de trajectoires émettrices.

Nous avons obtenu les projections de températures à la surface de l'eau de l'AR6 aux deux horizons de projection sur l'atlas interactif en ligne (GIEC, IPCC Interactive Atlas, 2021) réalisé en collaboration avec la NASA (*National Aeronautics and Space Administration*, en français Administration nationale de l'aéronautique et de l'espace). Celui-ci permet de sélectionner différentes grandeurs physiques, l'horizon de projection, la SSP, la *baseline* i.e. la période de référence à partir de laquelle les variations sont calculées, et offre également une visualisation sur le globe.

Le tableau suivant récapitule les augmentations de température globale (en °C) à la surface de l'eau attendues dans les trajectoires SSP5-8.5 et SSP2-4.5, aux deux horizons de projection, et détaille à la fois la médiane, l'écart-type et les principaux quantiles des distributions d'augmentation.

Horizon de projection	Scénario	Médiane	P25	P75	P10	P90	P5	P95	Ecart-type
Medium Term (2041-2060)	**SSP5-8.5**	**1,3**	**1**	**1,5**	**0,9**	**1,7**	**0,9**	**1,9**	**0,36**
Long Term (2081-2100)	**SSP5-8.5**	**2,9**	**2,1**	**3,4**	**2**	**3,8**	**1,9**	**4,2**	**0,79**
Medium Term (2041-2060)	SSP2-4.5	1	0,8	1,2	0,7	1,4	0,7	1,5	0,30
Long Term (2081-2100)	SSP2-4.5	1,6	1,2	1,9	1,1	2,1	1,1	2,3	0,43

TABLE 5 : PROJECTIONS D'AUGMENTATION GLOBALE DE TEMPERATURE DE L'EAU A HORIZONS 2021-2040 ET 2081-2020, PAR RAPPORT A LA PERIODE 1981-2010 (GIEC, AR6, 2021)

Nous allons construire à partir de ces projections deux jeux de test correspondant aux deux horizons, moyen terme et long terme. Nous nous plaçons dans le cadre de la trajectoire SSP la plus émettrice, i.e. SSP5-8.5, afin d'avoir l'estimation la plus prudente des pertes assurantielles sur ces deux périodes de projection. C'est pourquoi les données afférentes à cette trajectoire sont en gras dans le tableau ci-dessus.

Les projections présentées sont établies par rapport à la *baseline* 1981-2010. Nous avons volontairement choisi cette période de référence plus récente que la *baseline* 1850-1900, souvent utilisée par défaut, afin d'appliquer les projections d'augmentation aux données historiques les plus récentes de notre base de données qui ont *a priori* une qualité de donnée bien meilleure.

Comme nous ne disposons pas de projections différenciées par région sein de l'Atlantique Nord, nous faisons l'hypothèse simplificatrice que l'augmentation est homogène sur toute la zone d'étude.

Plusieurs options s'offrent alors à nous pour construire les deux jeux de test projetés.

La première consiste à appliquer, respectivement sur chacun des horizons, le choc médian aux températures de l'eau médianes (ou moyennes) observées entre 1981 et 2010, au sein de chaque carré spatial. Cette méthode a l'avantage d'être la plus simple et ne donne qu'une seule prédiction qui correspond à la médiane de chacune des périodes de projection. Toutefois, elle n'utilise pas la totalité de l'information disponible dans les données initiales, puisqu'elle ne considère que la moyenne et non le « faisceau » constitué par l'ensemble des années entre 1981 et 2010, ni dans les données de projection, car elle n'utilise que le choc médian et non les autres informations de la distribution du choc (quantiles et écart-type).

Une seconde option consiste donc à faire l'hypothèse que la distribution de choc est gaussienne, de moyenne égale à la médiane donnée – puisque pour une loi normale moyenne et médiane sont égales – et d'écart-type égal à l'écart-type donné par le GIEC et d'appliquer ce choc gaussien à chacune des années observées entre 1981 et 2010. Nous retenons cette seconde méthode qui permet d'exploiter toute la richesse de l'information statistique disponible et de conserver un caractère stochastique dans la projection des températures. En pratique, pour chacun des horizons, nous effectuons un tirage de loi normale de moyenne et d'écart-type fournis, et pour chaque année entre 1981 et 2010, que nous ajoutons aux températures de l'eau de l'année, pour chaque carré spatial. Ce tirage est en fait une matrice, de dimension 30 (années) x 948 (*features*). Chaque cellule de cette matrice correspond à un tirage différent de loi normale à une dimension de moyenne et d'écart-type donnés. Nous obtenons ainsi un faisceau de 30 réalisations d'années projetées crédibles pour chacune des périodes 2041-2060 et 2081-2100.

2.3.2. Entraînement sur la totalité de l'historique

Maintenant que nous disposons d'une base d'apprentissage complète et normalisée et des deux bases de test, nous allons entraîner les modèles sélectionnés sur l'historique complet (1851-2020) et les faire prédire sur chacun de ces jeux de test projetés. Nous obtiendrons alors une prédiction du nombre annuel moyen d'ouragans et d'ouragans majeurs à ces deux périodes, moyen et long terme, longues de vingt ans chacune.

Ces prédictions seront donc comparables davantage à une moyenne calculée sur un horizon passé, idéalement long de vingt ans au moins, par exemple les vingt dernières années, plutôt qu'un nombre d'ouragans sur une année précise. En effet, nous cherchons bien à prédire une tendance de la fréquence d'ouragans dans un monde plus chaud, et nous n'avons pas ici vocation à faire une prédiction précise pour une année, ce qui nécessiterait d'ailleurs de disposer de projections de températures précises par année, ce que nous n'avons pas.

En vue de la prédiction finale, nous entraînons donc les modèles sélectionnés sur une base d'apprentissage correspondant à la totalité de l'historique disponible 1851-2020. Nous repartons de la base initiale des températures de l'eau, nettoyée et filtrée sur la zone géographique d'intérêt, mais avant standardisation des variables explicatives et imputation des données manquantes.

Comme nous l'avons fait précédemment pour la calibration et la validation de ces mêmes modèles, nous avons procédé à l'imputation des données manquantes et au centrage-réduction des variables explicatives, en utilisant cette fois la totalité de l'information historique disponible sans risque de fuite de données (*leak*).

Nous avons pris soin d'appliquer les chocs, pour construire les jeux de test, sur les données historiques après l'imputation, pour ne pas avoir de valeurs manquantes, mais avant la standardisation, puisque les projections d'augmentation sont définies par rapport à des températures originales (non standardisées) en degré Celsius.

2.3.3. Prédiction des nombres annuels d'ouragans et d'ouragans majeurs

Nous présentons ci-dessous les résultats pour les modèles identifiés précédemment, donnant la meilleure prédiction sur le jeu de test, ainsi que pour le kNN pour information, d'abord pour le nombre annuel de *landfalls* d'ouragans puis pour le nombre d'ouragans majeurs, aux horizons moyen terme (MT) et long terme (LT). Par construction, chaque horizon est décrit par trente années. On présente ici la moyenne, l'écart-type et les principaux quantiles de la distribution des prédictions sur ces trente années. En outre, nous affichons également, à titre de référence, les mêmes indicateurs statistiques observés sur l'historique complet (1851-2020), sur l'historique *baseline* de nos projections du GIEC (1981-2010) et sur les vingt dernières années (2001-2020).

Stat.	1851-2020	1981-2010	2001-2020	Lasso MT	Lasso LT	Lasso *hurdle* MT	Lasso *hurdle* LT	GLM MT	GLM LT	kNN MT	kNN LT	GB MT	GB LT
count	170	30	20	30	30	30	30	30	30	30	30	30	30
mean	1,47	1,23	1,25	1,62	1,84	1,68	1,59	2,10	3,63	1,62	1,55	2,41	2,46
std	1,35	1,41	1,59	0,41	0,53	0,87	1,01	0,58	1,72	0,20	0,07	0,65	0,19
min	0,00	0,00	0,00	0,78	0,91	0,00	0,00	1,11	1,39	1,20	1,40	1,12	2,25
25%	0,00	0,00	0,00	1,33	1,47	1,77	0,36	1,67	2,12	1,50	1,50	2,04	2,44
50%	1,00	1,00	1,00	1,58	1,80	2,04	1,96	2,12	3,39	1,60	1,50	2,37	2,44
75%	2,00	2,00	2,00	1,85	2,11	2,16	2,27	2,42	4,91	1,70	1,60	2,78	2,45
max	6,00	6,00	6,00	2,55	3,05	2,49	2,83	3,44	8,18	2,30	1,70	3,81	3,41

TABLE 6 : RESULTATS DE PREDICTION DU NOMBRE D'OURAGANS AUX ETATS-UNIS A MOYEN (MT) ET LONG TERME (LT)

On peut tout d'abord remarquer que toutes les prédictions, que ce soit à MT ou LT, sont en moyenne au-dessus de la moyenne empirique sur l'historique complet 1851-2020 et *a fortiori* sur les vingt dernières années ou sur la *baseline*. Une augmentation de température de l'eau mène donc à une augmentation du nombre d'ouragans aux les Etats-Unis, aux deux horizons MT et LT.

En outre, la prédiction moyenne à LT est supérieure à la prédiction MT pour le Lasso, le GLM et le GB, ce qui semble *a priori* conforme à l'intuition, mais pas pour le Lasso zéro-modifié et le kNN.

Le kNN n'est pas un modèle adéquat pour notre problème car il a des difficultés à prédire sur les années projetées qui sont situées en dehors de son domaine d'apprentissage, puisqu'il ne raisonne que par ressemblance avec les observations dont il dispose. L'écart-type de ces prédictions semble anormalement faible. Le GB est confronté au même écueil, limité par son domaine d'apprentissage. Même si la prédiction LT est supérieure à la MT, son écart-type est également très faible, dans l'absolu et relativement au Lasso et au GLM.

Nous avons analysé en détail les prédictions année par année, pour chacun des horizons, du kNN et du GB. Il est particulièrement visible sur l'horizon LT que les prédictions de nombre annuel du kNN, respectivement du GB, sont quasi identiques au sein des trente années projetées, comme l'on peut le voir dans les résultats de prédiction détaillés présentés ci-dessous.

Année	Lasso MT	Lasso LT	EN MT	EN LT	Lasso hurdle MT	Lasso hurdle LT	GLM MT	GLM LT	GLM logit MT	GLM logit LT	GLM Poisson MT	GLM Poisson LT	GB MT	GB LT	KNN MT	KNN LT
1981	0.99	1.74	1.13	1.63	0.00	1.92	1.47	1.55	1.70	1.27	1.47	1.57	1.12	2.47	1.70	1.70
1982	1.48	0.91	1.42	1.27	2.05	1.44	1.11	1.55	1.62	1.23	1.11	1.64	2.69	2.45	1.50	1.60
1983	1.54	2.28	1.40	1.80	2.07	2.19	1.16	2.07	1.48	1.40	1.17	2.06	3.81	2.44	1.40	1.40
1984	2.18	1.72	1.93	1.75	2.29	2.21	1.72	5.51	1.82	2.33	1.72	5.28	2.32	2.38	1.60	1.50
1985	1.64	1.36	1.51	1.34	2.17	0.00	2.16	2.54	2.17	1.88	2.22	2.60	2.13	2.25	1.60	1.50
1986	1.69	1.46	1.63	1.68	2.02	2.07	2.56	8.18	2.43	4.04	2.65	8.47	2.38	2.43	1.60	1.60
1987	1.13	1.49	1.31	1.62	1.76	1.76	2.09	5.49	2.57	3.25	2.12	5.79	2.24	2.44	1.20	1.60
1988	1.91	2.73	1.70	2.18	2.19	2.49	1.81	3.57	1.90	3.89	1.79	3.59	3.03	3.41	1.70	1.70
1989	2.15	2.11	1.89	1.86	2.47	2.34	2.59	4.16	1.99	2.35	2.50	4.25	2.36	2.45	1.70	1.60
1990	1.55	1.94	1.49	1.60	2.07	0.00	2.43	2.27	2.22	2.24	2.48	2.27	2.25	2.45	1.60	1.50
1991	1.32	1.35	1.45	1.31	1.80	0.00	2.62	1.59	1.96	1.16	2.60	1.60	2.26	2.44	1.60	1.50
1992	1.52	1.82	1.53	1.71	2.02	1.93	2.30	3.11	2.57	2.52	2.37	3.32	3.34	2.31	1.60	1.60
1993	1.27	0.98	1.35	1.39	0.00	0.00	1.26	1.39	0.00	0.00	1.32	1.46	3.00	2.44	1.50	1.60
1994	1.20	1.99	1.34	1.70	2.01	2.13	2.38	2.92	2.57	2.53	2.40	3.05	2.00	2.47	1.50	1.50
1995	1.82	2.38	1.67	1.91	2.09	2.30	3.44	6.94	3.18	4.17	3.53	7.32	2.54	2.43	1.70	1.50
1996	1.86	1.78	1.67	1.59	2.37	2.26	2.32	2.04	2.11	1.87	2.29	2.10	2.58	2.44	1.70	1.50
1997	1.62	1.89	1.60	1.75	1.97	1.80	1.95	3.21	2.06	2.40	1.97	3.20	1.96	2.43	1.30	1.50
1998	1.83	2.14	1.75	1.82	2.01	2.37	2.86	2.92	2.18	2.39	2.96	3.01	3.62	2.44	1.60	1.50
1999	1.98	1.78	1.82	1.61	2.16	1.86	2.34	1.95	2.46	1.74	2.32	2.02	1.91	2.44	1.80	1.50
2000	1.16	2.88	1.43	2.25	1.76	2.83	1.76	4.84	1.82	2.06	1.78	4.65	2.52	2.44	1.50	1.60
2001	1.57	1.91	1.51	1.82	2.11	2.22	3.33	5.60	2.76	3.83	3.30	5.81	2.97	2.47	1.60	1.70
2002	0.78	1.07	1.09	1.25	0.00	0.00	1.66	3.69	0.00	2.67	1.71	3.76	1.98	2.53	1.50	1.50
2003	1.44	2.10	1.47	1.79	1.90	2.54	1.62	2.87	1.83	1.59	1.67	2.86	2.27	2.44	1.60	1.60
2004	2.55	3.05	2.13	2.38	2.39	2.74	2.19	4.93	2.04	4.16	2.20	5.05	2.47	2.44	2.30	1.60
2005	1.59	1.74	1.50	1.61	2.13	2.28	1.80	3.77	2.18	1.94	1.78	3.63	2.79	2.44	1.80	1.50
2006	1.36	1.49	1.31	1.42	0.00	0.00	1.94	1.81	1.87	1.32	1.97	1.80	1.24	2.42	1.50	1.50
2007	1.18	1.43	1.27	1.31	0.00	0.00	1.49	3.74	1.83	2.17	1.53	4.00	1.30	2.43	1.60	1.50
2008	2.16	2.40	1.85	1.96	2.16	1.90	2.60	5.52	2.73	3.57	2.61	5.62	2.76	2.44	1.60	1.50
2009	1.85	1.30	1.57	1.42	0.00	0.00	1.62	4.10	1.80	2.46	1.60	4.29	1.46	2.45	1.80	1.50
2010	2.31	2.11	1.89	1.90	2.49	1.98	2.35	4.99	2.37	2.85	2.32	5.20	2.96	2.44	2.00	1.60
Moyenne	**1.62**	**1.84**	**1.55**	**1.69**	**1.68**	**1.59**	**2.10**	**3.63**	**2.01**	**2.38**	**2.12**	**3.71**	**2.41**	**2.46**	**1.62**	**1.55**
Ecart-type	**0.41**	**0.53**	**0.24**	**0.28**	**0.87**	**1.01**	**0.58**	**1.72**	**0.67**	**1.01**	**0.58**	**1.78**	**0.65**	**0.19**	**0.20**	**0.07**

TABLE 7 : PREDICTIONS DES DIFFERENTS MODELES A MOYEN TERME (MT) ET LONG TERME (LT) A PARTIR DES PROJECTIONS D'AUGMENTATION DE TEMPERATURES DU GIEC PAR RAPPORT A LA PERIODE 1981-2010 DANS LA TRAJECTOIRE SSP5-8.5 (OBTENUES AVEC LA GRAINE ALEATOIRE SEED=1)

La valeur maximale relevée sur les données de température de l'eau projetées, en fonction des points géographiques et des mois, est quasi toujours supérieure à la valeur maximale constatée sur l'historique complet utilisé pour l'apprentissage, comme illustré en dernière ligne du tableau ci-dessous donnant les températures projetées à moyen terme, arbitrairement sur les 5 premiers *features* à titre d'exemple.

Année projetée	-47,5&7,5&1	-37,5&7,5&1	-32,5&7,5&1	-27,5&7,5&1	-22,5&7,5&1
1981	29,18	27,64	27,94	27,75	28,38
1982	29,11	28,11	28,58	28,35	29,03
1983	28,15	27,78	28,60	27,93	28,36
1984	28,31	28,15	28,29	28,84	29,71
1985	28,47	28,41	27,85	28,52	29,13
1986	28,26	27,69	28,73	28,63	28,35
1987	27,97	27,46	27,72	27,51	28,52
1988	29,24	29,27	28,27	28,84	29,77
1989	28,62	28,33	28,39	28,01	28,77
1990	28,97	29,52	28,56	27,96	28,51
1991	29,05	28,29	28,54	28,49	28,53
1992	28,55	28,60	28,22	28,70	28,31
1993	28,40	28,11	27,90	27,95	28,89
1994	28,83	29,26	28,65	28,33	28,58
1995	28,18	28,98	27,84	28,83	29,51
1996	29,35	29,98	29,71	28,97	29,49
1997	28,96	28,66	28,71	28,92	28,94
1998	29,43	28,42	28,42	28,56	28,42
1999	29,48	27,96	28,57	29,19	29,43
2000	29,26	29,39	28,72	28,70	28,60
2001	28,08	27,96	28,19	27,72	29,61
2002	29,03	28,80	28,67	29,01	28,80
2003	27,78	28,56	28,19	28,71	29,19
2004	29,16	28,54	28,68	29,11	28,16
2005	29,54	29,33	28,45	29,27	29,28
2006	29,33	29,41	29,02	28,75	29,32
2007	29,40	29,32	29,26	29,09	29,41
2008	29,06	28,42	28,18	28,57	28,62
2009	29,50	29,30	29,13	28,36	29,18
2010	28,83	29,05	28,39	28,12	29,23
Max	**29,54**	**29,98**	**29,71**	**29,27**	**29,77**
Max full histo	**28,54**	**28,95**	**28,64**	**29,08**	**28,32**
Différence	***3,5%***	***3,5%***	***3,7%***	***0,7%***	***5,1%***

TABLE 8 : EXTRAIT DES PROJECTIONS DE TEMPERATURE DE L'EAU A MOYEN TERME (MT) A PARTIR DES DONNEES DU GIEC

Ainsi, sur l'ensemble des *features*, la température maximale projetée à MT est en moyenne 3% plus élevée (avec des variations de températures maximales entre -15% et 13%) et 12% plus élevée à LT (avec des variations de températures maximales entre -6% et 26%).

C'est en effet une des limitations des algorithmes d'apprentissage supervisé qui ne parviennent pas efficacement à prédire en dehors de leur domaine d'apprentissage. Les modèles linéaires et linéaires généralisés ne présentent pas ce problème et ont au contraire, par nature, la capacité d'extrapoler à des valeurs en dehors du domaine d'entraînement. Ainsi, on privilégie pour la suite ces derniers qui ont également l'avantage d'être plus facilement lisibles et interprétables comme nous l'avons déjà mentionné.

Pour le GLM, on n'affiche ici que la version Binomiale Négative (avec $\alpha = 0{,}1$) car la version Poisson donne des résultats très proches et nous préférons, d'un point de vue statistique, le choix de la Binomiale Négative pour modéliser une telle variable aléatoire très volatile, comme expliqué plus haut. Le GLM donne des prédictions de nombre annuel d'ouragans notablement plus hautes que le modèle linéaire Lasso qu'il soit avec ou sans *hurdle*, et ce aux deux horizons de projection.

Le Lasso zéro-modifié fournit des prédictions à MT légèrement supérieures, à la fois en moyenne et sur les différents quantiles, à celles du Lasso classique, avec toutefois une valeur minimale nulle comme attendu du fait de la pondération par la probabilité de survenance (binaire) modélisée par une régression logistique. En revanche, sur l'horizon LT, il donne des prédictions bien plus basses, à la fois en moyenne et sur les différents quantiles. Deux effets peuvent expliquer les différences entre les résultats du modèle classique et de son homologue zéro-modifié.

Tout d'abord, les prédictions de probabilité de survenance données par la régression logistique ne sont pas identiques sur les deux horizons MT et LT, comme illustré sur le graphique ci-dessous.

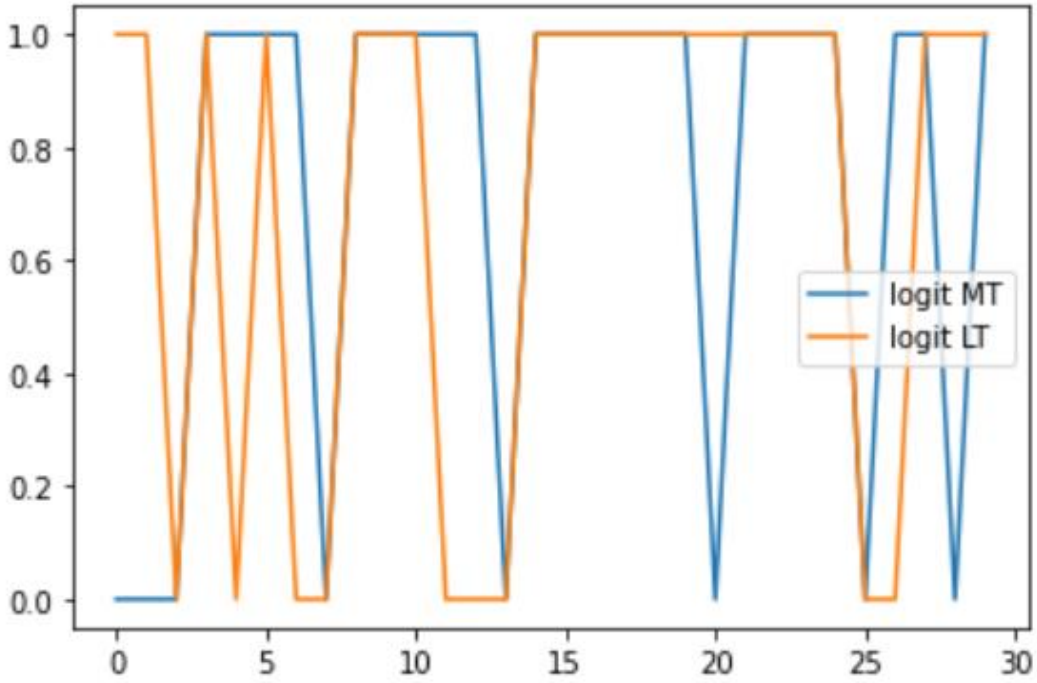

FIGURE 36 : PREDICTIONS DE LA REGRESSION LOGISTIQUE DES MODELES ZEROS-MODIFIES, AUX HORIZONS MOYEN ET LONG TERME

En effet, nous calibrons deux régressions logistiques distinctes entraînées toutes deux sur l'historique complet mais on utilise dans le premier cas la prédiction de cette régression sur les données de températures projetées à moyen terme et dans le second la prédiction à long terme. On remarque ici que la prédiction de la probabilité d'avoir au moins un *landfall* passe 8 fois par zéro à MT et 9 fois à LT et certains passages par zéro ne se font pas sur les mêmes années. Ainsi, les passages par zéro peuvent davantage coïncider avec des années où la prédiction Lasso est relativement élevée et le résultat final, égal au produit de la prédiction logistique et de la prédiction Lasso, peut donc globalement être impacté à la baisse en moyenne sur l'ensemble de la période de prédiction.

La seconde source de différence vient du fait que la régression Lasso au sein du modèle zéro-modifié est calibrée uniquement sur les observations présentant un nombre (variable à prédire) strictement positive.

Les résultats pour le nombre annuel d'ouragans majeurs atteignant les Etats-Unis sont synthétisés dans le tableau ci-dessous.

Stat.	1851-2020	1981-2010	2001-2020	Lasso MT	Lasso LT	Lasso *hurdle* MT	Lasso *hurdle* LT	GLM MT	GLM LT	GLM *hurdle* MT	GLM *hurdle* LT	kNN MT	kNN LT	GB MT	GB LT
count	170	30	20	30	30	30	30	30	30	30	30	30	30	30	30
mean	0,47	0,40	0,45	0,58	0,68	1,08	1,24	0,69	1,92	0,71	1,01	0,51	0,52	1,34	1,47
std	0,70	0,72	0,89	0,09	0,16	0,44	0,09	0,51	1,93	0,68	0,59	0,11	0,11	0,20	0,03
min	0,00	0,00	0,00	0,43	0,22	0,00	1,09	0,16	0,06	0,00	0,00	0,33	0,40	0,52	1,29
25%	0,00	0,00	0,00	0,53	0,57	1,18	1,17	0,32	0,70	0,00	0,72	0,40	0,42	1,25	1,47
50%	0,00	0,00	0,00	0,58	0,71	1,22	1,24	0,54	1,19	0,86	1,09	0,50	0,47	1,44	1,47
75%	1,00	1,00	0,25	0,63	0,78	1,29	1,29	1,00	2,40	1,26	1,31	0,60	0,60	1,47	1,47
max	3,00	3,00	3,00	0,80	0,95	1,38	1,42	2,50	6,95	1,92	2,00	0,73	0,80	1,48	1,47

TABLE 9 : RESULTATS DE PREDICTION DU NOMBRE D'OURAGANS MAJEURS AUX ETATS-UNIS A MOYEN (MT) ET LONG TERME (LT)

On peut également commencer par noter que tous les modèles donnent des prédictions en moyenne plus élevées que la moyenne empirique qu'elle soit calculée sur les vingt dernières années, sur la *baseline* ou sur l'historique complet.

De plus, tous les modèles fournissent une prédiction moyenne à LT supérieure à la prédiction moyenne à MT.

Analysons désormais les résultats plus en détail, modèle par modèle.

La régression Lasso zéro-modifié donne des résultats bien plus élevés que le Lasso classique à moyen et à long terme. A moyen terme, les prédictions semblent aussi plus dispersées, avec notamment au moins une année (valeur minimale) sans ouragan majeur, ce qui semble mieux capturer toute la dispersion possible sur cette variable. Elle avait également une MSE de test plus faible que le Lasso classique. Tout ceci nous conduit à privilégier le choix du Lasso zéro-modifié pour la prédiction du nombre d'ouragans majeurs avec un modèle linéaire.

Pour le GLM, on n'affiche ici aussi que la version Binomiale Négative (avec $\alpha = 0{,}1$), avec et sans *hurdle*.

Le GLM zéro-modifié donne à MT une moyenne proche du GLM classique mais avec un profil assez différent, avec une valeur nulle pour le quantile 25% et une valeur bien plus haute aux quantiles 50% et 75%. A LT en revanche, il prédit une moyenne nettement plus faible, tirée à la baisse à la fois par les prédictions les plus basses qui sont nulles mais aussi et surtout par des valeurs beaucoup plus basses dans la queue de distribution, visible sur la valeur maximale et du quantile 75%. Nous avons analysé ces écarts au regard des deux sources de différences décrites plus haut. Les prédictions plus basses à LT sont essentiellement dues à la calibration différente du GLM au sein du modèle zéro-modifié, qui est entraîné sur les seules observations strictement positives, au nombre de 63, sur 170 observations au total. Comme le montre la comparaison des courbes bleue et jaune dans le graphique ci-dessous, ceci engendre des prédictions très différentes sur les données projetées.

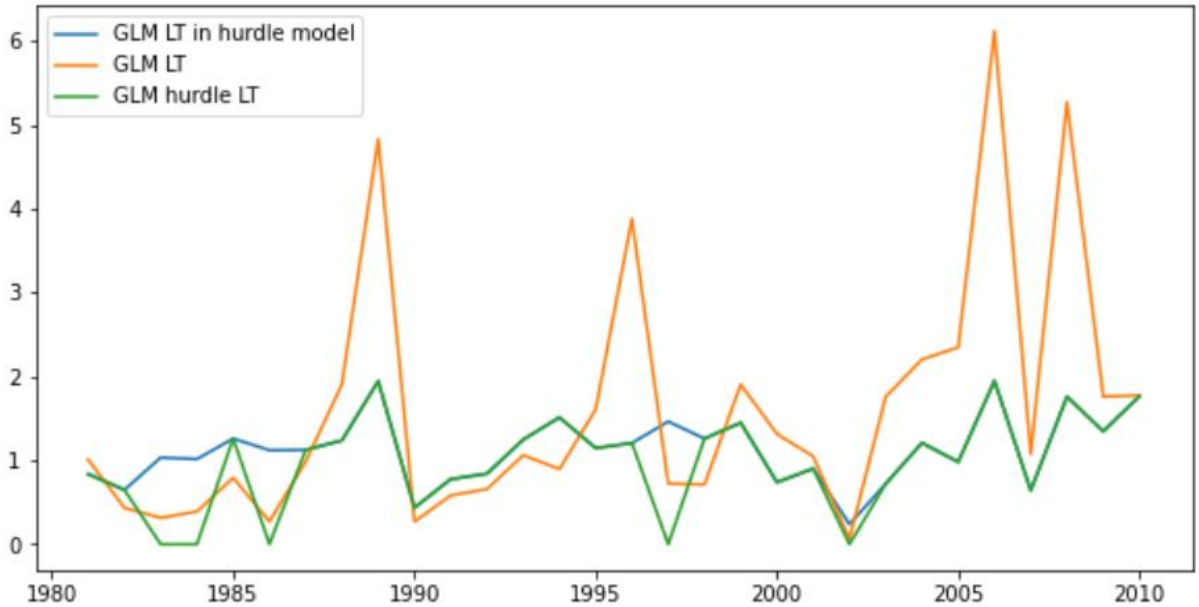

FIGURE 37 : COMPARAISON DES PREDICTIONS A LONG TERME DU NOMBRE D'OURAGANS MAJEURS DES MODELES GLM DANS LES VERSIONS AVEC ET SANS *HURDLE*

L'analyse des coefficients des variables explicatives dans les deux GLM permet de corroborer cette hypothèse et d'expliquer cette différence. Les deux tables ci-dessous décrivent à gauche le GLM classique et à droite le GLM zéro-modifié.

	coef	std err		coef	std err
-77.5&32.5&2	0.1724	0.144	**-77.5&32.5&2**	-0.0370	0.211
-77.5&37.5&12	0.0943	0.121	**-77.5&37.5&12**	0.0582	0.145
-42.5&12.5&5	0.0863	0.161	**-42.5&12.5&5**	0.1063	0.156
-57.5&17.5&8	0.1132	0.185	**-57.5&17.5&8**	0.1074	0.207
-77.5&37.5&2	0.1796	0.131	**-77.5&37.5&2**	0.0003	0.130
-22.5&27.5&5	0.0677	0.165	**-22.5&27.5&5**	0.0847	0.161
-52.5&17.5&8	-0.0795	0.234	**-52.5&17.5&8**	-0.1490	0.247
-22.5&17.5&7	0.3309	0.176	**-22.5&17.5&7**	0.0685	0.156
-62.5&22.5&1	0.3305	0.166	**-62.5&22.5&1**	0.0483	0.161
-77.5&32.5&4	-0.0733	0.155	**-77.5&32.5&4**	-0.0839	0.189
-32.5&27.5&12	-0.3335	0.137	**-32.5&27.5&12**	-0.0453	0.142
-57.5&22.5&12	-0.1952	0.132	**-57.5&22.5&12**	-0.1129	0.128
-67.5&22.5&12	-0.1979	0.104	**-67.5&22.5&12**	-0.0364	0.103
-42.5&37.5&5	-0.2210	0.133	**-42.5&37.5&5**	-0.0373	0.143
-62.5&22.5&12	0.0773	0.151	**-62.5&22.5&12**	0.0090	0.163
const	-1.0415	0.142	**const**	0.1230	0.158

FIGURE 38 : COMPARAISON DES COEFFICIENTS DES MODELES GLM AVEC ET SANS *HURDLE* UTILISES POUR PREDIRE LE NOMBRE D'OURAGANS MAJEURS A LONG TERME

On note en particulier que si la valeur de la constante est plus faible, et même négative pour le GLM classique, la plupart des coefficients ont des valeurs absolues plus faibles dans le GLM zéro-modifié. En particulier, les coefficients les plus élevés, autour de 0,33, ont des valeurs bien plus faibles, entre 0,05 et 0,06, tandis que les coefficients les plus négatifs sont moins fortement négatifs. Le coefficient du premier *feature* de la table a même une valeur positive dans le GLM classique et négative dans la version zéro-modifiée. Cela explique que les pics sont beaucoup moins marqués et notamment la valeur maximale, mais également la volatilité moindre des résultats du GLM zéro-modifié.

La pondération par le *logit* rajoute *in fine* un peu de volatilité multiplication par les prédictions de la régression logistique ne joue finalement que marginalement ici comme on le voit graphiquement entre la courbe bleue, avant pondération, et la courbe verte, après application du *logit*.

Que ce soit dans sa version avec ou sans *hurdle*, le GLM donne des prédictions à MT et à LT en moyenne plus basses que le Lasso zéro-modifié mais avec un écart-type plus élevé, donc une dispersion significativement plus grande qui est bien visible notamment sur la valeur du maximum relativement élevée et significativement plus élevée que celle du Lasso *hurdle*.

Enfin, concernant le kNN et GB, le même constat établi sur les prédictions de nombre d'ouragans reste valable : les prédictions ne semblent pas pertinentes du fait que les modèles travaillent en dehors de leur modèle d'apprentissage. Nous éliminons donc ces modèles pour la suite, dans notre application à l'estimation des pertes assurantielles totales.

Pour le besoin de cette application, par souci de lisibilité, nous décidons de retenir le modèle Lasso pour le nombre d'ouragans et le GLM zéro-modifié pour le nombre d'ouragans majeurs qui nous semblent les modèles les plus adaptés et performants au regard des différents critères considérés.

Le Lasso modélisant le nombre annuel de *landfalls* d'ouragans est défini par la sélection de variables explicatives et leurs coefficients associés suivants.

	coefficient
-62.5&37.5&4	0.000000
-57.5&37.5&4	0.019551
-77.5&32.5&4	0.000000
-62.5&37.5&3	0.083208
-77.5&32.5&2	0.000000
-22.5&17.5&7	0.175360
-77.5&27.5&2	0.150451
-62.5&22.5&1	0.100371
-32.5&12.5&7	0.096939
-17.5&12.5&4	0.054615
-62.5&22.5&12	-0.109264
-72.5&27.5&11	-0.000000
-67.5&22.5&11	-0.257748
-77.5&27.5&10	-0.071247
-72.5&27.5&10	-0.140791
intercept	1.470588

FIGURE 39 : COEFFICIENTS DU MODELE DE REGRESSION LINEAIRE LASSO RETENU POUR LA PREDICTION DU NOMBRE ANNUEL D'OURAGANS

On remarque que la sélection (optimale) utilisée par ce modèle contient 15 variables explicatives mais 4 coefficients sont nuls, ce qui est assez caractéristique du Lasso avec sa contrainte supplémentaire de minimisation de la valeur (en norme L1) des coefficients. En réalité, ce modèle permet donc une prédiction du nombre annuel d'ouragans à partir de 11 variables explicatives, et d'une constante (*intercept*).

Quant au GLM zéro-modifié, modélisant le nombre annuel de *landfalls* d'ouragans majeurs, il est défini par la sélection de variables explicatives et leurs coefficients associés suivants.

	coef	std err	z	P>|z|	[0.025	0.975]
-77.5&32.5&2	-0.0370	0.211	-0.175	0.861	-0.451	0.377
-77.5&37.5&12	0.0582	0.145	0.401	0.689	-0.226	0.343
-42.5&12.5&5	0.1063	0.156	0.681	0.496	-0.200	0.412
-57.5&17.5&8	0.1074	0.207	0.520	0.603	-0.298	0.512
-77.5&37.5&2	0.0003	0.130	0.002	0.998	-0.254	0.255
-22.5&27.5&5	0.0847	0.161	0.526	0.599	-0.231	0.400
-52.5&17.5&8	-0.1490	0.247	-0.604	0.546	-0.633	0.335
-22.5&17.5&7	0.0685	0.156	0.439	0.660	-0.237	0.374
-62.5&22.5&1	0.0483	0.161	0.300	0.764	-0.267	0.364
-77.5&32.5&4	-0.0839	0.189	-0.443	0.658	-0.455	0.287
-32.5&27.5&12	-0.0453	0.142	-0.319	0.750	-0.323	0.233
-57.5&22.5&12	-0.1129	0.128	-0.882	0.378	-0.364	0.138
-67.5&22.5&12	-0.0364	0.103	-0.355	0.723	-0.238	0.165
-42.5&37.5&5	-0.0373	0.143	-0.260	0.795	-0.318	0.244
-62.5&22.5&12	0.0090	0.163	0.055	0.956	-0.310	0.328
const	0.1230	0.158	0.779	0.436	-0.187	0.433

FIGURE 40 : COEFFICIENTS DU MODELE LINEAIRE GENERALISE (GLM) ZERO-MODIFIE RETENU POUR LA PREDICTION DU NOMBRE ANNUEL D'OURAGANS MAJEURS

3. Application actuarielle simplifiée à l'impact sur les pertes assurées du marché

Dans cette dernière partie, nous proposons une application concrète et simple au domaine de l'assurance et de la réassurance. Nous allons calculer les pertes assurantielles annuelles moyennes au niveau marché associées aux prédictions obtenues de nombres annuels moyens d'ouragans et d'ouragans majeurs atteignant les Etats-Unis, à horizon moyen et long terme.

3.1. Approche méthodologique

Pour ce faire, nous avons adopté une approche simplifiée où nous multiplions notre prédiction de nombre annuel d'ouragans, respectivement d'ouragans majeurs, ayant touché terre aux Etats-Unis par le coût moyen d'un tel événement pour le marché provenant du modèle RMS (Risk Management Models, Analytics, Software & Services, s.d.) version 18 (v18).

Nous avons extrait de l'outil RMS le fichier *Event Loss Table (ELT)* associé au région/péril Atlantique Nord/ouragans (*North Atlantic Hurricane* (NAHU)) et à l'exposition marché. Il s'agit d'un catalogue des événements stochastiques générés par RMS incluant pour chaque événement son coût moyen, pour l'exposition marché en question, son écart-type et les informations de *landfall*, à savoir si l'événement a touché terre et le cas échéant le ou les localisations de *landfall* et leurs intensités respectives exprimées en vitesse de vent sur une minute et en catégorie Saffir-Simpson. Nous présentons à titre illustratif un extrait des dix premières lignes de ce fichier ELT.

Event ID	Mean loss (USD)	Std deviation (USD)	Exposure (USD)	Rate	1st Landfall Bypassing	1st Landfall Category	1st Landfall one minute windspeed	1st Landfall State
1	530 671 615	757 407 367	442 709 339 454	10^{-10}	L	1	92	TX
2	2 331 507 631	1 903 268 542	527 220 608 400	10^{-10}	L	1	81	Antilles
3	4 385 053 464	3 021 260 415	432 922 806 129	10^{-10}	L	1	81	Turks and Caicos
4	6 434 825 976	2 970 440 555	1150 734 350 378	10^{-10}	L	1	81	FL
5	3 172 493 479	2 023 074 148	977 440 243 866	10^{-10}	L	2	104	Jamaica
6	945 968 000	1 460 352 001	638 194 599 723	10^{-10}	B	1	91	FL
7	36 448 586	74 548 291	77 224 808 788	10^{-10}	L	1	81	TX
8	13 512 193 214	4 693 363 388	1497 843 521 909	10^{-10}	L	3	115	GA
9	6 648 099 197	6 552 249 760	478 745 054 540	10^{-10}	L	2	104	TX
10	13 988 886 028	4 083 726 011	313 874 413 829	10^{-10}	L	3	127	MS

TABLE 10 : EXTRAIT DE L'ELT MARCHE DU RISQUE D'OURAGANS DANS L'ATLANTIQUE NORD DANS LE MODELE RMS V18

Cette table ELT ne contient que des événements causant des pertes assurées. Ainsi, une tempête qui passerait trop loin des côtes ou aurait une intensité trop faible ne serait pas incluse. La majorité des tempêtes provoquant des pertes assurées sont des tempêtes ayant fait *landfall*, au moins une fois au cours de leur trajectoire. Néanmoins, certains événements peuvent engendrer des dommages en passant très proches des côtes, sans toucher terre. Ces événements qualifiés de *Bypass* sont identifiables par la valeur « B » (par opposition à « L » pour *landfall*) du champ *1st landfall Bypassing*. Il s'agit ici d'un extrait de l'ELT où nous n'affichons que les informations relatives au premier landfall (*1st landfall*) mais l'ELT complet décline ces informations jusqu'au sixième *landfall*. Si une tempête ne touche terre qu'une seule fois au cours de sa trajectoire, seuls les champ *1st landfall* sont renseignés.

En ce qui nous concerne, nous avons filtré cette base sur les tempêtes ayant fait *landfall* au moins une fois aux Etats-Unis et conservé la catégorie du *landfall* le plus puissant en cas de *landfall* multiple. Sur cette sélection, nous avons calculé la perte moyenne, en dollars américains (USD), associée à chaque catégorie d'ouragan, comme présenté dans le tableau suivant.

Catégorie d'ouragans	Nombre d'événements (n)	Perte totale (S)	Perte moyenne (=S/n)
1	3 486	10 959 851 377 119	3 143 961 956
2	2 908	36 561 795 615 552	12 572 832 055
3	4 041	117 823 337 698 890	29 156 975 427
4	5 132	252 303 149 878 671	49 162 733 803
5	2 900	288 398 154 965 338	99 447 639 643

TABLE 11 : COUT MOYEN MARCHE PAR CATEGORIE D'OURAGANS DANS LE MODELE RMS V18

Comme nos prédictions sont au niveau agrégé ouragans totaux et ouragans majeurs, nous avons calculé le coût moyen d'un ouragan, d'un ouragan mineur et d'un ouragan majeur atteignant les Etats-Unis.

Catégorie d'ouragans	Nombre d'événements (n)	Perte totale (S)	Perte moyenne (=S/n)
1-2 (mineurs)	6 394	47 521 646 992 670	7 432 225 054
3-5 (majeurs)	12 073	658 524 642 542 899	54 545 236 689
Total	**18 467**	**706 046 289 535 569**	**38 232 863 461**

TABLE 12 : COUT MOYEN MARCHE D'UN OURAGAN MINEUR ET D'UN OURAGAN MAJEUR DANS LE MODELE RMS V18

3.2. Estimation dans le scénario le plus émetteur SSP5-8.5

Dans le tableau suivant, nous présentons les estimations de perte moyenne globale pour le marché de l'assurance, en milliard de dollars (Md$), en multipliant la prédiction moyenne obtenue dans le modèle Lasso pour le nombre d'ouragans et modèle GLM zéro-modifié pour le nombre d'ouragans majeurs par le coût moyen respectif. La prédiction du nombre d'ouragans mineurs est déduite des deux précédente par différence. Les résultats sont donnés aux horizons de projection moyen (MT) et long terme (LT) ainsi que, pour référence, aux trois mêmes périodes historiques utilisées plus haut pour la présentation des prédictions de nombre annuel.

Grandeur	1851-2020	1981-2010	2001-2020	MT	LT	MT vs 1851-2020	LT vs 1851-2020
Nombre d'ouragans (A)	1,47	1,23	1,25	1,62	1,84	10%	25%
Nombre d'ouragans majeurs (B)	0,47	0,40	0,45	0,71	1,01	51%	115%
Nombre d'ouragans mineurs (A-B)	1,00	0,83	0,80	0,91	0,83	-9%	-17%
Proportion d'ouragans majeurs	32%	32%	36%	44%	55%	37%	72%
Perte due aux ouragans mineurs (Md$)	7,43	6,19	5,95	6,75	6,18	-9%	-17%
Perte due aux ouragans majeurs (Md$)	25,67	21,82	24,55	38,85	55,23	51%	115%

TABLE 13 : ESTIMATION DES PERTES ASSURANTIELLES TOTALES A MOYEN ET LONG TERME DANS LA TRAJECTOIRE LA PLUS DEFAVORABLE SSP5-8.5

Pour rappel, on se place ici dans le cadre de la trajectoire SSP5-8.5 qui est la plus émettrice. Les résultats doivent donc *a priori*, espérons, être considérés comme une borne supérieure du nombre d'ouragans et d'ouragans majeurs attendus, et des pertes associées.

Ainsi, la perte totale pour le marché due aux ouragans majeurs serait 37% plus élevée à horizon 2041-2060 (MT) et 72% plus élevée à horizon 2081-2100 (LT) par rapport à l'historique de 1851 à 2020, dans cette trajectoire SSP la plus défavorable. C'est l'accroissement particulièrement marqué du nombre d'ouragans majeurs qui expliquerait l'essentiel de cette augmentation, faisant augmenter de 51% à MT et 115% à LT la perte marché dues aux ouragans majeurs, tandis que dans le même temps le nombre

d'ouragans mineurs serait en décroissance, -9% à MT et -17% à LT. La proportion d'ouragans majeurs passerait de 32% en moyenne depuis 1851, ou 36% dans les vingt dernières années, à 44% à MT et 55% à LT.

Ces résultats semblent cohérents avec deux constats établis par le (GIEC, Sixième rapport d'évaluation (AR6), 2021) et repris dans le résumé des travaux de recherche sur le lien entre changement climatique et ouragans élaboré par le GFDL (Knuston, 2021). D'une part, l'intensité des cyclones tropicaux est amenée à augmenter, avec un niveau de confiance moyen à élevé, de 1 à 10% selon les projections de modèle, dans un scénario à +2°C. D'autre part, la proportion d'ouragans les plus intenses, de catégorie 4 et 5, devrait également augmenter avec un niveau de confiance moyen à élevé. Toutefois, il est également précisé qu'il est moins évident que le nombre annuel moyen de ces ouragans les plus intenses augmente également étant donné que la plupart des études de modélisation concluent à une réduction, bien qu'en générale légère, du nombre total d'ouragans. C'est pour le coup un point de différence notable avec notre étude où tous nos modèles donnent des prédictions de nombre annuel d'ouragans en moyenne plus élevées que dans le passé, avec toutefois une diminution relative du nombre d'ouragans mineurs, le nombre d'ouragans majeurs augmentant bien plus fortement que le nombre total d'ouragans.

Néanmoins, nous devons garder en tête les nombreuses limites de notre étude, en particulier le fait que nous ne tenons compte que des températures de l'eau comme variables explicatives et que nous avons fait l'hypothèse forte d'un coût moyen constant dans le temps au sein d'une catégorie donnée d'ouragan.

3.3. Estimation dans le scénario intermédiaire SSP2-4.5

A titre illustratif et en guise de sensibilité de l'impact d'une augmentation globale de la température de l'eau sur le nombre d'ouragans, nous avons également estimé le nombre annuel moyen d'ouragans et d'ouragans majeurs atteignant les Etats-Unis avec ces mêmes modèles dans la trajectoire SSP2-4.5. Pour rappel, il s'agit d'un scénario plus optimiste projetant un pic d'émission en 2050 puis une réduction marquée et régulière au-delà, jusqu'en 2100 *a minima*. Comme présenté plus haut dans la Table 5, dans ce scénario, l'augmentation globale de température de l'eau projetée par le GIEC, par rapport à la période 1981-2010, est de +1°C à horizon MT et +1,6°C à LT, contre respectivement +1,3°C et +2,9°C dans SSP5-8.5.

Grandeur	1851-2020	1981-2010	2001-2020	MT	LT	MT vs 1851-2020	LT vs 1851-2020
Nombre d'ouragans (A)	1,47	1,23	1,25	1,59	1,68	8%	15%
Nombre d'ouragans majeurs (B)	0,47	0,44	0,40	0,60	0,90	27%	92%
Nombre d'ouragans mineurs (A-B)	1,00	0,79	0,85	0,99	0,78	-1%	-22%
Proportion d'ouragans majeurs	32%	36%	32%	37%	54%	17%	67%
Perte due aux ouragans mineurs (Md$)	7,43	5,86	6,32	7,39	5,81	-1%	-22%
Perte due aux ouragans majeurs (Md$)	25,67	24,24	21,82	32,52	49,19	27%	92%
Perte totale due aux ouragans (Md$)	33,10	30,11	28,14	39,90	55,01	21%	66%

TABLE 14 : ESTIMATION DES PERTES ASSURANTIELLES TOTALES A MOYEN ET LONG TERME DANS LA TRAJECTOIRE INTERMEDIAIRE SSP2-4.5

Dans cette trajectoire moins émettrice, le nombre d'ouragans et d'ouragans majeurs augmentent par rapport à la moyenne historique, mais moins fortement, 8% à MT et 15% à LT pour le nombre d'ouragans. L'augmentation est également plus faible sur les ouragans majeurs qui sont tout de même

projetés à 27% à MT et +92% à LT. Cela se traduit par une augmentation des pertes assurées de 21% à MT et 66% à LT, contre respectivement +51% et +115% dans le SSP5-8.5, toujours tirée à la hausse par les pertes associées aux ouragans majeurs.

Même dans cette trajectoire plus favorable, la dérive à la hausse des nombres d'ouragans et d'ouragans majeurs et de leurs pertes associées reste très marquée, ce qui s'explique par le fait que, bien que le pic d'émission soit atteint en 2050, la température globale continue d'augmenter jusqu'en 2100 comme illustré dans le graphique présenté plus haut Figure 35.

3.4. Incertitude et erreur stochastique

Comme dans toute modélisation, il faut garder en tête qu'il existe une incertitude autour des résultats, du fait des différentes limites des données et des modèles, que nous avons déjà évoquées, mais aussi tout simplement du fait que les résultats sont très souvent basés sur des tirages aléatoires de loi de distribution. C'est le cas ici, en particulier avec le tirage de loi normale qui est effectué pour construire les projections de températures constituant le jeu de test. Les résultats présentés plus haut sont fondés sur un certain tirage, que nous avons fixé pour avoir des résultats reproductibles. On parle souvent de graine aléatoire ou *seed* dans les langages de programmation, dont Python. Nous avons arbitrairement fixé le numéro de la graine aléatoire à 1.

Nous avons réalisé une étude de sensibilité en choisissant arbitrairement une autre graine aléatoire. Le tableau ci-dessous récapitule les résultats, en termes de nombres annuels et de pertes assurées marché, avec une graine aléatoire fixée à 2 (*seed*=2), notée *s2*.

Grandeur	MT (s2)	LT (s2)	MT (s1) vs MT (s2)	LT (s1) vs LT (s2)
Nombre d'ouragans (A)	1,59	1,78	-2%	-3%
Nombre d'ouragans majeurs (B)	0,71	0,84	0%	-17%
Nombre d'ouragans mineurs (A-B)	0,88	0,95	-3%	14%
Proportion d'ouragans majeurs	45%	47%	2%	-15%
Perte due aux ouragans mineurs (Md$)	6,54	7,04	-3%	14%
Perte due aux ouragans majeurs (Md$)	38,67	45,66	0%	-17%
Perte totale due aux ouragans (Md$)	45,21	52,70	-1%	-14%

TABLE 15 : ESTIMATION DE PERTES ASSURANTIELLES OBTENUES DANS LA TRAJECTOIRE SSP5-8.5 AVEC UNE AUTRE GRAINE ALEATOIRE (S2)

Par rapport aux résultats obtenus avec la graine aléatoire fixée à 1 (s1), les résultats avec cette autre graine aléatoire (s2) affichent des écarts entre -1% et 3% à MT et entre -3% et -17% à LT.
On exhibe ici qu'un autre cas particulier, qui n'aurait vocation à être généralisé. En particulier, on ne peut tirer de conclusion trop hâtive sur le fait que la sensibilité des résultats à la graine aléatoire serait plus élevée à LT. En revanche, on peut *a minima* en conclure que le caractère aléatoire des résultats induit une incertitude qui peut être au moins de l'ordre de 17% à LT.

Cet ordre de grandeur permet aussi de relativiser la précision obtenue grâce à nos différents modèles de prédiction. Ainsi, il n'est pas forcément pertinent par exemple de chercher à améliorer la MSE de validation de quelques pourcents en faisant appel à modèle plus perfectionné, ou en ayant recours à une recherche d'hyperparamètres optimaux plus exhaustive, alors que les résultats finaux seront de toute façon à considérer avec une incertitude de l'ordre de 15% à LT par nature même de la construction stochastique du jeu de test. Nous allons tenter de généraliser la quantification de cette incertitude.

Nous pouvons obtenir assez facilement un encadrement de l'incertitude stochastique liée à l'amplitude du choc appliqué aux températures de l'eau, à partir des projections du GIEC présentées en Table 5. En effet, cette table nous fournit, en plus du choc moyen, les valeurs des quantiles 5% et 95% du choc de température, respectivement aux horizons MT et LT. Nous pouvons en déduire les intervalles de confiance à 90% qui sont de la forme $[q_{5\%}\,;\,q_{95\%}]$ où $q_{5\%}$ est la valeur du quantile 5% et $q_{95\%}$ la valeur du quantile 95%. L'intervalle de confiance à 90% correspond à l'intervalle de valeurs dans lequel

la « vraie » valeur se situe avec une probabilité de 90%. Ainsi, les intervalles de confiance à 90% de l'augmentation de températures de l'eau par rapport à la période 1981-2010 sont [0,9 ; 1,9] à MT et [1,9 ; 4,2] à LT. Il nous suffirait alors de pouvoir associer à ces variations d'augmentation de température de l'eau, une estimation de l'impact sur les différentes grandeurs étudiées, nombres annuels et pertes assurées. C'est pourquoi, nous allons calculer, dans la partie suivante, la sensibilité des ces différentes grandeurs à une variation de température globale de l'eau de +1°C.

Il importe de préciser qu'on n'étudie ici, par simplification, que l'impact de l'erreur stochastique liée au tirage aléatoire de la loi normale servant aux projections de températures à MT et LT dans le cadre de la construction du jeu de test. En réalité, plus globalement dans notre approche de modélisation, il existe d'autres sources d'erreur stochastique en particulier dans la séparation des données en jeu d'apprentissage et de test (fonction *train_test_split*) l'instanciation et la calibration des modèles de régression *Gradient Boosting* qui dépendent d'un état aléatoire (*random_state*). Là aussi, afin d'avoir des résultats stables et reproductibles, nous avons pris soin de fixer ce paramètre, arbitrairement à la valeur 1. La sensibilité au *random_state* n'est pas facilement généralisable, contrairement au cas des tirages aléatoires de loi normale pour lequel nous disposons aisément des intervalles de confiance.

3.5. Sensibilité et élasticité

Il y a donc une variabilité des résultats du fait du caractère stochastique de l'augmentation appliquée aux différents *features*. Chacun est choqué avec un tirage différent mais issu de la même loi normale (même moyenne et écart-type). Ainsi, même s'il y a une dispersion du choc au sein des différentes *features*, en moyenne le choc appliqué est égal à la moyenne du choc MT, respectivement LT, donné par le GIEC et présenté en Table 5.

Aussi, nous pouvons définir une variation de nombre d'ouragans, en moyenne, associée à une augmentation de 1°C, en rapportant l'augmentation de nombre observée à MT, respectivement LT, aux chocs moyens de température, tels qu'indiqués par le GIEC.

On peut définir également l'élasticité du nombre total d'ouragans, du nombre d'ouragans majeurs et de la perte assurée totale par rapport à la température de l'eau. L'élasticité X d'une variable par rapport à la température de l'eau (SST) se calcule comme le rapport de la variation de la grandeur concernée sur la variation correspondante de SST, ce qui s'écrit :

$$Elasticité = \frac{\Delta X}{\Delta SST}$$

La variation de la grandeur X, notée ΔX, est calculée comme la variation relative de la moyenne de la grandeur X entre la période de projection MT, respectivement LT, et la moyenne sur la période de référence 1981-2010. A MT, par exemple, cette quantité s'exprime de la façon suivante :

$$\Delta X\ (MT) = \frac{E[X(MT)] - E[X(1981-2010)]}{E[X(1981-2010)]}$$

La variation de température de l'eau, notée ΔSST, est quant à elle calculée comme le rapport entre le choc de température moyen donné par le GIEC à MT, respectivement à LT, et la moyenne des températures observées sur tous les *features* entre 1981 et 2010, qui est égale à 24,66°C. A MT, dans la trajectoire et la graine aléatoire de référence pour cette étude (SSP5-8.5, *seed*=1), cette quantité vaut :

$$\Delta SST\ (MT) = \frac{E[SST(MT)] - E[SST(1981-2010)]}{E[SST(1981-2010)]} = \frac{1{,}3}{24{,}66} = 5\%$$

Nous présentons ci-après les sensibilités et élasticités des différentes grandeurs dans le scénario de tirage aléatoire de référence (*seed*=1). Nous les avons calculées à la fois à MT et LT puisque les résultats sont légèrement différents, mais l'ordre de grandeur demeure identique.

Grandeur	Sensi +1°C (MT)	Sensi +1°C (MT) (%)	Sensi +1°C (LT)	Sensi +1°C (LT) (%)	Elasti. (MT)	Elasti. (LT)	ΔX (MT)	ΔSST (MT)
Nombre d'ouragans (A)	0,3	18%	0,2	11%	6,0	4,2	31%	5%
Nombre d'ouragans majeurs (B)	0,2	34%	0,2	21%	14,8	13,0	78%	5%
Nombre d'ouragans mineurs (A-B)	0,1	6%	0,0	0%	1,7	0,0	9%	5%
Proportion d'ouragans majeurs	0,1	20%	0,1	14%	6,7	5,9	36%	5%
Perte due aux ouragans mineurs (Md$)	0,4	6%	0,0	0%	1,7	0,0	9%	5%
Perte due aux ouragans majeurs (Md$)	13,1	34%	11,5	21%	14,8	13,0	78%	5%
Perte totale due aux ouragans (Md$)	13,5	30%	11,5	19%	11,9	10,1	63%	5%

TABLE 16 : SENSIBILITE ET ELASTICITE DU NOMBRE D'OURAGANS ET DES PERTES ASSUREES EN FONCTION DE LA TEMPERATURE GLOBALE DE L'EAU

Ainsi, d'après notre modélisation, le nombre annuel moyen d'ouragans augmente de 18% par degré Celsius à la surface de l'eau, tiré par le nombre d'ouragans majeurs qui augmente de 36%. Les pertes assurées quant à elles augmenteraient en moyenne de 30%, soit 13 milliards de dollars, par degré Celsius.

Il faut toutefois bien noter que nous avons une certaine variabilité autour des valeurs de sensibilité et élasticité, à la fois selon la graine aléatoire choisie et selon l'horizon, MT ou LT, que l'on considère.

3.6. Résultats et intervalles de confiance

Désormais, nous disposons des intervalles de confiance à 90% des chocs de température (globale) de l'eau à MT et LT et sommes en outre capables d'associer à une variation de température les impacts estimés sur nos grandeurs d'intérêts, à savoir les nombres annuels de *landfalls* d'ouragans et d'ouragans majeurs et les pertes assurées correspondantes. Nous pouvons donc calculer des intervalles de confiance (IC) à 90% pour ces différentes grandeurs. La borne inférieure d'un tel intervalle correspond à la valeur obtenue pour une variation de température de l'eau égale à la valeur du quantile 5% indiquée par le GIEC, et la borne supérieure à la valeur du quantile 95%. Ces intervalles de confiance à 90% permettent donc *in fine* de refléter dans nos résultats l'incertitude autour des prévisions du GIEC, que nous avons présentées plus haut en Table 5.

Pour référence, nous rappelons aussi les valeurs obtenues dans le scénario central considéré, qui correspond à la graine aléatoire 1 (*seed*=1).

Enfin, nous affichons également les variations relatives par rapport à la moyenne sur l'historique complet, de 1851 à 2020, ce qui nous permet *in fine* d'avoir un encadrement des impacts relatifs sur ces grandeurs et non plus des valeurs uniques fondées sur un scénario précis de choc moyen et de graine aléatoire comme affiché plus haut en Table 13.

Grandeur	MT	LT	MT (IC)	LT (IC)	MT vs 1851-2020	LT vs 1851-2020
Nombre d'ouragans (A)	1,62	1,84	[1,5;1,8]	[1,63;2,12]	[0,02;0,22]	[0,11;0,44]
Nombre d'ouragans majeurs (B)	0,71	1,01	[0,62;0,86]	[0,8;1,29]	[0,31;0,82]	[0,7;1,74]
Nombre d'ouragans mineurs (A-B)	0,91	0,83	[0,89;0,94]	[0,83;0,83]	[-0,11;-0,06]	[-0,17;-0,17]
Proportion d'ouragans majeurs	0,44	0,55	[0,4;0,49]	[0,47;0,65]	[0,26;0,54]	[0,47;1,03]
Perte due aux ouragans mineurs (Md$)	6,75	6,18	[6,58;7,01]	[6,18;6,17]	[-0,11;-0,06]	[-0,17;-0,17]
Perte due aux ouragans majeurs (Md$)	38,85	55,23	[33,61;46,71]	[43,71;70,21]	[0,31;0,82]	[0,7;1,74]
Perte totale due aux ouragans (Md$)	45,60	61,41	[40,19;53,71]	[49,89;76,38]	[0,21;0,62]	[0,51;1,31]

TABLE 17 : INTERVALLES DE CONFIANCE (IC) A 90% DES RESULTATS DE PREDICTION A MOYEN ET LONG TERME DANS LA TRAJECTOIRE SSP5-8.5

Ainsi, on prévoit avec une probabilité de 90% un nombre d'ouragans 2% à 22% plus élevé à MT que la moyenne historique, et 11% à 44% plus élevé à LT, tiré à la hausse par le nombre d'ouragans majeurs. Ceci se traduirait par des pertes assurées en augmentation de 21% à 62% à MT et de 51% à 131% à LT.

Pour rappel, on quantifie et affiche ici uniquement l'erreur stochastique, ou erreur de modèle, liée aux tirages de loi normale servant à construire les projections de températures. C'est-à-dire nous n'estimons pas les autres sources d'erreurs stochastiques comme l'état du générateur aléatoire (*random_state*) des estimateurs *Gradient Boosting*.

Nous devons garder en tête qu'il existe, de plus, bien évidemment, une variabilité naturelle en fonction du modèle retenu, à la fois le type de modèle, sa paramétrisation et la sélection de *features* utilisée.

Conclusion

Dans cette étude, nous avons proposé une approche de modélisation et de quantification de l'influence du réchauffement climatique sur la survenance des ouragans affectant les Etats-Unis, à partir de méthodes de régressions statistiques classiques et d'algorithmes d'apprentissage supervisé. Nous avons fait le choix de prédire le nombre annuel de *landfalls* d'ouragans, respectivement d'ouragans majeurs, et d'utiliser les températures à la surface de l'eau (SST) dans l'Atlantique Nord entre 1850 et 2020, issues de la base *HadSST4* du *MetOffice*, comme seules variables explicatives. Nous avons considéré, pour notre cible, l'inventaire d'ouragans de 1851 à 2020 de la base *Atlantic HURDAT2* du *NHC*.

Tout d'abord, après avoir procédé à un prétraitement et une normalisation (centrage et réduction) des données explicatives, nous avons testé des modèles variés pour tenter de capturer au mieux cette relation entre températures de l'eau et survenance d'ouragans. Nous avons testé des modèles linéaires, puis des modèles linéaires généralisés (GLM) et enfin des algorithmes d'apprentissage supervisé comme les *k plus proches voisins* (kNN) et les forêts aléatoires avec *Gradient Boosting* (GB).

Etant confrontés à un problème de « grande dimension » avec un nombre relativement grand de variables explicatives, ou *features*, par rapport au nombre d'observations (170 années), nous avons procédé à une réduction de dimension, en définissant des sélections de variables explicatives fondées sur les corrélations linéaires (Pearson) et de rang (Spearman), en constituant différentes sélections de *features*. En pratique, pour chacun des modèles, nous avons testé différents jeux d'hyperparamètres et différentes sélections de *features* et conservé la paramétrisation et le jeu de *features* qui donnaient le meilleur résultat de prédiction i.e. l'erreur quadratique moyenne (MSE) de validation croisée la plus faible, tout en limitant le surapprentissage.

Pour les modèles linéaires et linéaires généralisés, nous avons également testé des variantes avec un *hurdle* à zéro. Ces modèles dits zéro-modifiés consistent à modéliser séparément, par une régression logistique, la probabilité d'avoir au moins un événement dans l'année, ce qui permet de mieux capturer les passages par zéro qui sont relativement fréquents lorsque l'on s'intéresse aux *landfalls*, en particulier aux *landfalls* d'ouragans majeurs dont la fréquence annuelle est de 0,47 sur l'historique.

Nous avons comparé le pouvoir prédictif de ces modèles, sur le critère retenu de la MSE. La régression linéaire Lasso, avec et sans *hurdle*, le GLM, avec et sans *hurdle*, et le *Gradient Boosting* sont ceux qui présentent la meilleure capacité de prédiction, i.e. les MSE de validation les plus faibles.

Puis, nous avons entraîné ces modèles présélectionnés sur la totalité des données historiques disponibles et avons analysé leurs prédictions sur deux jeux de données de test projetés, à moyen terme (2041-2060) et à long terme (2081-2100), obtenus à partir des prévisions de températures du (GIEC, Sixième rapport d'évaluation (AR6), 2021). Nous avons choisi comme trajectoire de référence la SSP5-8.5 qui est la plus émettrice, donc la plus conservatrice. Toutefois, nous avons réalisé une sensibilité en considérant la trajectoire intermédiaire SSP2-4.5 et, par ailleurs, calculé l'intervalle de confiance à 90% des prédictions de nombre annuel lié à l'incertitude existant autour de la valeur centrale des chocs de températures à moyen et long terme indiquée par le GIEC et utilisée pour la formation de nos deux jeux de test projetés.

Seuls les résultats des modèles linéaires et GLM ont été jugés acceptables, les modèles de *machine learning* rencontrant des difficultés à prédire en dehors de leur domaine d'apprentissage, sur des températures plus élevées que celles constatées dans l'historique. Le Lasso a été retenu pour prédire le nombre annuel de *landfalls* ouragans et le GLM zéro-modifié, pour le nombre de *landfalls* d'ouragans majeurs. Ces modèles présentent également l'avantage d'être transparents, facilement interprétables et d'avoir une expression analytique, contrairement aux modèles de *machine learning*.

Enfin, nous avons réalisé une application simplifiée au domaine de l'assurance, en calculant à partir des prédictions de nombre de *landfalls* d'ouragans et d'ouragans majeurs, une estimation des pertes totales assurées aux deux horizons de projection, moyen et long terme. Pour ce faire, nous avons multiplié les prédictions de nombres annuels moyens par le coût moyen respectif d'un tel événement pour le marché issu du logiciel de modélisation de catastrophes naturelles RMS version 18.

Dans la trajectoire SSP5-8.5, qui est la plus émettrice, le modèle retenu prédit à horizon moyen terme un nombre annuel moyen d'ouragans en légère augmentation, environ 10%, mais un nombre d'ouragans majeurs en augmentation de 51% en moyenne. Ainsi le nombre total d'ouragans projeté serait en baisse de 9% tandis que la proportion d'ouragans majeurs passerait de 32% à 44% (+37%) en moyenne. A long terme, le même modèle prédit une augmentation de 25% du nombre total d'ouragans tiré par une hausse de 115% juste sur les ouragans majeurs.

Ces ouragans étant le plus destructeurs, ils alourdiraient les pertes totales assurées de 38% à moyen terme et de 86% à long terme, toutes choses égales par ailleurs.

Il est important d'avoir en tête les limites de notre étude, certaines étant inhérentes aux données, d'autres à l'approche de modélisation et aux différentes hypothèses choisies.

Tout d'abord, notre modélisation est fondée sur un nombre relativement limité d'observations, 170 ans d'historique, ce qui est particulièrement peu pour des algorithmes d'apprentissage supervisé.

De plus, elle n'est fondée que sur les températures mensuelles à la surface de l'eau, dans l'Atlantique Nord. Si la température de l'eau est bien la principale grandeur physique d'influence de l'activité cyclonique, ce n'est pas la seule et elle ne suffit pas pour prédire précisément le nombre d'ouragans année par année, comme l'attestent nos résultats de MSE de validation. Le cisaillement vertical des vents notamment constitue une autre grandeur physique d'influence dans la formation des ouragans, mais elle n'est pas corrélée au réchauffement climatique et n'est pas disponible sur un historique aussi complet. Ainsi, le choix des températures de l'eau nous semble pertinent et acceptable pour l'objectif de prédiction recherché ici, à savoir décrire la tendance (moyenne) à moyen et long terme de la fréquence annuelle de *landfalls* d'ouragans, et par la suite des pertes assurées.

En outre, pour constituer notre base d'apprentissage de nombre annuel d'ouragans, nous avons éliminé les tempêtes tropicales n'atteignant pas la catégorie 1 sur l'échelle Saffir-Simpson, ce qui pourrait potentiellement mener à une sous-estimation du nombre futur d'ouragans puisque ces tempêtes pourraient, dans un monde plus chaud, voir leur intensité augmenter et passer le stade d'ouragan.

Au sein du traitement des données, préparatoire à la modélisation, nous avons fait des hypothèses sur le seuil de qualité de données (complétude) minimale pour conserver un *feature*, ainsi que sur l'imputation des données manquantes.

Sur la modélisation elle-même, nous avons choisi pour chaque famille de modèles, le modèle avec le jeu d'hyperparamètres et la sélection de *features* ayant le meilleur pouvoir prédictif, mesuré sur la métrique MSE.

Enfin, la prédiction finale est également dépendante des jeux de test utilisés que nous avons construits à partir des projections d'augmentation moyenne de température du GIEC, par rapport à la période 1981-2010, que nous avons appliquées – de façon stochastique au travers de tirages de loi normale - aux températures observées sur cette période de référence. L'incertitude autour de ces prévisions d'augmentation a toutefois été prise en compte et quantifiée, sous l'hypothèse que la distribution d'augmentation suit une loi normale de moyenne et écart-type définis dans (GIEC, Sixième rapport d'évaluation (AR6), 2021). Nous avons été ainsi en mesure de calculer les intervalles de confiance à 90% des prédictions de nombres annuels d'ouragans, respectivement d'ouragans majeurs, et de leurs pertes associées, aux deux horizons de projection. En l'absence d'information détaillée sur la distribution spatiale et la distribution mensuelle de l'augmentation de températures de l'eau, nous avons utilisé la même loi gaussienne pour la projection de toutes les variables explicatives.

En outre, il existe également des limites propres à notre application, à l'estimation des pertes assurées du marché. Celle-ci ne tient compte que d'un coût moyen, pas de l'ensemble de la distribution de coût par événement.

De plus, on fait l'hypothèse que le coût moyen est constant au cours du temps. Or, il est susceptible de varier en fonction de l'évolution de l'exposition assurée (valeur, répartition géographique, typologie…) mais également du fait de l'intensification, jugée très probable par le GIEC, du risque de submersion marine (*storm surge*) et des précipitations qui accompagnent les ouragans dans un monde plus chaud. Certaines études comme (Kossin, 2014) montrent également que le changement climatique pourrait entraîner une migration vers les pôles (*poleward migration*) des emplacements de formation et/ou d'intensité maximale des ouragans, effet que nous ne captons pas non plus dans notre étude puisque

nous ne nous modélisons que le nombre annuel de *landfalls* aux Etats-Unis et non la trajectoire complète ou la répartition spatiale des ouragans.

Une approche complémentaire intéressante, que nous n'avons pas explorée ici faute de temps, serait d'analyser les prédictions de nombres annuels faites par l'outil en licence libre *Prophet*, de *Meta*. Ce dernier permet de faire des projections à partir de données historiques et de tendances, concernant la variable à prédire et/ou les variables explicatives. Toutefois, cette approche aurait le défaut d'être relativement opaque, ne permettant pas d'identifier clairement le modèle utilisé, les hypothèses sous-jacentes, sa paramétrisation et le détail des coefficients le cas échéant.

Malgré ces incertitudes et limites de l'étude, il n'en demeure pas moins que le réchauffement climatique est avéré avec des conséquences déjà visibles et pour certaines « irréversibles pour des siècles et des millénaires » d'après le GIEC. Il est par exemple établi avec une probabilité très forte qu'il augmente la fréquence et l'intensité des épisodes de sécheresse et d'inondations. Pour la première fois, dans son sixième rapport, le GIEC décrit aussi une augmentation de l'activité cyclonique en lien avec le réchauffement climatique, que nous avons pu mettre en évidence et quantifier dans ce mémoire du moins pour le bassin Atlantique Nord. Une étude similaire sur un autre bassin, par exemple le Pacifique Est concernant les pays asiatiques ou le Pacifique Ouest concernant la côte ouest des Etats-Unis et du Mexique aboutirait probablement à la même conclusion.

Il est donc capital de continuer et de renforcer la lutte contre le réchauffement climatique. Dans cette lutte, chaque degré compte, comme le rappelle la communauté scientifique et comme on le voit encore dans notre étude sur la sensibilité des différentes grandeurs. D'après le (GIEC, Sixième rapport d'évaluation (AR6), 2021), par exemple un épisode de températures extrêmes de période de retour 10 ans a en moyenne 2,8 fois plus de chance de se produire dans le monde actuel à +1°C par rapport à la période 1850-1900 et 5,6 fois plus de chance dans un monde à +2°C. Pour un événement de période de retour 50 ans, ces probabilités sont multipliées en moyenne par un facteur 4,8 à +1°C et 13,9 à +2°C.

Parallèlement, nous devons mieux anticiper, comprendre et nous préparer à ces effets, en mettant notamment en place des stratégies d'adaptation afin d'être moins vulnérables, plus résilients, face au changement climatique. Les assureurs et réassureurs ont un rôle à jouer, tout d'abord en termes de prévention, en soutenant la recherche et la communauté scientifique et en sensibilisant les assurés et les différents acteurs économiques. De plus, en étant capables de mieux anticiper et évaluer l'augmentation des catastrophes naturelles, ils peuvent – et doivent – le refléter dans leur modélisation. En intégrant le réchauffement climatique dans leurs modèles de de tarification, envoyant ainsi un signal-prix cohérent aux assurés et les incitant par exemple à ne pas rebâtir dans les zones les vulnérables ou à mettre en place des mesures de prévention ou de protection. Dans le même temps, en le capturant aussi dans leur modèle interne, ou du moins dans leur calcul de capital requis, ils rehaussent leur exigence de fonds propres, ce qui leur garantit une solvabilité accrue et une plus grande résilience face au changement climatique, protégeant ainsi les assurés, et les acteurs économiques en général.

Pour finir sur une note d'optimisme, car nous en avons aussi besoin sur ce sujet brûlant : il est très peu probable qu'il existe une rétroaction négative d'après (Stanley B. Goldenberg, 2001). Autrement dit, une activité cyclonique accrue n'augmente pas en retour le réchauffement climatique contrairement à certains phénomènes naturels présentant des rétroactions, comme la fonte du pergélisol, également appelé *permafrost*, qui est causée par le réchauffement climatique et dégage des gaz à effet de serre accélérant à son tour le réchauffement climatique.

Bibliographie

Asthana, T., Krim, H., Sun, X., Roheda, S., & Xie, L. (2021). Atlantic Hurricane Activity Prediction: A Machine Learning Approach. *Atmosphere 2021, 12, 455*.

Best Track Data (HURDAT2). (2021). Récupéré sur NHC-NOAA: https://www.nhc.noaa.gov/data/

Emanuel, K. (2021). Atlantic tropical cyclones downscaled from climate reanalyses show increasing activity over past 150 years. *Nature*, Nat Commun 12, 7027. Récupéré sur https://www.nature.com/articles/s41467-021-27364-8.

EPA. (2021). *Climate Change Indicators: Tropical Cyclone Activity*. Récupéré sur https://www.epa.gov/climate-indicators/climate-change-indicators-tropical-cyclone-activity#ref5

Evan, A. T. (2012). Atlantic hurricane activity following two major volcanic eruptions. *Journal of geophysical research: Atmospheres*.

geopy. (s.d.). Récupéré sur Pypi.org: https://pypi.org/project/geopy/

GIEC. (2021). Récupéré sur IPCC Interactive Atlas: https://interactive-atlas.ipcc.ch/regional-information

GIEC. (2021). *Sixième rapport d'évaluation (AR6)*. Récupéré sur Premier Groupe de travail (WG1): https://www.ipcc.ch/report/sixth-assessment-report-working-group-i/

HadSST4. (2021). Récupéré sur MetOffice: https ://www.metoffice.gov.uk/hadobs/hadsst4/data/download.html

HRD. (2019, Novembre). *HURDAT2 format*. Récupéré sur NHC NOAA: https://www.nhc.noaa.gov/data/hurdat/hurdat2-format-nov2019.pdf

HRD. (2021). *HURDAT*. Récupéré sur AOML NOAA: https://www.aoml.noaa.gov/hrd/hurdat/All_U.S._Hurricanes.html

I4CE. (2021). *D'où viennent les cinq nouveaux scénarios du GIEC*. Récupéré sur I4CE: https://www.i4ce.org/dou-viennent-les-cinq-nouveaux-scenarios-du-giec-climat/

J.J. Kennedy, N. R. (2021, Décembre 10). *HadSST.4.0.1.0 Product User Guide*. Récupéré sur MetOffice: https ://www.metoffice.gov.uk/hadobs/hadsst4/HadSST.4.0.1.0_Product_User_Guide.pdf

Kennedy, J. R. (2019). An ensemble data set of sea surface temperature change from 1850 : the Met Office Hadley Centre HadSST.4.0.0.0 data set. *Journal of Geophysical Research: Atmospheres, 124*. Récupéré sur https ://do

Knuston, T. (2021). *Global Warming and Hurricanes, an overview of Research results*. Récupéré sur https ://www.gfdl.noaa.gov/global-warming-and-hurricanes/#statistical-relationships-between-ssts-and-atlantic-hurricanes

Kossin, J. E. (2014). The poleward migration of the location of tropical cyclone maximum intensity. *Nature*, 509, 349–352.

Météo France. (2020). Récupéré sur https://meteofrance.com/comprendre-la-meteo/le-vent/les-cyclones

Méteo France. (2022). Récupéré sur https://meteofrance.com/actualites-et-dossiers/actualites/planete/2021-5e-annee-la-plus-chaude-lechelle-du-globe-selon-copernicus

Mullahy, J. (1986). Specification and Testing of Some Modified Count Data Models. *Journal of Econometrics, 33, 341-365*.

MunichRe. (2018). Récupéré sur https://www.munichre.com/topics-online/en/climate-change-and-natural-disasters/natural-disasters/2017-year-in-figures.html

Pedregosa, F. a. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*. Récupéré sur scikit-learn.org: https://scikit-learn.org

project/global-land-mask. (s.d.). Récupéré sur pypi.org: https://pypi.org/project/global-land-mask/

Risk Management Models, Analytics, Software & Services. (s.d.). Récupéré sur RMS: https://www.rms.com/

Seabold, S. a. (2010). statsmodels: Econometric and statistical modeling with python. Dans *9th Python in Science Conference*. Récupéré sur statsmodels.org: https://www.statsmodels.org

Stanley B. Goldenberg, C. W.-N. (2001). The Recent Increase in Atlantic Hurricane Activity: Causes and Implications. *Science*.

SwissRe. (2021). *World Insurance Series*. Récupéré sur https://www.swissre.com/institute/research/sigma-research/World-insurance-series.html

The_Pandas_Development_team. (2020). *pandas-dev/pandas: Pandas.* Zenodo. Récupéré sur pandas.pydata.org: https://pandas.pydata.org/

UCAR. (2018). *Hurricanes and Climate Change*. Récupéré sur SCIED: https ://scied.ucar.edu/learning-zone/climate-change-impacts/hurricanes-and-climate-change

Van Rossum, G. &. (2009). *Python 3 Reference Manual.* Scotts Valley, CA: CreateSpace. Récupéré sur Python.org: https://www.python.org/

Wikipedia. (2021). *Démographie des Etats Unis*. Récupéré sur Wikipedia: https://fr.wikipedia.org/wiki/D%C3%A9mographie_des_%C3%89tats-Unis

Annexes

Annexe 1 : coefficients des modèles GLM testés pour la modélisation du nombre annuel d'ouragans

GLM avec loi Binomiale Négative et paramètre α=1 :

	coef	std err	z	P>|z|	[0.025	0.975]
-17.5&32.5&6	-0.1182	0.250	-0.473	0.636	-0.608	0.371
-17.5&12.5&3	0.0655	0.111	0.589	0.556	-0.153	0.284
-17.5&37.5&6	-0.0554	0.211	-0.262	0.793	-0.469	0.359
-22.5&37.5&8	0.1005	0.115	0.872	0.383	-0.125	0.326
-22.5&32.5&6	0.1549	0.158	0.980	0.327	-0.155	0.465
-57.5&37.5&4	0.2069	0.105	1.966	0.049	0.001	0.413
-57.5&22.5&8	0.1377	0.157	0.877	0.380	-0.170	0.446
-17.5&27.5&6	0.0369	0.174	0.212	0.832	-0.304	0.377
-32.5&7.5&5	0.0772	0.118	0.655	0.512	-0.154	0.308
-62.5&27.5&8	0.0488	0.146	0.333	0.739	-0.238	0.336
-62.5&27.5&12	-0.0456	0.129	-0.353	0.724	-0.299	0.208
-57.5&27.5&12	-0.0379	0.132	-0.287	0.774	-0.296	0.221
-57.5&32.5&12	0.0196	0.161	0.122	0.903	-0.296	0.335
-52.5&32.5&12	-0.1895	0.169	-1.121	0.262	-0.521	0.142
-67.5&22.5&11	-0.2164	0.088	-2.471	0.013	-0.388	-0.045
const	0.2617	0.099	2.654	0.008	0.068	0.455

GLM avec loi Binomiale Négative et paramètre α=2 :

	coef	std err	z	P>|z|	[0.025	0.975]
-17.5&32.5&6	-0.1509	0.297	-0.507	0.612	-0.734	0.432
-17.5&12.5&3	0.0638	0.132	0.482	0.630	-0.195	0.323
-17.5&37.5&6	-0.0413	0.250	-0.165	0.869	-0.532	0.449
-22.5&37.5&8	0.1021	0.136	0.748	0.454	-0.165	0.369
-22.5&32.5&6	0.1697	0.188	0.902	0.367	-0.199	0.539
-57.5&37.5&4	0.2203	0.126	1.755	0.079	-0.026	0.466
-57.5&22.5&8	0.1428	0.186	0.766	0.444	-0.223	0.508
-17.5&27.5&6	0.0652	0.209	0.313	0.755	-0.344	0.474
-32.5&7.5&5	0.0781	0.140	0.559	0.576	-0.196	0.352
-62.5&27.5&8	0.0495	0.175	0.283	0.777	-0.293	0.392
-62.5&27.5&12	-0.0525	0.156	-0.338	0.736	-0.358	0.252
-57.5&27.5&12	-0.0332	0.158	-0.210	0.834	-0.344	0.277
-57.5&32.5&12	0.0053	0.195	0.027	0.978	-0.376	0.387
-52.5&32.5&12	-0.1821	0.203	-0.896	0.370	-0.581	0.216
-67.5&22.5&11	-0.2704	0.113	-2.390	0.017	-0.492	-0.049
const	0.2550	0.116	2.194	0.028	0.027	0.483

GLM avec loi Binomiale Négative et paramètre α=0,1 :

	coef	std err	z	P>|z|	[0.025	0.975]
-17.5&32.5&6	-0.0813	0.202	-0.403	0.687	-0.477	0.315
-17.5&12.5&3	0.0673	0.091	0.743	0.457	-0.110	0.245
-17.5&37.5&6	-0.0714	0.173	-0.412	0.680	-0.411	0.268
-22.5&37.5&8	0.1013	0.094	1.073	0.283	-0.084	0.286
-22.5&32.5&6	0.1317	0.128	1.031	0.303	-0.119	0.382
-57.5&37.5&4	0.1902	0.085	2.237	0.025	0.024	0.357
-57.5&22.5&8	0.1323	0.128	1.035	0.300	-0.118	0.383
-17.5&27.5&6	0.0077	0.139	0.056	0.956	-0.265	0.280
-32.5&7.5&5	0.0729	0.096	0.757	0.449	-0.116	0.262
-62.5&27.5&8	0.0528	0.118	0.448	0.654	-0.178	0.284
-62.5&27.5&12	-0.0365	0.103	-0.355	0.723	-0.238	0.165
-57.5&27.5&12	-0.0451	0.105	-0.428	0.669	-0.251	0.161
-57.5&32.5&12	0.0272	0.127	0.214	0.830	-0.221	0.275
-52.5&32.5&12	-0.1907	0.134	-1.421	0.155	-0.454	0.072
-67.5&22.5&11	-0.1527	0.061	-2.503	0.012	-0.272	-0.033
const	0.2701	0.082	3.309	0.001	0.110	0.430

GLM avec loi de Poisson :

	coef	std err	z	P>|z|	[0.025	0.975]
-17.5&32.5&6	-0.0707	0.188	-0.377	0.706	-0.438	0.297
-17.5&12.5&3	0.0678	0.084	0.803	0.422	-0.098	0.233
-17.5&37.5&6	-0.0760	0.162	-0.469	0.639	-0.393	0.241
-22.5&37.5&8	0.1029	0.088	1.166	0.244	-0.070	0.276
-22.5&32.5&6	0.1217	0.119	1.026	0.305	-0.111	0.354
-57.5&37.5&4	0.1839	0.079	2.330	0.020	0.029	0.339
-57.5&22.5&8	0.1312	0.119	1.103	0.270	-0.102	0.364
-17.5&27.5&6	0.0007	0.128	0.005	0.996	-0.251	0.253
-32.5&7.5&5	0.0699	0.090	0.777	0.437	-0.106	0.246
-62.5&27.5&8	0.0562	0.109	0.514	0.607	-0.158	0.270
-62.5&27.5&12	-0.0326	0.095	-0.342	0.732	-0.219	0.154
-57.5&27.5&12	-0.0478	0.097	-0.491	0.623	-0.238	0.143
-57.5&32.5&12	0.0256	0.116	0.220	0.826	-0.202	0.253
-52.5&32.5&12	-0.1882	0.124	-1.522	0.128	-0.430	0.054
-67.5&22.5&11	-0.1336	0.053	-2.524	0.012	-0.237	-0.030
const	0.2729	0.077	3.555	0.000	0.122	0.423

GLM zéro-modifié avec loi Binomiale Négative :

	coef	std err	z	P>|z|	[0.025	0.975]
-17.5&32.5&6	-0.0357	0.328	-0.109	0.913	-0.678	0.606
-17.5&12.5&3	0.0135	0.144	0.094	0.925	-0.268	0.295
-17.5&37.5&6	0.0076	0.279	0.027	0.978	-0.540	0.555
-22.5&37.5&8	0.0546	0.157	0.349	0.727	-0.252	0.361
-22.5&32.5&6	0.0645	0.208	0.310	0.757	-0.343	0.472
-57.5&37.5&4	0.0413	0.154	0.268	0.789	-0.260	0.343
-57.5&22.5&8	0.1116	0.198	0.564	0.573	-0.276	0.499
-17.5&27.5&6	-0.0143	0.232	-0.062	0.951	-0.469	0.440
-32.5&7.5&5	0.0476	0.145	0.327	0.743	-0.237	0.333
-62.5&27.5&8	-0.0252	0.183	-0.138	0.891	-0.384	0.334
-62.5&27.5&12	0.0519	0.162	0.320	0.749	-0.266	0.370
-57.5&27.5&12	-0.0027	0.163	-0.016	0.987	-0.322	0.317
-57.5&32.5&12	0.0040	0.205	0.020	0.984	-0.397	0.405
-52.5&32.5&12	-0.1620	0.219	-0.741	0.459	-0.591	0.267
-67.5&22.5&11	-0.1102	0.120	-0.919	0.358	-0.345	0.125
const	0.6842	0.133	5.128	0.000	0.423	0.946

GLM zéro-modifié avec loi de Poisson :

	coef	std err	z	P>|z|	[0.025	0.975]
-17.5&32.5&6	-0.0218	0.186	-0.118	0.906	-0.386	0.342
-17.5&12.5&3	0.0101	0.082	0.123	0.902	-0.151	0.172
-17.5&37.5&6	-0.0070	0.159	-0.044	0.965	-0.319	0.305
-22.5&37.5&8	0.0606	0.090	0.675	0.500	-0.115	0.236
-22.5&32.5&6	0.0620	0.117	0.530	0.596	-0.167	0.291
-57.5&37.5&4	0.0237	0.087	0.271	0.787	-0.148	0.195
-57.5&22.5&8	0.1030	0.113	0.914	0.361	-0.118	0.324
-17.5&27.5&6	-0.0243	0.129	-0.188	0.851	-0.277	0.228
-32.5&7.5&5	0.0445	0.085	0.523	0.601	-0.122	0.211
-62.5&27.5&8	-0.0210	0.101	-0.207	0.836	-0.220	0.178
-62.5&27.5&12	0.0649	0.094	0.693	0.488	-0.119	0.248
-57.5&27.5&12	-0.0085	0.093	-0.091	0.927	-0.191	0.174
-57.5&32.5&12	0.0017	0.117	0.015	0.988	-0.227	0.231
-52.5&32.5&12	-0.1674	0.128	-1.312	0.189	-0.417	0.083
-67.5&22.5&11	-0.0824	0.057	-1.434	0.151	-0.195	0.030
const	0.6930	0.077	8.958	0.000	0.541	0.845