duas
Diplôme Universitaire des **Actuaires de Strasbourg**

UFR Mathématique Informatique Strasbourg

Mémoire présenté pour l'obtention du DUAS et l'admission à l'Institut des Actuaires

le 8 décembre 2023

Par : Kim POUILLY

Titre: Projection du risque tempête à horizon 2100

Confidentialité : ☒ NON ☐ OUI Durée : ☐ 1 an ☐ 2 ans ☐ 3 ans ☐ 4 ans ☐ 5 ans

Membres du jury de l'IA :

Entreprise :
ADDACTIS
Directeur de mémoire (entreprise) :

Membres du jury de l'Unistra :

Nom : Romain NOBIS

J. BERARD
P.-O. GOFFARD

Signature du responsable entreprise

Secrétariat : Mme Stéphanie Richard

Signature du candidat

Résumé

Les prochaines décennies seront marquées par de profonds changements structuraux ayant pour cause, au moins en partie, le changement climatique et ses impacts. Le monde de l'assurance et ses actuaires seront probablement des acteurs majeurs pour comprendre, analyser et apporter des solutions au problème. C'est dans cette perspective que se positionne cette étude qui vise à étudier l'évolution possible au cours de ce siècle d'un péril majeur affectant le continent européen, la tempête. L'étude sera restreinte au cas de la France métropolitaine et la question sera abordée en combinant des méthodes de modélisations actuarielles classiques, comme les modèles linéaires généralisés ou les modèles à inflation de zéros, avec des données météorologiques projetées issues de modèles climatologiques. Ces données seront extraites du projet EURO-CORDEX, un projet en lien avec le Programme Mondial de Recherches sur le Climat (PMRC), qui intègre les scénarios du Groupe d'Experts Intergouvernemental sur l'Évolution du Climat (GIEC) et les Modèles de Circulation Générale (GCM) ainsi que les Modèles Climatique Régionaux (RCM) des principales institutions climatiques européennes et françaises à l'instar de l'Institut Pierre-Simon Laplace (IPSL) ou du Centre National de Recherches Météorologiques (CNRM). Une approche multi-modèles et multi-scénarios a été retenue pour limiter l'incertitude inhérente des scénarios et des modèles.

Mots clefs: Changement Climatique, Modèle de Circulation Générale, GCM, Modèle Climatique Régional, RCM, projection, EURO-CORDEX, Modèle Linéaire Généralisé, GLM, Modèle à Inflation de Zéro, Assurance non-vie, IARD

Abstract

The coming decades will be marked by profound structural changes, caused at least in part by climate change and its impacts. The insurance industry and its actuaries are likely to play a key role in understanding, analyzing and finding solutions to this problem. It is in this perspective that this study is positioned, with the aim of examining the possible evolution over the course of this century of a major peril affecting the European continent : storms. The study will be restricted to metropolitan France, and the question will be addressed by combining classical actuarial modeling methods, such as generalized linear models or zero-inflated models, with projected meteorological data from climatological models. These data will be extracted from the EURO-CORDEX project, which is part of the World Climate Research Programme (WCRP) and integrates the scenarios of the Intergovernmental Panel on Climate Change (IPCC) and the General Circulation Models (GCM), as well as the Regional Climate Models (RCM) of the main European and French climate institutions, such as the Institut Pierre-Simon Laplace (IPSL) and the Centre National de Recherches Météorologiques (CNRM). A multi-model, multi-scenario approach was adopted to limit the inherent uncertainty of scenarios and models.

Keywords: Global Warming, General Circulation Model, GCM, Regional Climate Model, RCM, projection, EURO-CORDEX, Generalized Linear Model, GLM, Zero-Inflated Model, Non-life Insurance, P&C

Note de synthèse

Contexte et problématique

Les dommages causés par les tempêtes sont considérables, premièrement car ce sont des phénomènes qui se produisent sur de vastes zones, deuxièmement car leur intensité est variable et difficilement prévisible. Dans le domaine de l'assurance dommage en France, elles représentent le principal poste de sinistralité. Durant l'année 2022 par exemple, les coûts engendrés par des événements de grêle et de tempêtes se sont élevés à 6,4 milliards d'euros selon France Assureurs. Cette année 2022 a ainsi été estimée comme la pire depuis 1999 marquée par les mémorables tempêtes Lothar et Martin qui avaient généré un coût de 6,86 milliards d'euros.

Cette année marque également la tenue du deuxième exercice de stress-test de l'ACPR. L'objectif est de compléter le cadre méthodologique établi lors de l'exercice pilote de 2020 en encourageant les assureurs à maintenir leurs efforts dans l'intégration du risque climatique au sein de leurs états financiers et de leur gestion interne. Cet exercice vise à évaluer, d'une part, la réaction des assureurs face aux fluctuations de la sinistralité provoquées par le dérèglement climatique et, d'autre part, à anticiper les possibles répercussions sur leur solvabilité.

Il sera également question de tester la capacité des modèles à prendre en compte ces risques climatiques en exploitant les données disponibles et en identifiant les lacunes éventuelles, dans le but de les mesurer de la manière la plus appropriée possible. L'identification et l'extraction de données climatiques revêtent une importance cruciale dans ce contexte, et notre étude contribuera à renforcer les fondations nécessaires pour aborder ces questions climatiques.

Dans la conjoncture actuelle où le dérèglement climatique est de plus en plus tangible, une question intervient de manière instinctive : Allons-nous connaître un accroissement des événements climatiques extrêmes et en particulier une augmentation du nombre de tempêtes ?

Contenu du mémoire

Dans ce mémoire, nous essaierons alors de comprendre comment le risque de tempête serait susceptible d'affecter la France métropolitaine au cours du 21e siècle en nous basant sur les projections climatiques des météorologues en lien avec les scénarios mis en place par le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC).

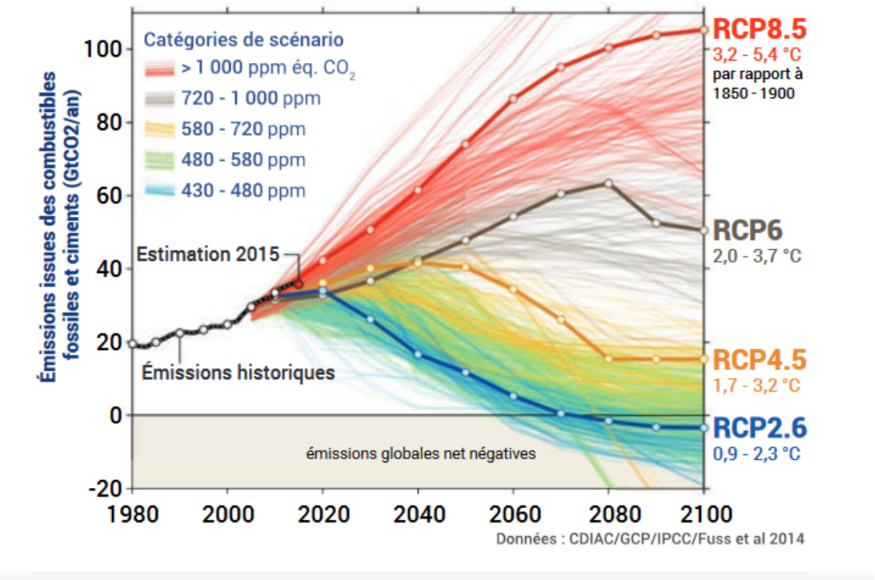

FIGURE 1 – Évolution des émissions entre 1980 et 2100, selon les différents scénarios disponibles. Les quatre scénarios sélectionnés dans le cadre du 5e rapport du Giec (RCP) sont mis en évidence. (Source : Global Carbon Project)

D'un point de vue assurantiel, l'objectif est de comprendre l'**évolution de la fréquence de sinistre à horizon 2100**. Nous étudierons pour cela la sinistralité tempête historique sur un portefeuille multirisque habitation en mesurant l'impact sur la fréquence de sinistre d'une multitude de paramètres météorologiques liés de manière directe ou indirecte à la tempête sans oublier les variables caractéristiques du bâtiment. Après l'identification des paramètres les plus importants, nous combinerons la projection des paramètres météorologiques avec notre modèle de fréquence ajusté pour en déduire une évolution de la fréquence de sinistre tempête au cours des prochaines années. Dans cette étude, les variables caractéristiques du bâtiment sont projetées de manière constante, seules les variables météorologiques fluctuent ce qui a pour effet d'isoler l'impact du changement climatique pour cette évolution de fréquence de sinistre.

Démarche pour obtenir la projection

Les différentes étapes de notre processus de modélisation peuvent être résumé par le schéma suivant :

FIGURE 2 – Méthode effectuée pour nos projections

La première étape a été de comprendre le phénomène tempête d'un point de vue physique afin d'identifier les paramètres météorologiques qui sont à l'origine des phénomènes de vents forts. Une tempête est une perturbation atmosphérique, plus précisément une dépression atmosphérique (une zone de basse pression), pouvant s'étendre sur une largeur de 2000 km où deux masses d'air, ayant des températures et un taux d'humidité différents, se confrontent. Cette confrontation génère des vents pouvant être extrêmes. Cette compréhension du phénomène nous a poussé à extraire des paramètres relatifs au vent, à la température, à la pression atmosphérique ainsi que d'autres paramètres atypiques comme l'énergie potentielle de convection disponible.

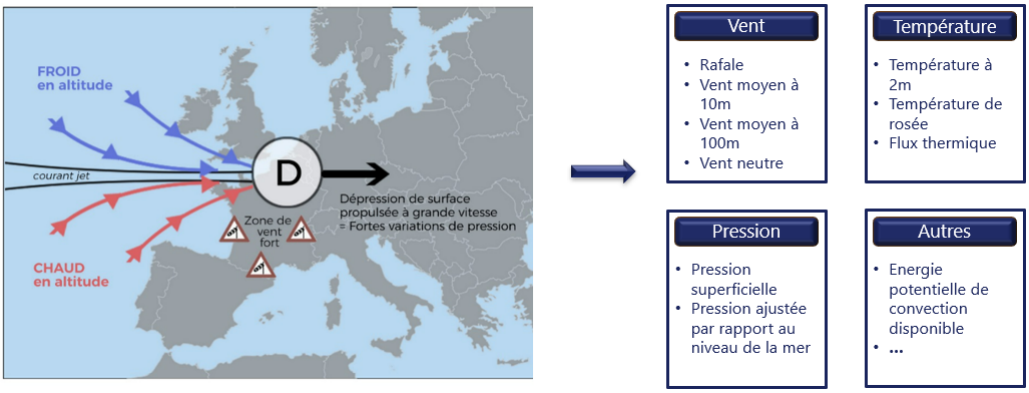

FIGURE 3 – Identification de variables météorologiques en lien avec la tempête

Il était ensuite nécessaire de s'approprier des notions climatiques pour pouvoir appréhender les modèles de circulation générale ainsi que les modèles climatiques régionaux construits par les climatologues. L'idée était de récupérer les sorties de ces modèles, en l'occurrence les variables météorologiques projetées selon différents scénarios, et de les intégrer dans nos modèles actuariels.

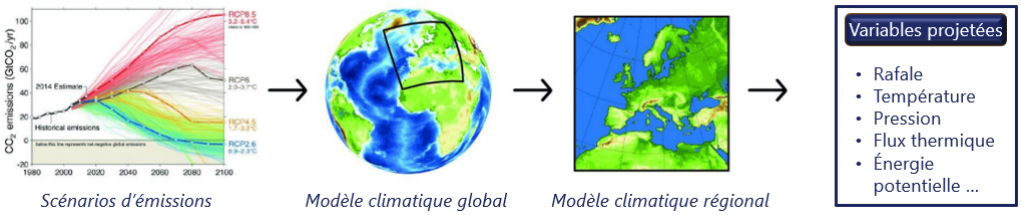

FIGURE 4 – Process d'extraction des variables météorologiques projetées

Un travail de recherche de sources de données a été entrepris afin de comprendre et d'identifier les plus pertinentes dans le contexte de cette étude. Divers critères ont été pris en compte pour évaluer quelles sources étaient les plus cohérentes pour obtenir des données météorologiques historiques et prospectives. Les deux tableaux suivants résument la démarche adoptée :

	Exhaustivité variables météos	Type de données	Temporalité	Résolution géographique	Gratuité
Météo-France	✓	réelles	✓	✓	✗
SYNOP	✓	réelles	✓	✗	✓
COPERNICUS (ERA5)	✓	réanalysées	✓	✓	✓

FIGURE 5 – Tableau récapitulatif des critères utilisés pour le choix de notre source de **données historiques**

	Exhaustivité variables météos	Type de données	Temporalité	Résolution géographique	Gratuité
DRIAS	✗	simulées	✓	✓	✓
NGFS	✗	simulées	✓	✓	✓
EURO-CORDEX	✓	simulées	✓	✓	✓

FIGURE 6 – Tableau récapitulatif des critères utilisés pour le choix de notre source de **données de projection**

La seconde étape a consisté à regrouper les données client (nous disposions ici d'un portefeuille assurantiel anonymisé en lien avec la sinistralité tempête) et les données annexes (météorologiques, densité de population et autres). Dans notre cas,

nous avons agrégé notre base de données par Canton x Mois x Année x Modalités des variables catégorielles caractéristiques du bâtiment les plus pertinentes. In fine, une modélisation d'individus canton avec des caractéristiques relatives au bâtiment est effectuée. Ce procédé d'agrégation nous a permis de concilier la disponibilité des données et la cohérence pour notre objectif de projection.

Il s'en est suivi une troisième étape d'analyse de la base de données à travers la mise en place de statistiques descriptives, d'identification des doublons et incohérences et de sélection de variables pertinentes. A la suite de cette étape, nous disposions d'une base de données prête à la modélisation. L'objectif étant d'analyser la fréquence de sinistre en combinant des modèles actuariels classiques et des données climatologiques, nous avons alors calibrer un modèle linéaire généralisé basé sur une loi binomiale négative. Le schéma suivant résume les différentes étapes qui ont permis d'aboutir au modèle calibré :

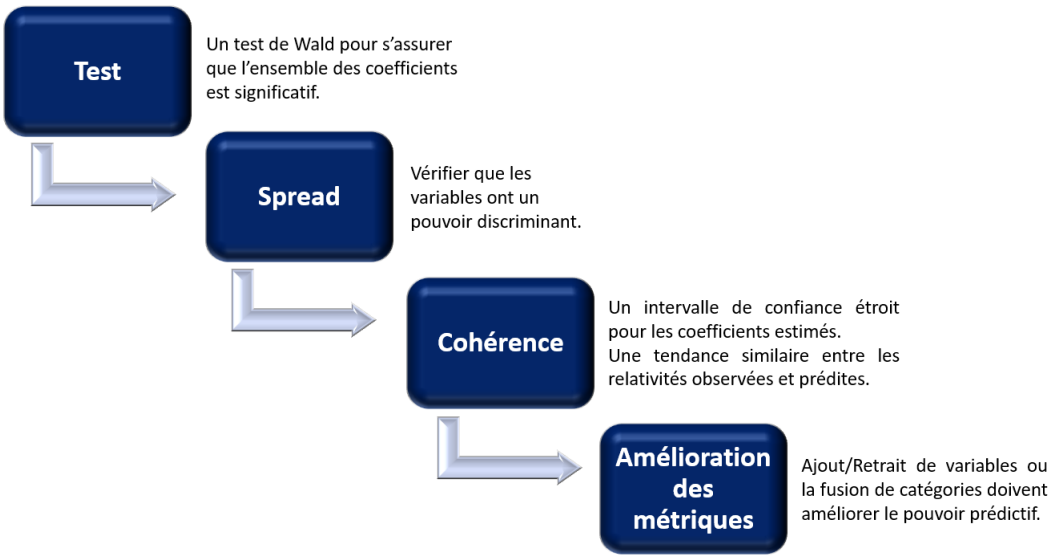

FIGURE 7 – Schéma utilisé pour calibrer les modèles

Puis pour tenter de pallier au déséquilibre des données lié à la prépondérance des lignes non sinistrées, nous avons décidé de nous intéresser aux modèles à inflation de zéros qui ont été spécifiquement conçu pour répondre à ce type de problématique. Il est important de noter que cette forte concentration de 0 est normale au vu du risque étudié, néanmoins dans une logique de régression le modèle pourrait avoir tendance à prédire une non-occurrence de sinistre d'où l'intérêt de mettre en place ce type de modèle pour challenger les modèles précédents.

Une fois le modèle établi, l'étape de projection de la fréquence de sinistre a été

initiée en combinant ce modèle et les données climatologiques en lien avec les scénarios du GIEC.

Pour limiter les incertitudes inhérentes aux modèles climatiques, une **approche multi-modèles et multi-scénarios** a été réalisée. Il était alors nécessaire dans un premier temps de sélectionner un ensemble de modèles et un ensemble de scénarios que nous souhaitions utiliser pour effectuer les projections.

Concernant les scénarios, nous avons décidé de retenir le RCP 4.5 qui a été retenu dans le dernier stress test de l'ACPR et le RCP 8.5 qui est le scénario extrême. Nous avons en revanche exclu les scénarios RCP 6.0, qui n'est plus utilisé dans le dernier rapport du GIEC, et le RCP 2.6 qui est jugé trop optimiste d'après la communauté scientifique.

Concernant les modèles climatiques, nous avons pris l'ensemble des couples GCM/RCM de la plateforme EURO-CORDEX qui calculaient les variables retenues pour le RCP 4.5 et le RCP 8.5. Nous n'avons pas appliqué de sélection en amont puisqu'il n'y pas *a priori* de modèle meilleur qu'un autre d'où l'idée de tous les combiner dans nos projections pour avoir les résultats les plus robustes possibles. Voici les différents couples sélectionnés :

GCM	RCM
CNRM-CERFACS-CNRM-CM5	CLMcom-CCLM4-8-17
CNRM-CERFACS-CNRM-CM5	KNMI-RACMO22E
CNRM-CERFACS-CNRM-CM5	SMHI-RCA4
ICHEC-EC-EARTH	CLMcom-CCLM4-8-17
ICHEC-EC-EARTH	KNMI-RACMO22E
ICHEC-EC-EARTH	SMHI-RCA4
IPSL-CM5A-MR	SMHI-RCA4
MPI-M-MPI-ESM-LR	CLMcom-CCLM4-8-17
MPI-M-MPI-ESM-LR	SMHI-RCA4

TABLE 1 – Listes des couples GCM/RCM sélectionnés pour effectuer les projections

Résultats du modèle

L'ajustement de notre modèle historique a été fait pour reproduire au mieux la sinistralité observée pour chaque mois et chaque région. Notre objectif final étant d'effectuer des projections à partir de ce modèle, nous souhaitions ainsi comprendre l'évolution géographique et temporelle du risque à travers les années. Nous présentons ci-dessous les écarts entre nos observations et nos prédictions :

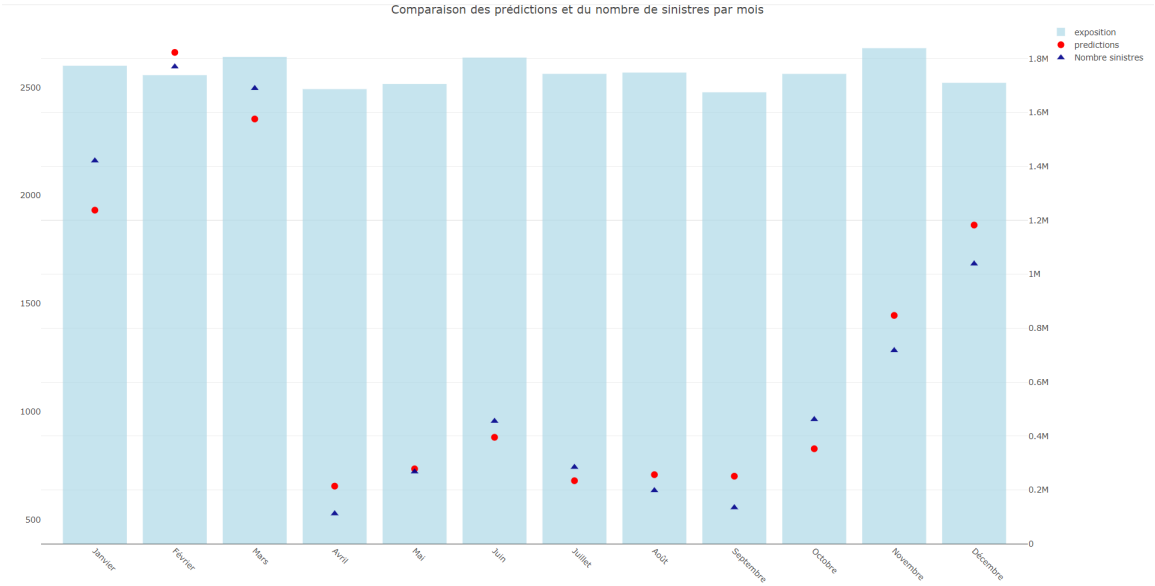

FIGURE 8 – Comparaison des prédictions et des observations à l'échelle mensuelle

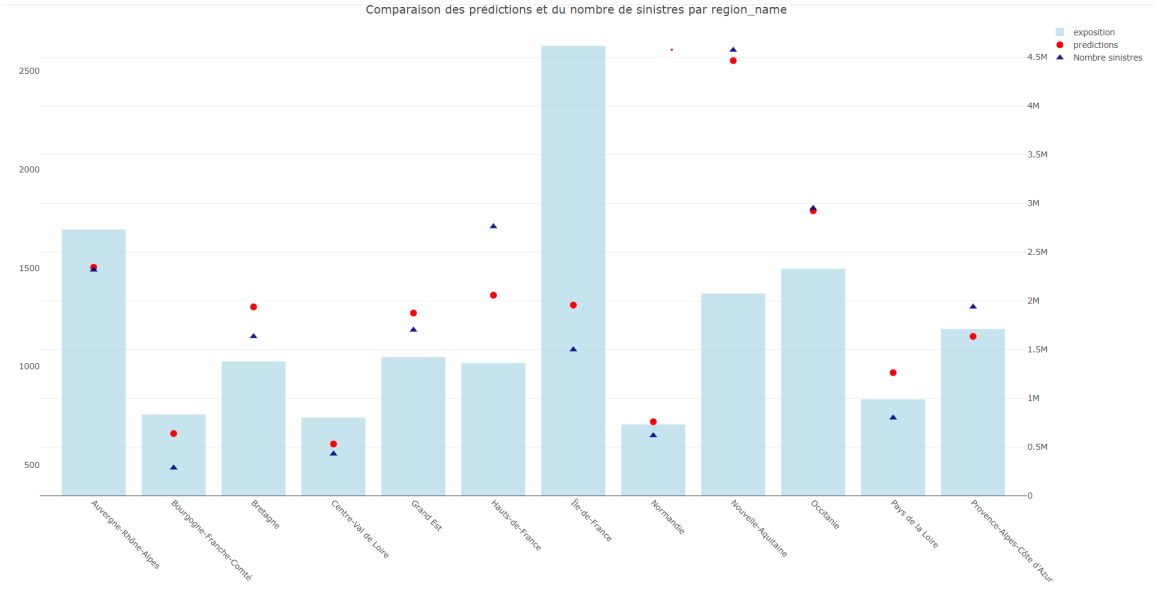

FIGURE 9 – Comparaison des prédictions et des observations à l'échelle régionale

Concernant les projections, nous avons tout d'abord remarqué qu'une différence de stabilité existait entre les scénarios. Dans une logique d'approche multi-scénarios et en comparant le résultat du nombre de sinistres projetés à l'échelle mensuelle et régionale pour le même modèle mais forcé par deux RCP différents, on remarque des effets beaucoup plus instables pour le RCP 8.5 :

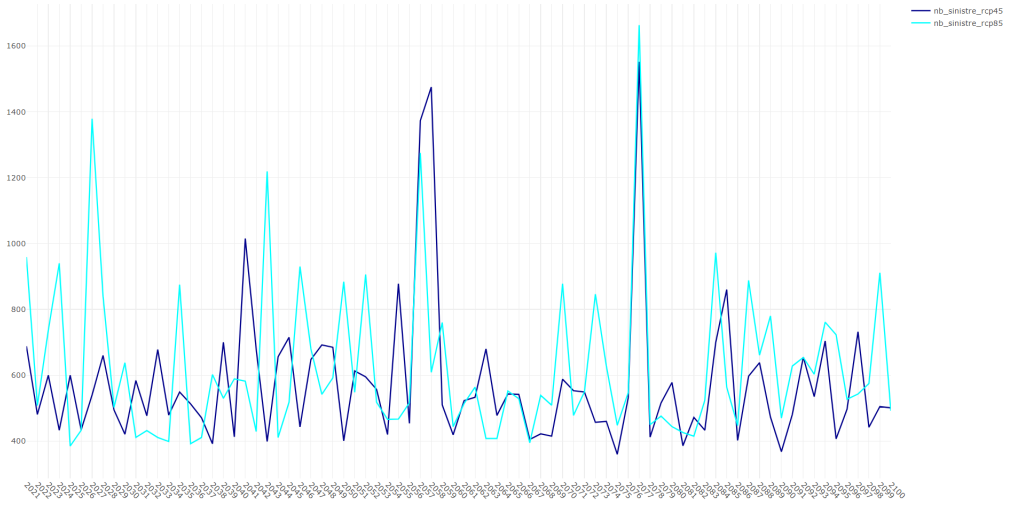

FIGURE 10 – Évolution du nombre de sinistres en Normandie pour un modèle du CNRM

Nous présentons un exemple de graphique qui montre ces différences en terme de nombre de pics pour les deux scénarios. L'exemple sélectionné prend le cas de la Normandie et d'un modèle du CNRM néanmoins nous retrouvons cet effet sur la majorité des modèles, des régions et des mois. Cette instabilité est cohérente et peut être expliqué par les effets des rétroactions qui peuvent être amplifiés dans le cas d'un réchauffement plus important. Nous pouvons ainsi déduire qu'un monde où la concentration des gaz à effet de serre augmente serait de moins en moins prévisible. Par ailleurs, nous voyons ici que le risque tempête n'augmenterait pas nécessairement mais les événements extrêmes pourraient être plus nombreux.

Nous avons ensuite effectué une analyse de l'évolution annuelle des sinistres à l'échelle nationale sous une vision multi-modèles et multi-scénarios. Nous avons tracé ces évolutions en effectuant la moyenne des sorties de chaque modèle encadré par les résultats des sorties maximales et minimales, pour chaque scénario :

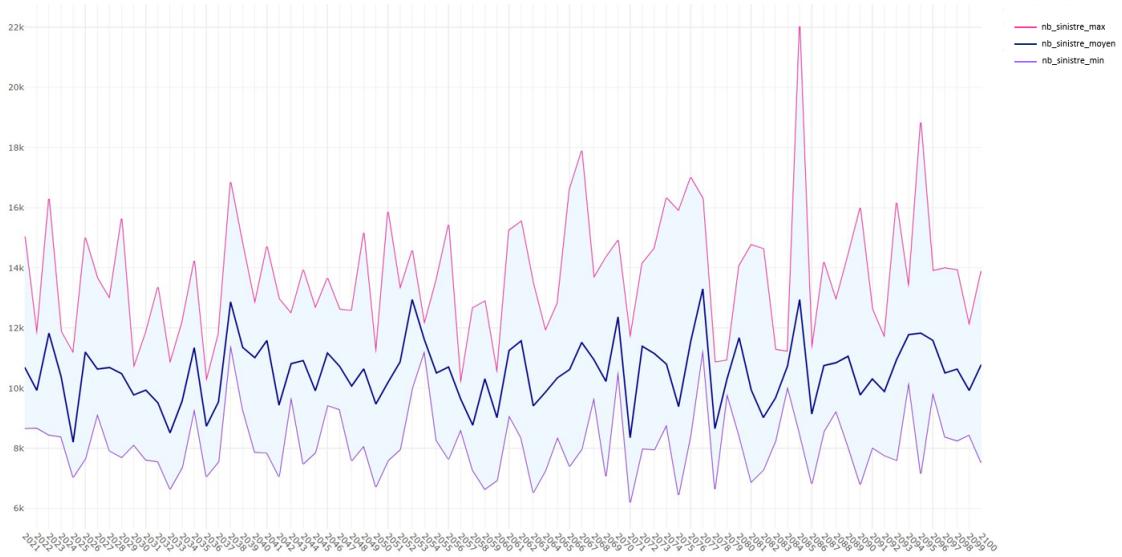

FIGURE 11 – Évolution du nombre de sinistres annuels en combinant l'ensemble des modèles pour le RCP 8.5

FIGURE 12 – Évolution du nombre de sinistres annuels en combinant l'ensemble des modèles pour le RCP 4.5

Nous n'observons pas de tendance particulière sur ces deux graphiques. Cette conclusion serait dans la lignée des études menées par COVÉA qui avait conclu dans leur livre blanc de janvier 2022 que le risque tempête ne subirait pas d'augmentation en termes de fréquence ou de sévérité à horizon 2050. Notre étude vient ainsi corroborer ces résultats.

Synthesis

Context and problem

The damage caused by storms is considerable, firstly because they occur over vast areas, and secondly because their intensity is variable and difficult to predict. In the French property and casualty insurance sector, storms are the main cause of claims. In 2022, for example, the costs generated by hail and storm events amounted to 6.4 billion euros, according to France Assureurs. The year 2022 was thus estimated to be the worst since 1999, marked by the memorable storms Lothar and Martin, which cost 6.86 billion euros.

This is also the year of the ACPR's second stress test exercise. It aims to complete the methodological framework of the 2020 pilot exercise by encouraging insurers to persevere in integrating climate risk into their financial statements and internal management. In other words, this exercise seeks to understand, on the one hand, insurers' behavior in the face of fluctuations in claims caused by climate change (via long-term scenarios) and, on the other hand, the possible repercussions on their solvency (via short-term scenarios). The aim will also be to test the ability of models to incorporate this type of risk, by exploiting available data and identifying missing data, in order to measure it as appropriately as possible. The notion of identifying and extracting climate data is crucial in the context of these climate issues, and our study will help to consolidate these foundations.

Hence, in the current context of increasingly tangible climate disruption, one question instinctively comes to mind : Are we going to see an increase in extreme climatic events and, in particular, an increase in the number of storms ?

Thesis' Content

In this thesis, we will try to understand how the risk of storms is likely to affect France in the 21st century. based on meteorologists' climate projections in relation to the scenarios by the Intergovernmental Panel on Climate Change (IPCC).

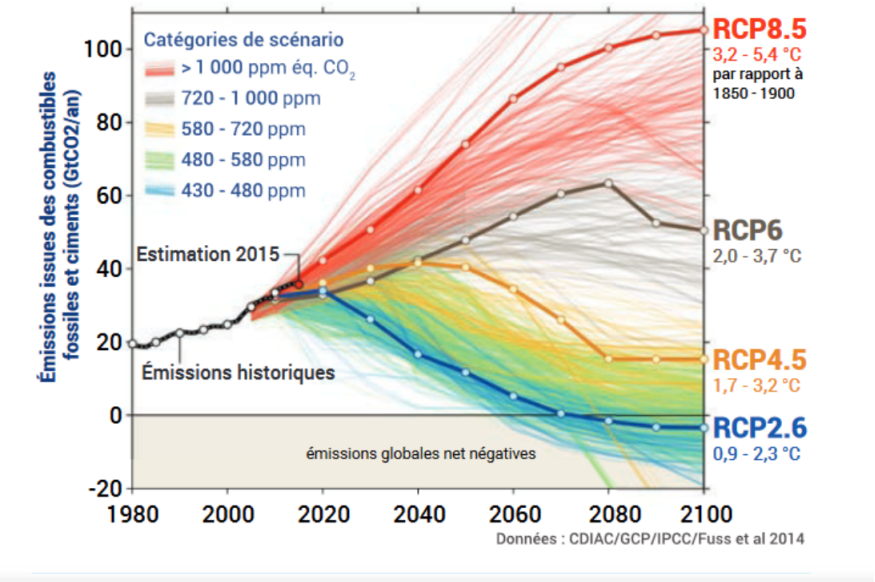

FIGURE 13 – Emissions trends between 1980 and 2100, according to the different scenarios available. The four scenarios selected for the 5th IPCC Report (RCP) are highlighted. (Source : Global Carbon Project)

From an insurance point of view, the aim is to understand the evolution of claims frequency by 2100. To this end, we will study the historical storm claims experience of a multi-risk home portfolio, measuring the impact on claims frequency of a multitude of meteorological parameters directly or indirectly linked to the storm, not forgetting variables characteristic of the building. Once we have identified the most important parameters, we will combine the projected meteorological parameters with our adjusted frequency model to deduce a trend in storm loss frequency over the next few years. In this study, the building's characteristic variables are projected as constant, with only the meteorological variables fluctuating, thus isolating the impact of climate change for this loss frequency trend.

Retained approach to obtain a projection

The various steps of our modeling process can be summarized by the following diagram :

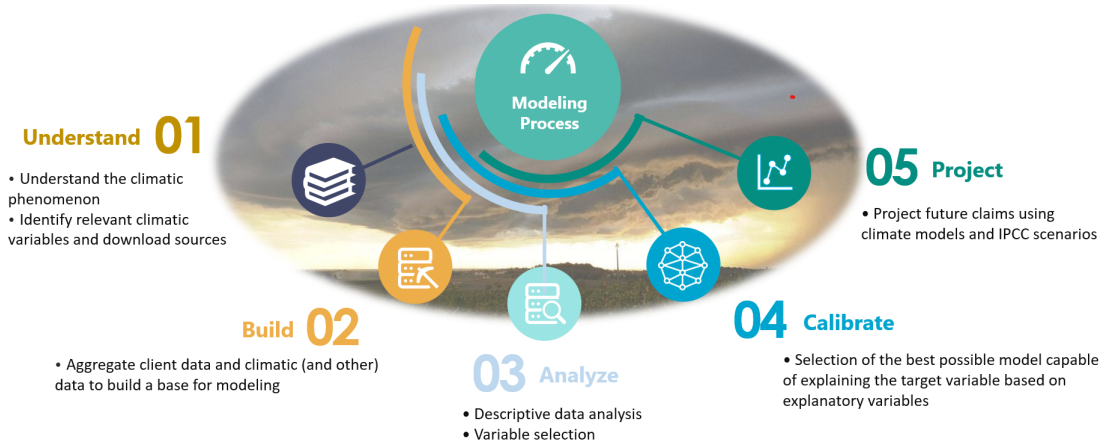

FIGURE 14 – Method performed for the projections

The first step was to understand the storm phenomenon from a physical perspective in order to identify the meteorological parameters that are the origin of strong wind events. A storm is an atmospheric disturbance, specifically an atmospheric depression (a low-pressure area), which can extend over a width of 2000 km where two air masses, with different temperatures and humidity levels, collide. This collision generates winds that can be extreme. This understanding of the phenomenon led us to extract parameters related to wind, temperature, atmospheric pressure, as well as other unique parameters such as convective available potential energy.

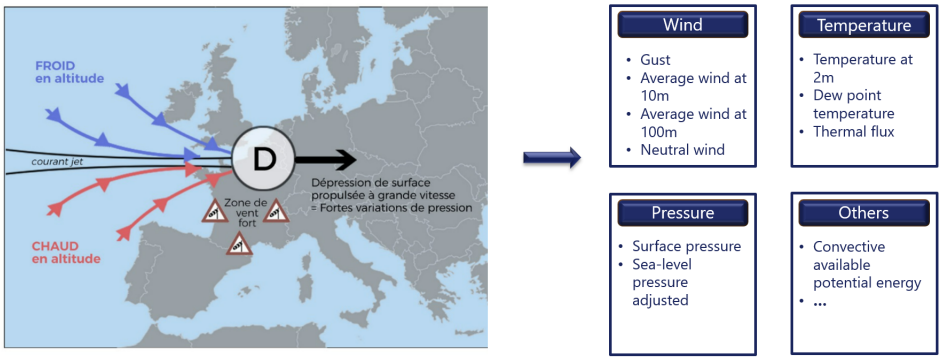

FIGURE 15 – Identification of meteorological variables related to the storm

It was then necessary to familiarize ourselves with climatic concepts in order to understand general circulation models as well as regional climate models constructed by climatologists. The idea was to retrieve the outputs of these models, specifically the projected meteorological variables under different scenarios, and integrate them into our actuarial models.

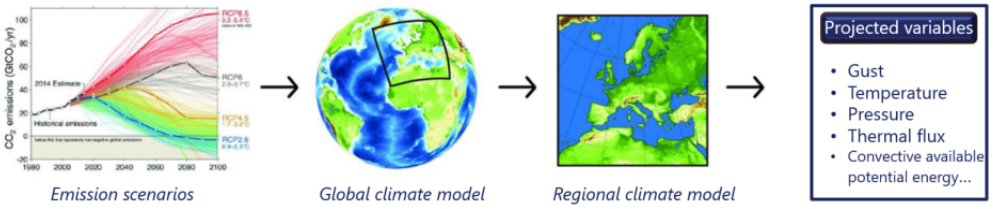

FIGURE 16 – Process of extracting projected meteorological variables

A research effort to identify relevant data sources was undertaken to understand and assess the most pertinent ones in the context of this study. Various criteria were considered to evaluate which sources were the most coherent for obtaining historical and prospective meteorological data. The following two tables summarize the approach adopted :

	Meteorogical variables comprehensiveness	Type of data	Temporality	Geographical resolution	Free of charge
Météo-France	✓	actual	✓	✓	✗
SYNOP	✓	actual	✓	✗	✓
COPERNICUS (ERA5)	✓	reanalyzed	✓	✓	✓

FIGURE 17 – Summary table of the criteria used for the selection of the **historical data source**

	Meteorogical variables comprehensiveness	Type of data	Temporality	Geographical resolution	Free of charge
DRIAS	✗	simulated	✓	✓	✓
NGFS	✗	simulated	✓	✓	✓
EURO-CORDEX	✓	simulated	✓	✓	✓

FIGURE 18 – Summary table of the criteria used for the selection of the **projection data source**

The second step involved consolidating client data (we had an anonymized insurance portfolio related to storm-related claims) and additional data (meteorological, population density, and others). In our case, we aggregated our database by Canton x Month x Year x Modalities of the most relevant categorical variables related to building characteristics. Ultimately, modeling individuals by Canton with building-related characteristics is performed. This aggregation process allowed us to reconcile data availability and consistency for our projection objective.

This was followed by a third step of database analysis through the implementation of descriptive statistics, identification of duplicates and inconsistencies, and selection of relevant variables. Following this step, we had a database ready for modeling. With the goal of analyzing claim frequency by combining traditional actuarial models and climatological data, we then calibrated a generalized linear model based on a negative binomial distribution. The following diagram summarizes the various steps that led to the calibrated model :

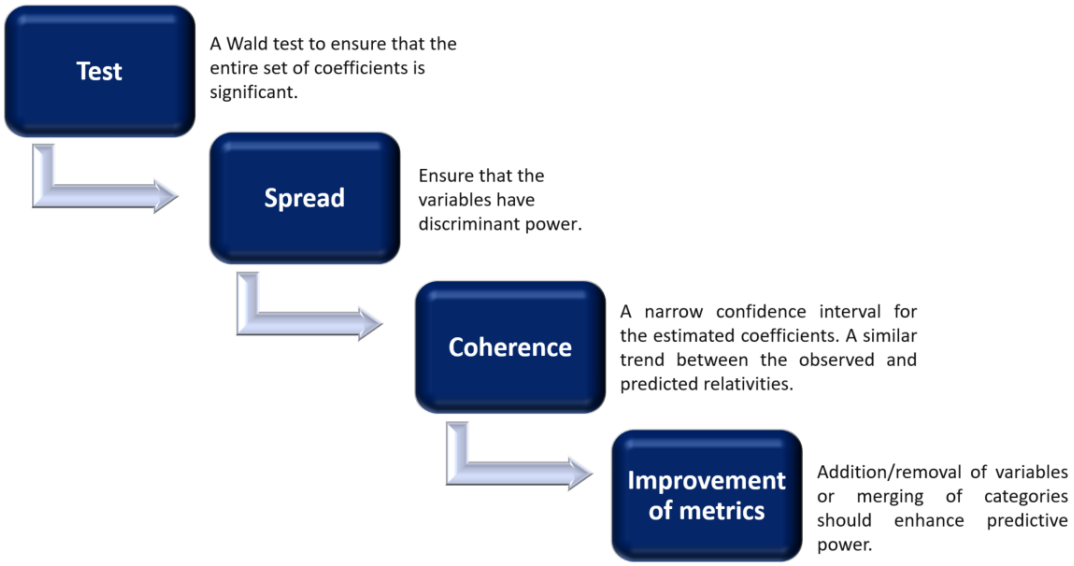

FIGURE 19 – Process used to calibrate the models

Then, in an attempt to address the data imbalance associated with the predominance of non-claims, we decided to explore zero-inflated models specifically designed to tackle this type of issue. It is important to note that this high concentration of zeros is normal given the studied risk ; however, in a regression context, the model might tend to predict a non-occurrence of claims. Hence, the interest in implementing this type of model to challenge the previous models.

Once the model was established, the step of projecting the claim frequency was initiated by combining this model with climatological data related to the scenarios provided by the IPCC.

To mitigate uncertainties inherent in climate models, a multi-model and multi-scenario approach was undertaken. It was necessary, initially, to select a set of models and a set of scenarios that we intended to use for conducting the projections.

Regarding the scenarios, we decided to adopt RCP 4.5, which was used in the latest stress test by the ACPR, and RCP 8.5, which represents the extreme scenario. However, we excluded the RCP 6.0 scenarios, no longer used in the latest IPCC report, and RCP 2.6, considered too optimistic by the scientific community.

Regarding climate models, we took all GCM/RCM pairs from the EURO-CORDEX platform that calculated the selected variables for RCP 4.5 and RCP 8.5. We did not apply upfront selection since there is no *a priori* better model than another, hence the idea of combining all of them in our projections for the most robust results possible. Here are the different pairs selected :

GCM	RCM
CNRM-CERFACS-CNRM-CM5	CLMcom-CCLM4-8-17
CNRM-CERFACS-CNRM-CM5	KNMI-RACMO22E
CNRM-CERFACS-CNRM-CM5	SMHI-RCA4
ICHEC-EC-EARTH	CLMcom-CCLM4-8-17
ICHEC-EC-EARTH	KNMI-RACMO22E
ICHEC-EC-EARTH	SMHI-RCA4
IPSL-CM5A-MR	SMHI-RCA4
MPI-M-MPI-ESM-LR	CLMcom-CCLM4-8-17
MPI-M-MPI-ESM-LR	SMHI-RCA4

Table 2 – List of selected GCM/RCM pairs for conducting the projections

Model results

The adjustment of our historical model was done to best replicate the observed claims for each month and each region. With our ultimate goal being to make projections based on this model, we aimed to understand the geographical and temporal evolution of the risk over the years. Below, we present the discrepancies between our observations and predictions :

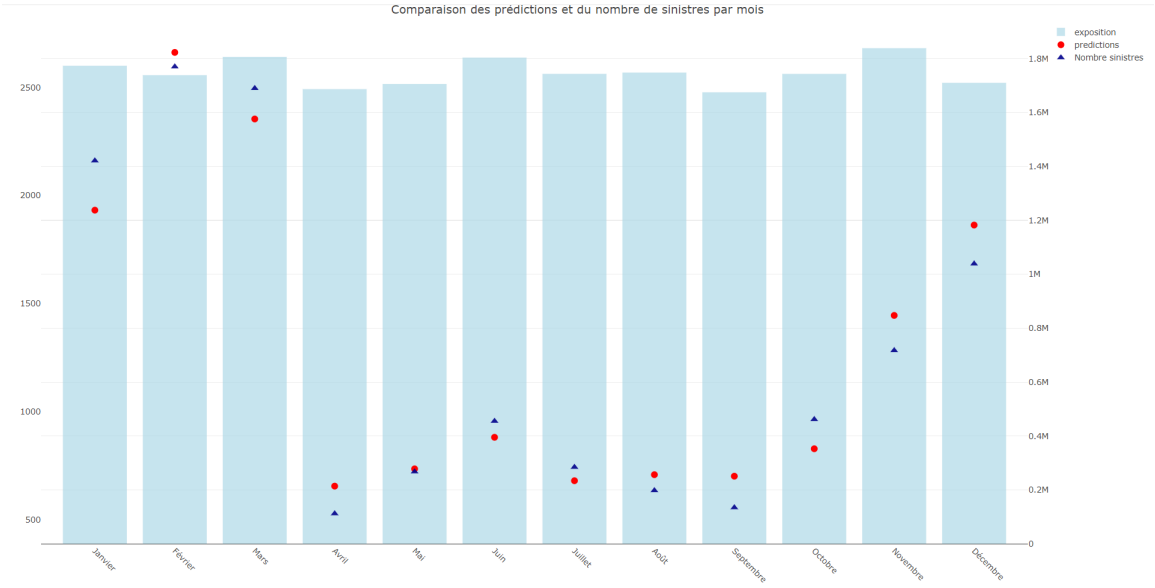

FIGURE 20 – Comparison of predictions and observations on a monthly scale

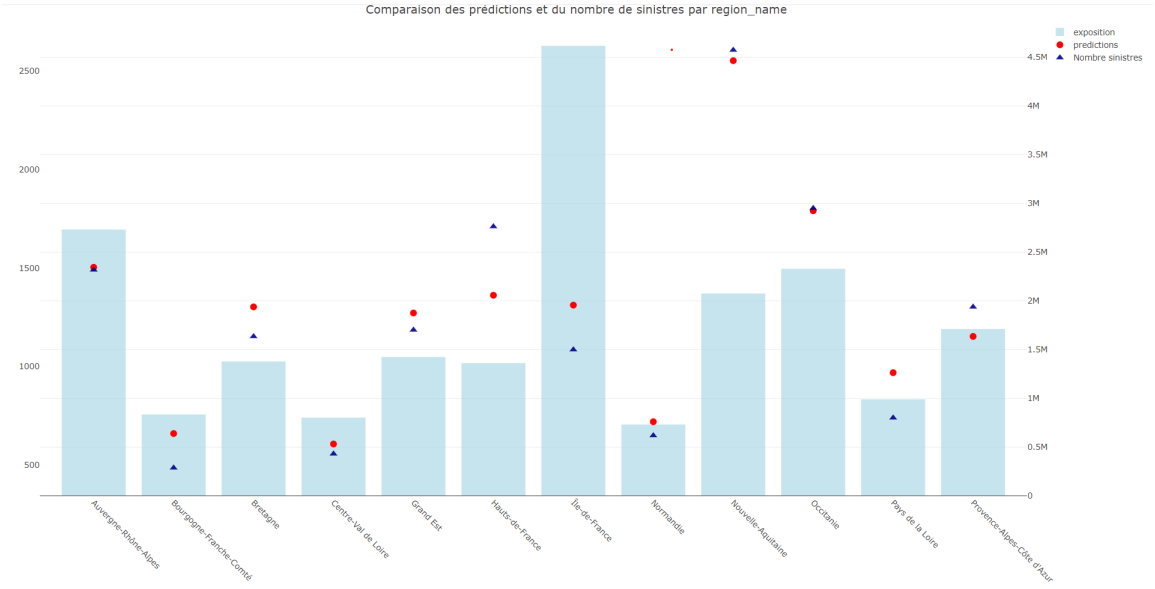

FIGURE 21 – Comparison of predictions and observations on a regional scale

Regarding the projections, we initially noticed a stability difference between the scenarios. In a multi-scenario approach and by comparing the projected number of claims on a monthly and regional scale for the same model but forced by two different RCPs, we observe much more unstable effects for RCP 8.5 :

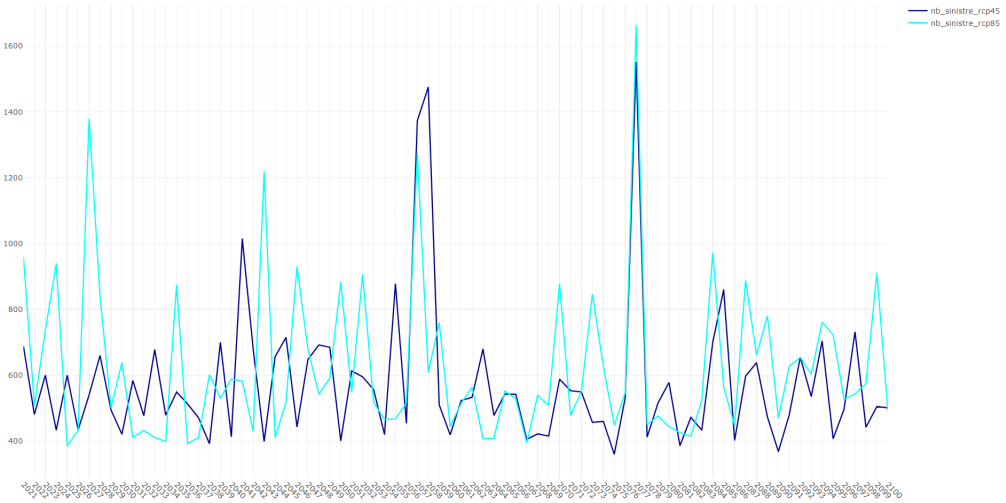

FIGURE 22 – Evolution of the number of claims in Normandie for a CNRM model

We present an example graph that illustrates these differences in terms of the number of peaks for the two scenarios. The selected example focuses on Normandie and a CNRM model ; however, we observe this effect across the majority of models, regions, and months. This instability is consistent and can be explained by retroaction effects that may be amplified in the case of more significant warming. Thus, we can infer that a world with an increasing concentration of greenhouse gases would become increasingly unpredictable. Additionally, we see here that the storm risk may not necessarily increase, but the frequency of extreme events could be higher.

We then conducted an analysis of the annual evolution of claims on a national scale with a multi-model and multi-scenario perspective. We plotted these trends by averaging the outputs of each model, framed by the results of the maximum and minimum outputs, for each scenario :

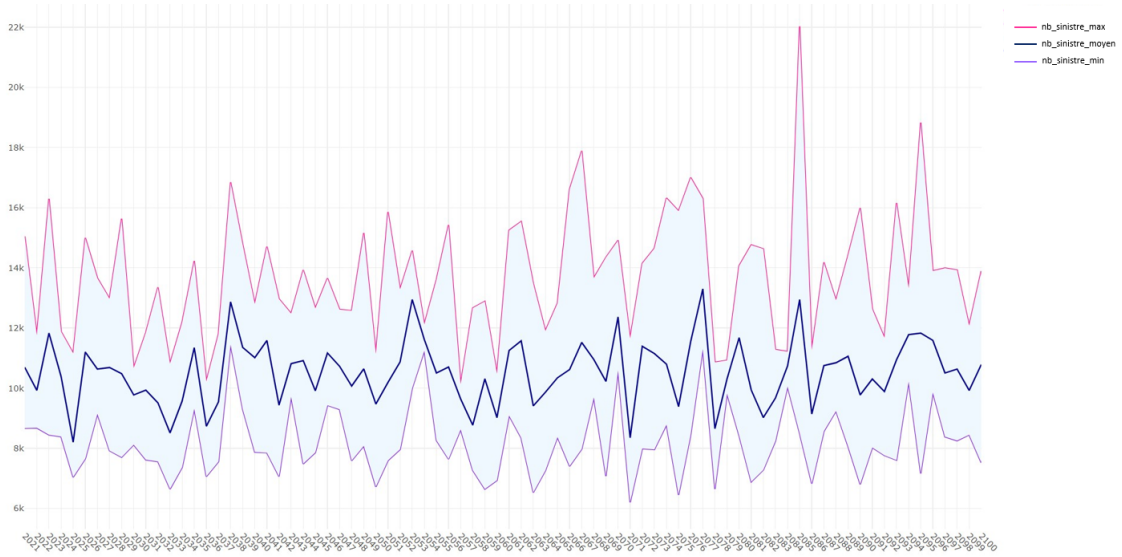

FIGURE 23 – Annual evolution of the number of claims by combining all models for RCP 8.5

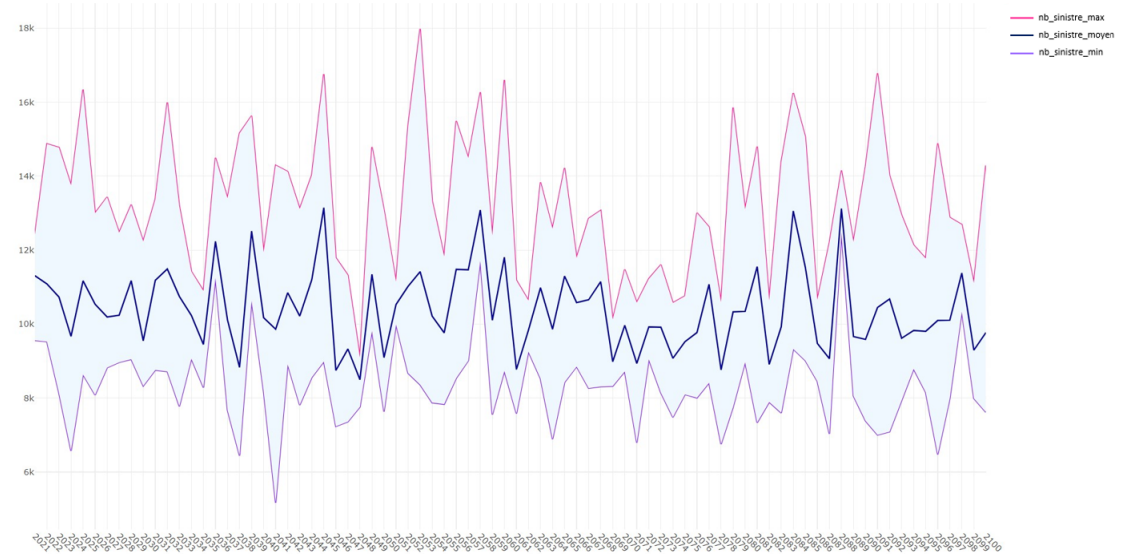

FIGURE 24 – Annual evolution of the number of claims by combining all models for RCP 4.5

We do not observe any particular trend in these two graphs. This conclusion aligns with the studies conducted by COVÉA, which concluded in their white paper of January 2022 that storm risk would not experience an increase in terms of frequency or severity by 2050. Our study thus corroborates these results.

Table des matières

!

Remerciements

Je remercie toute l'équipe d'ADDACTIS pour m'avoir encadré et suivi dans la réalisation de ce mémoire. Je pense tout particulièrement à Romain NOBIS, mon tuteur pour cette alternance, mais aussi à Auriol WABO-FOKA, Annabelle GARRIGUE et Benjamin POUDRET.

Je remercie mes professeurs et intervenants pour m'avoir transmis l'attrait pour les sciences actuarielles, pour le temps qu'ils m'ont accordé et pour les conseils qu'ils m'ont apportés.

Je remercie toutes les personnes qui ont participé et participent de manière directe ou indirecte à mon épanouissement, que ce soit mes parents, mes professeurs ou mes amis. Sans les nommer explicitement puisqu'ils se reconnaîtront, je leur en suis extrêmement reconnaissant et je n'oublie pas que c'est en partie grâce à eux que j'en suis arrivé là.

Introduction

Le continent européen dans son ensemble est soumis au risque "tempête", avec une concentration plus importante dans le nord du continent en raison du passage fréquent de perturbations atmosphériques. En France, ce risque est particulièrement élevé dans la partie nord du pays et sur les zones côtières. Selon Météo-France, il surviendrait en moyenne quinze tempêtes par an en France, dont une sur dix peut être qualifiée de "forte" selon leur propre critère, qui stipule qu'une tempête est considérée comme "forte" si au moins 20% des stations départementales enregistrent un vent maximal instantané supérieur à 100 km/h.

Les dommages causés par les tempêtes sont considérables, premièrement car ce sont des phénomènes qui se produisent sur de vastes zones, deuxièmement car leur intensité est variable et difficilement prévisible. Dans le domaine de l'assurance dommage en France, elles représentent le principal poste de sinistralité. Durant l'année 2022 par exemple, les coûts engendrés par des événements de grêle et de tempêtes se sont élevés à 6,4 milliards d'euros selon France Assureurs. Cette année 2022 a ainsi été estimés comme la pire depuis 1999 marquée par les mémorables tempêtes Lothar et Martin qui avaient généré un coût de 6,86 milliards d'euros.

Dans un contexte actuel où le dérèglement climatique est de plus en plus tangible, une question intervient de manière instinctive : Allons-nous connaître un accroissement des événements climatiques extrêmes et en particulier une augmentation du nombre de tempêtes ?

Dans ce mémoire, nous essaierons alors de comprendre comment ce risque serait susceptible d'affecter la France métropolitaine au cours du 21e siècle en nous basant sur les projections climatiques des météorologues en lien avec les scénarios mis en place par le Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC).

D'un point de vue assurantiel, l'objectif est de comprendre l'évolution de la fréquence de sinistre à horizon 2100. Nous étudierons pour cela la sinistralité tempête historique sur un portefeuille multirisque habitation en mesurant l'impact sur la

fréquence de sinistre d'une multitude de paramètres météorologiques liés de manière directe ou indirecte à la tempête sans oublier les variables caractéristiques du bâtiment. Après l'identification des paramètres les plus importants, nous combinerons la projection des paramètres météorologiques avec notre modèle de fréquence ajusté pour en déduire une évolution de la fréquence de sinistre tempête au cours des prochaines années.

Chapitre 1

Contexte et cadre de l'étude

1.1 Le phénomène météorologique

1.1.1 Les paramètres météorologiques qui gouvernent l'état atmosphérique

La Terre est entourée d'une couche de gaz appelée l'atmosphère, qui s'étend jusqu'à environ 800 km d'altitude. C'est dans cette enveloppe que naissent les phénomènes météorologiques qui influencent le climat. Pour évaluer l'état de l'atmosphère et anticiper ses perturbations, trois paramètres clés sont à utiliser :

— La **pression atmosphérique** correspondant à la pression exercée, sur une unité de surface, par la masse de la colonne d'air située à l'aplomb de cette surface. La pression diminue avec l'altitude selon un gradient de 3 hectopascals tous les 25m. On appelle **dépressions** les zones de basses pressions et **anticyclones** les zones de hautes pressions sachant que la pression de référence est de 1 015 hectopascals.

— La **température**, paramètre phare en météorologie qui varie en fonction des saisons, du positionnement géographique et des conditions climatiques.

— Le **taux d'humidité** qui indique la quantité de vapeur d'eau contenue dans l'air.

La survenance de vents violents qui engendrent les tempêtes s'explique grâce à l'étude de ses paramètres, qui seront donc cruciaux pour notre étude.

1.1.2 Définition

Une **tempête**[1] est une perturbation atmosphérique, plus précisément une dépression atmosphérique (une zone de basse pression), pouvant s'étendre sur une

1. Source : Les informations de cette section sont en partie extraites du dossier d'information sur les tempêtes du Ministère de la transition écologique indiqué dans la bibliographie

largeur de 2000 km où deux masses d'air, ayant des températures et un taux d'humidité différents, se confrontent. Cette confrontation génère des vents pouvant être extrêmes. Les météorologues évoquent le terme « tempête » lorsque les vents dépassent 89 km/h. A noter qu'en assurance la vitesse retenue pour définir une tempête est différente, elle sera dans ce cas de 100 km/h.
Les processus qui génèrent les circulations des masses d'air et qui mènent à la formation des tempêtes sont complexes. Nous les décrirons dans la prochaine partie de manière simplifiée pour l'hémisphère nord sachant que les processus sont identiques mais inversés dans l'hémisphère sud.

1.1.3 Processus de circulation des masses d'air

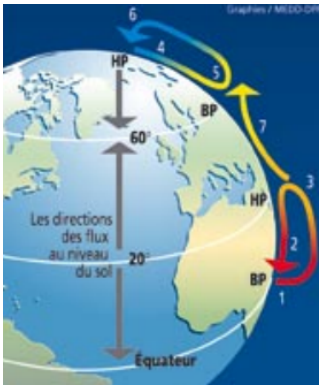

FIGURE 1.1 – Schéma sur la circulation des masses d'air

(1) Au niveau de l'équateur, l'air chauffé par le Soleil, monte en altitude, ce qui provoque une zone de basse pression au sol.
(2) L'air tropical se met alors en mouvement pour remplir cette dépression, tandis qu'en altitude, l'air se refroidit et redescend peu à peu jusqu'à atteindre le sol.
(3) Ce phénomène se déroule au niveau de la latitude 20°. Le cycle vertical que l'on vient de décrire forme comme un tube autour de la Terre qu'on nomme en météorologie, cellule de Hadley.
(4) Au niveau du pôle, une zone de haute pression est présente due à un air très froid et dense. Ce déséquilibre entraine des mouvements vers les zones tempérées au sud.
(5) L'air passant alors sur l'océan se charge en humidité et se réchauffe, ce qui provoque sa montée en altitude. Une zone de basse pression est alors créée au niveau de la latitude 60°.
(6) Une seconde boucle est générée lorsque l'air retourne vers le pôle.
(7) Au niveau des zones intermédiaires, que l'on nomme usuellement «zones tempérées», les différences de températures sont moins importantes. Les circulations verticales en boucles sont alors moins marquées. Il existe tout de même un courant principal sud-nord entre les hautes pressions tropicales et les basses pressions du cercle polaire, cependant les forces de Coriolis prennent le pas sur les forces ascendantes ou descendantes.
Les principales circulations d'air sont alors généralement observées dans un plan horizontal en basse altitude ce qui donnent naissance aux tempêtes dites des «latitudes tempérées». C'est ce type de tempête que nous allons étudier dans ce mémoire.

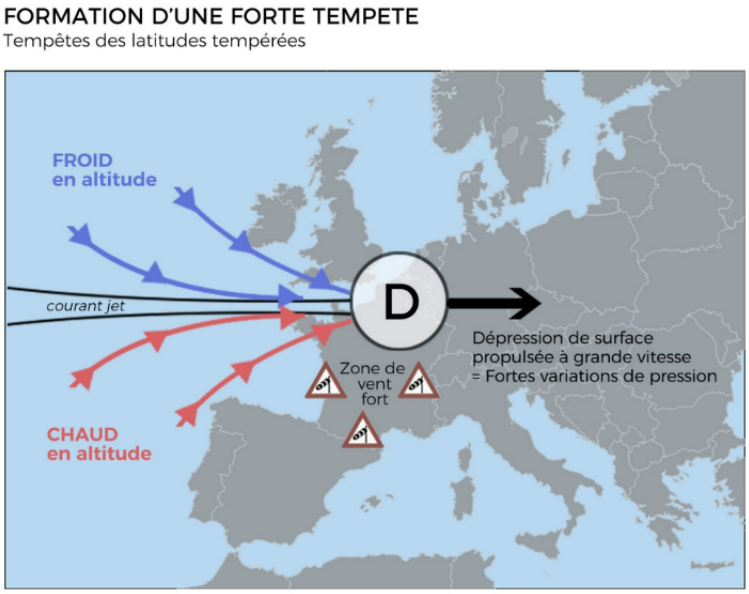

FIGURE 1.2 – Formation des tempêtes des latitudes tempérées (Source : livre METEO EXTREME p117)

1.1.4 La formation d'une tempête en Europe

La majorité des pertubations qui affectent le continent Européen se forme au niveau de l'Atlantique sur le front polaire, une surface séparant une zone d'air froid polaire et une zone d'air chaud tropical. On appelle **rail des dépressions**, le rail le long duquel se suivent les tempêtes, de l'océan vers notre continent.
Des cyclones tropicaux atlantiques sont aussi susceptibles de se transformer en tempête à la fin de leur parcours.
La mer Mediterranée peut aussi être le berceau de certaines dépressions qui affectent en premier lieu l'Espagne, l'Italie et le sud de la France.

En outre, il est important de préciser que les tempêtes des régions tempérées sont les plus fréquentes sur les **mois d'automne et d'hiver**. A cette période, les océans ont une température supérieure à la moyenne alors que l'air polaire est déjà froid ce qui crée un gradient de température relativement élevé entre deux masses d'air et donne ainsi naissance à une tempête. Par ailleurs, plus ce gradient est important plus la tempête est puissante. Ces tempêtes hivernales traversent l'Océan Atlantique et arrivent sur le continent européen en parcourant plusieurs milliers de kilomètres en quelques jours grâce à une vitesse de déplacement de l'ordre de 50 km/h. Ce genre de tempête traverse le territoire français en moyenne en trois jours.

1.1.5 Les conséquences climatiques des tempêtes

— Les tempêtes sont caractérisées en première instance par des **vents** anormalement élevés qui sont générés par l'inégalité des pressions résultant d'une différence de température des différentes masses d'air. Plus le gradient de pression entre la zone dépressionnaire et la zone anticyclonique est important, plus les vents seront violents. A noter que l'énergie d'un vent est proportionnelle au carré de sa vitesse. Pour exemple un vent de 200 km/h exercera une force quatre fois supérieure à un vent de 100km/h. D'autres facteurs sont susceptibles d'influencer la vitesse du vent, on pense notamment à la rotation de la terre (par l'intermédiaire de la force de Coriolis), à la courbure des trajectoires des courants aériens et aux frottements sur la surface terrestre.

En effet les forces de frottement influencent la vitesse des vents. Les bâtiments ou les reliefs pouvant accroitre ou diminuer cette vitesse. On notera qu'en règle générale, les rafales soufflant en mer sont supérieures aux rafales terrestres, sachant que l'intensité des vents diminue en fonction de l'avancement de la perturbation sur le continent. Ceci explique **l'exposition plus importante des zones littorales** au risque tempête.

Néanmoins, grâce à un écoulement de l'air atypique, le pouvoir destructeur des vents terrestres peut être sensiblement supérieur à celle des vents marins. On sait que le développement d'une tempête est dû aux interactions entre les couches de l'atmosphère inférieures et le **courant-jet/jet-stream** (zone de forts vents d'ouest, formant un ruban étroit situé dans des latitudes moyennes) et sa force est peu sensible à la surface du sol quand la dépression est présente. Ainsi, lorsque le courant-jet est assez puissant et qu'il est combiné à des phénomènes tourbillonnants d'une amplitude suffisante, les pertes de vitesse des vents terrestres peuvent être compensées voire même amplifiées.

— Nous avons ensuite **les vagues et les marées** qui sont susceptible de survenir suite au passage de tempêtes puissantes. En effet, des marées anormalement hautes sont observées sous l'effet conjugué du vent et d'un état dépressionnaire très marqué et on assiste parfois à un phénomène que l'on nomme **marée de tempête**, correspondant à une hausse temporaire du niveau de la mer, qui peut provoquer des dégâts considérables en allant parfois jusqu'à la subsmersion du littoral. L'exemple de la tempête Xynthia en 2010 est caractéristique de ce type de phénomène. A noter qu'il existe une relation entre la hauteur des vagues et la vitesse du vent, pour exemple un vent de 130 km/h a la capacité de provoquer des vagues d'une hauteur de 15 mètres environ.

— Nous avons pour finir les **pluies** potentiellement importantes lors de tempêtes et qui sont génératrices d'aléas importants comme les **inondations**, les

glissements de terrains et les **coulées boueuses**.

Pour notre étude, nous nous concentrerons exclusivement sur les vents

Nous comprenons par ces développements que la tempête est un phénomène météorologique extrêmement complexe. La genèse de ces événements fait l'objet de nombreuses recherches actuellement et notre compréhension des lois physiques qui régissent ces tempêtes, s'affine de plus en plus. Notre objectif est de pouvoir traduire ces connaissances météorologiques dans des *inputs* de modèles pour modéliser ce risque de la manière la plus fine possible.

Dans la partie suivante, nous présenterons le risque tempête d'un point vue réglementaire en assurance.

1.2 Le risque tempête en assurance

Les tempêtes sont responsables de la majorité du coût global des sinistres de la **garantie TGN** (Tempête, Grêle, Neige). Les tempêtes sont en effet à la fois les événements les plus fréquents et les plus coûteux de cette garantie. Néanmoins, d'un point de vue assurantiel les **tempêtes ne sont pas considérées comme des catastrophes naturelles**.

La **loi n°90-509 du 25 juin 1990** (extension de la **loi n°82-600 du 13 juillet 1982**) prévoit que les effets du vent dus aux tempêtes sont écartés du champ d'application du régime d'indemnisation des catastrophes naturelles. La loi n°82-600 précise quant à elle, le cadre dans lequel s'inscrit la procédure d'indemnisation des victimes de catastrophes naturelles, en se fondant sur le principe de solidarité nationale.

Les tempêtes relèvent ainsi d'une garantie spécifique volontaire de la part de l'assuré, alors annexée aux contrats classiques d'assurance, comme ceux concernant les dommages aux biens ou les pertes financières.

1.2.1 Le régime des tempêtes

La garantie contre les tempêtes, ouragans et cyclones est régie par l'**article L. 122-7** du Code des Assurances qui précise que « *les contrats d'assurance garantissant les dommages d'incendie ou tous autres dommages à des biens situés en France, ainsi que les dommages aux corps de véhicules terrestres à moteur, ouvrent droit à la garantie de l'assuré contre les effets du vent dû aux tempêtes, ouragans et cyclones, sur les biens faisant l'objet de tels contrats*».

Les contrats multirisques habitation (**MRH**) ainsi que les contrats d'assurance automobile incluent donc la garantie tempête. Le Conseil d'Etat a précisé que lorsque ces contrats sont souscrits, la garantie tempête est obligatoirement accordée.

En pratique, la garantie tempête est une extension de la garantie incendie d'un contrat d'assurance MRH par exemple.

Cependant, en matière d'automobile, les choses sont un peu différentes. En effet, l'obligation d'assurance prévue à l'article L. 211-1 du code des assurances ne concerne que la couverture de la responsabilité civile du conducteur vis-à-vis des tiers (couverture dite "au tiers"). La garantie "dommage" (vol, incendie, bris de glace, dommages au véhicule) d'un tel contrat est accessoire (généralement contenue dans les contrats d'assurance "Tous risques"). Si elle n'a pas été souscrite, les garanties "tempête" et "catastrophes naturelles" ne pourront pas être mobilisées.

Les biens garantis

Les biens garantis sont simplement ceux couverts par le contrat d'assurance. Pour cause, l'article L. 122-7 de code des assurances indique que sont exclus les dommages causés aux récoltes non engrangées (non abritées dans une grange), aux cultures, au cheptel vif hors bâtiments (animaux) et aux bois sur pied (particulier qui exploite lui-même les arbres qu'il a achetés).

Pour exemple dans un contrat d'assurance MRH, les biens garantis sont les bâtiments respectant les normes de construction (aménagements immobiliers, clôtures, garages, caves) ce qui exclut dans la majorité des cas les bâtiments et hangars de construction légère ainsi que le contenu (meubles ou objets).

Les événements couverts

La garantie couvre les effets du vent provoqués par une tempête, un ouragan ou un cyclone mais la loi ne définit pas ces termes. L'article L. 122-7 du Code des assurances, indique seulement que «*les effets du vent dû à un événement cyclonique pour lequel les vents maximaux de surface enregistrés ou estimés sur la zone sinistrée ont atteint ou dépasse 145 km/h en moyenne sur dix minutes ou 215 km/h en rafales*» relèvent des dispositions relatives à la garantie des catastrophes naturelles.

Le rôle revient donc aux assureurs de définir dans leur contrat, les conditions de mise en œuvre de cette garantie. En règle générale, les assureurs ne prennent en compte que les vents d'une intensité anormale (**plus de 100 km/h**), à l'origine de

nombreux dommages affectant des bâtiments de « bonne construction » (c'est-à-dire en mesure de résister à l'action habituelle des vents). Il faut, en plus, que ces dommages aient une ampleur exceptionnelle, ce qui sous-entend des **destructions nombreuses dans la commune** où se situent les biens sinistrés et dans les **communes environnantes**.

1.2.2 Comparaison des régimes tempêtes et catastrophes naturelles

Nous ne présenterons pas en détail le régime des catastrophes naturelles qui s'écarte de notre sujet. En revanche, nous dressons ci-après un tableau récapitulatif des deux régimes pour évoquer de manière simplifiée les différences entre les deux garanties :

	Tempêtes	**Catastrophes naturelles**
TEXTES	**L. 122-7** du code des assurances	**L.125-1 à L.125-6, A. 125-1 et A. 125-2** du code des assurances
QUELS CONTRATS ?	Tous les contrats garantissant les **dommages d'incendie** : habitation et automobile	Tous les contrats garantissant les dommages d'incendie et tous les autres **dommages à des biens** : habitation, automobile, exclusion des bateaux et avions
EVENEMENTS GARANTIS	Vent d'une intensité anormale	Intensité anormale d'un agent naturel (Arrêté interministériel de catastrophe naturelle)
DOMMAGES COUVERTS	Dommages survenus lors du sinistre et au cours des 48 heures suivantes, pertes indirectes si garanties au contrat, corporels : non	Dommages matériels directs, pertes indirectes : non, corporels : non
EXCLUSIONS	Récoltes non engrangées (culture, cheptel hors bâtiment) + Exclusions du contrat	Récoltes non engrangées (culture, cheptel hors bâtiment) + Exclusions du contrat

Table 1.1 – Comparaison des régimes d'assurance Tempêtes/Catastrophes naturelles

Dans la prochaine section, nous présenterons la philosophie sous-accente des modèles climatiques des météorologues. L'objectif étant d'acquérir des notions climatologiques suffisantes pour manipuler ces données de manière adéquate.

1.3 Les modèles climatiques liés aux projections météorologiques

Nous commençons cette partie par énoncer les définitions de termes courants employés par les climatologues et plus généralement par les experts du changement

climatique. Nous pourrons ensuite expliquer leurs modèles climatiques, les objectifs qu'ils essaient d'atteindre ainsi que les limites qui gravitent autour de ces processus complexes.

1.3.1 Définitions

Le climat

Selon l'Organisation Météorologique mondiale (OMM), le climat peut être défini comme la description statistique en termes de moyenne et de variabilité des quantités météorologiques pertinentes (température, humidité, vitesse de vent...) sur une période donnée. Cette période a été généralement définie comme étant de 30 ans.

Le changement climatique

Selon l'OMM, on entend par changement climatique une variation statistiquement significative du climat en termes d'état moyen ou de variabilité, qui persiste pendant une période prolongée (généralement des décennies ou plus). Dans les applications pratiques, des périodes de 30 ans sont fréquemment utilisées.

La variabilité du climat

La variabilité climatique est le terme employé pour décrire les variations du climat à toutes les échelles temporelles et spatiales, au-delà des événements météorologiques individuels. Cette variabilité est soit due à des processus naturels internes au système climatique, (variabilité interne), soit liée à des variations de facteurs externes qu'ils soient naturels ou d'origine anthropique (variabilité externe).

La **variabilité interne** résulte de processus chaotiques dans le système climatique et d'interactions non-linéaires entre ses composants (l'atmosphère, l'hydrosphère, la cryosphère, la biosphère...).

La **variabilité externe** inclut des facteurs extérieurs au système climatique. Il s'agit notamment de facteurs naturels tels que la variabilité solaire, les variations orbitales ou les éruptions volcaniques, mais aussi les forçages anthropiques tels que les émissions de gaz à effet de serre et d'aérosols dans l'atmosphère et les changements d'utilisation des sols.

Les modèles climatiques globaux et régionaux (GCM et RCM)

Les modèles numériques du climat sont utilisés pour projeter l'évolution future possible du système climatique ainsi que pour comprendre le système climatique lui-même. Ils sont construits sur des descriptions mathématiques des processus physiques qui régissent le système climatique (conservation de la quantité de mouvement, de la masse et de l'énergie, etc). Les solutions numériques des équations sous-jacentes sont obtenues à l'aide de supercalculateurs.

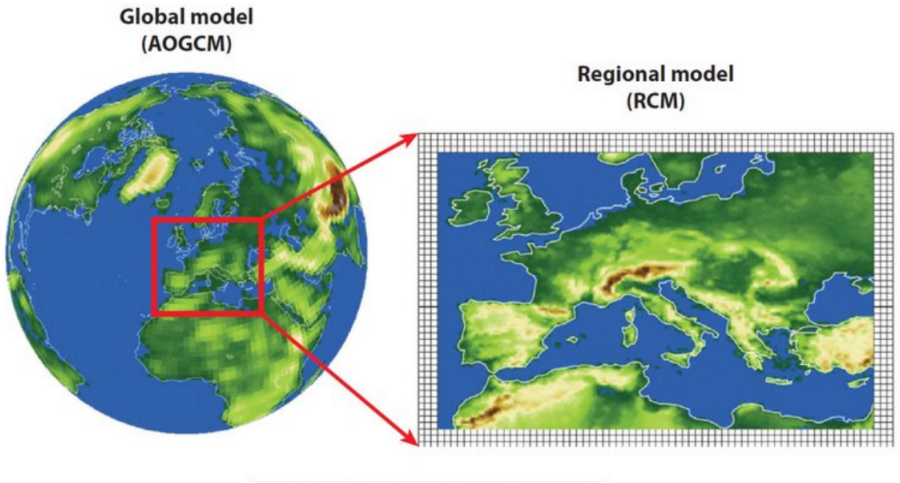

FIGURE 1.3 – Vue schématique du passage d'un GCM à un RCM (Source : Giorgi et al., 2015.)

Les modèles de circulation générale (MCG ou GCM en anglais pour *General Circulation Model*) sont des modèles climatiques numériques globaux utilisés pour étudier le changement climatique à l'échelle mondiale. Ils décrivent les différentes composantes du système terrestre ainsi que les interactions non linéaires et les rétroactions entre eux. Afin de simuler le climat passé, des valeurs mesurées sont utilisées comme données de forçage, alors que pour les projections futures, les valeurs de certains scénarios d'émissions sont utilisées.

En raison du grand nombre de points de données et de la grande complexité des GCM, leur intégration nécessite une grande quantité de ressources informatiques. La résolution de leur maille horizontale varie actuellement de 100 à 500 km et ils fournissent des résultats avec une fréquence temporelle de 6 heures. En raison de cette échelle horizontale et temporelle relativement grossière, les GCM sont insuffisants pour de nombreux aspects des estimations de la variabilité et du changement climatique à l'échelle régionale et locale. Par conséquent, la réduction d'échelle est nécessaire pour décrire les conséquences locales du changement global, ce qui peut être fait à l'aide d'une réduction d'échelle empirique-statistique ou grâce à la ré-

duction d'échelle dynamique à l'aide de modèles climatiques régionaux (RCM pour *Regional Climatic Model*).

Les RCM sont utilisés pour réduire l'échelle des simulations des GCM en utilisant les données de sortie des GCM comme conditions aux limites latérales (les frontières notamment). Les RCM sont généralement exécutés à une résolution horizontale de 10 à 50 km sur une région d'intérêt spécifique (par exemple, l'Europe dans le cas d'EURO-CORDEX, notre source de données principale que l'on évoquera dans les prochaines parties). Grâce à la résolution explicite des processus importants (les circulations d'air en montagne, les contrastes terre-océan) et des schémas de paramétrisation adaptés à des résolutions plus élevées, les RCM sont capables de fournir des caractéristiques plus détaillées du climat local.

Différence entre prédiction climatique et prévision météorologique

Les prévisions météorologiques sont élaborées à partir des résultats de modèles numériques simulant la circulation atmosphérique. Ces modèles atmosphériques sont capables de donner des résultats convenables jusqu'à une dizaine de jours environ. Sur ce genre de période, seule l'atmosphère influence le temps qu'il fait puisque les réservoirs lents tels que l'océan et la biosphère ont une cinétique beaucoup plus longue.

Les modèles météorologiques prennent les valeurs de températures de l'air, de l'humidité, de pression et de vitesse de vent à l'instant t et calculent le temps pour les jours suivants.

Les modèles climatiques, quant à eux, se basent sur les données météorologiques passées et actuelles puisque le climat est un système couplé où l'atmosphère, la cryosphère, la biosphère, les surfaces et eaux continentales et l'océan évoluent avec une cinétique relativement lente par l'intermédiaire d'échange d'énergie, de quantité de mouvement, d'eau et de composés chimiques.

Il est donc important de faire une distinction entre les concepts de météo et de climat qui se basent sur des modélisations différentes et sur une temporalité différente. Les météorologues travaillant sur des périodes assez courtes de quelques jours tandis que les climatologues utilisent des périodes d'au moins 30 ans pour construire leurs tendances.

Les scénarios climatiques

Les scénarios climatiques (ou projections climatiques) sont des représentations de divers états futurs possibles du système climatique, basées sur des simulations de

modèles numériques.

Ces modèles décrivent les processus et les interactions complexes qui affectent le système climatique en incluant des informations sur le forçage climatique anthropique à l'instar de facteurs comme le développement socio-économique, technologique, démographique et environnemental qui sont caractérisés dans les modèles climatiques par des changements équivalents en termes de concentrations de gaz à effet de serre ainsi que des changements dans l'utilisation et la couverture des terres.

Comme l'évolution future des facteurs anthropiques ne peut être connue à l'avance, leurs effets potentiels sont explorés à travers différents scénarios décrivant plusieurs voies possibles d'émission (et donc de concentration de gaz à effet de serre).

Lors d'une simulation climatique, le scénario d'émission choisi fournit des données de forçage en entrée pour le modèle climatique ce qui entraîne la réaction physique du système climatique à ce forçage anthropique futur particulier. En raison de ce caractère dépendant du forçage, les résultats des modèles climatiques ne sont pas interprétés comme des prévisions, mais comme des projections basées sur des scénarios.

Trois séries de scénarios d'émissions ont été utilisées dans les modèles de circulation générale (GCM). Ces scénarios ont servi de base aux quatre derniers rapports d'évaluation du GIEC (2001, 2007, 2014, 2021). Il s'agit des modèles dits SRES (*Special Report on Emissions Scenarios*), RCP (*Representative Concentration Pathways*) et SSP-RCP (*Shared Socio-economic Pathways*).

L'ensemble EURO-CORDEX est basé uniquement sur les scénarios RCP.

Les scénarios RCP

Les scénarios RCP ont été élaborés lors du cinquième rapport d'évaluation du GIEC en utilisant des modèles d'évaluation intégrée, des modèles climatiques et des modèles d'impact.

Quatre ensembles de scénarios ont été créés, nommés d'après leur forçage radiatif total (en W/m^2) en 2100 par rapport à 1750 : RCP 8.5, RCP 6.0, RCP 4.5 et RCP 2.6. Le RCP 8.5 représente une très forte émission de gaz à effet de serre, il traduit un forçage radiatif de 8,5 W/m^2, qui continue à augmenter même après 2100 ; le RCP4.5 et RCP6.0 sont des scénarios de stabilisation, ce qui signifie que les forçages se stabilisent à leur valeur donnée vers la fin du siècle ; et le RCP2.6 représente un scénario d'atténuation agressif avec des émissions futures négatives

considérables.

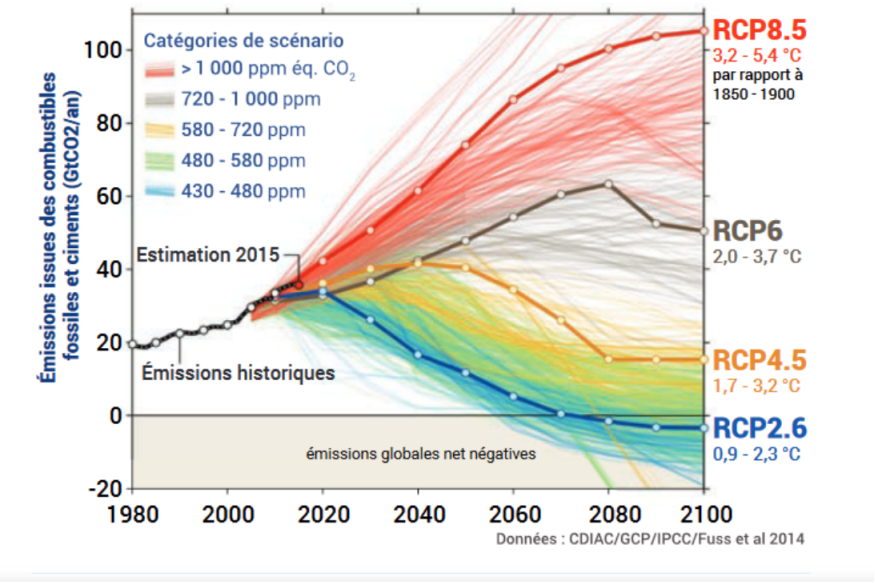

FIGURE 1.4 – Évolution des émissions entre 1980 et 2100, selon les différents scénarios disponibles. Les quatre scénarios sélectionnés dans le cadre du 5e rapport du Giec (RCP) sont mis en évidence. (Source : Global Carbon Project)

Selon le cinquième rapport du GIEC, le changement de température à la surface du globe d'ici à la fin du 21e siècle devrait rester inférieure à 2 °C par rapport à la période 1850-1900 (c'est-à-dire que l'objectif important de 2°C peut être maintenu), pour les scénarios RCP2.6 et RCP4.5, mais il est probable qu'il dépasse ce seuil pour les scénarios RCP6.0 et RCP8.5.

Le concept de forçage

Selon le GIEC, « le forçage radiatif mesure l'impact de certains facteurs affectant le climat sur l'équilibre énergétique du système couplé Terre/atmosphère. Le terme « radiatif » est utilisé car ces facteurs modifient l'équilibre entre le rayonnement solaire entrant et les émissions de rayonnements infrarouges sortant de l'atmosphère. Cet équilibre radiatif connu sous le nom d'**effet de serre** contrôle la température à la surface de la planète. Le terme forçage est utilisé pour indiquer que l'équilibre radiatif de la Terre est en train d'être déstabilisé. Le forçage radiatif est généralement quantifié comme le taux de transfert d'énergie par unité surfacique du globe, mesuré dans les hautes couches de l'atmosphère, et il est exprimé en Watt par mètre carré. Un forçage radiatif causé par un ou plusieurs facteurs est dit positif lorsqu'il entraîne un accroissement de l'énergie du système Terre/atmosphère et

donc le réchauffement du système. Dans le cas inverse, un forçage radiatif est dit négatif lorsque l'énergie va en diminuant, ce qui entraîne le refroidissement du système. Les climatologues sont confrontés au problème complexe d'identifier tous les facteurs qui affectent le climat, ainsi que les mécanismes de forçage, de quantifier le forçage radiatif pour chaque facteur et d'évaluer la somme des forçages radiatifs pour un groupe de facteurs. »

D'un point de vue climatique, les forçages correspondent ainsi aux inputs du modèle climatique. Dans les modèles n'incluant pas le cycle du carbone, la concentration du CO2 est fixée à l'entrée du modèle, ce qui peut être considéré comme un forçage. En revanche, dans les modèles incluant le cycle du carbone, les changements de concentration en CO2 seront fonction du climat et des changements liés à l'activité industrielle. Dans ce cas, les niveaux de CO2 seront des rétroactions et non des forçages. En réalité, presque tous les éléments constitutifs de l'atmosphère peuvent être considérés comme des rétroactions. Ainsi, définir un forçage dépend des rétroactions possibles dans le modèle.

Le concept de rétroaction

D'après le GIEC, une rétroaction climatique est une « interaction dans laquelle la perturbation d'une variable climatique provoque, dans une seconde variable, des changements qui influent à leur tour sur la variable initiale. Une rétroaction positive accentue la perturbation initiale, une rétroaction négative l'atténue. La perturbation initiale peut découler d'un forçage externe ou relever de la variabilité interne. »

1.3.2 Les limites de la modélisation climatique

Chaque réalisation du modèle climatique est une représentation incomplète de la réalité. Nous sommes limités, d'une part, par la puissance de calcul qui serait nécessaire pour travailler à toutes les échelles temporelles et spatiales, d'autre part, par notre compréhension des processus du système terrestre qui ne peuvent pas être simulés dans leur entièreté.

En effet, les processus du système climatique se produisent sur des échelles de temps qui vont des siècles à l'infra-journalier et des échelles spatiales allant de dizaines de milliers de kilomètres à moins d'un kilomètre, il est alors impossible de tous les intégrer dans les modèles. En outre, plusieurs phénomènes comme les cycles de vie des aérosols ne sont pas encore totalement compris et ne sont donc pas directement quantifiables.

Par ailleurs la configuration des modèles influence les résultats. Le nombre de

niveaux verticaux des modèles, les schémas numériques utilisés pour résoudre les équations ou la paramétrisations des modèles sont autant de facteurs participant à ces variations de résultats.

Les autres limites inhérentes aux projections climatiques sont l'incertitude des scénarios, car les scénarios RCP sont basés sur certaines hypothèses concernant l'avenir et la variabilité climatique interne.

En général, on peut dire que les modèles climatiques sont bons pour simuler l'état et les tendances du système climatique pour des tranches de temps et des régions importantes. Une attention particulière doit être apportée lorsqu'on utilise des RCM pour étudier des événements se produisant à de petites échelles temporelles et spatiales.

1.3.3 Utilisation des projections climatiques malgré ces difficultés

Il est recommandé d'utiliser le plus grand ensemble possible de projections différentes pour l'évaluation et l'application des résultats des modèles climatiques afin d'obtenir des résultats robustes. Seule une analyse d'ensemble permet de limiter les incertitudes inhérentes au modèle pour évaluer les résultats. Il existe principalement trois approches assez intuitives pour combler les incertitudes :

— L'approche **multi-modèles** qui consiste à combiner les sorties de plusieurs modèles autour d'un seul scénario d'émission.

— L'approche **multi-scénarios** qui se focalise sur plusieurs scénarios d'émission pour le même modèle climatique.

— L'approche **multi-paramètres** qui utilise différents schémas de paramétrage physique d'un même modèle.

Pour notre étude nous effectuerons une approche multi-modèles et multi-scénarios puisque nous n'avons pas la possibilité de gérer les paramètres physiques des modèles climatiques.

Dans la prochaine section, nous nous concentrerons sur le contexte réglementaire assurantiel mettant l'accent de manière croissante sur ces notions de risques climatiques en lien avec le réchauffement planétaire, ce qui a par ailleurs été l'élément déclencheur de notre côté pour mener ce type d'étude.

1.4 Le contexte réglementaire prépondérant sur ces questions sociétales et l'impact dans le domaine de l'assurance et de l'actuariat

L'accord de Paris sur le changement climatique impose à ses signataires de réduire les émissions de gaz à effet de serre dans le but de maintenir l'augmentation de la température mondiale bien en deçà de 2°C et de poursuivre les efforts pour la limiter à 1,5°C par rapport aux niveaux préindustriels.

Dans cette optique de limitation et de compréhension des risques en lien avec le changement climatique, les compagnies d'assurance et leurs actuaires se positionnent comme des acteurs susceptibles d'appréhender, de comprendre et d'évaluer les impacts qui découlent de ces risques, c'est pourquoi les superviseurs et les autorités de régulation comme l'ACPR (Autorité de Contrôle Prudentiel et de Résolution) et l'EIOPA *(European Insurance and Occupational Pensions Authority)*, engagent des actions concrètent pour parvenir à atteindre les objectifs fixés.

1.4.1 L'exercice pilote climatique 2020 de l'ACPR

De juillet 2020 à avril 2021, 9 groupes bancaires et 15 groupes d'assurance ont participé volontairement à un premier exercice pilote mené par l'ACPR visant à mesurer l'impact des risques climatiques sur leur activité. Ses conclusions sur le risque physique, objet de notre mémoire, ont été assez alarmantes puisque nous pourrions assister à une augmentation de la sinistralité d'un facteur 5 à 6 selon les départements français en combinant l'ensemble des catastrophes naturelles et les primes pourraient alors augmenter de 130% à 200% pour compenser les pertes.

Cet exercice a entre autres permis d'apporter une vision nouvelle dans l'évaluation globale des risques car il s'effectuait sur un horizon de 30 ans (de 2020 à 2050) pour prendre en compte la cinétique des effets du changement climatique, contrastant ainsi avec la durée habituelle de stress-test de 3 à 5 ans.

Les acteurs de la place ont ainsi pu appréhender les difficultés autour des données disponibles, des modèles de projection et des hypothèses établies. L'ACPR avait affirmé que ce genre d'exercice était amené à être reconduit en 2023 ce qui fera l'objet de notre prochaine section qui présentera l'exercice en cours.

1.4.2 Evolution de la réglementation concernant l'ORSA

L'ORSA *(Own Risk and Solvency Assessment)* est une série de processus qui forme un instrument d'analyse stratégique et décisionnelle ayant pour objectif d'évaluer

de manière continue et prospective le besoin global de solvabilité en lien avec le profil de risque particulier de chaque entité.

En septembre 2021, la Commission européenne a publié une proposition de révision de l'article 45 de la directive Solvabilité II suite au Pacte vert européen. Les nouvelles dispositions prévoient que les entreprises identifient toute exposition significative aux risques liés au changement climatique en évaluant l'impact de **scénarios de changement climatique à long terme** sur leur activité dans leur ORSA. Deux scénarios à long terme sont imposés par les autorités compétentes sachant que d'autres scénarios pourront être proposés par l'entreprise. Il y aura au minimum :

- Un scénario de changement climatique dans lequel l'augmentation de température mondiale reste inférieure à 2°C et dans l'idéal n'excède pas 1,5°C pour respecter les engagements de l'Union Européenne.
- Un scénario où la température mondiale excède les 2°C.

Nous assisterons ainsi à un changement d'horizon pour l'évaluation des risques liés au changement climatique. Cette évaluation s'effectuera sur le court terme (jusqu'en 2030), le moyen terme (jusqu'en 2050) et le long terme (jusqu'en 2100) contrastant avec l'horizon court terme habituel de l'ORSA.

Par ailleurs, l'EIOPA évoque que la première étape pour intégrer les risques climatiques dans l'ORSA réside dans le fait d'**évaluer leur matérialité**, en d'autres termes d'évaluer le caractère concret de ces risques. Des analyses quantitatives et qualitatives seront attendues. Afin de mener une telle évaluation de la matérialité, les étapes suivantes pourraient être envisagées.

Dans un premier temps, l'entreprise peut définir le contexte dans lequel elle serait exposée au changement climatique en spécifiant les domaines d'activité impactés, l'horizon temporel considéré et le contexte stratégique adopté.

Dans la deuxième étape, l'entreprise recherche quels sont les impacts possibles de ces risques sur ses produits d'assurance en évaluant son exposition relative. Dans cette étape une distinction peut être faite entre les risques physiques et les risques de transition. Les risques liés au changement climatique peuvent en effet être classés en deux catégories. D'après l'EIOPA, « **les risques de transition** sont les risques qui découlent de la transition vers une économie à faible émission de carbone et résiliente au changement climatique. Ils comprennent :

- Les risques politiques qui incluent les exigences en matière d'efficacité énergétique ou de mécanismes de tarification du carbone, ce qui augmenteraient le prix des combustibles fossiles.
- Les risques juridiques qui sanctionneraient les entreprises ayant un impact négatif sur le climat

- Les risques technologiques, dans le cas où une technologie ayant un impact moins néfaste sur la climat remplace une technologie plus néfaste pour le climat.
- Les risques liés au sentiment de marché, si les choix des consommateurs et des entreprises s'orientent vers des produits et des services moins nocifs pour le climat.
- Les risques liés à la réputation, dans le cas où des clients se détourneraient d'une entreprise si cette dernière à la réputation de nuire à l'environnement.

Les risques physiques sont des risques qui découlent des effets physiques du changement climatique. Ils comprennent :

- Les risques physiques aigus qui découlent d'événements météorologiques tels que les tempêtes, les inondations, les incendies...
- Les risques physiques chroniques qui découlent de changements climatiques à plus long terme, tels que les changements de température, l'élévation du niveau de la mer, la réduction de la disponibilité de l'eau... »

Dans la troisième et dernière étape, l'entreprise évalue la matérialité de ces risques sur son bilan en prenant en compte l'actif et le passif. L'entreprise indique son exposition, sa vulnérabilité et la probabilité qu'elle soit impactée par ce type de risque.

Avec l'ensemble de ces changements, les actuaires auront donc pour mission de maîtriser l'utilisation de scénarios climatiques et d'évaluer l'exposition aux risques climatiques sur plusieurs horizons de temps en intégrant ces nouvelles méthodologies dans le cadre de l'ORSA.

1.4.3 L'exercice climatique 2023

L'exercice climatique 2023 de l'ACPR a été lancé le 6 juillet dernier. Il vise à compléter le cadre méthodologique de l'exercice pilote de 2020, incitant les assureurs à poursuivre l'intégration du risque climatique au sein de leurs états financiers et de leur gestion interne, sans oublier l'évaluation de la vulnérabilité des assureurs français face au risque de changement climatique.

Dit autrement, cet exercice cherche à capter d'une part le comportement des assureurs face à des dérives de sinistralité engendrées par le dérèglement climatique (via les scénarios long terme), et d'autre part les possibles impacts sur la solvabilité de ces derniers (via les scénarios court terme). Il s'agira donc également de tester la capacité des modèles à intégrer ce type de risque, à savoir exploiter les données à disposition et identifier les données manquantes afin de pouvoir le mesurer de la façon la plus appropriée. L'objectif est donc prudentiel, mais également préventif.

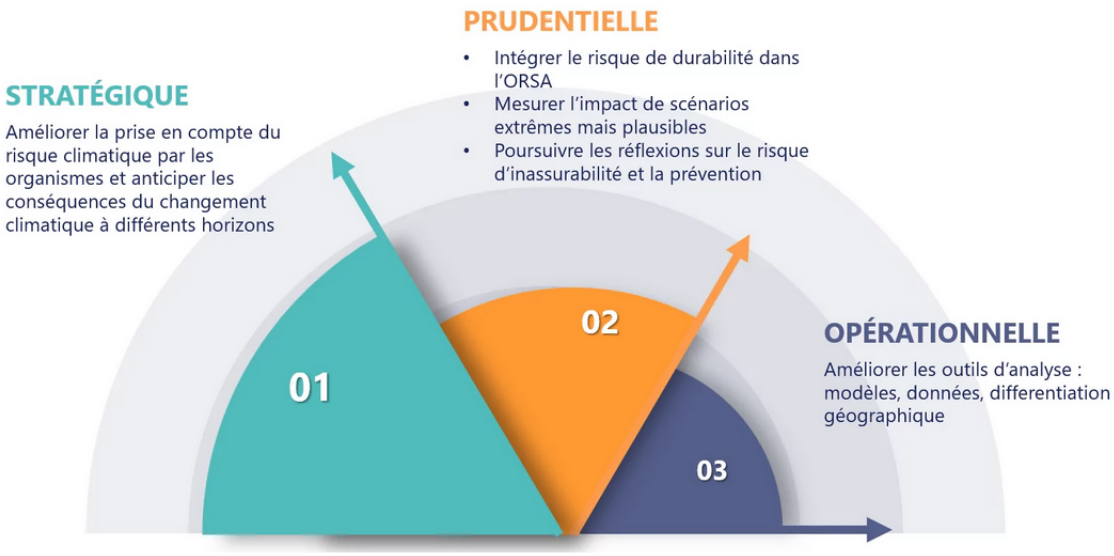

FIGURE 1.5 – Les objectifs de l'exercice climatique en trois dimensions (Source : ADDACTIS)

Des différences sont à souligner par rapport à l'exercice précédent. Pour notre étude, ce qu'il faut principalement retenir concerne l'**utilisation du scénario RCP 4.5** du GIEC pour évaluer les risques physiques aigus, ce qui contraste avec l'exercice pilote de 2020, dans lequel les risques physiques au passif étaient évalués sur la base du scénario RCP 8.5 (correspondant à une hypothèse de hausse des températures comprise entre 1,4°C et 2,6°C en 2050 contre 0,9°C et 2,0°C en 2050 pour le scénario RCP 4.5). Les raisons de ce choix se résume par le caractère plus plausible de ce scénario dans le contexte actuel.

Une autre caractéristique importante de cet exercice réside dans le caractère délégatoire des assureurs envers la CCR (Caisse Centrale de Réassurance) pour effectuer leur modélisation des risques physiques inclus dans le régime des catastrophes naturelles. La CCR leur founit, par type de péril, des estimations de sinistralité moyennes et celles correspondant au quantile à 98 %.

Nous anticipons le caractère obligatoire de ce type d'exercice pour les sessions suivantes. Les assureurs devront probablement effectuer leur propre modélisation lorsque l'ensemble du marché sera à jour sur ces sujets. L'intérêt de notre étude est de proposer une démarche de modélisation et de projection d'un péril inclu dans les risques physiques mais qui n'est pas modélisé par la CCR pour l'exercice climatique de cette année, n'étant pas considéré comme une catastrophe naturelle d'un point de vue assurantiel.

Synthèse

Dans ce chapitre le contexte et le cadre de l'étude ayant été introduits, tout d'abord par la définition du phénomène puis par sa traduction d'un point de vue assurantiel pour ensuite présenter les modèles de projections climatologiques et le contexte réglementaire actuel, nous passons désormais à la présentation des données météorologiques sur lesquelles se basent l'étude.

Chapitre 2

Présentation des données météorologiques et climatiques de l'étude

Une majeure partie de notre travail au préalable a consisté à effectuer une revue des données disponibles en Open Data les plus pertinentes pour notre problème. Nous étions à la recherche, d'une part de données météorologiques historiques pertinentes pour calibrer nos modèles et d'autre part, de données climatologiques futures pour pouvoir effectuer nos projections.

Nous avons au fil de nos recherches, analysé, rejeté ou conservé différentes sources de données météorologiques et climatologiques en fonction de plusieurs critères. Nous présenterons ici le fil conducteur et la démarche qui nous a permise de sélectionner les différentes sources de données retenues. Il est important de noter que cette démarche serait réplicable sur d'autres périls climatiques et sur d'autres études climatiques plus généralement.

Cette partie présente ainsi de manière succincte les données externes de l'étude. Les données internes relatives au portefeuille assurantiel seront présentées et analysées dans le chapitre suivant.

2.1 Les données météorologiques historiques

2.1.1 Les données de Météo-France

Météo-France est l'institution climatologique de référence en France et propose des données quotidiennes de l'ensemble des paramètres météorologiques pour toutes les stations françaises (554 stations). Dans l'idéal, nous aurions aimé avoir accès à

ces données puisque ce sont les meilleures concernant les relevés météorologiques historiques, néanmoins 200 000€ sont nécessaires pour avoir un accès illimité et annuel à l'ensemble des stations, c'est la raison pour laquelle nous avons écarté cette source de données.

Météo-France met néanmoins à disposition gratuitement des données d'observations issues des messages internationaux d'observation en surface (SYNOP) en lien avec l'Organisation Météorologique Mondiale (OMM). Des paramètres atmosphériques sont fournis à l'instar de la température, de l'humidité, des précipitations, de la force et direction du vent ainsi que certains paramètres observés comme la description des nuages ou de la visibilité. L'ensemble de ces paramètres sont délivrés par 42 stations en France métropolitaine avec un pas de temps de 3 heures et un historique datant de 1996.

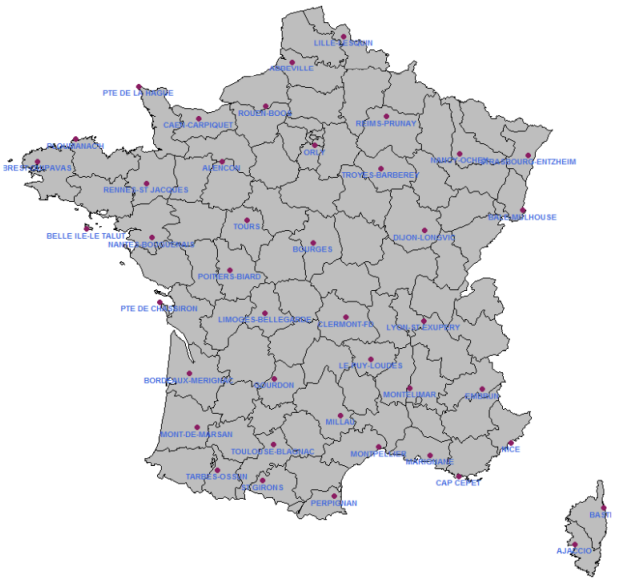

FIGURE 2.1 – Répartition des stations en libre accès de Météo-France

Le défaut majeur de cette source réside dans la couverture partielle du territoire par ces stations. En effet, la maille géographique est trop grossière et nous voyons que plusieurs départements sont dénués de stations. Notre étude étant faite à la maille canton, il nous est nécessaire d'obtenir des informations à une maille plus fine afin de minimiser les écarts entre la réalité historique et nos observations. Nous avons ainsi exclu cette source pour notre étude.

2.1.2 Les données de réanalyse ERA 5

A l'heure actuelle nous disposons d'une infrastructure météorologique assez complète qui permet de couvrir la quasi-totalité de la surface du globe. Néanmoins, les observations des stations météorologiques ne sont pas forcément réparties de manière uniforme, certains pays disposent de relevés à une maille géographique beaucoup plus fine. La réanalyse climatique permet de solutionner ce problème. Les données de réanalyse climatique sont générées en combinant les observations passées réelles issues de stations météorologiques avec les sorties de modèles climatiques créés à partir des lois physiques, pour avoir une vision complète et cohérente du climat passé.

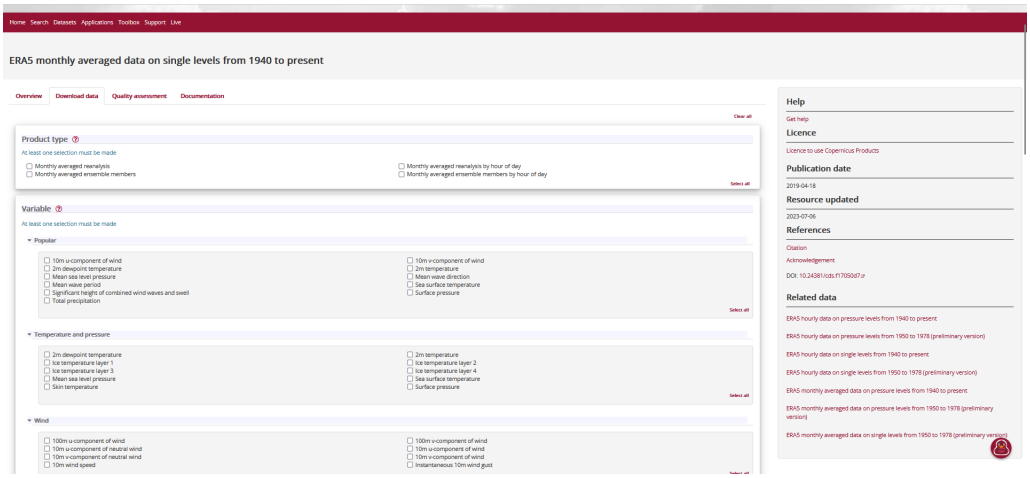

FIGURE 2.2 – Portail d'accès aux données du ERA5 (Source : Copernicus)

Pour simplifier notre explication, nous pouvons imaginer la Terre comme un gigantesque puzzle. Les pièces que nous avons correspondent aux observations météorologiques issues des stations météorologiques, des ballons-sonde, des avions, des bateaux ou des satellites. Nous pouvons alors visualiser la réanalyse climatique comme une machine intelligente qui a la faculté de combiner et traiter l'ensemble de ces informations pour déduire les pièces manquantes du puzzle à partir des lois physiques qui régissent les variations atmosphériques. Nous obtenons à la suite du processus des données sur de nombreux paramètres atmosphériques, de surface terrestre et d'état de la mer au plus proche de la réalité.

Dans notre cas, puisque nous n'avons pas eu accès aux données météorologiques de Météo France, nous avons décidé de travailler avec la réanalyse ERA 5 qui correspond à la dernière réanalyse climatique produite par le Centre Européen pour les Prévisions Météorologiques à Moyen Terme (CEPMMT). Ces données

disponibles sur le site du Copernicus (programme de l'Union européenne pour la mise à disposition de données de qualité) nous permettent d'avoir un accès gratuit et illimité à des observations météorologiques historiques cohérentes.

Plus précisément, les données ERA5 sont disponibles à des résolutions horaires et couvrent l'ensemble de la planète avec une grille régulière de 0,25 degré en latitude et en longitude (un point tous les 31 kilomètres environ). Ces données sont réparties en plusieurs paramètres météorologiques tels que la température de l'air, la pression atmosphérique, la vitesse et la direction du vent, l'humidité relative, les précipitations, l'évapotranspiration, le rayonnement solaire, la couverture nuageuse, entre autres.

Chaque paramètre est fourni à différentes altitudes verticales, généralement de la surface jusqu'à la limite de la troposphère, avec des niveaux prédéfinis. Les données sont interpolées à partir des observations et des sorties des modèles pour chaque grille et pour chaque intervalle horaire. L'ensemble de données est ensuite traité pour assurer la cohérence et la continuité des mesures dans le temps et dans l'espace.

ERA5 est considéré comme l'un des ensembles de données climatiques les plus complets et les plus fiables disponibles actuellement. Il est utilisé dans une grande variété de domaines de recherche, tels que la modélisation climatique, la météorologie, l'évaluation des ressources énergétiques renouvelables, la recherche sur le changement climatique, la gestion des ressources en eau, la prévision des conditions météorologiques extrêmes, et bien d'autres.

Dans notre cas, nous avons choisi cette source de données pour nos données météorologiques historiques puisqu'elle rassemblait l'ensemble des variables météorologiques pertinentes pour modéliser les tempêtes et qu'elle était gratuite.

Nous présentons dans la prochaine section, la façon dont nous avons sélectionné nos données météorologiques de projection.

2.2 Les données météorologiques projetées

2.2.1 Le DRIAS

Notre première source de modèles climatiques a été celle du DRIAS, un fournisseur de projections climatiques pour le territoire français financé par le Ministère de la Transition Ecologique, soutenu par Météo-France et par les principales insitutions climatologiques françaises (l'Institut Pierre-Simon Laplace (IPSL), le Centre Na-

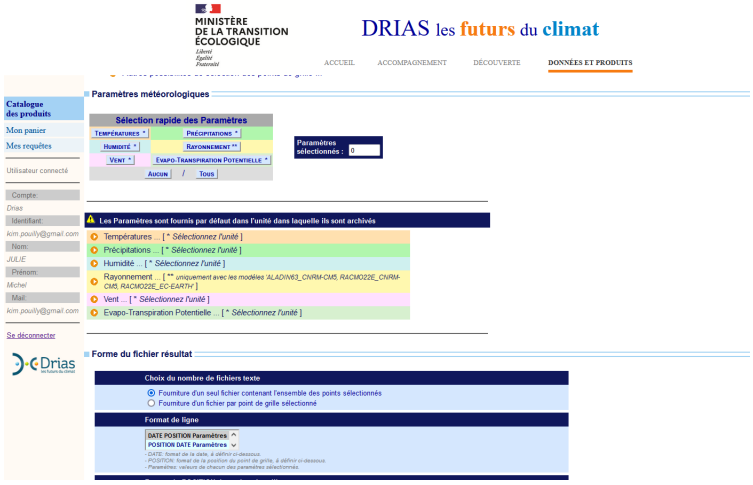

FIGURE 2.3 – Portail d'accès aux données du DRIAS (Source : DRIAS)

tional de Recherches Météorologiques (CNRM) et le Centre Européen de Recherche et de Formation Avancée en Calcul Scientifique (CERFACS)).

Par l'intermédiaire du portail mis à disposition il nous était possible de télécharger des projections de variables climatiques pour 3 scénarios de RCP différents (RCP 2.6, RCP 4.5, RCP 8.5) avec au sein de ces scénarios la possibilité de sélectionner différents modèles. Les variables climatiques disponibles sont les suivantes :

- Température minimale journalière
- Température maximale journalière
- Température moyenne journalière
- Précipitations totales
- Chute de neige à grande échelle
- Humidité
- Rayonnement visible incident à la surface
- Rayonnement infra-rouge incident à la surface
- Vitesse moyenne du vent
- Evapotranspiration potentielle

Ce dataset disponible en Open Data s'appuie en effet sur 12 couples GCM/RCM, produits dans le cadre de l'exercice international CMIP5 (Coupled Model Intercomparison Project Phase 5), qui ont servi à la rédaction du cinquième rapport du GIEC. Ces 12 couples ont été sélectionnés à partir d'une centaine de modèles effectués dans le cadre du projet Euro-Cordex (présenté par la suite). Plusieurs critères ont permis de sélectionner ces 12 couples parmi les autres modèles. Ils

ont privilégié des RCM issus des centres français de modélisation climatique, ils ont rejeté les couples GCM/RCM contenant une erreur connue et ils ont essayé d'optimiser la dispersion du changement climatique simulé sur la France par les couples sélectionnés. Le DRIAS propose ainsi 30 simulations du climat futur (12 projections RCP 8.5, dix projections RCP 4.5 et huit projections RCP 2.6) qui s'appuient sur les 12 couples GCM/RCM.

GCM	RCM	HISTO	RCP2.6	RCP4.5	RCP8.5
CNRM-CM5	Aladin63 V2	▬	▬	▬	▬
CNRM-CM5	Racmo22E v2	▬	▬	▬	▬
IPSL-CM5A-MR	WRF381P	▬		▬	▬
IPSL-CM5A-MR	RCA4	▬		▬	▬
HadGEM2-ES	RegCM4-6	▬	▬		▬
HadGEM2-ES	CCLM4-8-17	▬		▬	▬
EC-EARTH	Racmo22E v2	▬	▬	▬	▬
EC-EARTH	RCA4	▬	▬	▬	▬
MPI-ESM-LR	CCLM4-8-17	▬	▬	▬	▬
MPI-ESM-LR	REMO*	▬	▬	▬	▬
NorESM1-M	HIRHAM5 v3	▬		▬	▬
NorESM1-M	REMO**	▬	▬		▬

** REMO 2009 ; ** REMO 2015*

FIGURE 2.4 – Les 30 simulations du climat futur du jeu DRIAS-2020 basées sur les 12 couples GCM/RCM sélectionnés (Source : DRIAS)

L'avantage majeur de cette base réside dans la maille géographique très fine à laquelle les données sont disponibles. En effet, les modèles du DRIAS ont une résolution de 8km ce qui permet d'avoir en sortie 8981 points qui couvrent la France métropolitaine. Néanmoins, dans le cadre de notre étude sur les tempêtes, des projections de rafales étaient nécessaires or le DRIAS ne nous fournissait que la projection des vitesses moyennes de vent, c'est pourquoi nous avons décidé de nous tourner vers d'autres fournisseurs de données par la suite.

2.2.2 Le NGFS

Le Network for Greening the Financial System (NGFS) ou Réseau des banques centrales et des superviseurs pour rendre le système financier plus écologique, est un groupe de banques centrales et de superviseurs, créé en décembre 2017 lors du "One Planet Summit" de Paris. Ce groupe s'est engagé à promouvoir la stabilité financière dans le contexte du changement climatique et de la durabilité environnementale. L'un des principaux objectifs du NGFS est de fournir des informations aux acteurs financiers et aux décideurs politiques pour les aider à prendre des décisions éclairées en matière de climat.

Dans cette optique, le NGFS collabore avec des experts scientifiques et des ins-

titutions spécialisées pour collecter et analyser des données climatiques projetées (comme l'ISIMIP que nous présenterons dans le prochain paragraphe). Ces données incluent des projections sur les émissions de gaz à effet de serre, les températures mondiales, les événements météorologiques extrêmes et les impacts sur les écosystèmes. Ces projections aident à évaluer les risques climatiques futurs et à orienter les politiques et les investissements vers des actions de mitigation et d'adaptation. Le NGFS travaille également à développer des scénarios de transition climatique, qui permettent d'évaluer les conséquences de différentes trajectoires d'émission et d'orienter les politiques en conséquence. Ces scénarios sont utiles pour estimer les besoins en investissement dans les technologies bas-carbone, l'adaptation aux impacts du changement climatique et la réduction de la dépendance aux énergies fossiles.

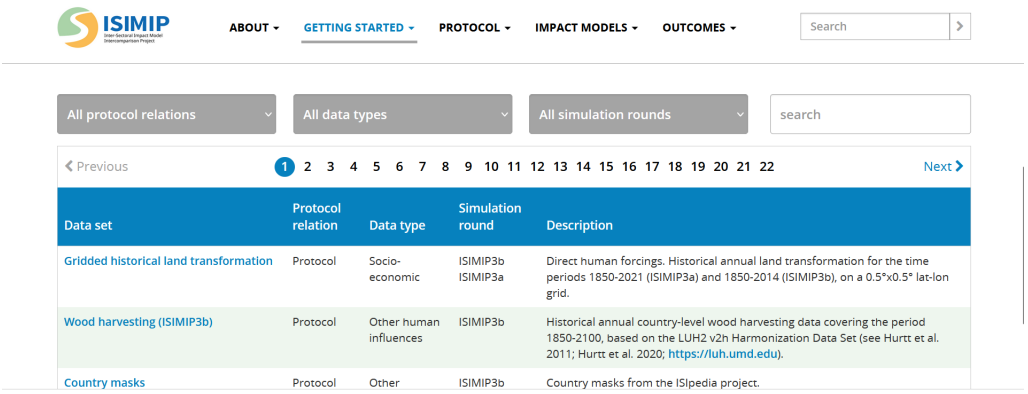

FIGURE 2.5 – Portail d'accès aux données du ISIMIP (Source : ISIMIP)

L'ISIMIP (Inter-Sectoral Impact Model Intercomparison Project) est une initiative de recherche collaborative qui vise à améliorer la compréhension des impacts du changement climatique sur différents secteurs clés. En réunissant des modèles climatiques, des modèles d'impact et des équipes de recherche du monde entier, ISIMIP facilite la comparaison des projections et des scénarios à l'échelle mondiale et régionale. L'objectif principal d'ISIMIP est de fournir des informations scientifiques précieuses pour les décideurs, les gestionnaires des ressources et les praticiens afin de les aider à prendre des décisions éclairées en matière d'adaptation au changement climatique, d'où la collaboration avec le NGFS. L'initiative se concentre sur une large gamme de secteurs, tels que l'agriculture, l'eau, l'énergie, la santé, les écosystèmes, les infrastructures et l'économie.

Cette source de données en lien avec le NGFS ne présentait pas l'ensemble des variables climatiques que nous souhaitions intégrer dans l'étude (notons aussi que cette source est plus appropriée pour des études sur les risques de transition où

des données en lien direct avec le carbone sont nécessaires). Une autre source de données présentée dans la partie suivante, à savoir l'EURO-CORDEX, a permis de combler cette lacune.

2.2.3 EURO-CORDEX

EURO-CORDEX est un projet international initié par la communauté scientifique pour améliorer les projections climatiques régionales en Europe. Il est né de la collaboration entre de nombreux instituts de recherche, universités et centres de modélisation climatique à travers l'Europe.

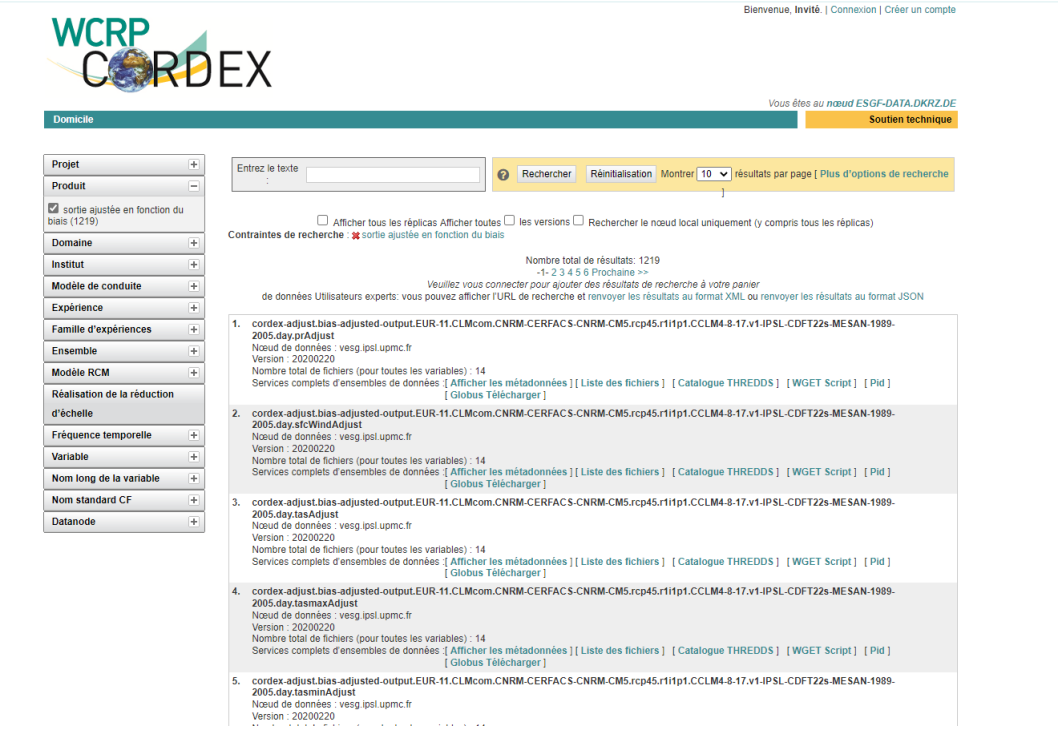

FIGURE 2.6 – Portail d'accès aux données d'EURO-CORDEX (Source : EURO-CORDEX)

L'origine d'EURO-CORDEX remonte au projet CORDEX (Coordinated Regional Downscaling Experiment) lancé par le Programme mondial de recherche sur le climat (PMRC) et l'Organisation météorologique mondiale (OMM). CORDEX visait à améliorer les projections climatiques régionales à l'échelle mondiale en encourageant la coordination internationale et la collaboration entre les communautés scientifiques régionales.

Dans le cadre de CORDEX, la branche régionale EURO-CORDEX a été créée spé-

cifiquement pour la région européenne. L'initiative EURO-CORDEX a été lancée en 2011 et a bénéficié d'un soutien important de diverses organisations scientifiques et environnementales.

EURO-CORDEX est soutenu par le Centre européen pour les prévisions météorologiques à moyen terme (CEPMMT), le Programme européen pour le climat (PEC) et d'autres organismes européens de recherche climatique. Il bénéficie également du soutien de nombreux partenaires institutionnels, notamment des instituts de recherche nationaux, des universités, des centres de modélisation et des services météorologiques nationaux.

Grâce à cette collaboration et à l'engagement des chercheurs à travers l'Europe, EURO-CORDEX a pu devenir **une source importante de données climatiques régionales de haute résolution** pour la communauté scientifique et les décideurs. Les données fournies par EURO-CORDEX sont utilisées dans de nombreuses études, rapports et initiatives pour évaluer les impacts du changement climatique en Europe et orienter les politiques d'adaptation.

Il utilise en l'occurrence des modèles climatiques régionaux (RCM) pour produire des projections climatiques à une résolution spatiale plus fine que les modèles climatiques globaux ce qui permet de capturer des caractéristiques régionales spécifiques, telles que les effets topographiques, les conditions côtières et les différences climatiques à petite échelle. Concrètement, les simulations EURO-CORDEX se concentrent sur des **mailles d'environ 12 km** (0,11 degré). Des simulations auxiliaires avec la résolution CORDEX standard d'environ 50 km (0,44 degré) sont également effectuées. Cela signifie que chaque point de la région étudiée représente une zone géographique de 10 km à 50 km de côté. Quant à la résolution verticale des modèles EURO-CORDEX qui est définie par le nombre de niveaux verticaux utilisés pour représenter l'atmosphère, elle varie généralement de **20 à 50 niveaux**, permettant une meilleure représentation des processus verticaux de l'atmosphère.

Par ailleurs, EURO-CORDEX intègre différents scénarios d'émissions de gaz à effet de serre afin de représenter différentes trajectoires d'évolution du climat. Ces scénarios sont alignés sur ceux utilisés par le GIEC et permettent d'évaluer les impacts climatiques en fonction des niveaux futurs d'émissions.

L'ensemble de ces facteurs combinés au fait que l'EURO-CORDEX **fournit l'ensemble des variables climatiques projetées disponible actuellement**, nous a poussé à prendre cette source de données pour nos projections.

2.3 Le format NetCDF

Dans ce paragraphe, nous effectuons une description rapide du **format des fichiers météorologiques**, à savoir le format NetCDF, format atypique pour un actuaire mais classique pour un climatologue.

L'ensemble des données météorologiques que nous avons téléchargé se trouvait en effet sous le format **NetCDF** (Network Common Data Form) (suffixe .nc), un format fortement utilisé en climatologie, en météorologie, en géologie et en océanographie. Il existe depuis 1988 et a été mis en place par Unidata, un organisme public américain.

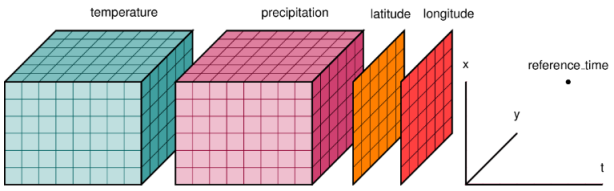

FIGURE 2.7 – Les tableaux multi-dimensionnels en NetCDF (Source : Zarr OGC 2020)

Les fichiers NetCDF sont basés sur une structure de données multidimensionnelle. Ils peuvent contenir des dimensions telles que le temps, la latitude, la longitude et les niveaux verticaux, permettant ainsi de représenter simultanément des données spatiales et temporelles.

Ces fichiers peuvent contenir différentes variables, chacune étant associée à des dimensions spécifiques. Par ailleurs, ces variables peuvent être de différents types, tels que des nombres réels, des entiers ou des caractères.

Les fichiers NetCDF peuvent également contenir des attributs, qui sont des métadonnées associées aux variables. On parle de fichiers "auto-descriptifs". Les attributs fournissent des informations supplémentaires sur les données, telles que les unités de mesure, les échelles, les commentaires, les conditions d'acquisition, les méthodes de traitement, le modèle climatique utilisé, les auteurs, les références bibliographiques, l'expérience spécifique, le système de coordonnée utilisé ainsi que le domaine géographique de la modélisation. L'ensemble de ces éléments aident à documenter et à contextualiser les données du fichier NetCDF.

De plus, NetCDF est un format binaire permettant de réduire leur taille de stockage. Un tel format permet aussi de diminuer le temps d'exécution pour la lecture ou l'écriture du fichier. Ces fichiers peuvent en outre être compressés, la compres-

sion étant applicable sur les variables individuelles ou sur l'ensemble du fichier, selon les besoins de l'utilisateur.

Concernant l'accès et l'utilisation de ces données, il est important de préciser que ces fichiers sont indépendants de la plate-forme qui les utilise et peuvent être lus et écrits par différents logiciels et langages de programmation, permettant une interopérabilité élevée entre les outils et les environnements utilisés pour l'analyse des données. Ces fichiers peuvent être ouverts par l'intermédiaire de librairies comme *xarray* sur Python ou *ncdf4* sur R par exemple.

Dans notre cas, nous avons principalement manipuler ces fichiers sur nos logiciels classiques de statistiques à savoir R et Python. Les fichiers NetCDF permettant un accès flexible aux données grâce à l'extraction sélective de variables, de dimensions ou de sous-ensembles de données spécifiques, nous avons tout d'abord conçu des algorithmes pour pouvoir reconstruire nos données météorologiques dans un tableau en 2 dimensions. Ces données étant initialement en 3 dimensions avec la longitude, la latitude et le temps.

Synthèse

Nous présentons pour résumer notre démarche, deux tableaux récapitulant les critères utilisés pour le choix des sources de nos données historiques et projetées :

- **Exhaustivité des variables météos :** indique si cette source contient l'ensemble des variables relatives aux tempêtes.
- **Type de données** : indique si les données proviennent de sources réelles, de réanalyse ou de simulation.
- **Temporalité :** indique si la source de données est capable de couvrir la période historique et projetée de la modélisation.
- **Résolution géographique :** indique si la maille géographique est assez fine afin d'avoir suffisamment de points couvrant la France.
- **Gratuité**

	Exhaustivité variables météos	Type de données	Temporalité	Résolution géographique	Gratuité
Météo-France	✓	réelles	✓	✓	✗
SYNOP	✓	réelles	✓	✗	✓
COPERNICUS (ERA5)	✓	réanalysées	✓	✓	✓

FIGURE 2.8 – Tableau récapitulatif des critères utilisés pour le choix de notre source de **données historiques**

	Exhaustivité variables météos	Type de données	Temporalité	Résolution géographique	Gratuité
DRIAS	✗	simulées	✓	✓	✓
NGFS	✗	simulées	✓	✓	✓
EURO-CORDEX	✓	simulées	✓	✓	✓

FIGURE 2.9 – Tableau récapitulatif des critères utilisés pour le choix de notre source de **données de projection**

Chapitre 3

Analyse des données

L'objectif de cette partie est de présenter le portefeuille assurantiel ainsi que les analyses et retraitements effectués qui ont permis de construire la base de données finale utilisée pour notre modélisation.

3.1 Construction de la base de modélisation

3.1.1 Présentation de la base initiale

La modélisation de la tempête nécessite au préalable de disposer d'un historique de sinistres causés par le péril. Nous avons eu à disposition une **base de données multirisque habitation (MRH) anonymisée** contenant **48 979 sinistres** relatifs à des sinistres tempêtes dans l'ensemble de la France métropolitaine exceptée la Corse, répartis de manière homogène sur le territoire. La base a 36 colonnes, chacune d'entre elle se référant à une variable précise et 7 087 300 lignes. La liste exhaustive des variables est présentée plus bas. A noter que toutes ces variables ne sont pas forcément en lien avec la survenance du risque tempête. C'est pour cette raison qu'un premier travail de tri s'impose. L'idée du tri est de retenir toute variable pouvant influer sur la survenance du risque tempête.

Nous énumérons ici les variables présentes dans notre base initiale :

1. *Contrat* : identifiant du contrat (variable caractère)
2. *Year* : année de couverture allant de **2016 à 2020** (variable numérique)
3. *Expo* : fréquence de présence dans le portefeuille sur une base annuelle (variable numérique)
4. *TypeHabitation* : le type de logement (variable catégorielle)
5. *CodePostal* : le code postal (variable numérique)
6. *CodeIris* : le code iris (variable numérique)
7. *CodeInsee* : le code INSEE (variable numérique)

8. *Charge* : la charge de sinistre observée (variable numérique)
9. *MoisDeSurvenance* : le mois de survenance du sinistre (variable numérique)
10. *ValMaison* : Valeur au mètre carré du bien immobilier résidentiel correspondant au prix moyen de l'immobilier normalisé par la surface habitable du bâtiment (variable numérique)
11. *ApeCode* : le code APE permettant d'identifier le secteur d'activité du travailleur indépendant ou de la société (variable caractère)
12. *ParcSurf* : la surface de la parcelle associée au bâtiment correspondant à la somme de toutes les surfaces des parcelles associées à un bâtiment (variable numérique)
13. *ConstPeriod* : Période de construction (variable numérique)
14. *EmpreinteSol* : l'empreinte au sol du bien correspondant à la surface projetée des limites extérieures du bâtiment (variable numérique)
15. *NbEtage* : le nombre d'étages du bâtiment. Cet attribut définit la quantité d'étages d'un bâtiment suivant la convention française où 0 signifie un bâtiment avec rez-de-chaussée et 1 signifie un bâtiment avec rez-de-chaussée plus premier étage (variable numérique)
16. *MaterToit* : le matériau de toit du bâtiment (variable catégorielle)
17. *FenetreToitPres* : indique la présence de fenêtre sur le toit (variable booléenne)
18. *TypeToit* : le type de toit du bâtiment (variable catégorielle)
19. *PropOcc* : Probabilité que le bâtiment soit occupé par leur propriétaire (variable numérique)
20. *ResPrinc* : Probabilité que le logement soit occupé en tant que résidence principale (variable numérique)
21. *Altitude* : l'altitude du plancher du bâtiment par rapport à la surface de la mer (variable numérique)
22. *Bati50m* : Nombre de bâtiment dans un rayon de 50 mètres (variable numérique)
23. *Dependance* : Présence d'une surface annexe (piscine ou autre) (variable booléenne)
24. *EauAltiDiff* : Différence d'altitude du point d'eau le plus proche par rapport à la surface de la mer (variable numérique)
25. *EauDist* : Distance au point d'eau le plus proche (variable numérique)
26. *HydrauliqueDist* : Distance à des stations hydrauliques (variable numérique)
27. *Mitoyennete* : Présence d'un mur mitoyen (variable booléenne)
28. *MurMitoyLong* : Longueur du mur mitoyen (variable numérique)

29. *SurfaceVeget* : Surface végétalisée (variable numérique)
30. *TypeSol* : Type de sol (variable catégorielle)
31. *ZoneUrbain* : Type de zone urbaine (variable catégorielle)
32. *MvtSolCouleeBoue* : Nombre de survenance mouvement de sol coulée de boue (variable numérique)
33. *MvtSolEboul* : Nombre de survenance d'éboulements (variable numérique)
34. *MvtSolEffondr* : Nombre de survenance d'effondrements (variable numérique)
35. *MvtSolErosion* : Nombre de survenance d'érosions des sols (variable numérique)
36. *MvtSolGliss* : Nombre de survenance de glissements de terrain (variable numérique)

Nous remarquons que **nous n'avons pas la date de sinistre mais seulement le mois** pour des raisons de confidentialité. Notre modélisation prendra en compte cet aspect important. Par ailleurs, nous avons la présence de variables assez atypiques qui ne sont pas incluses en règle générale dans les portefeuilles MRH. En effet, les variables *ValMaison*, *ApeCode*, *ParcSurf*, *EmpreinteSol*, *NbEtage*, *MaterToit*, *FenetreToitPres*, *TypeToit* et *Altitude* sont des variables qui proviennent d'un travail d'agrégation de multiples sources de données en Open Data réalisé en amont par une autre équipe.

3.1.2 Prétraitement des données et nettoyage de la base

Nous commençons par supprimer d'une part l'ensemble des variables qui de manière évidente ne sont pas liées à la sinistralité tempête et d'autre part les variables qui indiquent une maille géographique qui ne sera pas exploitée dans notre modélisation pour des raisons que nous développerons ultérieurement. Les variables *Contrat*, *CodePostal*, *CodeIris*, *PropOcc*, *ResPrinc*, *EauAltiDiff*, *EauDist*, *HydrauliqueDist*, *Mitoyennete*, *MurMitoyLong*, *TypeSol*, *MvtSolCouleeBoue*, *MvtSolEboul*, *MvtSolEffondr*, *MvtSolErosion* et *MvtSolGliss* sont alors supprimées.

Puis nous nous occupons des lignes en double relatives à des avenants au contrat ou à une erreur inhérente à la base. Nous supprimons ensuite les variables ayant un nombre de valeurs manquantes trop important ou présentant un niveau de confiance faible sur la validité de l'information qu'elles contiennent. Ce niveau de confiance nous a été indiqué par l'équipe qui nous a fourni la base de données. Nous supprimons ainsi les variables *ApeCode*, *ConstPeriod*, *FenetreToitPres*, *Bati50m*, *Dependance*, *SurfaceVeget* et *ZoneUrbain*.

Après ces deux opérations, nous avions la liste de variables suivantes : *Year*, *Expo*,

TypeHabitation, *CodeInsee*, *Charge*, *MoisDeSurvenance*, *ValMaison*, *ParcSurf*, *EmpreinteSol*, *NbEtage*, *MaterToit*, *TypeToit* et *Altitude*.

3.1.3 Jointure et méthode d'agrégation pour obtenir notre base finale

Pour compléter cette base initiale, nous avons dans un premier temps décidé de joindre les informations de deux bases annexes créées ou récupérées en amont :

- une base **DictionnaireCommuneCanton** faisant la correspondance entre le code INSEE de la commune et le code du canton correspondant sachant qu'en France nous avons environ 35 000 communes et 2 200 cantons. Cette base a été construite à partir des données que l'on retrouve sur l'INSEE. La jointure de cette base avec notre portefeuille a été immédiate puisque nous avions aussi le code INSEE présent dans la base initiale.
- une base intitulée **DataInsee** récupérée sur le site de l'INSEE nous permettant d'obtenir la variable *densité* par code commune. La jointure a là aussi été directe étant donné la présence du code commune dans les deux bases.

La construction de la base de données passe avant tout par le choix du pas de modélisation et de la granularité souhaitée. Dans notre étude, nous avons décidé d'effectuer une modélisation à la maille **Canton x Mois x Année** (les cantons étant des regroupements de communes). A noter que nous nous **limitons à la France métropolitaine**.

Ce choix a été appuyé par plusieurs facteurs. Premièrement, nous n'avions que le mois de survenance du sinistre et non pas la date précise, nous nous sommes donc adaptés à cette contrainte. Deuxièmement, nous avons choisi une maille Canton, qui est une maille géographique entre la commune et le Département car nous avions 2360 stations météorologiques pour nos données météorologiques historiques et 1868 cantons présents dans notre portefeuille. Ainsi, nous pouvions aisément faire un rapprochement entre nos stations et nos différents cantons (nous expliciterons cette méthode de rapprochement par la suite). Par ailleurs, dans une logique de projection, il est nécessaire de choisir une maille géographique cohérente i.e. ni trop fine ni trop large et nous pensons que le canton est le meilleur compromis.

Pour agencer notre base initiale suivant notre pas de modélisation, nous avons alors effectué les étapes suivantes :

1. Nous avons listé l'ensemble des cantons présents dans notre portefeuille.
2. Nous avons dupliqué chacun des cantons par 12 (le nombre de mois) et par 5 (notre nombre d'années historique).
3. Nous avons ensuite complété chaque ligne avec les variables présélectionnées en dissociant les modalités des variables catégorielles ou booléennes et en agré-

geant les variables numériques sur chaque ligne ayant subi ce regroupement **Canton x Mois x Année x Modalités des Variables catégorielles**. Par exemple, si l'on considère 2 cantons, 3 mois et 1 année, auxquelles on associe les variables assureurs "FenetreToitPres" et "ValMaison", notre base aura la forme suivante (à noter ici que la variable "ValMaison" a été retraitée puisque nous avions initialement la valeur moyenne au mètre carré des biens, nous avons alors multiplié cette valeur par la variable "EmpreinteSol" pour obtenir la valeur totale du bâtiment) :

FIGURE 3.1 – Schéma de la base agrégée suivant le pas de modélisation

Dans notre cas, nous avons agrégé notre base par canton, année, mois, type d'habitation, matériaux du toit et type de toit et nous avons selon cette agrégation sommer l'exposition et effectuer la moyenne des variables en lien avec l'altitude, la surface de la parcelle, la valeur du bâtiment, le nombre d'étages et l'empreinte au sol. Nous modélisons ainsi des **individus "canton"** sur un certain mois et une certaine année avec des caractéristiques relatives au type d'habitation, aux matériaux du toit et au type de toit. Notre base est désormais de 1 867 488 lignes.

Un inconvénient sur ce procédé d'agrégation des données réside dans la perte d'information concernant les caractéristiques des bâtiments à une maille ligne à ligne. En effet en sortie de cette agrégation, nous avons en réalité un dénombrement par Canton x Mois x Année des bâtiments ayant les caractéristiques des modalités des variables catégorielles ayant été utilisées dans l'agrégation. Néanmoins, l'avantage majeur qui nous a poussé à effectuer ce choix d'agrégation concerne la jointure avec les variables climatiques puisque grâce à cette agrégation, nous avons l'ensemble

des cantons sur l'ensemble des mois et des années qui ont subi ou non des sinistres auxquels on va pouvoir rattaché l'information climatique correspondante. Ce type de jointure permet de mettre l'accent sur les variables climatiques.

D'ailleurs, étant donné que les seules variables du portefeuille client n'expliquent pas la survenance de la tempête, il était alors indispensable d'étoffer nos informations à l'aide de variables dites météorologiques pour comprendre la sinistralité sous-jacente. Celles-ci sont téléchargeables au format de données NetCDF sur le site du programme d'observation de la Terre de l'Union européenne (Copernicus). Elles indiquent la valeur journalière (pour chaque heure) des variables météorologiques (vitesse du vent, humidité et autres) à des stations (points) selon des grilles régulières.

Pour construire cette base climatique historique, nous avons sélectionné et téléchargé une liste de paramètres météorologiques liés à la tempête d'après notre compréhension du phénomène. Tout d'abord des **paramètres relatifs à la vitesse du vent**, le vent étant la manifestation principal lors d'une tempête :

1. ***vent100mU*** : Ce paramètre est la vitesse horizontale de l'air se déplaçant vers l'est, à une hauteur de 100 mètres au-dessus de la surface de la Terre, mesuré en mètres par seconde (ms^{-1}).
2. ***vent100mV*** : Ce paramètre est la vitesse horizontale de l'air se déplaçant vers le nord, à une hauteur de 100 mètres au-dessus de la surface de la Terre (ms^{-1}).
3. ***vent10mU*** : Ce paramètre est la vitesse horizontale de l'air se déplaçant vers l'est, à une hauteur de 10 mètres au-dessus de la surface de la Terre (ms^{-1}).
4. ***vent10mV*** : Ce paramètre est la vitesse horizontale de l'air se déplaçant vers le nord, à une hauteur de 10 mètres au-dessus de la surface de la Terre (ms^{-1}).
5. ***vent10mNeutreU*** : Ce paramètre est la composante est du « vent neutre », à une hauteur de 10 mètres au-dessus de la surface de la Terre. Le vent neutre est calculé à partir de la contrainte de surface et de la longueur de rugosité correspondante en supposant que l'air est stratifié de manière neutre. Le vent neutre est plus lent que le vent réel dans des conditions stables, et plus rapide dans des conditions instables (ms^{-1}).
6. ***vent10mNeutreV*** : Ce paramètre est la composante nord du « vent neutre » (ms^{-1}).
7. ***rafale*** : Vent maximal de 3 secondes à 10 mètres au-dessus de la surface de la Terre (ms^{-1}).

Les composantes U et V indiquées ci-dessus doivent être combinées pour calculer la vitesse à la hauteur spécifique par le biais de cette formule :

$$vitesse = \sqrt{ventU^2 + ventV^2}$$

Une conversion des mètres par seconde en kilomètres par heure a aussi été appliquée en multipliant les vitesses par le facteur 3,6.

Nous avons par la même occasion téléchargé des **paramètres relatifs à la température** puisque l'on sait qu'une tempête est générée lorsque deux masses d'air de **températures** différentes se rencontrent d'où l'introduction de ces paramètres pour éventuellement déceler des variations lors de la survenance d'une tempête :

1. ***TempRosee2m*** : Ce paramètre est la température à laquelle l'air, à 2 mètres au-dessus de la surface de la Terre, devrait être refroidi pour que la saturation se produise. C'est une **mesure indirecte de l'humidité de l'air**. Ce paramètre a des unités en kelvin (K).
2. ***Temp2m*** : Ce paramètre est la température de l'air à 2m au-dessus de la surface de la terre (K).
3. ***Flux thermique sensible à la surface moyenne*** : Ce paramètre est le transfert de chaleur entre la surface de la Terre et l'atmosphère par les effets du mouvement turbulent de l'air (mais à l'exclusion de tout transfert de chaleur résultant de la condensation ou de l'évaporation). L'ampleur du flux de chaleur sensible est régie par la différence de température entre la surface et l'atmosphère, la vitesse du vent et la rugosité de la surface. Par exemple, l'air froid recouvrant une surface chaude produirait un flux de chaleur sensible important de la terre (ou de l'océan) dans l'atmosphère. La convention du CEPMMT pour les flux verticaux est positive à la baisse. Ce paramètre est mesuré en watt par mètre carré ($W.m^{-2}$).

Les températures mesurées en kelvin ont été converties en degrés Celsius (°C) en soustrayant 273,15.

Nous avons ensuite téléchargé des **paramètres relatifs à la pression**, la tempête étant un phénomène lié à un état dépressionaire :

1. ***PressionLevelSea*** : Ce paramètre est la pression (force par unité de surface) de l'atmosphère à la surface de la Terre, ajustée à la hauteur du niveau moyen de la mer. C'est une mesure du poids que tout l'air d'une colonne verticalement au-dessus d'un point de la surface de la Terre aurait, si le point était situé au niveau moyen de la mer. Il est calculé sur toutes les surfaces (terre, mer et eaux intérieures). Les cartes de la pression moyenne au niveau de la mer sont utilisées pour identifier l'emplacement des systèmes météorologiques à basse et haute pression, souvent appelés cyclones et anticyclones, d'où l'intérêt d'inclure cette variable dans le cadre de notre étude. Les unités de ce paramètre sont des pascals (Pa).

2. ***PressionSuperficielle*** : Ce paramètre est la pression de l'atmosphère à la surface de la terre, de la mer et des eaux intérieures. La pression superficielle est souvent utilisée en combinaison avec la température pour calculer la densité de l'air. La forte variation de pression avec l'altitude rend difficile la vision des systèmes météorologiques de basse et haute pression au-dessus des zones montagneuses, de sorte que la pression moyenne au niveau de la mer, plutôt que la pression de surface, est normalement utilisée à cette fin. Néanmoins nous avons tout de même décidé d'inclure ce paramètre que nous pourrons comparer ultérieurement avec la *Pression moyenne au niveau de la mer* pour juger de sa pertinence. Les unités de ce paramètre sont des Pascals (Pa).

Les pressions sont généralement mesurées en hectopascal (hPa) sachant que 1 hPa vaut 100 Pa, nous avons alors converti nos variables de pression en hPa en divisant leurs valeurs par 100.

La dernière variable que nous avons décidé d'intégrer à notre base climatique est l'*énergie potentielle de convection disponible* qui est un concept utilisé en météorologie pour évaluer la quantité d'énergie disponible dans l'atmosphère pour alimenter les mouvements convectifs (processus de transfert de chaleur qui se produisent dans les fluides en raison des variations de densité causées par les différences de température). Elle est principalement utilisée pour prévoir le développement de tempêtes et d'autres phénomènes météorologiques intenses. En effet, lorsqu'une partie de l'atmosphère est caractérisée par une énergie potentielle de convection disponible élevée, cela signifie qu'il y a une grande quantité d'énergie disponible pour alimenter la convection atmosphérique, ce qui peut favoriser le développement de tempêtes, d'orages violents ou d'autres phénomènes météorologiques intenses.

Plus précisément, cette variable prend en compte la différence d'énergie potentielle entre une parcelle d'air dans l'atmosphère et une parcelle d'air à un niveau de référence, généralement près de la surface terrestre. Cette différence d'énergie potentielle est le résultat de l'effet de la gravité sur l'air et de sa position verticale par rapport à sa position de repos. Lorsqu'une parcelle d'air est soulevée depuis la surface de la Terre, elle subit une expansion adiabatique (processus thermodynamique dans lequel un gaz se dilate et se refroidit en effectuant un travail mécanique sans échange de chaleur avec son environnement) à mesure qu'elle monte dans l'atmosphère, ce qui la refroidit. Si cette parcelle d'air est plus chaude et moins dense que l'air environnant à une certaine altitude, elle continuera de monter par convection. Cette variable mesure alors l'énergie potentielle acquise par cette parcelle d'air lorsqu'elle est soulevée.

En d'autres termes, l'énergie potentielle de convection disponible indique la quantité d'énergie que la parcelle d'air a accumulée en raison de sa différence de tem-

pérature par rapport à l'environnement. Plus cette différence de température est grande, plus l'énergie potentielle de convection est élevée. L'énergie potentielle de convection disponible est exprimée en joules par kilogramme (J/kg).

Les tempêtes étant provoquées par la rencontre de deux masses d'air de températures différentes, il était alors judicieux d'intégrer ce genre de variable traduisant de manière indirecte les phénomènes physiques engendrés par ces différences de température.

Une fois l'ensemble de ces variables téléchargées, il nous a fallu les agréger entre elles. Sachant que nous disposions uniquement du mois durant lequel s'est effectué le sinistre et que la variable mère pour la sinistralité tempête est la *rafale*, nous avons décidé de nous en servir comme fer de lance pour cette agrégation.

Concrètement, nous avons isolé pour chaque mois la rafale maximale ainsi que le jour, l'heure et la station météorologique où elle s'était produite, puis nous nous sommes servis de ces indices pour agréger l'ensemble des autres paramètres météorologiques sur ce même créneau. Nous avons ensuite associé chaque station ERA 5 au canton le plus proche en minimisant la distance euclidienne entre les coordonnées de la station et les coordonnées du centroïde du canton (nous avions 2360 stations pour 1868 cantons présents dans notre portefeuille) :

$$\min D = \sqrt{(longitude_{station} - longitude_{canton})^2 + (latitude_{station} - latitude_{canton})^2}$$

Nous avions au final les 1868 cantons (associé à une station ERA 5 et aux paramètres météorologiques correspondant) multiplié par 12 le nombre de mois (et notre maille d'agrégation), multiplié par 5 le nombre d'années correspondant à l'historique de notre portefeuille.

L'agrégation, qui nous a ainsi permis d'obtenir notre base finale de modélisation, a consisté à joindre notre portefeuille agrégé par Canton x Mois x Année et notre base climatique construite sur ce même format.

3.2 Analyses statistiques préliminaires

3.2.1 Analyse des corrélations

L'idée de cette section est d'effectuer un premier écrémage des variables explicatives pour repérer celles présentant des liens notables. Il est crucial de limiter les corrélations entre nos variables puisque la préservation des variables corrélées a une incidence sur l'évaluation des coefficients dans un modèle linéaire généralisé

par exemple. En effet, la présence d'une interdépendance entre les variables indépendantes impacte l'évaluation des coefficients dans les modèles de régression. Cela peut conduire à une surestimation des coefficients en termes absolus, à une augmentation de la variance, voire même à un renversement du signe du coefficient.

Coefficients de corrélation de Pearson

Pour mesurer le lien entre deux **variables continues** X et Y, nous utilisons dans un premier temps le rhô de Pearson noté ρ_p ici :

$$\rho_p(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\sigma_X \sigma_Y}$$

avec $Cov(X,Y)$ la covariance entre X et Y.

Ce coefficient permet de détecter la présence ou l'absence d'une **relation linéaire** entre deux variables quantitatives. Il est possible de démontrer que ce coefficient varie entre -1 et +1 et son interprétation est la suivante :

- si ρ_p est proche de 0, il n'y a pas de relation linéaire entre X et Y
- si ρ_p est proche de 1, il y a une relation linéaire positive entre X et Y
- si ρ_p est proche de -1, il y a une relation linéaire négative entre X et Y

Le signe de ρ indique donc le sens de la relation tandis que la valeur absolue de ρ indique l'intensité de la relation c'est-à-dire la capacité à prédire les valeurs de Y en fonction de celles de X.

De manière générale, le coefficient de Pearson est destiné à évaluer la liaison entre deux variables, X et Y, qui suivent une distribution gaussienne et qui ne contiennent pas de valeurs aberrantes. Ces conditions ne sont en pratique quasiment jamais remplies ainsi il est nécessaire d'être prudent quant à l'utilisation de ce coefficient qui peut conduire à des conclusions incorrectes quant à la présence ou l'absence d'une relation. Il convient également de souligner que l'absence d'une relation linéaire ne sous-entend pas forcément l'absence totale de lien entre les deux variables examinées.

On remarque dans notre cas, des corrélations importantes entre nos différentes variables relatives au vent ce qui n'est pas surprenant. La variable *rafale* a une corrélation de 0,749 avec la variable *vent100m*, de 0,687 avec *vent10m* et de 0,675 avec *vent10mNeutre*. De plus, les couples de variables (*vent100m*, *vent10m*) et (*vent10m*,*vent10mNeutre*) ont des corrélations proches de 1. Pour notre étude, nous privilégierons la variable *rafale* qui est la variable principale pour les phénomènes de vents violents, nous supprimerons alors les autres variables de vent qui sont fortement corrélées à la *rafale*.

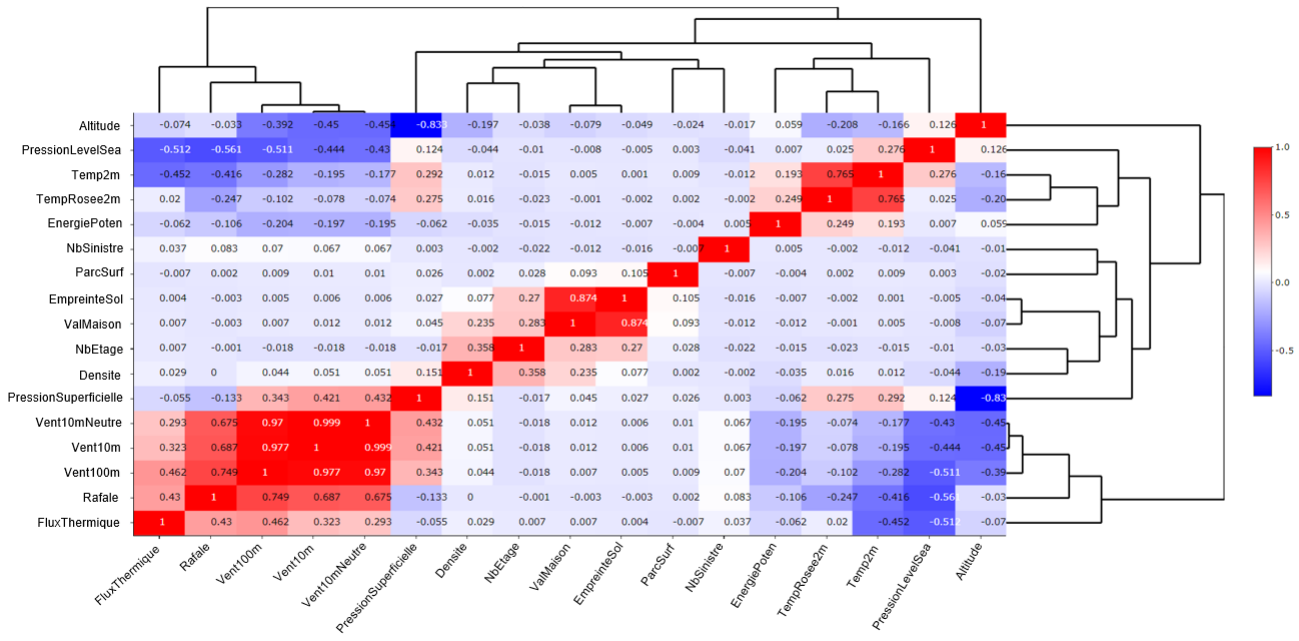

FIGURE 3.2 – Matrice de Corrélation de Pearson

Les variables *ValBatiment* et *EmpreinteSol* présente une corrélation très forte de 0,874, nous conservons alors la variable *ValBatiment* qui est plus parlante de manière générale.

Les variables *TempRosee2m* et *Temp2m* ont une corrélation de 0,765 et dans cette même idée d'explicabilité, nous conservons la variable *Temp2m*.

La dernière corrélation importante que nous relevons concerne les variables *Altitude* et *Pression superficielle* qui ont un coefficient de 0,833. La variable *Altitude* étant une variable du portefeuille et n'étant donc pas soumise à des variations, nous décidons d'opter pour celle-ci.

Coefficients de corrélation de Spearman

Pour compléter cette première étape de corrélation et pallier aux limites évoquées concernant le coefficient de Pearson, nous utilisons dans un second temps le coefficient de corrélation des rangs, également connu sous le nom de rhô de Spearman que l'on notera ρ_s ici. Il explore la présence d'une relation entre les classements des observations pour deux **variables continues** X et Y. Il nous permettra de confirmer ou d'infirmer notre première analyse effectuée avec le coefficient de Pearson.

Cette approche permet de détecter des liaisons monotones, qu'elles soient crois-

santes ou décroissantes, indépendamment de leur forme spécifique (linéaire, exponentielle, puissance, etc.). Ainsi, ce coefficient se révèle particulièrement précieux lorsque l'examen du nuage de point révèle une forme curviligne qui semble donc mal s'ajuster à une droite. Il convient de noter que le coefficient de Spearman sera préféré au coefficient de Pearson dans les situations où les distributions de X et Y présentent des dissymétries et/ou des valeurs exceptionnelles.

La base du coefficient de Spearman repose sur l'étude des différences de rangs entre les attributs des individus pour les caractères X et Y :

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n}(r(X_i)-r(Y_i))^2}{n^3-n}$$

où n est le nombre d'observations, $r(X_i)$ est les rang de X_i dans la distribution $X_1, ..., X_n$ et $r(Y_i)$ est les rang de Y_i dans la distribution $Y_1, ..., Y_n$.

Ce coefficient fluctue dans une plage s'étalent de -1 à +1 et sa signification est similaire à celle du coefficient de Pearson. Cependant, il a la particularité de mettre en lumière des relations non linéaires, qu'elles soient positives ou négatives.

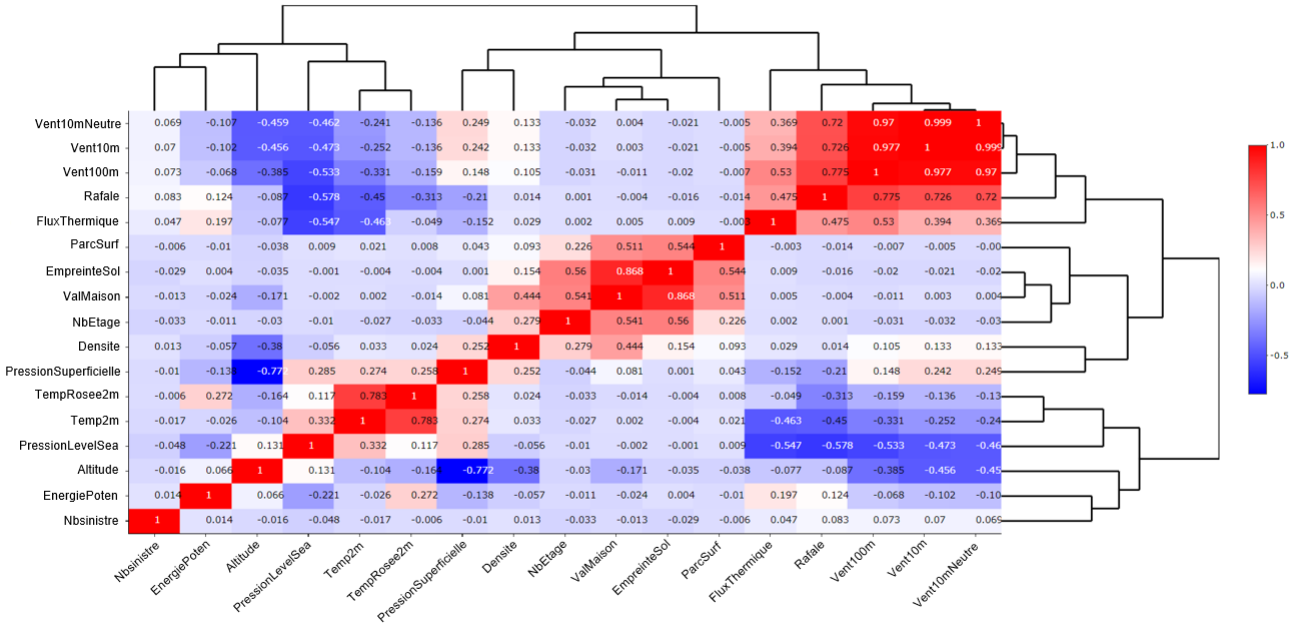

FIGURE 3.3 – Matrice de Corrélation de Spearman

Nous remarquons des coefficients élevés (supérieurs à un seuil arbitraire de 0,7) sur les mêmes variables que nous avions identifiées dans le cas précédent.

Notons tout de même qu'il sera important de garder en mémoire que des variables

comme la *Pression moyenne au niveau de la mer* présentent des coefficients non négligeables de -0,547 avec le *Flux thermique sensible à la surface moyenne* et de -0,578 avec la *rafale*, que le *Flux thermique sensible à la surface moyenne* et la *rafale* ont un coefficient de 0,475 et que le *Flux thermique sensible à la surface moyenne* et *Temp2m* ont un coefficient de -0,463. ces coefficients ne sont pas assez importants pour engendrer une suppression de variable en première instance mais nous serons vigilants sur ces couples de variables pour les prochaines analyses.

Le V de Cramer

Nous nous intéressons désormais aux **variables qualitatives** et pour mesurer le lien d'association entre deux variables qualitatives X et Y nous utilisons le test d'indépendance du χ^2 qui nous permettra ensuite de calculer le V de Cramer.

Le test du χ^2 consiste à tester l'hypothèse nulle selon laquelle les variables X et Y sont considérées comme indépendantes. Soit $X = X_1, ..., X_p$ et $Y = Y_1, ..., Y_q$ deux variables qualitatives possédant respectivement p et q modalités avec p et q supérieurs ou égaux à 2. Notons $n_{i,j}$ le nombre d'observations de modalités X_i et Y_j. Sous l'hypothèse nulle d'indépendance entre les deux variables X et Y, la répartition du nombre d'observations est donnée par :

$$e_{i,j} = \frac{n_{i,.} n_{.,j}}{n}$$

Pour accepter ou rejeter l'hypothèse d'indépendance, la statistique T entre les valeurs observées et théoriques est calculée :

$$T = \sum_{i,j} \frac{(n_{i,j} - e_{i,j})^2}{e_{i,j}}$$

Si l'hypothèse nulle est vérifiée alors $T \sim \chi^2_{(p-1),(q-1)}$. On compare cette statistique de test T au quantile d'ordre 1 - α de la loi du χ^2, α étant classiquement pris égal à 0.05. Dans le cas où T est supérieur ou égal α alors l'hypothèse nulle d'indépendance entre X et Y est rejetée avec un risque d'erreur de première espèce 5% et on décide de conserver l'hypothèse alternative qui affirme que X et Y sont considérées comme non indépendantes.

Pour quantifier le degré de dépendance entre les deux variables, le V de Cramer intervient. Sa formule est la suivante :

$$V = \sqrt{\frac{\chi^2}{n.(\min(p,q)-1)}}$$

La valeur de V est toujours comprise entre 0 et 1. La valeur limite 1 correspond au cas où l'une des variables peut s'écrire comme une fonction de l'autre. La valeur limite de 0 correspond au cas où les deux variables sont indépendantes. L'intensité de la relation entre les variables peut être abordé selon une approche par seuils :

- < 0.10 : la relation est nulle ou très faible
- ≥ 0.10 *et* < 0.20 : la relation est faible
- ≥ 0.20 *et* < 0.30 : la relation est moyenne
- ≥ 0.30 : la relation est forte

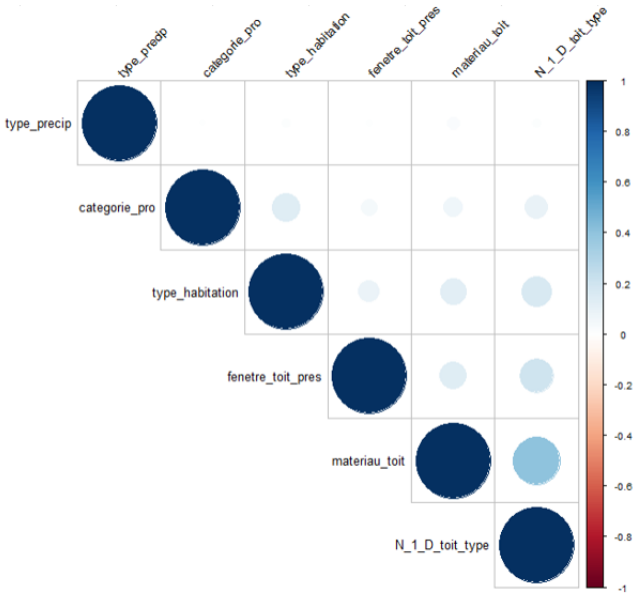

FIGURE 3.4 – Matrice de Corrélation du V de Cramer

D'après cette matrice le seul couple présentant une corrélation importante est le couple $MaterToit$ et $TypeToit$ ayant une corrélation de 0,407. Nous décidons néanmoins de conserver les deux variables qui sont des variables que nous considérons intéressantes pour la sinistralité tempête.

3.2.2 Tests statistiques

Nous souhaitions ici comparer les distributions des variables météorologiques pour le groupe des sinistrés et des non sinistrés. L'objectif étant d'exclure les variables ayant les mêmes distributions pour ces deux groupes distincts en stipulant que ces variables ne seraient pas pertinentes pour notre modélisation. Nous procédons dans un premier temps à un examen visuel en superposant les histogrammes des variables pour les deux groupes puis nous effectuons des tests statistiques pour passer de l'analyse qualitative à une analyse quantitative. Nous décidons d'utiliser deux tests non paramétriques qui ne reposent donc pas sur des hypothèses spécifiques concernant la forme de la distribution : le test de Kolmogorov Smirnov et le test U de Mann-Whitney.

Histogrammes

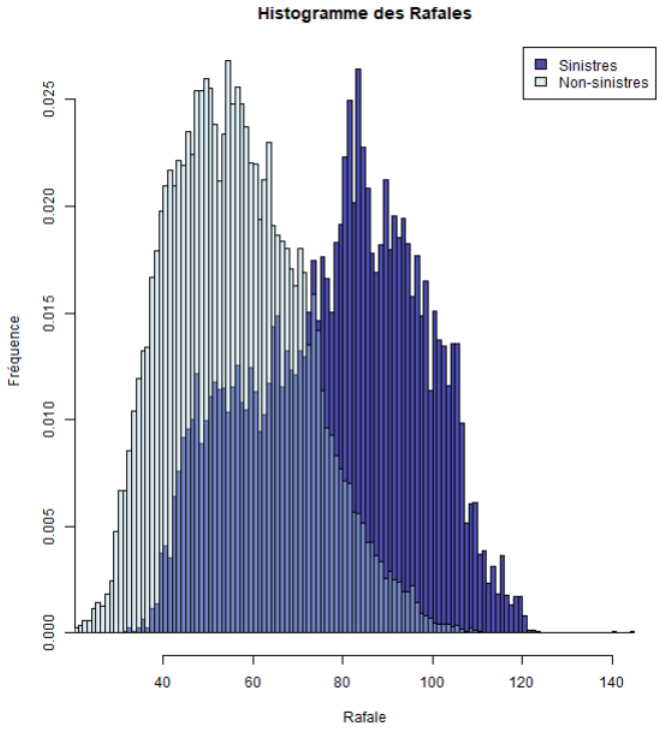

FIGURE 3.5 – Histogrammes des rafales pour les lignes sinistrées et non sinistrées

Nous avons une distinction claire entre ces deux histogrammes avec des pics de fréquence entre 40 et 60 km/h pour les lignes non sinistrées et un pic de fréquence aux environs des 90 km/h pour les lignes sinistrées. Nous remarquons tout de même deux phénomènes qui peuvent être étranges à première vue. Premièrement, nous avons des lignes sinistrées avec des rafales assez faibles entre 40 et 60 km/h. Ces lignes peuvent soit s'expliquer par des sinistres provoqués par des coups de vents et non de violentes rafales soit s'expliquer par une mauvaise approximation effectuée par notre jointure. Chaque bâtiment ne dispose pas d'une station météorologique à part entière, il est donc probable que la rafale qui a provoqué le sinistre ne soit pas celle récupérée durant notre jointure. Deuxièmement, on remarque que certaines lignes non sinistrées ont de violentes rafales supérieures à 80 km/h. Ces lignes peuvent s'expliquer par les variables caractéristiques du bâtiment qui engendrent une impossibilité de sinistre malgré de fortes rafales. Un exemple caricatural serait de visualiser un bunker. On comprend aisément que ce type de bâtiment ne pourrait subir de sinistre quelle que soit la violence du vent.

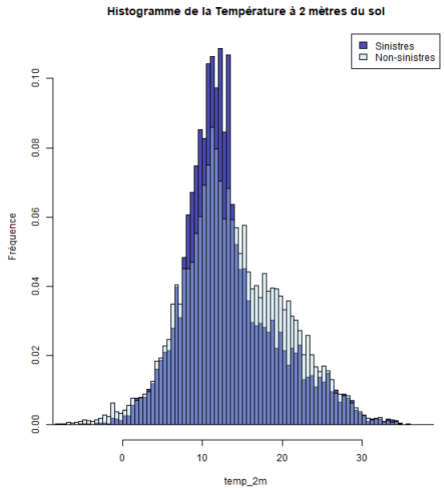

FIGURE 3.6 – Histogrammes des températures pour les lignes sinistrées et non sinistrées

L'analyse de ces histogrammes est moins évidente mais les deux distributions de température semblent être différentes avec un pic plus important aux alentours des 10°C pour les lignes sinistrées et une proportion plus importante entre 13°C et 26°C environ pour les lignes non sinistrées. Les tests statistiques nous permettront de trancher.

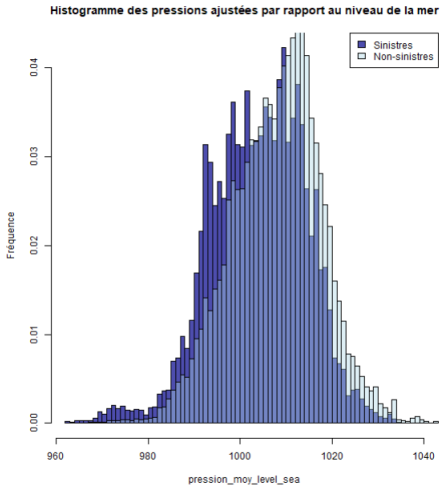

FIGURE 3.7 – Histogrammes des Pressions ajustées par rapport au niveau de la mer pour les lignes sinistrées et non sinistrées

Ici, il semble y avoir une translation de la distribution des pressions pour les lignes non sinistrées vers la droite ce qui indiquerait des pressions moins importantes

pour les lignes sinistrées ce qui est cohérent. En effet, la tempête est un phénomène dépressionnaire ce qui signifie qu'en règle générale, ce genre de phénomène survient lorsque la pression est plus faible que d'habitude.

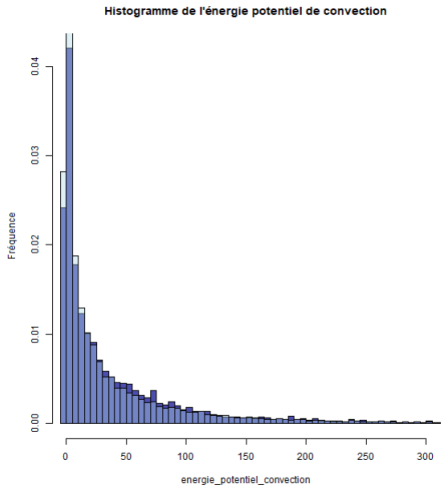

FIGURE 3.8 – Histogrammes de l'énergie potentielle de convection disponible pour les lignes sinistrées et non sinistrées

L'analyse de ces deux histogrammes superposés est assez complexe ici et laisse à penser que les deux distributions sont assez similaires, les tests statistiques nous permettront là aussi de prendre une décision.

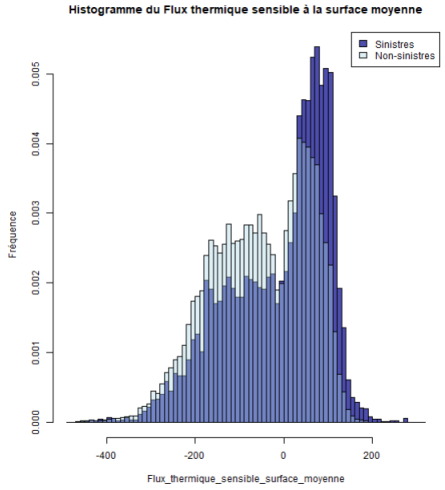

FIGURE 3.9 – Histogrammes du Flux thermique sensible à la surface moyenne pour les lignes sinistrées et non sinistrées

Nous observons ici des formes de distribution similaires mais avec des fréquences

différentes entre les lignes sinistrées et non sinistrées. En effet, nous avons une proportion plus importante pour les flux entre -200 et 0 $W.m^{-2}$ pour les lignes non sinistrées et une proportion plus importante entre 0 et 200 $W.m^{-2}$ pour les lignes sinistrées, ce qui laisse à penser que les deux distributions ne sont pas équivalentes. Les tests statistiques confirmeront cette hypothèse.

Test de Kolmogorov-Smirnov (bilatéral)

Le test de Kolmogorov-Smirnov (KS) est un outil statistique non paramétrique largement utilisé pour évaluer si deux échantillons proviennent de la même distribution continue. Conçu par Andrei Kolmogorov et Nikolai Smirnov, ce test est particulièrement utile pour comparer des échantillons dont on ne connait pas la forme exacte de la distribution.

Le test de KS se base sur la comparaison des fonctions de répartition empiriques des deux échantillons. La fonction de répartition empirique d'un échantillon est une courbe en escalier qui représente la proportion cumulée d'observations inférieures ou égales à une certaine valeur. En comparant ces courbes, le test détermine si les échantillons proviennent de la même distribution en évaluant la plus grande différence verticale entre les courbes.

Voici comment le test fonctionne :

1. **Hypothèse nulle et alternative :** Les hypothèses nulles et alternatives du test sont les suivantes :
   - Hypothèse nulle (H0) : Les échantillons proviennent de populations avec des distributions identiques.
   - Hypothèse alternative (H1) : Les échantillons proviennent de populations avec des distributions différentes.
2. **Calcul des fonctions de répartition empiriques :** On calcule les fonctions de répartition empiriques pour chaque échantillon. Pour rappel, cette fonction F_n pour n observations indépendantes et identiquement distribuées X_i est définit comme ceci :
   $$F_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{(-\infty,x]}(X_i)$$
   où $\mathbb{I}_{(-\infty,x]}$ est la fonction indicatrice valant 1 si $X_i \leq x$ et 0 sinon.
3. **Calcul de la statistique D :** On calcule l'écart maximum entre les fonctions de répartition empiriques. Pour un test bilatéral, la statistique du test s'écrit :
   $$D = \max_x |F_1(x) - F_2(x)|$$
4. **Calcul de la valeur p :** Une conversion entre la statistique D et la p-valeur est effectuée grâce à des tables précalculées ou un logiciel statistique, dans

notre cas R le fait automatiquement. Cette p-valeur représente la probabilité d'obtenir une différence aussi extrême que celle observée, si les échantillons venaient en fait de la même distribution.

5. **Interprétation du résultat :** Si la p-valeur est supérieure à un niveau de seuil (généralement 0,05) on ne rejette pas l'hypothèse nulle et on admet que les échantillons proviennent de la même distribution. Si la p-valeur est inférieur à ce niveau de seuil, on rejette H0 et on considère qu'il existe une différence statistiquement significative entre les deux distributions.

Dans notre cas, le test de Kolmogorv rejette l'hypothèse nulle avec une p-valeur proche de 0 pour l'ensemble des variables météorologiques indiquant ainsi qu'il existe une différence statistiquement significative entre les distributions des populations de chaque variable.

Pour corroborer les conclusions obtenues avec ce test en sachant que ce test amène souvent à rejeter H0 pour des échantillons de grande taille, nous utilisons le test U de Mann-Whitney.

Test U de Mann-Whitney (bilatéral)

Le test U de Mann-Whitney, également appelé le test des rangs de Wilcoxon-Mann-Whitney, est un test statistique non paramétrique utilisé pour déterminer si deux échantillons indépendants proviennent de populations ayant des distributions similaires. Il est utilisé lorsque les données ne suivent pas une distribution normale ou lorsque les échantillons sont de tailles différentes, ce qui est notre cas puisque nous avons beaucoup plus de lignes non sinistrées que sinistrées.

Voici comment le test U de Mann-Whitney fonctionne :

1. **Hypothèse nulle et alternative :** Les hypothèses nulles et alternatives du test sont les suivantes :
   - Hypothèse nulle (H0) : Les échantillons proviennent de populations avec des distributions identiques.
   - Hypothèse alternative (H1) : Les échantillons proviennent de populations avec des distributions différentes.
2. **Assignation des rangs :** Pour chaque observation dans les deux échantillons combinés, les observations sont classées de la plus petite à la plus grande. Les rangs sont attribués en fonction de leur position dans cet ordre. Lorsqu'il y a des ex-aequos dans les valeurs, deux approches sont possibles. Soit les rangs sont attribués de manière aléatoires pour les observations confondues. Soit les observations possédant des valeurs identiques se voient attribuer la moyenne de leurs rangs.

3. **Calcul de la somme des rangs :** La somme des rangs pour chaque échantillon est calculée. Cela représente la somme des rangs attribués aux observations dans chaque échantillon.
4. **Calcul de la statistique U :** La statistique de Mann et Whitney utilise la somme des rangs. Les quantités suivantes sont calculées :

   $$U_1 = S_1 - \frac{n_1(n_1+1)}{2} \; et \; U_2 = S_2 - \frac{n_2(n_2+1)}{2}$$

   où n_1 et n_2 sont la taille de l'échantillon 1 et 2 et S_1 et S_2 sont la somme des rangs de l'échantillon 1 et 2.
   Par convention, la statistique de Mann Whitney correspond à la plus petite quantité, soit :

   $$U = \min(U_1, U_2)$$

   Lorsque l'hypothèse nulle est vraie, l'espérance et la variance de U s'écrivent :

   $$\mathbb{E}(U) = \frac{1}{2}n_1n_2 \; et \; \mathbb{V}(U) = \frac{1}{12}(n_1 + n_2 + 1)n_1n_2$$

   La région critique du test correspond aux valeurs exagérément élevées ou exagérément faibles de U par rapport à son espérance.
5. **Calcul de la p-valeur :** En utilisant la statistique U calculée, une p-valeur est déterminée en fonction de la distribution nulle. Cette p-valeur indique la probabilité d'obtenir une statistique U aussi extrême que celle observée, sous l'hypothèse nulle.
6. **Décision statistique :** .Si la p-valeur est inférieure à un seuil de signification pré-défini (généralement 0,05), on peut conclure qu'il existe une différence statistiquement significative entre les distributions des deux groupes. Dans le cas contraire, on ne peut pas rejeter l'hypothèse nulle.

Ce test amène lui aussi à rejeter l'hypothèse nulle avec une p-valeur proche de 0 pour l'ensemble des variables météorologiques. Cette démarche ne nous aura donc pas permis de supprimer d'autres variables.

3.3 Méthode de catégorisation des variables continues

Nous décidons dans un premier temps de catégoriser l'ensemble de nos variables quantitatives pour nos modélisations. Pour ce faire nous utilisons la méthode des ***K-means***. L'algorithme des K-means est une technique de *clustering* largement utilisée en apprentissage automatique non supervisé pour regrouper des données similaires dans des groupes appelés *"clusters"*.

L'objectif principal du *K-means* est de partitionner un ensemble de données en K *clusters*, où K est un nombre préalablement spécifié par l'utilisateur. Chaque

clusters est représenté par un centre appelé "centroide", qui est le barycentre de tous les points du cluster. L'algorithme vise à minimiser la distance entre chaque donnée et leur centroïde respectif.

Voici les étapes détaillées de l'algorithme des K-means :

1. **Initialisation :** Choisir K centres de *cluster* initiaux. Cela peut être fait de différentes manières, en sélectionnant K points aléatoires parmi les données par exemple.
2. **Affectation des points aux clusters :** Pour chaque point de données, calculer la distance entre le point et chaque centroïde. Assigner le point au cluster dont le centroïde est le plus proche (généralement en utilisant la distance euclidienne). Ainsi, chaque point appartient maintenant à un cluster.
3. **Mise à jour des centroïdes :** Une fois que tous les points sont affectés à des clusters, calculer le nouveau centroïde de chaque cluster en prenant le barycentre de tous les points dans ce cluster. Cela déplace le centre du cluster vers le "centre de gravité" des points qui lui sont assignés.
4. **Répétition des étapes 2 et 3 :** Les étapes d'affectation et de mise à jour des centroïdes sont répétées de manière itérative jusqu'à ce qu'il n'y ait plus de changement dans l'affectation des points aux clusters ou jusqu'à ce qu'un critère prédéfini (comme un nombre maximum d'itérations) soit atteint.
5. **Convergence :** L'algorithme converge lorsque les affectations de points aux clusters cessent de changer, c'est-à-dire que les centroïdes restent les mêmes entre deux itérations consécutives.

L'algorithme K-means cherche ainsi à minimiser la somme des distances entre chaque point de données et le centroïde de son cluster respectif.

Par ailleurs, il existe plusieurs approches pour déterminer le nombre optimal de *clusters*, l'une d'entre elles consiste à **choisir un nombre de *clusters* de manière à ce que la part expliquée de l'inertie soit supérieure à 95%**. L'inertie expliquée étant calculée en divisant la variance interclasse par la variance totale. C'est cette méthode que nous appliquerons.

A la suite de ce processus l'ensemble de nos variables continues étaient donc catégorisées.

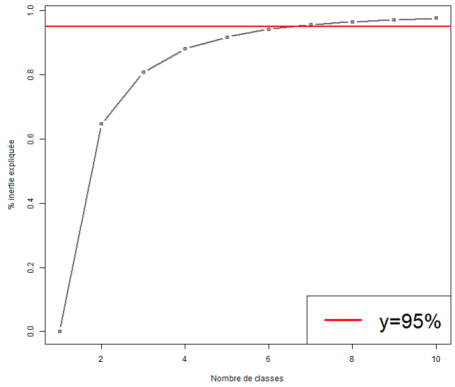

FIGURE 3.10 – Rapport de la variance interclasse sur la variance totale en fonction du nombre de classes pour trouver le K optimal

3.4 Test d'adéquation de lois pour notre variable cible

Pour mettre en place la modélisation, nous devons sélectionner une distribution de comptage qui correspond le mieux à nos données. Pour ce faire, nous effectuons un test d'adéquation entre les distributions de Poisson et Binomiale Négative et nos données observées en utilisant la méthode du Maximum de Vraisemblance. Ces deux lois étant classiquement utilisées pour des problématiques de comptage.

Étant donné l'écart considérable entre le nombre de lignes sans sinistre et le nombre de lignes avec un nombre de sinistres strictement positif, nous appliquons le logarithme du nombre d'occurrences pour permettre une interprétation plus aisée et simplifiée des résultats.

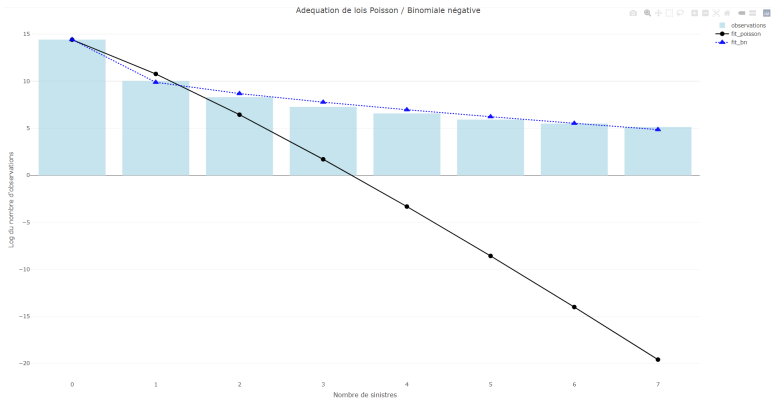

FIGURE 3.11 – Adéquation des lois de comptage

On remarque que la loi binomiale négative semble la plus adaptée à nos observations. La différence est particulièrement marquante lorsque le nombre de sinistre est strictement supérieur à 1 puisque dans ces situations la loi de Poisson a tendance à fortement sous- estimer la réalité. Nous opterons par la suite pour la loi binomiale négative pour effectuer nos modélisations.

Synthèse

Ce chapitre a permis de présenter le portefeuille assurantiel et nos méthodes d'agrégations des différentes bases de données pour permettre d'obtenir la base finale utilisée pour notre modélisation ainsi que différentes analyses statistiques préliminaires. Le prochain chapitre se concentrera sur le modèle mis en place.

Chapitre 4

Calibrage et validation du modèle historique

Notre objectif est d'effectuer un **modèle de fréquence** sur la sinistralité tempête en essayant de **reproduire au mieux la sinistralité observée pour chaque mois et chaque région**. Notre objectif final étant d'effectuer des projections à partir de ce modèle, nous souhaiterions ainsi **comprendre l'évolution géographique et temporelle du risque à travers les années**.

Nos modélisations sont des problématiques de régressions. Nous souhaitons modéliser la valeur d'une variable y qui est notre variable expliquée (le nombre de sinistres ici), en fonction des valeurs de m variables explicatives $x^{(1)}, ..., x^{(m)}$ (caractéristiques de notre portefeuille, variables météorologiques,...).

Généralement, on modélise les valeurs de y et $x = (x^{(1)}, ..., x^{(m)})$ à l'aide d'un couple de variables aléatoires (Y, X) avec $X = (X^{(1)}, ..., X^{(m)})$. L'objectif principal de nos régressions étant de déterminer l'espérance conditionnelle $\mathbb{E}(Y|X)$.

Pour effectuer cette modélisation, on se base sur un jeu de données (notre base finale agrégée) constitué de N observations conjointes des valeurs de y et de x , $[y_i, x_i = (x^{(1)}, ..., x^{(m)})]_{i=1,...,N}$. La modélisation consiste alors à expliciter le lien entre (Y, X) et notre base de données.

L'hypothèse habituellement effectuée consiste à modéliser les réponses $(y_i)_{i=1,...,N}$ comme étant issues d'une suite de variables aléatoires indépendantes $Y_i, ..., Y_N$ telles que $Y_i \sim Loi(Y|X = x_i)$.

De nombreuses méthodes de régression peuvent être appliquées pour notre problème, nous utliserons ici les Modèles Linéaires Généralisés (Generalized Linear Model - GLM) et les Modèles à inflation de zéro (Zero-inflated Model) de par leur

transparence en termes d'interprétation et leur historique important en actuariat non-vie.

4.1 Présentation théorique des GLM

4.1.1 Origine

Les GLM ont été formulés par John Nelder et Robert Wedderburn comme un moyen d'unifier divers autres modèles statistiques comme la régression linéaire, la régression logistique, la régression de Poisson, etc. Ce modèle est une généralisation de la régression linéaire ordinale.

La régression linéaire ordinale prédit la valeur attendue d'une quantité inconnue donnée (variable cible) comme une combinaison linéaire d'un ensemble de valeurs observées (les prédicteurs). Ce paramétrage implique qu'un changement constant dans un prédicteur conduit à un changement constant dans la variable cible. Ceci est approprié lorsque la variable cible peut varier indéfiniment dans un sens ou dans l'autre, ou plus généralement pour toute quantité qui ne varie que d'une quantité relativement faible par rapport à la variation des variables prédictives.

Toutefois, ces hypothèses ne conviennent pas à certains types de variables cibles. Par exemple, dans les cas où l'on s'attend à ce que la variable cible soit toujours positive et varie sur une large plage, les changements constants des prédicteurs pourraient entrainer des variations géométriques i.e. exponentielles, plutôt que des variations constantes.

Les GLM permettent d'apporter une solution dans ce type de situation en permettant aux variables cibles qui ont des distributions arbitraires (non gaussiennes potentiellement) et à une fonction arbitraire de la variable cible (la fonction de lien que nous étudierons par la suite) de varier linéairement avec les prédicteurs (plutôt que de supposer que la réponse elle-même doit varier linéairement).

4.1.2 Formulation

Un modèle linéaire généralisé [1] est constitué de trois composantes essentielles :

— une **composante systématique** : fonction affine des variables explicatives potentiellement encodées $x = (x^{(1)}, ..., x^{(m)})$

— une **composante aléatoire** : spécification du type de $Loi(Y|X)$ au sein de la famille exponentielle

1. Source : les informations de cette partie sont extraites du cours de Jean BERARD indiqué dans la bibliographie

— une **fonction de lien** : spécification de la relation entre $\mathbb{E}(Y|X)$ et la composante systématique du modèle.

Famille exponentielle : *Une loi de probabilité sur* $\mathbb{R}$ *(discrète ou continue) appartient à la famille exponentielle si elle possède une densité de la forme :*

$$f(y) = c(y, \phi).exp(\frac{y\theta - b(\theta)}{\phi})$$

où $\phi > 0$ *est le* ***"paramètre de dispersion"****,* $\theta \in I$ *est le* ***"paramètre naturel"*** *et* $b : I \subset \mathbb{R} \to \mathbb{R}$ *est supposée régulière sur l'intervalle* I*.*

Les lois Normale, Poisson, Gamma, Binomiale, Binomiale Négative, Tweedie et autres sont incluses dans cette famille paramétrique de distributions de probabilité qui possède des propriétés mathématiques intéressantes.

Expression pour l'espérance et la variance : *Si* Y *suit la loi associée à* $b(.)$*,* θ *et* ϕ*, on a :*

$$\mathbb{E}(Y) = b'(\theta) \text{ et } \mathbb{V}(Y) = \phi.b''(\theta)$$

remarquons que l'espérance ne fait intervenir que $b(.)$ *et* θ *et le paramètre* ϕ *a un effet multiplicatif sur la variance.*

Il est possible de reparamétrer le modèle en fonction de $\mu = \mathbb{E}(Y)$, en posant $\theta = (b')^{(-1)}(\mu)$ et $\mathbb{V}(Y) = \phi.v(\mu)$ où $v(\mu) = b''((b')^{(-1)}(\mu))$. Les trois éléments (μ, ϕ) et $v(.)$ caractérisent entièrement la loi de Y que l'on note alors $\mathcal{L}_v(\mu, \phi)$. La fonction $v(.)$ est appelée **fonction de variance**.

Expression pour l'espérance et la variance ici : *Si* Y *suit a loi associée à* (μ, ϕ) *et* $v(.)$ *i.e.* $Y \sim \mathcal{L}_v(\mu, \phi)$*, on a :*

$$\mathbb{E}(Y) = \mu \text{ et } \mathbb{V}(Y) = \phi.v(\mu)$$

Famille exponentielle d'une loi de Poisson $\mathcal{P}(\lambda)$: *Si* $Y \sim \mathcal{P}(\lambda)$ *alors la loi de* Y *se rattache à la famille exponentielle avec les caractéristiques suivantes :*

Caractéristiques	Poisson
Modèle	$Y \sim \mathcal{P}(\lambda)$
μ	λ
ϕ	1
$v(\mu)$	μ
θ	$log(\mu)$
$b(\theta)$	e^{θ}

Famille exponentielle d'une loi Binomiale Négative $\mathcal{NB}(k,p)$: *Si* $Y \sim \mathcal{NB}(k,p)$ *i.e. que pour tout entier* $n \geq 0$,

$$\mathbb{P}(Y=n) = \frac{\Gamma(n+k)}{\Gamma(k)n!}p^k(1-p)^n$$

Alors la loi de Y *se rattache à la famille exponentielle avec les caractéristiques suivantes :*

Caractéristiques	Binomiale Négative
Modèle	$Y \sim \mathcal{NB}(k,p)$
μ	$\frac{kp}{(1-p)}$
ϕ	1
$v(\mu)$	$\mu(1+\frac{\mu}{k})$
θ	$log(1-p)$
$b(\theta)$	$-k\ log(1-e^{\theta})$

Dnas un GLM, on suppose ainsi que la $Loi(Y|X=x)$ est donnée par une loi de la famille exponentielle caractérisée par les paramètres (μ,ϕ) et la fonction de variance $v(.)$ avec :

$$g(\mathbb{E}(Y|X=x) = g(\mu) = \eta = \sum_{i=1}^{m}\beta_i x^{(i)}$$

La fonction g est la **fonction de lien** du modèle, les nombres $(\beta_1,...,\beta_m)$ sont les **coefficients** du modèle estimés par Maximum de Vraisemblance. L'espérance conditionnelle $\mathbb{E}(Y|X=x)$ est donc donnée par :

$$\mathbb{E}(Y|X=x) = \mu = g^{-1}(\eta) = g^{-1}(\sum_{i=1}^{m}\beta_i x^{(i)})$$

Distribution	Fonction de lien
Poisson, Binomiale Négative	*Logarithme népérien*
Gamma, LogNormal	*Logarithme népérien*
Tweedie	*Logarithme népérien*
Binomial	*Logit,Probit,Cloglog*

TABLE 4.1 – Fonction de liens des principales distributions

Exemple classique avec une régression log-Poisson :

La variable cible Y est à valeurs dans $\mathbb{N}$ (comptage) et l'on suppose que :

$$Loi(Y|X=x) = \mathcal{P}(\lambda=\mu) \text{ avec } \mu = e^{\sum_{i=1}^{m}\beta_i x^{(i)}}$$

Ce modèle est un GLM avec $g(\mu) = log(\mu)$, $v(\mu) = \mu$ et $\phi = 1$. Du fait de la fonction de lien logarithmique, les variables explicatives $x^{(i)}$ ont un effet **multiplicatif** sur l'espérance de la loi de Poisson :

$$\mathbb{E}(Y|X = x) = \prod_{i=1}^{m} e^{\beta_i x^{(i)}}$$

On a $\mathbb{V}(Y|X = x) = \mathbb{E}(Y|X = x)$ pour tout x.

4.2 Présentation théorique des modèles à inflation de zéros

Pour tenter de pallier au déséquilibre des données, en l'occurrence la prépondérance des lignes non sinistrées, nous avons décidé de nous intéresser aux modèles à inflation de zéros qui ont été spécifiquement conçu pour répondre à ce type de problématique. Il est important de noter que cette forte concentration de 0 est normale au vu du risque étudié, néanmoins dans une logique de régression le modèle pourrait avoir tendance à prédire une non-occurrence de sinistre d'où l'intérêt de mettre en place ce type de modèle pour challenger les modèles précédents.

Il est tout d'abord nécessaire de faire la distinction entre les différentes raisons qui provoquent ces zéros :

— Premièrement, une ligne non sinistrée peut être due à l'impossibilité pour cette ligne de connaître un sinistre fort des caractéristiques des variables dépendantes. Dans notre cas, nous comprenons qu'un bunker par exemple, ne pourra en aucun cas être affecté par une tempête quelque soit la force du vent. Un exemple plus simple en assurance auto serait de dénombrer le nombre d'accidents de voiture. Les individus qui ne posséderaient pas de voiture seraient classés dans ce type de zéro. On les nomme **zéros structurels**.

— Deuxièmement, une ligne peut être non sinistrée malgré les caractéristiques de ces variables dépendantes qui auraient été susceptibles d'engendrer un sinistre. Ces zéros sont appelés les **zéros aléatoires**.

Pour répondre à la prédominance de l'inflation de zéros dans ce contexte, deux approches sont largement adoptées. D'une part, les **modèles Hurdle** adoptent une démarche en deux étapes pour la modélisation en supposant que les zéros sont exclusivement de nature structurelle, d'autre part les **modèles à inflation de zéros (ZI pour *Zero Inflated*)** incorporent une combinaison de modèles dans leur analyse et permettent de traiter les deux types de zéros simultanément. Dans notre étude nous privilégierons ainsi les modèles ZI.

4.2.1 Modèle Hurdle

Le modèle Hurdle se décompose en deux parties distinctes. La première partie aborde spécifiquement les cas où aucun sinistre ne se produit, tandis que la seconde

traite des cas où un nombre positif de sinistres est observé. Son fonctionnement est semblable à une combinaison de deux types de régression : tout d'abord, une régression logistique pour modéliser la probabilité d'occurrence de sinistres, suivie d'une régression de Poisson ou binomiale négative tronquée à gauche en 1 pour modéliser le nombre de sinistres dans les situations de sinistralité.

La distribution de Y dans le cas d'une régression de Poisson s'exprime ainsi, π étant la probabilité d'inflation de zéros :

$$\mathbb{P}(Y = y_i) \begin{cases} \pi & \text{si } y_i = 0 \\ (1-\pi)\frac{e^{-\lambda}\lambda^{y_i}}{y_i!(1-e^{-\lambda})} & \text{sinon.} \end{cases}$$

Ce type de modèle ne sera pas appliquée ici puisque l'origine des zéros est modélisée exclusivement par considération d'une provenance structurelle, ce qui ne correspond pas à notre problème. En effet, les deux types de zéros sont présents dans notre base. Le modèle suivant, qui combine à la fois les zéros structurels et ceux aléatoires, sera préféré.

4.2.2 Modèles à inflation de zéros

Ces modèles sont construits eux aussi grâce à une modélisation en deux parties, une fonction de masse en 0 et une fonction de répartition de comptage, mais qui cette fois-ci n'est pas tronquée à gauche, d'où la possibilité de modéliser des zéros aléatoires. Ainsi, deux types de zéros sont pris en compte, un zéro structurel estimé par une loi logit et un zéro aléatoire correspondant à un nombre de sinistre nul estimé par la loi de comptage.

La distribution de Y dans le cas d'une régression de Poisson Zéro Inflaté (*Zero Inflated Poisson (ZIP)*) s'exprime ainsi, π étant la probabilité d'inflation de zéros :

$$\mathbb{P}(Y = y_i) = \begin{cases} \pi + (1-\pi)e^{-\lambda} & \text{si } y_i = 0 \\ (1-\pi)\frac{e^{-\lambda}\lambda^{y_i}}{y_i!} & \text{sinon.} \end{cases}$$

On a $\mathbb{E}(Y) = (1-\pi)\lambda$ et $\mathbb{V}(Y) = \lambda(1-\pi)(1+\pi\lambda)$.

La distribution de Y dans le cas d'une régression Binomiale Négative Zéro Inflaté (*Zero Inflated Negative Binomial (ZINB)*) s'exprime ainsi, α étant un paramètre de sur-dispersion :

$$\mathbb{P}(Y = y_i) = \begin{cases} \pi + (1-\pi)(\frac{1}{1+\alpha\lambda})^{\alpha} & \text{si } y_i = 0 \\ (1-\pi)\frac{\Gamma(y+\frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha})y_i!}(\frac{\alpha\lambda}{1+\alpha\lambda})^{y_i}(\frac{1}{1+\alpha\lambda})^{\frac{1}{\alpha}} & \text{sinon.} \end{cases}$$

On a $\mathbb{E}(Y) = (1-\pi)\lambda$ et $\mathbb{V}(Y) = \lambda(1-\pi)(1+(\alpha+\pi)\lambda)$.

4.3 Outils utilisés pour calibrer nos modèles

Nous présentons ici la liste des outils utilisés pour ajuster nos GLM.

4.3.1 Test de WALD

Le test de WALD permet de vérifier que les coefficients estimés sont significativement différent de 0. La commande *glm* de R effectue systématiquement un test de WALD de l'hypothèse $\beta_j = 0$ pour chacun des coefficients $\beta_j, j = 0, ..., d$ estimés. Dans ce cas particulier, la p-valeur du test est simplement donnée par $\mathbb{P}(|Z| \geq |z_j|)$ où $Z \sim \mathcal{N}(0,1)$ et $z_j = \frac{\hat{\beta}_j}{\hat{\sigma}(\hat{\beta}_j)}$.

L'obtention d'un résultat statistiquement significatif pour le coefficient β_j correspond simplement au fait que l'hypothèse $\beta_j = 0$ produit une p-valeur de test suffisamment faible pour conduire à un rejet de cette hypothèse. Cependant le fait q'un modèle GLM ajusté donne lieu à des résultats de **test de WALD fortement significatifs pour chacun des coefficients ne garantit en aucun cas la qualité et la validité du modèle**. Dans notre cas, nous nous servons de ce test comme une mise en garde sur la pertinence d'un coefficient lorsque le résultat du test de WALD est non-significatif pour ce dernier.

4.3.2 Pouvoir discriminant des variables

Pour s'assurer que nos variables aient bien un impact non négligeable sur notre variable cible nous utilisons le ***Spread*** qui est calculé pour chaque variable catégorielle en effectuant le rapport du plus grand coefficient multiplicatif sur le plus petit relativement aux différentes modalités de la variable :

$$Spread(X_i) = \frac{\max\limits_i(\beta_i)}{\min\limits_i(\beta_i)}$$

L'idée est de vérifier que les *Spread* soient assez éloignés de 1. Si une variable a un *Spread* proche de 1, cela signifie que ses betas sont assez proches de 0 et par principe de parcimonie il est préférable de ne pas inclure la variable dans le modèle.

4.3.3 Cohérence des coefficients estimés

La cohérence des coefficients estimés est analysée sous deux angles. Nous nous assurons dans un premier temps que les coefficients soient estimés avec un étroit intervalle de confiance. Cet intervalle est calculé à partir de l'écart type fourni pour chaque coefficient par la fonction *glm* de R et des quantiles souhaités d'une loi normale. Dans notre cas, on calcule des intervalles de confiance bilatéraux symétrique de niveau de confiance (approché) de 95%.

Puis nous vérifions dans un second temps que la tendance des bêtas est similaire avec les relativités observées et prédites. Pour calculer les relativités, nous commençons tout d'abord par calculer la somme des sinistres observés et prédits normalisée par l'exposition pour chaque modalité de la variable catégorielle.

Les relativités observées et prédites sont ensuite calculées pour chaque modalité de la variable catégorielle en divisant les observations et les prédictions normalisées par modalité par les mêmes quantités en lien avec la modalité de référence.

Voici un exemple concret d'une variable en l'occurrence ici la *rafale* où les coefficients estimés respectent les critères évoqués ci-dessus :

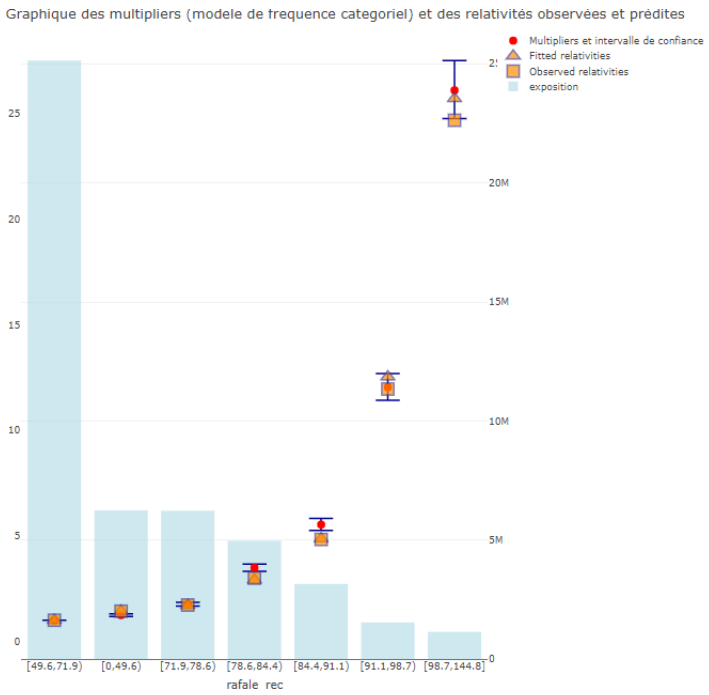

FIGURE 4.1 – Graphique des coefficients et des relativités associées pour la variable *rafale*

4.3.4 Comparaison entre nos observations et nos prédictions

Nous traçons pour chaque variable la somme des sinistres observés et prédits normalisée par l'exposition sur la base de test. L'objectif ici est de vérifier que nos variables sont cohérentes et que les écarts pour chaque modalité sont acceptables.

Voici un exemple concret de ce type de graphique sur une variable en l'occurrence ici la *densité* :

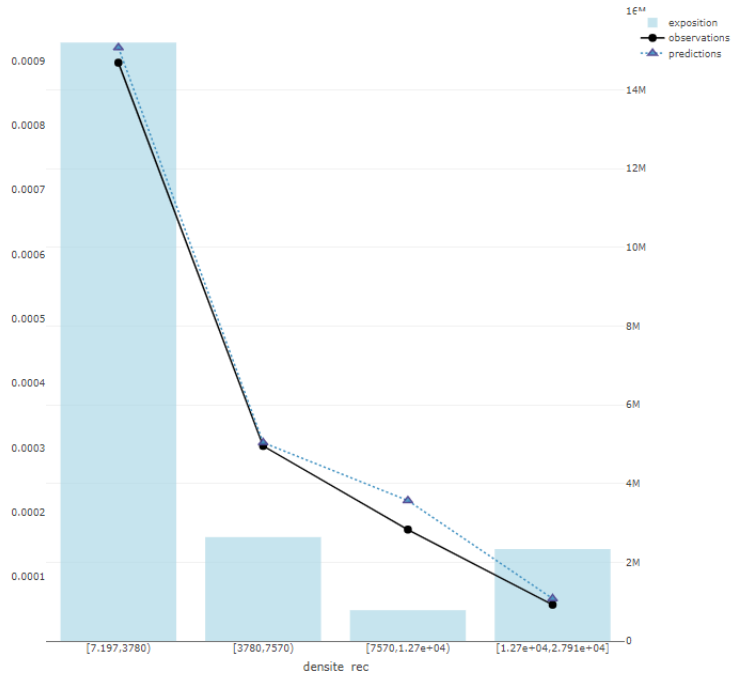

FIGURE 4.2 – Graphique de comparaison entre les observations et les prédictions pour la variable *densité*

4.3.5 Métriques de performance

L'idée de ces métriques est premièrement de vérifier que nos écarts entre nos prédictions et nos observations sont acceptables. Deuxièmement, nous les utilisons pour nous assurer que les ajustements effectués entre chaque version de modèles améliorent leur qualité d'ajustement et leur pouvoir prédictif.

Deviance

La déviance est une statistique qui compare la vraisemblance du modèle avec la vraisemblance d'un modèle saturé c'est à dire comprenant autant de variables explicatives que de données et prédisant donc parfaitement les données :

$$Deviance = 2log\frac{L_{modeleSature}}{L_{modele}} = 2 * (log(L_{modeleSature}) - log(L_{modele}))$$

La déviance est donc une statistique de qualité d'ajustement du modèle analogue à la somme des carrés des résidus dans les modèles linéaires classiques.

Plus la déviance est petite mieux le modèle prédit les données. Cependant cette métrique ne prend pas en compte le risque de sur-ajustement des données - *overfitting*. Lorsqu'on ajoute une variable explicative sans aucun lien avec la variable

dépendante (bruit aléatoire) la déviance diminuera en moyenne de 1 unité. Si la variable est informative, la déviance diminuera de plus d'une unité en moyenne.

AIC (Aikaike Information Criterion)

La déviance ne tient pas compte du problème de sur-ajustement des données (*overfitting*), une situation qui se produit lorsque l'on a trop de variables explicatives/paramètres à estimer par rapport au nombre de données.

Nous complétons alors nos critères de performance avec l'AIC, un critère d'information basé sur la vraisemblance qui pénalise les modèles avec trop de paramètres à estimer :

$$AIC = -2 * log(L) + 2k$$

où L est la vraisemblance du modèle et k le nombre de paramètres.

L'AIC seul est inutile. Il s'agit d'une mesure relative. Pour un jeu de données et un ensemble de modèles pour ce jeu de données, ceux qui ont un AIC plus petits sont considérés comme "meilleurs que les autres". Les AIC permettent également de comparer des modèles non emboîtés (i.e. dont les paramètres ne sont pas un sous-ensemble des paramètres de l'autre modèle).

RMSE (*Root Mean Square Error)*

La RMSE est une métrique de l'erreur qui mesure la différence entre les valeurs prédites par un modèle et les valeurs réelles. Elle calcule la racine carrée de la moyenne des écarts entre les prédictions et les observations :

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

où n est le nombre d'observations, y_i sont les valeurs réelles et $\hat{y}_i$ les valeurs prédites.

Plus la RMSE est faible, plus les prédictions du modèle sont proches des valeurs réelles, indiquant une meilleure performance du modèle. Notons que la RMSE est sensible aux erreurs importantes, car les erreurs sont élevées au carré dans la formule. Cela signifie que les valeurs aberrantes auront un impact plus important sur la RMSE.

Notre objectif est de créer un modèle qui reflète avec précision les tendances mensuelles et géographiques en matière de sinistralité. Pour effectuer des comparaisons entre nos différents modèles, nous agrégeons nos prédictions et nos observations sur ces échelles spatiales (régionales) et temporelles (mensuelles). Ensuite, nous calculons notre indicateur de performance afin d'identifier le modèle affichant les RMSE les plus faibles.

4.3.6 Méthode *stepwise*

Les méthodes *stepwise* (étape par étape) sont des techniques couramment utilisées dans la sélection de variables pour la construction de modèles de régression. Ces méthodes incluent le *forward stepwise*, le *backward stepwise* et le *stepwise* bidirectionnel (both). Elles visent à identifier les variables les plus importantes à inclure dans un modèle en ajoutant ou en supprimant séquentiellement des variables en fonction de leur contribution à la qualité du modèle.

Forward Stepwise Selection (Sélection ascendante)

Cette méthode commence avec un modèle nul (sans aucune variable) et ajoute itérativement les variables explicatives qui ont le plus grand impact sur la performance du modèle. À chaque étape, toutes les variables non encore incluses sont examinées et la variable qui améliore le plus le modèle en termes de critère de sélection (en l'occurrence ici l'AIC) est ajoutée. Le processus se répète jusqu'à ce qu'aucune variable supplémentaire n'améliore davantage la qualité du modèle.

Backward Stepwise Selection (Sélection descendante)

Cette méthode commence avec un modèle comprenant toutes les variables explicatives et supprime itérativement les variables ayant le moins d'impact sur la performance du modèle. À chaque étape, la variable dont la suppression améliore le plus le modèle en termes de critère de sélection est retirée. Le processus se répète jusqu'à ce que la suppression d'autres variables n'améliore pas suffisamment le modèle.

Stepwise Selection (*Both* - Sélection bidirectionnelle)

Cette méthode est une combinaison des deux approches précédentes. Elle commence avec un modèle partiel (peut-être le modèle nul) et peut soit ajouter soit supprimer des variables à chaque étape. À chaque étape, elle examine les effets d'ajout ou de suppression d'une variable et choisit la variable qui améliore le plus le modèle en fonction du critère de sélection. Le processus continue en alternant entre l'ajout et la suppression de variables jusqu'à ce que le modèle atteigne un point où l'ajout ou la suppression d'une variable n'améliore pas suffisamment le critère de sélection.

Ces méthodes sont utilisées pour automatiser la sélection des variables dans le but de simplifier le modèle tout en maintenant sa performance prédictive. Cependant, il est important de noter que les méthodes *stepwise* peuvent avoir des inconvénients, notamment la possibilité de sélectionner des variables non pertinentes ou d'ignorer des variables importantes. Cet inconvénient est dû au caractère itératif

de la méthode. D'autres méthodes comme la régression pénalisée Lasso ou l'exploration par un algorithme élaboré du type algorithme génétique pourraient être envisagées. Néanmoins dans notre cas, nous n'utiliserons ces méthodes *stepwise* qu'en arrière plan de nos critères précédent, la cohérence des coefficients estimés par exemple et leur spread étant privilégié ici.

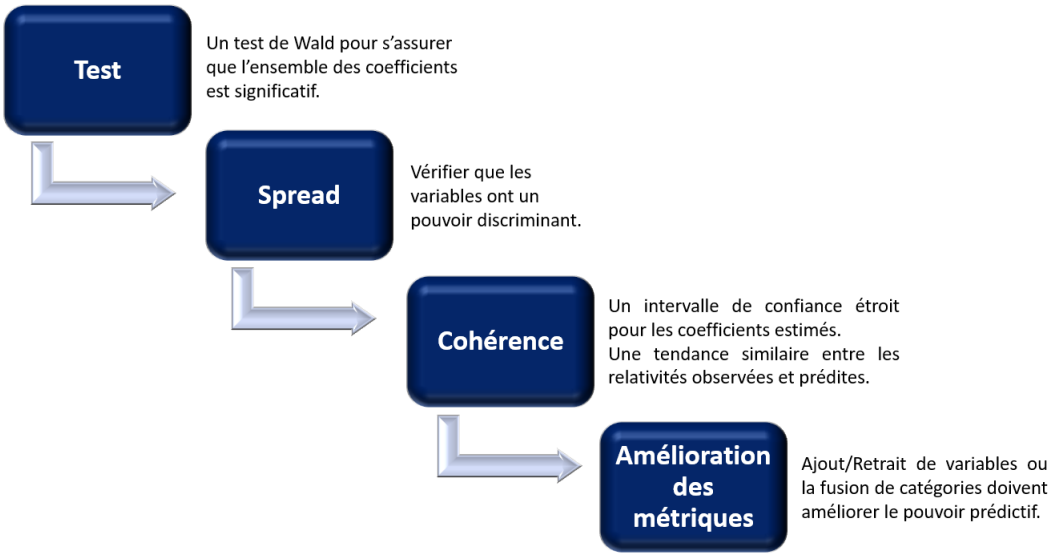

FIGURE 4.3 – Schéma utilisé pour calibrer les modèles

4.4 Application du GLM

Nous séparons notre base de données en une base d'entrainement (80% des données) et une base de test (20% des données) en effectuant cette répartition de manière homogène pour les sinistres de chaque année. Ce découpage nous permettra de vérifier que notre modèle n'est pas en sur-ajustement. Par ailleurs, nous mettons les modalités ayant la plus forte exposition en modalité de référence.

Le nombre de sinistres dans la base d'entrainement est de **34 265** et de **14 714** dans la base de test.

Suppression des variables non pertinentes

Voici la liste des variables du modèle initial :

Variables modèle 1
Altitude
Densite
NbEtage
ParcSurf
TypeHabitation
MaterToit
TypeToit
ValMaison
Temp2m
Rafale
Flux thermique sensible à la surface moyenne
PressionLevelSea
énergie potentielle de convection disponible

TABLE 4.2 – Liste des variables du modèle 1

Dans cette première version de modèle, l'ensemble des paramètres sont significatifs d'après le test de WALD. Nous observons ensuite la cohérence des coefficients estimés ainsi que les spreads. Deux variables attirent notre attention : *Flux thermique sensible à la surface moyenne* et *PressionLevelSea*.

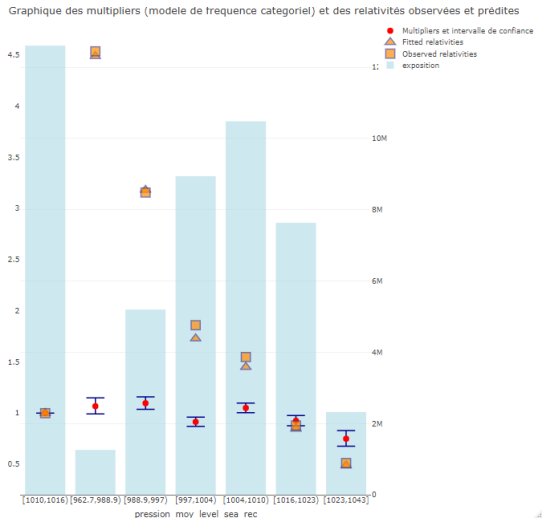

FIGURE 4.4 – Graphique des coefficients et des relativités associées pour la variable *PressionLevelSea*

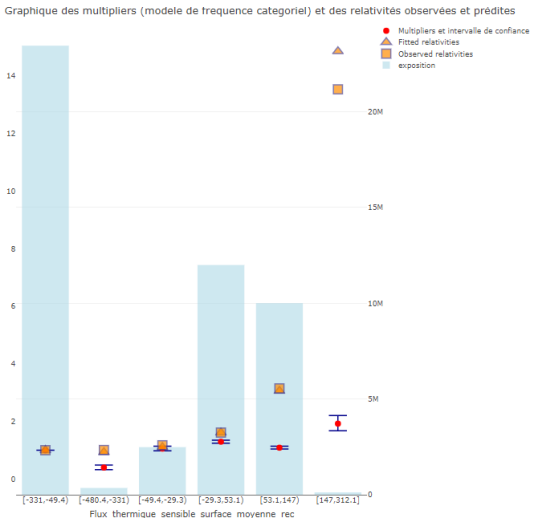

FIGURE 4.5 – Graphique des coefficients et des relativités associées pour la variable *Flux thermique sensible à la surface moyenne*

Ces variables possèdent des spreads assez faibles de plus on remarque que leurs coefficients ne reproduisent pas la tendance dessinée par les relativités. Par ailleurs, la variable *ValMaison* a elle aussi des coefficients avec une tendance contraire aux relativités. Nous décidons alors de les supprimer du modèle sachant que leur suppression améliore en plus les métriques du modèle (calculés sur la base d'entrainement pour le calibrage).

Métriques	Modèle de base	Modèle avec suppressions
AIC	167 608.4	167 049.7
Deviance	90 917.37	90 914.65
RMSE (mois)	263.0225	238.1888
RMSE (régions)	232.07	221.3414
Sinistres prédits	15 274.07	15 148.54

TABLE 4.3 – Métriques de comparaison des modèles

Fusion de catégories

Voici la liste des variables du modèle suivant :

Variables modèle 2
Altitude
Densite
NbEtage
ParcSurf
TypeHabitation
MaterToit
TypeToit
Temp2m
Rafale
énergie potentielle de convection disponible

TABLE 4.4 – Liste des variables du modèle 2

Dans cette seconde version, nous effectuons des **regroupement de catégories** en associant les catégories ayant des bêtas estimés similaires et une sinistralité observée proche, en prenant garde que nos modifications améliorent les métriques du modèle :

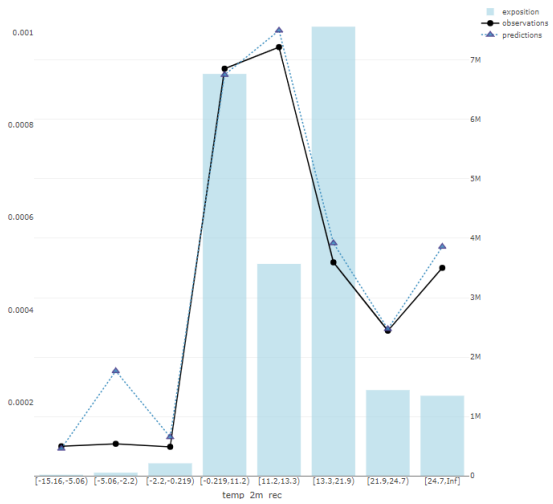

FIGURE 4.6 – Graphique de comparaison entre les observations et les prédictions pour la variable *Temp2m*

Pour la variable *Temp2m*, nous remarquons que les trois premières catégories présentent un risque observée similaire, nous décidons alors de les regrouper.

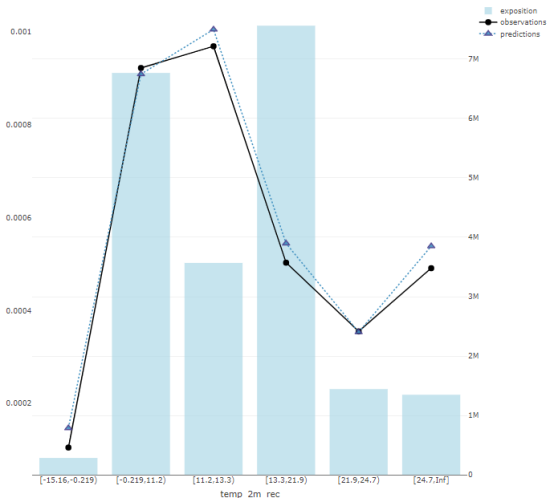

FIGURE 4.7 – Graphique de comparaison entre les observations et les prédictions pour la variable *Temp2m* recatégorisée

Le regroupement a porté ses fruits et permet une adéquation plus fine entre les observations et les prédictions pour les premières catégories.

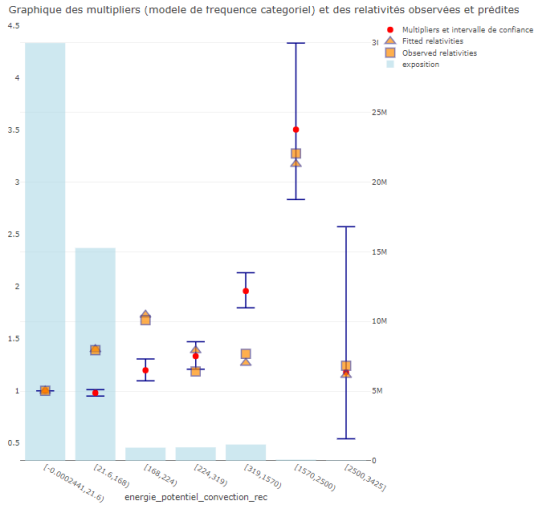

FIGURE 4.8 – Graphique des coefficients et des relativités associées pour la variable *énergie potentielle de convection disponible*

Pour la variable *énergie potentielle de convection disponible*, nous remarquons que les deux dernières catégories présentent une exposition quasi nulle, nous regroupons alors les trois dernières catégories pour obtenir une exposition plus intéressante.

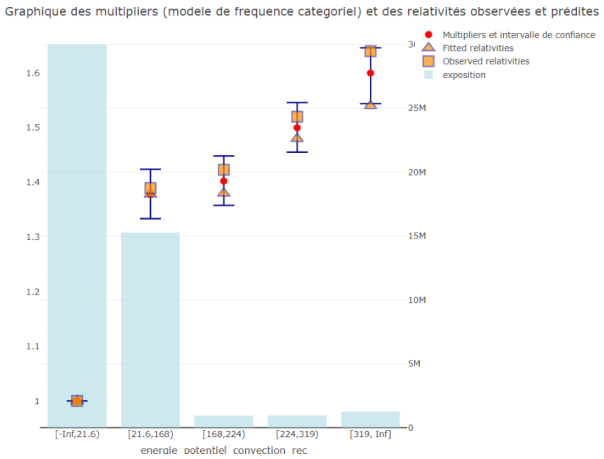

FIGURE 4.9 – Graphique des coefficients et des relativités associées pour la variable *énergie potentielle de convection disponible recodée*

Les autres variables ont un comportent cohérent vis à vis des coefficients et de leurs relativités et la comparaison entre les observations et les prédictions ne montrent pas d'écarts aberrants. **L'ensemble des graphiques concernant les autres variables sont joints en Annexe**.

Métriques	Modèle avec suppressions	Modèle avec suppressions et recatégorisations
AIC	167 049.7	167 012.2
Deviance	90 914.65	90 912.43
RMSE (mois)	238.1888	235.3915
RMSE (régions)	221.3414	216.4492
Sinistres prédits	15 148.54	15141.5

TABLE 4.5 – Métriques de comparaison des modèles

Comparaison avec un modèle GLM Poisson

Nous effectuons, à titre indicatif, une comparaison avec le modèle GLM Poisson. Nous observons comme attendu une dégradation des performances avec la loi de Poisson :

Métriques	Modèle Binomiale Négative	Modèle Poisson
AIC	167 012.2	213 913
Deviance	90 912.43	166 778.6
RMSE (mois)	235.3915	307.4622
RMSE (régions)	216.4492	372.3801
Sinistres	15 141.5	15 590.91

TABLE 4.6 – Métriques de comparaison des modèles

4.5 Validation du GLM

4.5.1 Dépendance entre variables explicatives - analyse des VIF (*Variance Inflation Factor*)

Les facteurs d'inflation de la variance sont un outil permettant d'évaluer *a posteriori* l'impact de la dépendance entre variables explicatives sur la précision de l'estimation des coefficients. Pour une variable i, la VIF se calcule par la formule suivante :

$$VIF_i = \frac{1}{1-R_i^2}$$

où R_i^2 est le coefficient de détermination de la régression de la variable i par l'ensemble des autres variables explicatives du modèle.

On considère qu'une VIF supérieure à 5 indique la présence d'une multicolinéarité entre la variable en question et les autres et son utilisation est donc néfaste pour l'ajustement du modèle.

Dans notre modèle GLM Binomiale Négative, la variable **TypeHabitation** a la VIF la plus élevée avec une valeur d'environ 2,2. Par conséquent, nous pouvons faire l'approximation que la colinéarité entre nos différentes variables est négligeable.

4.5.2 Analyse des résidus

Les résidus constituent une mesure de l'écart entre la valeur de l'espérance modélisée pour la réponse $\mu_i^{est.}$ et la valeur observée y_i de celle-ci. Dans l'idéal, on souhaiterait disposer de résidus $r_1, ..., r_n$ qui lorsque le modèle étudié est valide, constituent des réalisations de variables aléatoires $R_1, ..., R_N$:

- centrées i.e. $\mathbb{E}(R_i) = 0$
- de variance unité i.e. $\mathbb{V}(R_i) = 1$
- de loi normale $\mathcal{N}(0, 1)$
- indépendantes

On peut alors éprouver la validité du modèle, en examinant la conformité des tracés produits vis-à-vis de ces hypothèses [2] :

- Vérifier le caractère centré des résidus le long des divers tracés aide à vérifier l'absence de sur ou de sous estimation systématique de l'espérance.
- Pour des résidus centrés, vérifier la constance à 1 de la variance le long des divers tracés aide à vérifier l'abcence de sur ou de sous estimation systématique de la variance.

2. Source : ces informations sont extraites du cours de Jean BERARD indiqué dans la bibliographie

- Pour des résidus centrés de variance unité, vérifier la normalité de la distribution le long des divers tracés aide à vérifier le caractère approprié de la loi utilisée par le modèle.

Afin de disposer de résidus qui, si le modèle est valide se comportent approximativement comme des variables aléatoires centrées, indépendantes et de variance constante on introduit les résidus de Pearson :

$$R_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{v(\hat{\mu}_i)/w_i}}$$

où Y_i est l'observation, $\hat{\mu}_i$ la prédiction et w_i l'exposition pour la ligne i.

Théoriquement, il est possible de démontrer que ces résidus ne peuvent suivre une loi normale même lorsque le modèle est valide et que l'on dispose d'un jeu de données important, excepté lorsque l'exposition totale associée à chaque ligne i du jeu de données tend vers $+\infty$. Pour se rapprocher de cette situation, on peut considérer des *résidus agrégés* (*crunched residuals*) dans lesquels on mesure l'écart entre l'espérance modélisée et la réponse totale observée pour un groupe de données G représentant une exposition totale suffisante. On peut définir le résidus de Pearson agrégé sur le groupe G comme ceci :

$$R_G^P = \frac{\sum_{i \in G} w_i y_i - \sum_{i \in G} w_i \hat{\mu}_i}{\sqrt{\sum_{i \in G} w_i v(\hat{\mu}_i)}}$$

Nous traçons ainsi ce type de résidus en ayant recours à 500 groupes sur lesquels seront calculés les résidus agrégés. Il s'agit d'un compromis permettant d'avoir à la fois une exposition substantielle par paquet, et un nombre raisonnable de résidus à examiner. De plus, on calcule deux seuils de valeurs calculés par approximation normale : un seuil ayant une probabilité 95% de ne pas être dépassé, individuellement pour chaque résidu, et un seuil global ayant une probabilité d'environ 95% de n'être dépassé par aucun des résidus. Une transformation logarithmique est aussi appliquée afin de faciliter la visualisation grâce à un meilleur étalement des résidus :

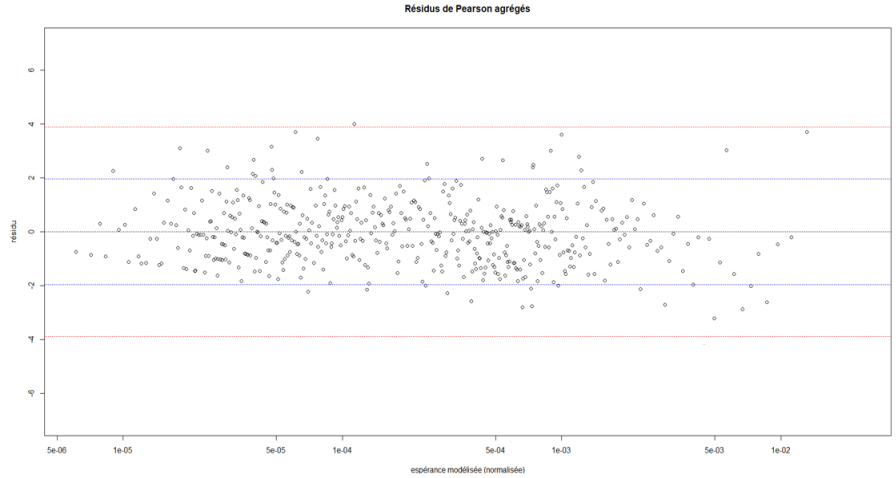

FIGURE 4.10 – Graphique des résidus de Pearson agrégés

Nous n'observons quasiment pas de points aberrants ou de forme suspecte sur ce nuage de points, les hypothèses semblent donc valides. Nous effectuons pour finir une comparaison entre la fonction de répartition empirique des résidus et la fonction de répartition de la loi $\mathcal{N}(0,1)$ pour vérifier l'hypothèse de normalité :

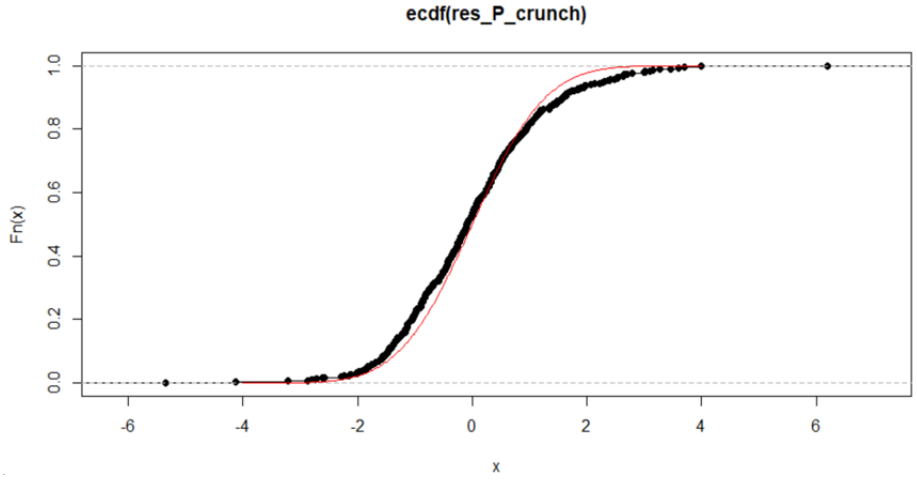

FIGURE 4.11 – Comparaison des fonctions de répartition

Nous observons ici un léger décalage entre les deux fonctions de répartition mais qui reste acceptable pour affirmer que la loi choisie pour effectuer notre modélisation est cohérente.

4.5.3 Lift curve

Nous subdivisons ici nos prédictions en classe de risques croissant d'expositions similaire et nous comparons ces prédictions avec nos observations. Nous traçons ces graphiques sur la base d'entrainement et la base de test, l'objectif étant d'avoir des tendances similaires et des écarts minimes entre la courbe des observés et des prédits.

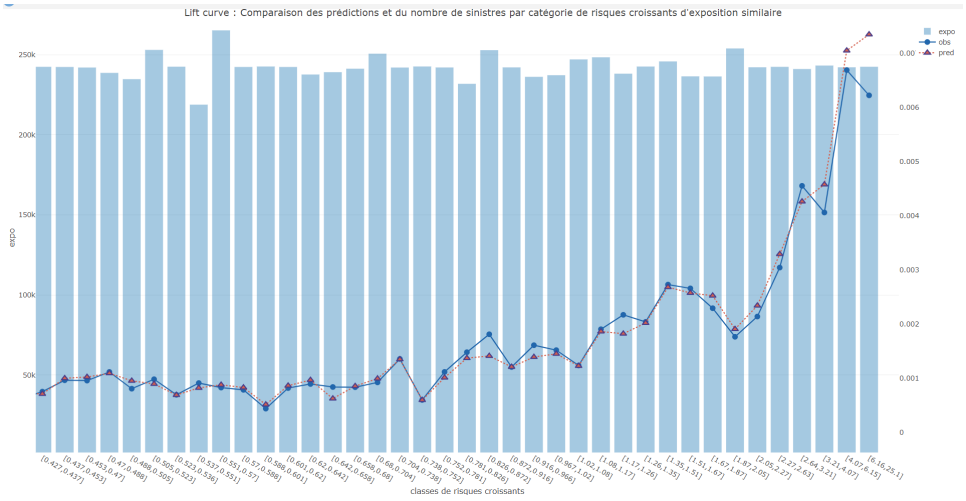

FIGURE 4.12 – Lift curve sur la base d'entrainement

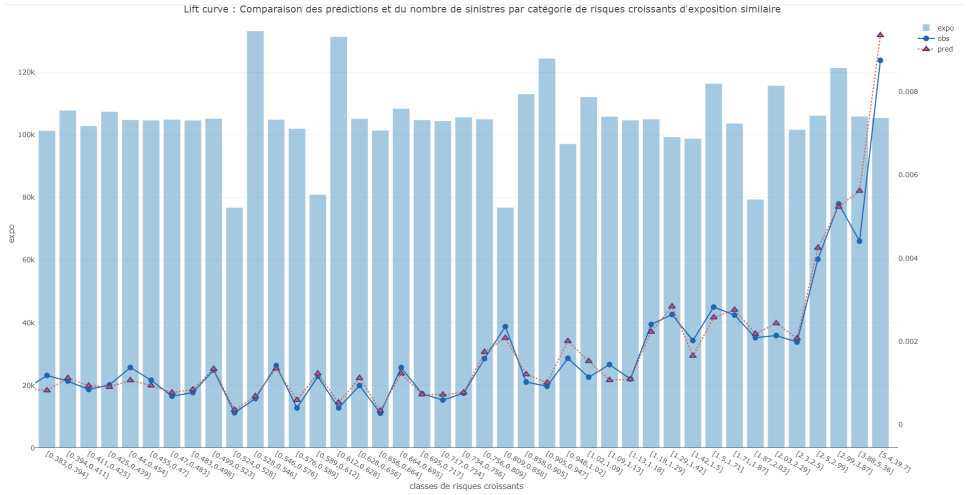

FIGURE 4.13 – Lift curve sur la base de test

Nous ne remarquons pas d'écarts aberrants entre nos prédictions sur la base d'entrainement et nos prédictions sur la base de test. Nous n'avons pas non plus de

sur-ajustement ou de sous-ajustement au vu de ces graphiques. Nous en concluons que notre modèle est cohérent.

4.6 Amélioration grâce à une régression Binomiale Négative Zéros Inflatés

Nous implémentons désormais un modèle ZINB et leur implémentation nécessite au préalable de faire un choix concernant les variables à inclure dans la partie logit. Pour effectuer ce choix, nous décidons d'effectuer une régression logistique avec comme variable cible, une variable binaire indiquant la survenance ou non d'un sinistre. L'idée étant d'analyser les variables les plus pertinentes en fonction de leurs coefficients et de leur significativité. L'ensemble des variables ayant un impact non négligeable dans cette régression logistique, nous incluons alors la même combinaison de variables dans la composante logit et dans la composante GLM. L'avantage d'avoir cette même combinaison implique que les coefficients estimés sont similaires entre le modèle GLM et le modèle ZINB.

Métriques	Modèle Binomiale Négative	Modèle ZINB
AIC	167 012.2	165 127.8
RMSE (mois)	235.3915	162.953
RMSE (régions)	216.4492	131.0178
Sinistres	15 141.5	14 902.4

Table 4.7 – Métriques de comparaison des modèles

Ce modèle améliore les performances prédictives et sera donc sélectionné par la suite pour effectuer nos projections. Voici les graphiques comparatifs par département et par région entre nos prédictions et nos observations sur la base test. Ces mêmes prédictions ont aussi été effectuées sur la base d'entrainement pour vérifier la similarité des écarts et que nous ne sommes pas en situation de sur-ajustement.

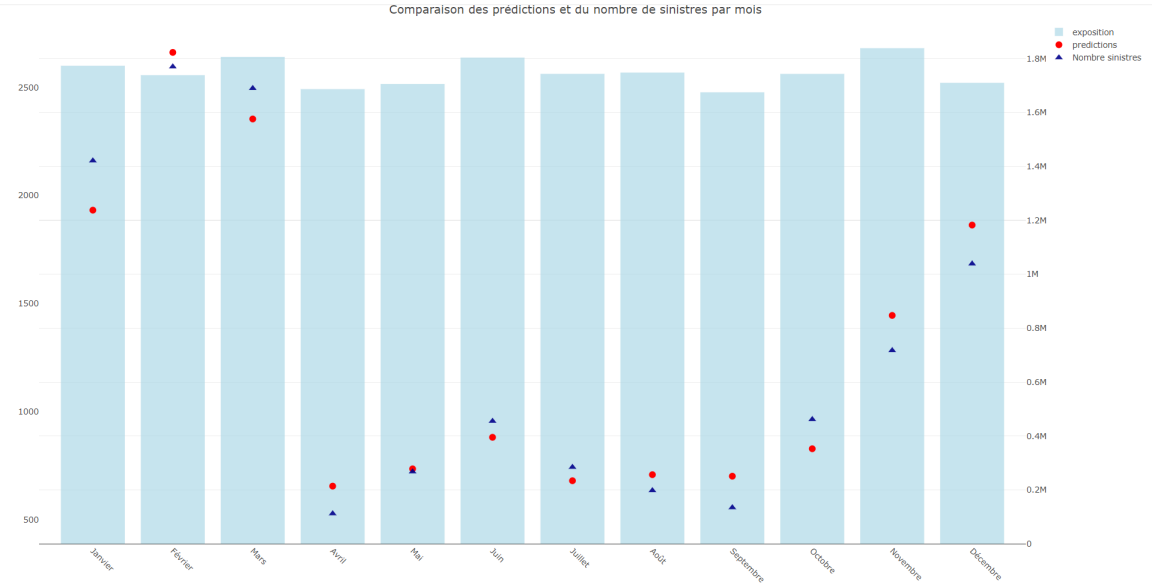

FIGURE 4.14 – Comparaison des prédictions et des observations à l'échelle mensuelle

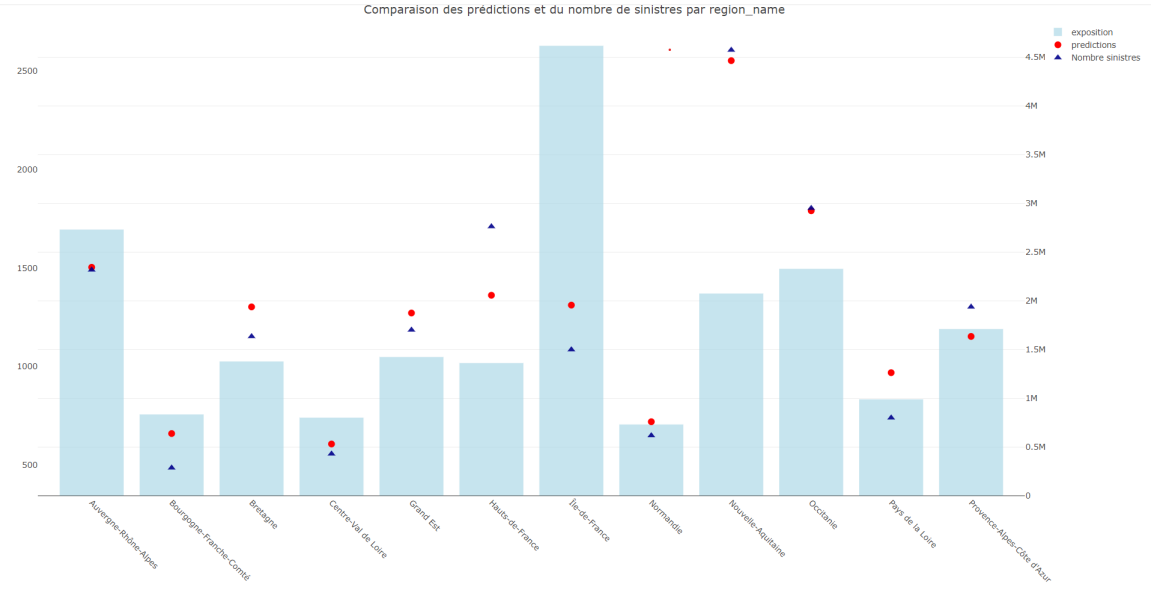

FIGURE 4.15 – Comparaison des prédictions et des observations à l'échelle régionale

Synthèse

Cette partie a permis de présenter la démarche adoptée pour calibrer le modèle historique ayant pour objectif de reproduire le plus précisément possible, la fréquence de sinistre régionale et mensuelle. Une attention particulière a été portée sur l'analyse de la pertinence des variables et en particulier celles météorologiques, qui seront à la source de nos projections.

La méthode employée ici reste perfectible et plusieurs limites sont à présenter. Premièrement, l'ensemble des variables a été catégorisé. Un modèle intégrant des splines ou des polynômes pourrait être mis en place ne serait-ce qu'à titre de comparaison. Les premiers essais n'étaient pas concluants sur l'amélioration des résultats au regard de la complexité intégrée de ces méthodes. Deuxièmement, des études complémentaires pourraient être effectuées pour comprendre l'origine des différences entre nos observations et nos prédictions par région ou par mois pour éventuellement déceler un schéma qui engendre ces écarts. Les résultats étant globalement satisfaisants, nous avons décidé de passer directement à l'étape de projection.

Chapitre 5

Projection du risque tempête

Dans cette partie, nous présentons notre approche pour projeter la fréquence de sinistralité tempête à partir de notre modèle historique et des données de projection des modèles climatiques. L'objectif est d'analyser l'évolution géographique et temporelle de cette fréquence à horizon 2100.

5.1 Présentation des données de projection

Pour limiter les incertitudes inhérentes aux modèles climatiques, nous avons vu dans le chapitre 1 que l'idéal était d'effectuer une **approche multi-modèles et multi-scénarios**. Il est donc nécessaire dans un premier temps de sélectionner un ensemble de modèles et un ensemble de scénarios que nous souhaitons utiliser pour effectuer les projections.

Concernant les **scénarios**, nous avons décidé de retenir le **RCP 4.5** qui a été retenu dans le dernier stress test de l'ACPR et le **RCP 8.5** qui est le scénario extrême. Nous avons en revanche exclu les scénarios RCP 6.0, qui n'est plus utilisé dans le dernier rapport du GIEC, et le RCP 2.6 qui est jugé trop optimiste d'après la communauté scientifique.

Concernant les **modèles climatiques**, nous avons pris l'ensemble des couples GCM/RCM présents à la fois pour le RCP 4.5 et pour le RCP 8.5. Nous n'avons pas appliqué de sélection en amont puisqu'il n'y pas **a priori** de modèle meilleur qu'un autre d'où l'idée de tous les combiner dans nos projections pour avoir les résultats les plus robustes possibles. Voici les différents couples sélectionnés :

GCM	RCM
CNRM-CERFACS-CNRM-CM5	CLMcom-CCLM4-8-17
CNRM-CERFACS-CNRM-CM5	KNMI-RACMO22E
CNRM-CERFACS-CNRM-CM5	SMHI-RCA4
ICHEC-EC-EARTH	CLMcom-CCLM4-8-17
ICHEC-EC-EARTH	KNMI-RACMO22E
ICHEC-EC-EARTH	SMHI-RCA4
IPSL-CM5A-MR	SMHI-RCA4
MPI-M-MPI-ESM-LR	CLMcom-CCLM4-8-17
MPI-M-MPI-ESM-LR	SMHI-RCA4

TABLE 5.1 – Listes des couples GCM/RCM sélectionnés pour effectuer les projections

5.2 Méthode de projection

5.2.1 Mise en place d'un algorithme de traitement des fichiers NetCDF

La première étape consiste à télécharger l'ensemble des données climatiques sur le site d'EURO-CORDEX et à les mettre en forme pour pouvoir les exploiter dans nos logiciels statistiques.

Les fichiers sont initialement en trois dimensions avec la latitude, la longitude et le temps. L'idée est alors d'extraire la variable météorologique et de la transposer dans un tableau classique en deux dimensions contenant quatre colonnes : la date, la longitude, la latitude et la valeur du paramètre météorologique en lien avec cette date et ce couple de coordonnées. Nous devons pour ce faire dupliquer chaque date ainsi que les couples de coordonnées, correspondant aux points dans la grille de résolution, et extraire les paramètres météorologiques liés à la date et aux points en question.

date	latitude	longitude	rafale
2021-01	45.95388	5.399719	21.24233
2021-02	45.95388	5.399719	21.55686
2021-03	45.95388	5.399719	21.87842
2021-04	45.95388	5.399719	19.83147
2021-05	45.95388	5.399719	15.36266
2021-06	45.95388	5.399719	17.79186
2021-07	45.95388	5.399719	15.02899
2021-08	45.95388	5.399719	19.47023
2021-09	45.95388	5.399719	18.73149
2021-10	45.95388	5.399719	25.74673
2021-11	45.95388	5.399719	24.52311
2021-12	45.95388	5.399719	21.33951

FIGURE 5.1 – Exemple d'une base retranscrite dans un tableau en deux dimensions

Dans le cas d'EURO-CORDEX, les fichiers ont une particularité puisqu'ils sont fournis dans un **système de coordonnées rotationnelles**. Ce système permet de retranscrire de manière plus fine l'aspect sphérique du globe. Une grille en coordonnées classiques est néanmoins incluse dans le fichier originel pour pouvoir effectuer la correspondance entre les deux systèmes de coordonnées.

5.2.2 Agrégation des différentes variables météorologiques projetées

Après avoir téléchargé l'ensemble des variables projetées pour l'ensemble des modèles et scénarios jusqu'à l'horizon souhaité (ici 2100), une étape d'agrégation est désormais nécessaire pour constituer des bases climatiques projetées qui seront au nombre de 18 puisque nous avons 9 couples GCM/RCM et 2 scénarios. Les variables météorologiques projetées sont celles retenues par notre modèle historique :

— la **rafale**

— la **température**

— l'**énergie potentielle de convection disponible**

Pour effectuer cette agrégation, nous utilisons le même critère qui nous a permis d'agréger les variables météorologiques de notre base historique : la rafale maximale mensuelle est identifiée ainsi que le jour auquel elle s'est produite. Puis nous utilisons ce jour pour extraire les valeurs relatives des températures et des énergies potentielles pour enfin accoler ces trois bases par l'intermédiaire de ces dates communes et des points qui composent la base.

Pour finir, nous rapprochons chaque point, identifié par sa latitude et sa longitude, des coordonnées du centroïde du canton le plus proche en minimisant la distance euclidienne séparant ces deux points et en ne prenant en compte que les cantons présents dans notre portefeuille assurantiel.

5.2.3 Création de la base assurantielle projetée

Nous commençons par isoler les lignes de l'année 2020 (la dernière année présente dans le portefeuille) pour effectuer une projection avec les dernières conditions disponibles de notre portefeuille. Ce choix a été effectué pour mesurer uniquement l'impact du changement climatique. En effet, dans ces projections, seules les variables météorologiques fluctuent.

Nous remplaçons ensuite année après année les variables météorologiques historiques par celles projetées grâce aux différentes bases construites précédemment. Nous effectuons cette procédure pour chaque modèle et chaque scénario.

5.2.4 Calcul des projections

L'obtention des résultats de projection est assez simple puisqu'une fois notre modèle historique calibré, nous n'avons plus qu'à utiliser ce dernier en remplaçant notre base d'entrainement par nos bases assurantielles projetées et utiliser la fonction *predict* de R par exemple pour obtenir l'espérance prédite du nombre de sinistres.

5.3 Analyse et interprétation des résultats

5.3.1 Une différence de stabilité entre les scénarios

Dans une logique d'approche multi-scénarios et en comparant le résultat du nombre de sinistres projetés à l'échelle mensuelle et régionale pour le même modèle mais forcé par deux RCP différents, on remarque des effets beaucoup plus instables pour le RCP 8.5 :

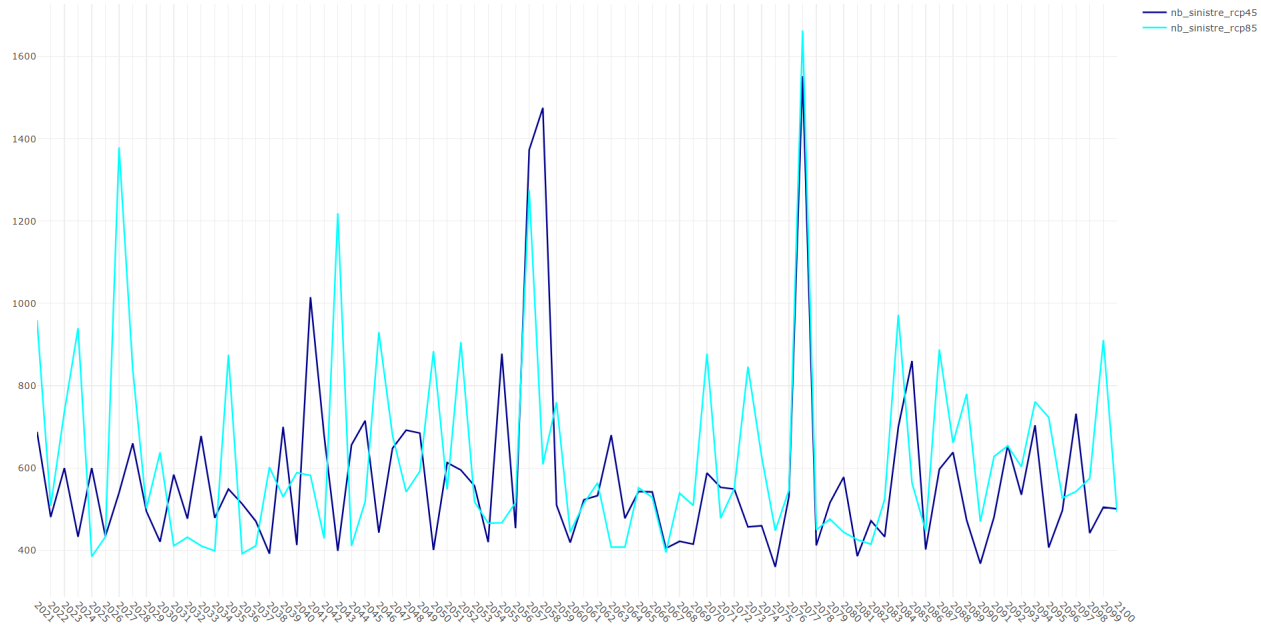

FIGURE 5.2 – Évolution du nombre de sinistres en Normandie pour un modèle du CNRM

Nous présentons un exemple de graphique qui montre ces différences en terme de nombre de pics pour les deux scénarios. L'exemple sélectionné prend le cas de la Normandie et d'un modèle du CNRM néanmoins nous retrouvons cet effet sur la majorité des modèles, des régions et des mois. Cette instabilité est cohérente et peut être expliqué par les effets des rétroactions qui peuvent être amplifiés dans le cas d'un réchauffement plus important. Nous pouvons ainsi déduire qu'un monde où la concentration des gaz à effet de serre augmente serait de moins en moins

prévisible. Par ailleurs, nous voyons ici que le risque tempête n'augmenterait pas nécessairement mais les événements extrêmes pourraient être plus nombreux.

5.3.2 Analyse de l'évolution annuelle des sinistres à l'échelle nationale sous une vision multi-modèles et multi-scénarios

Dans une logique d'approche multi-modèles et multi-scénarios, nous souhaitions tout d'abord visualiser l'évolution annuelle du nombre de sinistres d'un point de vue globale sur la France métropolitaine et pour ce faire nous avons tracé ces évolutions en effectuant la moyenne des sorties de chaque modèle encadré par les résultats des sorties maximales et minimales, pour chaque scénario :

FIGURE 5.3 – Évolution du nombre de sinistres annuels en combinant l'ensemble des modèles pour le RCP 8.5

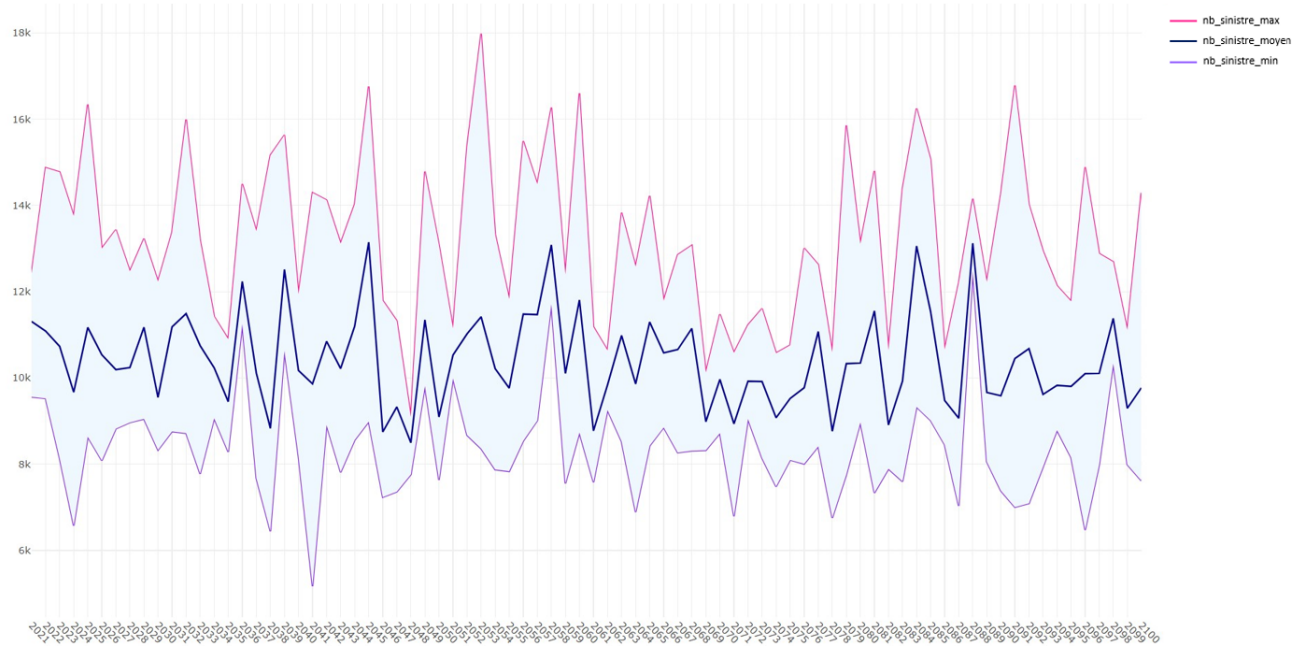

FIGURE 5.4 – Évolution du nombre de sinistres annuels en combinant l'ensemble des modèles pour le RCP 4.5

Nous n'observons pas de tendance particulière sur ces deux graphiques. Cette conclusion serait dans la lignée des études menées par COVÉA qui avait conclu dans leur livre blanc de janvier 2022 que le risque tempête ne subirait pas d'augmentation en termes de fréquence ou de sévérité à horizon 2050 d'après l'ensemble des modèles qu'ils avaient sélectionnés.

GCM	RCM
CNRM-CERFACS-CNRM-CM5	SMHI-RCA4
ICHEC-EC-EARTH	DMI-HIRHAM5-V1
ICHEC-EC-EARTH	KNMI-RACMO22E
IPSL-CM5A-MR	KNMI-RACMO22E
MPI-M-MPI-ESM-LR	CLMcom-CCLM4-8-17
MOHC-HadGEM2-ES	CLMcom-CCLM4-8-17

TABLE 5.2 – Listes des couples GCM/RCM sélectionnés par COVÉA

Néanmoins après une analyse séparée des modèles, on remarque que certains présentent une tendance haussière que ça soit en terme d'augmentation du nombre de sinistres mais aussi en terme d'augmentation des pics de rafales ici défini comme une rafale dépassant 100 km/h :

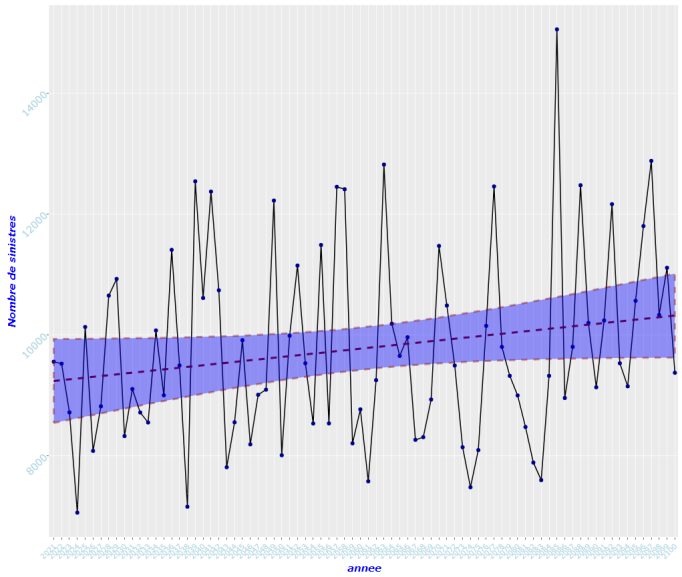

FIGURE 5.5 – Évolution du nombre de sinistres annuels pour le modèle couplé (CNRM-CERFACS-CNRM-CM5,KNMI-RACMO22E) sous le RCP 4.5

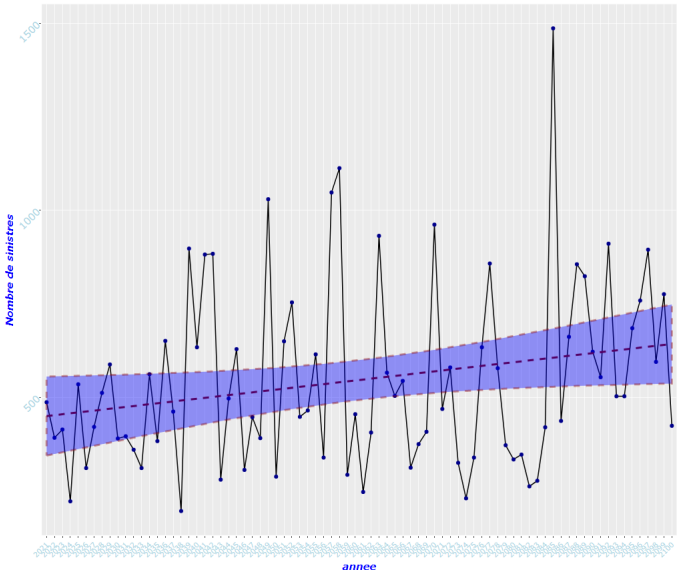

FIGURE 5.6 – Évolution du nombre de pics de rafales annuels pour le modèle couplé (CNRM-CERFACS-CNRM-CM5,KNMI-RACMO22E) sous le RCP 4.5

On retrouve cette tendance haussière à la fois sur les prédictions de sinistres de notre modèle et sur les projections des pics de rafales du modèle climatique en question, ce qui permet d'affirmer que cette tendance n'est pas propre à notre modélisation. Il est important de remarquer que ce modèle n'était pas présent

dans la série choisie par COVÉA c'est la raison pour laquelle nous pourrions avoir des conclusions qui diffèrent en fonction des modèles.

Toute la difficulté réside dans cette pluralité de modèles climatiques construits avec un paramétrage différent qui en sortie produisent des résultats différents. Nous avons présenté ci-dessus un modèle présentant une tendance haussière mais d'autres modèles ne présentent pas de tendance voire une tendance décroissante d'où l'absence de tendance sur la moyenne de la combinaison des différents résultats qui reste plus ou moins stable au fil des années.

5.3.3 Analyse de l'évolution annuelle des sinistres à l'échelle régionale et mensuelle sous une vision multi-modèles et multi-scénarios

Pour affiner notre analyse, nous sommes passés à une échelle plus réduite en observant la sinistralité projetée au niveau des régions et des mois. L'idée ici est de vérifier que le risque tempête n'est pas amené à se déplacer d'un point de vue géographique ou temporelle.

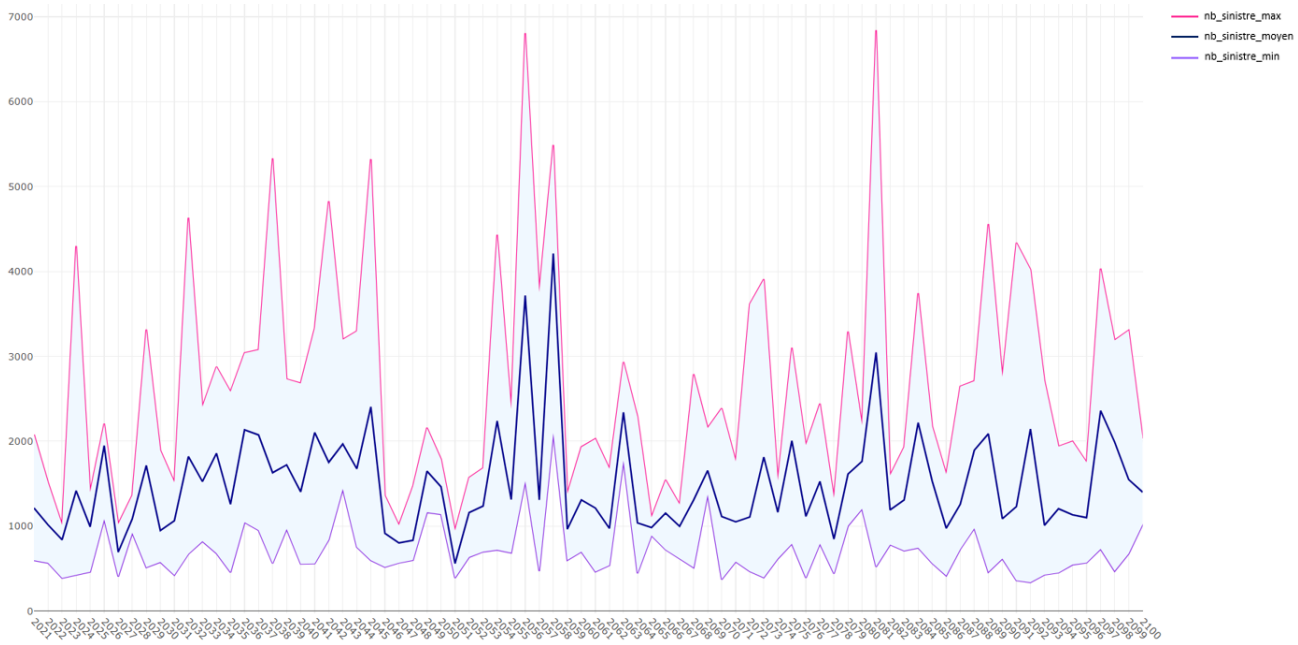

FIGURE 5.7 – Évolution du nombre de sinistres annuels sur le mois de Janvier en combinant l'ensemble des modèles pour le RCP 4.5

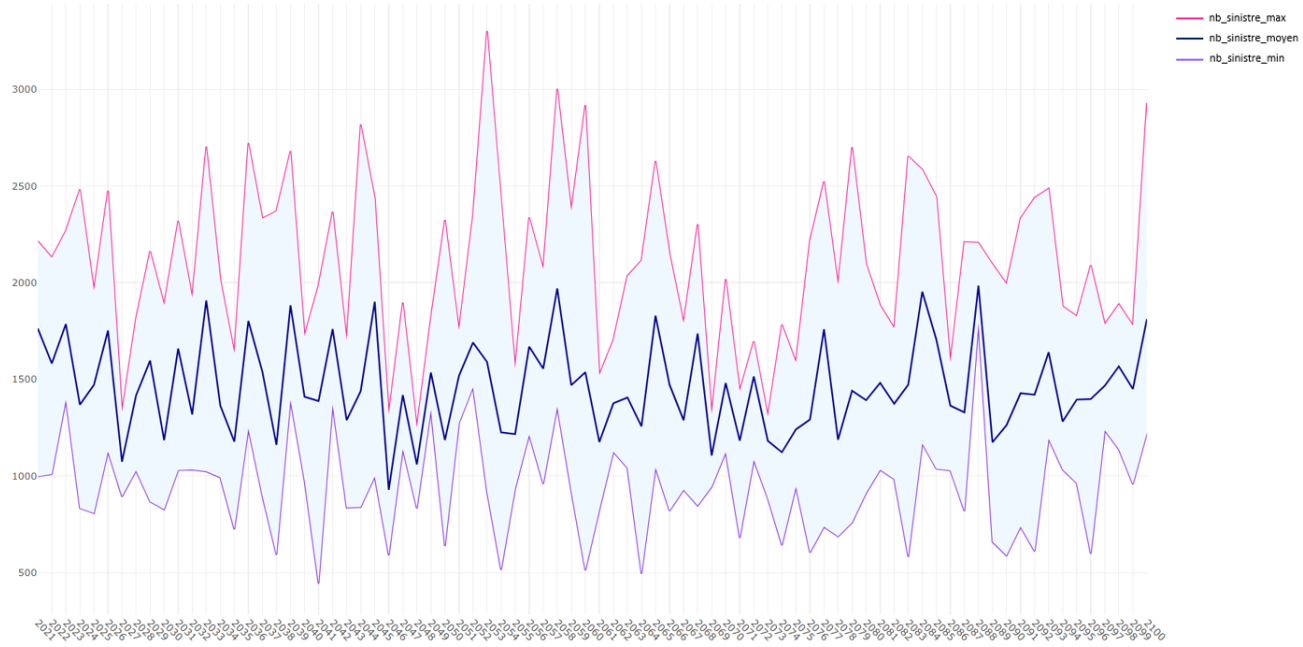

FIGURE 5.8 – Évolution du nombre de sinistres annuels sur la région de Bretagne en combinant l'ensemble des modèles pour le RCP 4.5

Nous affichons ci-dessus le graphique du mois de Janvier qui est le plus affecté par des tempêtes historiquement et la région de Bretagne qui est zone fréquemment affectée par des courants de vents importants. L'allure des graphiques ci-dessus est similaire pour l'ensemble des régions et l'ensemble des mois : il n'y a pas à première vue de tendance qui se dessine et la variabilité est importante au fil des années. Notons tout de même que la borne maximum pour la Bretagne semble légèrement haussière.

Ces conclusions rejoignent celles de Météo-France et de Covéa et indiquent que le nombre de tempêtes n'est a priori pas significativement lié au changement climatique. Nous poursuivrons néanmoins nos analyses par la suite en essayant d'aborder la question sous d'autres angles voire d'autres méthodes puisque comme nous l'avons évoqué ci-dessus, certains modèles climatiques mènent à des conclusions parfois opposées et il serait intéressant de comprendre l'origine de ces différences.

Conclusion

L'objectif de cette étude a été de présenter une démarche permettant d'intégrer les notions de changement climatique dans les modélisations actuarielles régulièrement employées. Ce type de modélisation est encore peu fréquent et commence à voir le jour notamment avec l'avènement d'une réglementation de plus en plus insistante concernant l'intégration de ces notions dans les processus ORSA.

Une difficulté régulièrement soulignée repose sur la mobilisation de données fiables et pertinentes, d'une part concernant le portefeuille assurantiel, d'autre part concernant les variables climatiques historiques et prospectives. Si les portefeuilles sont propres à chaque assureur et peuvent difficilement faire l'objet d'une amélioration substantielle, excepté via l'amélioration de processus internes de récolte et de fiabilisation des données, les données climatiques peuvent en revanche être le vecteur d'une source d'amélioration significative des modélisations effectuées.

Ces travaux ont donc tout d'abord permis de faire un bilan concernant les données actuellement disponibles et exploitables, tant d'un point de vue historique mais aussi prospectif. Les points forts et les points faibles relatifs aux objectifs de notre étude ont donc été identifiés pour obtenir in fine une source unique pour les données historiques à savoir ERA 5 et une autre source, elle aussi unique, pour les données futures à savoir EURO-CORDEX.

Concernant la modélisation, il nous a fallu réfléchir à une méthode pour construire notre base de données, qui soit à la fois cohérente pour la modélisation historique et pour l'objectif de projection de la sinistralité, c'est la raison pour laquelle nous avons effectué une modélisation à l'échelle du canton. Il existait, par ailleurs, un avantage relatif au nombre de stations météorologiques historiques dans ERA 5 quasiment similaire au nombre de cantons présents en France métropolitaine. Cette maille géographique est assez fine pour identifier la sinistralité tempête et permet de ne pas avoir une maille de résolution trop fine pour des données de projection qui se veulent initialement extraites de modèles climatiques globaux. Concernant le pas de temps, retenu au niveau de mois, il nous était imposé par la disponibilité des données fournies.

L'étape suivante a consisté à construire un modèle de fréquence de sinistre optimisé pour s'ajuster au mieux à la fréquence de sinistre mensuelle et régionale. Ces deux critères ont été sélectionnés pour comprendre l'évolution temporelle et géographique du risque au cours du siècle. Le pas de temps mensuel est assez naturel au regard de la méthode de construction de la base de données. Le pas de temps régional quant à lui se justifie par le cadre des projections qui nécessite une maille géographique importante pour analyser les résultats. Les modèles climatiques régionaux d'EURO-CORDEX ont une résolution de 12 km et notre modélisation est effectuée à une maille canton, mais l'analyse des résultats sur une maille plus importante permet de lisser les résultats et de limiter les incertitudes tant vis à vis de la modélisation que vis à vis des données climatologiques futures.

Une fois le calibrage du modèle effectué, nous avons alors effectué des projections de la fréquence de sinistre en combinant notre modèle avec les projections des paramètres météorologiques issus des modèles climatiques liés aux scénarios du GIEC. Pour limiter les incertitudes et la divergence des résultats entre les différents modèles et scénarios, nous avons effectué une approche multi-modèles et multi-scénarios qui se veut plus robuste.

Les premiers résultats présentés ici indiquent que la sinistralité tempête ne suivrait pas *a priori* de tendance à la hausse ou à la baisse liée au changement climatique néanmoins la survenance d'événements extrêmes pourrait être plus fréquente. Ces résultats sont en concordance avec ceux obtenus par les institutions phares comme Météo-France.

Notre modélisation et notre approche sont toutefois perfectibles et continueront d'être développées prochainement, nous souhaitons évoquer ici les limites actuelles de l'étude.

Pour commencer, nous ne disposions que du mois où s'est déroulé le sinistre. Le rapprochement météorologique consistant alors à associer au sinistre la rafale maximale du mois en question dans le canton où le bâtiment est instauré suivi des autres paramètres météorologiques peut être discuté notamment dans le cas où plusieurs tempêtes seraient survenues dans le même canton sur le même mois. La modélisation a ensuite été effectuée en catégorisant l'ensemble des variables. La présentation détaillée d'un modèle incluant des variables quantitatives ajustées par des polynômes ou des splines serait à envisager pour affirmer la pertinence de notre choix initial. Par ailleurs, nous nous sommes pour le moment concentré sur la modélisation de la fréquence mais l'étape suivante consisterait à effectuer une modélisation du coût pour éventuellement intégrer par la suite des scénarios de réassurance liés à la charge de sinistres.

Enfin, nous ne disposions que de 5 années d'historique de sinistres et il était alors difficile de comparer de manière quantitative notre nombre de sinistres historiques avec celui prédit pour les années futures. La possibilité d'une prépondérance de sinistres dans certaines zones, relative non pas à la sinistralité en elle-même, mais à notre période historique restreinte, pourrait fausser nos coefficients d'augmentation ou de diminution de nombre de sinistres. Cet effet a été observé dans nos analyses récentes. Une idée pour contourner ce problème serait de récupérer un portefeuille avec un historique de sinistre plus important.

Annexe

Graphiques des différentes variables du modèle final pour l'ajustement du GLM

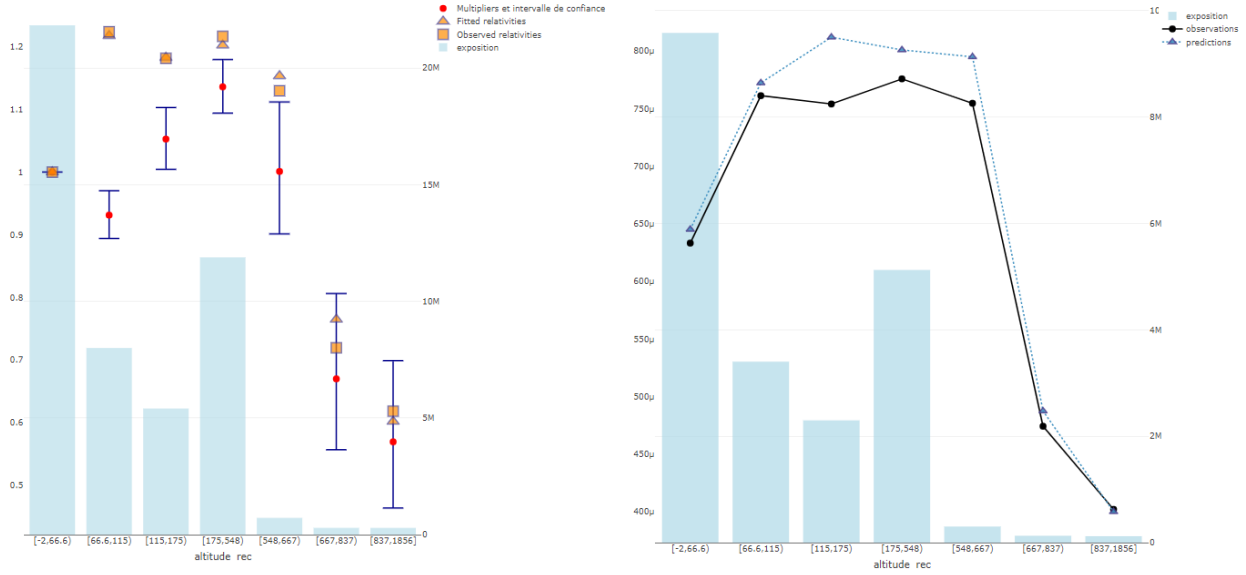

FIGURE 9 – Graphiques des coefficients et des relativités associées ainsi qu'une comparaison entre les observations et les prédictions pour la variable *Altitude*

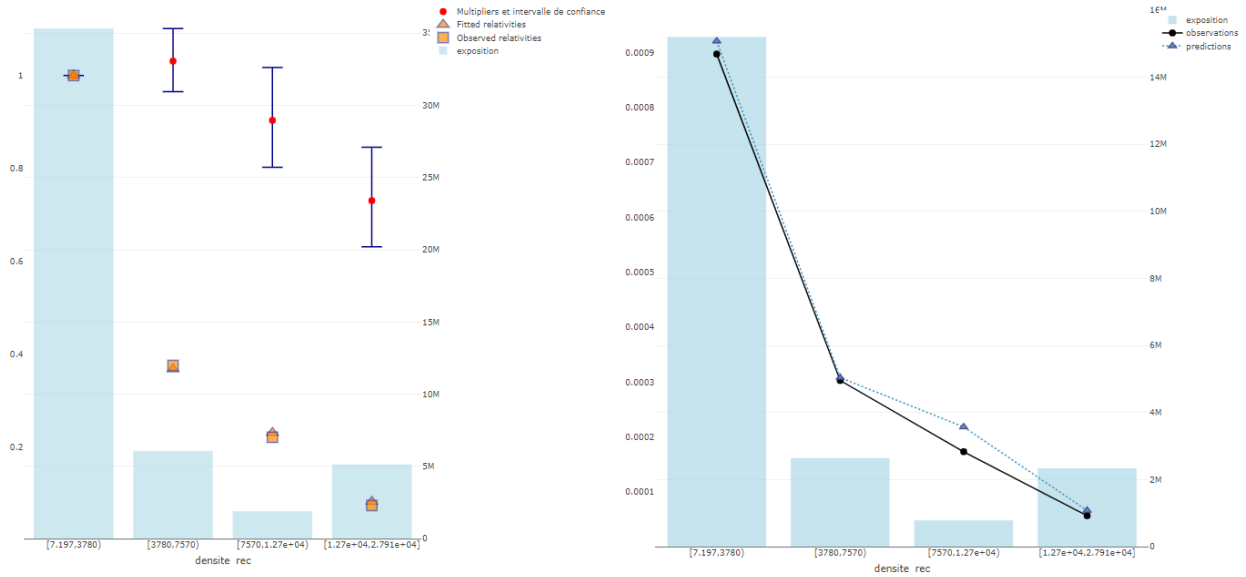

FIGURE 10 – Graphiques des coefficients et des relativités associées ainsi qu'une comparaison entre les observations et les prédictions pour la variable *Densite*

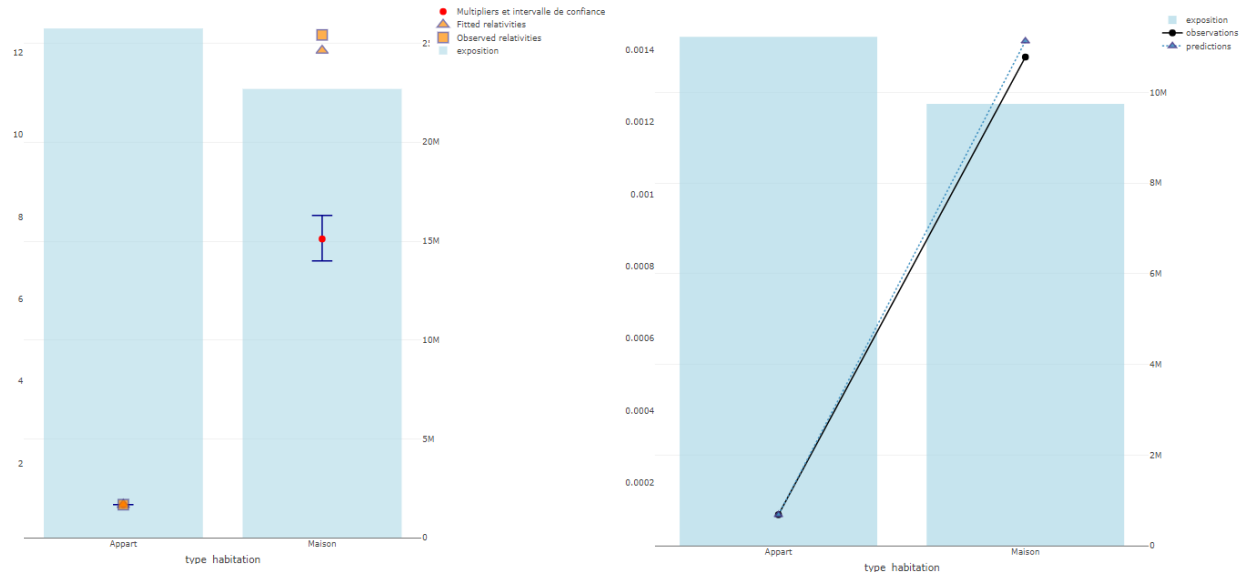

FIGURE 11 – Graphiques des coefficients et des relativités associées ainsi qu'une comparaison entre les observations et les prédictions pour la variable *TypeHabitation*

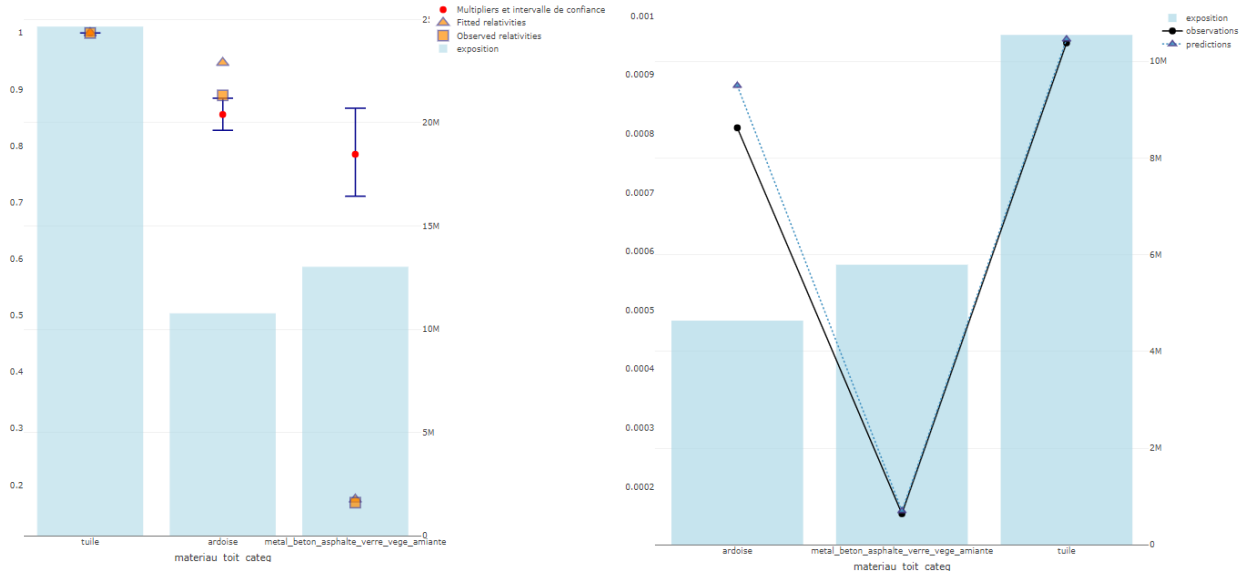

FIGURE 12 – Graphiques des coefficients et des relativités associées ainsi qu'une comparaison entre les observations et les prédictions pour la variable *MateriauToit*

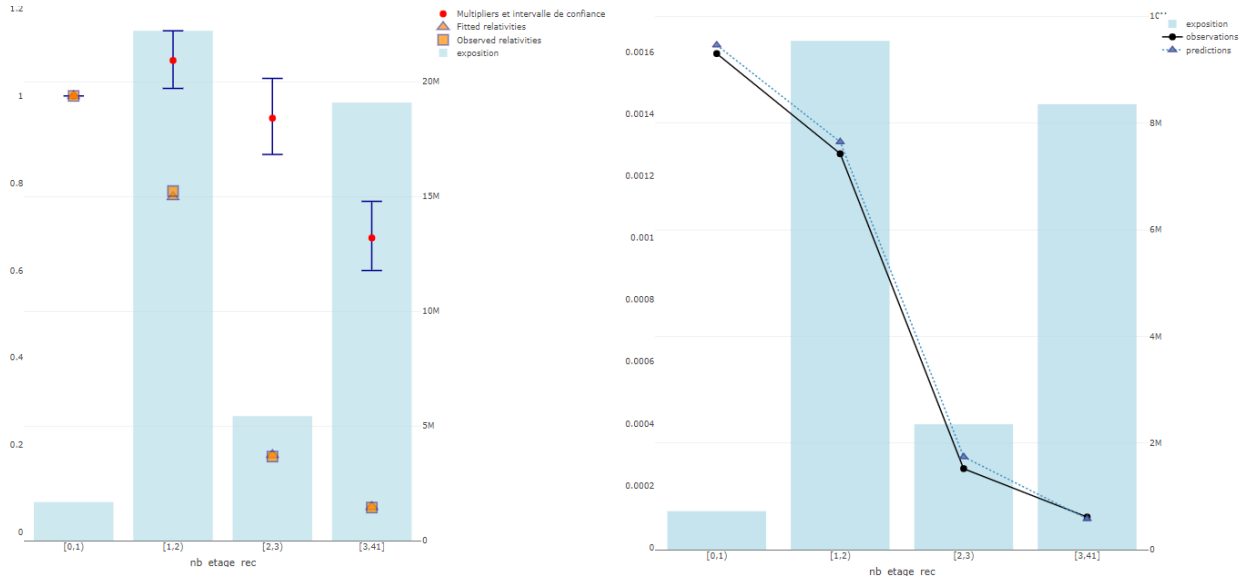

FIGURE 13 – Graphiques des coefficients et des relativités associées ainsi qu'une comparaison entre les observations et les prédictions pour la variable *NbEtage*

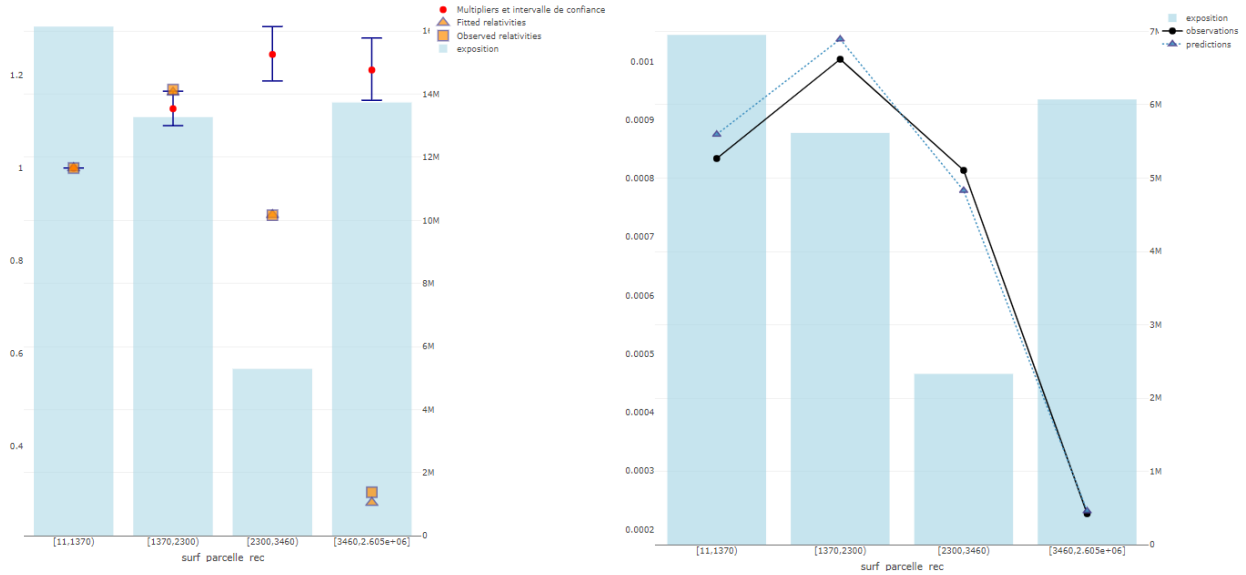

FIGURE 14 – Graphiques des coefficients et des relativités associées ainsi qu'une comparaison entre les observations et les prédictions pour la variable *ParcSurf*

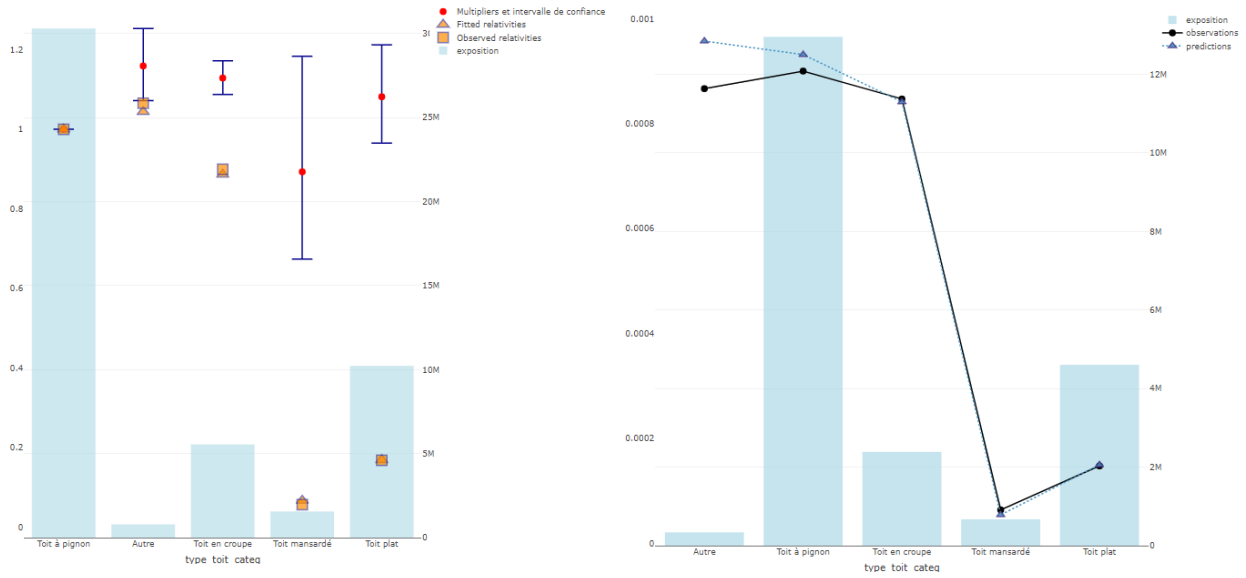

FIGURE 15 – Graphiques des coefficients et des relativités associées ainsi qu'une comparaison entre les observations et les prédictions pour la variable *TypeToit*

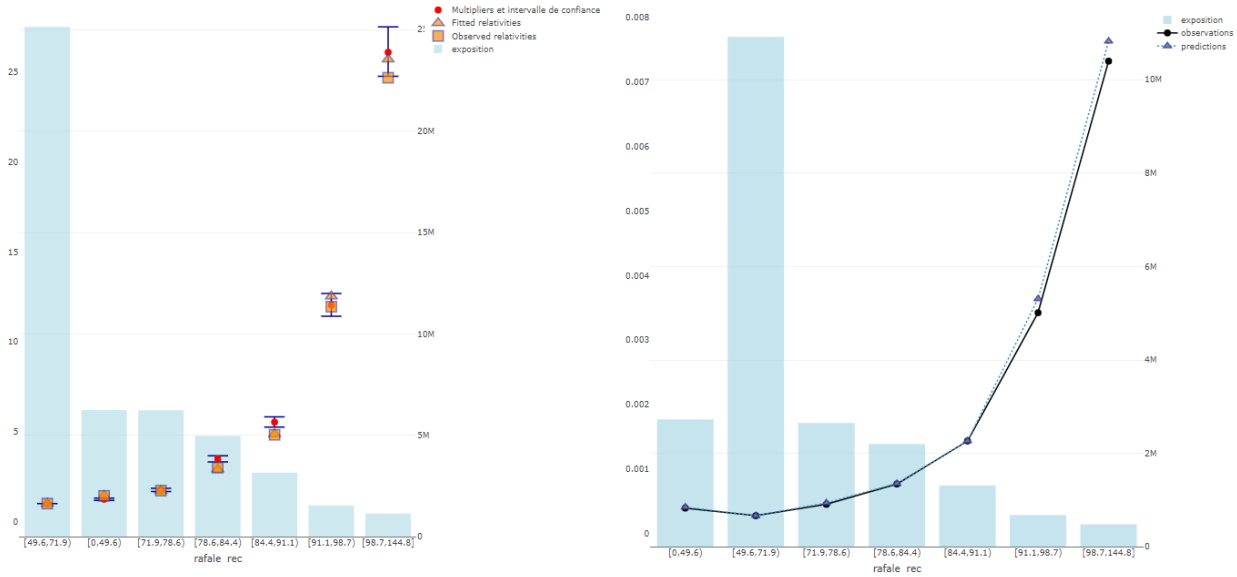

FIGURE 16 – Graphiques des coefficients et des relativités associées ainsi qu'une comparaison entre les observations et les prédictions pour la variable *Rafale*

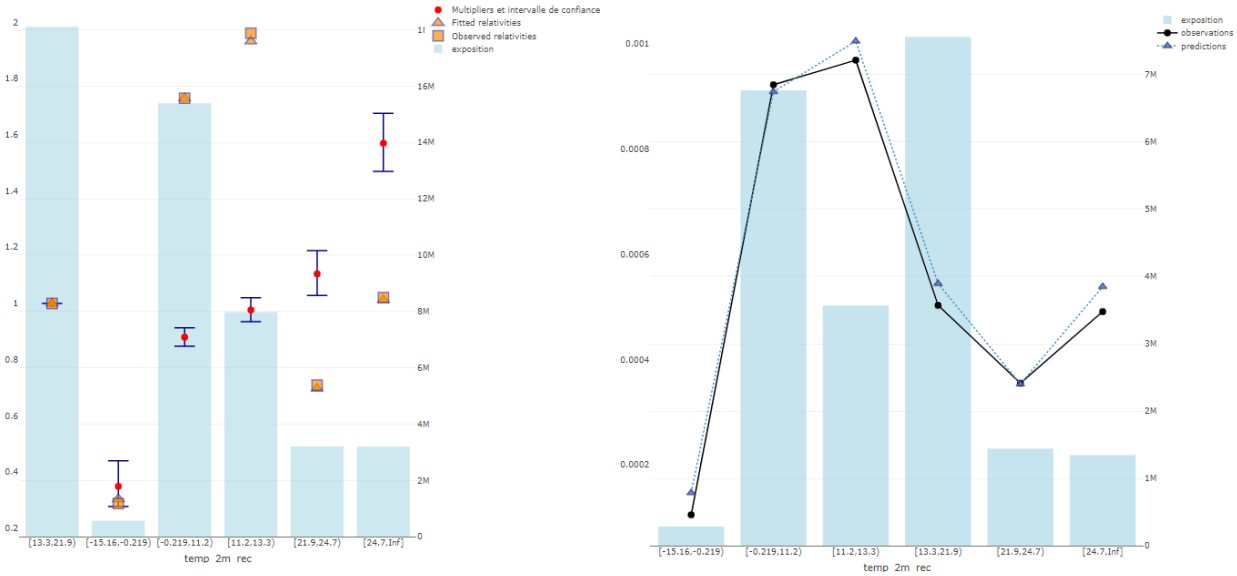

FIGURE 17 – Graphiques des coefficients et des relativités associées ainsi qu'une comparaison entre les observations et les prédictions pour la variable *Temp2m*

Bibliographie

[1] MINISTÈRE DE L'ÉCOLOGIE ET DU DÉVELOPPEMENT DURABLE (2002). *Les tempêtes, dossier d'information.*

[2] DENUIT M. et CHARPENTIER A. (2004). *Mathématiques de l'assurance non-vie - Tome I, Principes fondamentaux de la théorie du risque. Economica.*

[3] FESER F. et al. (2011). *Regional Climate Models Add Value to Global Model Data - A Review and Selected Examples. Bulletin of the American Meteorological Society 92 : 1181-1192.*

[4] PINTO J.G. et al. (2012). *Loss potentials associated with European windstorms under future climate conditions. Institute for Geophysics and Meteorology, University of Cologne.*

[5] DI LUCA A. et al. (2013). *Potential for added value in temperature simulated by high-resolution nested RCMs in present climate and in the climate change signal. Climate Dynamics 40, 443-464.*

[6] BAN N. et al. (2014). *Evaluation of the convection-resolving regional climate modeling approach in decade-long simulations. Journal of Geophysical Research : Atmospheres 119.*

[7] CHARPENTIER A. (2014). *Computational Actuarial Science with R. CRC Press.*

[8] JACOB D. et al. (2014). *EURO-CORDEX : new high-resolution climate change projections for European impact research. Regional Environmental Change 14 : 563-578.*

[9] MORNET A. (2015). *Contributions à l'évaluation des risques en assurance tempête et automobile. Thèse de doctorat, ISFA et Université Claude Bernard Lyon I.*

[10] DIALLO A. O. (2017). *Inférence statistique dans des modèles de comptage à inflation de zéros. Applications en économie de la santé. Thèse de doctorat, INSA Rennes et Université Gaston Berger de Saint-Louis.*

[11] JAMES G. et al. (2017). *An Introduction to Statistical Learning with Applications in R, 2e édition, Springer.*

[12] BEDI N. (2018). *Modélisation du risque de tempête en France Métropolitaine. Mémoire d'actuariat, ISUP.*

[13] ACPR (2019). *Les assureurs français face au risque de changement climatique, Analyses et synthèses, n°102.*

[14] LANGEVIN N. (2019). *Modélisation de la sinistralité tempête, apport de l'Open Data et du Machine Learning. Mémoire d'actuariat, ENSAE.*

[15] ACPR (2020). *Scénarios et hypothèses principales de l'exercice pilote climatique, Direction d'études et d'analyse des risques.*

[16] ALLEN T. et al. (2020). *Climate-Related Scenarios for Financial Stability Assessment : an Application to France. Rapport technique Banque de France.*

[17] CHRISTENSEN O.B. et al. (2020). *CORDEX Archive Design, Version 3.2.*

[18] FRANCE ASSUREURS (2020). *Impact du changement climatique sur l'assurance à l'horizon 2050.*

[19] BENESTAD R. et al. (2021) *Guidance for EURO-CORDEX climate projections data use.*

[20] EIOPA (2021). *Consultation paper on Application guidance on running climate change materiality assessment and using climate change scenarios in the ORSA. Rapport technique.*

[21] GIEC (2021). *Sixième rapport d'évaluation (AR6). Premier Groupe de travail (WG1).*

[22] SHAO Z. (2021). *Impact du changement climatique sur le risque de grêle pour un portefeuille d'assurance français à horizon 2050. Mémoire d'actuariat, EURIA.*

[23] SOUBEYROUX J-M. et al. (2021). *Les nouvelles projections climatiques de référence DRIAS 2020 pour la métropole.*

[24] ACPR (2022). *La gouvernance des risques liés au changement climatique dans le secteur de l'assurance, Analyse et Synthèse. N°102.*

[25] BERARD J. (2022). *Modèles Linéaires Généralisés. Cours dispensé au DUAS (2ème année).*

[26] COVEA (2022). *Changement climatique & Assurance : Quelles conséquences sur la sinistralité à horizon 2050 ?*

[27] EIOPA (2022). *Methodological principles of insurance stress testing - climate change component. EIOPA-BOS-21/579.*

Table des figures

Liste des tableaux