Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Corentin BOYEAU

Titre du mémoire : Évaluation de l'impact du changement climatique sur le risque inondation en France métropolitaine

Confidentialité : ☑ NON ☐OUI (Durée : ☐ 1 an ☐2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus.

Membres présents du jury de la filière : Signature :	Entreprise : Nom : GROUPAMA Signature :
	Directeur de mémoire en entreprise Nom : Marc BAGARRY Signature :
Membres présents du jury de l'Institut des Actuaires : Signature :	Invité : Nom : Signature :
	Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels (après expiration de l'éventuel délai de confidentialité) Signature du responsable entreprise : Signature du candidat :

Résumé

Mots clés : inondations, changement climatique, modèles CAT, GLM, machine learning, taux de destruction, données externes

Les inondations dévastatrices de juillet 2021 qui ont affecté l'Europe de l'Ouest avec un bilan humain de plus de 200 morts et près de 11 milliards d'euros de dommages assurés nous rappellent le danger de telles catastrophes et nous amènent à nous interroger sur l'évolution de leur sévérité face au changement climatique. Le but de ce mémoire est d'évaluer la sensibilité des territoires au risque inondation en développant une vision suffisamment robuste permettant d'identifier les risques les plus vulnérables que ce soit aussi bien au climat actuel qu'au climat futur.

Pour acquérir cette robustesse dans les modélisations, le seul historique sinistre ne suffit pas et il a été nécessaire de passer par des modèles catastrophes naturelles basés aussi bien sur l'historique que sur la simulation d'évènements fictifs. Cependant, la création de tels modèles en interne étant relativement complexe, il a fallu utiliser ceux provenant d'agences de modélisations externes. Une telle approche implique de n'avoir aucune mainmise sur le modèle et empêche toute étude d'impact du changement climatique.

C'est pourquoi il a été décidé de réaliser un modèle en interne basé sur des modèles linéaires généralisés, ainsi que sur des méthodes de *machine learning*, en s'entraînant directement sur les résultats du modèle catastrophe naturelle fourni par le modélisateur. Ces modélisations internes ont préalablement nécessité un grand travail d'enrichissement du portefeuille avec des données externes potentiellement explicatives du risque inondation, telles que la distance au cours d'eau, l'altitude, les précipitations, l'imperméabilité du sol ou encore le coefficient de pente.

Les indicateurs de précipitations renseignés en entrée de notre modèle nécessitent à la fois de disposer d'une vision historique, mais également de projections climatiques futures. De cette manière, il est ainsi possible d'étudier à la fois l'état historique du risque inondation en utilisant en entrée du modèle les précipitations actuelles puis d'étudier l'évolution de la sinistralité en faisant varier les indicateurs de précipitations selon différents horizons de temps et différents scénarios de changement climatique.

Abstract

Keywords: floods, climate change, CAT model, GLM, machine learning, destruction rate, external data

The devastating floods of July 2021 that affected Western Europe with more than 200 deaths and nearly 11 billion euros in insured damages reminds us of the dangers of such disasters and leads us to question the evolution of their severity with respect to climate change. The aim of this thesis is to evaluate the sensitivity of territories to flood risk by developing a sufficiently robust vision to identify the most vulnerable risks of both the current and future climate.

To acquire this robustness in the modeling, the loss history alone is not sufficient, and it was necessary to use natural catastrophe models based on both historical and simulated fictitious events. However, as the creation of such in-house models is relatively complex, it was necessary to use models from external modeling agencies. With such an approach, we have no control over the model itself, which is not suitable for studying the impact of climate change.

Therefore, it was decided to build an in-house model based on generalized linear models, as well as machine learning methods, by directly training these models on the results of the CAT model provided by the modeler. These in-house models required extensive work to enrich the portfolio with external data, such as distance to the river, altitude, precipitation, soil impermeability or the slope coefficient, that could potentially explain the flood risk.

The precipitation indicators entered in our model require both a historical vision and future climate projections. As such, it is possible to study both the historical state of flood risk by using current precipitation as an input to the model, then to study the evolution of losses by varying the precipitation indicators according to different time horizons and different climate change scenarios.

Note de synthèse

Contexte et démarche de l'étude

En tant que premier risque naturel en France, les inondations constituent une composante majeure de la sinistralité des catastrophes naturelles. Si leur prise en charge est effectuée par le biais d'une surprime identique sur l'ensemble des contrats d'assurance dommages (+12% pour les habitations, +6% pour les véhicules) via le régime spécifique d'indemnisation des catastrophes naturelles, leur modélisation n'est cependant pas à négliger et le changement climatique nous amène encore davantage à surveiller ces risques. En effet, alors que le sixième rapport du GIEC (Groupe d'experts Intergouvernemental sur l'Évolution du Climat) devrait paraitre en 2022, les précédents rapports ont déjà démontré le rôle que joue l'Homme dans ces évolutions et la hausse globale des températures d'ici 2100 ne fait plus aucun doute.

Il est ainsi primordial de mettre en place des méthodes permettant d'évaluer l'impact de cette évolution du climat sur la sévérité des évènements climatiques et donc sur la sinistralité qui en découle. Afin de commencer à sensibiliser les acteurs de l'assurance à cette question, l'ACPR (Autorité de contrôle prudentiel et de résolution) a récemment mis en place l'exercice « pilote climatique » dont l'objectif était d'évaluer les risques associés au changement climatique et auquel Groupama a participé. C'est dans ce contexte que l'on souhaitait approfondir ces travaux et notamment ceux sur la partie inondation, pour laquelle Groupama avait déjà amorcé un projet fin 2019 afin de mieux appréhender ce risque.

Dans la continuité de ce projet, l'objectif est donc à la fois d'être en mesure d'évaluer la vulnérabilité de chacun de nos sites assurés au climat présent, mais également d'y intégrer une vision au climat futur. C'est dans cette optique que nous avons cherché à mettre en place un modèle inondation et dont la démarche de construction est détaillée sur le schéma ci-dessous :

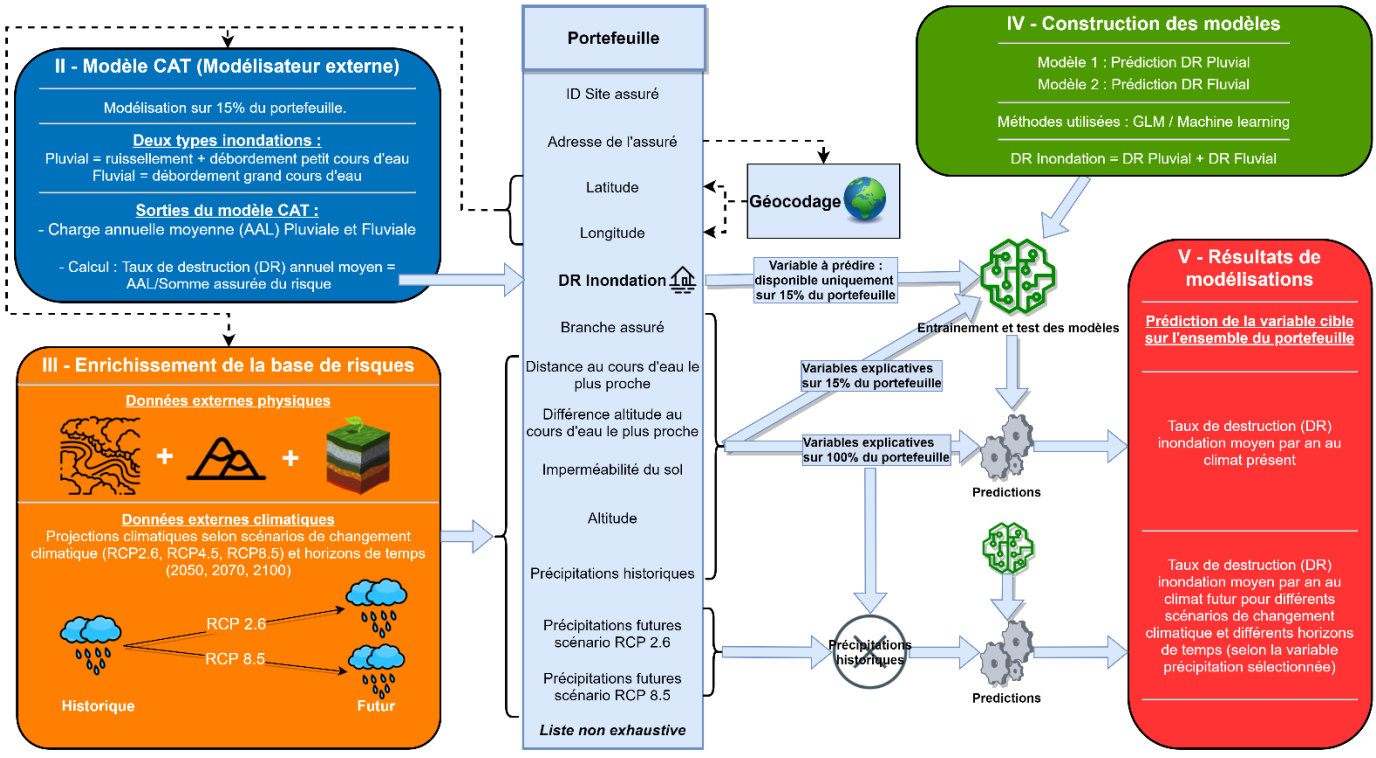

Figure 1 - Récapitulatif de la démarche du mémoire

La modélisation des catastrophes naturelles

Lorsqu'il s'agit d'évaluer les pertes financières probables que peut représenter un risque, l'approche fréquence-coût est souvent retenue, notamment dans le cadre de la tarification des risques de masse telle qu'en auto ou en habitation. Cependant pour la modélisation des catastrophes naturelles cette approche n'est pas adaptée, et ce notamment, car la faible profondeur d'historique dont les assureurs disposent ne permet pas d'obtenir une distribution de perte assez robuste pour représenter fidèlement la diversité des évènements climatiques et le caractère exceptionnel de certains.

C'est dans ce contexte qu'interviennent les modèles catastrophes naturelles (modèles CAT) qui s'appuient notamment sur la modélisation physique des évènements, puis sur la traduction de ces évènements en pertes financières. Cela permet de compléter la vision historique avec des scénarios inondations extrêmes qui n'ont pas été nécessairement observés dans le passé proche et qui ne sont donc pas présents dans notre base sinistre. La modélisation physique des évènements étant très complexe, Groupama, comme une majorité d'assureurs et de réassureurs, achète des résultats de modélisations auprès de modélisateurs extérieurs.

Ces modèles permettent pour chacun de nos sites assurés de déduire une perte annuelle moyenne (AAL) pour le risque inondation (hors submersions marines), en apportant un détail supplémentaire selon le type d'inondations, fluvial ou pluvial, que l'on définit de la manière suivante :

- Le risque fluvial : les principaux cours d'eau débordent ou sortent de leur lit et inondent les zones environnantes.
- Le risque pluvial : des précipitations excessives et importantes provoquent une inondation indépendamment du débordement d'un plan d'eau majeur (inondation par ruissellement). Le débordement des rivières et ruisseaux mineurs est également inclus dans ce risque.

Dans notre étude, on ne modélisera pas directement la perte annuelle moyenne, mais plutôt le taux de destruction annuel moyen défini comme :

$$Taux\ de\ destruction\ annuel\ moyen = \frac{Charge\ annuelle\ moyenne\ (AAL)}{Somme\ assurée}$$

L'utilisation seule de ce modèle n'est cependant pas suffisante, et ce pour deux raisons principales. Tout d'abord, les modélisations ont été effectuées sur seulement 15% du portefeuille, on ne dispose donc pas du taux de destruction inondation sur l'ensemble de nos risques. De plus, nous n'avons aucun accès au modèle et il nous est donc impossible de le modifier afin d'évaluer l'impact du changement climatique.

C'est pourquoi nous avons choisi de réaliser un modèle en interne basé sur des modèles linéaires généralisés, ainsi que sur des méthodes de *machine learning*, en s'entraînant directement sur les taux de destruction obtenus grâce au modèle CAT. Ainsi, on pourra généraliser les taux de destruction sur l'ensemble du portefeuille et l'on aura la mainmise sur le modèle afin de faire des études de sensibilité face au changement climatique.

Mise en place du modèle : enrichissement de la base de risques

Avant de commencer les modélisations, l'objectif est de récupérer des variables physiques et climatiques qui pourraient nous aider à caractériser le risque inondation. Nos modèles seront ainsi alimentés d'un côté par des variables déjà présentes dans nos bases telles que la branche assurée, la somme assurée ou le nombre d'étages, et de l'autre par des variables enrichies dans cette partie et qui seront calculées à partir des coordonnées géographiques du risque. Cette partie nous a ainsi permis de récupérer un total de 29 variables (détaillées en annexe 4) et qui peuvent être toutes classées dans les catégories suivantes :

- La distance au cours d'eau le plus proche
- L'altitude du site assuré
- La différence d'altitude entre le site assuré et le cours d'eau le plus proche
- L'imperméabilité du sol
- Un indicateur d'accumulation de l'eau dans la zone
- Le coefficient de pente au niveau du risque
- L'indice d'humidité topographique
- Des indicateurs de précipitations simulés aux conditions historiques
- Des indicateurs de précipitations simulés aux conditions futures (selon différents scénarios de changement climatique et différentes périodes)

Les indicateurs de précipitations sont obtenus à partir de modèles climatiques régionaux (RCM) produits dans le cadre du projet CORDEX *(**CO**ordinated **R**egional climate **D**ownscaling **EX**periment)*. Ces modèles reposent sur des systèmes d'équations basés sur des lois de la physique, de la chimie et de la dynamique des fluides afin de reproduire au mieux le comportement du climat. Ils permettent d'obtenir des simulations climatiques aux conditions historiques sur la période 1976-2005 (période disponible la plus récente dans les simulations historiques CORDEX), mais également aux conditions futures sur les périodes 2021-2050, 2041-2070 et 2071-2100 et selon trois scénarios d'émissions de gaz à effet de serre dits RCP (pour *Representative Concentration Pathway*). Le scénario RCP 2.6 correspond au scénario le plus optimiste avec une baisse progressive des émissions à partir de 2050, le scénario RCP 8.5 est le plus pessimiste et correspond à une poursuite des émissions au rythme actuel et le scénario RCP 4.5 correspond à l'entre-deux avec une stagnation des émissions avant la fin du XXIe siècle.

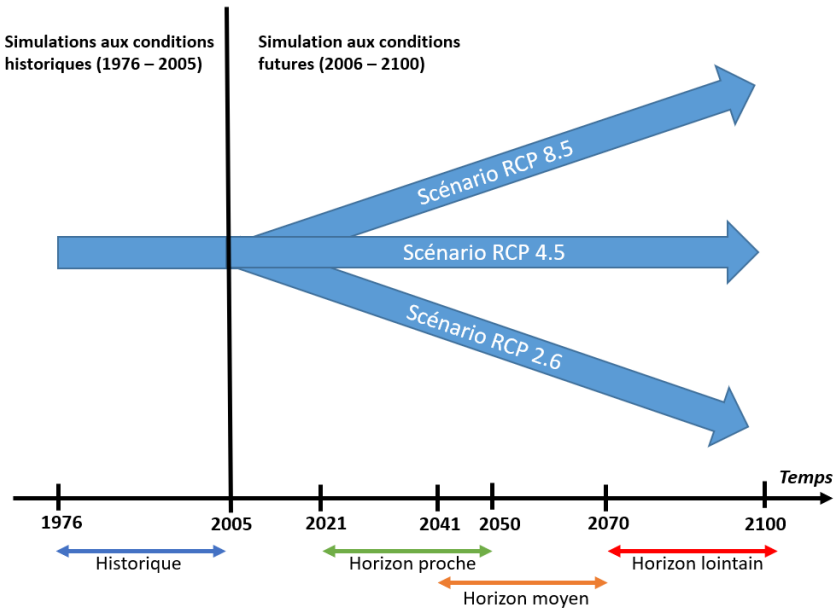

Figure 2 - Structure des modélisations : période et scénarios considérés

Construction d'un modèle prédictif de la sinistralité inondation

Maintenant que nous disposons d'une base de modélisation enrichie de variables climatiques et physiques, l'objectif est désormais d'établir un lien entre ces variables et le taux de destruction issue du modèle CAT. Le modèle entraîné utilisera les indicateurs de précipitations aux conditions historiques. Cependant lors de l'étape de prédiction dans la partie suivante, on pourra utiliser les indicateurs de précipitations aux conditions futures pour prédire le taux de destruction inondation qui en découle. On pourra ainsi évaluer l'évolution de la vulnérabilité en comparant le taux de destruction obtenu à partir des précipitations historiques avec celui obtenu à partir des précipitations futures (selon la période et le scénario RCP considérés).

Les modélisations se sont faites en deux étapes. Une première modélisation pour la prédiction du taux de destruction annuel moyen associé au risque pluvial et une deuxième modélisation pour le risque fluvial. Les deux risques étant considérés indépendants par le modélisateur, le taux de destruction annuel moyen inondation totale peut être simplement obtenu en faisant la somme de celui des deux risques en question.

Afin d'éviter les problèmes de convergence dans nos modèles, on effectue une première sélection de variables à partir de diverses mesures (Pearson, Spearman et Kendall) nous permettant de sélectionner les variables les plus pertinentes et d'éliminer celles étant trop corrélées entre elles. Les variables sélectionnées pour le risque pluvial sont la distance au cours d'eau le plus proche, le quantile à 99% des précipitations journalières, l'indice d'humidité topographique, ainsi que le coefficient de Manning. Pour le risque fluvial, on sélectionne la distance à la rivière large la plus proche, la différence d'altitude à la rivière large la plus proche, la distance à la rivière et l'imperméabilité du sol. De plus, pour ces deux modèles, on intègre également des variables présentes dans le portefeuille et qui ont été utilisées directement par le modélisateur, à savoir, la branche assurée, le taux d'engagement contenu (part que représente le contenu dans le total de la somme assurée), le taux d'engagement perte d'exploitation ainsi que le nombre d'étages (pour le portefeuille immeuble).

Par la suite, la sélection du modèle pour le risque pluvial s'est faite par minimisation de la racine de l'erreur quadratique moyenne (RMSE) obtenue par validation croisée sur l'ensemble d'entraînement. Il est apparu que parmi les quatre modèles considérés, à savoir le GLM gamma, la régression Bêta, le *gradient boosting* et le modèle de forêt aléatoire, c'est ce dernier qui a permis de minimiser le RMSE et qui a donc été sélectionné.

Concernant les modélisations sur le risque fluvial, celles-ci ont été plus complexes. En effet, il est apparu sur la base d'entraînement que 95% des sites assurés avaient un taux de destruction moyen nul, c'est-à-dire qu'ils n'étaient pas exposés à ce risque. Une première étape de classification a donc été nécessaire pour identifier les sites assurés exposés à ce risque de ceux qui ne le sont pas. Le modèle optimal a été sélectionné par maximisation du score F1. Parmi les trois modèles essayés, à savoir le GLM binomial, le *gradient boosting* et le modèle de forêt aléatoire, c'est encore ce dernier qui a été sélectionné. Ensuite, un modèle de régression est appliqué sur les risques qui sont exposés au risque fluvial selon le modèle de classification. La sélection est faite de la même manière que pour le risque pluvial et c'est le GLM gamma qui permet cette fois-ci de minimiser l'erreur quadratique moyenne.

D'autres indicateurs ont également été considérés pour évaluer nos modèles, tels que la corrélation de Spearman entre la prédiction effectuée et les résultats fournis par notre modélisateur. Il apparait que la corrélation obtenue pour le risque pluvial est d'environ 33%, contre 50% pour le risque fluvial. Le modèle ainsi développé présente trois avantages principaux par rapport aux modèles CAT, à savoir l'autonomie, la transparence et la facilité d'exécution, ce qui rend son utilisation idéale pour la gestion interne du risque inondation.

Exploitation des modèles et impact du changement climatique

Au climat présent, la prédiction des modèles sur l'ensemble du portefeuille a fait apparaitre qu'il s'agissait principalement des régions du sud et sud-est de la France qui sont les plus vulnérables au risque inondation.

Au climat futur, il est tout d'abord intéressant de s'intéresser à l'évolution de l'aléa climatique, avant de remplacer les précipitations historiques par les précipitations futures dans notre modèle. L'indicateur considéré est le quantile à 99% des précipitations journalières qui est l'une des variables explicatives du modèle pluvial, c'est donc cette variable qui va être modifiée pour étudier l'impact du changement climatique. Parmi les 12 modèles CORDEX utilisés, on présente ci-dessous les résultats obtenus en sélectionnant le quantile à 95% de ces modèles afin d'établir une vision la plus prudente possible. On constate que la mise en place d'actions politiques fortes menées dès à présent (scénario RCP 2.6) aurait peu d'impacts sur l'évolution des précipitations extrêmes à l'horizon 2050. C'est principalement à l'horizon 2100 que la différence entre les scénarios est la plus flagrante.

VUE "PESSIMISTE" (QUANTILE 95%)	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100
	Auvergne-Rhône-Alpes	10%	12%	12%	10%	14%	17%	12%	14%	22%
	Bourgogne-Franche-Comté	10%	10%	12%	10%	15%	19%	14%	18%	29%
	Bretagne	9%	9%	9%	10%	11%	17%	10%	19%	32%
	Centre-Val de Loire	9%	11%	11%	9%	13%	15%	11%	15%	26%
	Corse	14%	15%	13%	9%	14%	15%	14%	15%	12%
	Grand Est	9%	10%	11%	11%	13%	19%	14%	20%	33%
	Hauts-de-France	9%	11%	11%	10%	12%	14%	11%	18%	34%
	Île-de-France	8%	11%	12%	9%	11%	15%	11%	17%	27%
	Normandie	9%	10%	10%	11%	12%	16%	12%	20%	32%
	Nouvelle-Aquitaine	10%	11%	11%	9%	11%	13%	9%	12%	18%
	Occitanie	11%	11%	14%	9%	11%	15%	10%	12%	14%
	Pays de la Loire	9%	10%	10%	8%	11%	15%	9%	17%	27%
	Provence-Alpes-Côte d'Azur	7%	12%	11%	8%	9%	15%	9%	12%	14%
	Total général	**10%**	**11%**	**12%**	**9%**	**12%**	**16%**	**11%**	**15%**	**23%**

Tableau 1 - Projections d'évolution des précipitations extrêmes par région par rapport à la période historique

Notre modèle nous permet désormais de traduire cette évolution de l'aléa climatique en une évolution de la sinistralité. Remarquons cependant que le modèle fluvial ne dépend étonnamment d'aucune variable climatique, l'impact du changement climatique sur ce risque sera donc considéré nul et l'impact sur le risque inondation total pourrait s'avérer sous-estimé. Seul le risque pluvial est donc impacté par le changement du climat selon nos modèles. Finalement, les résultats obtenus nous indiquent une augmentation de la sinistralité inondation sur notre portefeuille d'ici 2050 de +11% pour le scénario RCP 8.5. En 2021, la FFA a fait la même conclusion avec une évolution de +11% d'ici 2050. En 2018 la CCR établissait une vision plus pessimiste avec +38% d'augmentation.

Scénario changement climatique	Source étude	Année Publication	Horizon futur modélisé	Évolution[1] de la sinistralité globale inondation[2] à l'horizon 2XXX	Détails par types d'inondations	Évolution de la sinistralité à l'horizon 2XXX
RCP 8.5	Groupama[3]	2022	2100	+18%	Pluvial[4]	+34%
					Fluvial[4]	NA[5]
	Groupama	2022	2050	+11%	Pluvial	+21%
					Fluvial	NA
	FFA	2021	2050	+11%	Pas de distinction entre types d'inondations	
	CCR	2018	2050	+38%	Ruissellement	+50%
					Débordement	+24%
RCP 4.5	Groupama	2022	2100	+13%	Pluvial	+26%
					Fluvial	NA
	Groupama	2022	2050	+9%	Pluvial	+18%
					Fluvial	NA
	CCR	2015	2050	+20%	Pas de distinction entre types d'inondations	
RCP 2.6	Groupama	2022	2100	+9%	Pluvial	+18%
					Fluvial	NA
	Groupama	2022	2050	+9%	Pluvial	+18%
					Fluvial	NA

(1) Évolution dû à la variation de l'aléa climatique (sans prise en compte de l'inflation ni de l'évolution future de la répartition des risques sur le territoire)
(2) Inondations hors submersions marines
(3) Par prudence on considérera l'évolution globale donnée par le quantile à 95% de l'étude Groupama
(4) Risque Pluvial = Ruissellement et Débordement de ruisseaux/rivières - Risque Fluvial = Débordement de fleuves
(5) Pas de résultats concluants sur l'évolution du risque fluvial : on considère une évolution nulle pour le calcul de l'évolution globale inondation

Tableau 2 - Impact du changement climatique sur le risque inondation selon différentes études

Remerciements

Je souhaite adresser mes remerciements à l'ensemble des personnes ayant contribué, de près ou de loin, à la réalisation de ce mémoire.

Tout d'abord, à mon tuteur, Marc Bagarry, pour m'avoir fait confiance dans la réalisation de mon alternance au sein de la Direction Réassurance de Groupama, ce qui m'a ainsi permis de travailler sur des sujets passionnants que sont les catastrophes naturelles et le changement climatique. Je tiens également à le remercier pour son expertise et toutes les brillantes idées qu'il a pu apporter tout au long de la construction de cette étude.

Plus généralement, je souhaite remercier l'ensemble de l'équipe pour m'avoir intégré si rapidement malgré la distance imposée par le confinement. Je tiens particulièrement à remercier Vincent Noel pour son aide et ses précieux conseils, Pascal Deramez pour son soutien sur nos différentes missions communes, Michael Vlaskovski pour ses corrections et Nasrine Mbaraka pour sa bonne humeur tout au long de mon année passée au sein du service.

Je souhaite aussi remercier mon tuteur académique, Nicolas Bousquet, pour ses différents conseils dans la mise en place de l'étude.

Un grand merci également à l'ensemble de mes relecteurs pour leurs retours pertinents, aussi bien sur la forme que sur le fond du mémoire.

Enfin, je souhaite remercier ma famille et mes amis pour leur soutien tout au long de mes années d'études, notamment ma mère pour m'avoir soutenu tout ce temps et mon père pour tous ses encouragements. Une pensée également pour Sandrine, qui m'a guidé vers l'actuariat et sans qui ce mémoire n'existerait surement pas.

Confidentialité

Dans un souci de confidentialité, les montants de pertes présentés dans ce mémoire, aussi bien les montants de pertes historiques enregistrés par le Groupe que ceux simulés par le modélisateur ont été modifiés.

Table des matières

Introduction

En tant que premier risque naturel en France, les inondations constituent une composante majeure de la sinistralité des catastrophes naturelles et leur modélisation est primordiale pour assurer une gestion optimale de ces risques. C'est pourquoi, à la suite des inondations ayant touché le sud de la France en fin d'année 2019, Groupama a lancé un vaste projet pour mieux appréhender la compréhension de ce risque.

L'objectif premier est d'être en mesure d'identifier les portefeuilles et zones géographiques les plus exposés afin de réduire la sensibilité du groupe aux phénomènes d'inondations, que ce soit par le biais d'actions de préventions ou de politiques de souscriptions. Le second objectif vise à étudier l'impact du changement climatique sur le risque inondation et ses répercussions sur l'évolution de la sinistralité. Ces dernières années ont été particulièrement marquées par des évènements inondations majeurs, à l'image des inondations dévastatrices de juillet 2021 qui ont affecté l'Europe de l'Ouest et notamment l'Allemagne, la Belgique, le Luxembourg et les Pays-Bas et pour lesquels les dégâts ont été les plus lourds, avec un bilan humain de plus de 200 morts et près de 11 milliards de dommages assurés. Ces inondations causées par un niveau de précipitations record nous amènent à nous interroger sur l'évolution des précipitations avec le changement climatique et sur l'impact qui en découle sur la fréquence et l'intensité des futurs évènements inondations.

Pour répondre à ces objectifs, ce mémoire s'articule autour de la construction d'un modèle qui nous permettra dans un premier temps d'évaluer le coût annuel moyen du risque inondation pour chaque site assuré et dans un second temps nous permettra de traduire la variation de l'aléa climatique (i.e des précipitations) en une variation de la sinistralité. On divisera l'étude en cinq grandes parties :

Dans un premier temps, la première partie vise à introduire le contexte de l'étude en évoquant notamment les chiffres de l'inondation en France, les rapports du GIEC sur l'évolution du climat, ainsi que l'état actuel des travaux menés en assurance sur le changement climatique.

La deuxième partie s'intéresse à la modélisation classique des catastrophes naturelles. Ces modèles catastrophes naturelles sont effectués par des modélisateurs extérieurs et permettent de quantifier les pertes en se basant sur des simulations d'évènements inondations. On se basera sur les résultats de ce modèle pour la construction de notre modèle construit en partie 4.

Par la suite, on présentera l'ensemble des travaux ayant permis d'enrichir la base de risques avec des variables potentiellement explicatives du risque inondation : distance au cours d'eau, imperméabilité du sol, précipitations, etc. On évoquera également plus en détail les modèles climatiques CORDEX permettant d'obtenir des projections d'évolution des précipitations avec le changement climatique.

La partie 4 sera dédiée à la construction du modèle inondation. On utilisera des modèles linéaires généralisés, ainsi que des modèles de *machine learning*, qui seront entraînés sur les sorties du modèle catastrophe naturelle étudié en partie 2.

Enfin, la dernière partie vise tout d'abord à présenter les résultats obtenus après application du modèle sur l'ensemble du portefeuille. Puis, les modèles seront relancés, en faisant varier les indicateurs de précipitations renseignés en entrée du modèle selon les projections fournies par les modèles climatiques CORDEX. On disposera ainsi de nouveaux résultats de sinistralité obtenus selon des hypothèses d'évolution du climat, et ce pour différentes périodes futures et différents scénarios de changement climatique.

I. Contexte

1) Cadre de l'étude

En tant que troisième assureur en assurance dommages en 2020, Groupama Assurances Mutuelles (GMA) est particulièrement exposé aux risques climatiques. Né au XIXe siècle de la volonté d'agriculteurs souhaitant protéger leurs terres contre différents aléas, GMA est aujourd'hui l'un des assureurs majeurs du marché français. C'est dans ce contexte que fin 2019, à la suite des nombreuses inondations qui ont touché le sud de la France que GMA a lancé des travaux pour mieux appréhender et contrôler ce risque. Ainsi la prise en compte du changement climatique dans ces travaux est un axe majeur dans l'étude de la sensibilité du groupe au risque inondation.

De plus, étant donné le caractère extrême des risques étudiés et en l'occurrence ceux liés au risque inondation, les travaux réalisés au cours de ce mémoire ont été faits au sein de la Direction Réassurance de Groupama. Au fil des années la Direction Réassurance a déjà pu réaliser de nombreuses études climatiques notamment sur les tempêtes et a pu ainsi développer une expertise sur ces sujets en collaboration avec de nombreux acteurs extérieurs comme les courtiers en réassurance, les réassureurs, les agences de modélisation et les instituts météorologiques français.

Comme la plupart des assureurs de premier plan du marché, GMA agit en tant que réassureur de ses entités, et réassure donc notamment ses 9 caisses régionales de France Métropolitaine, ainsi que les 2 caisses d'outre-mer. L'étude effectuée dans ce mémoire se concentrera uniquement sur la France Métropolitaine et donc sur les 9 caisses de GMA, ainsi que sur les risques du GAN, autre filiale du groupe.

Figure 3 - Caisses régionales métropolitaines de Groupama

2) Le risque inondation

a. Description du risque inondation

Afin de mieux comprendre les variables utiles à la modélisation du risque inondation, il est primordial de comprendre ce risque et ses différentes composantes. L'inondation est une submersion temporaire, rapide ou lente, d'une zone habituellement hors de l'eau, qu'elle qu'en soit l'origine, elles peuvent être par ruissellement, submersion marine, débordement de cours d'eau ou encore par remontée de nappe.

À chaque inondation, on peut associer un niveau d'aléa, lié à plusieurs éléments :

- La hauteur maximale de submersion
- La durée de submersion
- La vitesse d'écoulement
- La période de retour
- La soudaineté
- L'ampleur

Ces éléments dépendent également de la typologie de l'évènement inondation considéré, on en distingue plusieurs :

Types d'inondation	**Phénomène**	**Caractéristiques**
Crues lentes de plaine	Débordement de cours d'eau à la suite de précipitations répétées et prolongées.	Lentes : apparaissent en plusieurs jours/heures. De longue durée : persiste pendant un à plusieurs jours.
Crues rapides et torrentielles	Débordement de cours d'eau à la suite de précipitations intenses.	Montée des eaux rapides, principalement en zone montagneuse avec un accroissement de la vitesse d'écoulement du cours d'eau.
Ruissellement	Eau qui ne peut plus s'infiltrer dans le sol à la suite de précipitations intenses.	Souvent en milieu urbain dû notamment à l'artificialisation des sols.
Submersion marine	Élévation du niveau de la mer à la suite de conditions météorologiques défavorables (pleine mer, tempête…).	Inondations souvent dévastatrices. Affecte les zones côtières.
Remontée de nappe	Montée du niveau de la nappe phréatique jusqu'à la surface du sol.	Inondation progressive et lente.

Tableau 3- Les différents types d'inondations (source : georisques.gouv)

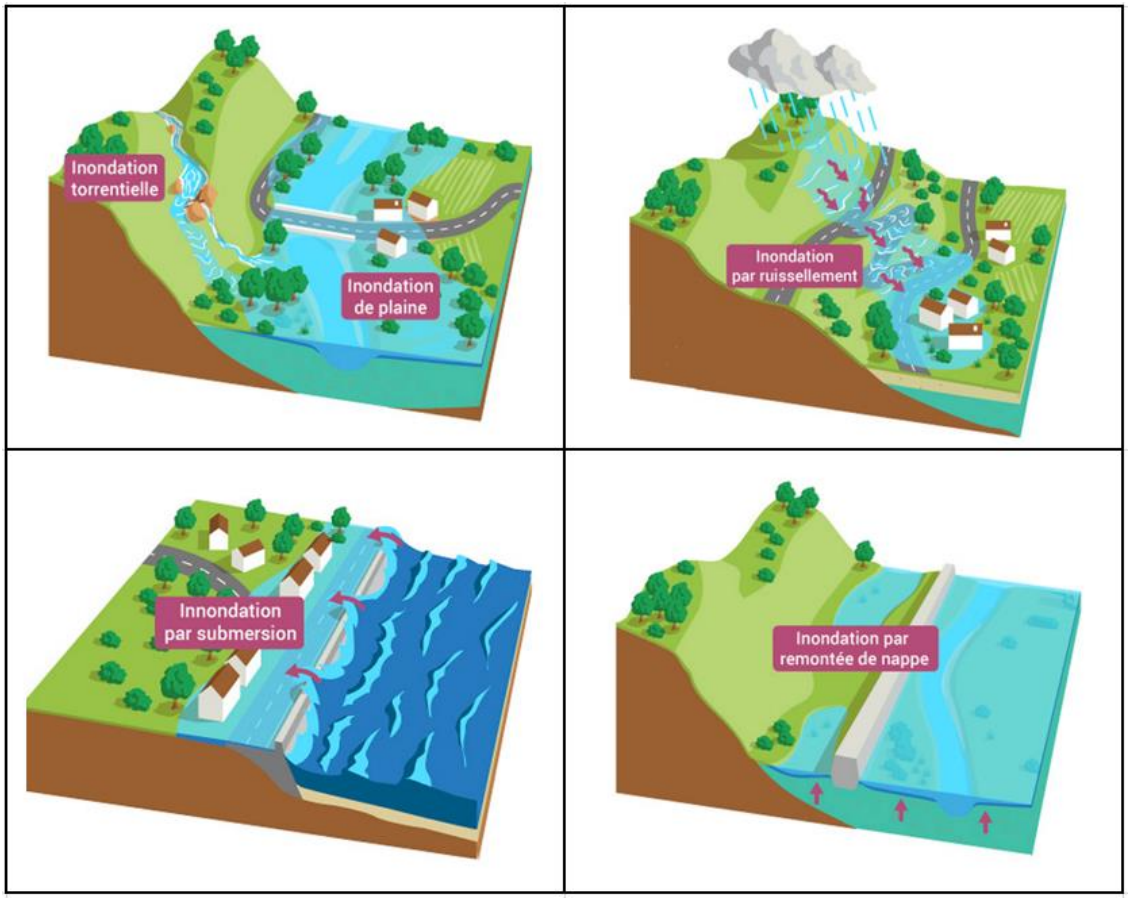

Figure 4- Illustration des différents types d'inondations (source : EauFrance)

On s'intéressera par la suite uniquement à la modélisation des 3 premiers types d'inondations, à savoir les inondations par débordement de cours d'eau (torrentielle ou de plaine) et par ruissellement. À noter cependant qu'il est toujours compliqué de totalement séparer ces différents évènements, ceux-ci pouvant en effet être liés. Par exemple, le ruissellement et la submersion marine peuvent contribuer voire causer un débordement de cours d'eau, débordement qui peut lui-même causer une remontée de nappe phréatique par exemple.

Ces phénomènes sont tous causés par un aléa météorologique à savoir des houles de forte intensité ou niveau de la mer élevé dans le cas de la submersion marine et des précipitations en intensité ou en durée pour le reste. L'étude des précipitations et de leur évolution avec le changement climatique sera donc primordiale dans la suite de l'étude.

b. Les chiffres de l'inondation en France

En France métropolitaine, on estime que 16,8 millions d'habitants sont exposés aux différentes conséquences des inondations par débordement de cours d'eau. Depuis 1982 le coût moyen annuel de la sinistralité (Non-Auto) inondation s'élève à 554 millions d'euros, contre 475 millions d'euros pour la sécheresse depuis 1989 faisant des inondations le premier risque naturel en France.

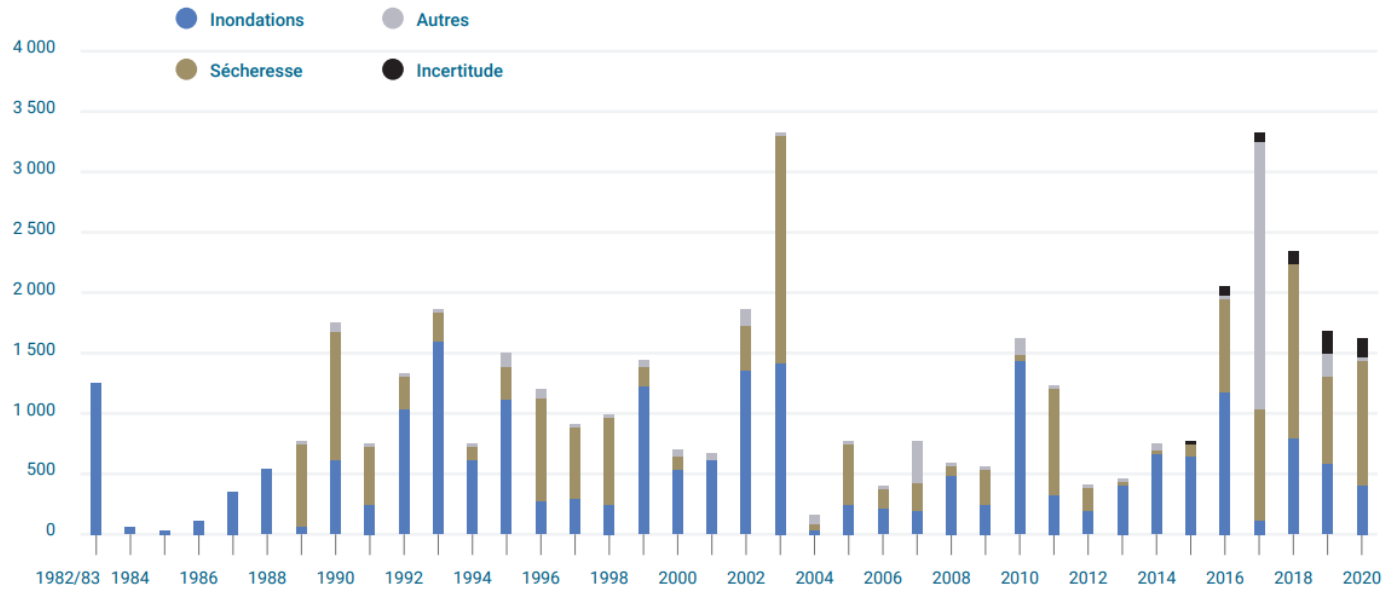

Figure 5 - La sinistralité catastrophes naturelles non-auto de 1982 à 2020 en millions d'euros 2020 (source : CCR)

La sinistralité catastrophe naturelle annuelle s'élève en moyenne à 1044 millions d'euros avec une récente sinistralité particulièrement dévastatrice, dépassant en effet la barre des 1500 millions d'euros depuis cinq ans. L'étude de ces catastrophes et de leur évolution avec le changement climatique apparait donc comme indispensable. Notons cependant une part plus importante de la sécheresse ces dernières années. Pour l'année 2020, la sécheresse représente 72% du coût de l'exercice contre 27% pour l'inondation.

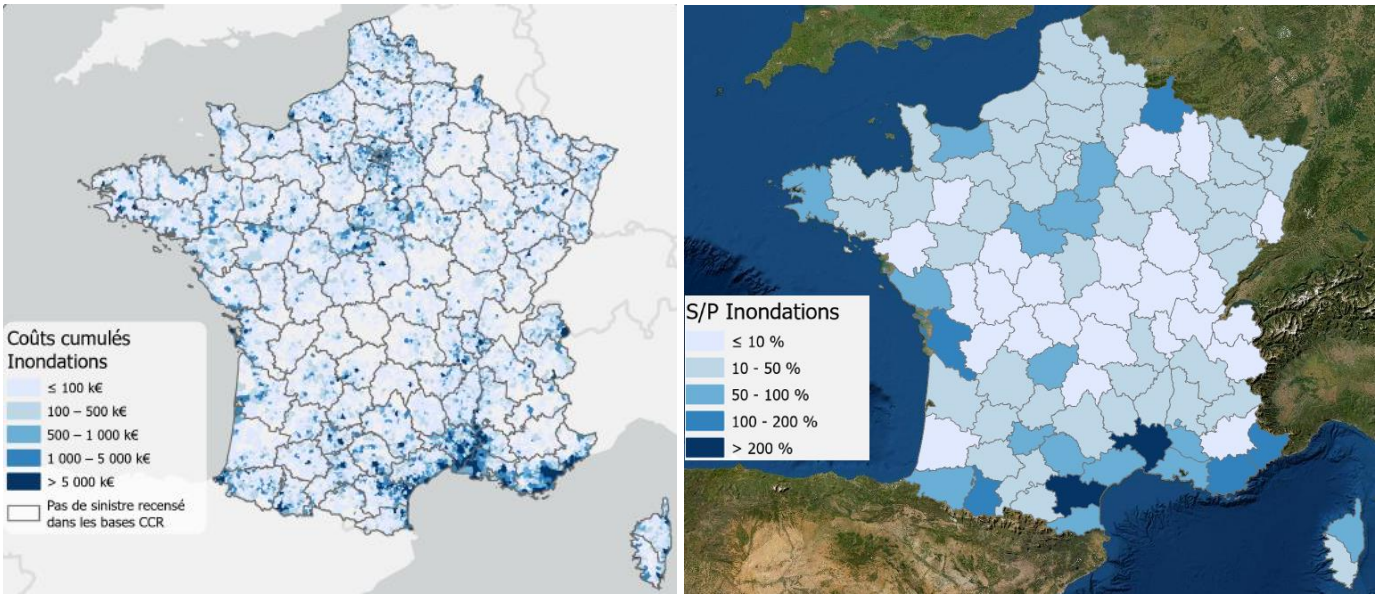

Figure 6- Coûts inondations cumulés par commune et S/P inondations par département de 1995 à 2018 (source : CCR)

Historiquement il apparait que le sud de la France et particulièrement la côte méditerranéenne est la plus touchée par les inondations avec de nombreuses communes dépassant les 5 millions d'euros de coût inondation dans la période 1995 à 2018. En calculant le cumul des sinistres sur le cumul des primes sur la période considérée, deux départements ressortent avec un ratio deux fois supérieur aux primes catastrophes naturelles perçu par la Caisse Centrale de Réassurance (CCR), à savoir le Gard et l'Aude. Cinq autres départements ressortent également comme particulièrement sinistrés avec des ratios entre 100 et 200%, le Var, les Alpes-Maritimes, les Hautes-Pyrénées, la Charente-Maritime ainsi que les Ardennes. À noter que la prime considérée est la prime acquise corrigée des variations du taux de surprime Cat Nat et du taux de prélèvement dans le cadre du fond de prévention des risques

naturels majeurs (FPRNM). Pour mieux comprendre ce dernier point, faisons un point sur le régime Cat Nat en France et les éventuelles conséquences que pourrait avoir une augmentation de la sinistralité due au changement climatique.

c. La réassurance des catastrophes naturelles en France

Le régime d'indemnisation des catastrophes naturelles (régime CAT-NAT) a été créé par la loi du 13 juillet 1982 et a permis de pallier un manque de couverture des risques naturels qui n'étaient jusqu'alors que très peu assurés. Ce régime se traduit par une extension obligatoire sur les contrats d'assurance dommages. Celle-ci donne lieu au paiement d'une prime additionnelle et unique, qui ignore donc le degré d'exposition, traduisant ainsi la logique de solidarité fondée sur l'alinéa 12 du préambule de la Constitution de 1946 : « La Nation proclame la solidarité et l'égalité de tous les Français devant les charges qui résultent des calamités nationales ». Cette surprime est actuellement de :

- 12% pour les contrats MRH (multirisque habitation)
- 6% pour les contrats d'assurance de véhicule

En contrepartie de leur obligation de couverture les assureurs peuvent se réassurer auprès de la CCR (Caisse centrale de réassurance). C'est l'entreprise de réassurance publique, qui est donc détenue par l'État. Elle permet de réassurer dans le cadre du régime CAT-NAT tout assureur qui lui en fait la demande et réalise ainsi une mutualisation de l'ensemble des risques à travers la couverture des portefeuilles des différents assureurs. Plus précisément, la CCR réassure les assureurs privés avec un traité quote-part à 50% et un traité *stop-loss* sur les risques conservés par l'assureur après application du traité quote-part. Le traité *stop-loss* proposé par la CCR a la particularité d'avoir une portée illimitée (il n'y a donc pas de plafond) grâce à la garantie de l'État. En effet, l'État peut intervenir en dernier recours afin d'éviter toute défaillance du système.

Attention cependant la réassurance auprès de la CCR ne peut se faire uniquement si le régime CAT-NAT est activé. Pour le déclencher il faut d'abord que le bien endommagé soit couvert par un contrat d'assurance dommage et également que la commune dans lequel il se trouve soit reconnue en état de catastrophe naturelle par arrêté interministériel.

Cette reconnaissance repose sur le caractère anormal de l'intensité du phénomène naturel considéré, mais suivant le phénomène, la définition d'intensité anormale peut être ambiguë. Par exemple pour les vents, des vitesses supérieures à 145 km/h en moyenne sur 10 minutes ou supérieures à 215 km/h en rafales sont considérées comme anormales. De plus, certains périls sont exclus d'office puisqu'il existe déjà des garanties d'assurance permettant d'indemniser les pertes qu'ils occasionnent : c'est le cas de la grêle ou de la neige par exemple. On peut résumer le fonctionnement via le schéma suivant :

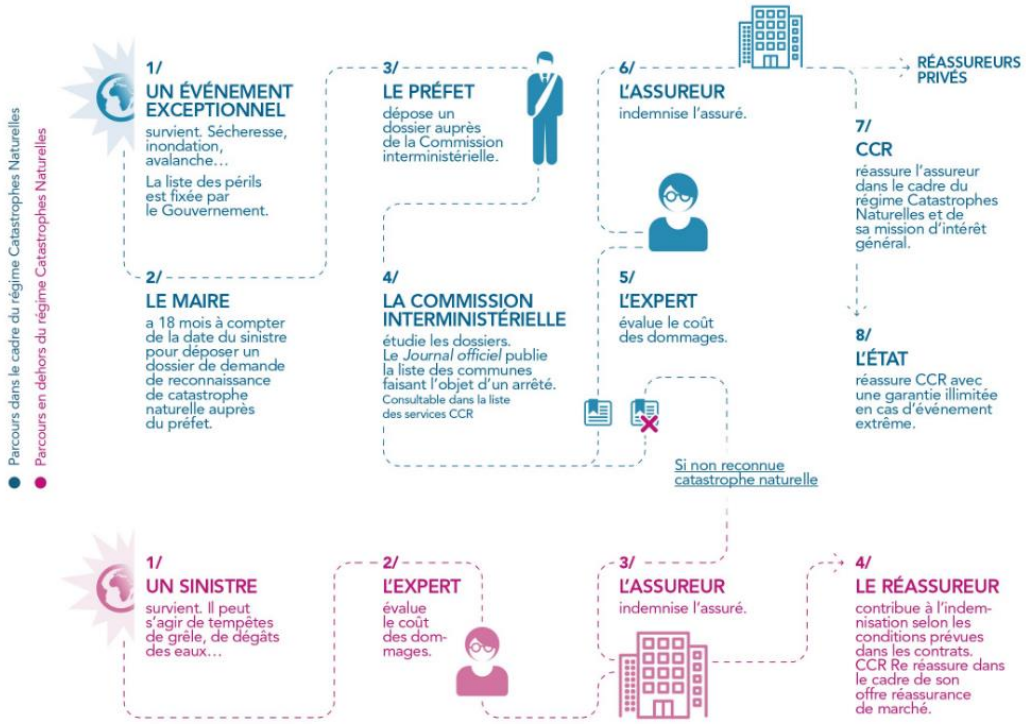

Figure 7- Fonctionnement du régime CAT NAT (source : CCR)

En 2020, le total des primes Cat Nat est estimé à 1,72 milliard d'euros, soit une augmentation de 2,5% par rapport à 2019. Globalement, l'évolution n'a cessé de croître depuis 1982, s'expliquant en grande partie par l'évolution de l'assiette sur laquelle elles sont calculées, c'est-à-dire l'augmentation des primes dommages auto et dommages aux biens sur le marché français. Cette augmentation s'explique également par l'évolution du taux de surprime hors auto, passant de 5,5% à 9% en 1985, puis de 9% à 12% en 2000. L'étude des conséquences du changement climatique sur la sinistralité des catastrophes naturelles pourrait de nouveau remettre en question ce taux de surprime Cat Nat. Notons également que depuis 1995, une partie des primes Cat Nat permet d'alimenter le Fonds de prévention des risques naturels majeurs (FPRNM) ou fonds Barnier servant à financer des actions de prévention et d'expropriation de biens exposés à un risque naturel majeur.

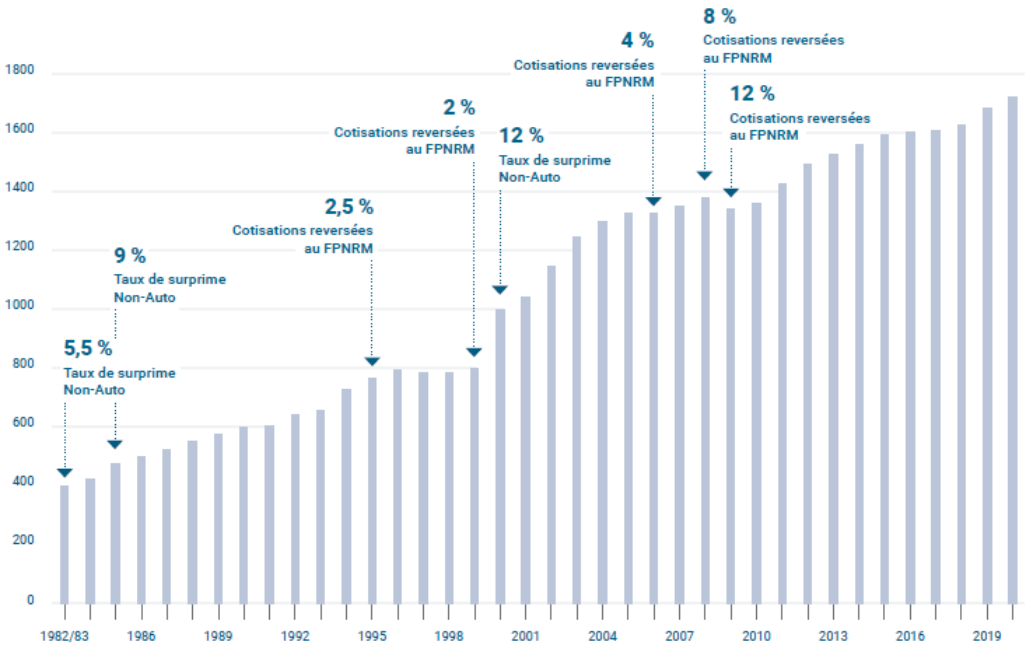

Figure 8 - Évolution des primes Cat Nat depuis 1982 (source : CCR)

3) La problématique du changement climatique

a. Les rapports du GIEC

En matière d'étude sur l'évolution du climat, ses impacts et ses causes, le GIEC (Groupe d'experts Intergouvernemental sur l'Évolution du Climat) est devenu une référence depuis maintenant plus de 30 ans. Cet organisme intergouvernemental est ouvert à l'ensemble des pays membres de l'Organisation des Nations unies (ONU) et compte actuellement 195 pays soit la quasi-totalité des pays du monde.

Les rapports du GIEC mettent en lumière les connaissances les plus avancées sur le sujet et permettent également d'identifier les solutions existantes pour limiter l'ampleur du réchauffement et de s'adapter aux différents changements attendus. Ces rapports sont effectués sous le contrôle de trois groupes de travail différents, le premier vise à étudier l'évolution du climat, le second s'intéresse à la vulnérabilité des systèmes socio-économiques et naturels aux changements climatiques et le troisième s'occupe des solutions envisageables pour atténuer les changements climatiques.

Dans l'attente du sixième rapport du GIEC prévu officiellement en 2022, mais dont les premières fuites dévoilées par l'AFP annoncent que « le pire est à venir », on peut s'intéresser au cinquième rapport publié en 2014. Ce rapport a notamment vu apparaitre l'introduction de quatre scénarios d'émissions de gaz à effet de serre dits RCP (pour *Representative Concentration Pathway*) permettant aux climatologues d'effectuer différentes projections climatiques selon différentes évolutions possibles du comportement à venir des sociétés humaines. Concrètement, ces scénarios sont effectués en fonction de l'évolution du forçage radiatif, exprimé en W/m2, il est défini comme un changement dans la différence entre le rayonnement entrant et sortant, dû par exemple à la concentration des gaz à effet de serre. Un forçage radiatif positif indique donc un réchauffement du système climatique considéré.

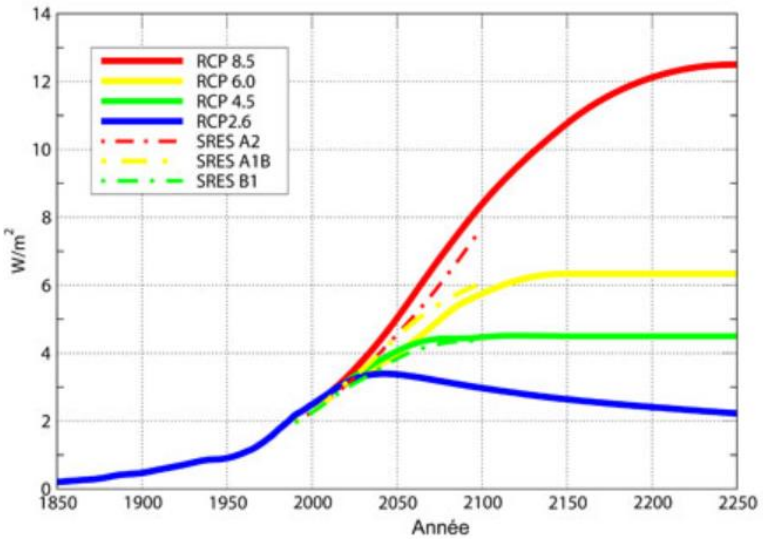

Figure 9- Évolution du bilan radiatif de la terre ou « forçage radiatif » en W/m2 sur la période 1850-2250 selon les différents scénarios (source : Météo France)

Notons ainsi que le scénario RCP 8.5 correspond au scénario le plus pessimiste dans lequel aucune politique climatique n'est menée, tandis que le scénario 2.6 est le plus optimiste avec des politiques rapides de décroissance des gaz à effet de serre et un réchauffement global inférieur à 2°C en 2100 (par rapport aux températures préindustrielles) ce qui correspondrait aux objectifs des accords de Paris (COP21).

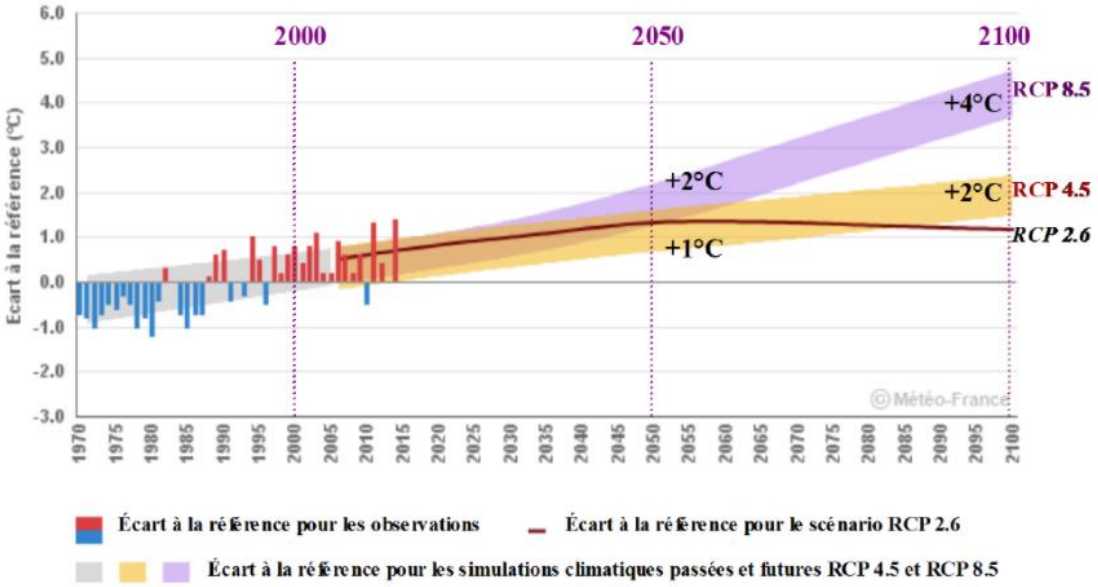

Figure 10- Température moyenne annuelle en France métropolitaine : écart à la référence 1976 - 2005 (source : Météo France)

Il est à l'heure actuelle difficile de déterminer une tendance concernant le scénario vers lequel on se dirige. Selon Météo France, en agissant maintenant le scénario 4.5 semble le plus probable avec une hausse de 2,2°C en moyenne d'ici 2100, tandis qu'en tardant trop le scénario 8.5 sera à envisager avec une hausse des températures de 3,9°C.

Scénario	Évolution des émisions de gaz à effet de serre	Évolution probable des températures (2046 - 2065)	Évolution probable des températures (2081 - 2100)
RCP 8.5	Poursuite de l'augmentation de gaz à effet de serre au rythme actuel. **Scénario le plus pessimiste.**	1,4 à 2,6°C	2,6 à 4,8°C
RCP 4.5	Scénario avec stabilisation des émissions avant la fin du XXIe siècle à un niveau faible.	0,9 à 2,0 °C	1,1 à 2,6°C
RCP 2.6	Scénario à tres faibles émissions avec un point culminant avant 2050. **Scénario le plus optimiste.**	0,4 à 1,6°C	0,3 à 1,7°C

Tableau 4 - Résumé des scénarios du GIEC (Source : GIEC)

Enfin, ce rapport estime comme étant probable (i.e avec une probabilité de 66%) l'augmentation de l'intensité ou de la durée des sécheresses et comme très probable (i.e avec une probabilité de 90%) l'augmentation de la fréquence, de l'intensité ou du nombre des épisodes de précipitations abondantes. Face à ces défis environnementaux, les assureurs ont un rôle central à jouer et la prise en compte du changement climatique dans la gestion des risques est primordiale.

b. Vers une prise de conscience dans l'assurance : l'exercice pilote climatique de l'ACPR

L'exercice pilote climatique conduit par l'ACPR (Autorité de contrôle prudentiel et de résolution) est un exercice d'évaluation des risques associés au changement climatique mené de juillet 2020 à avril 2021. Il réunissait une majorité des groupes bancaires et organismes d'assurance du marché français avec 9 organismes bancaires et 15 organismes d'assurance, dont Groupama (85% du total du bilan bancaire et 75% du total du bilan des assureurs).

L’exercice visait à réaliser une projection du bilan pour différents horizons de temps jusqu’en 2050 (2025, 2035, 2040, 2050) en mettant en place une approche granulaire sur 55 secteurs d’activités et pour 3 scénarios de transitions différents, un scénario de référence correspondant à une transition ordonnée permettant de satisfaire les engagements de l’accord de Paris, et deux autres scénarios, l’un correspondant à une transition retardée et l’autre à une transition rapide.

L'analyse de l’impact de l’augmentation de la fréquence et de l’intensité des catastrophes naturelles sur l’activité dommages aux biens a été réalisée avec la collaboration de la CCR qui fournissait aux assureurs l’augmentation des sinistres par département basé sur la base des expositions des assureurs au maillage communal. La CCR fournissait aux assureurs des taux d’évolutions de sinistralité aux différents horizons et pour différents périls (tous périls, inondation, submersion marine et sécheresse), selon un modèle basé sur le scénario RCP 8.5 du GIEC et dont les résultats ont été publiés en 2018 par la CCR.

Les résultats des modélisations ont montré que la sinistralité des branches entrant dans le cadre du régime Cat Nat augmente de 174% entre 2019 et 2050. Concernant le seul risque inondation, la carte ci-dessous présente les résultats par département. Il apparait par exemple que dans le Gard, la sinistralité inondation s’élevait à 7,86 euros par habitant en 2019 et devrait augmenter sur la période 2020 – 2050 dans une fourchette comprise entre 115 et 196%. À noter que ces évolutions dépendent de la seule sinistralité 2019, ainsi les régions en rouge foncé sont souvent celles dont les montants de sinistres sont aujourd’hui les plus faibles.

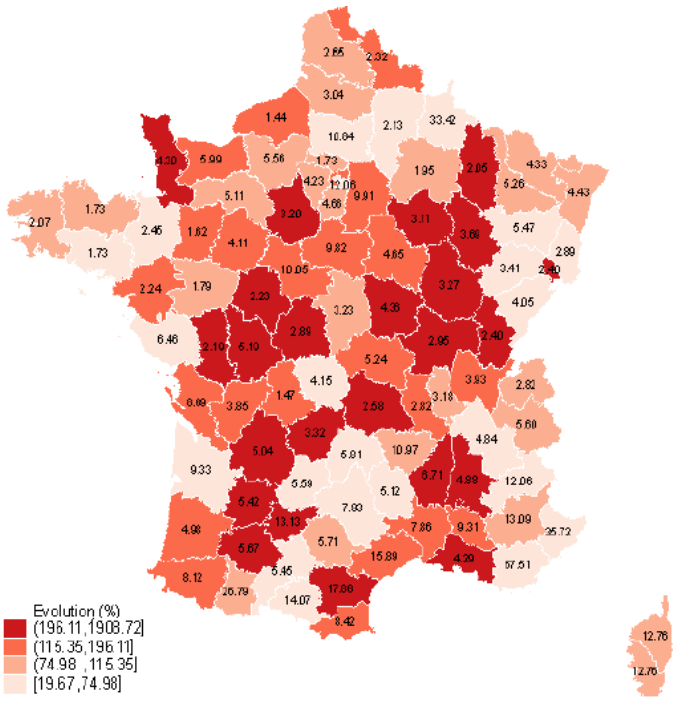

Figure 11- Sinistralité inondation modélisée sur la période 2019-2050 (source : ACPR)

C’est dans ce contexte que la Direction Réassurance de Groupama, qui a participé à l’exercice pour la partie Cat Nat, souhaitait dans la continuité de ces travaux approfondir l’étude de l’impact du changement climatique et notamment son impact sur le risque inondation.

4) Historique des travaux réalisés sur le risque inondation et changement climatique

La Direction Réassurance de Groupama a déjà effectué des travaux sur le changement climatique avec le mémoire d'Arnaud Dalleau *« Évaluation de l'impact du changement climatique sur le risque cyclonique dans les Antilles »* publié en 2021. Les modèles CORDEX utilisés dans ce mémoire seront également introduits par la suite.

Les mémoires sur le risque inondation sont nombreux, on peut citer :

- *Création d'un modèle de tarification du risque inondation* (2019) par Rimen Ayoub qui a utilisé une approche par GLM pour la modélisation.
- *Modélisation stochastique des inondations en France et Applications en Réassurance* (2017) par Hamza El Hassani, dans lequel un modèle CAT inondations a notamment été créé.
- *Cartographie du risque inondation* (2016) par Elie Dadoun dans lequel une cartographie du risque inondation en Turquie est effectuée.

L'approche utilisée dans le mémoire sera totalement différente des mémoires précédemment cités. La partie modélisation du mémoire sera construite autour d'un modèle catastrophe existant proposé par un modélisateur extérieur à Groupama, l'objectif sera de créer un modèle en interne en l'entraînant non pas sur la sinistralité comme cela peut être fait classiquement avec une approche fréquence coût, mais en l'entraînant directement sur les résultats du modèle CAT fournit par le modélisateur. Par la suite, on pourrait faire varier les variables explicatives selon des hypothèses de changement climatique, notamment du point de vue des variables sur les précipitations permettant ainsi d'avoir une traduction de la variation de l'aléa climatique en une évolution de la sinistralité.

II. La modélisation des catastrophes naturelles

1) Fonctionnement

a. Une nécessité d'utilisation de modèles catastrophes naturelles

Lorsqu'il s'agit d'évaluer les pertes financières probables que peut représenter un risque, l'approche fréquence-coût est souvent retenue, notamment dans le cadre de la tarification des risques de masse tels qu'en auto ou MRH. Cependant dans le cadre de la modélisation des catastrophes naturelles cette approche n'est pas adaptée, et ce pour plusieurs raisons.

- Elle suppose tout d'abord une indépendance entre les observations, ce qui n'est pas le cas étant donné le caractère concentré des catastrophes qui surviennent et qui touchent donc plusieurs risques d'une même zone.
- On doit également vérifier une hypothèse d'indépendance entre la fréquence des sinistres et leur coût. Encore une fois, cette hypothèse ne semble pas vérifiée, les catastrophes climatiques de forte amplitude sont rares, tandis que les catastrophes de plus faible amplitude apparaissent avec des fréquences souvent plus élevées.
- Étant donné le caractère exceptionnel des évènements inondations et la faible profondeur d'historique dont les assureurs disposent il n'est pas possible d'obtenir une distribution de perte assez robuste pour représenter fidèlement la diversité des évènements climatiques et le caractère exceptionnel de certains. Dans le cadre de Solvabilité 2 notamment, estimer une perte survenant en moyenne tous les 200 ans avec un historique de sinistralité de 10 ans n'est pas envisageable avec des méthodes d'estimation classique.

C'est dans ce contexte qu'interviennent les modèles catastrophes naturelles (modèles CAT) qui s'appuient notamment sur la modélisation physique des évènements, puis sur la traduction de ces évènements en pertes financières. Permettant ainsi de prendre en compte la corrélation qui existe entre les risques, tout en complétant la vision historique avec des scénarios inondations extrêmes qui n'ont pas été nécessairement observés dans le passé proche et qui ne sont donc pas présents dans notre base sinistre. La modélisation physique des évènements étant très complexe, Groupama, comme une majorité d'assureurs et de réassureurs, achète des résultats de modélisations auprès de modélisateurs extérieurs. Il en existe actuellement trois, à savoir AIR, EQECat et RMS. Ces modèles servent aussi bien à la tarification des programmes de réassurance, qu'à l'évaluation et à la gestion du risque catastrophe.

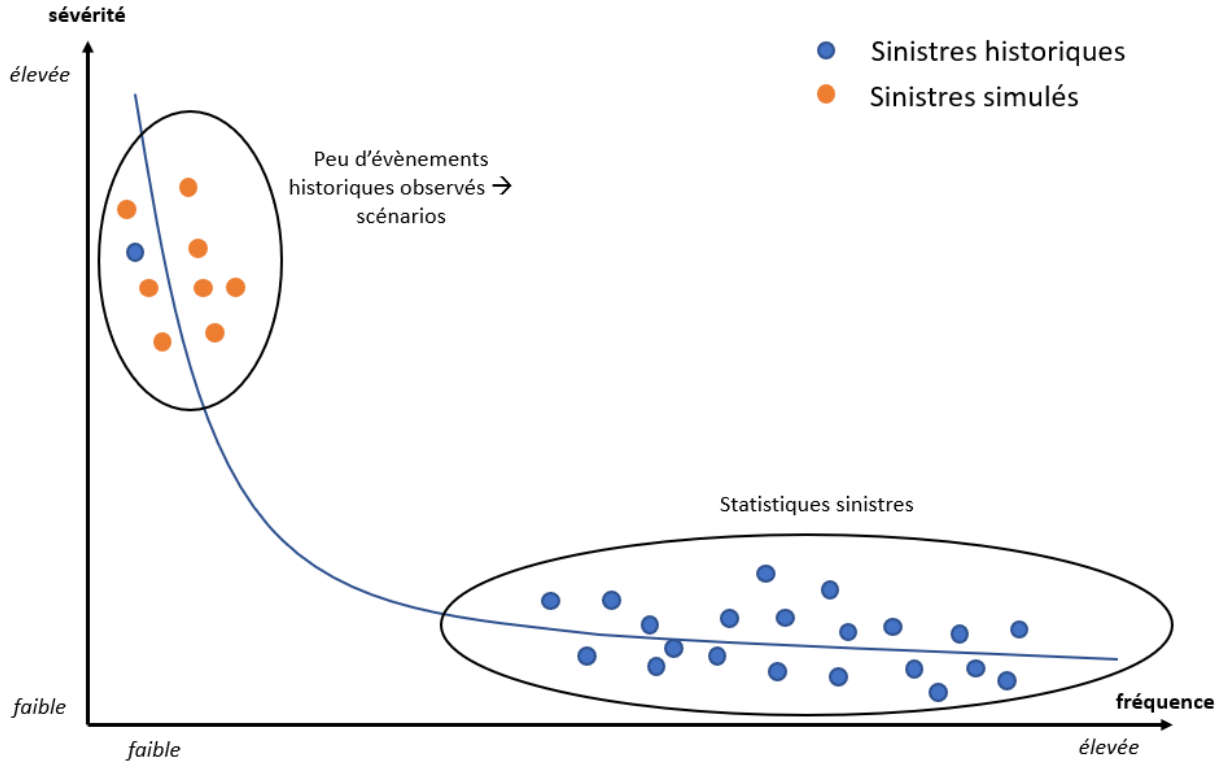

Figure 12 - schéma : intérêt de la simulation de scénarios probables

On corrige la problématique de la faible profondeur d'historique en introduisant des scénarios inondations fictifs, mais probables, afin de représenter le plus fidèlement possible la distribution des pertes, et notamment pour les évènements extrêmes qui sont de faibles fréquences.

b. Structure d'un modèle catastrophe

Les modèles CAT sont construits autour de quatre modules différents ayant chacun leur rôle.

Tout d'abord, le module exposition contient la base de données des risques de l'assureur, permettant ainsi de transmettre aux autres modules les caractéristiques des biens assurés, que ce soit du point de vue physique et géographique afin d'alimenter le module vulnérabilité, ou bien du point de vue des conditions d'assurances pour alimenter le module financier.

Concernant les caractéristiques physiques et géographiques, la première information importante est la géolocalisation, c'est-à-dire l'affectation d'une latitude et d'une longitude à chaque risque, qui est primordiale étant

donné la forte dépendance entre l'emplacement du risque et la fréquence/sévérité des catastrophes naturelles. Cette information peut être également calculée directement par les modèles à partir de l'adresse lorsque l'assureur ne fournit pas l'information.

On retrouve également le type de bien, une maison aura davantage tendance à être touchée par le risque inondation qu'un appartement par exemple. La branche assurée est aussi importante et des courbes de vulnérabilité différentes sont appliquées en fonction des branches, des bâtiments industriels bénéficient probablement d'une meilleure protection contre le risque inondation qu'un simple bâtiment résidentiel par exemple.

On retrouve ensuite la somme assurée (ou engagements), elle correspond à la somme garantie par le contrat d'assurance. Elle est décomposée en trois parties : la valeur du bâtiment assuré, la valeur du contenu et la valeur des pertes d'exploitation garanties.

En parallèle, le module aléa joue un rôle majeur et beaucoup plus complexe, celui de construire un catalogue d'évènements sur plusieurs milliers d'années, en l'occurrence 50 000 années dans le modèle que nous avons utilisé. Les caractéristiques de ces évènements sont calibrées à partir de lois de probabilité sur l'historique afin d'assurer la pertinence en matière de fréquence, d'intensité et de localisation des évènements fictifs simulés. Les caractéristiques de la zone géographique sont également utilisées, précipitations, températures ou encore typologie du sol.

Les deux précédents modules sont par la suite utilisés pour le module vulnérabilité permettant d'évaluer les dégâts causés par les évènements simulés. Les modélisateurs utilisent pour cela des courbes de vulnérabilités reliant la sévérité de l'évènement et le taux de destruction du bâtiment. Ces courbes dépendent également des différentes caractéristiques évoquées pour le module exposition.

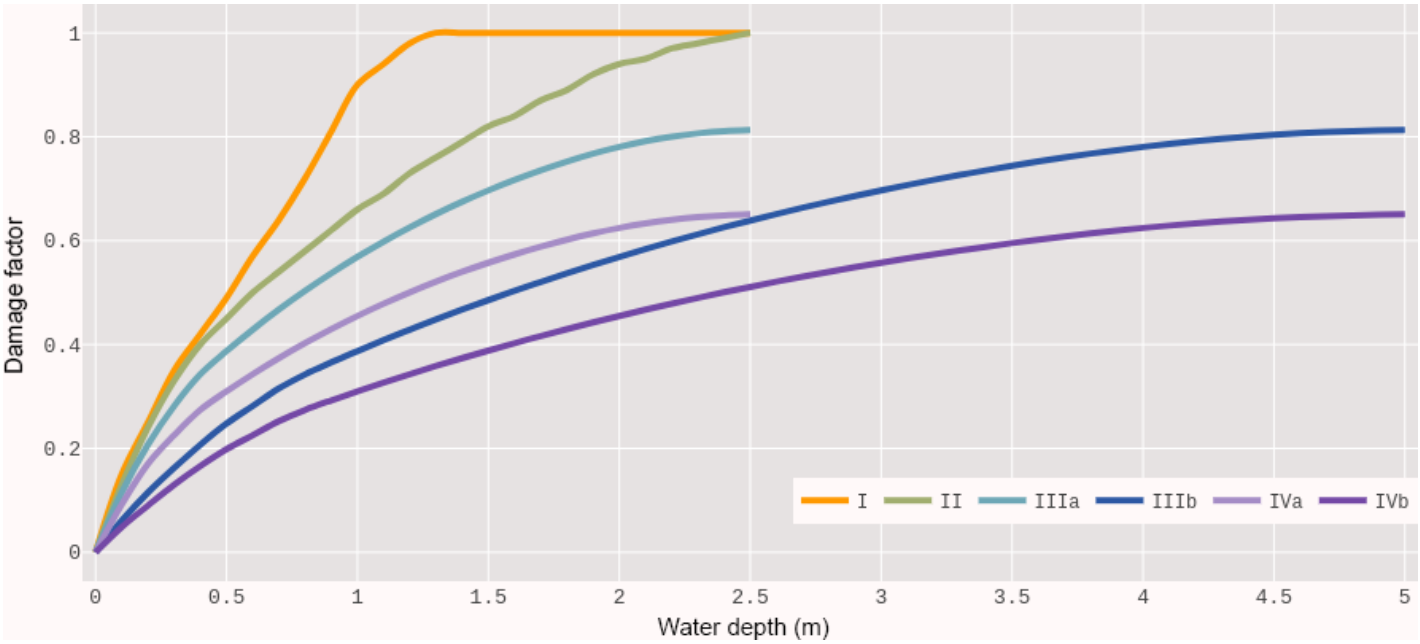

Figure 13 - Exemple de courbe de vulnérabilité : taux de destruction en fonction de la hauteur d'eau (Source : Natural Hazards and Earth System Sciences (NHESS))

Enfin, le module financier relie les taux de destruction aux dommages assurés pour en déduire une perte financière sur laquelle on applique également les différentes conditions d'assurance, afin d'obtenir une perte nette de conditions d'assurance.

On peut résumer la structure des modèles CAT avec le schéma ci-dessous.

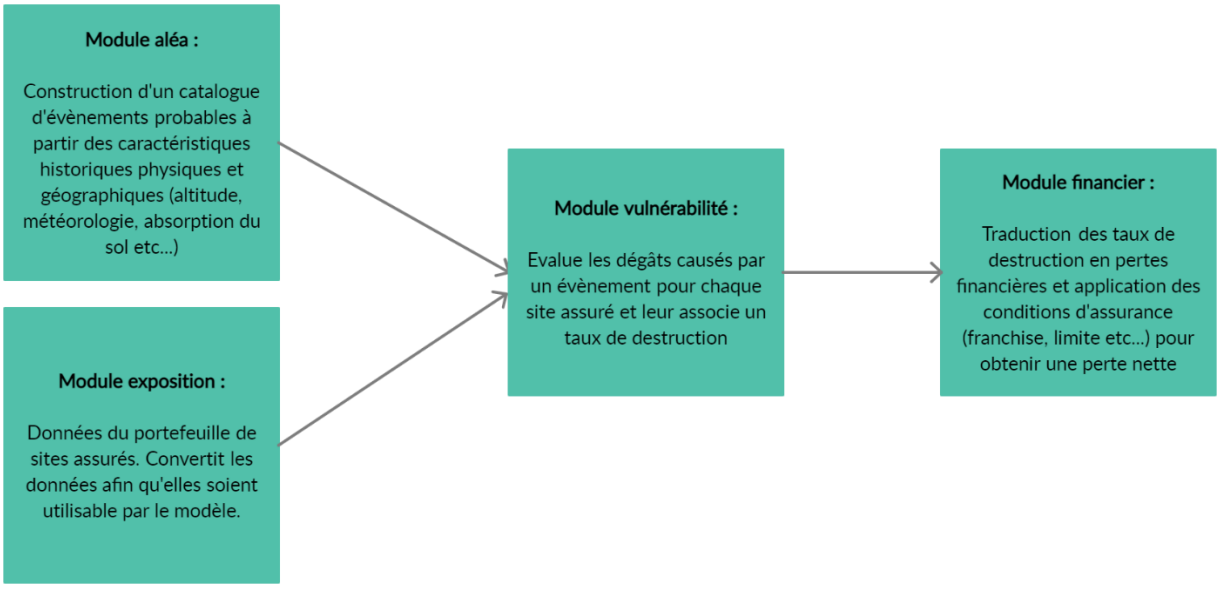

Figure 14 - schéma : structure d'un modèle CAT

c. Les sorties du modèle

Une fois le module financier terminé, plusieurs indicateurs sont ensuite calculés afin de permettre à l'utilisateur du modèle d'exploiter les résultats et d'évaluer l'exposition du portefeuille au risque considéré. On introduit pour cela trois indicateurs majeurs, à savoir l'AAL, l'AEP et l'OEP.

- L'AAL pour *Annual Average Loss* est la perte annuelle moyenne. En notant $X_1, X_2, \dots, X_n$ les variables aléatoires correspondant aux pertes annuelles sur n années modélisées (en l'occurrence 50 000 dans notre cas). On a simplement :

$$AAL = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- L'AEP pour *Aggregate Exceedance Probability* est la probabilité que la somme des pertes annuelles dépasse un certain seuil. En notant $Y_1, Y_2, \dots, Y_N$ les variables aléatoires correspondant aux pertes annuelles par évènement dans l'année. On obtient :

$$AEP(x) = \mathbb{P}\left(\sum_{i=1}^{N} Y_i > x\right)$$

- L'OEP pour *Occurrence Exceedance Probability* est la probabilité que la perte annuelle maximale par évènement dépasse un certain seuil. D'où :

$$OEP(x) = \mathbb{P}(\max(Y_1, \dots, Y_N) > x)$$

Réciproquement, les deux derniers indicateurs permettent de représenter un niveau de confiance donné en fonction de la perte associée. Plutôt que d'exprimer le niveau de confiance directement en probabilité avec AEP(x), on l'exprime régulièrement en période de retour avec 1/AEP(x). Plusieurs périodes de retour sont souvent fournies par les modélisateurs avec des périodes de retour qui vont de 2 ans (donc une probabilité de dépasser le seuil de perte de 1/2), jusqu'à 5000 ans (donc une probabilité de dépasser le seuil de perte de 1/5000). Pour une année (AEP) ou un évènement donné (OEP), une période de retour de k années signifient qu'une année ou qu'un évènement de telle ampleur se produit en moyenne une fois sur k années consécutives. Du côté du modélisateur, pour déterminer le montant de perte associé à la période de retour de k années, on représente la série des pertes sur les 50 000 années de modélisations, et l'on sélectionne le quantile d'ordre $1 - \frac{1}{k}$.

À noter que ces indicateurs sont fournis avec une vision *ground up loss* correspondant à la sinistralité brute et une vision *gross loss* correspondant à la sinistralité nette après application des conditions d'assurance lors du module financier.

Enfin, dans le cas spécifique du risque inondation, ces indicateurs sont fournis selon 3 visions :

- Le risque fluvial : les principaux cours d'eau débordent ou sortent de leur lit et inondent les zones environnantes.
- Le risque pluvial : des précipitations excessives et importantes provoquent une inondation indépendamment du débordement d'un plan d'eau majeur (inondation par ruissellement). Le débordement des rivières et ruisseaux mineurs est également inclus dans ce risque.
- Le risque combiné : combinaison du risque fluvial et pluvial. Les deux risques étant supposés indépendants on obtient que : AAL combiné = AAL fluvial + AAL pluvial.

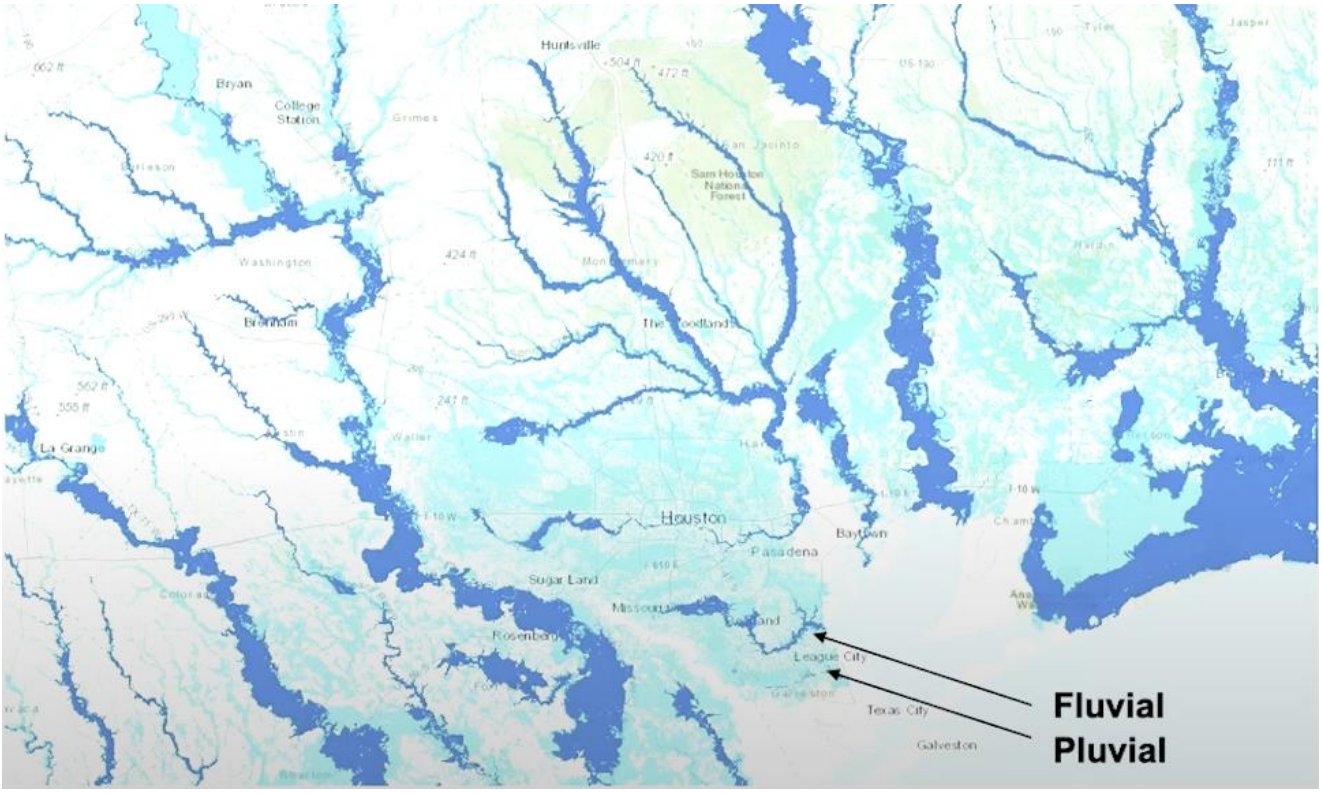

Figure 15 - Distinction entre le risque fluvial et pluvial : empreintes inondation (Source : RMS)

2) Application sur les données GMA

a. Données modélisées

L'objectif pour notre étude étant de prédire la charge annuelle inondation pour chacun des risques de notre base, l'AAL, l'AEP et l'OEP nous étaient donc fournis à un maillage individuel pour chaque site assuré. On disposait également d'une vision plus globale avec ces indicateurs par zone IRIS (découpage infracommunal effectué par l'INSEE) et un maillage global sur l'ensemble de la base pour chaque branche assurée.

Cependant, nous ne pouvions pas faire modéliser l'ensemble de la base et il fallait donc choisir la partie de la base que l'on souhaitait utiliser pour entraîner nos modèles inondation. On pouvait ainsi créer notre modèle sur cette sous-partie puis utiliser le modèle ainsi créé pour prédire les charges inondations pour les autres risques qui n'ont pas été envoyés au modélisateur. On a ainsi sélectionné 2500 zones IRIS réparties aléatoirement sur le territoire parmi les 50 000 existantes pour chacune de nos branches assurées.

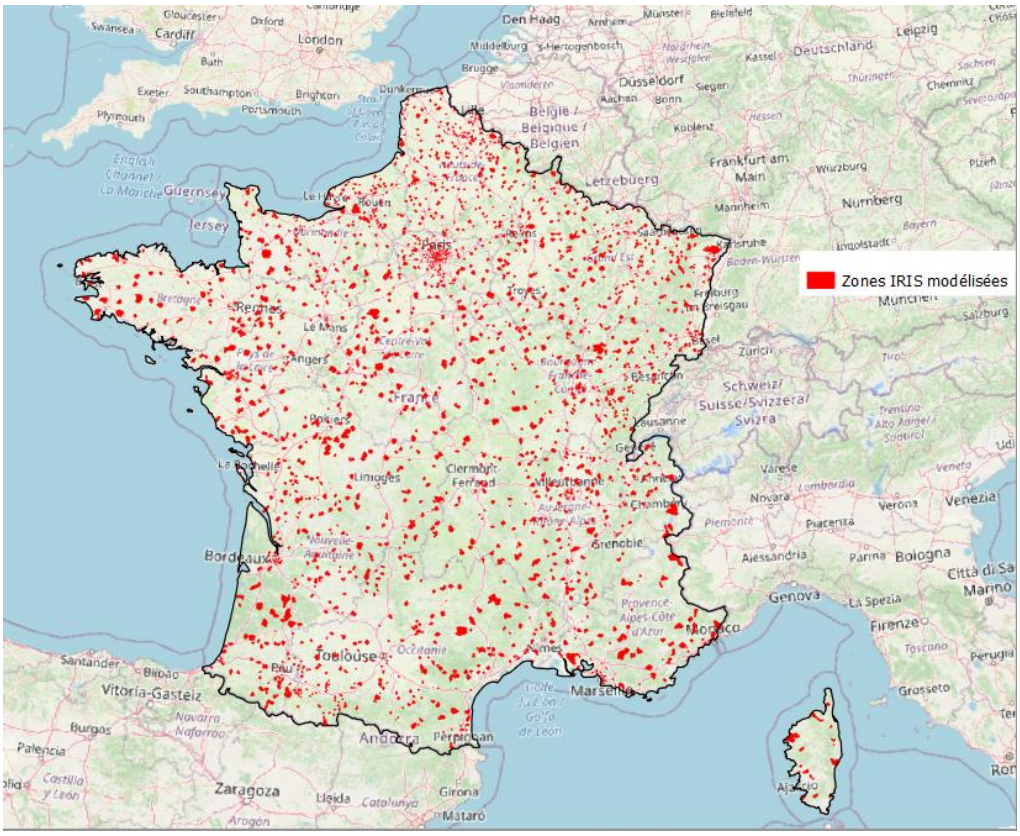

Figure 16 - Répartition des Zones IRIS modélisées

b. Observations sur les résultats

Il est difficile d'exploiter directement les résultats étant donné qu'on ne traite qu'un faible échantillon de notre portefeuille, sur des zones réparties sur l'ensemble du territoire. On peut cependant avoir un premier aperçu des zones qui ressortent particulièrement à risques. Pour gommer l'effet induit par la somme assurée sur la charge annuelle moyenne inondation, on considère plutôt le taux de destruction calculé simplement de la manière suivante :

$$Taux\ de\ destruction\ inondation = \frac{AAL\ inondation}{Engagements}$$

L'AAL considéré est l'AAL combiné incluant ainsi à la fois le risque dit « fluvial » et le risque « pluvial ».

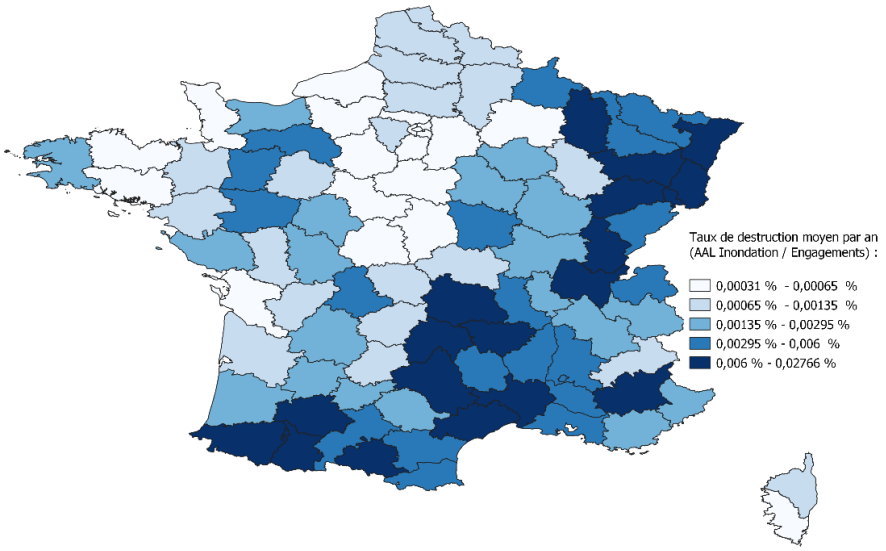

Figure 17 - Taux de destruction annuel moyen par département selon échantillon modélisé- branche habitation individuelle

Les zones à risques semblent globalement se concentrer autour des chaînes de montagnes davantage exposées aux précipitations régulières. Le Sud-est ressort également ce qui est cohérent avec la forte sinistralité inondation observée ces dernières années selon la CCR, comme étudié en première partie. À noter cependant que cette carte est construite uniquement avec un faible échantillon de zones IRIS de chaque département et qu'elle n'est donc pas nécessairement représentative de l'ensemble du territoire, mais permet cependant d'en avoir un premier aperçu. On pourra disposer d'une vue plus globale une fois le modèle inondation crée en partie IV.

En complément des AAL nous ayant permis de calculer ces taux de destruction, on dispose également des courbes AEP qu'on peut utiliser pour avoir une vue plus précise qu'une simple moyenne, et ainsi réaliser des études uniquement sur de faibles périodes de retour. On représente ci-dessous la courbe AEP obtenue en sortie des modélisations.

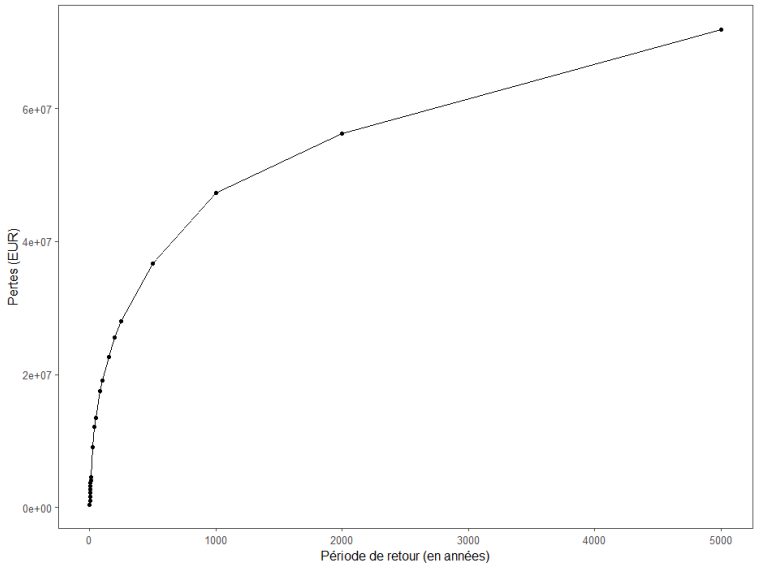

Figure 18 - Courbe AEP toute branche sur l'échantillon de risques modélisé

3) Cohérence des simulations avec la sinistralité observée

Afin de s'assurer que le modèle du prestataire extérieur est de bonne qualité et fournit une vision réaliste de la charge inondation, il est important de passer par une étape de comparaison entre les simulations effectuées et la sinistralité inondation observée. Cependant, il est souvent très compliqué de mettre en place une méthode robuste permettant avec certitude de s'assurer de la qualité d'un modèle, à partir uniquement de la sinistralité passée. En effet, si l'on décide d'utiliser un modèle CAT, c'est justement car la sinistralité passée ne suffit pas pour avoir une vision globale du risque inondation et ne permet pas de prendre en compte les évènements extrêmes qui n'ont pas été forcément observés ces dernières années.

Pour la comparaison, on utilise la sinistralité de 2010 à 2019 sur les zones IRIS modélisées (la submersion marine n'étant pas modélisée par le modélisateur, la tempête Xynthia de 2010 sera retirée de l'étude). Nous disposons ainsi d'un historique de 10 ans afin d'évaluer 50 000 années de simulations. Plutôt que de comparer directement la moyenne empirique des pertes annuelles sur ces 10 années avec celle sur les 50 000, on s'intéresse plutôt à la distribution de ces pertes, la moyenne étant très sensible aux valeurs extrêmes qui n'ont pas été observées dans le passé proche. On met alors en place la méthode ci-dessous :

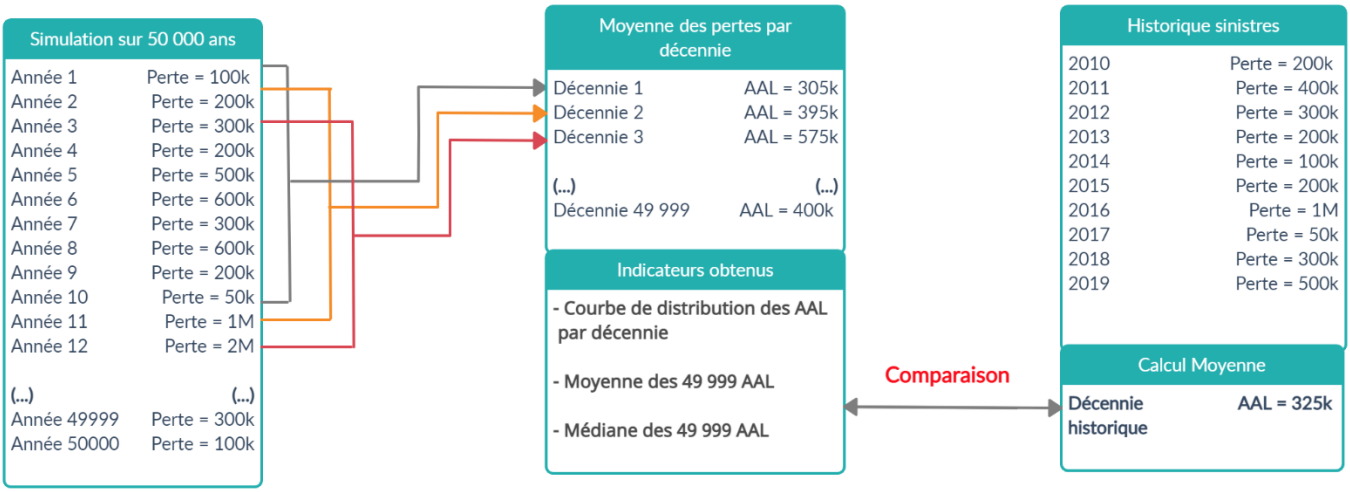

Figure 19 - Schéma explicatif du processus de comparaison entre les simulations et l'historique sinistre disponible

Du côté modélisation, l'idée est de rééchantillonner les 50 000 années simulées à l'aide d'une fenêtre mobile avec incrément de 1 an. On obtient ainsi 49 999 décennies pour lesquelles on dispose pour chacune d'une moyenne empirique des pertes sur 10 ans (l'AAL). On peut alors représenter la courbe de distribution des AAL par décennie selon les simulations et la comparer avec l'AAL observée sur la dernière décennie. Cette méthode permet de s'assurer de la bonne calibration du modèle et des courbes de vulnérabilité avec la typologie des risques de Groupama. Par exemple s'il apparait que 95% des décennies modélisées ont un AAL supérieur ou à l'inverse inférieur à la décennie historique observée alors qu'en parallèle cette décennie ne semblait pas particulièrement extrême selon les observations, cela pourrait nous amener à nous interroger sur la bonne calibration du modèle.

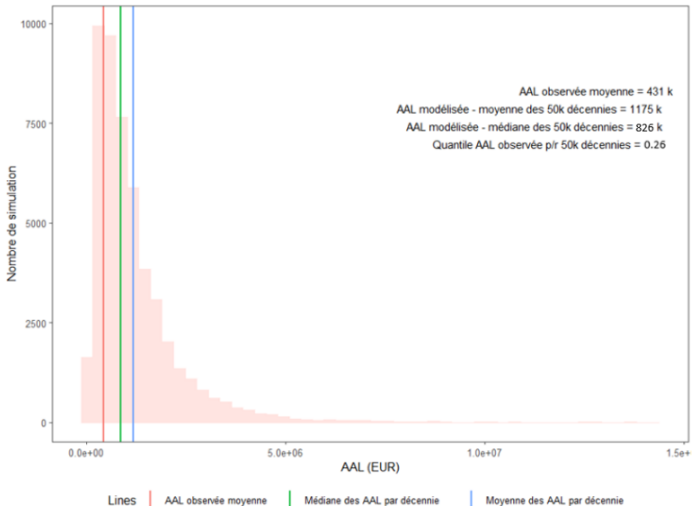

Figure 20 - Répartition de la charge annuelle moyenne modélisée et comparaison avec la charge observée

D'après les modélisations il apparait que 74% des décennies ont une charge annuelle moyenne supérieure à ce qui a été observé de 2010 à 2019. Ainsi la dernière décennie observée a été plutôt favorable selon le modèle avec un AAL au niveau du premier quartile de la série des pertes. En s'intéressant de plus près à l'historique de sinistralité fourni par la CCR, il apparait en effet que la dernière décennie a une charge annuelle moyenne près de 30% inférieure à ce qui a été observé dans la période 1989 – 2008. Ainsi un AAL historique sur les dix dernières années inférieures à la médiane selon le modèle, ne semble pas totalement incohérent au vu du faible historique dont on dispose.

Date	AAL (M €)
1989 - 2008	~ 600
2009 - 2019	~ 466
1989 - 2019	~ 546

Tableau 5 - Charge annuelle inondation non-auto sur le marché français avec exclusion de la tempête Xynthia (Source : CCR)

Finalement, il apparait que les modélisations semblent cohérentes avec l'historique sinistre dont nous disposons. On décide ainsi de se baser sur ces résultats de modélisation afin d'alimenter la variable à expliquer de notre modèle, à savoir l'AAL inondation par risque. De cette manière, on évitera de se baser sur notre base sinistre inondation avec une profondeur trop faible et on essayera de capter la dépendance géographique de chaque site assuré en introduisant diverses variables propres à la géolocalisation du risque, telles que la distance au cours d'eau, l'altitude ou encore la pluviométrie.

III. Mise en place du modèle : enrichissement de la base de risques

Avant de commencer les modélisations, l'objectif de cette partie est de récupérer des variables physiques et climatiques qui pourraient nous aider à caractériser le risque inondation. Nos modèles seront ainsi alimentés d'un côté par des variables déjà présentes dans nos bases telles que la branche assurée, la somme assurée ou le nombre d'étages, et de l'autre par des variables enrichies dans cette partie et qui seront calculées à partir de la géolocalisation du risque. Il est à noter que l'ensemble des variables exposées dans cette partie ne seront pas nécessairement présentes dans le modèle final, et que le caractère explicatif de ces variables sera étudié dans la partie IV. La finalité de cette partie est donc de rassembler un maximum de variables potentiellement intéressantes pour nos modèles, quitte à ce que certaines s'avèrent finalement non utilisées.

1) Calcul des variables physique

a. Distance au cours d'eau le plus proche

La première variable qui nous semblait intéressante pour caractériser le risque inondation est la distance au cours d'eau le plus proche, étant donné que lors de leur débordement ce sont souvent les habitations les plus proches qui sont les plus touchées. Cependant, une des premières difficultés était de différencier les différents types de cours d'eau, se trouver au bord d'un grand fleuve ou d'un petit ruisseau ne présente pas le même type de risque. Également, le modélisateur nous fournit un AAL fluvial ainsi qu'un AAL pluvial, l'un caractérisant des débordements de cours d'eau majeurs et l'autre davantage de ruisseaux et de petites rivières. Il était donc important de pouvoir différencier le type de cours d'eau. Pour cela, on dispose de plusieurs sources de données donnant une catégorisation des cours d'eau.

Les premières sources de données utilisées sont les données OpenStreetMap (OSM). Il s'agit d'un projet collaboratif qui met à disposition des données géographiques sous licence libre. Les cours d'eau proposés ont l'avantage de se baser directement sur les données satellites et suivent ainsi de manière très précise le tracé des cours d'eau. De plus, on distingue quatre catégories différentes, une catégorie *river* pour les rivières larges, une catégorie *stream* pour les ruisseaux et petites rivières, une catégorie *canal* pour les canaux navigables artificiels et finalement une catégorie *drain* pour les drains artificiels permettant l'évacuation de l'eau.

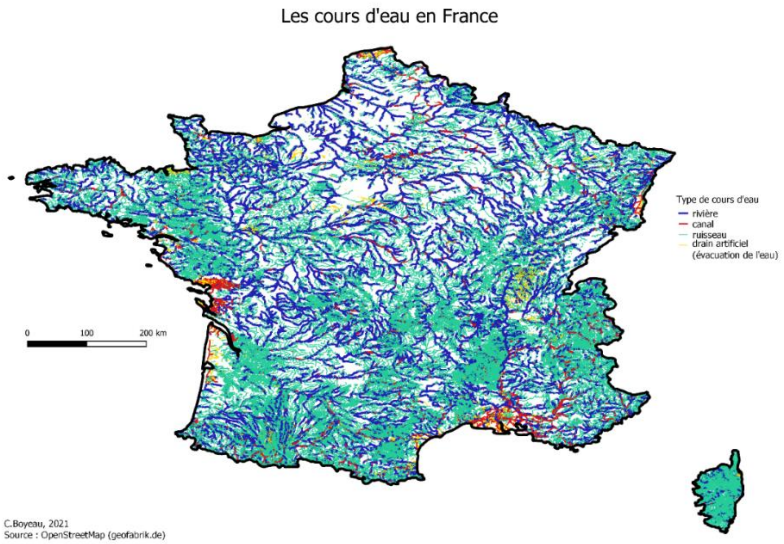

Figure 21 - Les cours d'eau en France Métropolitaine (Source : OpenStreetMap)

La deuxième source de donnée utilisée provient des tronçons hydrographiques de la BD Topage, il s'agit du nouveau référentiel hydrographique français mis à disposition par Sandre (le Service d'administration nationale des données et référentiels sur l'eau), en remplacement de la BD Carthage, et vise à être plus précise et plus exhaustive. Les tronçons hydrographiques désignent le découpage le plus fin d'un réseau hydrographique et peut composer aussi bien un cours d'eau que traverser un plan d'eau, la base fournit par Sandre dispose également d'une information sur la largeur de ces tronçons en considérant 3 catégories : inférieur à 15 mètres de largeur, entre 15 et 50 mètres ou bien plus de 50 mètres. Le risque fluvial considéré par le modélisateur étant basé sur les cours d'eau majeurs, cette information pouvait être pertinente pour ce risque. On obtient la carte ci-dessous après retraitement des données, en sélectionnant uniquement les tronçons d'origine naturelle, ainsi que les tronçons suffisamment longs, pour écarter notamment les simples plans d'eau larges.

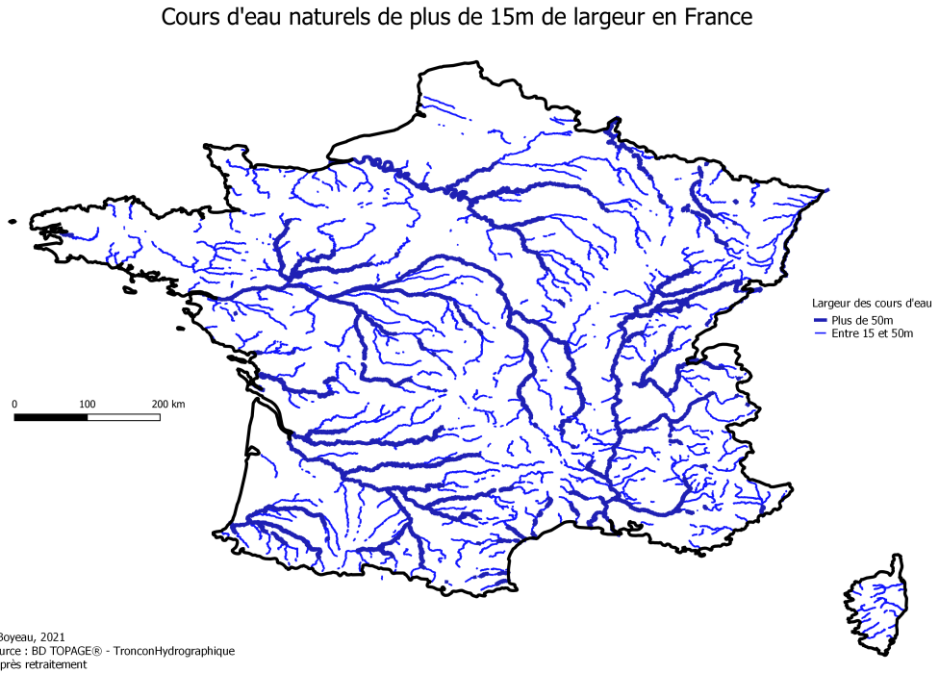

Figure 22 - Cours d'eau naturels de plus de 15m de largeur en France Métropolitaine (Source : BD TOPAGE)

Pour chaque site assuré, on calcule le point du cours d'eau le plus proche, c'est-à-dire sa latitude et sa longitude, dont on en déduit ensuite la distance avec le risque, on trouvera plus de détails sur ce calcul en annexe 1. Les variables suivantes ont été calculées :

- Distance au cours d'eau le plus proche (source : toutes classes OSM)
- Distance à la rivière la plus proche (source : classe *river* OSM)
- Distance au ruisseau le plus proche (source : classe *stream* OSM)
- Distance à la rivière moyenne la plus proche (source : classe 15-50m BD Topage)
- Distance à la rivière large la plus proche (source : classe >50m BD Topage)

Évidemment, la corrélation entre ces variables est très forte et l'ensemble ne sera pas utilisé dans les modèles, mais l'objectif est de repérer celles qui auront le plus grand pouvoir explicatif.

b. Altitude du risque

L'altitude du risque (c'est-à-dire la hauteur entre le point et le niveau de la mer) peut également être une variable intéressante à calculer, on peut s'imaginer que plus un point est en hauteur et moins il est susceptible d'être inondé. Pour calculer cette hauteur, on se base sur un modèle numérique d'élévation (MNE) fourni par l'agence européenne de l'environnement (EEA : European Environment Agency). Ce modèle est l'un des plus précis disponible en licence libre, capturant les élévations de terrains à des intervalles de 1 seconde d'arc soit environ tous les 30 mètres. Concrètement, on dispose d'un fichier *raster*, c'est-à-dire d'une grille composée de pixels d'environ 30 mètres par 30 mètres, chacun donnant l'altitude au centre du pixel.

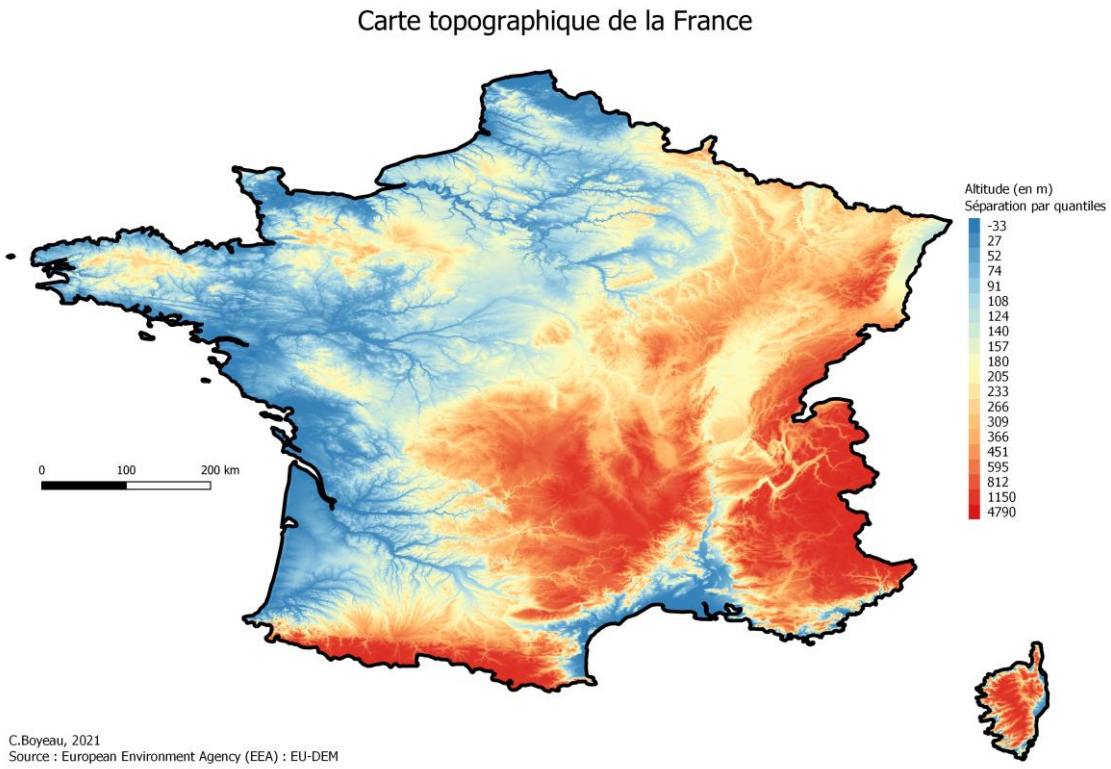

Figure 23 - Carte topographique de la France Métropolitaine

c. Différence d'altitude entre le cours d'eau le plus proche et le risque

Si l'altitude peut être une information intéressante, l'information seule n'est probablement pas suffisante, étant donné qu'un cours d'eau peut également se trouver en haute altitude. C'est pourquoi on calcule également l'altitude du point de cours d'eau le plus proche afin de déterminer ensuite la différence d'altitude entre le site assuré et le cours d'eau. On peut trouver ci-dessous un exemple de calcul de variables pour deux risques fictifs. Par exemple, l'un se trouve à Montmartre à une altitude de 124 mètres à presque trois kilomètres de la Seine se trouvant elle-même à une altitude de 30 mètres, on en déduit donc une différence d'altitude de 94 mètres.

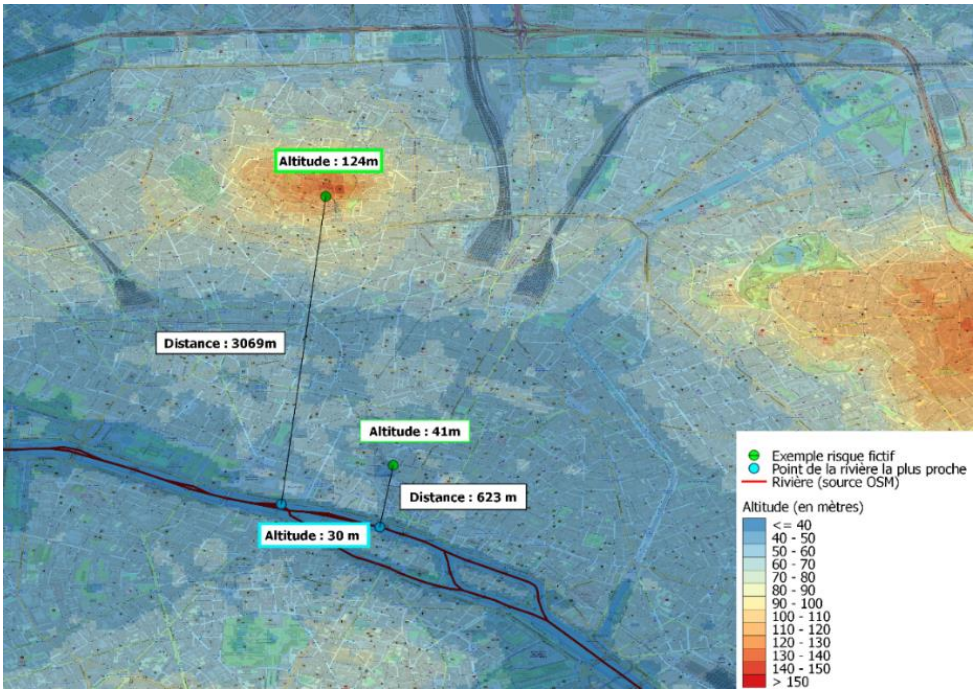

Figure 24 - Exemple de calcul de variables sur des sites assurés fictifs (rivière la plus proche, altitude, différence d'altitude)

De même que pour la distance au cours d'eau, on calcule cette variable pour les différentes sources dont on dispose :

- Différence d'altitude au cours d'eau le plus proche (source : toutes classes OSM)
- Différence d'altitude à la rivière la plus proche (source : classe *river* OSM)
- Différence d'altitude au ruisseau le plus proche (source : classe *stream* OSM)
- Différence d'altitude à la rivière moyenne la plus proche (source : classe 15-50m BD Topage)
- Distance d'altitude à la rivière large la plus proche (source : classe >50m BD Topage)

d. Coefficient de Manning

Une autre variable importante dans l'analyse du risque inondation concerne l'imperméabilisation du sol, qui peut grandement influencer la vitesse et le volume de ruissellement de l'eau. Si les sols ont normalement une certaine capacité d'absorption des crues, l'urbanisation grandissante des territoires et la bétonisation des sols qui en découle, empêche ce phénomène d'absorption et provoque ainsi davantage d'inondations par ruissellement ou aggrave simplement les phénomènes de crues.

Afin de prendre en compte cette caractéristique du sol, la première variable calculée est le coefficient de Manning. Ce coefficient provient de la formule de Manning-Strickler donnant une estimation de la vitesse moyenne d'un liquide s'écoulant sur une surface libre (c'est-à-dire une surface pour laquelle le fluide ne remplit pas complètement la section). La formule est la suivante :

$$V = \frac{1}{n} R^{\frac{2}{3}} S_f^{\frac{1}{2}}$$

Où V est la vitesse d'écoulement (en m/s), R le rayon hydraulique (en m), S_f la pente hydraulique (en m/m) et enfin n le coefficient de Manning. Ainsi, le coefficient de Manning est inversement proportionnel à la vitesse d'écoulement, plus celui-ci est petit et plus la vitesse d'écoulement est importante et donc moins le sol est susceptible d'absorber l'excédent d'eau. Pour attribuer à chacun des risques un coefficient de Manning, on s'est basé sur une étude menée par des chercheurs dans laquelle est calculé un coefficient de Manning selon l'occupation

des sols donnée par la base Corine Land Cover (CLC). On trouvera en annexe 3 un descriptif rapide de cette base. On remarque ainsi que les surfaces artificielles ont des coefficients de Manning allant de 0.013 à 0.025, tandis que les surfaces agricoles ont des coefficients plus élevés, entre 0.03 et 0.08. L'eau circule donc moins vite au niveau des surfaces agricoles.

LABEL1	LABEL2	LABEL3	Mannings n
1 Artificial surfaces	1.1 Urban fabric	1.1.1 Continuous urban fabric 1.1.2 Discontinuous urban fabric	0.013
	1.2 Industrial, commercial and transport units	1.2.1 Industrial or commercial units 1.2.2 Road and rail networks and associated land 1.2.3 Port areas 1.2.4 Airports	0.013
	1.3 Mine, dump and construction sites	1.3.1 Mineral extraction sites 1.3.2 Dump sites 1.3.3 Construction sites	0.013
	1.4 Artificial, non-agricultural vegetated areas	1.4.1 Green urban areas 1.4.2 Sport and leisure facilities	0.025
2 Agricultural areas	2.1 Arable land	2.1.1 Non-irrigated arable land 2.1.2 Permanently irrigated land 2.1.3 Rice fields	0.03
	2.2 Permanent crops	2.2.1 Vineyards 2.2.2 Fruit trees and berry plantations 2.2.3 Olive groves	0.08
	2.3 Pastures	2.3.1 Pastures	0.035
	2.4 Heterogeneous agricultural areas	2.4.1 Annual crops associated with permanent crops 2.4.2 Complex cultivation patterns 2.4.3 Land principally occupied by agriculture, with significant areas of natural vegetation 2.4.4 Agro-forestry areas	0.04 0.04 0.05 0.06
3 Forest and semi natural areas	3.1 Forests	3.1.1 Broad-leaved forest 3.1.2 Coniferous forest 3.1.3 Mixed forest	0.1
	3.2 Scrub and/or herbaceous vegetation associations	3.2.1 Natural grasslands 3.2.2 Moors and heathland 3.2.3 Sclerophyllous vegetation 3.2.4 Transitional woodland-shrub	0.04 0.05 0.05 0.06
	3.3 Open spaces with little or no vegetation	3.3.1 Beaches, dunes, sands 3.3.2 Bare rocks 3.3.3 Sparsely vegetated areas 3.3.4 Burnt areas 3.3.5 Glaciers and perpetual snow	0.025 0.035 0.027 0.025 0.01
4 Wetlands	4.1 Inland wetlands	4.1.1 Inland marshes 4.1.2 Peat bogs	0.04
	4.2 Maritime wetlands	4.2.1 Salt marshes 4.2.2 Salines 4.2.3 Intertidal flats	0.04
5 Water bodies	5.1 Inland waters	5.1.1 Water courses 5.1.2 Water bodies	0.05
	5.2 Marine waters	5.2.1 Coastal lagoons 5.2.2 Estuaries 5.2.3 Sea and ocean	0.07

Tableau 6 - Valeurs moyennes du coefficient de Manning basées sur les données de couverture terrestre du Corine Land Cover (Source : Papaioannou et al., 2018)

e. Imperméabilité des sols

La deuxième source de données utilisée pour capter l'imperméabilité du sol est une base de données fournie par l'EEA sous format *raster* (grille) disponible à très haute résolution par pixels de 10 mètres par 10 mètres. Elle permet ainsi d'avoir une vision beaucoup plus précise de l'artificialisation des sols en comparaison de ce qui a pu être calculé avec le coefficient de Manning. L'EEA met à disposition un coefficient d'imperméabilité du sol variant de 0% pour les sols les plus perméables, souvent d'origine naturelle, à 100% pour ceux totalement imperméables, souvent d'origine artificielle. On peut par exemple remarquer ci-dessous la prédominance des surfaces imperméables en région parisienne.

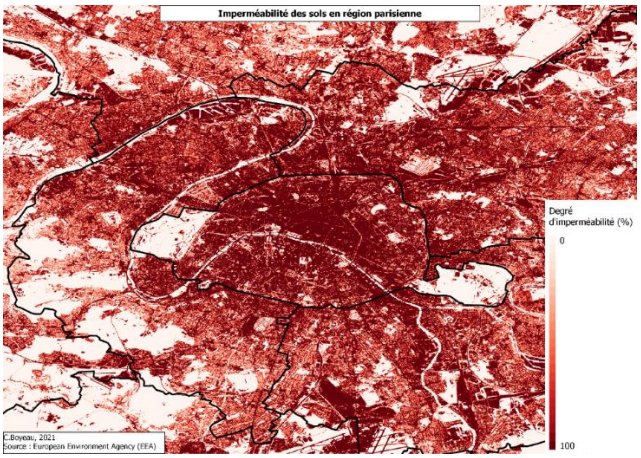

Figure 25 - Exemple : imperméabilité des sols en région parisienne (Source : EEA)

2) Variables liées au modèle d'écoulement de l'eau à huit directions (D8)

Les données topographiques ne permettent pas uniquement de calculer des altitudes et peuvent être utilisées afin de calculer de nombreuses variables pertinentes dans le cas du risque inondation, tel que des zones d'accumulation de l'eau ou encore des coefficients de pentes. Pour calculer ces variables, il est nécessaire d'introduire un modèle d'écoulement, généralement appelé 'modèle de flux à huit directions (D8)' basé sur l'approche de Jensen et Domingue introduite en 1988. Afin de raccourcir les temps de calcul, nous n'utiliserons pas le fichier topographique de l'EEA, mais plutôt le modèle d'élévation MERIT DEM qui possède une résolution inférieure et donc qui permet d'accélérer significativement les temps de calcul (résolution de 3 secondes d'arc contre 1 seconde d'arc pour les données de l'EEA)

a. Principe : Calcul des directions de flux

Cette approche repose sur le calcul de la direction potentielle de l'eau au niveau d'un pixel donné. Pour rappel, les fichiers topographiques sont fournis au format *raster*, c'est-à-dire que le fichier est une grille composée de pixels, ayant chacun une valeur d'altitude exprimée en mètres. Pour chaque pixel, celui-ci est entouré de huit autres pixels, quatre pixels adjacents et quatre pixels en diagonale. Parmi ces huit pixels, le modèle considère simplement que l'eau s'écoulera dans la direction de celui qui a la plus faible altitude. Une fois le pixel identifié, il suffit d'encoder la direction avec un numéro selon la correspondance ci-dessous.

Figure 26 - Encodage des directions de flux (Source : Arcgis.com)

La figure ci-dessous présente un exemple de calcul sur un échantillon. Le modèle prend bien en entrée un *raster* donnant l'altitude du terrain et calcule en sortie un *raster* de mêmes dimensions possédant dans chaque pixel l'encodage associé à la direction que devrait emprunter l'eau.

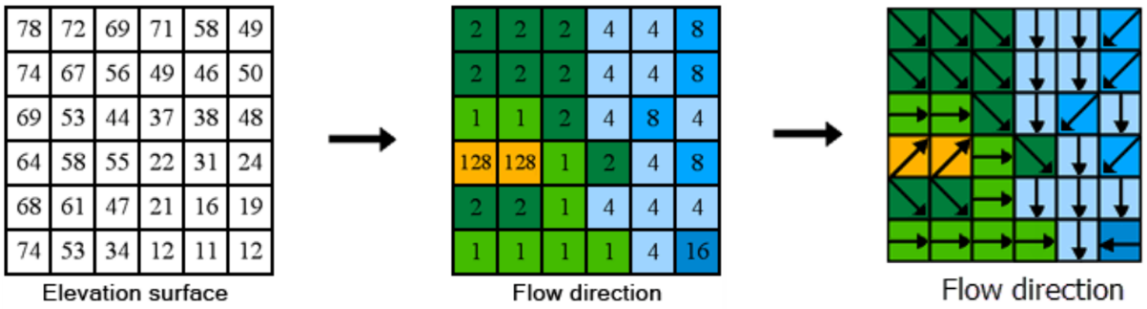

Figure 27 - Exemple de calcul de direction de flux (Source : Arcgis.com)

À noter également qu'avant d'exécuter cet algorithme il est souvent nécessaire de corriger le fichier topographique en « remplissant » les fosses. Une fosse désigne un pixel pour lequel l'ensemble des pixels qui l'entoure possède une altitude supérieure. Il est donc impossible pour le modèle D8 de déterminer une direction de flux, d'autant plus que ces fosses sont souvent causées par des erreurs dans le *raster* d'altitude. Il faut « remplir » ces fosses en augmentant leur élévation au niveau du terrain adjacent de plus faible altitude.

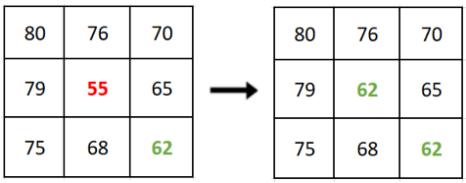

Figure 28 - Exemple de suppression de fosses

La direction de l'eau en tant que telle au niveau d'un risque particulier et donc d'un unique pixel, sans considération de l'ensemble du *raster* n'apporte aucune information intéressante et c'est pourquoi cette variable ne sera pas utilisée en tant que variable explicative des modèles. Cependant, cette information de la direction est primordiale pour le calcul des variables qui vont suivre.

b. Zones d'accumulations

Le calcul des zones d'accumulation permet d'utiliser la direction des flux afin d'identifier les zones vers lesquelles l'eau est le plus susceptible de s'écouler. On calcule pour cela un indicateur d'accumulation potentielle de l'eau, qu'on appelle surface contributive. La surface contributive d'une cellule est définie comme le nombre de pixels adjacents qui se déversent dans cette cellule, auquel on vient ajouter leur propre surface contributive. Lorsqu'aucune cellule adjacente ne se déverse dans la cellule considérée alors sa surface contributive par défaut est nulle. À noter qu'en fonction des logiciels, la valeur par défaut de la surface contributive est de 1 et non de 0. C'est d'ailleurs la valeur de 1 qui sera utilisé par la suite, notamment pour le calcul de l'indice d'humidité topographique qui sera introduit par la suite. On fait donc une somme cumulée de l'ensemble des surfaces contributives en partant du point supérieur jusqu'au point inférieur.

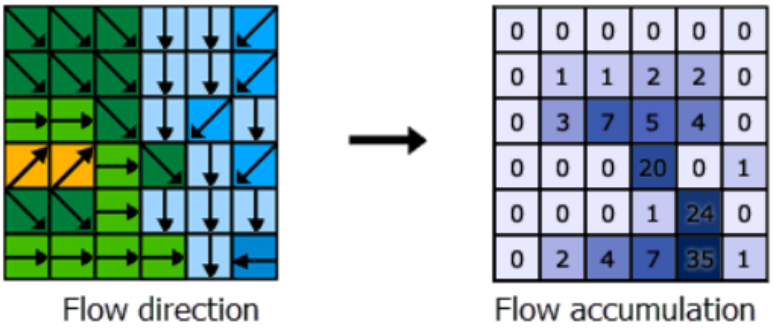

Figure 29 - Exemple de calcul d'accumulation de flux (Source : Arcgis.com)

De cette manière, on peut identifier les crêtes pour lesquelles la surface contributive est nulle et à l'inverse identifier les canaux d'écoulement qui ont des surfaces contributives élevées et vers lesquels l'eau serait davantage susceptible de s'accumuler et de faire des dégâts. Cette information peut donc être intéressante pour repérer les sites assurés à risque et l'on ajoute donc la surface contributive associée à chaque risque dans notre base de modélisation.

En représentant les surfaces contributives les plus élevées comme sur la figure ci-dessous, on peut identifier un réseau de drainage complet qui s'avère recouvrir une majorité des cours d'eau fournis par OpenStreetMap. De plus, on remarque de manière générale que plus la surface contributive est élevée et plus le cours d'eau correspondant est dense et large. Tandis qu'à l'inverse les zones d'accumulation plus faible telles qu'entre 50 et 500 correspondent davantage à des petits ruisseaux voir ne recouvrent aucun cours d'eau existant. Dans ce cas, cela permet d'identifier des canaux d'écoulement naturels non recouverts par de l'eau, mais potentiellement plus susceptibles d'en recevoir en cas de forte pluie par exemple.

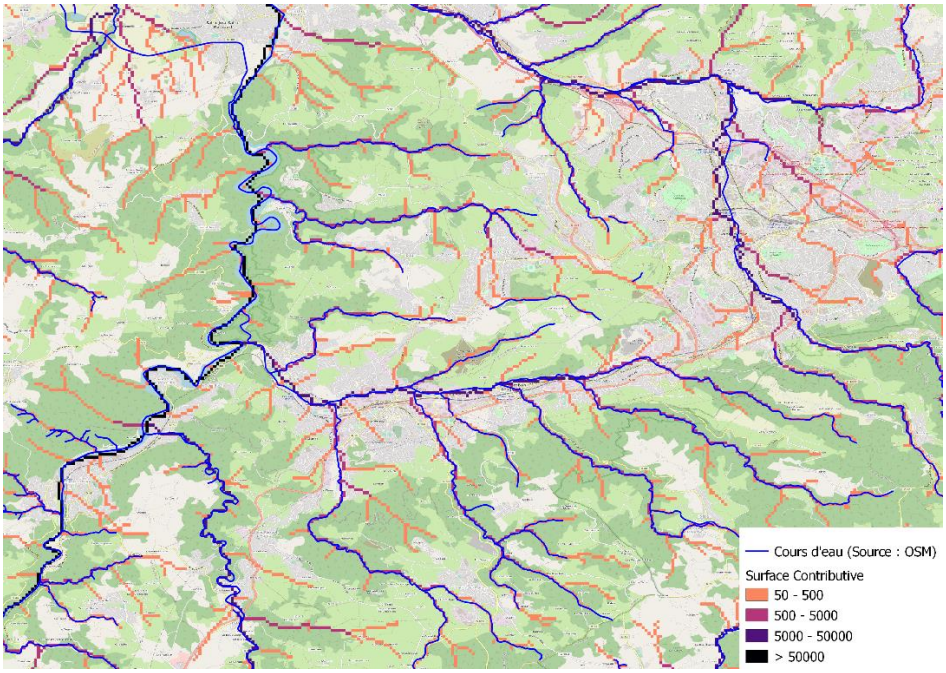

Figure 30 - Comparaison des cours d'eau OpenStreetMap et le réseau de drainage obtenu à partir des surfaces contributives

c. Distance au réseau de drainage le plus proche

Le modèle permet également de calculer une distance au cours d'eau le plus proche, mais exprimé non pas en mètres comme effectué dans la partie précédente, mais en comptant le nombre de cellules séparant le point de cours d'eau le plus proche et le site assuré, en suivant les directions d'écoulement de l'eau. Le fait de suivre les directions d'écoulement de l'eau pourrait éventuellement permettre à la distance calculée d'être plus cohérente avec la réalité comparée à une simple distance en mètres à vol d'oiseau. Les cours d'eau considérés sont cette fois-ci calculés à partir des surfaces contributives. On considère quatre réseaux de drainage distinct : l'ensemble des cellules pour lesquelles la surface contributive est supérieure à 50000, 5000, 500 puis 50. On a représenté ci-dessous un réseau de rivières obtenu à partir du seuil de 5000 et le *raster* de distance associé. On remarque par rapport à la figure ci-dessus que le réseau obtenu avec ce seuil de 5000 ne recouvre pas l'ensemble du réseau de rivières d'OpenStreetMap, mais pourrait s'avérer être une variable intéressante pour la prédiction du risque fluvial qui considère uniquement les cours d'eau larges.

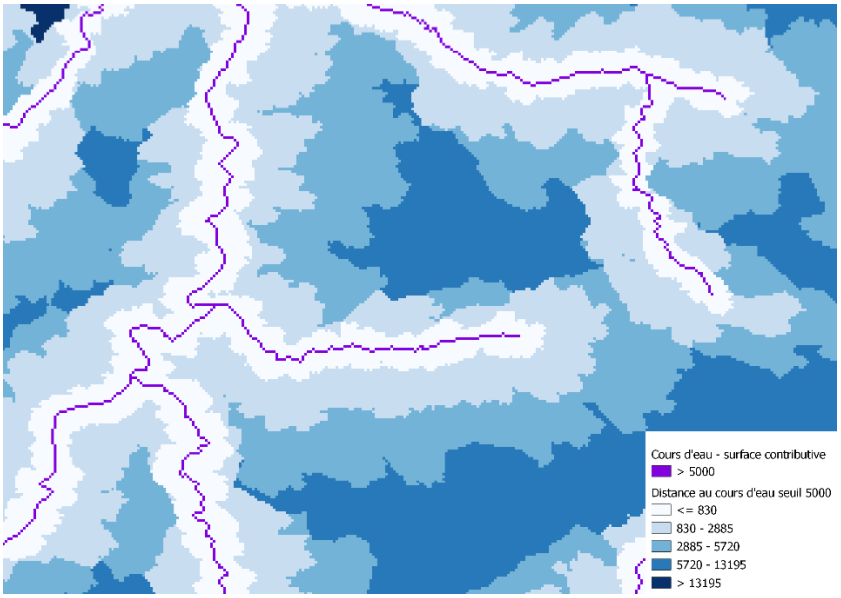

Figure 31- Distance au point du réseau de drainage le plus proche avec seuil de 5000

Les variables suivantes ont ainsi été calculées :

- Distance au réseau de drainage le plus proche avec seuil > 50000
- Distance au réseau de drainage le plus proche avec seuil > 5000
- Distance au réseau de drainage le plus proche avec seuil > 500
- Distance au réseau de drainage le plus proche avec seuil > 50

d. Pente

La pente au niveau du site assuré peut également être une information intéressante dans le cadre du risque inondation. Le calcul du coefficient d'inclinaison d'une pente (déclivité) se fait simplement comme étant :

$$Déclivité = \frac{\Delta h}{d}$$

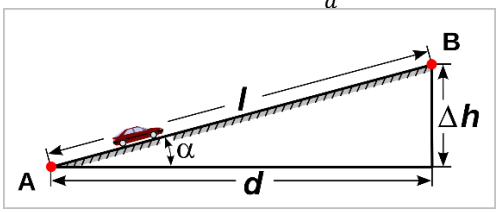

Figure 32 - Calcul de la déclivité d'une pente

Dans notre cas il suffit donc de calculer à partir du *raster* de directions de flux, la différence entre l'altitude de la cellule à partir de laquelle on part, avec l'altitude de la cellule vers laquelle on se déverse selon les directions de flux. Par la suite, il reste à diviser cette différence par la distance (en mètres) entre les deux centres des cellules. On a représenté ci-dessous le *raster* obtenu sur la France Métropolitaine. Logiquement, les chaînes de montagnes ressortent avec des pentes plus importantes.

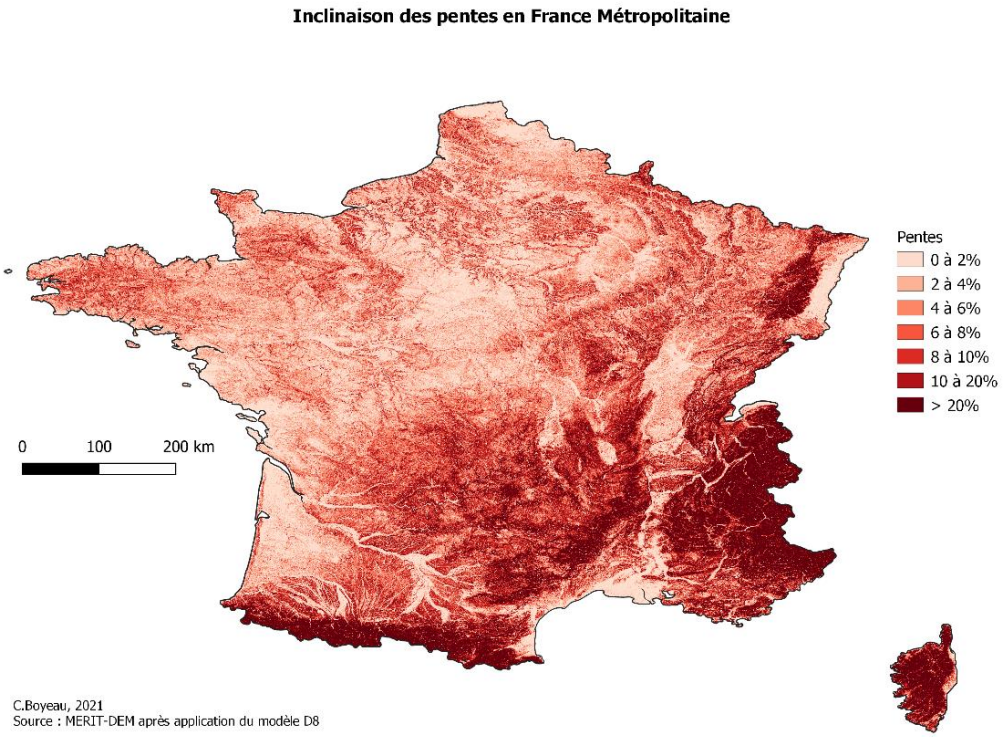

Figure 33 - Inclinaison des pentes en France Métropolitaine

e. Indice d'humidité topographique

L'indice d'humidité topographique (ou *Topographic Wetness Index – TWI*) est un indicateur d'humidité du sol permettant d'estimer où l'eau est susceptible de s'accumuler. L'indice est une fonction de la pente et de la zone de contribution en amont que l'on a appelé surface contributive :

$$TWI = \ln \frac{a}{\tan b}$$

Où :
a = surface contributive (en m^2)
b = inclinaison de la pente (en radians)

Pour plus de visibilités, on a représenté ci-dessous le calcul en se focalisant sur une zone géographique restreinte en l'occurrence le sud-est de la France. Logiquement, il apparait par exemple que les plans d'eau en bord de mer ont les indices d'humidité les plus élevés.

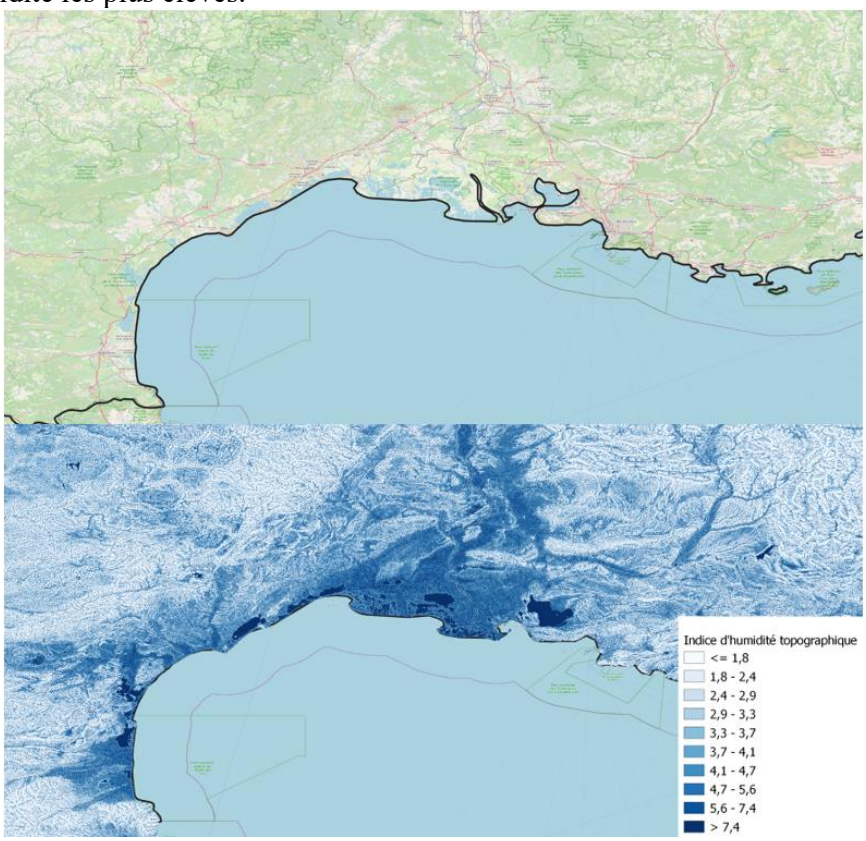

Figure 34 - Indice d'humidité topographique sur la côte méditerranéenne

3) Calcul des données climatiques : simulations DRIAS issues du projet CORDEX

Les données climatiques sont primordiales pour nos modélisations. Tout d'abord, car les scénarios construits par les modélisateurs dans le module aléa dépendent en partie de ces variables, c'est donc une information primordiale pour la compréhension des zones à risques. De plus, c'est principalement ce paramètre qui risque d'être impacté par le changement climatique. Il nous faut donc à la fois une vision des précipitations au climat actuel, mais également au climat futur afin d'ajuster nos modèles à ces changements et d'en évaluer les impacts sur la sinistralité future.

a. La modélisation du changement climatique : GCM et RCM

Afin d'étudier l'impact du changement climatique, étant donné qu'il n'est pas possible de se baser sur le passé récent pour faire des projections, les scientifiques utilisent des GCM. Les modèles climatiques globaux, usuellement appelés GCM pour *Global Climate Model* sont des modèles créés par la communauté scientifique dans le but de reproduire le plus fidèlement possible le comportement du climat terrestre. Ces modèles reposent sur des systèmes d'équations basés sur des lois de la physique, de la chimie et de la dynamique des fluides afin de reproduire au mieux le comportement du climat. Les GCM sont construits sur des grilles tridimensionnelles, les trois dimensions étant la latitude, la longitude et l'altitude. On peut visualiser cette grille ainsi que des exemples d'interactions physiques prises en compte sur le schéma ci-dessous.

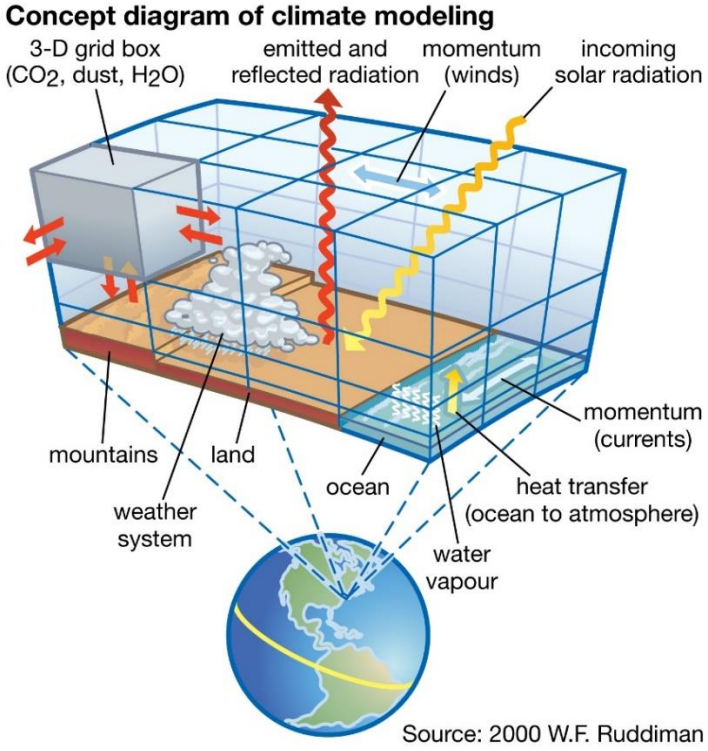

Figure 35 - Schéma d'ensemble d'un modèle climatique

Ces modèles sont globalement construits selon les étapes suivantes :

- Sélection des paramètres physiques considérés comme suffisant pour caractériser le système climatique et pour l'objectif de modélisation, en l'occurrence l'évolution long terme du climat (température moyenne de départ, pression, humidité, gaz à effets de serre au cours du temps, salinité de l'eau, etc.)
- Détermination du maillage considéré. Les RCM récents sont souvent construits sur des grilles d'environ 100 kilomètres de côté.
- On exprime les relations d'un compartiment à un autre par des lois physiques (par exemple les équations que l'atmosphère doit respecter en permanence, comme la conservation de l'énergie)
- On programme le modèle (conditions initiales, équations, etc.)
- On détermine les conditions initiales choisies précédemment pour chaque point de grille
- On lance le modèle : le modèle détermine sur la base des équations et des interactions entre l'ensemble des compartiments comment vont évoluer les paramètres de chaque point de grille (températures, précipitations, humidité, rayonnement, vent …) selon un intervalle de temps régulier (toutes les heures, tous les jours, tous les mois ...).

Une fois le modèle construit il suffit donc dans un premier temps de lancer le modèle selon des conditions initiales historiques. Puis de le relancer en modifiant ces conditions initiales (teneur en gaz à effet de serre au cours du temps par exemple) selon des hypothèses de changement climatique données par les scénarios du GIEC. Finalement, il sera possible de comparer l'évolution du climat entre la simulation aux conditions historiques et les simulations faites selon différents scénarios de changement climatique.

Cependant, ces modèles permettent surtout d'étudier l'évolution des paramètres météorologiques à grande échelle et la maille de 100 kilomètres peut être rapidement limitante lorsqu'il s'agit d'étudier des phénomènes à un point de vue plus local comme on souhaite le faire. De plus, les GCM captent difficilement les phénomènes extrêmes tels que des vents violents ou des précipitations intenses souvent liés à des phénomènes de petite échelle. Pour affiner les résultats des modèles climatiques globaux, les climatologues mettent au point des modèles de climat régionaux (RCM pour *Regional Climate Models*). La descente d'échelle effectuée par les RCM n'est effectuée que sur une partie du globe, l'Europe par exemple, et offre ainsi de plus hautes résolutions spatiales.

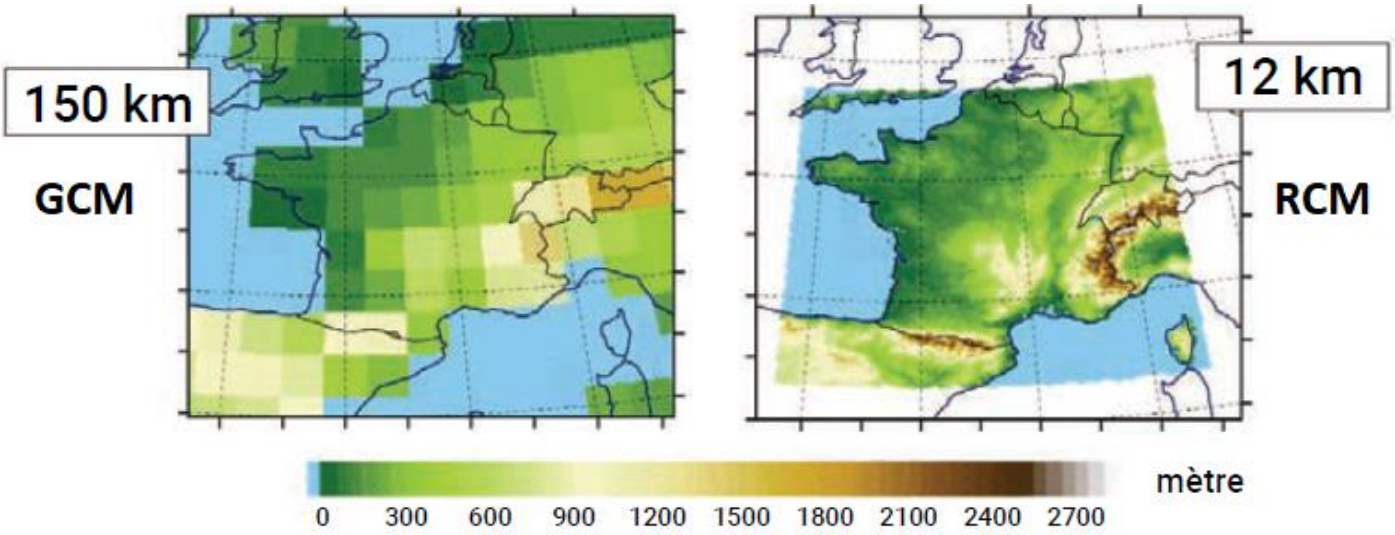

Figure 36 - Descente d'échelle dynamique depuis la modélisation globale jusqu'à la modélisation régionale (source : DRIAS)

b. Source des données : projet CMIP5 et CORDEX

Le projet d'intercomparaison des modèles couplés (CMIP pour *Coupled model intercomparison project*) est un projet lancé dans le cadre du programme mondial de recherche sur le climat (PMRC ou WCRP pour *World Climate Research Programme*). Ce projet a pour objectif de mettre en place une coordination scientifique et technique entre des dizaines de centres climatiques à travers le monde, dans le cadre de l'élaboration de modèles climatiques globaux. Il permet ainsi de mettre à disposition un ensemble de GCM différents, en y associant de nombreux résultats de recherches, par exemple concernant l'incertitude due à l'imperfection des modèles dans l'estimation du changement climatique. La cinquième phase du projet (CMIP5) a servi de base au cinquième rapport du GIEC paru en 2014.

Dans le même esprit, le projet CORDEX *(**CO**ordinated **R**egional climate **D**ownscaling **EX**periment)*, également mis en place dans le cadre du PMRC, permet d'assurer une coordination scientifique des laboratoires à l'échelle régionale afin d'élaborer un cadre commun pour la production de modèles climatiques régionaux (RCM) et d'en assurer leur standardisation (critères sur les variables nécessaires en sortie, sur le maillage spatial et temporel …).

Le programme Euro-CORDEX est la branche européenne du projet et rassemble une dizaine de laboratoires tels que Météo France ou l'Institut Pierre Simon Laplace (IPSL).

c. Simulations DRIAS-2020

Le projet DRIAS (**D**onner accès aux scénarios climatiques **R**égionalisés français pour l'**I**mpact et l'**A**daptation de nos **S**ociétés et environnement) est un projet soutenu par le ministère de la Transition écologique avec l'appui scientifique de l'IPSL, du CNRM et du Cerfacs, dont l'objectif principal est de faciliter l'accès et l'utilisation aux modèles climatiques régionaux sur le territoire français. Concrètement, plusieurs actions sont effectuées par les scientifiques du projet DRIAS :

- Sélection de GCM parmi l'ensemble CMIP5 et de couples GCM/RCM parmi l'ensemble Euro-CORDEX
- Application d'une méthode de correction de biais des modèles à partir des données historiques
- Calculs d'indicateurs à partir des données brutes (exemple : calcul du nombre moyen de jours de pluie par an à partir des précipitations quotidiennes sur la période considérée)
- Simplification du téléchargement des données et limitation à la France (Choix de la zone géographique, des indicateurs voulus, de la période considérée …)

d. Sélection des modèles climatiques et description des variables brutes obtenues

Pour notre étude, il est important de ne pas se baser uniquement sur un unique modèle climatique pour appréhender au mieux les différentes sources d'incertitude. Pour cela les experts du projet DRIAS ont constitué un ensemble de modèles selon différents critères :

- Choisir uniquement des modèles appliqués sur au moins deux scénarios d'émissions parmi les trois considérés à savoir le RCP 2.6, le RCP 4.5 et le RCP 8.5.
- Prise en compte des RCM jugés plus réaliste sur l'Europe
- Diversification des RCM
- Sélection préférentielle pour les modèles issus des centres climatiques français (Aladin et WRF)
- Limitation à 12 couples GCM/RCM afin de faciliter les traitements pour les utilisateurs
- Rejet des couples GCM/RCM présentant une erreur connue
- Optimisation de la dispersion du changement climatique simulé par les couples sélectionnés
- Sélection des couples ayant une cohérence physique entre les modèles GCM et RCM

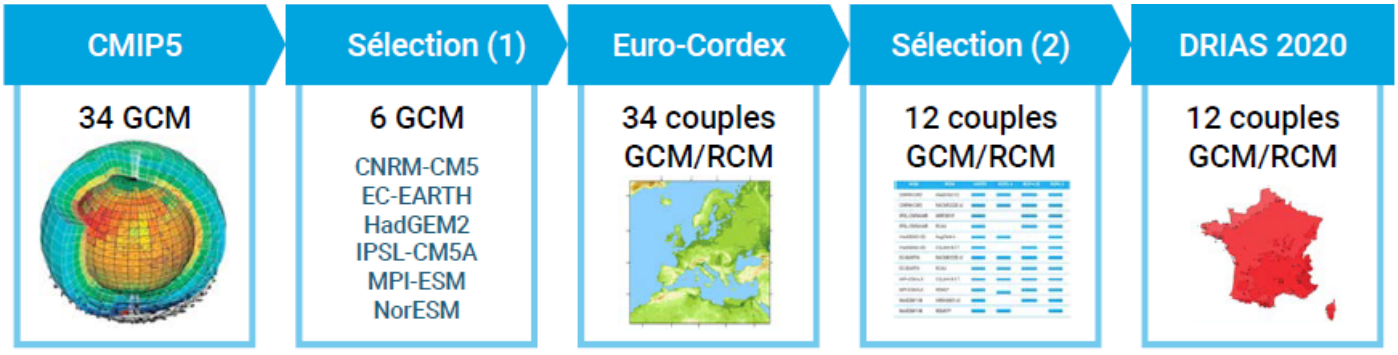

Figure 37 - Processus de sélection des modèles pour le jeu DRIAS-2020 (source : DRIAS)

On retrouve sur la table ci-dessous le détail des couples GCM/RCM sélectionnés. On possède donc :

- 12 modèles simulés aux conditions historiques
- 8 modèles simulés selon le scénario RCP 2.6
- 10 modèles simulés selon le scénario RCP 4.5
- 12 modèles simulés selon le scénario RCP 8.5

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*

Tableau 7 - Simulations sélectionnées pour le jeu DRIAS-2020 (Source : DRIAS)

Chacun de ces modèles climatiques permet de simuler l'évolution de différentes variables à un pas de temps quotidien. Les variables disponibles concernent notamment la température, les précipitations, l'humidité, les rayonnements et le vent. On s'intéressera dans notre étude aux précipitations totales quotidiennes fournies en millimètres.

Les simulations en conditions historiques seront étudiées sur la période de référence 1976-2005, la période de 30 ans d'études est une base standard pour les analyses climatiques et correspond à la période disponible la plus récente dans les simulations historiques Euro-CORDEX.

Les simulations en conditions futures selon les différents scénarios sont effectuées sur la période 2006-2100. On dispose donc pour chacun des modèles et chacun des scénarios, les précipitations quotidiennes simulées sur ces 95 années. Afin de travailler sur une base comparable aux conditions historiques, on calculera nos indicateurs de précipitations selon trois périodes de 30 ans : une période horizon proche (2021-2050), une période horizon moyen (2041-2070) et enfin une période horizon lointain (2071-2100).

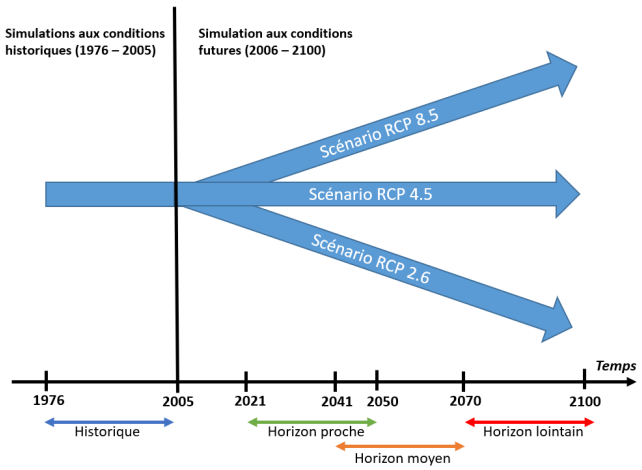

Figure 38 - Structure des modélisations : période et scénarios considérés

e. Correction de biais : méthode quantile-quantile

La méthode quantile-quantile consiste à élaborer une fonction permettant de ramener la distribution des valeurs simulées aux conditions historiques avec celles des valeurs réellement observées. En effet, les variables obtenues aux conditions historiques ne sont que des simulations calculées à partir des paramètres météorologiques historiques rentrés en condition initiale du modèle, mais la façon de se comporter du modèle une fois les conditions initiales renseignées a probablement différé de ce qui a pu être réellement observé. C'est pourquoi il est nécessaire d'effectuer un travail de mise en cohérence de la distribution simulée avec celle observée, puis d'appliquer ces mêmes transformations pour les simulations aux conditions futures.

Pour mettre en œuvre cette méthode, il est nécessaire de disposer de données de référence historique sur la période 1976-2005 considérée. Pour cela, la base de données Safran a été utilisée, elle est produite par Météo-France et est constituée de données horaires sur la France métropolitaine avec une résolution de 8 kilomètres, comparable à la maille de 12 kilomètres des données Euro-CORDEX. On associe à chaque point de grille des modèles RCM, le point de grille de la base Safran le plus proche.

Ainsi, on dispose pour chaque point de grille de 8 kilomètres et pour chaque jour entre 1976 et 2005, du cumul des précipitations quotidiennes, selon l'historique SAFRAN et selon le modèle RCM considéré. On applique la méthode quantile-quantile ci-dessous indépendamment pour chaque point de grille :

Soit Q^k_{histo} le quantile d'ordre k de la distribution du cumul des précipitations quotidiennes selon l'historique et $Q^k_{modèle}$ le quantile d'ordre k de la distribution des précipitations quotidiennes selon le RCM choisi. On calcule ces quantiles par pas de 1%. Pour tout k entre 0 et 1 par pas de 0,01 on calcule le coefficient de correction suivant :

$$Correc_k = \begin{cases} 0 & si\ Q^k_{modèle} = 0 \\ \dfrac{Q^k_{histo}}{Q^k_{modèle}} & si\ Q^k_{modèle} \neq 0 \end{cases}$$

Notons $CP(j)$ le cumul des précipitations du jour j obtenu en sortie du RCM.
Soit l'ordre k associé à CP(j) tel que k vérifie : $Q^k_{modèle} \leq CP(j) \leq Q^{k+0,01}_{modèle}$ (Sauf si CP(j) $\leq Q^{0,01}_{modèle}$ alors k =0.01)

Notons $CP_{corrigé}(j)$ le cumul des précipitations après correction du biais, tel que :

$$CP_{corrigé}(j) = CP(j) * Correc_k$$

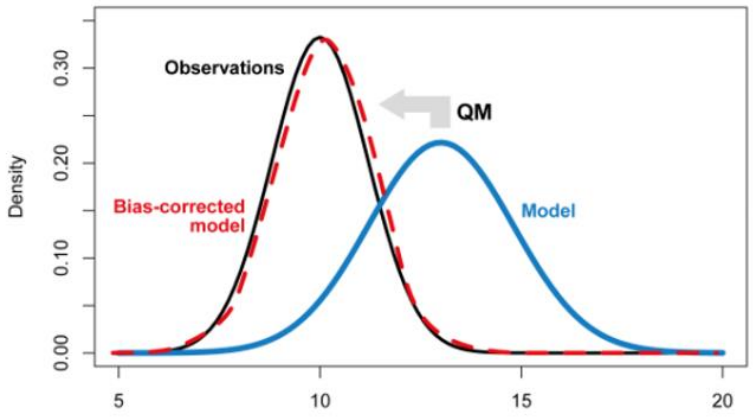

Figure 39 – Exemple de correction du modèle aux conditions historique par méthode quantile-quantile (Source : MeteoSwiss)

Cette méthode consiste en réalité à considérer que le modèle est capable de prédire la distribution des variables climatiques, mais pas la valeur exacte de chaque quantile. Pour les simulations aux conditions futures, on réutilise le même coefficient de correction afin d'appliquer les mêmes transformations que sur le modèle aux conditions historiques. On fait donc ici une hypothèse forte de stationnarité selon laquelle la fonction de transfert (représentée par les coefficients de correction) calibrée sur la période du passé reste valable dans le futur.

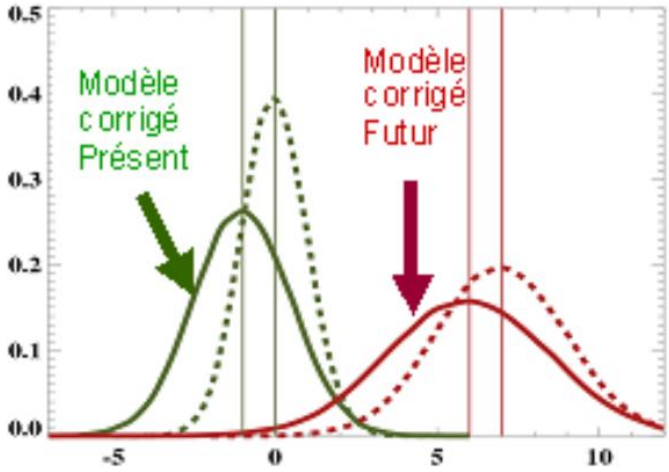

Figure 40 - Exemple de correction des RCM historiques et futurs (Source : DRIAS)

L'ensemble de modèles DRIAS-2020 est basé sur la méthode ADAMONT largement inspirée de la méthode quantile-quantile. Elle permet de prendre en compte les différents types de temps en considérant les centiles par saisons et par régimes de temps, et non pas sur la totalité de la période comme avec la méthode quantile-quantile. Les régimes de temps sont définis par le CNRM de la manière suivante : 'La circulation atmosphérique de grande échelle aux latitudes tempérées est caractérisée par des fluctuations des courants-jets entre différents états quasi stationnaires qu'on appelle les régimes de temps'. Leur prise en compte permet d'augmenter les chances de respecter l'hypothèse forte de stationnarité, en restant dans le même type de temps et dans la même saison.

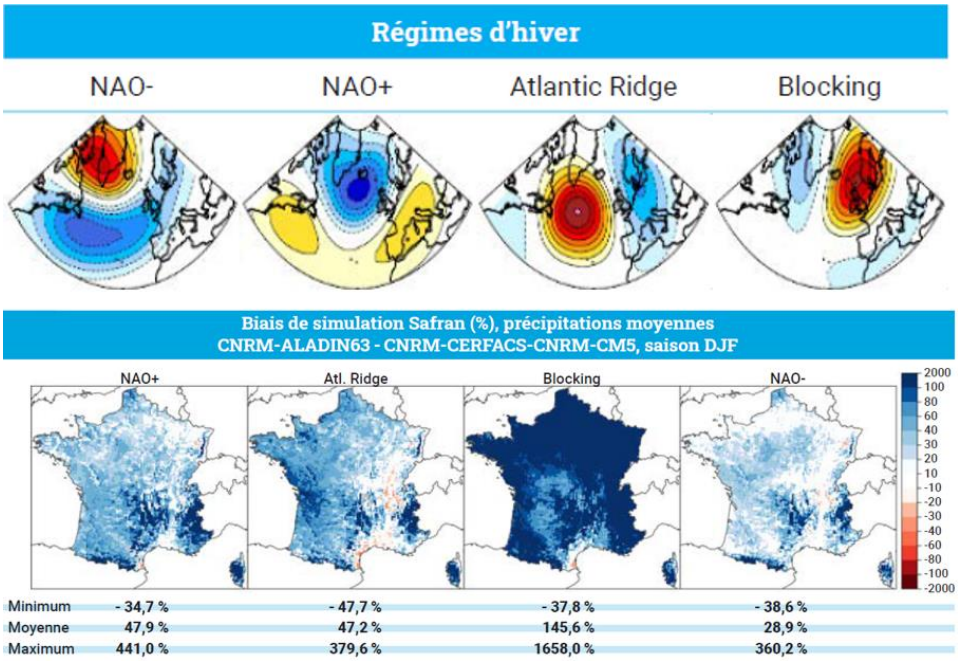

Figure 41 – Schéma des régimes d'hiver et biais de simulation pour ces régimes sur la saison décembre/janvier/février (Source : DRIAS)

f. Indicateurs climatiques calculés

Une fois les données brutes corrigées des biais, la dernière étape reste le calcul d'indicateurs climatiques pour les précipitations. En effet, on dispose seulement pour le moment du cumul des précipitations quotidiennes pour chaque jour sur une période donnée, on aimerait créer des indicateurs représentatifs des comportements moyens ou extrêmes, afin de pouvoir étudier leur évolution avec le changement climatique.

On calcule ces indicateurs sur des périodes de 30 ans, à savoir 1976-2005 pour les modèles aux conditions historiques, et trois périodes distinctes pour les modèles aux conditions futures (horizon proche : 2021-2050, horizon moyen : 2041-2070, horizon lointain : 2071-2100) et pour chacun des trois scénarios RCP 2.6, 4.5 et 8.5.

Notons CP_i le cumul des précipitations du jour i (i allant de 1 à N, N étant le nombre de jours totaux sur la période de 30 ans étudiée). On calcule les indicateurs suivants :

- Moyenne des précipitations quotidiennes de la période (mm) :

$$CPM = \frac{1}{N}\sum_{i=1}^{N} CP_i$$

- Nombre de jours de pluie moyen par an (jours) :

$$Nb_{1mm} = Nombre\ de\ jours\ de\ pluie\ moyen\ par\ an\ pour\ lesquels\ CP_i \geq 1mm$$

- Précipitations moyennes des jours pluvieux (mm) :

$$CPP = \frac{1}{Nb_{1mm}}\sum_{i=1}^{N} CP_i \quad si\ CP_i \geq 1mm$$

- Nombre de jours de fortes précipitations moyen par an (jours) :

$$Nb_{20mm} = Nombre\ de\ jours\ moyen\ par\ an\ pour\ lesquels\ CP_i \geq 20mm$$

- Précipitations quotidiennes intenses (mm) :

$$CP90 = 90eme\ centile\ des\ précipitations\ quotidiennes\ sur\ la\ période$$

- Précipitations quotidiennes extrêmes (mm) :

$$CP99 = 99eme\ centile\ des\ précipitations\ quotidiennes\ sur\ la\ période$$

- Période de sécheresse maximale (jours) :

$$Nb_S = Max(Nombre\ de\ jours\ consécutifs\ pour\ lesquels\ CP_i < 1mm)$$

- Nombre maximum de jours pluvieux consécutifs (jours) :

$$Nb_P = Max(NNombre\ de\ jours\ consécutifs\ pour\ lesquels\ CP_i \geq 1mm)$$

- Fraction des précipitations quotidiennes intenses (%) :

$$F_{90} = 100 * \frac{\sum_{i=1}^{N} CP_i \ \ si \ CP_i > CP90}{\sum_{i=1}^{N} CP_i}$$

Une fois l'ensemble de ces indicateurs calculés pour chacun des 12 couples GCM/RCM vient la question du choix du modèle climatique à considérer pour enrichir notre base de risque et alimenter notre futur modèle inondation. Pour cela, on décide de considérer l'ensemble de ces couples, en sélectionnant pour chaque point de grille et pour chaque indicateur, la médiane, le quantile à 5% et le quantile à 95% des 12 valeurs de l'indicateur (obtenu à partir de chacun des 12 modèles). À noter que pour le scénario RCP 4.5 on ne dispose que de dix modèles, et huit modèles pour le RCP 2.6, on effectue donc les quantiles sur dix et huit indicateurs dans ces cas respectifs. On effectue le calcul des quantiles par une interpolation linéaire. Par exemple dans le cas de 12 indicateurs la médiane correspond au quantile d'ordre 0,5 et est obtenue par interpolation linéaire entre la statistique d'ordre 6 (quantile d'ordre $\frac{6-1}{12-1} \approx 0{,}45$) et la statistique d'ordre 7 (quantile d'ordre $\frac{7-1}{12-1} \approx 0{,}55$).

On peut visualiser ci-dessous un exemple avec 12 valeurs comme c'est le cas pour les simulations aux conditions historiques et pour le scénario RCP 8.5. Cette méthode nous permettra d'avoir une vision médiane des modèles, mais également une vision extrême des projections climatiques nous permettant de considérer les simulations les plus optimistes et pessimistes pour chaque point de grille.

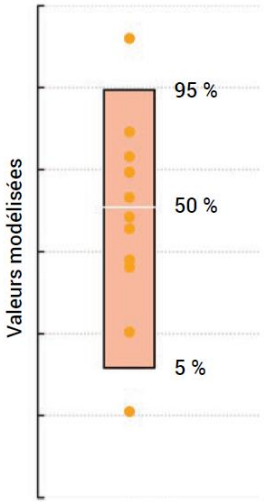

Figure 42 - Schéma : calcul de la médiane, du quantile à 5% et 95% des indicateurs climatiques (Source : DRIAS)

On trouvera ci-après les indicateurs précipitations obtenus en faisant la médiane des simulations aux conditions historiques. Ce sont ceux que l'on utilisera pour notre modèle inondation, on associe donc à chacun des risques de notre base, les indicateurs correspondant à leur géolocalisation. On attribue également les indicateurs obtenus sur les simulations aux conditions futures selon les différents scénarios et horizons, ce qui nous permettra dans la dernière partie de rejouer notre modèle inondation avec ces nouvelles valeurs.

L'ensemble des variables calculées dans cette partie sont résumées en annexe 4.

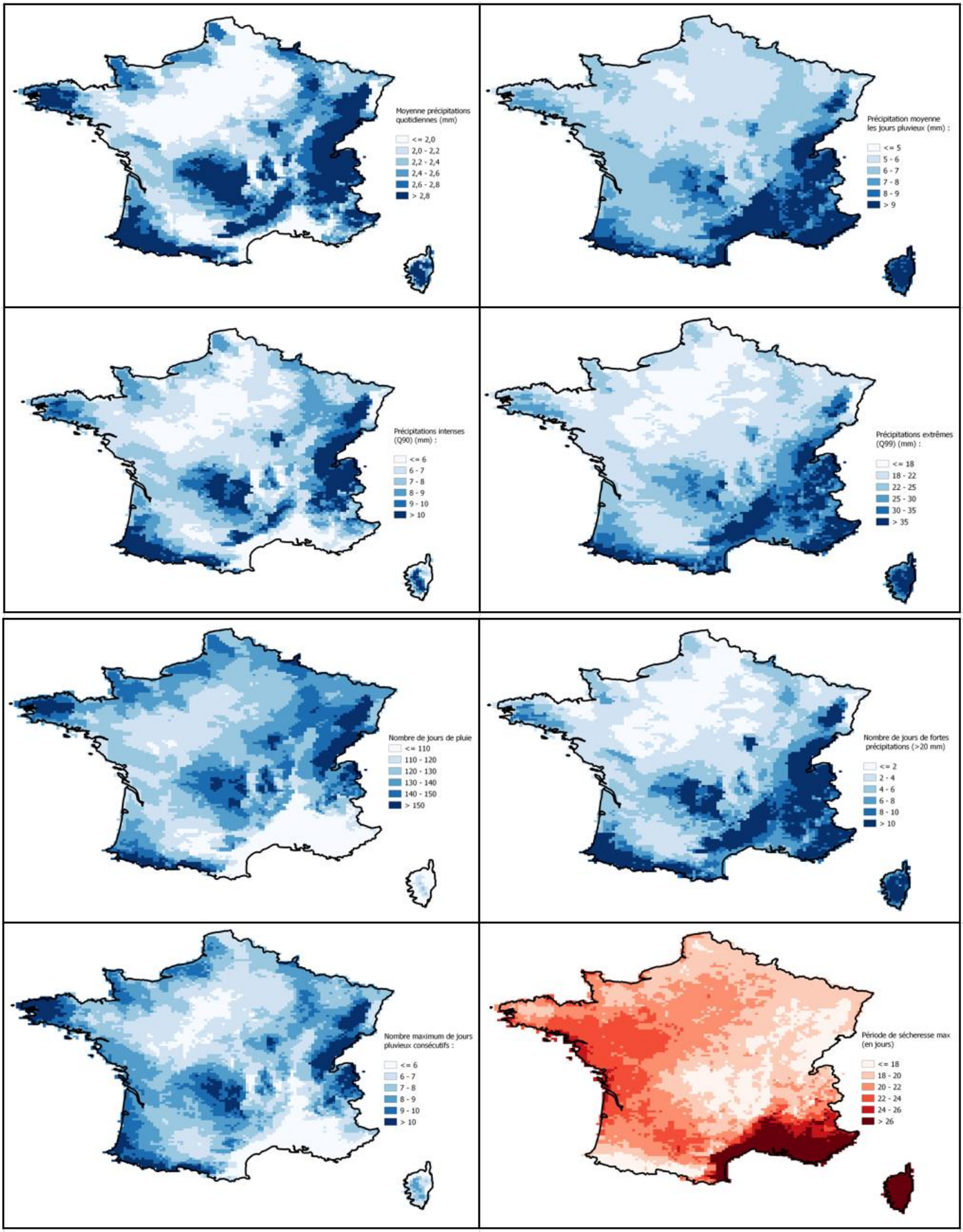

Figure 43 - Indicateurs précipitations selon médiane des simulations historiques 1976-2005 de l'ensemble DRIAS-2020

IV. Construction d'un modèle prédictif de la sinistralité inondation

Maintenant que nous disposons d'une base de modélisation enrichie de variables climatiques et physiques, l'objectif est désormais d'établir un lien entre ces variables et les charges inondations fournies par le modélisateur. On disposera ainsi d'un modèle permettant de faire la traduction de l'aléa climatique en une charge de sinistralité. Finalement, on pourra dans la dernière partie utiliser l'évolution de l'aléa climatique selon les projections du climat présentées précédemment pour en déduire un impact sur la sinistralité future.

1) Préparation des modélisations

a. Variable à modéliser : taux de destruction

Pour nos modélisations, on a décidé de faire la prédiction des taux de destruction moyens par an pour le risque « fluvial » et « pluvial ». Ils permettent de représenter le potentiel destructeur des inondations pour chaque site assuré, en corrigeant les charges moyennes par an avec la somme assurée. On obtient donc :

$$Taux\ de\ destruction\ moyen\ par\ an - risque\ fluvial = \frac{AAL\ fluvial}{Somme\ assurée}$$

$$Taux\ de\ destruction\ moyen\ par\ an - risque\ pluvial = \frac{AAL\ pluvial}{Somme\ assurée}$$

Cet indicateur ne prend cependant pas en compte la distinction entre les sommes assurées relatives au bâtiment, au contenu ou aux pertes d'exploitation (PE). Afin de prendre en compte la différence de vulnérabilité entre les différentes garanties assurées, on introduit deux variables supplémentaires qui rentreront en compte dans les variables explicatives de nos modèles :

$$Taux\ d'engagement\ contenu = \frac{Somme\ assurée\ contenu}{Somme\ assurée}$$

$$Taux\ d'engagement\ PE = \frac{Somme\ assurée\ PE}{Somme\ assurée}$$

b. Préparation des bases d'entraînement, de test et validation croisée

Pour effectuer nos modélisations, on entraînera nos modèles sur la base de risques présentée en deuxième partie et qui a été envoyée au modélisateur externe. Cette base représente environ 250 000 risques répartis sur 2500 zones IRIS choisies aléatoirement sur le territoire français. Elle servira donc d'exemple à nos modèles lors de la phase d'apprentissage.

Une fois le modèle optimal sélectionné, il est nécessaire de disposer de données afin d'évaluer la performance finale du modèle sur des risques qu'il n'a jamais vus. Pour cette phase, on dispose d'autres données envoyées aux modélisateurs dans le cadre d'une étude de Groupama sur des zones identifiées en interne comme étant à risques. Cette base représente un total d'environ 750 000 sites assurés. On a fait le choix de ne pas utiliser ces données pour la phase d'apprentissage étant donné que les risques présents dans cette base ont été au préalable identifiés par Groupama comme étant à risque d'inondation. La typologie de ces sites assurée n'est donc pas représentative de

l'ensemble de notre base, ce qui aurait pu se traduire par une mauvaise adaptation du modèle sur de nouveaux exemples, moins risqués, et donc une potentielle surestimation de la charge inondation totale. C'est pourquoi on a préféré effectuer nos modèles sur un échantillon sélectionné aléatoirement sur le territoire. Les zones communes entre la base aléatoire et la base de zones à risques ont été conservées dans la base d'entraînement et retirées dans la base de test afin de n'avoir aucune zone IRIS en commun entre ces deux bases.

Cependant, la base de test sert uniquement à étudier la performance de notre modèle à la fin du processus, on ne peut donc pas l'utiliser afin de choisir le modèle, les variables et les hyperparamètres optimaux. Pour réaliser ces choix, il est d'usage de comparer la performance de ces modèles :

- L'approche la plus simple est de tester la performance du modèle directement sur la table d'entraînement. Cependant, le risque est de provoquer du surapprentissage (*overfitting*) : le modèle pourrait trop bien s'adapter aux données d'entraînement en capturant des fluctuations aléatoires, rendant le modèle inefficace face à de nouvelles données. Le modèle perd alors sa capacité de généralisation.

- Une autre approche est donc de séparer notre jeu de données une deuxième fois. D'un côté une table d'entraînement et de l'autre une table de validation sur laquelle est effectuée la phase de prédiction permettant d'évaluer la performance des modèles. Toutefois, cette séparation et les scores obtenus dépendront énormément de la manière dont on a séparé les données et cela ne permet pas de profiter de l'ensemble du jeu de données.

- L'approche finale que nous avons utilisée est donc la validation croisée (*cross validation*). L'idée est de séparer nos données en n blocs, cinq par exemple. Sur les cinq, un est utilisé pour évaluer la performance, et les autres permettent au modèle de s'entraîner. On répète ce processus cinq fois en changeant à chaque fois le bloc permettant le test. On obtient cinq scores de performance dont on fait la moyenne pour obtenir un score de cross-validation. C'est ce score qui nous permettra de comparer les modèles.

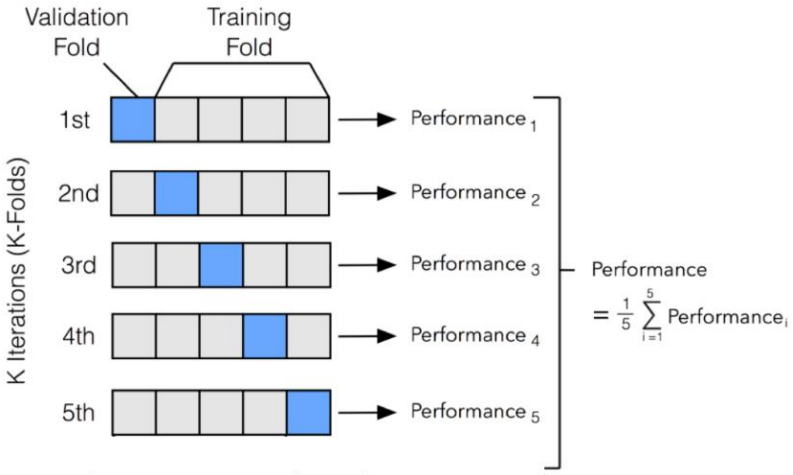

Figure 44 - Schéma de fonctionnement de la validation croisée (Source : Kaggle)

Pour notre étude, on a effectué la séparation en dix blocs. Celle-ci a été faite aléatoirement à partir des zones IRIS, ainsi à chaque itération on s'entraîne sur les sites assurés de 2250 zones IRIS, et l'on évalue les performances sur les 250 zones IRIS restantes. Cela permet de tester nos modèles sur des zones géographiques différentes de la phase d'apprentissage et évite ainsi tout risque de surapprentissage.

c. Scores de performances utilisés pour la régression

Afin d'effectuer cette validation croisée, on a besoin pour chaque étape d'un indicateur de performance permettant de comparer les modèles. Pour cela, on introduit trois indicateurs différents :

- La racine de l'erreur quadratique moyenne (ou RMSE pour *Root Mean Squared Error*) est la racine de la moyenne du carré des résidus (différence entre les valeurs observées et les valeurs estimées par le modèle). Cette mesure nous donne une indication sur la précision de notre modèle et permet d'évaluer la distance entre les données observées et prédites. À noter que cet indicateur dépend de l'échelle et n'est donc comparable que pour des prédictions issues d'un même ensemble de données. En notant y_i la valeur de la $i^{ème}$ observation du jeu de validation et $\widehat{y_i}$ la valeur prédite par le modèle pour la $i^{ème}$ observation, on obtient :

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

  Cet indicateur sera l'indicateur principal utilisé. Cependant, on basera également nos décisions selon les deux indicateurs suivants, répondants à d'autres problématiques métiers.

- S'il est important de fournir une bonne estimation pour chaque site assuré individuellement, le modèle sera également utilisé pour des analyses macro et on veut donc s'assurer que la charge globale prédite est cohérente avec les observations. Pour cela on calcule simplement le taux d'évolution entre la charge annuelle moyenne inondation prédite sur l'ensemble du jeu de validation et la charge annuelle inondation observée.

$$Score\ AAL = \frac{\left|\widehat{AAL_i} - AAL_i\right|}{AAL_i} \quad avec\ AAL_i = y_i * Engagement_i$$
$$\widehat{AAL_i} = \widehat{y_i} * Engagement_i$$

  On applique une valeur absolue, car une moyenne globale sera effectuée à l'issue de la validation croisée, on souhaite éviter que les effets d'une surestimation de charge soient contrebalancés avec les effets d'une sous-estimation.

- Enfin en dehors de l'aspect purement quantitatif on veut s'assurer que le modèle a une bonne capacité pour fournir un classement cohérent entre les risques, afin d'identifier ceux étant les plus à risques et ceux qui le sont moins. Pour cela, on calcule le coefficient de corrélation de Spearman entre les taux de destruction prédits et ceux observés. Ce coefficient permet de calculer la corrélation non pas entre les valeurs prises par les variables, mais entre leur rang. On introduira plus en détail cette mesure dans la partie suivante.

À noter qu'au sein de chaque modèle, d'autres indicateurs pourront être utilisés pour effectuer la sélection des variables. Dans le cas des modèles linéaires généralisés, on introduira des critères d'informations tel que l'AIC.

d. Scores de performances utilisés pour la classification

Le modèle de prédiction du taux de destruction pour le risque fluvial nécessitera une première étape de classification. En effet, 95% des taux de destruction sur la base d'entraînement présentent une charge nulle et ne sont donc pas exposés au risque fluvial qui concerne uniquement les crues de grands fleuves. On introduira donc une variable binaire « *Exposed_fluvial* » égale à 1 si le taux de destruction moyen est strictement positif et égal à 0 sinon.

Cette première étape de modélisation nécessitera des indicateurs spécifiques différents de la régression. On utilisera notamment une matrice de confusion qui permet d'identifier :

- Le nombre de vrais positifs (VP ou TP) : prédiction positive et valeur réelle positive
- Le nombre de faux positifs (FP) ou erreur de type I : prédiction positive et valeur réelle négative
- Le nombre de vrais négatifs (VN ou TN) : prédiction négative et valeur réelle négative
- Le nombre de faux négatifs (FN) ou erreur de type II : prédiction négative et valeur réelle positive

Dans notre situation, la classe positive correspondra aux risques exposés au risque fluvial (*Exposed_fluvial* = 1) et la classe négative à ceux qui ne le sont pas (*Exposed_fluvial* = 0)

		Classe réelle	
		-	+
Classe prédite	-	True Negatives *(vrais négatifs)*	False Negatives *(faux négatifs)*
	+	False Positives *(faux positifs)*	True Positives *(vrais positifs)*

Figure 45 - Matrice de confusion (source : Openclassrooms)

À partir de cette matrice, on peut en déduire différents indicateurs :

- $Spécificité = \frac{VN}{VN+FP}$: indiquera la part des risques non exposés au risque fluvial qui ont été correctement prédits

- $Sensibilité\ ou\ Rappel = \frac{VP}{FN+VP}$: indiquera la part des risques exposés au risque fluvial qui ont été correctement prédits

- $Valeur\ prédictive\ positive\ ou\ Précision = \frac{VP}{VP+FP}$: indiquera la part des risques prédits comme étant à risque fluvial qui le sont réellement

- $Valeur\ prédicitive\ négative = \frac{VN}{VN+FN}$: indiquera la part des risques prédits comme n'étant pas à risque fluvial qui le sont réellement

- $Accuracy = \frac{VP+VN}{VP+FP+VN+FN}$: indiquera la part de prédictions correctes sur le total des prédictions faites.

L'*accuracy*, parfois appelé exactitude en français, est une mesure de performance très régulièrement utilisée pour évaluer les modèles de classification vu qu'elle indique simplement le pourcentage de bonnes prédictions. Cependant, cet indicateur peut être trompeur dans le cas de classes déséquilibrées comme c'est le cas pour notre jeu de données. En effet sur la base de test par exemple, environ 83% de nos risques ne sont pas exposés au risque fluvial. Ainsi un modèle naïf qui ne prédirait que des classes négatives (non exposés au risque fluvial) aurait un score d'exactitude de 83%, ce qui pourrait nous amener à penser que le modèle est bon alors qu'il n'a en réalité aucun pouvoir prédictif.

C'est pourquoi on préfère souvent utiliser la précision et le rappel afin d'optimiser les modèles dans le cas de classes déséquilibrées. D'un côté, la précision permet de minimiser le taux d'erreur parmi les risques prédits positifs et de l'autre le rappel permet de maximiser la détection de cas positifs. Pour trouver un compromis entre ces deux mesures, on introduit une moyenne harmonique entre le rappel et la précision :

$$F_1 = \frac{2}{\frac{1}{rappel} + \frac{1}{précision}} = 2 * \frac{précision * rappel}{précision + rappel}$$

La précision et le rappel ayant le même numérateur et des dénominateurs différents, on préfère moyenner leur inverse afin de les considérer sur le même dénominateur, plutôt que d'utiliser une simple moyenne arithmétique qui aurait moins de sens. C'est donc ce score F_1 que l'on cherchera à maximiser par la suite pour le choix de notre modèle de classification.

e. Autres indicateurs pour la classification : AUC, courbe ROC, courbe PR…

Les modèles de classification peuvent être utilisés directement pour prédire la classe, mais permettent également d'accéder à la probabilité d'appartenir à chacune des classes. Ainsi, dans le cas d'une classification binaire considérons la probabilité d'appartenir à la classe « 1 », par défaut le modèle prédira « 1 » si la probabilité est supérieure au seuil de 0.5 et 0 lorsqu'elle est inférieure. Cependant, il est possible de faire varier ce seuil afin d'optimiser les indicateurs considérés.

L'idée de la courbe ROC (*Receiver Operating Characteristic*) est de faire varier le seuil de classification de 1 à 0 et de représenter le taux de vrais positifs (TPR), en fonction du taux de faux positifs (FPR) pour chacun des seuils considérés. Le taux de vrais positifs est également connu sous le nom de sensibilité et le taux de faux positifs peut être calculé comme : (1 - spécificité). Un classifieur qui coïnciderait avec la diagonale indiquerait un modèle qui n'apporte aucun pouvoir prédictif et qui est donc équivalent à un modèle aléatoire où l'on attribue la classe au hasard étant donné qu'on aurait un taux de vrais positifs égal au taux de faux positifs. À l'inverse, les classifieurs qui donnent des courbes plus proches du coin supérieur gauche indiquent une meilleure performance, donnant en effet le plus possible de vrais positifs avec le moins possible de faux positifs. Si la distance euclidienne au coin supérieur gauche peut être choisie pour choisir le seuil optimal, nous avons décidé de considérer uniquement la minimisation du score F1.

Il peut être utile de résumer la performance de chaque classifieur en une seule mesure. Une approche courante consiste à calculer l'aire sous la courbe ROC, abrégée en AUC (*Area Under Curve*). Elle est équivalente à la probabilité qu'une classe positive choisie au hasard soit classée plus haut qu'une classe négative choisie au hasard. Un modèle qui classe au hasard les individus aurait donc un AUC de 50% (ce qui correspond à l'aire sous la diagonale).

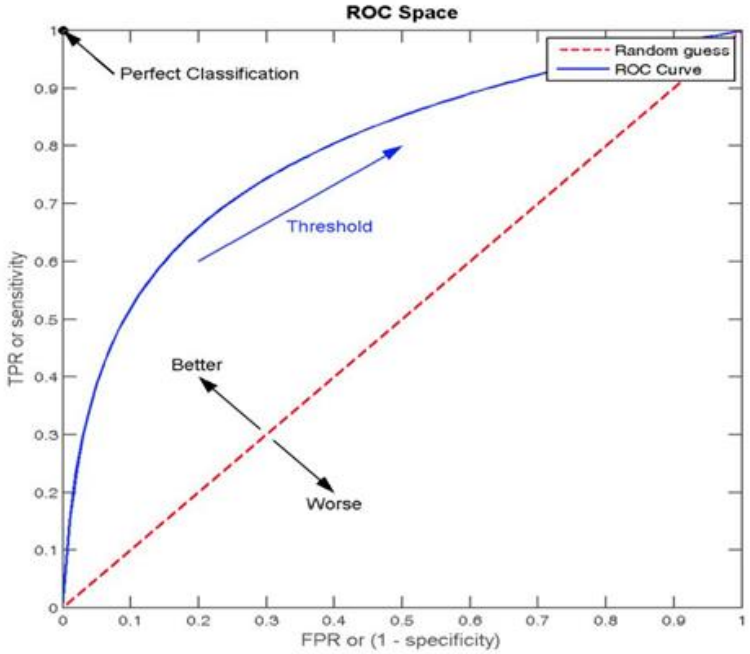

Figure 46 - Illustration d'une courbe ROC

Selon le même principe que la courbe ROC, on représentera également une courbe précision-rappel (ou courbe PR) qui indique la variation de la précision en fonction du rappel pour différents seuils de probabilité. Un modèle avec une classification parfaite est représenté par un point de coordonnées (1,1). Un classifieur obtenu aléatoirement serait une ligne horizontale sur le graphique avec une précision qui est proportionnelle au nombre d'exemples positifs dans l'ensemble de données, soit environ 17% sur notre ensemble de test.

2) Présélection des variables

Grâce à l'enrichissement effectué dans la partie III, on dispose d'une trentaine de variables liées au risque inondation et qui pourraient s'avérer utile pour la prédiction des taux de destruction. On rappelle que ces variables sont résumées en annexe 4. Afin de simplifier les calculs et d'éviter les problèmes de convergence dans nos modèles linéaires généralisés, on effectue une première sélection de variables à partir de diverses mesures nous permettant de sélectionner les variables les plus pertinentes et d'éliminer celles étant trop corrélées entre elles. Les méthodes utilisées dans cette partie sont communément appelées en *Machine Learning* des « *filter methods* », c'est-à-dire que l'on opère directement sur le jeu de données à l'aide d'une certaine mesure statistique que l'on doit spécifier. Il existe d'autres méthodes telles que les « *wrapper methods* », qu'on utilisera par la suite au sein de chaque modèle et pour lesquelles la sélection est guidée par le résultat du modèle (par exemple à partir des performances en validation croisée).

a. Coefficient de Pearson

Le coefficient de Pearson, aussi appelé coefficient de corrélation linéaire, évalue la relation linéaire entre deux variables. On calcule le coefficient de Pearson entre deux variables X et Y supposées suivre une loi normale :

$$r = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum(X - \bar{\bar{X}})(Y - \bar{Y})}{\sqrt{\sum(X - \bar{\bar{X}})^2 (Y - \bar{Y})^2}}$$

Le coefficient est compris entre -1 et 1. Plus le coefficient est proche de 1, plus la relation linéaire positive entre les variables est forte et plus il est proche de -1 et plus la relation linéaire négative est forte. Un indicateur proche de 0 indique des variables indépendantes. Ce coefficient est cependant très sensible aux valeurs extrêmes qui sont très présentes dans notre jeu de données. C'est pourquoi on attribuera plus d'importance aux deux corrélations de rangs introduites ci-dessous.

b. Coefficient de Spearman

Les coefficients de rangs peuvent s'avérer utiles pour capter des relations monotones qui ne seraient pas forcément affines. En effet, ces corrélations se basent sur le rang des variables X et Y et non directement sur leur valeur. Le coefficient rho de Spearman est introduit de manière similaire au r de Pearson :

$$\rho = \frac{cov(R_X, R_Y)}{\sigma_{R_X}\sigma_{R_Y}} = \frac{\sum_{i=1}^{n}(R(X_i) - \overline{R(X)})(R(Y_i) - \overline{R(Y)})}{\sqrt{\sum_{i=1}^{n}\left(R(X_i) - \overline{R(X)}\right)^2 \sum_{i=1}^{n}\left(R(Y_i) - \overline{R(Y)}\right)^2}} = 1 - \frac{6\sum_{i=1}^{n}\left(R(X_i) - R(Y_i)\right)^2}{n(n^2-1)}$$

avec n le nombre de pairs et $R(X_i)$ le rang de X_i allant de 1 *pour le maximum* à *n pour le minimum*

L'interprétation est similaire que pour le coefficient de Pearson, un rho égal à 1 désigne un classement identique entre les deux variables, un classement inverse lorsque celui-ci est égal à -1 et un classement indépendant lorsqu'il est proche de 0.

c. Coefficient de Kendall

On introduit finalement un dernier coefficient de corrélation qui agit également sur les rangs. Pour le calculer il est nécessaire de décompter le nombre de paires concordantes c'est-à-dire lorsque $x_i < x_j \; et \; y_i < y_j \left(ou \; x_i > x_j \; et \; y_i > y_j\right)$, ainsi que le nombre de paires discordantes c'est-à-dire lorsque $x_i < x_j \; et \; y_i > y_j \; (ou \; x_i > x_j \; et \; y_i < y_j)$. On obtient finalement le tau de Kendall déduit de la manière suivante :

$$\tau = \frac{n_c - n_d}{n_c + n_d} = \frac{n_c - n_d}{n(n-1)/2}$$

$$avec \; n_c = nombre \; de \; paires \; concordantes \; et \; n_d = nombre \; de \; paires \; discordantes$$

Ce coefficient est également compris entre –1 et 1 peut s'interpréter de la même manière que le coefficient de Spearman.

d. Visualisation des données

Dans un premier temps, avant de passer à l'analyse des corrélations, on souhaite avoir une première idée des variables potentiellement explicatives. Pour ce faire et pour davantage de clarté dans nos représentations on décide de transformer nos variables continues en variables binaires établies selon un certain seuil. Pour le risque fluvial, considérons la variable « *Exposed_fluvial* » déjà introduite dans la partie IV-1-d, égale à 1 lorsque le taux de destruction est strictement positif et à 0 sinon.

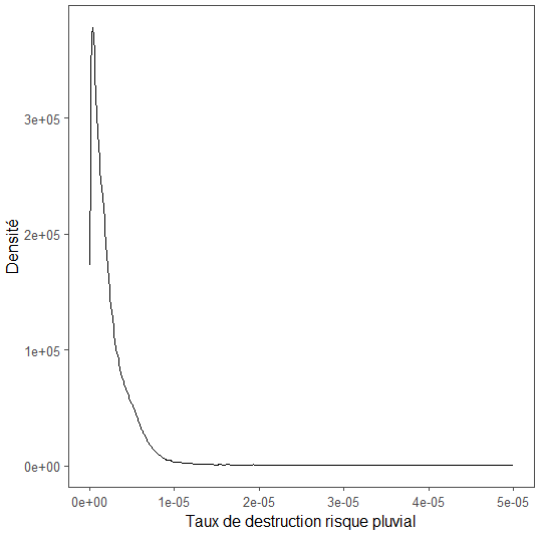

Figure 47 - Répartition du taux de destruction moyen pour le risque pluvial sur la base d'entraînement

Pour le risque pluvial, considérons la variable « *Exposed_pluvial* » que l'on calcule à partir du seuil de 0,00001, qui est donc égale à 1 lorsque le taux de destruction pluvial est supérieur à ce seuil et 0 sinon. 93% des sites assurés ont un taux de destruction inférieur à ce seuil, les 7% restants étant étalés sur l'intervalle]0,00001; 0,013] et qui correspond aux valeurs extrêmes de notre échantillon. On souhaite donc déterminer les variables permettant d'expliquer le caractère extrême de ces risques.

Pour la visualisation, on utilisera des boîtes à moustache dont l'interprétation est rappelée en annexe 5, et qui nous permet de comparer la différence de distribution des variables en fonction que les taux de destruction soient extrêmes (*Exposed_pluvial* = 1) ou non (*Exposed_pluvi*al = 0). Les résultats pour le risque fluvial sont renseignés en annexe 5.

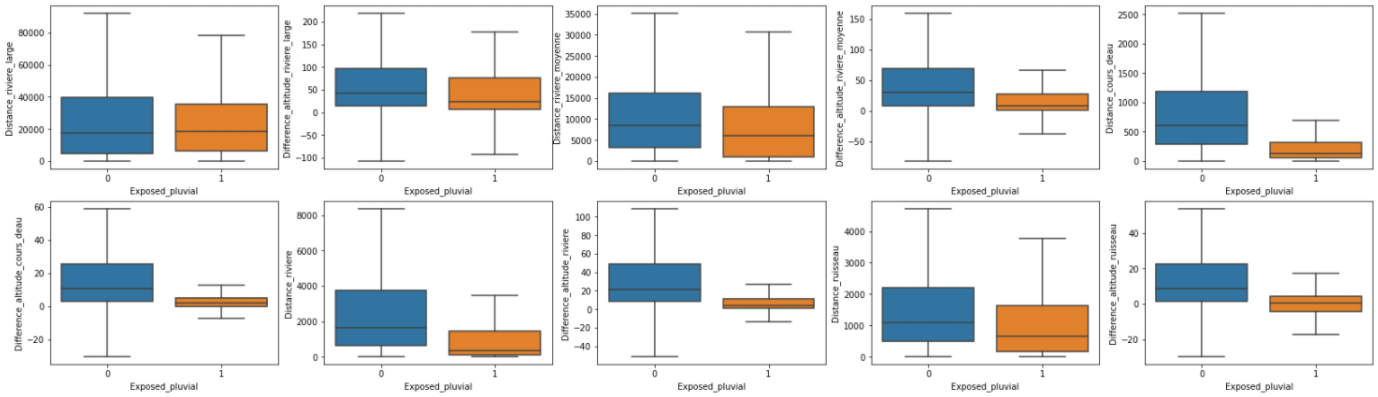

Figure 48 - Boîte à moustaches en fonction des valeurs de Exposed_pluvial pour les variables liées aux distances aux cours d'eau

Plusieurs variables ressortent clairement, notamment la distance au cours d'eau, la distance à la rivière la plus proche, la différence d'altitude au cours d'eau, la différence d'altitude à la rivière la plus proche ou encore la différence d'altitude au ruisseau. Logiquement, il apparait que les risques avec les taux de destruction les plus extrêmes (*Exposed_pluvial* = 1) ont des distributions beaucoup plus concentrées autour de 0. Ainsi plus un risque est proche d'un cours d'eau et plus son taux de destruction annuel moyen pour le risque pluvial est important.

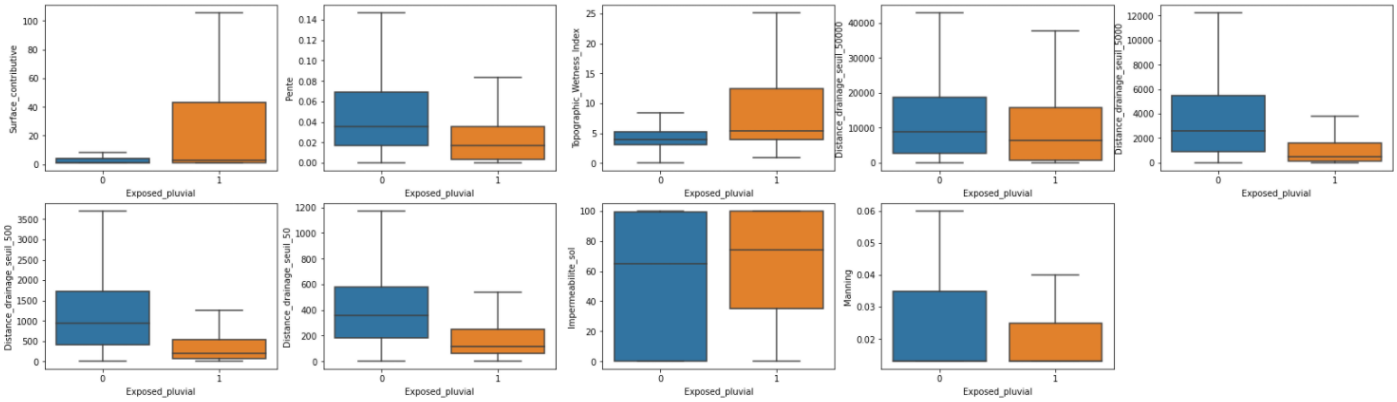

Figure 49 - Boîte à moustaches en fonction des valeurs de Exposed_pluvial pour les variables liées au modèle d'écoulement de l'eau D8 et à l'imperméabilité des sols

Les variables liées au modèle d'écoulement de l'eau D8 semblent particulièrement intéressantes. Il apparait que plus la surface contributive est importante et plus le taux de destruction l'est également, ce qui est cohérent étant donné qu'une forte surface contributive désigne des zones dans lesquelles l'eau est plus susceptible de s'accumuler. Également, on observe que plus le coefficient de pente est faible et plus le taux de destruction est important, l'eau n'étant que potentiellement passagère dans les zones avec de fortes pentes. L'indice d'humidité topographique, qui est fonction des deux précédentes variables est donc logiquement corrélé positivement avec le taux de destruction. On observe pour finir que plus le coefficient de Manning est faible (ce qui est équivalent à une forte imperméabilité du sol) et plus le taux de destruction est important.

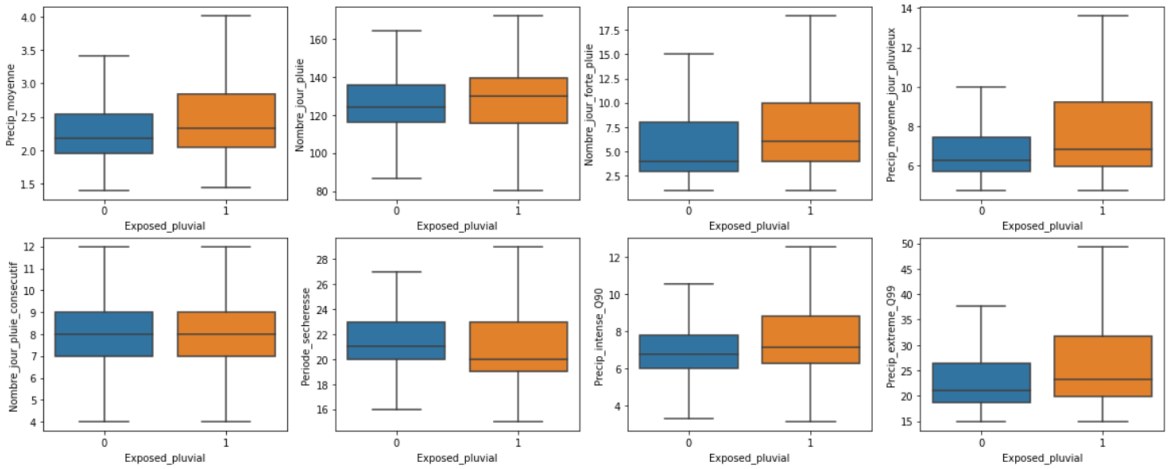

Figure 50 - Boîte à moustaches en fonction des valeurs de Exposed_pluvial pour les variables liées aux précipitations

Trois variables climatiques semblent particulièrement corrélées avec les taux de destruction du risque pluvial, à savoir le nombre de jours de forte pluie, la valeur des précipitations extrêmes (quantile à 99%) et les précipitations moyennes les jours pluvieux.

e. Résultats

Afin de formaliser ces premières observations, on s'intéresse donc aux trois mesures de corrélations introduites précédemment. Cette étape va également permettre d'éliminer les variables trop corrélées entre elles et qui risquent d'impacter la performance de nos modèles et d'ajouter une complexité inutile. Une fois les différentes mesures agrégées on obtient le tableau ci-dessous :

cols	Pearson	Rang Pearson	Spearman	Rang Spearman	Kendall	Rang Kendall	Rang Moyen
Distance_cours_deau	7.9 %	7.0	17.2 %	1.0	11.5 %	2.0	3.333333
Nombre_jour_forte_pluie	8.8 %	4.0	15.5 %	3.0	11.0 %	3.0	3.333333
Precip_extreme_Q99	8.0 %	6.0	15.3 %	5.0	10.5 %	5.0	5.333333
Difference_altitude_riviere	5.3 %	14.0	16.4 %	2.0	10.9 %	4.0	6.666667
Difference_altitude_cours_deau	5.7 %	12.0	14.7 %	7.0	9.7 %	7.0	8.666667
Difference_altitude_ruisseau	6.5 %	9.0	14.3 %	8.0	9.5 %	9.0	8.666667
Topographic_Wetness_Index	17.0 %	1.0	12.4 %	13.0	8.2 %	13.0	9.000000
Distance_riviere	5.8 %	11.0	14.1 %	9.0	9.5 %	8.0	9.333333
Distance_drainage_seuil_500	6.7 %	8.0	14.1 %	10.0	9.4 %	10.0	9.333333
Distance_drainage_seuil_50	10.0 %	2.0	11.2 %	14.0	7.4 %	14.0	10.000000
Precip_moyenne_jour_pluvieux	6.2 %	10.0	12.9 %	12.0	8.8 %	11.0	11.000000
Manning	0.1 %	30.0	15.3 %	4.0	11.7 %	1.0	11.666667
Distance_drainage_seuil_5000	1.8 %	23.0	14.7 %	6.0	9.9 %	6.0	11.666667
Difference_altitude_riviere_moyenne	3.3 %	18.0	13.3 %	11.0	8.8 %	12.0	13.666667
Pente	4.5 %	17.0	9.7 %	15.0	6.3 %	16.0	16.000000

Tableau 8 - Classement des 15 premières variables les plus significatives selon différentes mesures statistiques pour le taux de destruction moyen par an du risque pluvial

Comme observé précédemment la distance au cours d'eau ressort logiquement dans les variables les plus pertinentes. Le risque pluvial comprend notamment les crues de petites rivières et de ruisseaux, il est donc cohérent de voir ressortir cette variable pour laquelle les cours d'eau considérés sont en majorité des ruisseaux ou des petites rivières.

Concernant les variables climatiques, il s'agit du nombre de jours de forte pluie et des précipitations extrêmes qui ressortent. Ce résultat semble assez cohérent avec la première partie dans laquelle on a vu que les crues rapides et torrentielles, ainsi que les inondations par ruissellement étaient notamment provoquées par des pluies intenses.

Enfin, l'indice d'humidité topographique introduit dans la partie III apparait également dans les variables les plus intéressantes. Celui-ci est obtenu à partir du modèle d'écoulement de l'eau D8 et c'est l'indicateur le plus complet qui a été calculé à partir de cet algorithme. Il est en effet fonction du coefficient de pente et de la surface contributive, eux-mêmes obtenus à partir des directions de flux.

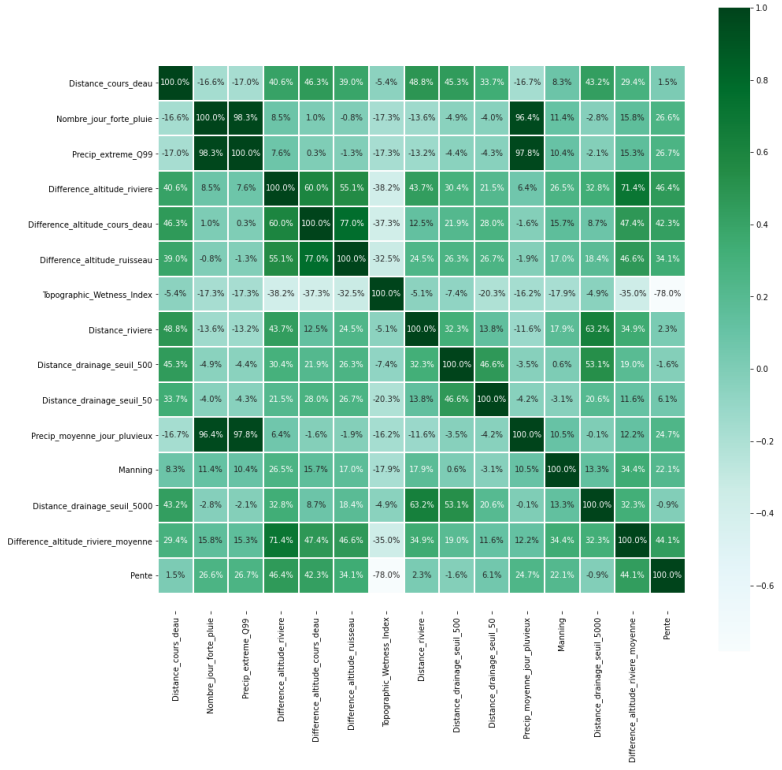

Figure 51 - Matrice de corrélation de Spearman du top 15 des variables pour le risque pluvial

La matrice ci-dessus nous donne les corrélations de Spearman entre les différentes variables. On décide de sélectionner les variables par ordre d'importance et de retirer celles avec une corrélation supérieure à 30% avec des variables déjà sélectionnées. On conserve finalement :

- La distance au cours d'eau le plus proche
- Le nombre de jours de forte pluie et le quantile à 99% des précipitations (précipitations extrêmes). Malgré la forte corrélation entre les deux, on testera chacun des modèles avec ces deux variables séparément afin de sélectionner celle qui fournira les meilleurs résultats.
- L'indice d'humidité topographique
- Le coefficient de Manning

De plus, en complément de ces variables, on ajoute d'autres variables dont on sait qu'elles ont été utilisées directement par le modélisateur, à savoir :

- La branche assurée
- Le taux engagement contenu
- Le taux engagement pertes d'exploitation
- Le nombre d'étages (disponible uniquement pour le portefeuille Immeuble, le reste étant renseigné à 0).

On répète ensuite le même processus pour le risque fluvial :

cols	Pearson	Rang Pearson	Spearman	Rang Spearman	Kendall	Rang Kendall	Rang Moyen
Difference_altitude_riviere_large	3.1 %	3.0	24.0 %	4.0	19.4 %	4.0	3.666667
Distance_riviere_large	1.5 %	11.0	26.6 %	1.0	21.5 %	1.0	4.333333
Distance_riviere	2.6 %	5.0	23.6 %	5.0	19.1 %	5.0	5.000000
Difference_altitude_riviere	1.8 %	8.0	22.9 %	6.0	18.5 %	6.0	6.666667
Difference_altitude_ruisseau	3.7 %	2.0	17.9 %	9.0	14.4 %	9.0	6.666667
Distance_drainage_seuil_50000	1.2 %	16.0	25.8 %	2.0	20.8 %	2.0	6.666667
Difference_altitude_riviere_moyenne	1.3 %	15.0	24.6 %	3.0	19.9 %	3.0	7.000000
Topographic_Wetness_Index	4.3 %	1.0	15.5 %	11.0	12.5 %	11.0	7.666667
Distance_cours_deau	2.0 %	7.0	15.7 %	10.0	12.7 %	10.0	9.000000
Pente	1.4 %	14.0	19.2 %	7.0	15.5 %	7.0	9.333333
Manning	2.9 %	4.0	11.1 %	15.0	10.3 %	14.0	11.000000
Distance_drainage_seuil_5000	0.5 %	19.0	18.5 %	8.0	14.9 %	8.0	11.666667
Impermeabilite_sol	1.6 %	9.0	12.6 %	13.0	10.7 %	13.0	11.666667
Difference_altitude_cours_deau	1.5 %	12.0	14.8 %	12.0	12.0 %	12.0	12.000000
Distance_drainage_seuil_500	1.5 %	10.0	11.5 %	14.0	9.3 %	15.0	13.000000

Tableau 9 - Classement des 15 premières variables les plus significatives selon différentes mesures statistiques pour le taux de destruction moyen par an du risque fluvial

La distance à la rivière large la plus proche et la différence d'altitude avec celle-ci ressortent comme les variables les plus corrélées avec notre variable cible ce qui semble cohérent étant donné que le risque fluvial concerne les débordements de grands fleuves. Notons cependant l'absence de variables concernant les précipitations dans ce classement. Les débordements de grands fleuves étant souvent provoqués par de longues pluies qui peuvent s'étendre sur plusieurs jours, il est étonnant de ne voir ressortir aucune variable liée, telle que le nombre de jours de pluie consécutif par exemple. Cette absence de corrélation significative avec des variables climatiques nous amène donc à limiter notre future étude du changement climatique au risque pluvial, sensible aux pluies extrêmes.

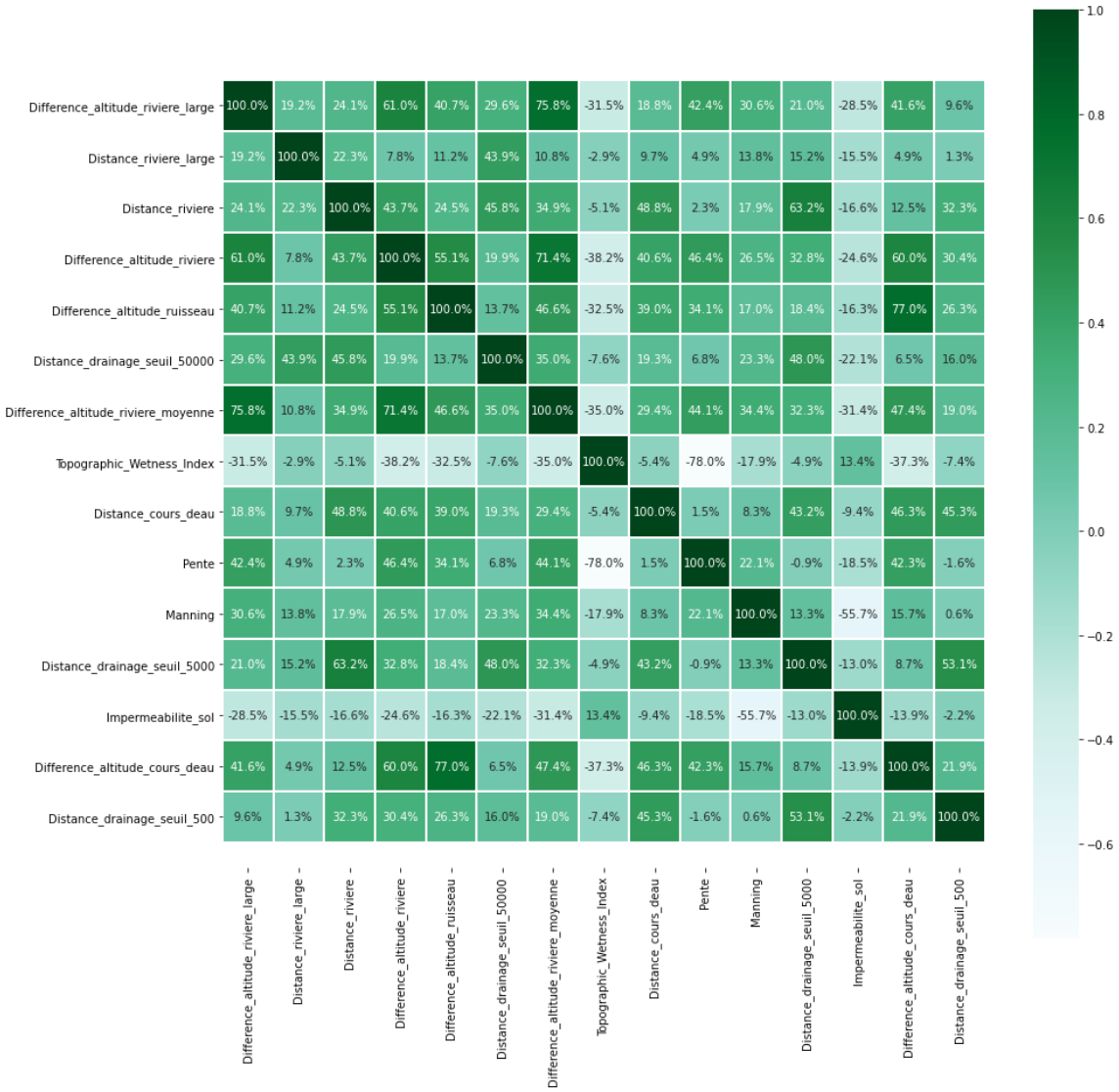

Figure 52 - Matrice de corrélation de Spearman du top 15 des variables pour le risque fluvial

Tout comme pour le risque pluvial, on sélectionne les variables par ordre d'importance en ne sélectionnant pas celles ayant une trop forte corrélation avec des variables déjà sélectionnées, d'où le choix final :

- La différence d'altitude à la rivière large la plus proche
- La distance à la rivière large la plus proche
- La distance à la rivière la plus proche
- L'imperméabilité du sol

De plus, en complément de ces variables, on ajoute d'autres variables dont on sait qu'elles ont été utilisées directement par le modélisateur, à savoir :

- La branche assurée
- Le taux engagement contenu
- Le taux engagement pertes d'exploitation
- Le nombre d'étages (disponible uniquement pour le portefeuille Immeuble, le reste étant renseigné à 0).

3) Les modèles linéaires généralisés (GLM)

L'objectif désormais est d'établir un lien entre les variables explicatives sélectionnées ci-dessus et la variable à expliquer dont on veut prédire les issues, en l'occurrence le taux de destruction moyen par an pour le risque pluvial et fluvial. Pour cela, on testera dans un premier temps les modèles linéaires généralisés.

a. Introduction

Les modèles linéaires généralisés (ou GLM pour *Generalized Linear Model*) sont devenus des références en assurance et notamment en tarification. Ils sont préférés aux modèles linéaires classiques qui demandent trop d'hypothèses difficilement vérifiées. Tout d'abord, les modèles linéaires supposent que la variable à expliquer est une fonction linéaire des variables explicatives, ce qui est rarement le cas en assurance. De plus, utiliser un modèle linéaire pour prédire des variables positives telles que des taux de destruction n'est pas adapté à notre distribution et on risquerait de prédire des taux négatifs, ce qui au passage ne vérifierait pas l'hypothèse selon laquelle les erreurs ϵ_i doivent être distribuées aléatoirement selon une loi normale. Enfin la variance σ^2 des erreurs doit être supposée constante (hypothèse d'homoscédasticité) ce qui est également rarement vérifié.

Pour remédier à ces limites, on introduit alors les GLM, définis selon l'équation suivante :

$$g(E[Y]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

Avec :

- $Y = (Y_i)_{i=1,\dots,n}$ un vecteur de $\mathbb{R}^n$ contenant les variables à expliquer sur nos n observations. Il s'agit de la composante aléatoire du modèle dont la densité est supposée appartenir à une loi de la famille exponentielle, qu'on introduira par la suite.

- Pour i allant de 1 à n : $(X_{1i}, \dots, X_{pi})$ un p-uplet contenant l'ensemble des p variables explicatives pour l'observation i.

- $\beta_0, \dots, \beta_p$ les paramètres de la régression estimés grâce à la méthode du maximum de vraisemblance. La combinaison linéaire des paramètres et des variables explicatives représente la composante déterministe du modèle.

- Une fonction de lien g strictement monotone et dérivable qui définit la relation entre la variable à expliquer et la composante déterministe.

b. La famille exponentielle

La loi de Y doit appartenir à la famille exponentielle, c'est-à-dire qu'il faut trouver $\theta \in \mathbb{R}$ (paramètre canonique), $\phi \in \mathbb{R}$ (paramètre de dispersion), a une fonction définie sur $\mathbb{R}^*$, b une fonction définie sur $\mathbb{R}$ et deux fois dérivables et c une fonction définie sur $\mathbb{R}^2$ tel que la densité de Y peut s'écrire sous la forme :

$$f(y|\theta, \phi) = \exp\left[\frac{y\theta - b(\theta)}{a(\theta)} + c(y, \phi)\right]$$

Dans ce cas, les moments de Y peuvent d'exprimer en fonction des différents paramètres :

$$E[Y] = \mu = b'(\theta)$$
$$Var[Y] = \sigma^2 = b''(\theta) * a(\phi)$$

Ainsi le paramètre canonique θ dépend de $E[Y] = \mu$ et donc des paramètres de régression β. On a en effet : $\theta(\mu(\beta)) = b'^{-1}(\mu(\beta)) = b'^{-1}\left(g^{-1}\left(\sum_{k=1}^{p}\beta_k x_k\right)\right)$.

De nombreuses lois usuelles appartiennent à la famille exponentielle, on peut notamment citer les lois de Poisson souvent utilisées pour modéliser les fréquences de sinistres, ainsi que les lois Gamma souvent utilisées pour modéliser le coût des sinistres.

Distribution de Y_i	θ_i	ϕ	$a_i(\phi)$	$b(\theta_i)$	$c(y_i, \phi)$
Normale$(\mu_i ;\sigma^2)$	μ_i	σ^2	ϕ	$\frac{\theta^2}{2}$	$-\frac{1}{2}\left\{\frac{y^2}{\sigma^2} + log(2\pi\sigma^2)\right\}$
Poisson(μ_i)	$\log(\mu_i)$	1	ϕ	$\exp(\theta_i)$	$-\log y!$
Binomiale$\frac{1}{m_i}(m_i ;\mu_i)$	$\log\left(\frac{\mu_i}{1-\mu_i}\right)$	$\frac{1}{\mu_i}$	ϕ	$\log(1+\exp\theta_i)$	$\log\binom{m_i}{m_i y_i}$
Gamma$(\mu_i ;\alpha)$	$\frac{-1}{\mu_i}$	α^{-1}	ϕ	$-\log(-\theta)$	$\alpha\log(\alpha y) - \log y - \log\Gamma(\alpha)$
Inverse Gaussienne$(\mu_i ;\sigma^2)$	$\frac{-1}{2\mu_i^2}$	σ^2	ϕ	$-(-2\theta)^{1/2}$	$-\frac{1}{2}\left\{\log(2\pi\phi y^3) + \frac{1}{\phi y}\right\}$

Tableau 10 - Exemples de familles exponentielles et leurs paramètres

c. Fonction de lien

Une fois que l'on a déterminé à quelle distribution de la famille exponentielle notre variable à expliquer correspond, il faut ensuite y associer une fonction de lien, qui doit être strictement monotone et dérivable. Les fonctions de lien les plus utilisées sont les suivantes :

- La fonction identité telle que $g(x) = x$, ce qui revient au modèle linéaire classique :

$$E[Y] = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

- La fonction logarithme népérien telle que $g(x) = \ln(x)$, ce qui donne un modèle multiplicatif, souvent apprécié en tarification :

$$E[Y] = \exp(\beta_0) * \exp(\beta_1 X_1) * \ldots * \exp(\beta_p X_p)$$

- La fonction logit telle que $g(x) = \ln\left(\frac{x}{1-x}\right)$, ce qui donne :

$$E[Y] = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}$$

- La fonction inverse telle que $g(x) = \frac{1}{x}$, ce qui donne :

$$E[Y] = \frac{1}{\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$$

d. Estimation des paramètres de la régression

Les paramètres à estimer sont les paramètres β et ϕ. ϕ étant un paramètre de nuisance qui n'influence pas la maximisation de la vraisemblance en β on ne s'attardera pas sur ce point, mais il peut être estimé par maximum de vraisemblance tout comme les paramètres β.

Pour rappel, étant donné une certaine loi de probabilité Y et des observations $(y_1, y_2, \dots, y_n)$, alors la vraisemblance quantifie la probabilité que les observations proviennent d'un échantillon de la loi de Y. Ainsi plus la vraisemblance est proche de 0 et moins l'adéquation est bonne, c'est pourquoi on cherche à maximiser cette vraisemblance, afin d'en déduire les paramètres optimaux. On obtient simplement la probabilité d'observer l'échantillon avec le produit d'observer chacune des réalisations, d'où la vraisemblance notée L et définie par :

$$L(\beta) = \prod_{i=1}^{n} f(y_i|\theta_i(\mu_i(\beta)), \phi)$$

En pratique, il est souvent plus simple de passer par le logarithme népérien afin de transformer le produit en somme :

$$l(\beta) = \ln\left(\prod_{i=1}^{n} f(y_i|\theta_i, \phi)\right) = \sum_{i=1}^{n} \ln\big(f(y_i|\theta_i, \phi)\big)$$

Les paramètres optimaux, qu'on appelle estimateur du maximum de vraisemblance et qu'on note $\hat{\beta}$ sont les paramètres qui permettent de maximiser la fonction de log vraisemblance, d'où :

$$\hat{\beta} = \underset{\beta}{\operatorname{argmax}}\, l(y, \beta)$$

En pratique, on résout ce problème par des méthodes itératives tel que l'algorithme de Newton-Raphson.

Une fois les paramètres estimés on peut finalement passer à l'étape de prédiction de l'espérance pour l'ensemble de nos observations i :

$$\widehat{\mu_i} = \mu_i(\hat{\beta}) = g^{-1}\left(\sum_{k=1}^{p} \widehat{\beta}_k x_{k,i}\right)$$

e. Sélection de variables et critères d'information

Il existe plusieurs indicateurs permettant de mesurer la qualité de nos modèles sur la base des différences entre observations et estimations.

Tout d'abord, le calcul de la déviance permet de comparer notre modèle avec le modèle dit saturé. Le modèle saturé est un modèle avec la même distribution et la même fonction de lien, mais qui possède autant de paramètres que d'observations et qui estime donc exactement les données, d'où $\widehat{\mu_i^S} = y_i$. On définit la déviance comme deux fois la différence entre la log-vraisemblance maximisée du modèle saturé et celle du modèle sélectionné, soit :

$$D = 2\left(L(\hat{\beta}_S) - L(\hat{\beta})\right)$$

Avec $\hat{\beta}_S$ l'estimateur du maximum de vraisemblance du modèle saturé, d'où $L(\hat{\beta}_S)$ la log-vraisemblance maximisée :

$$L(\hat{\beta}_S) = \prod_{i=1}^{n} f(y_i|\theta_i(\mu(\hat{\beta}_S)), \phi) = \prod_{i=1}^{n} f(y_i|\theta_i(\widehat{\mu_i^S}), \phi) = \prod_{i=1}^{n} f(y_i|\theta_i(y_i), \phi)$$

$L(\hat{\beta})$ étant donc la log-vraisemblance maximisée obtenue sur le modèle sélectionné :

$$L(\hat{\beta}) = \prod_{i=1}^{n} f(y_i|\theta(\mu(\hat{\beta})), \phi) = \prod_{i=1}^{n} f(y_i|\theta(\widehat{\mu_i}), \phi)$$

Plus la déviance est faible et plus les deux log-vraisemblances sont proches, traduisant ainsi un ajustement de bonne qualité. Asymptotiquement, la déviance suit une loi du Khi-2 à n-p-1 degré de liberté, son espérance est donc n-p-1. On peut également comparer cette déviance avec la déviance du modèle nul, qui est le modèle composé uniquement de la constante sans les variables explicatives. Définissons alors le pseudo R^2 ou R^2 de McFadden comme :

$$R^2_{McFadden} = 1 - \frac{D}{D_0}$$

Tout comme le R^2 dans le cas linéaire, cet indicateur est borné entre 0 et 1 et sert à évaluer la qualité de l'ajustement du modèle, il représente la part de la déviance expliquée par le modèle. Plus l'indicateur est proche de 1 et plus l'amélioration du modèle apportée par les variables explicatives est importante.

Il existe d'autres indicateurs permettant de mesurer la qualité du modèle. On peut définir notamment les critères d'informations qui permettent de trouver un équilibre entre qualité du modèle et complexité en pénalisant l'ajout de paramètres, évitant ainsi l'ajout de variables peu significatives. Il existe deux principaux critères, l'AIC (Akaïke information criterion) et le BIC (Bayesian information criterion) définis tels que :

$$AIC = -2 * L + 2 * p$$
$$BIC = -2 * L + \ln(n) * p$$

Avec $L = L(\hat{\beta})$ la log-vraisemblance maximisée du modèle, p le nombre de paramètres du modèle et n le nombre d'observations. L'objectif est donc de minimiser ces deux critères, pour ce faire il peut être intéressant de s'intéresser à la combinaison de variables qui permet d'obtenir un AIC ou BIC minimal. Il existe pour cela deux principales méthodes :

- La méthode ascendante (*forward selection*) : on démarre l'algorithme avec une unique variable en ajoutant à chaque étape la variable permettant d'obtenir le plus faible AIC ou BIC. L'algorithme s'arrête lorsqu'il n'y a plus de variables à rajouter ou bien si l'ajout d'une nouvelle variable fait augmenter l'AIC ou BIC.

- La méthode descendante (backward selection) : on démarre l'algorithme avec l'ensemble des variables et on retire à chaque étape celle qui permet d'obtenir le plus faible AIC ou BIC. L'algorithme s'arrête lorsqu'il n'y a plus de variables à enlever ou bien si le retrait d'une nouvelle variable fait augmenter l'AIC ou BIC.

Dans notre étude, on se focalisera principalement sur une méthode descendante en utilisant l'AIC.

f. Le cas de la régression Bêta

On s'intéresse dans cette partie à la régression Bêta qui repose sur la loi du même nom et dont la densité est définie sur [0,1]. Elle est régulièrement utilisée pour la modélisation de taux ou de proportions définies sur ce même intervalle. On décide de dédier une sous-partie à ce cas dans la mesure où la théorie des modèles linéaires généralisés ne peut pas être directement appliquée avec la loi bêta qui ne fait pas partie de la famille exponentielle naturelle (*Natural exponential family :* NEF). La fonction de densité de la loi $Beta(\alpha, \beta)$ est donnée par :

$$f(y;\alpha,q) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(q\beta)} y^{\alpha-1}(1-y)^{\beta-1} \qquad \forall y \in]0;1[$$

$$Avec\ p,\beta > 0\ et\ \Gamma(.)\ la\ fonction\ Gamma$$

Ferrari and Cribari-Neto (2004) ont proposés une paramétrisation différente en posant $\mu = \frac{\alpha}{\alpha+\beta}$ et $\phi = \alpha + \beta$:

$$f(y;\mu,\phi) = \frac{\Gamma(\phi)}{\Gamma(\mu\phi)\Gamma\big((1-\mu)\phi\big)} y^{\mu\phi^{-1}}(1-y)^{(1-\mu)\phi^{-1}} \qquad \forall y \in]0;1[$$

$$Avec\ 0 < \mu < 1\ et\ \phi > 0$$

On peut montrer que : $E[Y] = \mu\ \ et\ \ Var[Y] = \frac{\mu(1-\mu)}{1+\phi}$

On appelle ϕ le paramètre de précision, plus il augmente et plus la variance diminue et ϕ^{-1} le paramètre de dispersion.

Soit $y_1, \dots, y_n$ un échantillon tel que $y_i \sim B(\mu_i, \phi)$. Le modèle de régression Bêta est défini comme :

$$g(\mu_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,k} = x_i^T \beta = \eta_i$$

Avec $g :]0,1[\to \mathbb{R}$ une fonction de lien strictement monotone, deux fois dérivable et surjective.

On peut montrer que la fonction de log-vraisemblance s'écrit sous la forme $l(\beta,\phi) = \sum_{i=1}^{n} l_i(\mu_i,\phi)$ avec :

$$l_i(\mu_i,\phi) = \log\big(\Gamma(\phi)\big) - \log\big(\Gamma(\mu_i\phi)\big) - \log(\Gamma(1-\mu_i)\phi) + (\mu_i\phi - 1)\log(y_i) + \big((1-\mu_i)\phi - 1\big)\log(1-y_i)$$

Tout comme avec les GLM on estime les paramètres du modèle par maximum de vraisemblance. Une extension de ce modèle a été introduite par Simas et al. (2010) afin de modéliser le paramètre de précision ϕ qu'on ne considère plus comme constant pour toutes les observations. Le modèle est désormais construit autour de deux équations :

$$g(\mu_i) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,k} = x_i^T \beta = \eta_i$$
$$g(\phi_i) = \gamma_0 + \gamma_1 z_{i,1} + \gamma_2 z_{i,2} + \cdots + \gamma_p z_{i,h} = z_i^T \gamma = \eta_i$$

$$Avec\ k + h < n$$

L'estimation des coefficients se fait toujours par maximum de vraisemblance à partir de la fonction $l(\beta,\phi) = \sum_{i=1}^{n} l_i(\mu_i,\phi_i)$.

4) Notions et modèles de machine learning

a. Arbre de décision (CART)

Les arbres de régression et de classification (*Classification and regression trees* ou CART) sont des méthodes d'apprentissage statistique utilisées comme leur nom l'indique à la fois pour des problématiques de régression ou de classification. On peut donc les utiliser à la fois pour la prédiction de variables numériques (régression) ou qualitatives (classification). On n'utilisera pas directement ces modèles, mais ils servent de base à d'autres modèles de *machine learning* tels que les forêts aléatoires ou le *gradient boosting*.

La figure ci-dessous présente un exemple d'arbre de décision. L'ensemble des observations sur la base d'entraînement est tout d'abord réuni à la racine de l'arbre puis chaque division sépare chaque nœud en deux nœuds fils selon une certaine condition binaire de la forme « $X_i \leq a$? » lorsque X_i est une variable ordinale ou « $X_i = modalité$? » lorsque X_i est une variable nominale (pas de relations d'ordre entre les modalités). Il faut ainsi déterminer à chaque étape, la variable X_i et le seuil « a » (ou la modalité) permettant de séparer les données en deux groupes les plus homogènes possibles au sens d'un critère à préciser. Une fois le partitionnement terminé (selon une certaine règle à préciser également), chaque nœud terminal de l'arbre est appelé feuille et représente une sous-division de nos observations qui respecte les différentes conditions déterminées à chaque nœud qui précède.

Une fois l'arbre maximal construit la dernière étape est l'élagage de cet arbre, qui consiste à chercher le meilleur sous arbre dans le but de réduire la complexité de l'algorithme et d'éviter le surapprentissage. Pour l'étape de prédiction de la variable à expliquer, une certaine valeur est déterminée à chaque feuille, soit la moyenne des observations associées à cette feuille dans le cas d'une régression, soit la modalité la plus représentée parmi les observations de cette feuille dans le cas d'une classification. C'est donc cette valeur qui sera prédite en sortie du modèle.

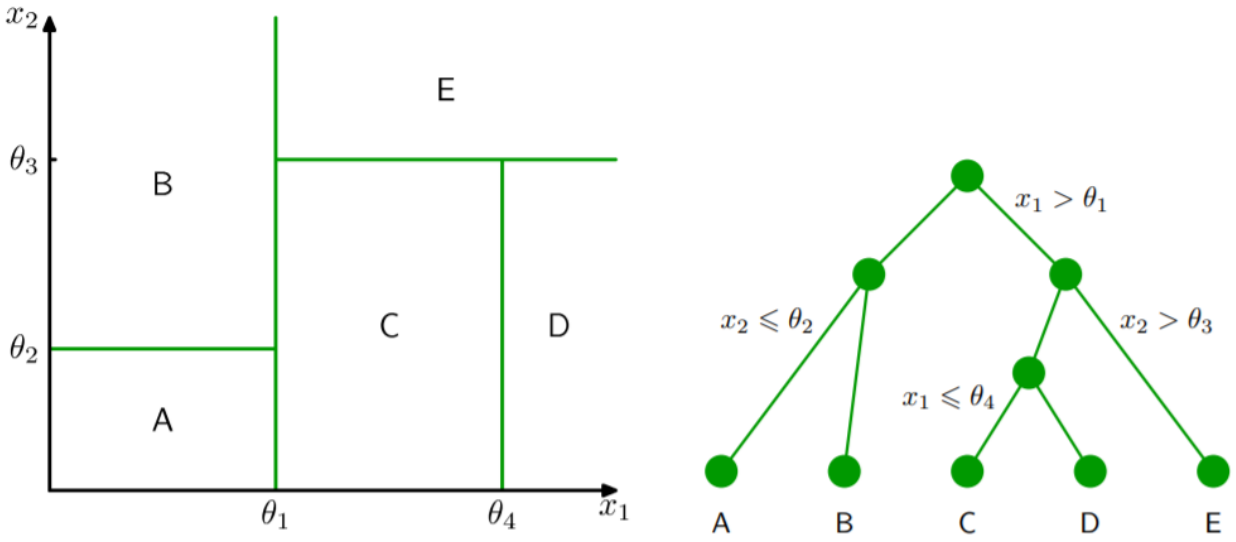

Figure 53 – Schéma d'un arbre de décision avec deux variables explicatives

Le processus décrit ci-dessus nécessite donc trois précisions :

- Un critère permettant d'identifier la division optimale à chaque nœud
- Une règle permettant de déterminer la fin du partitionnement
- Une méthode permettant d'effectuer l'élagage de l'arbre

Considérons que l'on souhaite étudier la variable à expliquer Y sur n individus, qui est soit quantitative réelle, soit qualitative à m modalités (modalités que l'on notera $\tau_l \; avec \; l = 1, \dots, m$). De plus, on dispose d'observations sur p variables explicatives X_j (quantitative ou qualitative).

On souhaite dans un premier temps déterminer un critère d'homogénéité permettant de séparer les observations en deux groupes les plus homogènes possibles au sens de la variable à expliquer. Pour cela on introduit une fonction d'hétérogénéité D qui doit vérifier deux conditions. Tout d'abord, elle doit être nulle lorsque le nœud est homogène. C'est-à-dire lorsque tous les individus appartiennent à la même modalité dans le cas d'une classification, ou s'ils prennent tous la même valeur de Y dans le cas d'une régression. De plus, cette fonction se doit d'être maximale lorsque les valeurs de Y sont très dispersés ou équiprobables.

Soit N un nœud et notons N_D et N_G les sous-divisions droites et gauches obtenues. Considérons toutes les divisions possibles du nœud N parmi l'ensemble des couples $(X_j, seuil)$, le couple optimal sélectionné est celui qui permet de rendre minimal la somme des hétérogénéités des nœuds fils, soit $D_{N_G} + D_{N_D}$. Ce qui revient également à maximiser à chaque étape de la construction de l'arbre la grandeur : $D_N - (D_{N_G} + D_{N_D})$.

Dans le cas d'une variable cible Y quantitative on définit la fonction d'hétérogénéité comme égal à la variance, soit :

$$D_N = \frac{1}{|N|} \sum_{i \in N} (y_i - \bar{y}_N)^2$$

$$avec \; |N| \; l'effectif \; du \; noeud \; N$$

Le problème revient donc à minimiser la variance intragroupe ou encore :

$$\frac{|N_G|}{|N|} \sum_{i \in N_G} (y_i - \bar{y}_{N_G})^2 + \frac{|N_D|}{|N|} \sum_{i \in N_D} (y_i - \bar{y}_{N_D})^2$$

Lorsque la variable cible est qualitative, la fonction d'hétérogénéité est soit définie comme l'entropie, soit comme la concentration de Gini. On utilisera dans notre cas la concentration de Gini, définie comme :

$$D_N = \sum_{l=1}^{m} p_N^l \, (1 - p_N^l)$$

$$o\grave{u} \; p_N^l \; est \; la \; proportion \; de \; la \; modalit\acute{e} \; \tau_l \; de \; Y \; dans \; le \; noeud \; K$$

On effectue donc cette minimisation de la somme des hétérogénéités étape par étape. On définit ensuite un critère d'arrêt souvent défini simplement lorsque le nœud est homogène, ou autrement dit lorsque la fonction d'hétérogénéité est nulle. Il est également possible de définir un seuil d'arrêt lorsque le nombre d'observations que contient la feuille est trop faible.

La procédure précédente nous a permis d'obtenir un arbre maximal $A_{\max}$. On cherche à déterminer un arbre optimal entre l'arbre trivial possédant une seule feuille et l'arbre maximal qui risque de nous conduire à une situation de surapprentissage. On cherche donc un compromis entre l'adéquation aux données et la complexité de l'arbre. Afin de réduire le nombre d'arbres à considérer, on passe par un algorithme itératif nous permettant dans un premier temps de considérer une suite emboîtée de sous-arbres.

Soit A un arbre et notons F_A son nombre de feuilles, on définit sa qualité d'ajustement comme la somme des fonctions d'hétérogénéité sur chaque feuille de l'arbre, soit :

$$D(A) = \sum_{f=1}^{F_A} D_f$$

On déduit de cette grandeur la complexité de l'arbre en ajoutant une pénalisation pour chaque feuille d'arbre supplémentaire :

$$C(A) = D(A) + \gamma * F_A$$

On effectue l'algorithme itératif à partir de cette formule :

- On pose tout d'abord $\gamma = 0$. L'arbre permettant de minimiser $C(A)$ est donc $A_{\max}$
- Par la suite, on augmente γ jusqu'à ce que l'arbre permettant de minimiser $C(A)$ ne soit plus l'arbre maximal, mais un arbre A_{F_A-1} ne possédant plus que $F_A - 1$ feuilles
- On recommence l'opération jusqu'à ce que l'arbre permettant de minimiser $C(A)$ soit l'arbre trivial A_1

On obtient donc une suite emboîtée d'arbres :

$$A_{F_A} \supset A_{F_A} - 1 \supset \cdots \supset A_1$$

Finalement, on choisit l'arbre A_{opt} parmi ces sous-arbres par un processus de validation croisée.

b. Forêt aléatoire

Une forêt aléatoire ou *random forest* est un modèle d'apprentissage automatique qui repose en partie sur une méthode de *bagging* appliqué à des arbres de décisions. Le *bagging* introduit par Breiman en 1996 est la contraction de *boostrap aggregating* que l'on peut traduire par « rééchantillonnage et agrégation ». L'idée générale est de tirer indépendamment k observations aléatoires avec remise parmi la base de n observations ($k \leq n$), c'est la composante « *bootstrap* » du bagging. Certaines observations peuvent donc être répétées. On réitère ce processus m fois et on obtient alors m échantillons que l'on va utiliser pour entraîner m modèles, tels que des arbres de décisions. On obtient ainsi m prédictions différentes que l'on va agréger soit par une moyenne dans le cas de la régression, soit par un vote majoritaire dans le cas de la classification (la modalité la plus prédite parmi les m classifieurs est sélectionnée).

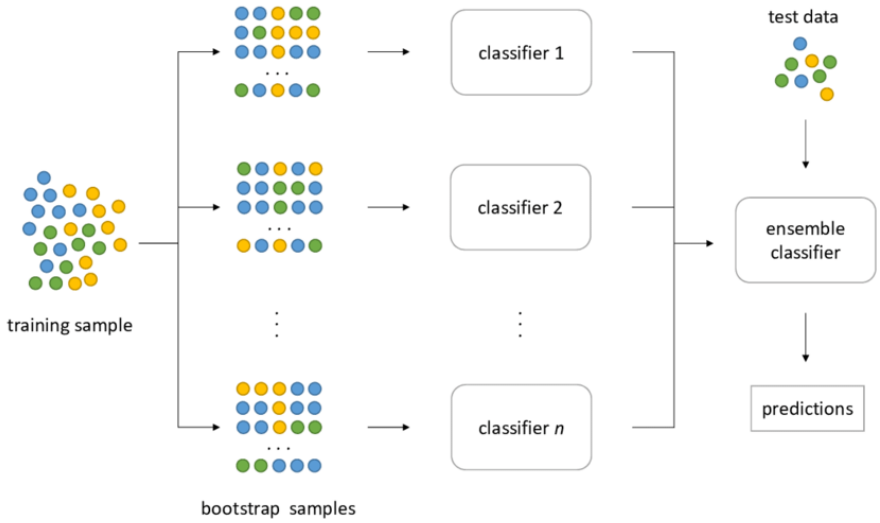

Figure 54 – Schéma explicatif du bagging

Le modèle de forêt aléatoire repose en grande partie sur le *bagging* (appliqué aux arbres de décisions), mais possède cependant une composante supplémentaire dans la construction des arbres. Ce modèle introduit en effet un tirage aléatoire des variables. Ainsi, à chaque nœud des arbres de décisions, plutôt que de trouver la division optimale en agissant sur l'ensemble des variables, on considère uniquement m variables tirées aléatoirement et sans remise parmi l'ensemble des variables explicatives. Par défaut, dans notre cas, la valeur de m sera égale à la racine carrée du nombre de variables explicatives total, mais il sera également possible d'optimiser ce paramètre. À noter également qu'à la différence d'un simple algorithme CART, les arbres ne sont pas élagués dans le cas d'une forêt aléatoire.

L'algorithme de forêt aléatoire met également à disposition une estimation de son erreur de généralisation qu'on appelle l'erreur « *Out-Of-Bag »* (OOB) ou erreur « hors sac » pour désigner les observations qui seraient en dehors de l'échantillon *bootstrap*. L'idée est assez simple, on considère une observation (x_i, y_i) et on s'intéresse uniquement aux arbres pour lesquels cette observation est « *Out-Of-Bag* », c'est-à-dire qu'elle n'a pas été tirée dans l'échantillon *bootstrap.* On calcule par la suite les prédictions associées à cette observation pour chacun de ces arbres et on agrège les résultats de manière classique (moyenne ou vote majoritaire). On effectue cette opération pour chacune des observations et on calcule finalement l'erreur associée en faisant l'erreur quadratique moyenne dans le cas d'une régression ou en retournant la proportion d'observations mal classées en classification.

L'erreur OOB permet également de fournir un indicateur d'importance entre les variables. Concrètement en considérant la variable m_i, on calcule tout d'abord l'erreur OOB de l'arbre de manière classique, puis on la recalcule en permutant aléatoirement les observations de la variable m_i. Logiquement, si la variable est importante l'erreur après permutation de ces observations devrait être beaucoup plus grande. En faisant la moyenne des écarts sur tous les arbres, si l'écart moyen est faible cela signifie que la variable a peu d'importance.

Ce modèle présente de nombreux avantages par rapport à l'algorithme CART, en étant beaucoup moins instable, moins sensible au surapprentissage et présente généralement de meilleures performances. Cependant contrairement à un simple arbre de décision, l'agrégation d'un ensemble d'arbres rend l'interprétabilité du modèle beaucoup plus compliqué, d'où un certain effet boîte noire, qui rend ce genre de modèles souvent rarement utilisés en tarification.

c. Gradient Boosting Machines (GBM) et eXtreme Gradient Boosting (XGBoost)

Tout comme les forêts aléatoires avec le *bagging*, le renforcement de gradient ou *Gradient Boosting Machines* est un modèle d'agrégation d'arbres basé sur le *boosting*. Cette méthode adopte un principe similaire au *bagging*, l'objectif étant de construire une famille de modèles qui sont ensuite agrégés. La principale différence provient notamment de la manière dont est construite la famille de modèles. En effet, si le *bagging* repose sur une construction parallèle de chaque modèle, le *boosting* est construit de manière récurrente.

L'idée de base est de s'adapter au modèle précédent en donnant plus de poids aux observations mal ajustées ou mal prédites. L'algorithme de boosting basique est l'*AdaBoost.* Dans un premier temps, les poids de chaque observation sont initialisés à $\frac{1}{n}$ (où n est le nombre d'observations). Par la suite, ce poids reste inchangé si l'observation est bien classée et croît proportionnellement au défaut d'ajustement dans le cas inverse. Enfin, on agrège les résultats de l'ensemble des modèles en associant un poids c_m à chaque classifieur, calculé à partir du taux d'erreur du modèle et permettant d'attribuer davantage d'importance aux modèles les plus performants.

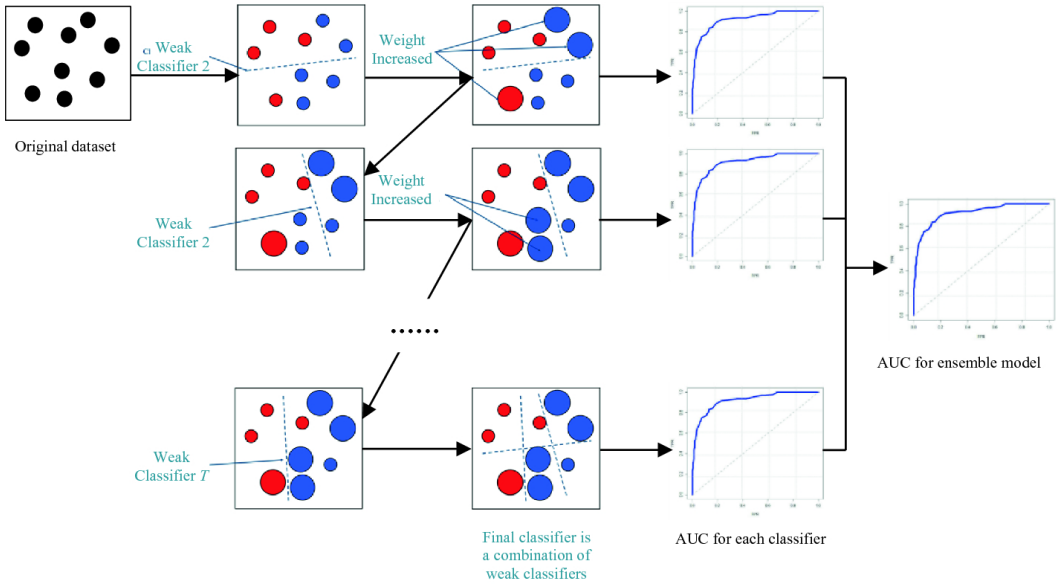

Figure 55 - Schéma explicatif du boosting

De la même façon que pour *AdaBoost*, l'algorithme de *gradient boosting* vise à construire une séquence de modèles de sorte que chaque modèle ajouté à la combinaison apparaisse comme un pas vers une meilleure solution. À la différence que dans le cas du renforcement de gradient, ce pas est franchi dans la direction du gradient de la fonction de perte, ce qui permet d'améliorer les propriétés de convergence. Notons l une fonction de perte convexe et différentiable dont l'objectif est par définition d'évaluer l'écart entre les valeurs prédites et les observations. Ce modèle est donc basé sur un algorithme de descente de gradient dont l'objectif est pour rappel de déterminer un minimum local en partant d'un point aléatoire et en se déplaçant vers la plus forte pente.

Le modèle se présente comme une approximation de la fonction f :

$$\hat{f}_m(x) = \hat{f}_{m-1}(x) + \gamma_m \delta(x_i)$$

$$= \hat{f}_{m-1}(x) - \gamma_m \sum_{i=1}^{n} \nabla_{f_{m-1}} l\left(y_i, f_{m-1}(x_i)\right)$$

À chaque étape on cherche donc à calculer le meilleur pas de descente γ tel que :

$$\min_{\gamma} \sum_{i=1}^{n} \left[l\left(y_i, f_{m-1}(x_i) - \gamma \frac{\partial l\left(y_i, f_{m-1}(x_i)\right)}{\partial f_{m-1}(x_i)} \right) \right]$$

On peut résumer l'algorithme global pour la régression de la manière suivante :

Algorithm 5 *Gradient Tree Boosting* pour la régression

Soit x_0 à prévoir
Initialiser $\widehat{f_0} = \arg\min_{\gamma} \sum_{i=1}^{n} l(y_i, \gamma)$
for $m = 1$ à M **do**
  Calculer $r_{mi} = -\left[\frac{\partial l(y_i, f(\boldsymbol{x}_i))}{\partial f(\boldsymbol{x}_i)}\right]_{f=f_{m-1}}$; $i = 1, \ldots, m$
  Ajuster un arbre de régression δ_m aux couples $(\boldsymbol{x}_i, r_{mi})_{i=1,\ldots,n}$
  Calculer γ_m en résolvant : $min_{\gamma} \sum_{i=1}^{n} l(y_i, f_{m-1}(\boldsymbol{x}_i) + \gamma \delta_m(\boldsymbol{x}_i))$.
  Mise à jour : $\widehat{f_m}(\boldsymbol{x}) = \widehat{f}_{m-1}(\boldsymbol{x}) + \gamma_m \delta_m(\boldsymbol{x})$
end for
Résultat : $\widehat{f_M}(\boldsymbol{x}_0)$.

Figure 56 - Algorithme de Gradient Boosting pour la régression (Source : WikiStat)

Lors de notre étude, on n'utilisera pas directement le *Gradient Boosting,* mais plutôt le eXtreme Gradient Boosting souvent abrégé comme XGBoost et qui est une version particulière du modèle décrit ci-dessus. Une des différences réside dans la construction des arbres de décisions qui sont cette fois-ci élagués contrairement à ce qui est fait dans le renforcement de gradient. De plus, de nombreux calculs sont parallélisés permettant de réduire significativement les temps de calcul.

d. Sélection de variables : Élimination récursive de variables *(RFE)*

L'algorithme d'élimination récursive de variables (*Recursive Feature Elimination* ou RFE) est un algorithme permettant de sélectionner le sous-ensemble de variables optimales permettant de maximiser le score de validation croisée sur l'indicateur de performance considéré. Dans le cas de la régression, on utilisera le RMSE, et dans le cas de la classification, le score F1. L'algorithme est construit de la manière suivante :

- On calcule le score de validation croisée sur l'ensemble de nos variables.
- On enlève la variable la moins importante selon l'indicateur d'importance intégré dans le modèle (l'algorithme n'est donc pas applicable pour les modèles ne présentant aucun coefficient d'importance comme le modèle des k plus proches voisins par exemple).
- On réitère les étapes précédentes en recalculant le score de validation croisée sans la variable que l'on vient d'enlever.
- On s'arrête une fois qu'il ne reste plus qu'une variable.
- On sélectionne pour finir le sous-ensemble de variables donnant le meilleur score de validation croisée.

e. Sélection de paramètres : Grid Search

Une fois les variables optimales sélectionnées pour chaque modèle, la méthode *Grid Search* nous permettra de sélectionner les paramètres optimaux. L'algorithme se décompose de la manière suivante :

- On définit préalablement un ensemble de valeurs pour les paramètres que l'on souhaite tester.
- Pour chacune des combinaisons possibles de valeurs de paramètres, on calcule le score de validation croisée des modèles.
- Enfin, on sélectionne la combinaison de paramètres fournissant le meilleur score de validation croisée.

5) Résultats de modélisation sur le risque pluvial

Maintenant que l'on a décrit les différents modèles et méthodes qui seront utilisés par la suite, commençons par déterminer le modèle optimal pour la prédiction du taux de destruction moyen par an pour le risque pluvial. À noter que les résultats qui seront présentés ci-dessous seront effectués avec la variable précipitation extrême, mais que le même processus sera effectué pour la variable nombre de jour de forte pluie et que les résultats obtenus à partir des deux variables seront comparés lors de la synthèse des résultats.

a. Distribution des taux de destruction du risque pluvial

Avant de passer aux modélisations, on peut s'intéresser à la distribution des taux de destruction. Une des difficultés pour la prédiction de ces taux est la forte concentration des risques avec de faibles taux, puis la grande dispersion

pour les risques davantage exposés au risque pluvial. Ainsi chaque erreur de modélisation entraînera potentiellement des conséquences importantes sur l'erreur quadratique moyenne, très sensible aux valeurs extrêmes.

1%	5%	10%	25%	50%	75%	90%	95%	99%
4.156814e-08	1.572390e-07	2.898480e-07	7.133069e-07	1.701912e-06	3.615043e-06	6.728059e-06	3.483353e-05	1.046057e-03

Tableau 11 - Quantiles des taux de destruction moyens par an pour le risque pluvial - base d'entraînement

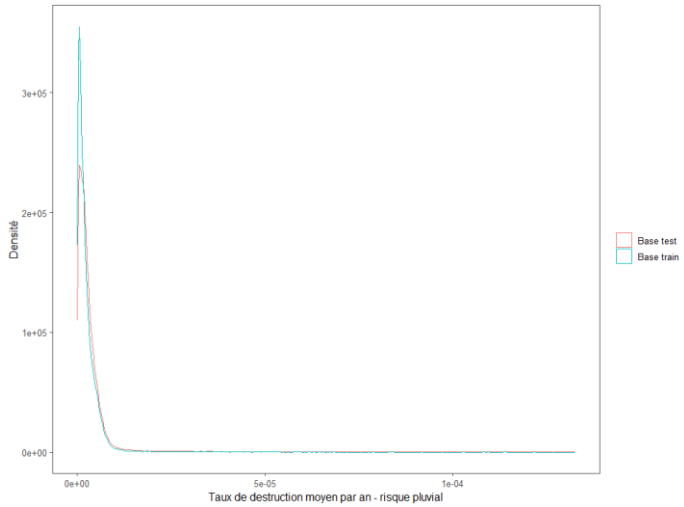

Figure 57 - Densité du taux de destruction moyen par an du risque pluvial

b. Résultats du GLM Gamma

Le premier modèle étudié est le modèle linéaire généralisé appliqué pour la distribution Gamma associée à la fonction de lien log. Cette distribution est bien adaptée à nos données, étant donné que la fonction Gamma est définie uniquement sur les réels strictement positifs.

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-5.244  -2.172  -1.737  -1.258  28.304

Coefficients:
                                Estimate Std. Error  t value Pr(>|t|)
(Intercept)                   -1.296e+01  8.807e-02 -147.118  < 2e-16 ***
Nombre_etages                 -4.060e-01  5.421e-02   -7.489 6.98e-14 ***
Distance_cours_deau           -4.423e-04  1.687e-05  -26.220  < 2e-16 ***
Topographic_Wetness_Index      1.689e-01  4.567e-03   36.990  < 2e-16 ***
Manning                       -2.141e+00  1.166e+00   -1.836 0.066300 .
Precip_extreme_Q99             6.636e-02  2.722e-03   24.384  < 2e-16 ***
Taux_engt_contenu              6.222e-01  6.176e-02   10.075  < 2e-16 ***
Taux_engt_PE                  -6.894e-01  2.535e-01   -2.720 0.006532 **
Portefeuille_1                 3.440e-01  2.393e-01    1.438 0.150523
Portefeuille_2                 1.410e-01  9.646e-02    1.462 0.143757
Portefeuille_3                 4.784e-01  1.486e-01    3.219 0.001285 **
Portefeuille_4                 9.249e-02  1.749e-01    0.529 0.596886
Portefeuille_5                 8.137e-01  1.955e-01    4.163 3.15e-05 ***
Portefeuille_6                -1.447e+00  2.096e-01   -6.905 5.05e-12 ***
Portefeuille_7                 2.632e-01  3.276e-01    0.803 0.421729
Portefeuille_8                 1.256e+00  2.968e-01    4.231 2.33e-05 ***
Portefeuille_9                -2.409e-01  7.040e-02   -3.421 0.000623 ***
Portefeuille_10               -2.108e-02  4.984e-02   -0.423 0.672358
Portefeuille_11               -2.653e-01  1.019e+00   -0.260 0.794667
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 78.84458)

    Null deviance: 1558147  on 244110  degrees of freedom
Residual deviance: 1165278  on 244092  degrees of freedom
AIC: -5239851

Number of Fisher Scoring iterations: 17
```

Tableau 12 - Résultats du GLM Gamma sur les taux de destruction du risque pluvial

Une majorité des variables est significative à 95% selon le test de Student à l'exception du coefficient de Manning qui dispose d'une p-valeur de 0,066 ainsi que différentes modalités de portefeuille. Notons que par confidentialité, nous n'exprimerons pas explicitement le nom des différents portefeuilles. Le taux d'engagement perte

d'exploitation ressort également avec une p-valeur supérieure aux autres, mais est tout de même significatif au seuil de 1%.

Variable_enlevee	Deviance	Pseudo_R2	AIC
Aucune	1165278	0.2521389	-5239851
Manning	1165630	0.2519130	-5239745
Taux_engt_PE	1165722	0.2518537	-5239717
Nombre_etages	1168927	0.2497967	-5238737
Taux_engt_contenu	1173335	0.2469681	-5237394
Portefeuille	1174090	0.2464832	-5237184
Precip_extreme_Q99	1216796	0.2190748	-5224369
Distance_cours_deau	1264665	0.1883535	-5210470
Topographic_Wetness_Index	1328444	0.1474207	-5192627

Tableau 13- Première étape de la procédure backward AIC pour le modèle GLM Gamma - taux de destruction pluvial

Pour la sélection des paramètres optimaux, on met en place une méthode descendante par AIC comme décrite dans la partie IV.3. La méthode nous indique que le GLM permettant de minimiser le critère est le modèle disposant de toutes les variables. En effet, l'algorithme s'arrête dès la première étape comme on peut le voir sur le tableau ci-dessus, peu importe la variable qui est retirée, l'AIC augmente. Notons tout de même les trois variables ayant la plus grande influence sur l'AIC et sur le pseudo-R2 lorsqu'on les retire. À savoir l'indice d'humidité topographique pour lequel le pseudo-R2 passe de 25,2 à 14,7% lorsqu'on le retire du modèle, la distance au cours d'eau provoquant une baisse de presque 7% du pseudo-R2 lorsqu'on la retire et enfin les précipitations extrêmes causant une baisse de 25,2 à 21,9%. On conserve donc l'ensemble des variables et on obtient les résultats suivants :

Model	Cross_val_RMSE	Cross_val_Total_AAL_diff	Cross_val_Spearman	TestSet_RMSE	TestSet_Total_AAL_diff	TestSet_Spearman
Gamma	0.0003522422	0.4553059	0.3272618	0.0003731653	0.07032336	0.2920786

Tableau 14 - Résultats des indicateurs de performance sur le modèle GLM Gamma optimal - taux de destruction pluvial

Les colonnes 2 à 4 nous indiquent les scores obtenus par validation croisée. En l'état la racine de l'erreur quadratique moyenne n'est pas vraiment interprétable, et prendra tout son sens lors de la comparaison des différents modèles entre eux.

Concernant le score d'AAL nous indiquant la différence de charge annuelle modélisée sur l'assiette totale entre la prédiction et l'observation, on obtient un score de 45,5%. Concrètement, en valeur absolue, la différence entre la charge annuelle moyenne inondation prédite et celle observée (c'est-à-dire celle qui est donnée par nos modélisateurs) est en moyenne de 45,5% sur les dix itérations de validation croisée.

Le coefficient de Spearman entre les valeurs prédites et celles observées et de 32,7%. Le modèle fait donc mieux que l'aléatoire, mais fournit un score qui reste relativement faible.

c. Résultats de la régression Bêta

La régression à partir de la distribution Bêta est à priori assez bien adaptée à nos données étant donné qu'elle est définie sur l'intervalle]0; 1[ce qui est parfait pour l'étude du taux de destruction pluvial qui est compris dans ce domaine de définition.

```
Pearson residuals:
     Min       1Q   Median       3Q      Max
 -0.7668  -0.4762  -0.4533  -0.4215 138.7661

Coefficients (mean model with logit link):
                                Estimate Std. Error   z value Pr(>|z|)
(Intercept)                   -1.063e+01  1.045e-02 -1016.933  < 2e-16 ***
Distance_cours_deau           -1.888e-05  1.850e-06   -10.205  < 2e-16 ***
Topographic_Wetness_Index      2.364e-02  4.887e-04    48.377  < 2e-16 ***
Manning                       -7.513e-01  1.269e-01    -5.918 3.25e-09 ***
Precip_extreme_Q99             1.259e-02  2.928e-04    42.988  < 2e-16 ***
Taux_engt_contenu              2.348e-02  6.724e-03     3.492 0.000479 ***
Taux_engt_PE                  -7.935e-02  2.761e-02    -2.874 0.004055 **
Portefeuille_1                 2.458e-01  2.598e-02     9.460  < 2e-16 ***
Portefeuille_2                 2.547e-01  1.050e-02    24.257  < 2e-16 ***
Portefeuille_3                 3.264e-01  1.608e-02    20.303  < 2e-16 ***
Portefeuille_4                 2.019e-01  1.903e-02    10.605  < 2e-16 ***
Portefeuille_5                 4.358e-01  2.100e-02    20.755  < 2e-16 ***
Portefeuille_6                -2.077e-01  2.319e-02    -8.958  < 2e-16 ***
Portefeuille_7                 2.302e-01  3.559e-02     6.467 9.97e-11 ***
Portefeuille_8                 4.660e-02  1.213e-02     3.841 0.000123 ***
Portefeuille_9                 1.417e-01  7.694e-03    18.423  < 2e-16 ***
Portefeuille_10                2.551e-01  5.465e-03    46.678  < 2e-16 ***
Portefeuille_11                8.665e-02  1.118e-01     0.775 0.438469

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  5445.34      25.85   210.7   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 2.583e+06 on 19 Df
Pseudo R-squared: 0.2195
Number of iterations: 107 (BFGS) + 19 (Fisher scoring)
```

Tableau 15 – Résultats du GLM Bêta sur les taux de destruction du risque pluvial

L'ensemble des variables est significatif au seuil de 1% selon le test de Wald à l'exception du portefeuille numéro 11. Le coefficient de précision ϕ introduit dans la partie IV.3 n'est ici pas modélisé par la régression et on utilise donc un coefficient constant.

Variable_enlevee	AIC
Aucune	-5166436
Taux_engt_PE	-5166430
Taux_engt_contenu	-5166426
Manning	-5166405
Distance_cours_deau	-5166343
Precip_extreme_Q99	-5164662
Portefeuille	-5163202
Topographic_Wetness_Index	-5160568

Tableau 16 - Première étape de la procédure backward AIC pour le modèle GLM Bêta - taux de destruction pluvial

Le modèle permettant de minimiser l'AIC est le modèle entraîné avec l'ensemble des variables. On retrouve ci-dessous les résultats obtenus en les conservant toutes. Le coefficient de Spearman est particulièrement bon avec un score atteignant les 50%, cependant le score d'AAL semble très élevé avec un score de 74%. On surestime donc largement la charge inondation totale en moyenne sur la validation croisée.

Model	beta_RMSE_cv	beta_Total_AAL_diff_cv	beta_Sperman_score_cv	beta_RMSE_test	beta_Total_AAL_diff_test	beta_Sperman_score_test
Beta	0.000349712	0.7410971	0.5006585	0.0003517979	0.1375932	0.3725111

Tableau 17 - Résultats des indicateurs de performance sur le modèle GLM Bêta optimal - taux de destruction pluvial

En complément de la moyenne μ on peut effectuer un autre modèle de régression Bêta en faisant également la prédiction du coefficient de précision ϕ, on pourra ainsi comparer à l'issue des modélisations les résultats des deux méthodes.

```
Pearson residuals:
     Min       1Q   Median       3Q      Max
 -1.0817  -0.5216  -0.4759  -0.4278 233.6557

Coefficients (mean model with logit link):
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                   -1.327e+01  1.853e-02 -716.212  < 2e-16 ***
Nombre_etages                 -3.958e-01  9.205e-03  -43.005  < 2e-16 ***
Distance_cours_deau           -2.846e-04  2.193e-06 -129.797  < 2e-16 ***
Topographic_Wetness_Index      1.754e-01  1.077e-03  162.867  < 2e-16 ***
Manning                       -2.247e+00  2.326e-01   -9.662  < 2e-16 ***
Precip_extreme_Q99             6.698e-02  5.774e-04  116.000  < 2e-16 ***
Taux_engt_contenu              5.784e-01  1.278e-02   45.271  < 2e-16 ***
Taux_engt_PE                  -6.448e-01  5.162e-02  -12.491  < 2e-16 ***
Portefeuille_1                 4.611e-01  5.117e-02    9.012  < 2e-16 ***
Portefeuille_2                 3.122e-01  1.901e-02   16.422  < 2e-16 ***
Portefeuille_3                 6.384e-01  2.991e-02   21.343  < 2e-16 ***
Portefeuille_4                 2.551e-01  3.621e-02    7.044 1.87e-12 ***
Portefeuille_5                 9.636e-01  3.910e-02   24.647  < 2e-16 ***
Portefeuille_6                -1.308e+00  4.667e-02  -28.019  < 2e-16 ***
Portefeuille_7                 3.820e-01  6.932e-02    5.511 3.56e-08 ***
Portefeuille_8                 1.360e+00  5.499e-02   24.724  < 2e-16 ***
Portefeuille_9                -6.681e-02  1.491e-02   -4.481 7.43e-06 ***
Portefeuille_10                1.046e-01  1.042e-02   10.032  < 2e-16 ***
Portefeuille_11               -1.151e-01  2.113e-01   -0.545    0.586

Phi coefficients (precision model with log link):
                                Estimate Std. Error  z value Pr(>|z|)
(Intercept)                    1.190e+01  2.155e-02  551.962  < 2e-16 ***
Nombre_etages                  4.958e-01  1.159e-02   42.771  < 2e-16 ***
Distance_cours_deau            4.909e-04  3.145e-06  156.103  < 2e-16 ***
Topographic_Wetness_Index     -2.024e-01  1.214e-03 -166.690  < 2e-16 ***
Manning                        1.527e+00  2.750e-01    5.553 2.81e-08 ***
Precip_extreme_Q99            -7.198e-02  6.699e-04 -107.452  < 2e-16 ***
Taux_engt_contenu             -7.628e-01  1.494e-02  -51.075  < 2e-16 ***
Taux_engt_PE                   7.553e-01  6.065e-02   12.452  < 2e-16 ***
Portefeuille_1                -1.727e-01  5.923e-02   -2.916  0.00355 **
Portefeuille_2                 7.416e-02  2.257e-02    3.286  0.00102 **
Portefeuille_3                -2.695e-01  3.527e-02   -7.639 2.18e-14 ***
Portefeuille_4                 6.796e-02  4.233e-02    1.605  0.10839
Portefeuille_5                -5.689e-01  4.621e-02  -12.311  < 2e-16 ***
Portefeuille_6                 1.443e+00  5.340e-02   27.023  < 2e-16 ***
Portefeuille_7                -8.130e-02  8.050e-02   -1.010  0.31255
Portefeuille_8                -1.448e+00  6.686e-02  -21.651  < 2e-16 ***
Portefeuille_9                 4.191e-01  1.731e-02   24.217  < 2e-16 ***
Portefeuille_10                3.605e-01  1.215e-02   29.679  < 2e-16 ***
Portefeuille_11                3.879e-01  2.469e-01    1.571  0.11627
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 2.637e+06 on 38 Df
Pseudo R-squared: 0.1375
Number of iterations: 129 (BFGS) + 12 (Fisher scoring)
```

Tableau 18 – Résultats du GLM Bêta (mu et phi) sur les taux de destruction du risque pluvial

À l'exception de certains portefeuilles l'ensemble des variables semblent significatives selon le test de Wald, que ce soit pour les coefficients liés à la moyenne μ ou au coefficient de précision ϕ. Les résultats de ce modèle sur l'ensemble des variables apparaissent comme ayant de plus mauvais scores RMSE et de Spearman, mais surestiment moins la charge globale inondation. On décide finalement de privilégier l'erreur quadratique moyenne et donc de conserver le modèle utilisant un coefficient de précision constant.

Model	beta_phi_RMSE_cv	beta_phi_Total_AAL_diff_cv	beta_phi_Sperman_score_cv	beta_phi_RMSE_test	beta_phi_Total_AAL_diff_test	beta_phi_Sperman_score_test
Beta phi	0.0003534736	0.4932673	0.3530316	0.000375469	0.05081976	0.294674

Tableau 19 - Résultats des indicateurs de performance sur le modèle GLM Bêta phi optimal - taux de destruction pluvial

d. Résultats du XGBoost

Le modèle essayé par la suite est un modèle de boosting de gradient. À noter que pour ce modèle, les variables catégorielles ne sont pas prises en charge, il a donc fallu encoder la variable portefeuille selon un encodage *One Hot* consistant simplement à rajouter autant de colonnes qu'il y a de catégories en plaçant un « 1 » dans la colonne associée à la bonne branche assurée et un « 0 » dans les autres.

Pour la sélection des variables, on utilise cette fois-ci la méthode d'élimination récursive de variables présentée dans la partie IV.4. Comme on peut le voir, le score RMSE le plus faible est atteint à l'étape 16 lorsque l'on retire la variable coefficient de Manning. Ce modèle est donc constitué des variables restantes, à savoir le taux d'engagement contenu, la distance au cours d'eau, la valeur de précipitations extrêmes et l'indice d'humidité topographique.
Cependant, il apparait d'après les scores que retirer des variables a une incidence négative sur le score d'AAL ainsi que sur le coefficient de Spearman. On décide donc de tester les deux méthodes, en optimisant les paramètres d'un côté pour le modèle conservant toutes les variables et de l'autre pour le modèle optimal composée d'uniquement ces quatre variables.

Var_remove	Step	RMSE_cv	Total_AAL_Diff_cv	Spearman_score_cv
Aucune	1	0.0003550859	52.237 %	23.237 %
Portefeuille_4	2	0.0003550805	52.2467 %	23.2343 %
Portefeuille_7	3	0.0003541425	55.9933 %	23.9244 %
Portefeuille_8	4	0.0003541425	55.9933 %	23.9244 %
Portefeuille_11	5	0.0003541425	55.9933 %	23.9244 %
Portefeuille_6	6	0.0003592077	59.7833 %	23.304 %
Nombre_etages	7	0.0003556103	67.3314 %	23.0303 %
Portefeuille_1	8	0.0003570924	65.5535 %	23.5335 %
Portefeuille_2	9	0.0003543852	71.4013 %	23.3639 %
Portefeuille_3	10	0.0003559469	67.1261 %	22.6379 %
Taux_engt_PE	11	0.0003523727	64.9745 %	22.613 %
Portefeuille_5	12	0.0003577005	66.0623 %	22.6781 %
Portefeuille_9	13	0.0003524898	61.3166 %	23.7544 %
Portefeuille_10	14	0.0003510607	58.4659 %	22.4932 %
Portefeuille_12	15	0.0003529202	66.177 %	22.0054 %
Manning	16	0.0003524497	69.4342 %	22.2342 %
Taux_engt_contenu	17	0.0003541762	80.3575 %	20.6696 %
Distance_cours_deau	18	0.0003645343	84.8325 %	16.4003 %
Precip_extreme_Q99	19	0.0003591932	75.6139 %	10.6725 %

Tableau 20 - Algorithme de Recursive Feature Elimination appliqué sur le XGBoost risque pluvial

On commence tout d'abord par optimiser le modèle composé de l'ensemble des variables. Pour la recherche des meilleurs paramètres, on utilise l'algorithme *GridSearch* présenté dans la partie IV.4.
On trouve ci-dessous un exemple d'application de l'algorithme en faisant varier le paramètre de profondeur maximum des arbres *max_depth* ainsi que le taux d'apprentissage *eta.* À noter que davantage de valeurs de ces paramètres ont été testées, mais qu'on affiche ici celles qui ont fourni les meilleurs résultats pour plus de lisibilités. Ainsi dans notre cas une profondeur d'arbre de 8 et un coefficient eta de 0.05 permet de minimiser notre RMSE et ce sont les paramètres qui sont donc choisis pour ce modèle.

nrounds	max_depth	eta	RMSE_cv	Total_AAL_Diff_cv	Spearman_score_cv
500	8	0.05	0.0003439913	0.5169507	0.2701866
500	12	0.05	0.0003459048	0.5572078	0.2777733
500	10	0.05	0.0003471070	0.4972329	0.2721229
500	8	0.10	0.0003482824	0.5400867	0.2685089
500	10	0.10	0.0003490213	0.5841085	0.2653987
500	12	0.10	0.0003496803	0.5219579	0.2691177
500	8	0.15	0.0003513367	0.5669707	0.2686404
500	10	0.15	0.0003520759	0.5123393	0.2659527
500	12	0.15	0.0003544985	0.5227634	0.2602715

Tableau 21 - Recherche des meilleurs paramètres pour le XGBoost toutes variables

Les résultats obtenus pour ce modèle sont assez convaincants et fournissent le plus faible RMSE obtenu jusque-là tout en maintenant un score AAL relativement faible par rapport aux autres modèles et un coefficient de Spearman à 27%.

Model	GB_all_RMSE_cv	GB_all_Total_AAL_diff_cv	GB_all_Sperman_score_cv	GB_RMSE_test	GB_Total_AAL_diff_test	GB_Sperman_score_test
XGB toutes variables	0.0003439913	0.5169507	0.2701866	0.0003533154	0.06581361	0.2704108

Tableau 22 - Résultats des indicateurs de performance sur le modèle XGBoost toutes variables optimal - taux de destruction pluvial

On effectue ensuite le même processus sur le modèle de boosting entraîné uniquement sur les quatre variables optimales. On choisit 6 comme profondeur maximale de l'arbre et un coefficient eta de 0.05 selon l'algorithme de *GridSearch* :

nrounds	max_depth	eta	RMSE_cv	Total_AAL_Diff_cv	Spearman_score_cv
500	6	0.05	0.0003448124	0.6061714	0.2469704
500	6	0.10	0.0003466815	0.6281912	0.2470072
500	8	0.05	0.0003472852	0.5976761	0.2716689
500	8	0.10	0.0003484072	0.6187265	0.2724474
500	10	0.05	0.0003485522	0.5505397	0.2749939
500	10	0.10	0.0003489509	0.5516379	0.2715067
500	4	0.05	0.0003493188	0.6662277	0.2248211
500	4	0.10	0.0003501237	0.7253023	0.2284642
500	2	0.05	0.0003586221	0.9242772	0.1800029
500	2	0.10	0.0003620389	0.9241753	0.1710300

Tableau 23 - Recherche de meilleurs paramètres pour le XGBoost après RFE

En entraînant notre modèle sur les trois variables sélectionnées précédemment et en utilisant les paramètres optimaux déterminés ci-dessus on obtient les résultats suivants :

Model	GB_RMSE_cv	GB_Total_AAL_diff_cv	GB_Sperman_score_cv	GB_RMSE_test	GB_Total_AAL_diff_test	GB_Sperman_score_test
XGB RFE	0.0003448124	0.6061714	0.2469704	0.0003511347	0.1675651	0.2537538

Tableau 24 - Résultats des indicateurs de performance sur le modèle XGBoost après RFE optimal - taux de destruction pluvial

Il apparait finalement que le modèle donnant les meilleurs indicateurs est le modèle optimisé avec l'ensemble des variables. Il donne en effet un RMSE et un score AAL plus faible, ainsi qu'une corrélation de Spearman plus élevée.

e. Résultats de la forêt aléatoire

Enfin, le dernier modèle considéré est une forêt aléatoire. Pour la sélection de variables, on utilise également un algorithme d'élimination récursive de variables.

Var_remove	Step	RMSE_cv	Total_AAL_Diff_cv	Spearman_score_cv
Aucune	1	0.0003408781	56.2188 %	38.1469 %
Nombre_etages	2	0.0003424758	59.5042 %	38.4926 %
Taux_engt_PE	3	0.0003432878	59.2522 %	38.7581 %
Manning	4	0.0003450794	65.0339 %	39.1809 %
Portefeuille	5	0.0003489070	67.8058 %	34.7624 %
Taux_engt_contenu	6	0.0003493729	76.055 %	29.1201 %
Precip_extreme_Q99	7	0.0003611373	89.3274 %	18.3157 %
Distance_cours_deau	8	0.0004286591	85.2779 %	7.2265 %

Tableau 25 - Algorithme de Recursive Feature Elimination appliqué sur le Random Forest - taux de destruction pluvial

Le modèle optimal est le modèle conservant l'ensemble des variables, puisqu'il permet en effet de minimiser le RMSE ainsi que le score d'AAL, tout en ayant quasiment le meilleur score de Spearman. On décide donc de conserver l'ensemble des variables. Notons tout de même les cinq variables qui semblent avoir le plus d'impact sur la dégradation de la corrélation de Spearman, à savoir la variable portefeuille, le taux d'engagement contenu,

les précipitations extrêmes, la distance au cours d'eau ainsi que l'indice d'humidité topographique qui est l'ultime variable restante en sortie d'algorithme.

mtry	max_depth	num.trees	RMSE_cv	Total_AAL_Diff_cv	Spearman_score_cv
0	0	500	0.0003408781	0.5621884	0.3814687
4	0	500	0.0003468016	0.6647166	0.3918788
0	4	500	0.0003378189	0.7678506	0.3181121
4	4	500	0.0003378774	0.8232656	0.2984332
0	8	500	0.0003374656	0.6841055	0.3505086
4	8	500	0.0003397750	0.6975323	0.3256874
0	12	500	0.0003388688	0.6239061	0.3659344
4	12	500	0.0003431642	0.6727856	0.3594247

Tableau 26 - Recherche de meilleurs paramètres pour le Random Forest - taux de destruction pluvial

Il apparait que le RMSE minimal est obtenu pour une profondeur maximale de 8 et une valeur de *mtry (*nombre de variables testées à chaque division) par défaut, qui est renseigné comme étant la racine du nombre de variables explicatives. Cependant, ce choix de paramètres dégrade fortement le score d'AAL ainsi que le score de Spearman. On décide donc de finalement utiliser les paramètres par défaut qui permettent d'obtenir un bon compromis entre les trois indicateurs avec le plus faible score d'AAL et quasiment la meilleure corrélation de Spearman parmi les différents paramètres testés.

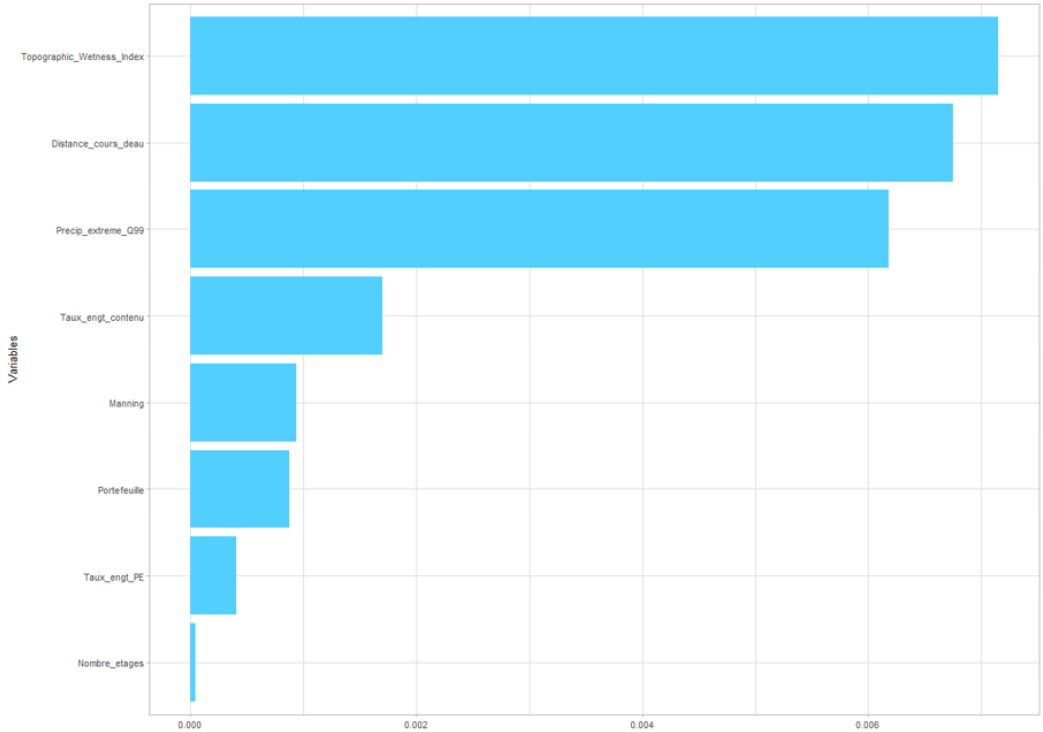

Figure 58 - Influence relative des variables du Random Forest - taux de destruction pluvial

On a représenté ci-dessus l'apport de chacune des variables dans la forêt aléatoire. On remarque que la variable sur l'indice d'humidité topographique est celle qui apporte le plus d'informations, suivie de près par la distance au

cours d'eau et par les précipitations extrêmes. Notons tout de même l'apport non négligeable du taux d'engagement contenu, du coefficient de Manning et de la variable portefeuille dans le modèle.

Model	Cross_val_RMSE	Cross_val_Total_AAL_diff	Cross_val_Spearman	TestSet_RMSE	TestSet_Total_AAL_diff	TestSet_Spearman
Random Forest	0.0003407000	0.5834091	0.3809876	0.0003528151	0.11566119	0.3306166

Tableau 27 - Résultats des indicateurs de performance sur le modèle Random Forest - taux de destruction pluvial

f. Synthèse des résultats et choix du modèle

Avant de s'intéresser au tableau final présentant l'ensemble des modèles et leurs scores associés, on souhaite s'intéresser à la distribution de nos variables afin de s'assurer que l'allure des taux de destruction prédits est similaire aux taux de destruction fournis par notre modélisateur.

On remarque directement que les deux modèles les plus fidèles à la distribution de la base de test sont les deux modèles de machine learning, à savoir le boosting de gradient et la forêt aléatoire. Les autres modèles ont des densités beaucoup moins représentatives de la réalité et notamment le modèle de régression Bêta qui fournissait un coefficient de Spearman intéressant, mais qui est cependant totalement décalé et qui surestime grandement la plupart des taux de destruction.

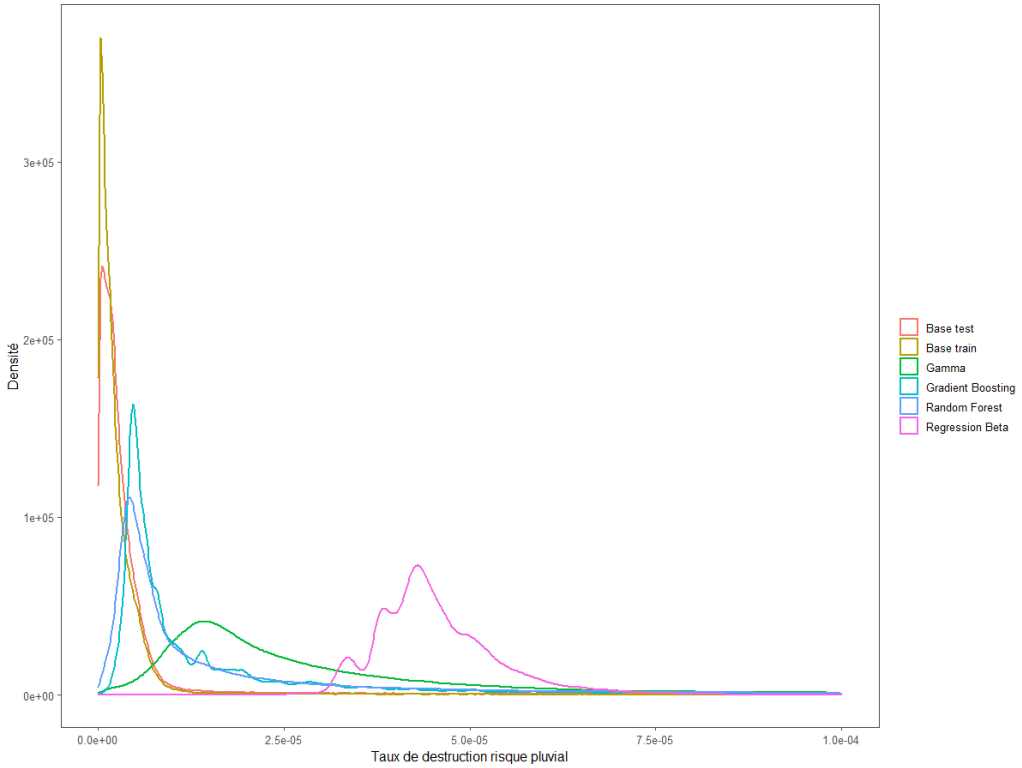

Figure 59 – Courbe de distribution des taux de destruction modélisés sur la base de test et ceux observés – risque pluvial

On obtient des résultats similaires en représentant la courbe de distribution des résidus, les deux modèles impliquant des arbres de décisions possèdent les pics de densité les plus importants au niveau des faibles valeurs d'erreur.

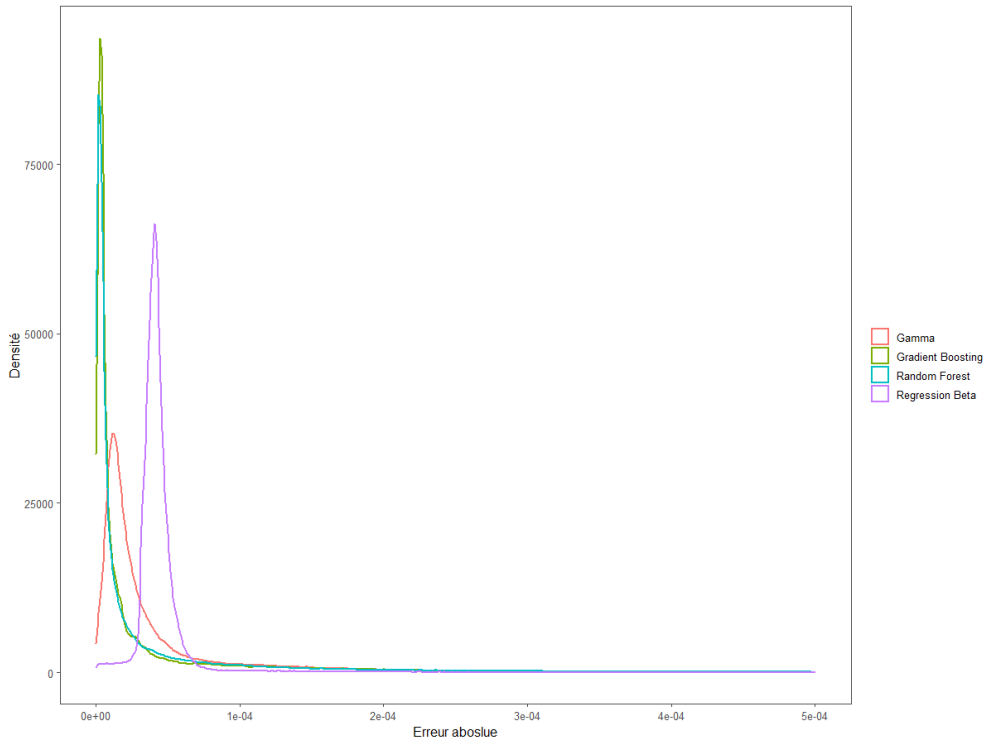

Figure 60 - Courbe de distribution de la valeur absolue des résidus – risque pluvial

Ainsi, au vu de ces courbes de distribution, ainsi qu'en se basant sur le tableau récapitulatif ci-dessous on décide de sélectionner pour nos modélisations un modèle de forêt aléatoire qui permet de minimiser le RMSE par rapport aux autres modèles. De plus, ce modèle permet d'obtenir la deuxième meilleure corrélation de Spearman tout en ayant un score d'AAL relativement faible par rapport à ce qu'on peut obtenir sur la régression Bêta.

Model	Cross_val_RMSE	Cross_val_Total_AAL_diff	Cross_val_Spearman	TestSet_RMSE	TestSet_Total_AAL_diff	TestSet_Spearman
Random Forest	0.0003407000	0.5834091	0.3809876	0.0003528151	0.11566119	0.3306166
Gradient Boosting	0.0003439913	0.5169507	0.2701866	0.0003533154	0.06581361	0.2704108
Regression Beta	0.0003497120	0.7410971	0.5006585	0.0003517979	0.13759318	0.3725111
Gamma	0.0003522422	0.4553059	0.3272618	0.0003731653	0.07032336	0.2920786

Tableau 28 - Synthèses des résultats de modèles - taux de destruction pluvial - avec variable précipitations extrêmes

Enfin, nous avions fait le choix dans la partie de présélection des variables de faire exactement les mêmes processus de sélections de modèles, de variables et de paramètres en le faisait d'un côté pour la variable précipitations extrêmes et de l'autre pour la variable nombre jour de forte pluie. Finalement, au vu du tableau ci-dessous qui présente les résultats pour la variable nombre de jour de forte pluie, il apparait que la variable précipitations extrêmes semble fournir de meilleurs résultats, on décide donc de sélectionner cette dernière.

Model	Cross_val_RMSE	Cross_val_Total_AAL_diff	Cross_val_Spearman	TestSet_RMSE	TestSet_Total_AAL_diff	TestSet_Spearman
Random Forest	0.0003418475	0.9772993	0.3812585	0.0003539772	0.12633679	0.322887
Gradient Boosting	0.0003492653	0.7380426	0.2790522	0.0003558311	0.10372548	0.271702
Regression Beta	0.0003497024	0.7394301	0.4984336	0.0003517633	0.13853446	0.375785
Gamma	0.0003527632	0.4287957	0.3239399	0.0003819330	0.09307292	0.291592

Tableau 29- Synthèses des résultats de modèles - taux de destruction pluvial - avec variable nombre de jour de forte pluie

6) Résultats de modélisation sur le risque fluvial

Cette partie vise à présenter les différents résultats de modèles obtenus pour la prédiction de la variable concernant le taux de destruction annuel moyen pour le risque fluvial. Cette étape de prédiction se fait en deux étapes, une première étape de classification et une deuxième étape de régression dépendant des résultats de la première étape. Le processus complet est détaillé ci-dessous.

a. Méthodologie de prédiction

Soit $X_i \in \{0,1\}$ la variable aléatoire binaire qu'on a également appelée *Exposed_fluvial* qui permet d'identifier si le risque est exposé au risque fluvial ou non. Soit $\widehat{X}_\iota$ la prédiction de cette variable obtenue par le biais du modèle de classification qui sera sélectionné dans la partie suivante.

Soit $Y_i \in]0,1]$ la variable continue qui correspond au taux de destruction moyen par an pour le risque fluvial, uniquement pour les risques exposés à ce risque (autrement dit pour les risques ayant un taux de destruction non nul). Soit $\widehat{Y}_\iota$ la prédiction de cette variable obtenue par le biais du modèle qui sera sélectionné dans cette partie. L'entraînement sera fait uniquement sur les risques de la base d'entraînement ayant un taux de destruction strictement positif. Sur cette base 95% des risques ont un taux de destruction nul (*Exposed_fluvial* = 0), on effectue donc l'entraînement de ces modèles sur seulement 5% de la base d'entraînement, soit environ 12 500 risques sur un total de 250 000 risques.

Soit $Z_i \in [0,1]$ la variable du taux de destruction moyen par an pour le risque fluvial qui inclut les taux de destruction nuls. La prédiction de cette variable sera effectuée à partir des prédictions effectuées précédemment à partir du modèle de classification pour X_i ainsi que pour le modèle de régression pour Y_i.

$$\widehat{Z}_\iota = \widehat{X}_\iota * \widehat{Y}_\iota$$

À noter que les indicateurs de performances calculés dans la partie régression (RMSE, score d'AAL et de Spearman), seront calculés à partir de $\widehat{Z}_\iota$ directement et non de $\widehat{Y}_\iota$.

Pour rappel, 95% des taux de destruction pour le risque fluvial sont nuls, il est donc indispensable de passer par cette première étape de classification avant d'effectuer notre régression. De plus, on représente ci-dessous la distribution des 12500 taux de destruction strictement positifs de notre base d'entraînement. Globalement, le constat est le même que pour la distribution des taux pour le risque pluvial avec une grande concentration des faibles taux puis une queue de distribution très étendue. Pour plus de lisibilités, on a limité la courbe des abscisses au quantile à 95% des taux de destruction strictement positif soit à environ 0,0017.

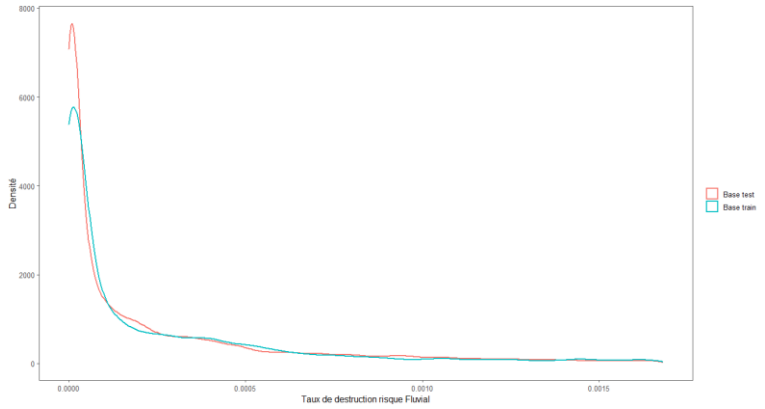

Figure 61 - Densité du taux de destruction moyen par an du risque fluvial (avec exclusion des taux nuls)

b. Synthèses des résultats et choix du modèle de classification

Contrairement à la partie pluviale, on ne présentera pas ici le choix des paramètres et variables des différents modèles, la méthode étant sensiblement la même, on pourra trouver le détail du processus en annexe 6.

On présente ci-dessous le récapitulatif des résultats pour l'ensemble des trois modèles. Le modèle de forêt aléatoire ainsi que celui de boosting de gradient ont des résultats très similaires contrairement à la régression logistique qui présente des résultats largement inférieurs.

Pour la prédiction de la variable *Exposed_fluvial*, on décide alors de sélectionner le modèle permettant de maximiser le score F1, à savoir la forêt aléatoire. En faisant la moyenne sur les dix itérations de la validation croisée on obtient en effet une précision de 43%, un rappel 64% et un score F1 d'environ 51%. En calculant ces indicateurs sur la prédiction obtenue avec la base de test on obtient le même classement dans les modèles avec un score F1 de 59% pour le *random forest*, 58% pour le boosting de gradient et 43% pour la régression logistique. Comme détaillé en annexe 6, le choix des paramètres avec l'algorithme GridSearch a fait apparaitre que ce score optimal de la forêt aléatoire était obtenu avec un nombre d'arbres de 100 ainsi qu'une profondeur maximale d'arbre de 12.

Modèle	Seuil_opti_cv	Cross_val_precision	Cross_val_rappel	Cross_val_f_score	TestSet_precision	TestSet_rappel	TestSet_f_score
Random Forest	0.234	0.4282011	0.6367665	0.5057002	0.5242377	0.6664884	0.5868660
Gradient Boosting	0.248	0.4271718	0.6252551	0.5009043	0.5157330	0.6658199	0.5812440
GLM binomial logit	0.185	0.3026263	0.5008373	0.3710373	0.4193827	0.4331640	0.4261619

Tableau 30 - Résultats des indicateurs de performance sur l'ensemble des modèles – Exposed_fluvial

Concernant l'importance des variables, la distance à la rivière large, la différence d'altitude à la rivière large ainsi que la distance à la rivière ressortent particulièrement. À l'inverse, le taux d'engagement pertes d'exploitation, ainsi que le nombre d'étages sur le portefeuille immeuble, semblent avoir peu d'incidence sur l'exposition au risque fluvial.

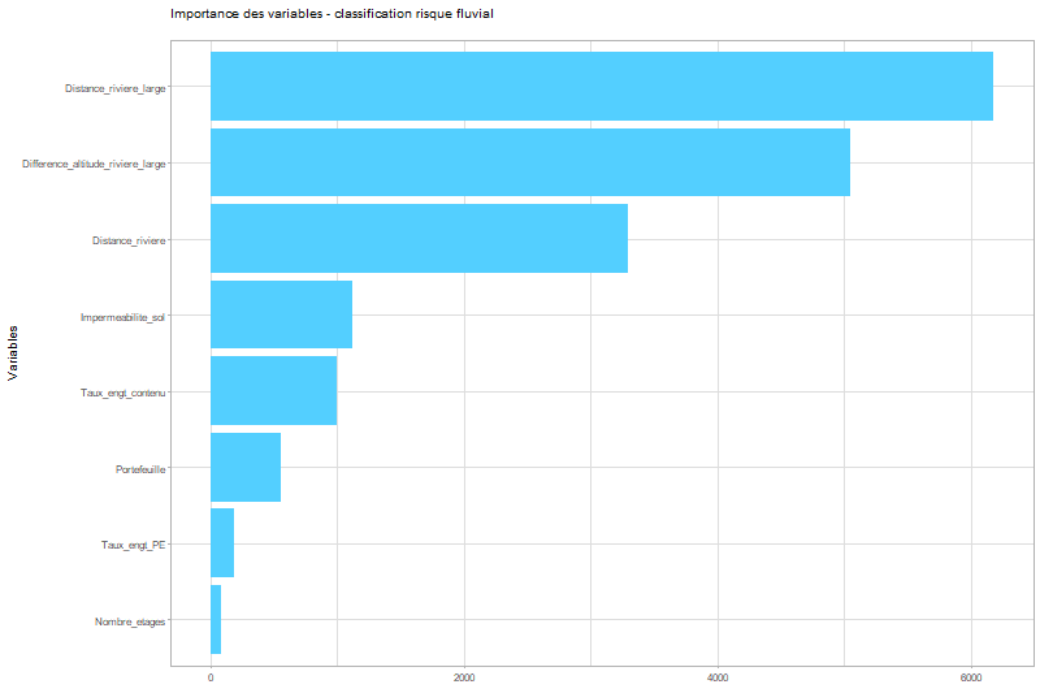

Figure 62 - Influence relative des variables du Random Forest – classification risque fluvial

On peut également comparer les résultats des modèles en traçant leurs courbes ROC et PR obtenues sur la base de test. Ces représentations sont intéressantes pour mesurer l'apport prédictif des modèles en comparaison d'une simple prédiction aléatoire représentée par le tracé violet. Pour rappel, comme introduit dans la partie IV.1, ces courbes représentent l'évolution de différents indicateurs en faisant varier le seuil de classification de 0 à 1. On a représenté par un point noir le seuil de classification choisi lors de la validation croisée, en maximisant la moyenne des scores F1. Les points rouges désignent les seuils optimaux qu'il aurait fallu prendre pour maximiser le score F1 sur la base de test. On observe que pour les deux modèles d'arbres de classification, le seuil choisi semble satisfaisant vu qu'il est assez proche du seuil optimal.

De plus, il apparait comme on a pu le voir dans le tableau ci-dessus que les courbes associées à la régression logistique sont davantage rapprochées des courbes obtenues par tirage aléatoire, indiquant ainsi un pouvoir prédictif inférieur aux autres modèles.

Un autre indicateur intéressant est l'aire sous la courbe (AUC) ROC. Il est de 86,4% sur le modèle de forêt aléatoire et de 87,1% pour le boosting de gradient, que l'on peut donc comparer aux 50% obtenus lors d'un tirage aléatoire. Concrètement, cela signifie qu'en sélectionnant deux risques aléatoirement, l'un exposé au risque fluvial, l'autre non, la probabilité d'identifier correctement le risque exposé au risque fluvial parmi les deux est de 86,4% pour le *random forest* contre 87,1% pour le *xgboost*.

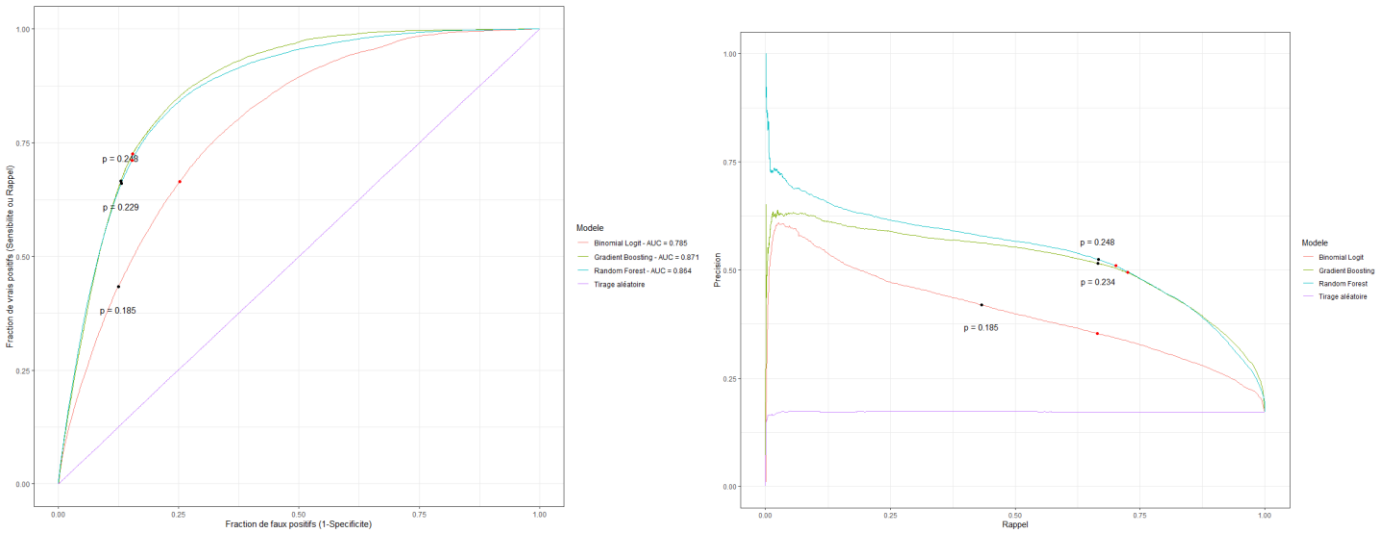

Figure 63 - Courbe ROC et PR sur la base de test pour l'ensemble des modèles – risque fluvial

Enfin, on peut établir un bilan des indicateurs de classification obtenus sur la base de test avec le modèle de forêt aléatoire en représentant sa matrice de confusion. Dans l'ensemble, on observe une exactitude (*accuracy*) globale de 83,9% et donc une part d'erreur de 16,1%. Cependant comme indiqué en partie IV.1 l'exactitude n'est pas l'indicateur le plus adapté dans le cas de classes déséquilibrées, et qu'en considérant simplement un modèle naïf qui prédirait uniquement des « 0 », l'exactitude obtenue serait de 82,8%. Pour mieux visualiser le pouvoir prédictif de notre modèle, considérons par exemple le score F1 qui n'est pas influencé par la présence de classes déséquilibrées. Le score obtenu est 58,7%, à noter que par mesure de comparaison, en effectuant une prédiction totalement aléatoire, on obtient un score F1 de 25,5%, soit près de deux fois inférieur au score obtenu avec le modèle sélectionné.

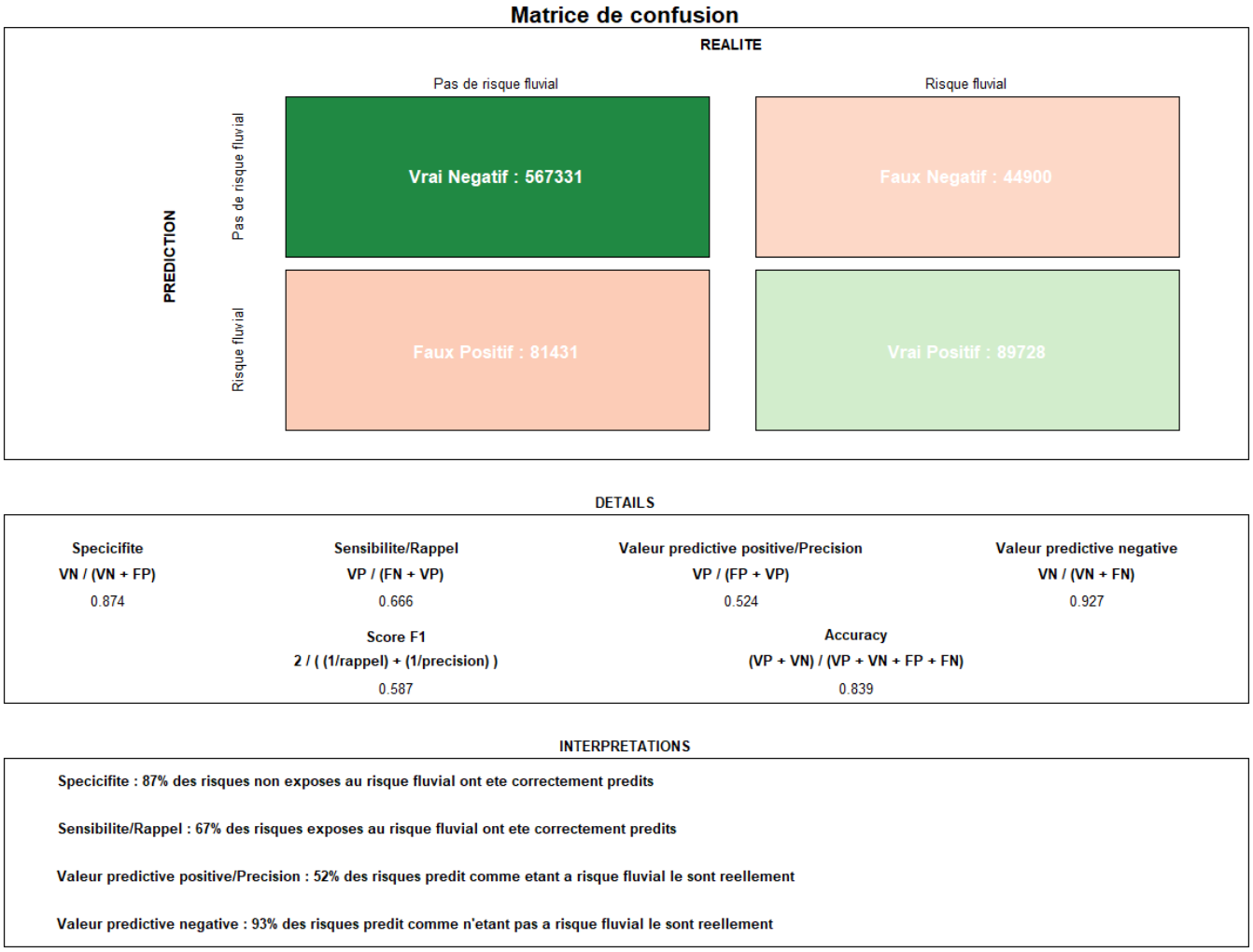

Figure 64 - Matrice de confusion obtenue avec le modèle Random Forest - Exposed_fluvial

c. Synthèses des résultats et choix du modèle de régression

Dans un deuxième temps, on peut entraîner nos modèles de régression sur les risques ayant un taux de destruction strictement positif puis tester ces modèles sur les risques considérés comme positifs selon le modèle de classification précédemment sélectionné. On trouvera plus de détails sur la construction des modèles en annexe 7. Les modèles considérés sont les mêmes que pour le taux de destruction du risque pluvial, à savoir le GLM Gamma, la régression Beta, le modèle *xgboost* et enfin le modèle de forêt aléatoire. Avant de passer à l'étape de comparaison des indicateurs de performance entre les modèles, on décide tout d'abord de comparer la distribution des taux de destruction prédits en comparaison et ceux fournis par notre modélisateur. Il semble que contrairement au risque pluvial aucune des distributions n'est réellement convaincante. Le modèle de forêt aléatoire prédit un trop grand nombre de faibles taux, tandis que le boosting de gradient semble totalement irrégulier, surement dû au faible nombre de variables sélectionnées pour ce modèle. Encore une fois, la distribution de la régression bêta est beaucoup trop excentrée au niveau des valeurs extrêmes et le modèle gamma semble donc cette fois le plus fidèle à la distribution des données.

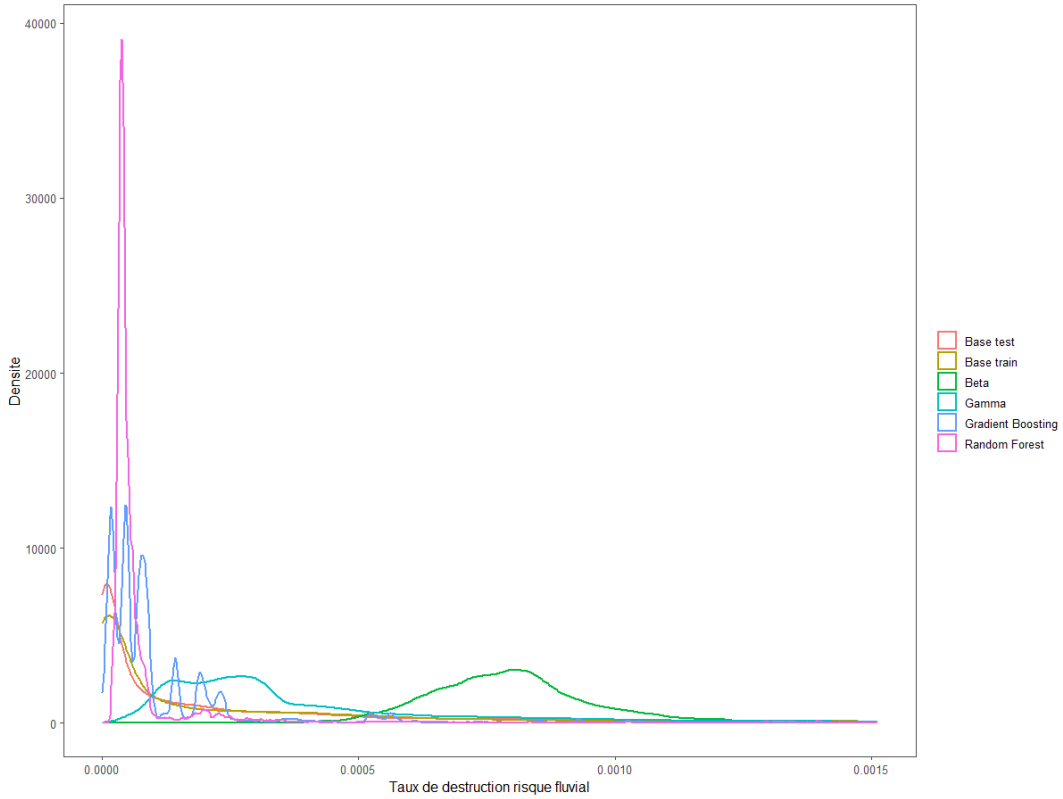

Figure 65 - Courbe de distribution des taux de destruction moyens par an – taux de destruction risque fluvial

Le constat est le même en s'intéressant aux indicateurs de performance et c'est le modèle GLM gamma qui ressort avec le RMSE le plus faible, on décide donc de sélectionner ce modèle. De plus, c'est également ce modèle qui permet de minimiser le score d'AAL parmi les quatre. Le score de Spearman quant à lui est globalement stable entre les différents modèles et est principalement lié à l'effet de la classification précédente. Notons que le coefficient de Spearman est meilleur que pour le risque pluvial, vu qu'il atteint quasiment les 50%. Il apparait également que l'on semble surestimer encore une fois la charge totale de près de 50% en moyenne sur les dix itérations de validation croisée.

Model	Cross_val_RMSE	Cross_val_Total_AAL_diff	Cross_val_Spearman	TestSet_RMSE	TestSet_Total_AAL_diff	TestSet_Spearman
Gamma	0.001198327	0.4857650	0.4917637	0.001198234	0.5183526	0.4978091
Beta	0.001199956	1.5136707	0.4930183	0.001222359	1.9073025	0.5019801
Random Forest	0.001252056	0.8550581	0.4915053	0.001164390	-0.6047074	0.5072335
Gradient Boosting	0.001310219	0.7572280	0.4919843	0.001361898	-0.3838808	0.5031777

Tableau 31 - Synthèses des résultats de modèles - taux de destruction fluvial

Ainsi la modélisation du risque fluvial s'est faite en deux étapes. Tout d'abord une première étape de classification avec un *random forest* permettant d'identifier si un risque est exposé au risque fluvial ou non, puis une deuxième étape de régression à l'aide d'un GLM gamma permettant de prévoir le taux de destruction des sites assurés exposés au risque fluvial.

V. Exploitation des modèles et impact du changement climatique

Maintenant que l'on dispose de modèles permettant de déterminer pour chacun des sites assurés un taux de destruction moyen par an et donc une charge annuelle moyenne, aussi bien pour le risque « fluvial » que « pluvial », on souhaite désormais tester le modèle sur l'ensemble du portefeuille afin d'en déduire des résultats sur les régions et portefeuilles les plus exposés au risque inondation.

Dans un deuxième temps, on appliquera les projections climatiques Cordex à notre modèle afin d'en déduire un impact sur la sinistralité future. Étant donné que les résultats de modélisation sur le risque fluvial ne dépendent pas du climat, on limitera l'étude du changement climatique au seul risque pluvial pour lequel les sorties du modèle dépendent des précipitations extrêmes (quantile à 99% des précipitations journalières).

1) Résultats sur la base de test et comparaison avec les sorties du modélisateur

Avant de passer à l'analyse sur l'ensemble du portefeuille, on souhaite tout d'abord analyser les résultats sur la base de test afin de pouvoir comparer les sorties avec les informations fournies initialement par le modélisateur. Pour rappel, la base de test est composée d'environ 750 000 risques choisis préalablement par Groupama comme étant à fort risque inondation. Les résultats sur cette base ne sont donc pas représentatifs de l'ensemble du portefeuille et sont présentés uniquement dans un objectif de comparaison avec les sorties du modélisateur.

a. Résultats sur le risque pluvial

On remarque sur les cartographies ci-dessous que notre modèle a tendance à surestimer les charges dans le nord de la France et notamment dans les départements de l'Orne et de l'Eure qui ressortent en bleu foncé. Le sud et nord-est de la France semblent mieux modélisés et ressortent comme davantage touchés que la partie nord-ouest. Notons particulièrement les Pyrénées-Atlantiques, la région Occitanie, certains départements des Alpes et des Vosges qui ressortent comme particulièrement touchés dans les deux modélisations.

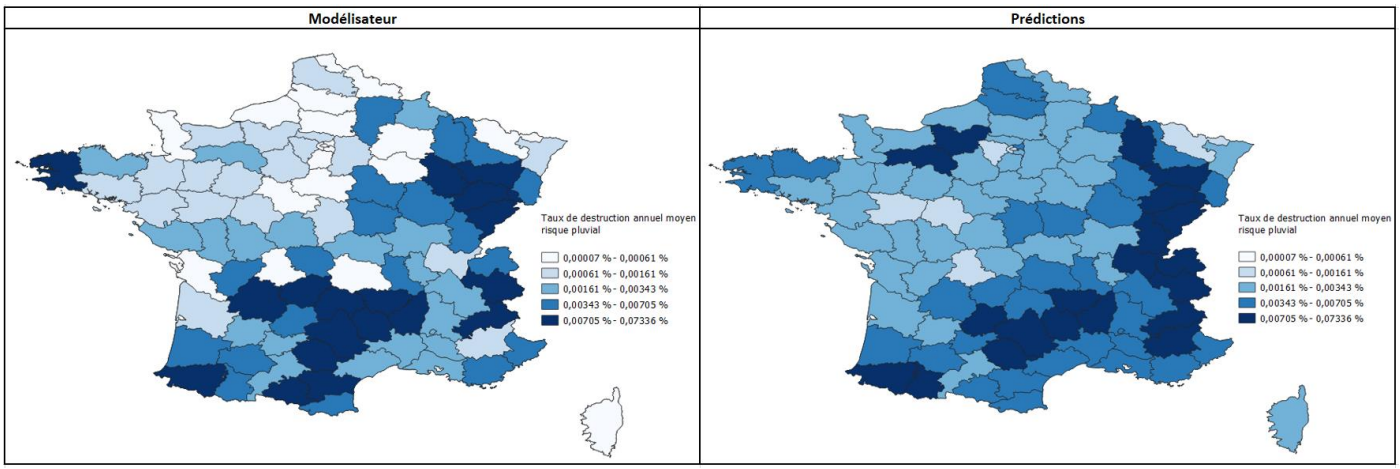

Figure 66 - Résultats du modèle pluvial par département sur la base de test

On s'intéresse par la suite aux résultats par portefeuille. On observe que le portefeuille A apparait comme particulièrement touché par les inondations avec un taux de destruction de 0.008% selon le modélisateur, soit le portefeuille le plus touché selon eux, contre 0.006% selon notre modèle, faisant de ce portefeuille le plus impacté par les inondations derrière le B. Notons également une modélisation un peu trop élevée sur le portefeuille F avec un taux de destruction de 0.004% selon notre modèle contre seulement 0.002% selon le modélisateur. Enfin, la

modélisation du portefeuille H semble quant à elle globalement bonne avec un taux égal à 0.005% dans les deux cas.

Portefeuille	Taux destruction pluvial modélisateur	Taux destruction pluvial prédicition	Classement modélisateur	Classement prediction
Branche A	0.006%	0.005%	2	4
Branche B	0.005%	0.007%	5	1
Branche C	0.005%	0.005%	4	3
Branche D	0.008%	0.006%	1	2
Branche F	0.002%	0.004%	7	6
Branche G	0.001%	0.002%	8	7
Branche H	0.005%	0.005%	3	5
Branche I	0.003%	0.002%	6	8
Total général	**0.004%**	**0.004%**		

Tableau 32 - Comparaison des résultats par portefeuille entre le modélisateur et les prédictions internes – risque pluvial

Par la suite, on décide de comparer la courbe de Lorenz obtenue avec notre modèle et celle obtenue par le modélisateur. La courbe de Lorenz permet de représenter la distribution d'une variable aléatoire (revenus, patrimoine, sinistralité, etc.) au sein d'une population et permettait initialement à celui qui l'a introduit, l'économiste Max O.Lorenz, de représenter les inégalités salariales. Concrètement, on représentera dans notre cas, la fonction de répartition qui associe à chaque quantile la part de sinistralité causée par celui-ci. Elle associe donc la part Y de sinistralité causée par une fraction X des sites assurés.

Cette courbe est très intéressante dans la gestion des risques et permet de visualiser quelle proportion des risques est à l'origine de quelle proportion des inondations. On décide donc de la comparer avec la courbe obtenue par le modélisateur afin de s'assurer de la cohérence entre les deux résultats. Par exemple en considérant l'avant-dernier trait, il apparait que 87,5% des sites assurés sont responsables de 12,5% de la sinistralité selon notre modèle contre seulement 3% selon le modélisateur. Ainsi à l'inverse, cela signifie que les 12,5% des sites assurés les plus à risques causent 87,5% de la charge annuelle moyenne pour le risque pluvial selon les prédictions internes contre 97% selon le modélisateur. Si l'inégalité est flagrante quel que soit le modèle considéré, notre modèle a tendance à davantage répartir la charge inondation entre les risques et le côté extrême de certains risques est moins marqué que du côté du modélisateur.

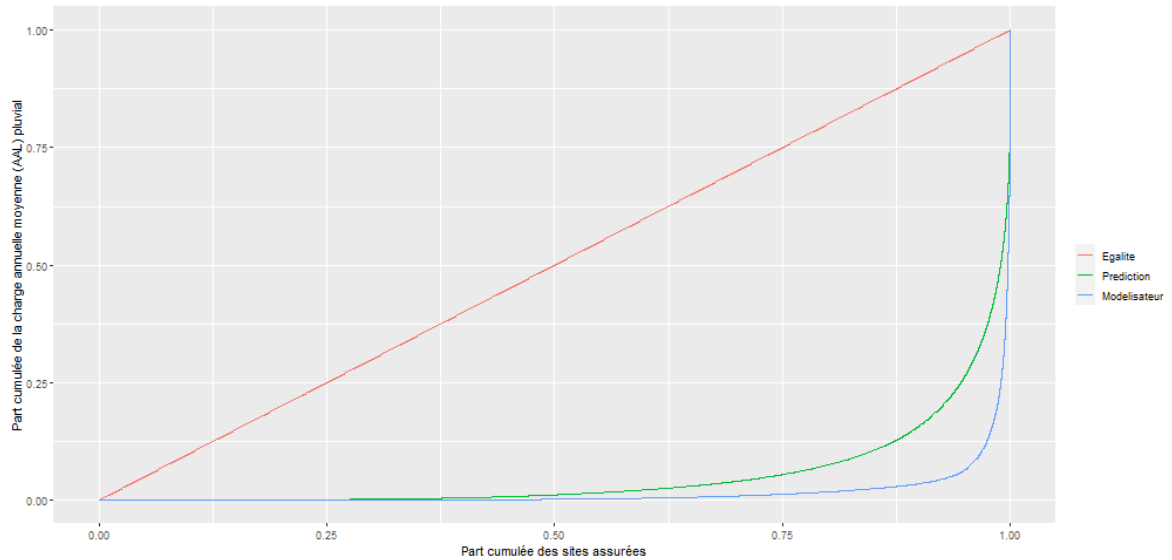

Figure 67 – Comparaison modélisateur vs prédiction des courbes de Lorenz appliquées à la distribution de l'AAL pluvial

b. Résultats sur le risque fluvial

Concernant le risque fluvial, le constat est globalement le même que pour le risque pluvial, il semble que la charge inondation globale est sur-modélisée et que davantage de départements ressortent en bleu foncé. Notons tout même une certaine cohérence et que la quasi-totalité des départements en bleu foncé côté modélisateur le sont également sur notre modèle. Il s'agit principalement des départements traversés par des fleuves à savoir la Loire pour les départements de la région Pays de Loire et Centre-Val de Loire, la Seine pour l'Île-de-France, la Garonne pour le Sud-Ouest et le Rhône pour le Sud-Est. À noter que côté modélisateur les départements du Sud-Est traversés par le Rhône semblent étonnamment peu touchés par le risque fluvial. Enfin du côté de la Bretagne, de la Normandie, et du nord de la France les deux modèles sont cohérents et modélisent une faible charge fluviale dans ces zones.

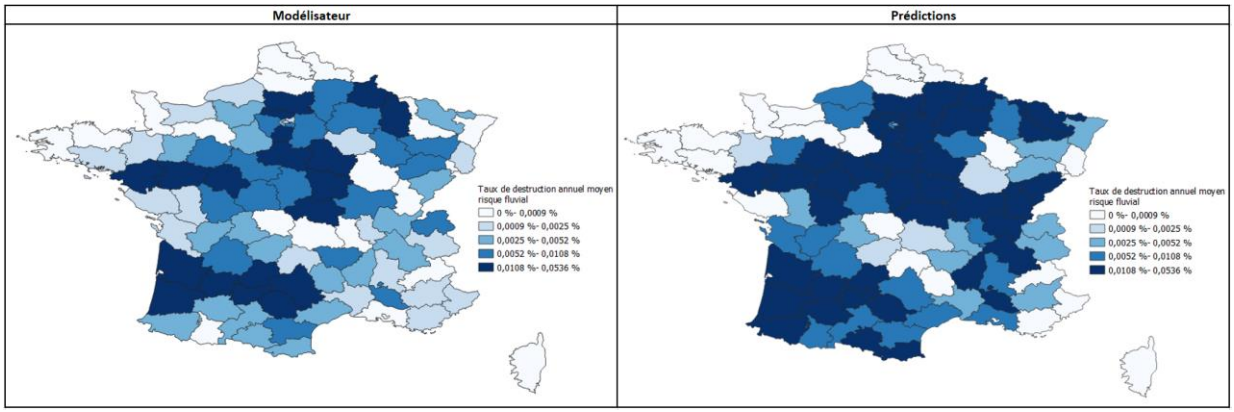

Figure 68 - Résultats du modèle fluvial par département sur la base de test

La modélisation par portefeuille est très satisfaisante et le classement des portefeuilles entre eux est pratiquement le même entre les deux modèles. On observe cependant que les taux de destruction sont surestimés par notre modèle, quels que soient les portefeuilles. Notamment sur le portefeuille I pour lequel la surestimation est près de trois fois supérieure à ce qu'indique notre modélisateur.

Portefeuille	Taux destruction fluvial modélisateur	Taux destruction fluvial prédicition	Classement modélisateur	Classement prediction
Branche A	0.024%	0.026%	2	2
Branche B	0.004%	0.007%	6	6
Branche C	0.006%	0.011%	5	4
Branche D	0.007%	0.010%	4	5
Branche F	0.002%	0.006%	7	7
Branche G	0.002%	0.003%	8	8
Branche H	0.009%	0.012%	3	3
Branche I	0.113%	0.329%	1	1
Total général	**0.006%**	**0.009%**		

Tableau 33 - Comparaison des résultats par portefeuille entre le modélisateur et les prédictions internes – risque fluvial

Également, tout comme pour le risque pluvial, notre modèle a tendance à davantage répartir la charge inondation entre les risques. On peut en effet le voir sur la courbe de Lorenz ci-dessous, la courbe verte qui représente notre prédiction est plus proche de la diagonale correspondant à une égalité parfaite dans la sinistralité inondation entre les risques. Enfin la partie classification de la modélisation fluviale est clairement visible sur la courbe de Lorenz étant donné que plus de 80% des risques ont une charge fluviale nulle, ce qui se traduit par une courbe horizontale sur l'axe des abscisses jusqu'à ce niveau.

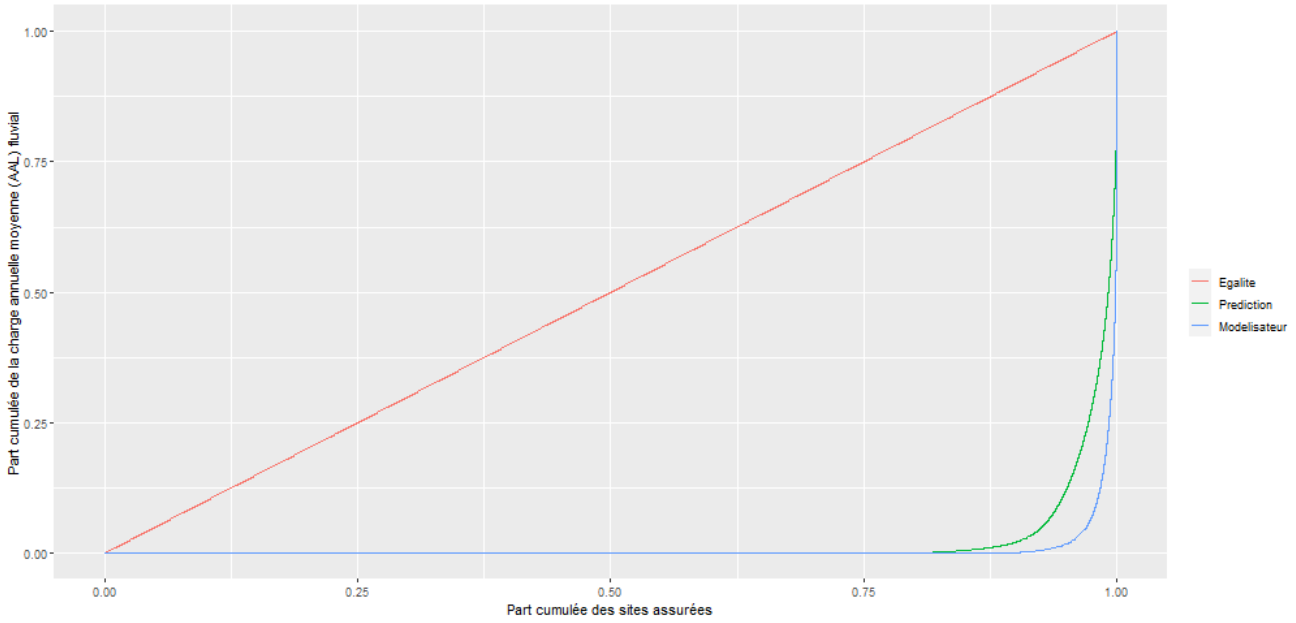

Figure 69 - Comparaison modélisateur vs prédiction des courbes de Lorenz appliquées à la distribution de l'AAL fluvial

c. Résultats globaux risque combiné

Le tableau ci-dessous présente les écarts de modélisations au global pour chaque région. On remarque dans l'ensemble qu'il est très difficile de retrouver exactement les mêmes résultats que le modélisateur et il apparait que l'on surestime de 50% la charge globale fluviale et de 12% la charge pluviale soit une surestimation globale de la charge inondation de +35%.

Sur le risque fluvial, cinq régions sont surmodélisées de plus de 100% à savoir la région Auvergne-Rhône-Alpes, Bourgogne-Franche-Comté, Île-de-France, Normandie et Provence-Alpes-Côte d'Azur. De plus, les deux modèles concluent sur une non-exposition au risque fluvial des sites assurés se trouvant en Corse dans notre base de test. À noter cependant que la Corse est peu représentée dans notre base de test, ce qui explique aussi la forte surestimation pour le risque pluvial. Sur ce risque, quatre autres régions apparaissent à plus de 100%, à savoir la région Centre-Val de Loire, Hauts-de-France, Île-de-France, Normandie et Pays de la Loire. Au total sur le risque combiné, seul trois régions s'avèrent surestimées de plus du double : la Corse, l'Île-de-France, et la Normandie.

Régions	Ecart de prédiction - risque fluvial	Ecart de prédiction - risque pluvial	Ecart de prédicition total - risque combiné
Auvergne-Rhône-Alpes	114%	61%	88%
Bourgogne-Franche-Comté	241%	2%	95%
Bretagne	-37%	-6%	-14%
Centre-Val de Loire	24%	148%	32%
Corse	Non exposé	420%	420%
Grand Est	98%	0%	39%
Hauts-de-France	41%	127%	78%
Île-de-France	105%	247%	116%
Normandie	115%	517%	209%
Nouvelle-Aquitaine	11%	-40%	-7%
Occitanie	41%	-9%	20%
Pays de la Loire	28%	148%	36%
Provence-Alpes-Côte d'Azur	119%	55%	82%
Total général	50%	12%	35%

Tableau 34 – Écarts de modélisations par région selon les trois visions pluviale/fluviale/combinée

2) Résultats sur l'ensemble du portefeuille

On peut désormais s'intéresser à l'analyse des résultats en appliquant le modèle sur l'ensemble des risques afin d'en déduire les régions et portefeuilles sur lesquels appliquer d'éventuelles décisions de souscriptions liées à la gestion du risque inondation. On commencera par étudier les résultats sur le risque pluvial, puis sur le risque fluvial et enfin sur le risque combiné qui désigne la totalité du risque inondation étudié dans ce mémoire (c'est-à-dire pluvial et fluvial regroupé).

a. Résultats sur le risque pluvial

La carte ci-dessous représente les taux de destruction moyens par an pour le risque pluvial par département. De manière générale, les résultats semblent plutôt cohérents et c'est principalement le sud et l'est de la France qui ressortent comme particulièrement à risques d'inondations, ce qui correspond également aux zones dans lesquelles on observait le plus de précipitations extrêmes dans les cartographies de la partie III.3.f.

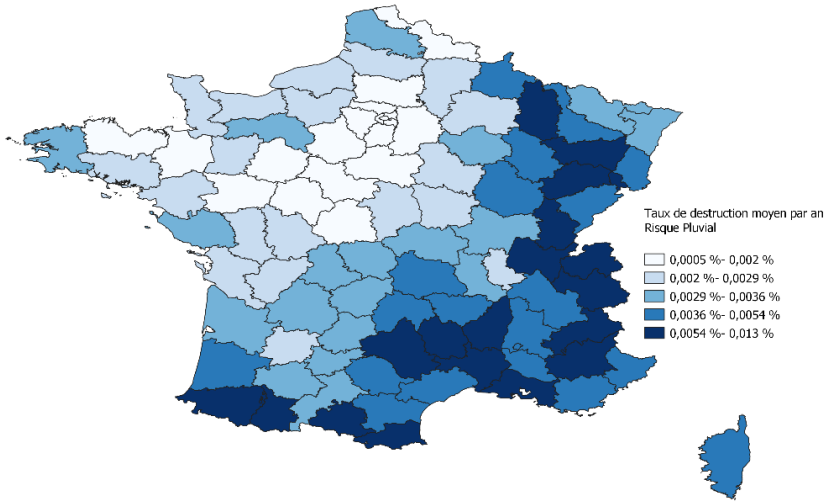

Figure 70 - Taux de destruction moyen par an pour le risque pluvial modélisé par département

Logiquement, ce sont les régions présentes dans ces territoires qui ressortent comme les plus exposées, on retrouve dans l'ordre, la Corse, la région Provence-Alpes-Côte d'Azur, l'Auvergne-Rhône-Alpes et l'Occitanie. Ces quatre régions se démarquent clairement, avec plus de 0.005% de taux de destruction annuel moyen pour le risque pluvial. À noter que la côte méditerranéenne est la région la plus touchée par les inondations selon l'historique de la CCR présenté dans la partie I, il est donc rassurant de voir ressortir cette région en deuxième position de notre modèle pluvial.

Régions	Taux de destruction moyen par an - risque pluvial	Rang
Corse	0.0053%	1
Provence-Alpes-Côte d'Azur	0.0052%	2
Auvergne-Rhône-Alpes	0.0051%	3
Occitanie	0.0050%	4
Bourgogne-Franche-Comté	0.0043%	5
Grand Est	0.0041%	6
Nouvelle-Aquitaine	0.0034%	7
Pays de la Loire	0.0024%	8
Normandie	0.0024%	9
Bretagne	0.0024%	10
Hauts-de-France	0.0022%	11
Centre-Val de Loire	0.0018%	12
Île-de-France	0.0011%	13
Total général	**0.0036%**	

Tableau 35 - Classement des régions les plus exposées au risque inondation pluvial

Enfin en étudiant les résultats par portefeuille, on remarque que la branche D ressort comme la plus touchée par le risque inondation avec 0.0052% de taux de destruction, loin devant la branche G, moins vulnérable à ce risque et qui présente un taux de destruction de 0.0018%

Portefeuilles	Taux de destruction moyen par an - risque pluvial	Rang
Branche D	0.0052%	1
Branche B	0.0048%	2
Branche F	0.0045%	3
Branche C	0.0041%	4
Branche I	0.0039%	5
Branche A	0.0038%	6
Branche H	0.0037%	7
Branche G	0.0018%	8
Total général	**0.0036%**	

Tableau 36 - Classement des portefeuilles les plus exposés au risque inondation pluvial

b. Résultats sur le risque fluvial

La prédiction du taux de destruction annuel moyen pour le risque fluvial passe par une première étape de classification afin d'identifier dans un premier temps les sites assurés exposés à ce risque. On représente sur les cartes ci-dessous des points selon quatre couleurs distinctes. En bleu les vrais positifs, c'est-à-dire les risques que notre modèle a prédits comme étant exposé au risque fluvial et qui le sont réellement (selon le modélisateur). En vert les vrais négatifs, c'est-à-dire les risques que notre modèle à prédit comme n'étant pas exposé au risque fluvial et qui en effet, ne le sont pas. En rouge et orange, l'ensemble des erreurs que l'on a faites, en rouge les faux positifs, notre modèle considère que le risque est exposé au risque fluvial alors qu'il ne l'est pas, en orange les faux négatifs, notre modèle considère que le site assuré n'est pas exposé au risque fluvial alors qu'il l'est.

Par exemple, les résultats à Nantes sont très satisfaisants et nous permettent d'approcher une exactitude de presque 100%, seuls de rares sites assurés ressortent en orange et rouge. Les résultats à Paris sont légèrement moins bons et il semble que l'on ait tendance à sous-estimer le nombre de risques exposés au risque fluvial étant donné que de nombreux sites apparaissent en orange (faux négatifs). Sur l'ensemble de la Bretagne, très peu de risques sont exposés au risque fluvial et notre modèle semble bien capter ce phénomène. Enfin à Bordeaux les résultats sont peu satisfaisants, selon le modélisateur aucun des sites assurés dans la ville n'est exposé au risque fluvial, une observation qui n'est pas captée par notre modèle, ce qui induit donc une grande quantité de faux positifs aux abords de la Garonne. Un résultat qui peut paraitre étonnant du côté de notre modélisateur, si en effet Bordeaux a historiquement bénéficié d'une coïncidence heureuse des marées qui ont permis d'atténuer les crues, le site gironde.gouv.fr pointe tout de même une dizaine d'inondations depuis le début du XXe siècle. Notons le plus récemment la crue du 31 janvier 2014 qui a notamment inondé le quartier de la Bastide dont certains risques apparaissent pourtant en faux positifs sur notre cartographie (quartier rive droite de la Garonne).

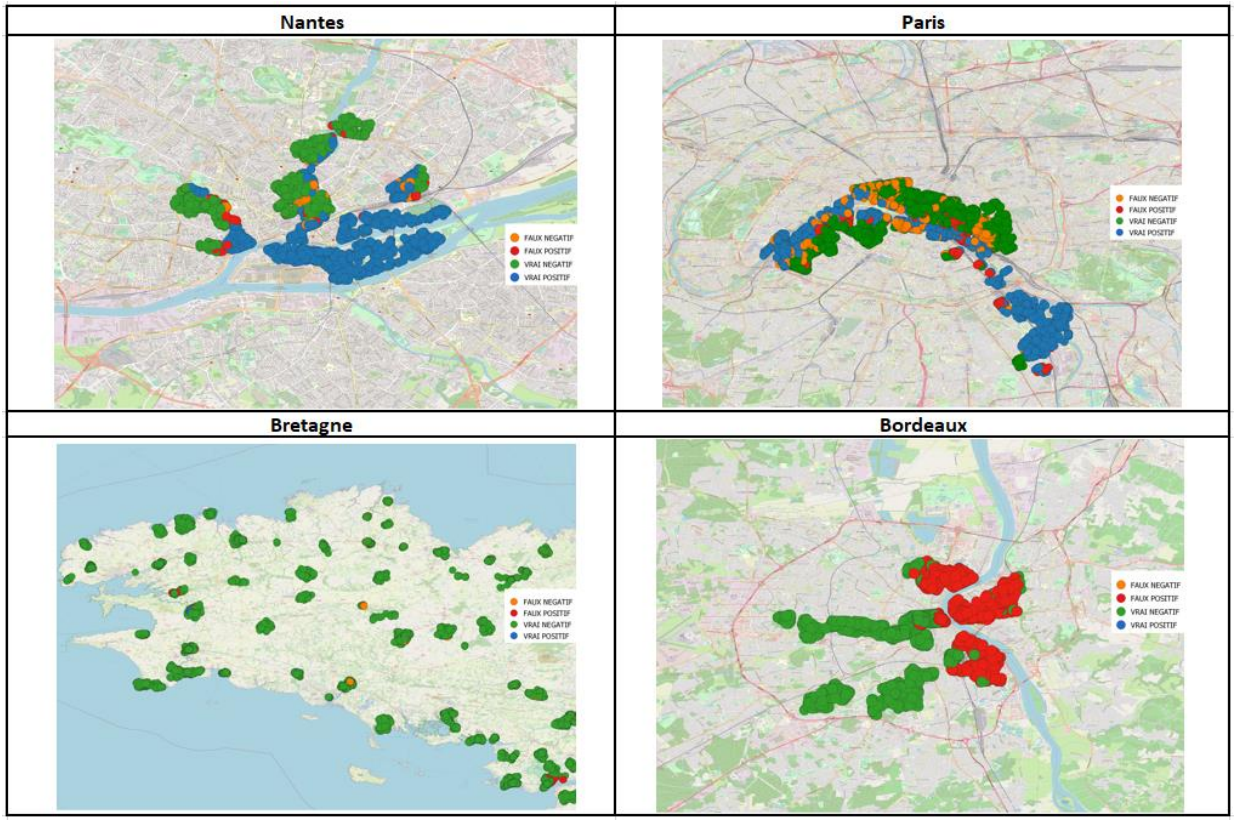

Figure 71 - Exemple de résultats de la classification sur l'exposition au risque fluvial pour différents territoires

Par la suite, on peut étudier les résultats finaux après application de la classification puis de la régression sur l'ensemble des risques. D'un point de vue de la répartition géographique, les résultats sont assez similaires avec ce qu'on a déjà pu observer sur la base de test. Dans l'ensemble, les départements touchés sont surtout ceux traversés par de grands fleuves.

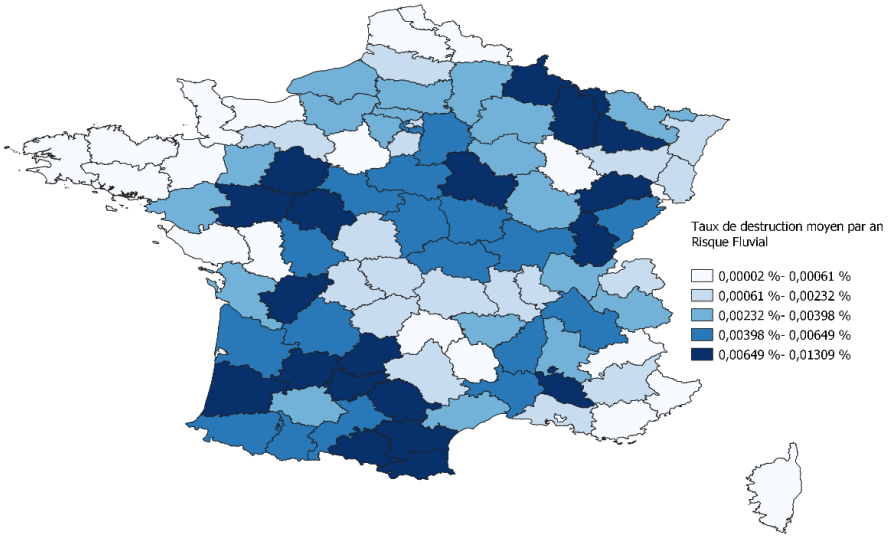

Figure 72 - Taux de destruction moyen par an pour le risque fluvial modélisé par département

Les résultats par région sont cohérents avec la carte par département. Le Sud-ouest, traversé notamment par la Garonne et la Dordogne, ressort comme particulièrement touché avec la région Occitanie et Nouvel-Aquitaine qui ont respectivement un taux de destruction moyen par an de 0.0062% et 0.0052%. On retrouve également les régions Centre-Val de Loire et Bourgogne-Franche-Comté respectivement traversées par la Loire et la Seine et qui atteignent des taux de destruction de 0.0054% et 0.0053%.

Régions	Taux de destruction moyen par an - risque fluvial	Rang
Occitanie	0.0062%	1
Bourgogne-Franche-Comté	0.0054%	2
Centre-Val de Loire	0.0053%	3
Nouvelle-Aquitaine	0.0052%	4
Pays de la Loire	0.0045%	5
Grand Est	0.0033%	6
Île-de-France	0.0032%	7
Auvergne-Rhône-Alpes	0.0028%	8
Normandie	0.0019%	9
Provence-Alpes-Côte d'Azur	0.0017%	10
Hauts-de-France	0.0011%	11
Bretagne	0.0002%	12
Corse	0.0000%	13
Total général	**0.0035%**	

Tableau 37 – Classement des régions les plus exposées au risque inondation fluvial

Enfin en ce qui concerne les résultats par portefeuille, on retrouve tout comme le risque pluvial, un portefeuille G moins vulnérable au risque inondation avec un taux de destruction près de trois fois inférieur au portefeuille H qui occupe la troisième place derrière les portefeuilles I et A.

Portefeuilles	Taux de destruction moyen par an - risque fluvial	Rang
Branche I	0.0491%	1
Branche A	0.0059%	2
Branche H	0.0041%	3
Branche D	0.0038%	4
Branche C	0.0034%	5
Branche F	0.0026%	6
Branche B	0.0020%	7
Branche G	0.0014%	8
Total général	**0.0035%**	

Tableau 38 - Classement des portefeuilles les plus exposés au risque inondation fluvial

c. Résultats sur le risque inondation combiné (fluvial et pluvial)

Maintenant que l'on s'est intéressé aux résultats indépendamment pour le risque pluvial et fluvial, dressons un bilan du risque inondation et des régions les plus touchées par ce risque selon nos modélisations. Tout d'abord, il apparait que les deux risques semblent peser un poids équivalent dans la charge inondation globale avec un taux de destruction générale sur l'ensemble du territoire de 0.0036% pour le risque pluvial, contre 0.0035% pour le risque fluvial.

Géographiquement, on remarque que la partie nord-ouest de la France semble moins touchée par les phénomènes inondations avec des taux de destruction moyens par an inférieur à 0.004% pour la plupart des départements. Pour le reste le sud et l'ouest de la France ont pratiquement tous des taux de destruction supérieurs à 0.006% annuels, avec une concentration de département ayant des taux supérieurs à 0.011% dans le sud-ouest de la France.

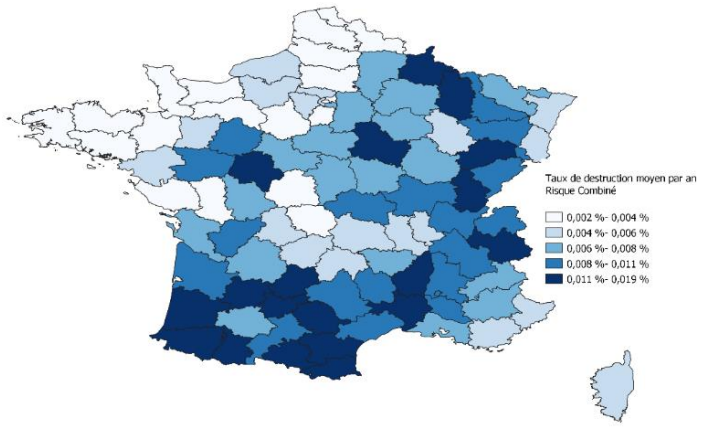

Figure 73 - Taux de destruction moyen par an pour le risque inondation (combiné) modélisé par département

Logiquement il ressort que les régions les plus exposées au risque inondation se trouvent au sud ou à l'ouest de la France. La région Occitanie occupe la première place avec 0.0112% de taux de destruction annuel moyen, contre 0.0097% pour la région Bourgogne-Franche-Comté et 0.0086% pour la Nouvelle-Aquitaine. À l'inverse, on retrouve en dernière position les régions présentent dans la partie nord-ouest de la France, avec notamment la Bretagne qui occupe la dernière position avec un taux de destruction moyen plus de quatre fois inférieur à la région Occitanie.

Régions	Taux de destruction moyen par an - risque combiné	Rang
Occitanie	0.0112%	1
Bourgogne-Franche-Comté	0.0097%	2
Nouvelle-Aquitaine	0.0086%	3
Auvergne-Rhône-Alpes	0.0079%	4
Grand Est	0.0074%	5
Centre-Val de Loire	0.0072%	6
Pays de la Loire	0.0070%	7
Provence-Alpes-Côte d'Azur	0.0070%	8
Corse	0.0053%	9
Île-de-France	0.0043%	10
Normandie	0.0043%	11
Hauts-de-France	0.0033%	12
Bretagne	0.0026%	13
Total général	**0.0071%**	

Tableau 39 - Classement des régions les plus exposées au risque inondation (combiné)

Au niveau des portefeuilles, les branches I, A et D apparaissent comme particulièrement vulnérables au risque inondation. À l'inverse, la branche G s'avère avoir des taux de destruction au moins deux fois inférieurs aux autres portefeuilles.

Portefeuilles	Taux de destruction moyen par an - risque combiné	Rang
Branche I	0.0529%	1
Branche A	0.0097%	2
Branche D	0.0089%	3
Branche H	0.0077%	4
Branche C	0.0075%	5
Branche F	0.0071%	6
Branche B	0.0068%	7
Branche G	0.0032%	8
Total général	**0.0071%**	

Tableau 40 - Classement des portefeuilles les plus exposés au risque inondation (combiné)

Enfin, la courbe de Lorenz nous indique que le risque pluvial touche davantage de sites assurés que le risque fluvial. En effet, les phénomènes de ruissellement notamment sont plus susceptibles de toucher n'importe quel site assuré, contrairement au risque fluvial qui touche uniquement des risques proches de grands fleuves. Au global, selon nos modélisations, il semble que 75% de la charge de sinistralité modélisée pour l'inondation est portée par uniquement 4% du portefeuille, constituant ainsi un résultat particulièrement intéressant pour la gestion du risque inondation.

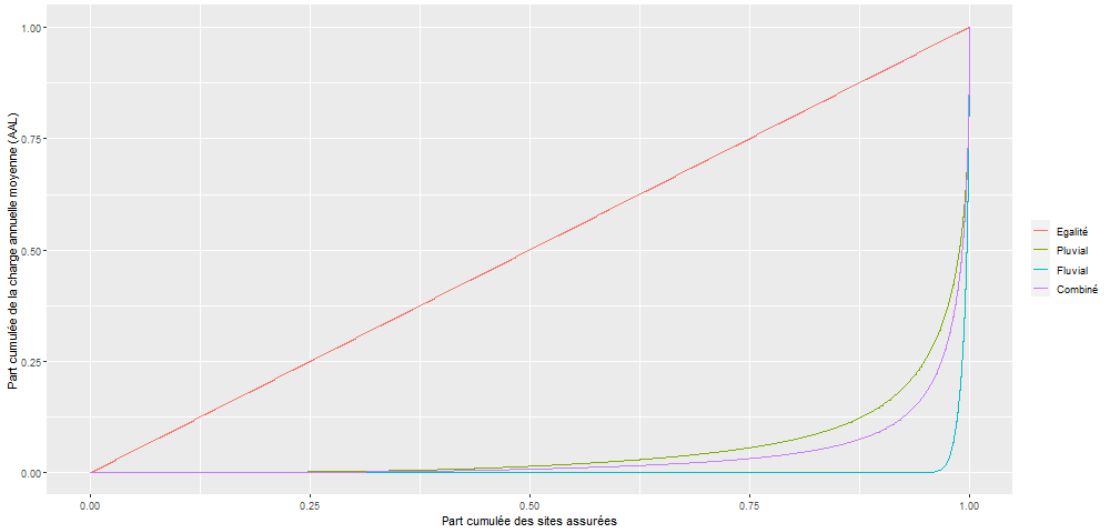

Figure 74 - Courbes de Lorenz appliquées à la distribution de l'AAL pluvial, fluvial et combiné

3) Évolution des précipitations avec le changement climatique

Maintenant que l'on dispose d'une vision plus claire des zones à risques d'inondations sur la période actuelle, l'objectif par la suite est d'étudier l'impact du changement climatique sur la charge annuelle moyenne. Pour cela, cette partie vise dans un premier temps à étudier l'évolution des précipitations avec le changement climatique.

Comme étudié dans la partie III, les données de projections climatiques CORDEX fournissent en supplément des simulations sur la période historique, des projections selon différents scénarios de changement climatique (RCP 2.6, RCP 4.5 et RCP 8.5) et selon différents horizons de temps (horizon proche H1 : 2021-2050, horizon moyen H2 : 2041-2070, horizon lointain H3 : 2071 – 2100).

De plus grâce au jeu de données DRIAS-2020, on dispose de douze modèles CORDEX différents, permettant pour chaque couple (RCP, horizons de temps) de calculer trois visions distinctes obtenues par un calcul de quantiles sur chaque point de grille. Une première vision qu'on qualifiera d'« optimiste » qui correspond au quantile à 5% de ces modèles, une vision en considérant la médiane des modèles et une vision « pessimiste » en sélectionnant le quantile à 95% des modèles. Cette méthode nous permettra de prendre en considération un nombre important de projections climatiques sans avoir à faire de choix arbitraire parmi ces modèles.

a. Résultats généraux sur l'évolution des précipitations

De manière générale, il semble que les précipitations journalières moyennes vont quasiment rester inchangées avec le changement climatique. On observe que ce changement sera encore moins important pour le scénario RCP 8.5 pour lequel la moyenne des précipitations devrait évoluer de seulement 1% d'ici 2100 contre 6% pour le scénario RCP 2.6.

MEDIANE DES MODELES	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100
	Auvergne-Rhône-Alpes	4%	7%	6%	3%	3%	4%	4%	3%	0%
	Bourgogne-Franche-Comté	6%	6%	6%	4%	5%	6%	6%	6%	7%
	Bretagne	2%	2%	3%	3%	0%	2%	2%	3%	4%
	Centre-Val de Loire	4%	4%	4%	3%	3%	4%	4%	4%	4%
	Corse	3%	6%	7%	-1%	0%	-1%	-4%	-2%	-11%
	Grand Est	5%	6%	6%	5%	5%	7%	6%	7%	10%
	Hauts-de-France	6%	6%	5%	4%	4%	5%	4%	7%	12%
	Île-de-France	6%	5%	5%	4%	4%	4%	4%	5%	7%
	Normandie	5%	3%	4%	3%	3%	4%	4%	5%	7%
	Nouvelle-Aquitaine	4%	5%	6%	3%	2%	3%	1%	1%	-3%
	Occitanie	2%	5%	7%	1%	0%	1%	0%	-1%	-8%
	Pays de la Loire	4%	4%	4%	4%	1%	3%	2%	4%	4%
	Provence-Alpes-Côte d'Azur	3%	8%	6%	1%	-1%	2%	0%	-1%	-8%
	Total général	4%	**5%**	**6%**	**3%**	**2%**	**4%**	**3%**	**3%**	**1%**

Tableau 41 - Projections d'évolution des précipitations journalières moyennes par région selon le jeu de données DRIAS-2020

Cependant, cela ne signifie pas que la typologie des précipitations ne va pas changer. En effet, on observe que le nombre de jours de pluie devrait diminuer avec le changement climatique ce qui pourrait notamment aggraver les phénomènes de sécheresse. Si l'indicateur est globalement stable pour le scénario RCP 2.6 avec une légère augmentation de 1% d'ici 2100, la baisse est de -3% à horizon 2100 pour le scénario RCP 4.5 et atteint les -9% pour le scénario 8.5.

MEDIANE DES MODELES	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100
	Auvergne-Rhône-Alpes	1%	1%	1%	-1%	-2%	-2%	-2%	-4%	-10%
	Bourgogne-Franche-Comté	1%	1%	1%	0%	-2%	-1%	-1%	-3%	-7%
	Bretagne	-1%	0%	0%	-2%	-5%	-4%	-3%	-4%	-10%
	Centre-Val de Loire	1%	1%	1%	0%	-3%	-3%	-2%	-3%	-9%
	Corse	-2%	1%	2%	-4%	-5%	-6%	-6%	-8%	-18%
	Grand Est	1%	1%	1%	0%	-1%	-1%	0%	-2%	-5%
	Hauts-de-France	1%	2%	1%	0%	-2%	-1%	-1%	-2%	-4%
	Île-de-France	1%	1%	1%	-1%	-3%	-2%	-2%	-3%	-7%
	Normandie	1%	0%	1%	-1%	-3%	-2%	-1%	-3%	-6%
	Nouvelle-Aquitaine	0%	1%	2%	-1%	-3%	-3%	-2%	-5%	-12%
	Occitanie	0%	1%	2%	-2%	-4%	-4%	-3%	-6%	-14%
	Pays de la Loire	0%	0%	1%	-1%	-4%	-4%	-2%	-4%	-10%
	Provence-Alpes-Côte d'Azur	0%	2%	1%	-3%	-4%	-4%	-5%	-6%	-15%
	Total général	0%	**1%**	**1%**	**-1%**	**-3%**	**-3%**	**-2%**	**-4%**	**-9%**

Tableau 42 - Projections d'évolution du nombre de jours de pluie par région selon le jeu de données DRIAS-2020

Géographiquement, la différence entre les territoires est relativement faible. On peut cependant noter un effet plus marqué dans le sud de la France, pour lequel le nombre de jours de pluie est déjà faible, avec une diminution de près de 15% en Provence-Alpes-Côte d'Azur et de 14% en Occitanie. À l'inverse, la diminution est plus faible dans le Nord-est pour lequel les diminutions sont de l'ordre de 5%.

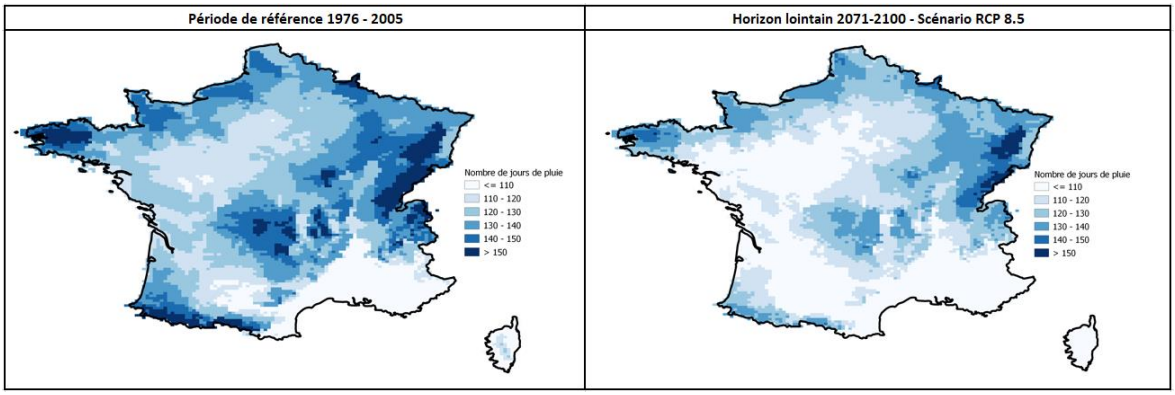

Figure 75 – Différence de nombre de jours de pluie en France entre la période de référence et la période 2071-2100 RCP8.5 selon la médiane des modèles DRIAS-2020

Enfin on représente ci-dessous l'évolution du nombre de jours de forte pluie (précipitations supérieures à 20mm au niveau du point de grille) pour la vision médiane. Contrairement au nombre de jours de pluie, les fortes pluies vont significativement augmenter avec une hausse de 15% d'ici 2100 pour le scénario RCP 2.6, 18% pour le scénario RCP 4.5 et enfin 23% pour le scénario 8.5.

MEDIANE DES MODELES	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100
	Auvergne-Rhône-Alpes	8%	13%	11%	6%	9%	12%	9%	11%	12%
	Bourgogne-Franche-Comté	17%	19%	18%	15%	21%	26%	22%	27%	40%
	Bretagne	16%	18%	16%	17%	19%	29%	19%	30%	48%
	Centre-Val de Loire	30%	33%	29%	28%	34%	45%	32%	42%	68%
	Corse	6%	9%	12%	1%	2%	2%	-2%	1%	-8%
	Grand Est	17%	19%	19%	18%	23%	35%	27%	35%	59%
	Hauts-de-France	24%	27%	26%	24%	31%	38%	26%	46%	81%
	Île-de-France	25%	30%	28%	26%	30%	42%	30%	50%	76%
	Normandie	25%	23%	26%	23%	31%	38%	24%	40%	64%
	Nouvelle-Aquitaine	13%	15%	16%	11%	14%	17%	9%	14%	20%
	Occitanie	6%	10%	14%	3%	4%	6%	4%	4%	-1%
	Pays de la Loire	22%	24%	21%	22%	26%	39%	21%	38%	56%
	Provence-Alpes-Côte d'Azur	5%	10%	9%	3%	2%	7%	3%	3%	-3%
	Total général	12%	**15%**	**15%**	**10%**	**13%**	**18%**	**12%**	**17%**	**23%**

Tableau 43 - Projections d'évolution du nombre de jours de forte pluie par région selon le jeu de données DRIAS-2020

La disparité entre les territoires est ici davantage marquée que pour le nombre de jours de pluie. Une disparité qui semble même s'accentuer au fur et à mesure du temps et plus le scénario considéré est pessimiste. En effet dans un premier temps, le sud de la France, déjà très touché par les précipitations intenses, semble épargné par cette augmentation avec des évolutions d'ici 2100 pour le scénario RCP 8.5 de -3% et -1% respectivement pour les régions Provence-Alpes-Côte d'Azur. Les régions moins touchées jusque-là tel que les Hauts-de-France où l'Île-de-France vont subir des augmentations beaucoup plus importantes, de l'ordre de +80% d'ici 2100.

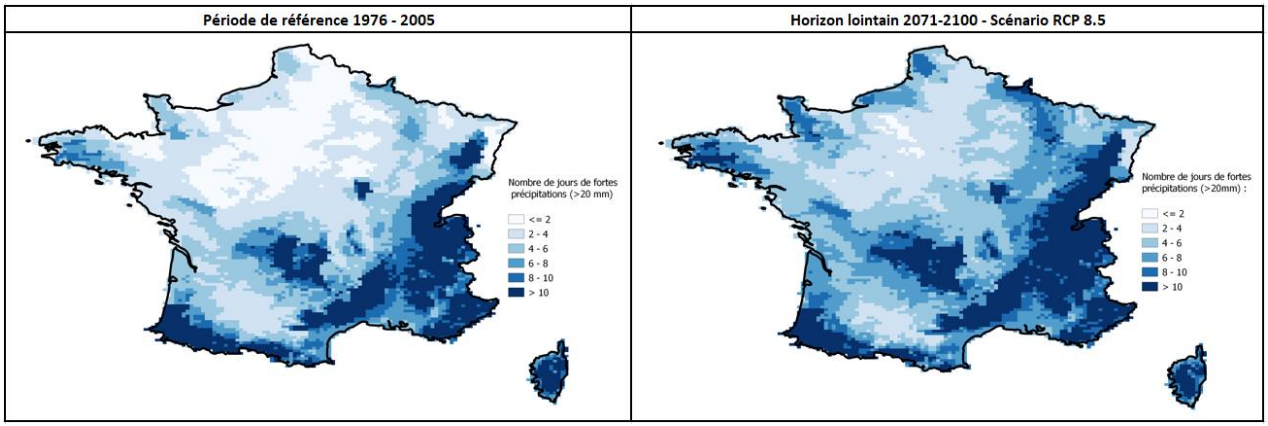

Figure 76 - Différence du nombre de jours de fortes précipitations en France entre la période de référence et la période 2071-2100 RCP8.5 selon la médiane des modèles DRIAS-2020

On remarque donc que si les précipitations moyennes vont certes rester quasi stables cela s'explique notamment par une diminution du nombre de jours de pluie qui va être compensée par des précipitations plus intenses durant ces mêmes jours de pluie. On trouvera plus de détails sur l'évolution des autres indicateurs de précipitations en annexe 8 et notamment l'évolution des précipitations moyennes les jours pluvieux qui permet de confirmer cette hypothèse.

b. Détails sur l'évolution des précipitations extrêmes

Pour notre étude, on va s'intéresser plus spécialement au quantile à 99% des précipitations, c'est-à-dire la variable « précipitations extrêmes » qui a été utilisée dans la forêt aléatoire pour la prédiction du risque pluvial. C'est donc cette variable qui sera modifiée pour réaliser les projections de sinistralité future.
On remarque tout d'abord, à part quelques exceptions dans certaines zones, que l'ensemble des cartes sont rouges. Cela signifie une augmentation des précipitations extrêmes quels que soient les horizons de temps et quels que soient les scénarios climatiques, y compris dans le scénario RCP2.6, supposé le plus optimiste, et pour lequel il faudrait rapidement mener des politiques très fortes permettant une réduction des gaz à effet de serre et un réchauffement inférieur à 2°C d'ici 2100.

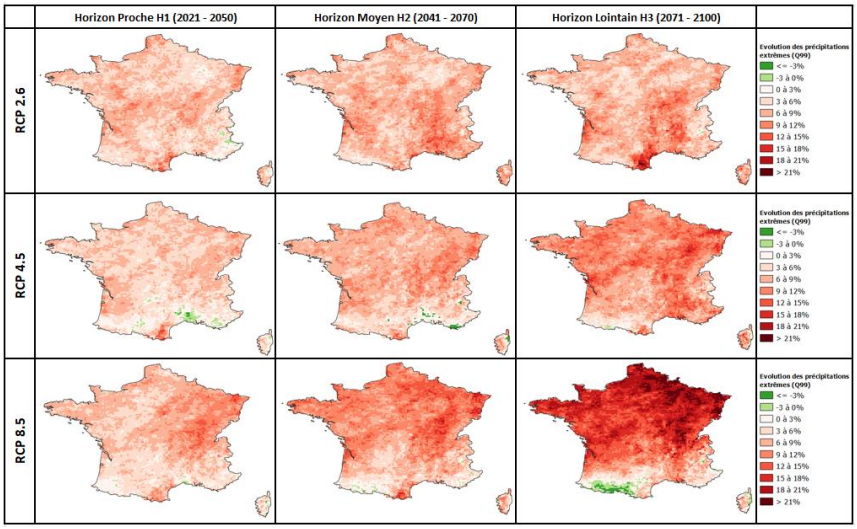

Figure 77 - Évolution du quantile à 99% des précipitations (i.e variable "précipitations extrêmes") selon la vision médiane des modèles climatiques du jeu DRIAS-2020

Dans le cas où aucune politique forte n'est menée, ce qui correspond au chemin qui est suivi pour le moment et donc au scénario RCP 8.5, on remarque qu'il s'agit principalement du nord de la France qui est touché, partie jusque-là moins exposée aux précipitations extrêmes. Notons cependant l'est de la France autour des Vosges et du Jura, déjà très exposé à ce risque selon les modélisations précédentes et qui devrait voir augmenter encore davantage ses précipitations extrêmes, avec une hausse d'environ 20%. De manière générale, on remarque sur les cartographies ci-dessous que les précipitations extrêmes devraient rester tout de même largement concentrées dans le sud et l'est de la France.

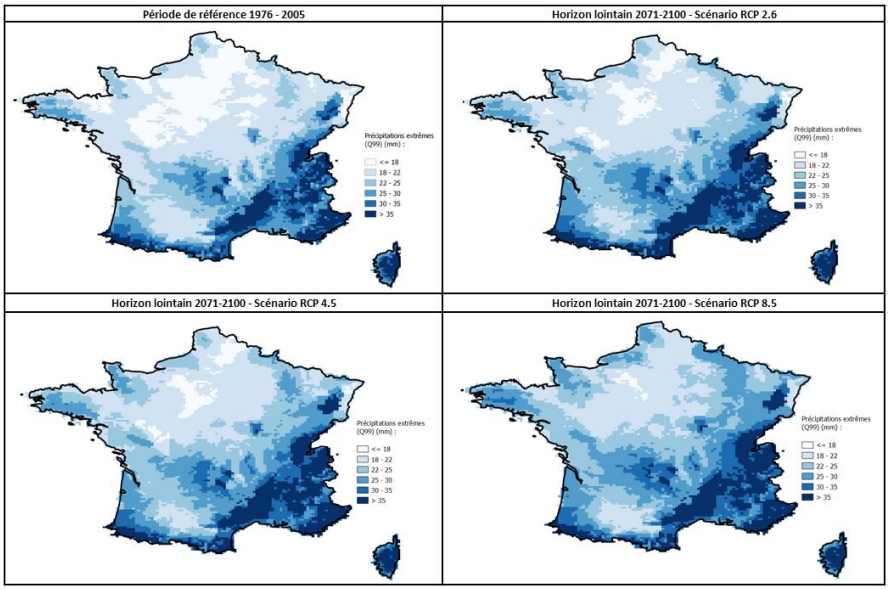

Figure 78 - Différence du quantile à 99% des précipitations en France entre la période de référence et la période 2071-2100 selon les trois scénarios RCP (vue médiane des modèles DRIAS-2020)

Les cartes ci-dessus ont été réalisées uniquement à partir de la vision médiane des modèles, on peut se représenter l'ensemble des visions sur le tableau ci-dessous, découpé selon les régions françaises.

On remarque une nouvelle fois que l'ensemble des pourcentages sur le total général sont positifs, y compris pour la vision « optimiste » réalisée à partir du quantile à 5% des modèles climatiques. Pour le scénario RCP 2.6, il apparait que les augmentations sont très faibles voir nulles au fil des horizons de temps signe d'une certaine stagnation d'ici la fin du siècle. Pour les deux autres scénarios, l'évolution des précipitations ne semble pas stagner et augmente en continu jusqu'en 2100, à l'exception de la vision « optimiste » pour laquelle les différences entre les scénarios sont moins claires. On remarque également que le scénario RCP 8.5 présente une plus forte variabilité entre les régions. Par exemple, pour la vision médiane sur la période 2071 à 2100 les évolutions varient de 4 à 20% pour le scénario RCP 8.5, et de seulement 6 à 9% pour le scénario 2.6. Enfin à l'horizon 2050 la différence entre les scénarios est très faible avec une augmentation pour la vision médiane selon les scénarios RCP2.6, 4.5 et 8.5 respectivement de +6%, +5% et +6%.

	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100
VUE "PESSIMISTE" (QUANTILE 95%)	Auvergne-Rhône-Alpes	10%	12%	12%	10%	14%	17%	12%	14%	22%
	Bourgogne-Franche-Comté	10%	10%	12%	10%	15%	19%	14%	18%	29%
	Bretagne	9%	9%	9%	10%	11%	17%	10%	19%	32%
	Centre-Val de Loire	9%	11%	11%	9%	13%	15%	11%	15%	26%
	Corse	14%	15%	13%	9%	14%	15%	14%	15%	12%
	Grand Est	9%	10%	11%	11%	13%	19%	14%	20%	33%
	Hauts-de-France	9%	11%	11%	10%	12%	14%	11%	18%	34%
	Île-de-France	8%	11%	12%	9%	11%	15%	11%	17%	27%
	Normandie	9%	10%	10%	11%	12%	16%	12%	20%	32%
	Nouvelle-Aquitaine	10%	11%	11%	9%	11%	13%	9%	12%	18%
	Occitanie	11%	11%	14%	9%	11%	15%	10%	12%	14%
	Pays de la Loire	9%	10%	10%	8%	11%	15%	9%	17%	27%
	Provence-Alpes-Côte d'Azur	7%	12%	11%	8%	9%	15%	9%	12%	14%
	Total général	**10%**	**11%**	**12%**	**9%**	**12%**	**16%**	**11%**	**15%**	**23%**
	Régions	**RCP2.6**			**RCP4.5**			**RCP8.5**		
		2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100
MEDIANE DES MODELES	Auvergne-Rhône-Alpes	6%	8%	8%	5%	7%	9%	7%	9%	12%
	Bourgogne-Franche-Comté	7%	8%	8%	6%	9%	11%	9%	11%	17%
	Bretagne	6%	6%	6%	6%	7%	10%	7%	11%	17%
	Centre-Val de Loire	6%	7%	6%	5%	7%	10%	6%	9%	15%
	Corse	5%	8%	8%	4%	6%	7%	3%	6%	4%
	Grand Est	5%	6%	7%	7%	8%	12%	9%	12%	20%
	Hauts-de-France	6%	7%	7%	6%	8%	10%	7%	11%	20%
	Île-de-France	6%	7%	7%	6%	7%	10%	7%	11%	17%
	Normandie	6%	6%	7%	6%	8%	11%	7%	11%	17%
	Nouvelle-Aquitaine	6%	7%	8%	5%	7%	8%	5%	7%	11%
	Occitanie	6%	7%	9%	3%	5%	7%	5%	6%	4%
	Pays de la Loire	6%	7%	6%	6%	7%	11%	6%	10%	15%
	Provence-Alpes-Côte d'Azur	4%	8%	7%	3%	4%	9%	4%	5%	5%
	Total général	**6%**	**7%**	**7%**	**5%**	**7%**	**9%**	**6%**	**9%**	**12%**
	Régions	**RCP2.6**			**RCP4.5**			**RCP8.5**		
		2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100
VUE "OPTIMISTE" (QUANTILE 5%)	Auvergne-Rhône-Alpes	2%	5%	4%	-1%	2%	3%	2%	3%	3%
	Bourgogne-Franche-Comté	3%	5%	4%	2%	5%	5%	5%	7%	10%
	Bretagne	4%	3%	3%	3%	4%	6%	3%	6%	8%
	Centre-Val de Loire	2%	2%	3%	3%	3%	5%	2%	5%	9%
	Corse	-1%	2%	5%	-1%	1%	0%	-5%	-2%	-10%
	Grand Est	2%	3%	3%	3%	5%	7%	6%	7%	12%
	Hauts-de-France	2%	2%	1%	2%	4%	5%	2%	6%	12%
	Île-de-France	2%	3%	3%	4%	2%	4%	3%	6%	10%
	Normandie	4%	2%	4%	2%	4%	5%	3%	6%	10%
	Nouvelle-Aquitaine	3%	4%	4%	3%	3%	4%	1%	2%	5%
	Occitanie	1%	2%	5%	-3%	-2%	0%	-3%	-3%	-7%
	Pays de la Loire	3%	3%	3%	2%	4%	6%	2%	4%	7%
	Provence-Alpes-Côte d'Azur	-1%	3%	3%	-2%	-1%	3%	0%	-2%	-9%
	Total général	**2%**	**3%**	**4%**	**1%**	**2%**	**4%**	**1%**	**3%**	**3%**

Tableau 44 - Projections d'évolution des précipitations extrêmes par région selon le jeu de données DRIAS-2020

4) Impact du changement climatique sur la sinistralité inondation

a. Impact sur la sinistralité pluvial

On peut désormais relancer nos modèles à partir des variables précipitations projetées selon les différents scénarios de changement climatique, afin d'en déduire un impact sur la sinistralité. Notons que l'ensemble des taux d'évolution de sinistralité présentés dans cette partie ne tiennent pas compte de la future potentielle inflation et supposent une stabilité de répartition du portefeuille sur le territoire à travers le temps. Ces évolutions dépendent donc uniquement de la variation de l'aléa climatique.

Les résultats d'évolution de la sinistralité pour le risque d'inondation pluvial ont été réalisés après avoir fait varier la variable précipitations extrêmes selon les projections présentées précédemment. On retrouve dans un premier temps des cartographies similaires à ce qui a été présenté dans la partie V.3, à l'exception que cette fois on présente une évolution de la charge moyenne de sinistralité et non une évolution de l'aléa climatique. Les résultats restent cependant très proches que ceux de la variation de l'aléa climatique pour lesquels les plus fortes augmentations de précipitations extrêmes étaient observées dans le nord de la France. On observe notamment pour le scénario RCP 8.5 projeté sur la période 2071-2100 qu'une majorité des départements du nord de la France ressortent avec des évolutions de sinistralité quasiment toutes supérieures à +20%. Enfin, on remarque tout de même que, quels que soient les scénarios, certains départements ressortent avec une légère diminution de leur sinistralité.

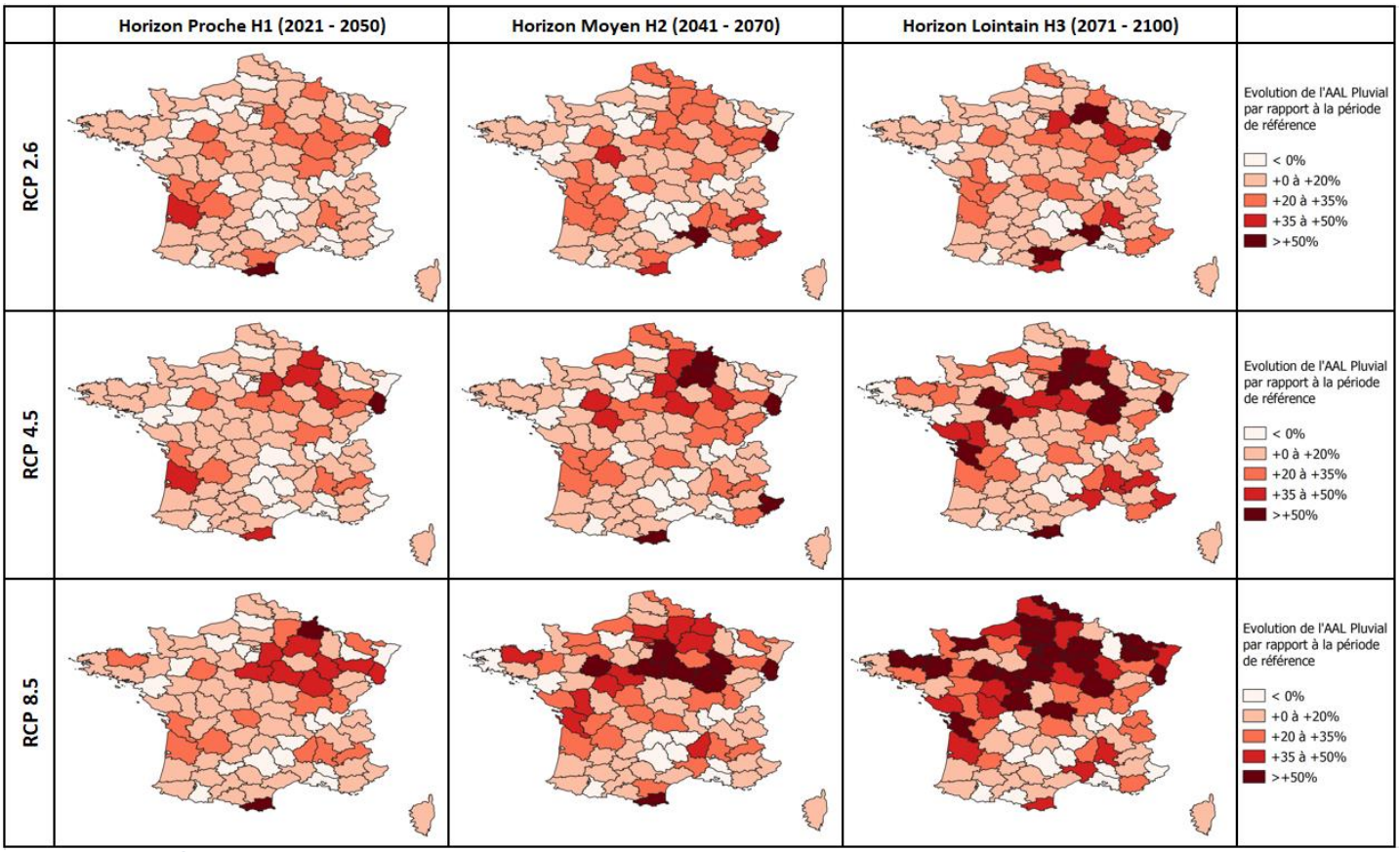

Figure 79 - Évolution de la charge annuelle moyenne de sinistralité pour le risque pluvial selon les différents scénarios et différents horizons de temps (vue médiane des modèles)

Le tableau ci-dessous nous permet d'avoir une vue globale de l'impact du changement climatique par région, par horizon de temps et selon les trois visions : « optimiste », médiane et « pessimiste » :

VUE "PESSIMISTE" (QUANTILE 95%)	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100
	Auvergne-Rhône-Alpes	13%	17%	13%	13%	18%	21%	17%	18%	37%
	Bourgogne-Franche-Comté	28%	30%	27%	25%	37%	28%	30%	29%	35%
	Bretagne	17%	12%	12%	14%	14%	23%	16%	25%	37%
	Centre-Val de Loire	23%	36%	33%	31%	35%	38%	35%	37%	85%
	Corse	20%	30%	31%	12%	25%	20%	35%	19%	25%
	Grand Est	24%	29%	32%	24%	31%	39%	40%	43%	49%
	Hauts-de-France	18%	22%	26%	26%	26%	34%	23%	38%	84%
	Île-de-France	22%	48%	40%	19%	29%	62%	27%	66%	73%
	Normandie	14%	17%	14%	21%	27%	35%	23%	58%	43%
	Nouvelle-Aquitaine	17%	20%	17%	19%	16%	25%	14%	23%	19%
	Occitanie	18%	18%	24%	11%	15%	22%	21%	23%	24%
	Pays de la Loire	11%	15%	8%	6%	13%	42%	5%	41%	40%
	Provence-Alpes-Côte d'Azur	18%	-1%	-3%	24%	19%	4%	9%	9%	1%
	Total général	**18%**	**19%**	**18%**	**18%**	**21%**	**26%**	**21%**	**27%**	**34%**

MEDIANE DES MODELES	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100
	Auvergne-Rhône-Alpes	4%	9%	11%	8%	9%	13%	7%	12%	17%
	Bourgogne-Franche-Comté	19%	20%	22%	20%	26%	27%	25%	30%	28%
	Bretagne	8%	8%	8%	9%	8%	11%	9%	15%	27%
	Centre-Val de Loire	11%	19%	13%	13%	20%	29%	17%	30%	39%
	Corse	17%	6%	9%	9%	11%	12%	2%	14%	7%
	Grand Est	12%	17%	23%	18%	24%	25%	23%	28%	40%
	Hauts-de-France	12%	20%	19%	11%	22%	20%	13%	25%	55%
	Île-de-France	12%	13%	15%	11%	12%	31%	10%	38%	75%
	Normandie	6%	8%	6%	6%	11%	14%	7%	19%	39%
	Nouvelle-Aquitaine	18%	18%	14%	16%	17%	19%	14%	17%	17%
	Occitanie	7%	17%	15%	5%	3%	11%	6%	11%	9%
	Pays de la Loire	-2%	0%	-2%	0%	3%	17%	0%	13%	30%
	Provence-Alpes-Côte d'Azur	-1%	5%	4%	1%	25%	1%	4%	2%	1%
	Total général	**9%**	**13%**	**13%**	**10%**	**14%**	**16%**	**11%**	**16%**	**23%**

VUE "OPTIMISTE" (QUANTILE 5%)	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100
	Auvergne-Rhône-Alpes	-2%	1%	5%	-4%	-1%	5%	-1%	2%	2%
	Bourgogne-Franche-Comté	13%	15%	10%	5%	20%	14%	13%	18%	24%
	Bretagne	11%	7%	6%	6%	9%	10%	6%	9%	8%
	Centre-Val de Loire	1%	5%	6%	6%	8%	10%	4%	15%	21%
	Corse	-3%	-2%	8%	0%	-2%	3%	-7%	6%	-10%
	Grand Est	1%	6%	5%	7%	9%	15%	13%	20%	29%
	Hauts-de-France	5%	1%	3%	4%	5%	13%	4%	17%	25%
	Île-de-France	0%	4%	3%	5%	5%	6%	4%	10%	26%
	Normandie	4%	5%	9%	2%	5%	7%	10%	8%	16%
	Nouvelle-Aquitaine	11%	17%	13%	12%	12%	14%	0%	4%	10%
	Occitanie	-3%	2%	8%	-5%	-3%	0%	-4%	-2%	-3%
	Pays de la Loire	-3%	-1%	0%	-2%	-5%	-2%	-4%	-2%	2%
	Provence-Alpes-Côte d'Azur	28%	2%	-1%	13%	18%	3%	17%	14%	-15%
	Total général	**5%**	**5%**	**6%**	**3%**	**6%**	**7%**	**4%**	**8%**	**8%**

Tableau 45 – Impact du changement climatique sur la charge annuelle moyenne inondation pour le risque pluvial par région

On remarque notamment :

- La charge de sinistralité globale par an devrait à priori augmenter dans les années à venir, y compris lorsque l'on considère le scénario 2.6 avec la vision la plus optimiste (quantile à 5% des modèles).

- À horizon 2050, la différence entre les scénarios est très faible entre les scénarios RCP2.6, 4.5 et 8.5 avec une augmentation respective de +9%, +10% et +11%. Ainsi, même en agissant maintenant les effets sur la sinistralité d'ici 2050 ne devrait pas présenter une grande différence.

- Selon la vision médiane, sur la période 2071-2100, la charge de sinistralité inondation pour le risque pluvial devrait augmenter en moyenne de 13% pour le scénario RCP 2.6, de 16% pour le scénario RCP 4.5 et de 23% pour le scénario RCP 8.5. En étudiant la vision conservatrice qui considère le quantile à 95% des modèles, ces chiffres passent respectivement à 18, 26 et 34%

- En considérant la vision prudente pour le scénario RCP 2.6, la charge de sinistralité semble plus ou moins stagner entre la période 2041-2070 et la période 2071-2100 avec une charge totale passant de 19 à 18%.

Ce n'est pas le cas pour les scénarios RCP 4.5 et 8.5, qui voient leur charge de sinistralité augmenter respectivement de 21 à 26% et de 27 à 34%.

- Il semble que les régions du nord de la France sont celles qui seront les plus impactées par le changement climatique sur le risque inondation pluvial, avec une augmentation de la sinistralité annuelle qui devrait augmenter de 75% pour la région Île-de-France et de 55% pour les Hauts-de-France, selon les projections d'ici 2100 pour le scénario RCP8.5. La région Grand Est, déjà particulièrement vulnérable aux inondations voit sa sinistralité augmenter de 40% d'ici 2100 selon la vision médiane des modèles.

- Les régions Provence-Alpes-Côte d'Azur et Occitanie déjà très touchées par les inondations pluviales devraient subir une plus faible hausse de sinistralité, avec une évolution ne dépassant pas les +10% selon la vision médiane d'ici 2100 pour le scénario 8.5.

Malgré une forte augmentation de la sinistralité inondation dans le nord de la France, on remarque sur le tableau ci-dessous que le classement des régions les plus exposées au risque d'inondation pluvial demeurera quasiment le même dans le futur, avec cependant une variabilité plus faible entre les régions. Pour le scénario RCP 8.5 , la région Grand Est devrait passer en deuxième place du classement, juste derrière la région Auvergne-Rhône-Alpes avec une augmentation du taux de destruction moyen par an de 0.0041% à 0.0058% en 2100. On note également la stagnation de la région Provence-Alpes-Côte d'Azur qui devrait conserver une exposition au risque inondation similaire à l'horizon 2100, ce qui la fait passer de la deuxième place au climat présent à la sixième place au climat futur.

	Taux de destruction moyen par an - risque pluvial			
Territoires Caisses Régionales	Période de référence	RCP2.6 : 2071 -2100	RCP4.5 : 2071 -2100	RCP8.5 : 2071 -2100
Corse	0.0053%	0.0057%	0.0059%	0.0056%
Provence-Alpes-Côte d'Azur	0.0052%	0.0055%	0.0053%	0.0053%
Auvergne-Rhône-Alpes	0.0051%	0.0057%	0.0058%	0.0060%
Occitanie	0.0050%	0.0057%	0.0055%	0.0054%
Bourgogne-Franche-Comté	0.0043%	0.0053%	0.0055%	0.0055%
Grand Est	0.0041%	0.0051%	0.0052%	0.0058%
Nouvelle-Aquitaine	0.0034%	0.0038%	0.0040%	0.0040%
Pays de la Loire	0.0024%	0.0024%	0.0028%	0.0032%
Normandie	0.0024%	0.0026%	0.0028%	0.0034%
Bretagne	0.0024%	0.0026%	0.0026%	0.0030%
Hauts-de-France	0.0022%	0.0027%	0.0027%	0.0035%
Centre-Val de Loire	0.0018%	0.0021%	0.0024%	0.0026%
Île-de-France	0.0011%	0.0012%	0.0014%	0.0019%
Total général	**0.0036%**	**0.0041%**	**0.0042%**	**0.0044%**

Tableau 46 - Taux de destruction moyen par an pour le risque pluvial - période historique et projections selon vision médiane des modèles

Pour finir, il ne semble pas y avoir de différence majeure entre les différents portefeuilles, ces risques devraient subir une augmentation d'environ 13% pour le scénario RCP 2.6, 16% pour le scénario RCP 4.5 et 23% pour le scénario RCP 8.5. Seul le portefeuille I semble se démarquer des autres avec des augmentations significativement inférieures à la moyenne globale.

Portefeuille	RCP2.6 : 2071-2100	RCP4.5 : 2071-2100	RCP8.5 : 2071-2100
Branche A	13%	17%	22%
Branche B	8%	13%	21%
Branche C	13%	14%	22%
Branche D	11%	14%	21%
Branche F	11%	9%	23%
Branche G	19%	26%	26%
Branche H	13%	16%	22%
Branche I	2%	4%	12%
Total général	**13%**	**16%**	**23%**

Tableau 47 - Évolution de la sinistralité par portefeuille pour différents scénarios de changement climatique (vue médiane des modèles)

b. Impact sur la sinistralité globale

Comme évoqué précédemment, l'impact du changement climatique sur le risque fluvial n'a pas pu être effectué étant donné qu'aucune variable sur les précipitations n'a été utilisée pour modéliser ce risque. L'absence de corrélations significatives entre la charge annuelle fluviale donnée par le modélisateur et les indicateurs de précipitations nous amène à considérer que l'impact sera moindre que pour les phénomènes liés au risque « pluvial ». On considérera donc que l'impact du changement climatique sur ce risque est nul pour le calcul de l'évolution globale de la charge inondation. Cependant, cette hypothèse est loin d'être vérifiée et il semble très peu probable que l'impact du changement climatique sur ce risque soit totalement nul.

Tout d'abord, l'absence de corrélations avec les indicateurs de précipitations est assez étonnante et nous amène à nous interroger sur la pertinence des sorties obtenues auprès de notre modélisateur étant donné que les crues de grands fleuves sont principalement dues à des précipitations répétées et prolongées sur plusieurs jours. Une évolution de ces phénomènes dans le futur aura donc à priori un impact sur la sinistralité pour le risque fluvial.

Cependant dans un deuxième temps, l'annexe 8 nous indique que l'évolution de la variable « Nombre de jours de pluie consécutifs » avec le changement climatique devrait quasiment rester stable, ce qui irait dans le sens d'une faible évolution de la sinistralité des inondations causées par les crues de grands fleuves (risque fluvial).

VUE "PESSIMISTE" (QUANTILE 95%)	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100	2021 - 2050	2041 - 2070	2071 - 2100
	Auvergne-Rhône-Alpes	8%	10%	7%	8%	11%	12%	10%	11%	24%
	Bourgogne-Franche-Comté	13%	14%	13%	12%	17%	13%	14%	13%	16%
	Bretagne	15%	11%	11%	12%	13%	20%	14%	22%	34%
	Centre-Val de Loire	6%	9%	9%	8%	9%	10%	9%	10%	22%
	Corse	20%	30%	30%	12%	25%	20%	34%	18%	25%
	Grand Est	13%	16%	18%	14%	17%	21%	22%	24%	27%
	Hauts-de-France	12%	15%	23%	17%	18%	27%	16%	31%	60%
	Île-de-France	4%	17%	10%	4%	11%	19%	6%	20%	18%
	Normandie	7%	9%	8%	12%	14%	19%	12%	34%	25%
	Nouvelle-Aquitaine	7%	8%	7%	8%	6%	10%	6%	9%	8%
	Occitanie	9%	8%	11%	5%	7%	10%	10%	10%	11%
	Pays de la Loire	4%	6%	3%	2%	5%	15%	2%	16%	15%
	Provence-Alpes-Côte d'Azur	12%	0%	-1%	17%	14%	3%	6%	7%	1%
	Total général	**9%**	**10%**	**9%**	**9%**	**11%**	**13%**	**11%**	**14%**	**18%**
MEDIANE DES MODELES	**Régions**	**RCP2.6**			**RCP4.5**			**RCP8.5**		
		2021 - 2050	**2041 - 2070**	**2071 - 2100**	**2021 - 2050**	**2041 - 2070**	**2071 - 2100**	**2021 - 2050**	**2041 - 2070**	**2071 - 2100**
	Auvergne-Rhône-Alpes	2%	5%	6%	5%	6%	8%	5%	7%	11%
	Bourgogne-Franche-Comté	8%	9%	10%	9%	12%	12%	12%	15%	13%
	Bretagne	7%	6%	7%	8%	7%	9%	8%	13%	26%
	Centre-Val de Loire	3%	5%	3%	3%	5%	8%	5%	8%	10%
	Corse	16%	6%	8%	8%	10%	11%	2%	13%	10%
	Grand Est	7%	9%	13%	10%	14%	14%	12%	16%	22%
	Hauts-de-France	8%	13%	13%	8%	14%	14%	8%	20%	40%
	Île-de-France	2%	4%	3%	2%	3%	8%	2%	13%	22%
	Normandie	3%	4%	3%	3%	5%	7%	3%	10%	21%
	Nouvelle-Aquitaine	8%	8%	6%	7%	8%	8%	6%	7%	7%
	Occitanie	4%	8%	7%	2%	2%	5%	3%	6%	5%
	Pays de la Loire	0%	0%	-1%	0%	1%	6%	0%	5%	11%
	Provence-Alpes-Côte d'Azur	-1%	4%	3%	1%	17%	2%	3%	1%	0%
	Total général	**4%**	**7%**	**6%**	**5%**	**7%**	**8%**	**6%**	**9%**	**12%**
VUE "OPTIMISTE" (QUANTILE 5%)	**Régions**	**RCP2.6**			**RCP4.5**			**RCP8.5**		
		2021 - 2050	**2041 - 2070**	**2071 - 2100**	**2021 - 2050**	**2041 - 2070**	**2071 - 2100**	**2021 - 2050**	**2041 - 2070**	**2071 - 2100**
	Auvergne-Rhône-Alpes	-1%	1%	3%	-4%	-1%	3%	-1%	1%	1%
	Bourgogne-Franche-Comté	6%	7%	4%	3%	9%	7%	6%	8%	11%
	Bretagne	9%	6%	5%	5%	8%	10%	6%	8%	7%
	Centre-Val de Loire	0%	1%	2%	2%	2%	3%	1%	4%	6%
	Corse	-2%	-2%	8%	-1%	-2%	3%	-7%	5%	-11%
	Grand Est	0%	3%	3%	4%	5%	8%	7%	11%	16%
	Hauts-de-France	4%	1%	2%	3%	4%	9%	3%	11%	21%
	Île-de-France	-1%	1%	1%	1%	1%	1%	0%	2%	7%
	Normandie	1%	3%	5%	1%	2%	3%	9%	4%	8%
	Nouvelle-Aquitaine	5%	7%	6%	5%	5%	6%	0%	2%	4%
	Occitanie	-1%	1%	4%	-2%	-1%	0%	-2%	-1%	-2%
	Pays de la Loire	0%	0%	1%	0%	-1%	0%	-1%	0%	1%
	Provence-Alpes-Côte d'Azur	19%	1%	-1%	10%	12%	1%	11%	9%	-10%
	Total général	**3%**	**3%**	**3%**	**1%**	**3%**	**4%**	**2%**	**4%**	**4%**

Tableau 48 - Impact du changement climatique sur la charge annuelle moyenne inondation (risque combiné) par région (en considérant l'évolution du risque fluvial nulle dans le temps)

Le tableau ci-dessus présente l'évolution de la charge inondation globale en considérant d'un côté l'augmentation de la charge pluviale tel que présentée dans la partie qui précède et en considérant de l'autre côté une charge fluviale qui reste stable dans le temps et selon les différents scénarios.

Les résultats par territoires diffèrent légèrement par rapport à l'analyse faite pour le seul risque pluvial étant donné que la part que représente le risque fluvial sur le total de la sinistralité inondation n'est pas la même en fonction des régions. Ainsi la région Île-de-France, dans laquelle le risque fluvial constitue une part importante de la charge totale et qui présentait une des plus fortes évolutions sur le seul risque pluvial, avec +73% d'ici 2100 selon la vision prudente des modèles pour le scénario RCP8.5, voit son évolution réduite à +18% sur le risque inondation total. Dans le même temps la région Grand Est, dont la charge annuelle pluviale représente une large part de la sinistralité inondation, et pour laquelle son évolution était de 49% sur la sinistralité pluviale, voit son évolution réduite à +27% pour la vision prudente.

Au global sur l'ensemble du territoire, étant donné que la charge pluviale et fluviale représente à chacune d'elle environ 50% de la charge globale inondation, les évolutions présentées ci-dessus sont donc à peu de choses près deux fois inférieures aux évolutions sur le seul risque pluvial. Par exemple en considérant la vision prudente des modélisations (quantile à 95%), on déduit une augmentation de la charge inondation globale d'ici 2100 de près de +9% pour le scénario RCP2.6, +13% pour le scénario 4.5 et de +18% pour le scénario 8.5. Pour la vision médiane des modèles ces augmentations passent à respectivement +6, +8 et +12% à horizon 2100.

5) Comparaison avec les autres études effectuées sur le changement climatique

Ces dernières années trois études majeures ont été menées pour tenter d'évaluer l'évolution de la sinistralité future des catastrophes naturelles due au changement climatique :

- Une étude de la CCR en 2015, sur l'évolution de la sinistralité à horizon 2050 pour les périls sécheresse, inondation et submersion marine. Cette étude utilise des projections climatiques qui reposent sur le scénario RCP4.5. Le modèle climatique utilisé est le modèle ARPEGE de Météo France.

- Une nouvelle étude similaire de la CCR en 2018 se basant cette fois sur le scénario RCP8.5 et dont les résultats ont servi de base à la partie catastrophe naturelle du pilote climatique de l'ACPR. Pour cette raison, on se basera directement sur cette étude plutôt que sur les conclusions de l'ACPR.

- Une étude de la FFA (Fédération Française de l'assurance) en 2021, sur l'évolution de la sinistralité à horizon 2050 pour les périls sécheresse, inondation, submersion marine et tempête. Cette étude utilise des projections climatiques qui reposent sur le scénario RCP8.5. Deux modèles climatiques sont utilisés, un provenant de l'Institut Pierre Simon Laplace (IPSL) et un second de l'Institut Max-Planck de météorologie (MPI).

- Le mémoire ci-présent sur l'évolution de la sinistralité de Groupama à l'horizon 2050, 2070 et 2100 pour le péril inondation pour les scénarios RCP2.6, 4.5 et 8.5. Pour rappel, douze modèles climatiques régionaux ont été utilisés dans notre étude. Les modèles climatiques globaux utilisés provenant de six laboratoires distincts, à savoir, le CNRM (Centre Nationale de Recherche Météorologique) qui est une unité de recherche de Météo France et du CNRS, l'IPSL, le centre Hadley qui fait partie du service national britannique de météorologie (*Met Office*), le *European EC-Earth consortium*, l'Institut Max-Planck de météorologie (MPI) situé en Allemagne et enfin le *Norwegian Climate Consortium* (NCC).

Les méthodologies utilisées dans ces études sont assez différentes et il est donc intéressant de comparer les résultats obtenus pour chacune. Tout d'abord, l'étude de la FFA est effectuée à partir d'une analyse de corrélations entre les indicateurs climatiques et les indemnisations versées par le passé par les assureurs. Il ressort de cela que l'indicateur climatique le plus corrélé à la sinistralité est le quantile à 90% du cumul des précipitations. Le quantile à 99% des précipitations extrêmes a également été retenu par la FFA pour l'analyse des évènements extrêmes, c'est le même indicateur qui est utilisé dans notre modèle sur le risque pluvial.

Les deux études de la CCR utilisent un modèle catastrophe comme celui utilisé par notre modélisateur et dont le fonctionnement général est détaillé en partie II. Le module aléa est alimenté par 400 années de données climatiques simulées au climat actuel (autour de 2000) et l'évolution est déduite en relançant le modèle pour 400 années au climat futur (autour de 2050) selon les scénarios RCP4.5 pour l'étude de 2015 et RCP8.5 pour l'étude de 2018.

Finalement, on a utilisé dans ce mémoire les résultats de modélisations d'un modèle catastrophe sur 250 000 risques répartis aléatoirement sur le territoire Français. Afin d'obtenir des résultats sur l'ensemble du portefeuille, on a entraîné divers modèles de *machine learning* sur ces risques grâce à un enrichissement préalable de notre base avec des variables permettant d'expliquer la vulnérabilité au risque inondation, telles que le quantile à 99% du cumul des précipitations quotidiennes. Enfin, le modèle ainsi créé a été relancé en faisant varier cette dernière variable selon les projections de changement climatique pour chaque horizon de temps, scénarios et selon la méthode d'agrégation des modèles de climat (quantile à 5%, médiane ou quantile à 95% de ces modèles). Pour la comparaison avec les autres études, on utilisera la vision conservatrice de notre étude en sélectionnant le quantile à 95% des douze modèles.

Scénario changement climatique	Source étude	Année Publication	Horizon futur modélisé	Évolution[1] de la sinistralité globale inondation[2] à l'horizon 2XXX	Détails par types d'inondations	Évolution de la sinistralité à l'horizon 2XXX
RCP 8.5	Groupama[3]	2022	2100	+18%	Pluvial[4]	+34%
					Fluvial[4]	NA[5]
	Groupama	2022	2050	+11%	Pluvial	+21%
					Fluvial	NA
	FFA	2021	2050	+11%	Pas de distinction entre types d'inondations	
	CCR	2018	2050	+38%	Ruissellement	+50%
					Débordement	+24%
RCP 4.5	Groupama	2022	2100	+13%	Pluvial	+26%
					Fluvial	NA
	Groupama	2022	2050	+9%	Pluvial	+18%
					Fluvial	NA
	CCR	2015	2050	+20%	Pas de distinction entre types d'inondations	
RCP 2.6	Groupama	2022	2100	+9%	Pluvial	+18%
					Fluvial	NA
	Groupama	2022	2050	+9%	Pluvial	+18%
					Fluvial	NA

(1) Évolution dû à la variation de l'aléa climatique (sans prise en compte de l'inflation ni de l'évolution future de la répartition des risques sur le territoire)
(2) Inondations hors submersions marines
(3) Par prudence on considérera l'évolution globale donnée par le quantile à 95% de l'étude Groupama
(4) Risque Pluvial = Ruissellement et Débordement de ruisseaux/rivières - Risque Fluvial = Débordement de fleuves
(5) Pas de résultats concluants sur l'évolution du risque fluvial : on considère une évolution nulle pour le calcul de l'évolution globale inondation

Tableau 49 - Impact du changement climatique sur le risque inondation selon différentes études, scénarios de changement climatique et horizons de temps

Le tableau ci-dessus présente les différents résultats pour chacune des études. Il semble que l'étude de la CCR est la plus pessimiste avec une augmentation de la sinistralité de +38% pour le scénario RCP 8.5 et de +20% pour le scénario RCP 4.5. Cette différence de près de 18 points de pourcentage entre les deux scénarios nous amène à nous questionner. En effet, l'une des remarques que l'on a pu faire dans l'étude de l'évolution de l'aléa climatique à horizon 2050 était la faible différence entre les scénarios. Pour rappel, on avait déduit une augmentation du nombre de jours de forte pluie de +10% à horizon 2050 selon le scénario RCP4.5 contre +12% pour le scénario RCP8.5. De même, l'évolution des précipitations extrêmes passe de +5% à +6% entre les deux scénarios. Dans notre étude,

cela se traduit par une faible différence de sinistralité à l'horizon 2050 entre les scénarios, ce qui diffère donc du résultat obtenu par la CCR.

L'étude de la FFA est beaucoup plus optimiste et conclue sur une augmentation de la sinistralité inondation à horizon 2050 à +11%.

Enfin, notre étude établit une évolution identique aux résultats obtenus par la FFA, mais reste assez loin des conclusions de la CCR, même en considérant la vision conservatrice de notre étude. On conclut en effet à une augmentation de +11% d'ici 2050 pour le RCP 8.5 contre +38% selon la CCR, ainsi qu'une évolution de +9% sur le scénario RCP 4.5 contre +20% selon la CCR. Cette potentielle sous-estimation de l'évolution peut s'expliquer par les résultats peu concluants sur l'impact des précipitations dans les modélisations du risque fluvial, qui nous ont amenés à considérer un impact nul du changement climatique sur ce risque. On remarque la faible différence de résultats entre les différents scénarios à horizon 2050, avec une augmentation identique pour le scénario RCP 2.6 et RCP 4.5 à +9%, et seulement deux points de pourcentage en plus pour le 8.5 à +11%. Pour finir, il est intéressant de noter l'évolution entre l'horizon 2050 et l'horizon 2100 fournie par notre étude. Le scénario RCP 2.6 ne présente aucune évolution entre 2050 et 2100 et la charge semble stagner à +9% par rapport aux années 2000. Le scénario RCP 4.5 voit sa charge passer de +9% à 13%. Finalement, le scénario RCP 8.5 possède l'écart le plus important avec une évolution qui passe de +11% à +18%.

Conclusion

En résumé, le but de ce mémoire était d'évaluer la sensibilité du territoire au risque inondation en développant une vision suffisamment robuste permettant d'identifier les risques les plus vulnérables, que ce soit aussi bien au climat actuel, qu'au climat futur en prenant en compte le changement climatique. Nos modèles étant basés sur des sorties de modèles catastrophes naturelles déjà existants, il a été nécessaire de réaliser les prédictions selon la même typologie d'inondations, en considérant d'un côté les inondations pluviales concernant les phénomènes de ruissellement et de crues de rivières et ruisseaux et de l'autre les inondations fluviales regroupant les crues de fleuves et grandes rivières.

Afin de gommer l'effet de la somme assurée sur la charge annuelle moyenne, la prédiction s'est faite sur le taux de destruction annuel moyen pour le risque pluvial et fluvial. Le risque fluvial ne touchant qu'une faible part du portefeuille a nécessité d'établir une première étape de classification par forêt aléatoire afin de distinguer les risques exposés à ce phénomène de ceux qui ne le sont pas. Pour l'étape de régression, un modèle linéaire généralisé gamma a été utilisé pour la partie fluviale et une forêt aléatoire pour la partie pluviale. Ces deux modèles ont nécessité l'utilisation de variables disponibles en interne telles que la branche assurée, le taux d'engagements contenu et pertes d'exploitation, ainsi que le nombre d'étages pour le portefeuille immeuble. De plus, un grand travail d'enrichissement de la base de risque a été nécessaire. Pour le risque pluvial, ce sont les variables concernant l'indice d'humidité topographique, la distance au cours d'eau, le quantile à 99% des précipitations et le coefficient de Manning qui ont été retenues comme variables explicatives. Pour le risque fluvial, les variables distance à la rivière large, distance à la rivière, différence d'altitude à la rivière large et imperméabilité du sol ont été retenues.

L'intégration des modèles climatiques CORDEX provenant de l'ensemble DRIAS-2020 a permis d'obtenir une vision historique des précipitations extrêmes ainsi qu'une vision future selon différents horizons de temps. Les résultats de modèles obtenus grâce aux simulations historiques de précipitations ont permis de déduire une forte vulnérabilité au risque inondation des régions du sud et sud-est de la France.

Les modélisations CORDEX au climat futur font apparaitre que quel que soit le scénario de changement climatique et y compris en considérant les modèles les plus optimistes, la hausse des précipitations extrêmes semble inévitable à horizon 2050. De plus, on constate que la mise en place d'actions politiques fortes menées dès à présent (scénario RCP 2.6) aurait peu d'impacts sur l'évolution des précipitations extrêmes à l'horizon 2050. En effet, en considérant les modèles les plus pessimistes, la hausse des précipitations extrêmes d'ici 2050 devraient être de l'ordre de +10% pour le scénario RCP 2.6 contre +11% pour le scénario d'inaction politique RCP 8.5.

En appliquant ces projections à notre modèle inondation les constatations sont les mêmes, avec une évolution de la sinistralité inondation globale d'ici 2050 de +9% pour le scénario RCP 2.6 contre +11% pour le scénario RCP 8.5. La différence entre ces scénarios est plus flagrante à l'horizon 2100, vu que l'on constate une stagnation de la sinistralité pour le scénario RCP 2.6 qui reste à +9% tandis que la hausse passe à +18% pour le scénario RCP 8.5.

Le modèle ainsi développé présente de nombreux avantages en comparaison des modèles catastrophes naturelles que l'on peut acheter auprès des modélisateurs externes. Tout d'abord, celui-ci nous permet une plus grande autonomie étant donné qu'il ne nous est pas nécessaire de passer par un intermédiaire à chaque fois que l'on souhaite obtenir des résultats. De plus, son fonctionnement est beaucoup plus transparent, ce qui permet ainsi de directement analyser les variables importantes dans la détermination de la charge inondation. Enfin, sa facilité et rapidité d'exécution est indispensable pour l'intégrer dans un processus interne de gestion du risque.

Cependant, l'ensemble de ces résultats est à nuancer, et ce pour plusieurs raisons. Tout d'abord, le modèle développé tente de reproduire les résultats d'un modèle catastrophe naturelle dont la construction est beaucoup plus complexe qu'un simple GLM ou modèle de *machine learning*. Un tel modèle est donc difficilement reproductible avec ce type de méthodes et la nécessité de l'utilisation des modèles catastrophes naturelles en réassurance n'est pas compromise par les résultats de ce mémoire. En effet, la correspondance entre notre modèle et le modèle CAT est loin d'être parfaite, avec un coefficient de Spearman entre les deux sorties de 33% sur le risque pluvial et de 50% sur le risque fluvial, ce qui reste au-dessus d'un tirage aléatoire et ses 0% de corrélations, mais encore loin d'un classement parfait entre les deux modèles qui devrait atteindre les 100%.

De plus, le fait que les variables précipitations ne ressortent pas dans les modèles du risque fluvial est étonnant étant donné que les crues de fleuves se font souvent sur plusieurs jours à la suite d'intempéries prolongées. Ce résultat peut être expliqué par des failles dans le modèle catastrophe naturelle sur lequel on se base qui n'utiliserait pas les précipitations dans leur modélisation du risque fluvial ou encore plus simplement par des indicateurs de précipitations utilisés qui ne seraient pas assez pertinents et qui n'arrivent pas à capter les territoires touchés par ces phénomènes de pluies prolongées. Ce résultat implique que l'impact du changement climatique modélisé pour le risque fluvial est nul. Or la charge fluviale représente 50% de la charge inondation totale ce qui pourrait provoquer une sous-estimation de l'impact du changement climatique sur le risque inondation dans notre étude. Cependant ce résultat peut être simplement la conséquence que le risque fluvial s'avère moins sensible au changement climatique, ce qui est également une observation faite dans l'étude de la CCR, qui conclut sur une augmentation deux fois moins importante du risque de débordement face au risque de ruissellement.

Il serait pour finir intéressant d'affiner davantage nos modèles en enrichissant la base de risques avec d'autres variables explicatives telles que les débits des fleuves proches de nos sites assurés, en utilisant les données de stations hydrométriques. De plus, de nombreuses communes sont dotées de digues de protections contre les inondations ce qui impacte fortement la vulnérabilité inondation des risques avoisinants. Enfin, il serait utile de développer notre propre modèle catastrophe naturelle afin d'étudier les différences de résultats obtenus entre les deux méthodologies.

Bibliographie

Richard Laganier (2006) Territoires, inondation et figures du risque : La prévention au prisme de l'évaluation, L'Harmattan

Caisse Centrale de Réassurance (2021) *Les Catastrophes naturelles en France Bilan 1982-2020*

GIEC, 2013: Résumé à l'intention des décideurs, Changements climatiques 2013: Les éléments scientifiques. Contribution du Groupe de travail I au cinquième Rapport d'évaluation du Groupe d'experts intergouvernemental sur l'évolution du climat [sous la direction de Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S. K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex et P.M. Midgley]. Cambridge University Press, Cambridge, Royaume-Uni et New York (État de New York), États-Unis d'Amérique.

Ministère de l'Environnement et de la Lutte contre les changements climatiques du Canada : Aspects hydrauliques pour l'analyse et la conception des réseaux de drainage; 41p.

Papaioannou, George & Efstratiadis, Andreas & Vasiliades, Lampros & Loukas, Athanasios & Papalexiou, Simon Michael & Koukouvinos, A. & Tsoukalas, Ioannis & Kossieris, Panagiotis. (2018). An Operational Method for Flood Directive Implementation in Ungauged Urban Areas

Jenson, S. K. et J. O. Domingue. 1988. "Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis." Photogrammetric Engineering and Remote Sensing 54 (11): 1593-1600.

Soubeyroux, Bernus, Corre, Drouin, Dubuisson, Etchevers, Gouget, Josse, Kerdoncuff, Samacoits, Tocquer (2020). Les nouvelles projections climatiques de référence DRIAS 2020 pour la métropole

Déqué M. et al., 2007 : Frequency of precipitation and temperature extremes over France in an anthropogenic scenario : model results and statistical correction according to observed values. Global and Planetary Change. 57 : 16-26.

RAKOTOMALALA R., 2015 : Pratique de la Régression Logistique, Université Lyon 2, support de cours

BESSE P. [2003] Pratique de la modélisation statistique, Université de Toulouse, support de cours

Ferrari SLP, Cribari-Neto F (2004). "Bêta Regression for Modelling Rates and Proportions." Journal of Applied Statistics, 31(7), 799–815.

Cribari-Neto, F., & Zeileis, A. (2010). Bêta Regression in R. Journal of Statistical Software, 34(2), 1–24

Bertrand, Frédéric; Meyer, Nicolas; Beau-Faller, Michèle; El Bayed, Karim; Namer, Izzie-Jacques; Maumy-Bertrand, Myriam. Régression Bêta PLS. Journal de la société française de statistique, Tome 154 (2013) no. 3, pp. 143-159.

Simas AB, Barreto-Souza W, Rocha AV (2010). "Improved Estimators for a General Class of Bêta Regression Models." Computational Statistics & Data Analysis, 54(2), 348–366.

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Chapman & Hall, (1984). "Classification and Regression Trees."

Friedman, J. (1999). Greedy Function Approximation – À Gradient Boosting Machine

Caisse Centrale de Réassurance (2015). Modélisation de l'impact du changement climatique sur les dommages assurés dans le cadre du régime Catastrophes Naturelles ; 35p.

Caisse Centrale de Réassurance (2018). Conséquences du changement climatique sur le coût des catastrophes naturelles en France à horizon 2050; 32p.

Fédération Française de l'Assurance (2021). Impact du changement climatique sur l'assurance à l'horizon 2050 ; 32p.

Annexes

Annexe 1 – Géocodage et coordonnées géographiques

L'étape d'enrichissement des variables passe par une première étape primordiale de géocodage qui vise à affecter des coordonnées géographiques (latitude et longitude) à une adresse postale. La majorité des données externes qui seront utilisées dans ce mémoire sont des données géographiques qui sont exprimées sous forme de latitude/longitude. Pour faire le rapprochement entre un site assuré et ces données, il est donc indispensable d'affecter une longitude et une latitude à chacun de nos risques afin de pouvoir les positionner sur le globe terrestre. La longitude et la latitude correspondent aux angles mesurés depuis le centre de la Terre vers un point de surface :

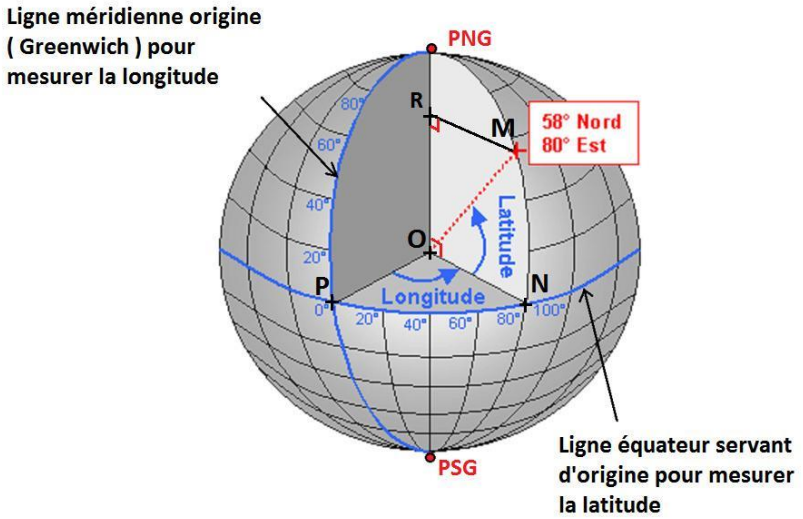

Figure 80 – Les coordonnées géographiques (latitude/longitude) sur la Terre

On représente le globe terrestre par une sphère de centre O et l'on considère un repère formé horizontalement par l'équateur et verticalement par le méridien de Greenwich. On peut ainsi déterminer pour tout point M de la surface de la Terre sa latitude et longitude associée. On appelle l'ensemble des lignes « horizontales » des parallèles, qui ont toutes la même latitude, et l'ensemble des lignes « verticales » des méridiens, qui ont donc tous la même longitude.

Dans notre étude, le processus de géocodage a été effectué à l'aide du géocodeur mis à disposition gratuitement par le gouvernement à travers l'API Adresse. Cette API permet de faire le lien entre l'adresse renseignée dans nos bases et une des adresses présentes dans la base d'adresse nationale (BAN). Cette base de données vise à recenser l'intégralité des adresses du territoire français ainsi que leurs coordonnées géographiques. Elle comprend actuellement près de 25 millions d'adresses et 250 000 lieux-dits.

Annexe 2 – Calcul distance entre deux points géolocalisés

Une fois le point du cours d'eau le plus proche du cours calculé par le logiciel de cartographie (QGIS), on souhaite calculer la distance entre ce point et le site assuré considéré. Pour ce faire, on utilise la formule de haversine qui permet de calculer la distance minimale entre deux points sur la surface d'un cercle (distance du grand cercle). Considérons deux points A et B sur la sphère dont on veut connaître la distance minimale qui les sépare et qui ont pour latitude ϕ_A, ϕ_B et pour longitude λ_A, λ_B.
On introduit tout d'abord les grandeurs a et c définies par :

$$a = \sin^2\left(\frac{\phi_B - \phi_A}{2}\right) + \cos(\phi_A).\cos(\phi_B).\sin^2\left(\frac{\lambda_B - \lambda_A}{2}\right)$$

$$c = 2.\arctan\left(\frac{\sqrt{a}}{\sqrt{1-a}}\right)$$

À noter que les latitudes et longitudes doivent être exprimées en radians et non en degrés (multiplication par $\frac{\pi}{180}$ pour passer de degrés à radians).

Finalement, la distance entre les points A et B est :

$$d = R * c$$
$$avec\ R = 6\,378\,137\ mètres\ le\ rayon\ de\ la\ Terre$$

Annexe 3 – Corine Land Cover

Corine Land Cover (CLC) est une représentation biophysique de l'occupation des sols réalisée sur 39 pays dans le cadre du programme européen de surveillance des terres de Copernicus, piloté par l'Agence européenne pour l'environnement. Elle est réalisée sur des zones d'au moins 25 hectares avec une méthodologie commune à l'ensemble des pays européens et calculée principalement grâce à l'interprétation visuelle des images satellites. La nomenclature est répartie en 44 postes divisés selon cinq grandes familles d'occupation du territoire :

- Territoires artificialisés
- Territoires agricoles
- Forêts et milieux semi-naturels
- Zones humides
- Surfaces en eau

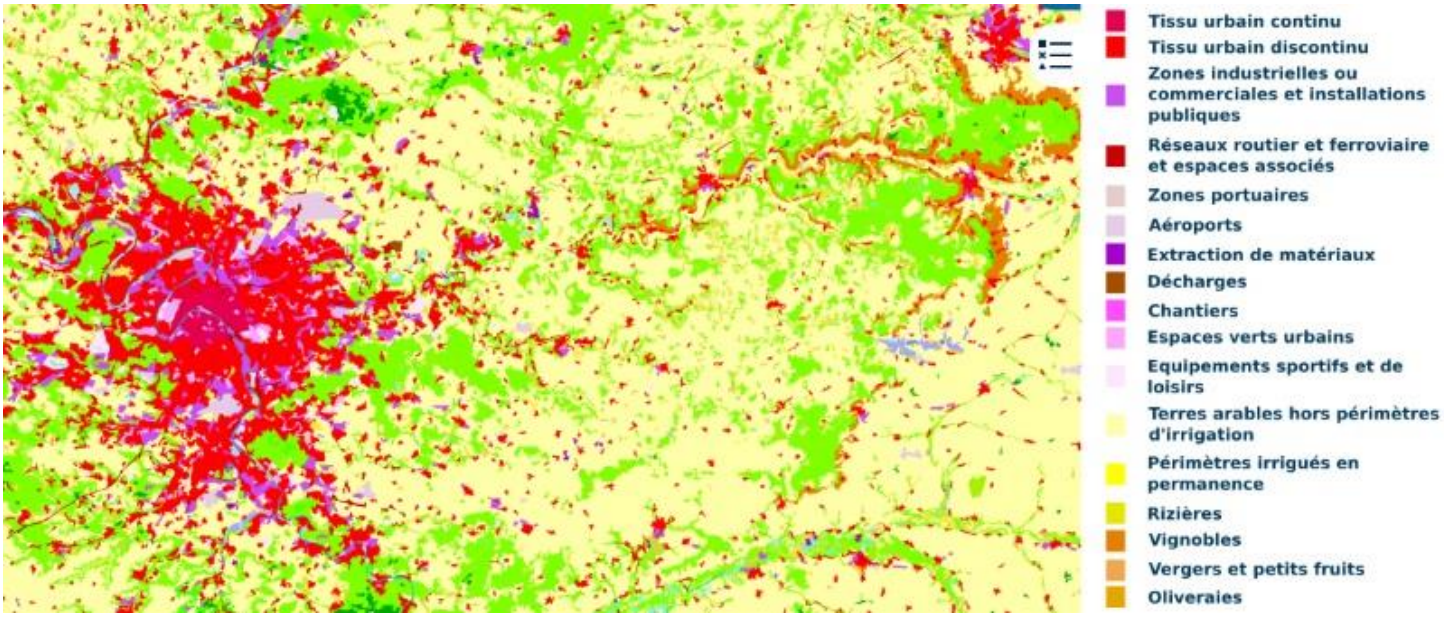

Figure 81 - Corine Land Cover en région parisienne

Annexe 4 – Récapitulatif des variables

Le tableau ci-dessous présente l'ensemble des variables considérées dans l'étude. On compte un total de trois variables cibles qu'on cherche à prédire, ainsi que 33 variables potentiellement explicatives dont 29 issues de l'enrichissement des variables et qui permettent d'alimenter nos modèles. À noter que l'ensemble des 9 variables climatiques est disponible également au climat futur selon trois horizons de temps (2021-2050, 2041-2070, 2071-2100), trois scénarios de changement climatique (RCP 2.6, 4.5, 8.5) ainsi que trois méthodes de calcul (quantile à 5%, 50%, 95%), ce qui fait un total de 243 variables au climat futur (9*3*3*3). On présente ci-dessous la source des données, les calculs effectués ainsi qu'une description rapide de chaque variable :

	Source	Calculs	Nom de la variable	Description de la variable
Target	Modélisateur externe	Calculs à partir de l'AAL (charge annuelle moyenne)	Destruction_rate_Pluvial	Taux de destruction (AAL/Engagements) risque pluvial
			Destruction_rate_Fluvial	Taux de destruction (AAL/Engagements) risque fluvial
			Exposed_fluvial	Egal à 1 si taux de destruction fluvial > 0. Egal à 0 sinon
Variables explicatives	Groupama (données internes)		Portefeuille	Portefeuille assuré (Résidentiel maison, Résidentiel appartemment, Commercial etc...)
			Taux_engt_PE	Part de l'engagement perte d'exploitation sur la somme assurée totale
			Taux_engt_contenu	Part de l'engagement contenu sur la somme assurée totale
			Nombre_etages	Nombre d'étages (disponible uniquement sur le portefeuille immeuble)
	EU-DEM (EEA)	Extraction raster R	Altitude	Altitude du risque (hauteur du risque par rapport au niveau de la mer)
	OpenStreetMap (OSM)	Calcul point le plus proche QGIS puis fonction distance R	Distance_cours_deau	Distance au cours d'eau le plus proche
			Distance_riviere	Distance à la rivière la plus proche
			Distance_ruisseau	Distance au ruisseau le plus proche
	EU-DEM et OSM	Calcul altitude point le plus proche puis différence avec le risque	Difference_altitude_cours_deau	Différence d'altitude entre le risque et le cours d'eau le plus proche
			Difference_altitude_riviere	Différence d'altitude entre le risque et la rivière la plus proche
			Difference_altitude_ruisseau	Différence d'altitude entre le risque et le ruisseau le plus proche
	BD TOPO	Calcul point le plus proche QGIS puis fonction distance R	Distance_riviere_large	Distance au fleuve ou rivière large le plus proche (largeur > 50m)
			Distance_riviere_moyenne	Distance au fleuve ou rivière moyen le plus proche (largeur entre 15 et 50m)
	EU-DEM et BD TOPO	Calcul altitude point le plus proche puis différence avec le risque	Difference_altitude_riviere_large	Différence d'altitude entre le risque et la rivière large la plus proche
			Difference_altitude_riviere_moyenne	Différence d'altitude entre le risque et la rivière moyenne la plus proche
	CLC et Papaioannou et al., 2018	Calcul zone CLC du risque sur QGIS	Manning	Coefficient de Manning issue de la formule de Manning-Strickler
	EEA	Extraction raster R	Impermeabilite_sol	Degré d'imperméabilisation des sols allant de 0 à 100%
	MERIT-DEM	Modèle d'écoulement de l'eau D8 (via algorithme TauDEM sur QGIS)	Surface_contributive	Zones d'accumulation des eaux
			Pente	Coefficient de pente au niveau du risque
			Topographic_Wetness_Index	Indice d'humidité topographique (TWI) au niveau du risque
			Distance_drainage_seuil_50	Distance au réseau de drainage le plus proche avec surface contributive > 50
			Distance_drainage_seuil_500	Distance au réseau de drainage le plus proche avec surface contributive > 500
			Distance_drainage_seuil_5000	Distance au réseau de drainage le plus proche avec surface contributive > 5000
			Distance_drainage_seuil_50000	Distance au réseau de drainage le plus proche avec surface contributive > 50000
	Simulations DRIAS-2020 (issues des simulations Euro-CORDEX)	Extraction raster R	Precip_moyenne	Moyenne des précipitations quotidiennes de la période (période référence : 1976 - 2005)
			Nombre_jour_pluie	Nombre de jours de pluie (> 1mm) moyen par an
			Precip_moyenne_jour_pluvieux	Précipitations moyennes des jours pluvieux
			Nombre_jour_forte_pluie	Nombre de jours de fortes précipitations (>20 mm) moyen par an
			Precip_intense_Q90	Précipitations quotidiennes intenses (quantile à 90%)
			Precip_extreme_Q99	Précipitations quotidiennes extrêmes (quantile à 99%)
			Periode_secheresse	Période de sécheresse maximale (jours consécutifs sans pluie < 1mm)
			Nombre_jour_pluie_consecutif	Nombre maximum de jours pluvieux consécutifs (> 1mm)
			Fraction_precip_intense	Part des précipitations intenses (>quantile à 90%) sur le cumul total des précipitations

Tableau 50 - Récapitulatif des variables utilisées pour les modélisations inondations

Annexe 5 – Visualisation des variables pour le risque fluvial

On a présenté en partie IV.2 la visualisation des variables pour le risque pluvial. La visualisation est faite à partir d'une variable de classification binaire effectuée à partir d'un seuil. On compare la distribution de la variable explicative selon que la variable cible soit supérieure ou inférieure à ce seuil. Cela permet d'avoir une première idée sur la distribution de la variable cible. Pour effectuer cette comparaison, on utilise une boîte à moustache (ou

box plot en anglais) permettant de représenter facilement le profil d'une série statistique et donc le fonctionnement est rappelé ci-dessous :

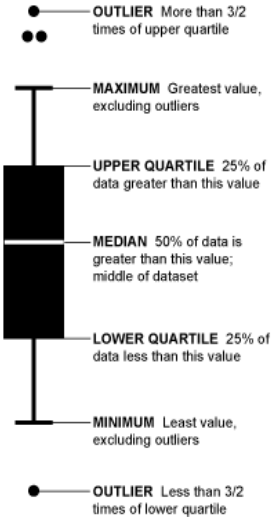

Figure 82 - Fonctionnement d'une boîte à moustache (box plot)

On peut ainsi facilement visualiser la médiane, le troisième quartile, le premier quartile, le maximum (hors valeurs aberrantes) et le minimum (hors valeurs aberrantes). Notons que pour plus de lisibilités nous avons exclu les valeurs aberrantes (*outliers)* des graphes ci-dessous.

Pour rappel, la variable *Exposed_fluvial* est égale à 0 si le risque a une charge annuelle pour le risque fluvial nulle et égale à 1 si celle-ci est strictement positive. On compare donc la distribution des variables selon que le risque est exposé au risque fluvial ou non.

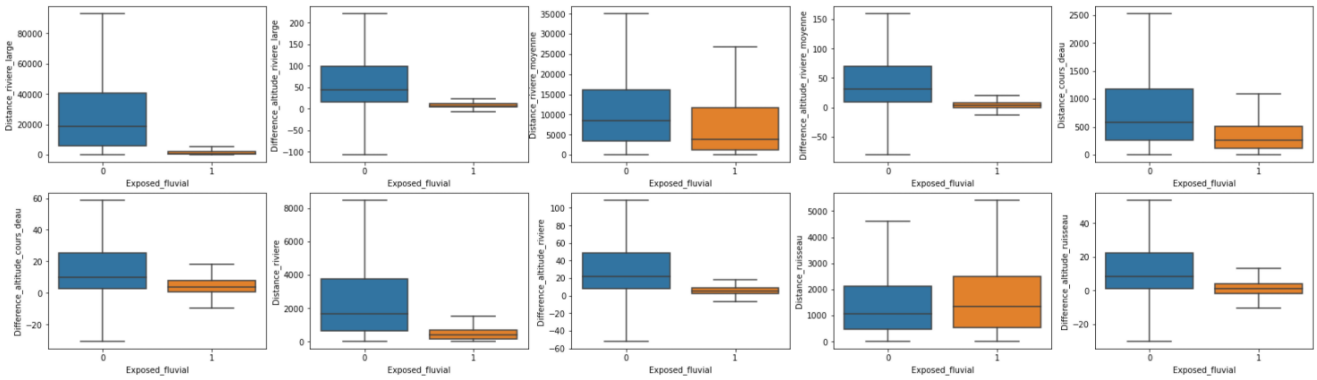

Figure 83 - Boîte à moustaches en fonction des valeurs de Exposed_fluvial pour les variables liées aux distances aux cours d'eau

Dans un premier temps, il apparait que de nombreuses boîtes à moustaches sont beaucoup plus denses et concentrées autour de 0, et ce pour les risques qui sont exposés au risque fluvial. Le risque fluvial désigne l'ensemble des évènements liés au débordement de fleuves et il est donc rassurant de voir ressortir notamment les variables concernant la distance à la rivière large la plus proche et la distance à la rivière ainsi que les différences d'altitude associées. À l'inverse la variable distance au ruisseau semble répartie de façon équivalente entre les risques exposés au risque fluvial et ceux qui ne le sont pas, ce qui est cohérent avec la définition du risque fluvial qui ne prend pas en compte ce type de cours d'eau.

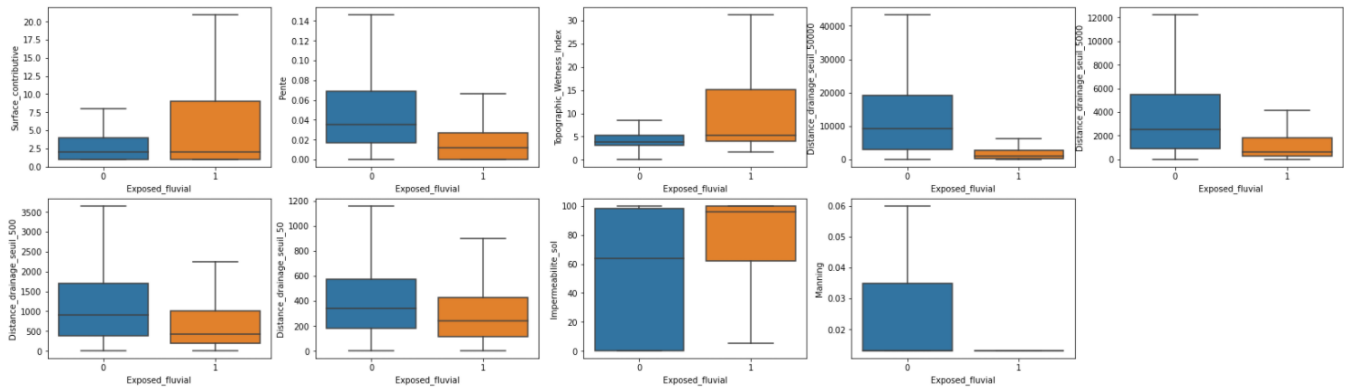

Figure 84 - Boîte à moustaches en fonction des valeurs de Exposed_fluvial pour les variables liées au modèle d'écoulement de l'eau D8 et à l'imperméabilité des sols

Les variables calculées à l'aide du modèle d'écoulement de l'eau D8 semblent également avoir des distributions très différentes en fonction que le risque soit exposé ou non au risque fluvial. Par exemple, il apparait que les risques exposés ont un indice d'humidité topographique qui est de manière générale beaucoup plus important que pour ceux non exposés à ce risque. La distance au réseau de drainage est cohérente avec la définition du risque fluvial, c'est en effet le seuil 50000 qui semble le plus pertinent. C'est le seuil le plus élevé que l'on a considéré et qui permet donc de capter uniquement les réseaux de drainage dans lesquels l'eau est la plus susceptible de s'accumuler, à savoir principalement les grandes rivières et fleuves. L'observation sur l'imperméabilité des sols est la même que pour le risque pluvial, plus l'imperméabilité du sol est élevée et plus le site assuré est susceptible d'être exposé au risque fluvial.

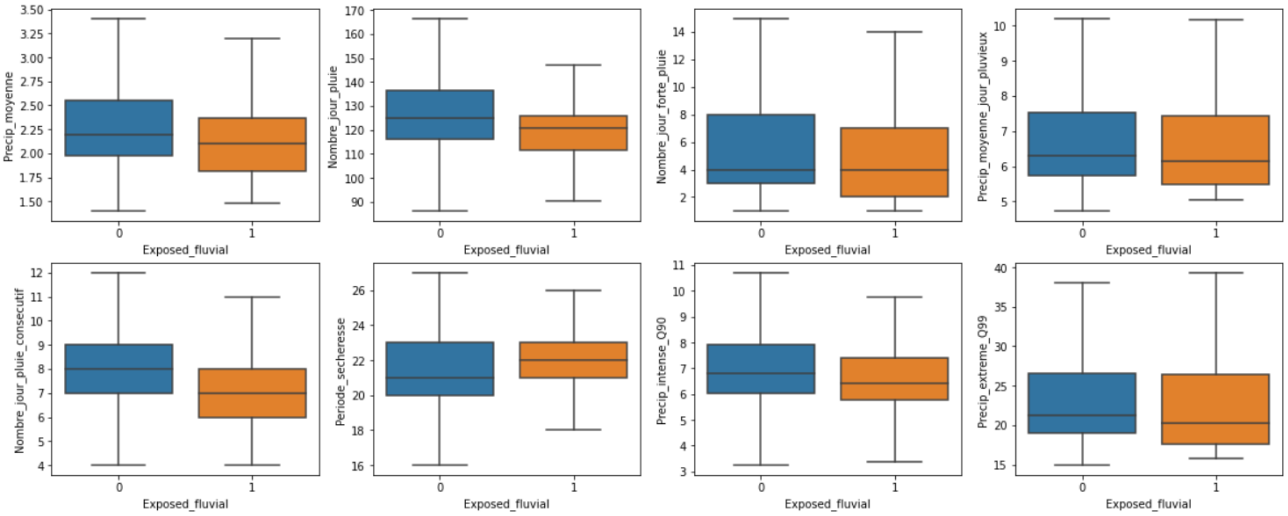

Figure 85 - Boîte à moustaches en fonction des valeurs de Exposed_fluvial pour les variables liées aux précipitations

Enfin, les variables climatiques calculées à partir du cumul des précipitations quotidiennes semblent avoir des résultats beaucoup moins nets que pour les variables précédentes, ce qui est cohérent avec le fait qu'elles ne soient pas ressorties dans le top 15 des variables présentées en partie IV.2. La variable qui présente le plus de différence entre les risques exposés au risque fluvial ou non est la variable calculant le nombre maximum de nombre de jours de pluie consécutifs. Cependant, il apparait que le nombre de jours de pluie consécutifs maximum est plus faible pour les sites assurés exposés à ce risque que pour les autres. Pourtant le risque fluvial est composé principalement de crues lentes de plaines, souvent causées par des précipitations répétées et prolongées, il est donc étonnant de voir apparaitre une corrélation négative entre le nombre de jours de pluie consécutif et l'exposition à ce risque.

Annexe 6 – Résultats détaillés de modélisation pour la partie classification fluviale

Cette annexe vise à présenter le détail des résultats, ainsi que le processus de choix des variables et des paramètres pour les différents modèles de classification qui permettent d'étudier si le risque est exposé au risque fluvial.

a. Résultats de la régression logistique

Le premier modèle essayé un GLM binomial avec une fonction de lien logit. Les résultats sont résumés dans le tableau ci-dessous.

```
Call:
glm2(formula = y_train ~ ., family = binomial(link = logit),
    data = data_train_opt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0410  -0.3245  -0.1187  -0.0204   5.1809

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                     -1.868e+00  4.339e-02 -43.058  < 2e-16 ***
Nombre_etages                    6.013e-02  2.297e-02   2.618  0.00884 **
Distance_riviere                -1.180e-03  1.754e-05 -67.257  < 2e-16 ***
Distance_riviere_large          -4.405e-05  7.881e-07 -55.900  < 2e-16 ***
Difference_altitude_riviere_large -3.023e-03  9.416e-05 -32.108  < 2e-16 ***
Impermeabilite_sol               6.789e-03  3.376e-04  20.107  < 2e-16 ***
Taux_engt_contenu                2.580e-01  3.072e-02   8.397  < 2e-16 ***
Taux_engt_PE                     1.538e-01  1.395e-01   1.102  0.27029
Portefeuille_1                   1.671e-01  1.324e-01   1.262  0.20692
Portefeuille_2                   3.829e-01  6.149e-02   6.228 4.74e-10 ***
Portefeuille_3                   3.552e-01  8.398e-02   4.229 2.34e-05 ***
Portefeuille_4                   4.900e-01  8.575e-02   5.714 1.10e-08 ***
Portefeuille_5                   2.947e-01  1.037e-01   2.841  0.00450 **
Portefeuille_6                   2.557e-01  1.122e-01   2.279  0.02264 *
Portefeuille_7                   3.959e-01  1.592e-01   2.486  0.01291 *
Portefeuille_8                   1.640e-01  1.331e-01   1.232  0.21789
Portefeuille_9                   7.377e-01  4.500e-02  16.395  < 2e-16 ***
Portefeuille_10                  2.331e-01  4.064e-02   5.736 9.71e-09 ***
Portefeuille_11                  1.959e+00  4.699e-01   4.170 3.05e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 98634  on 244716  degrees of freedom
Residual deviance: 72408  on 244698  degrees of freedom
AIC: 72446

Number of Fisher Scoring iterations: 9
```

Tableau 51 - Résultats de la régression logistique pour la variable Exposed_fluvial

La variable perte d'exploitation n'est pas significative au seuil de 5% selon le test de Wald ainsi que les portefeuilles 1 et 8. Pour la sélection de variables, on peut répéter un algorithme descendant en utilisant l'AIC :

Variable_enlevee	Etape	Deviance	Pseudo_R2	AIC
Toutes variables	0	72407.74	0.2658926	72445.74
Taux_engt_PE	1	72408.94	0.2658804	72444.94

Tableau 52 – Résumé de la procédure backward AIC pour la régression logistique – Exposed_fluvial

Les résultats sont cohérents avec le test de Wald, le fait de retirer la variable taux d'engagement perte d'exploitation permet de diminuer l'AIC en conservant un pseudo R2 quasiment identique. On décide donc de retirer cette variable. On obtient finalement les résultats suivants :

Modèle	Seuil_opti_cv	Cross_val_precision	Cross_val_rappel	Cross_val_f_score	TestSet_precision	TestSet_rappel	TestSet_f_score
GLM binomial logit	0.185	0.3026263	0.5008373	0.3710373	0.4193827	0.433164	0.4261619

Tableau 53 - Résultats des indicateurs de performance sur le modèle de régression logistique – Exposed_fluvial

Le seuil de classification permettant d'optimiser le score F1 par validation croisée est de 0.185. On obtient un score F1 de 37% avec une précision de 30% et un rappel de 50%.

b. Résultats du *XGBoost*

Tout comme pour le risque pluvial, on utilise l'algorithme d'élimination récursive de variables pour faire notre sélection. À chaque itération, on recherche le seuil optimal permettant de maximiser le score F1 pour chaque sous-ensemble de variables. L'indicateur maximal est obtenu à l'étape 5, on retire donc l'information des portefeuilles 1,7,8 et 11 qui n'ont aucun pouvoir prédictif et qui font baisser l'indicateur de performance.

Var_remove	Step	Meilleur_seuil	precision_cv	rappel_cv	f_score_cv
Aucune	1	0.199	0.4030938	0.6523375	0.4925373
Portefeuille_1	2	0.201	0.4048561	0.6507642	0.4934088
Portefeuille_7	3	0.206	0.4066049	0.6434055	0.4925659
Portefeuille_8	4	0.206	0.4066049	0.6434055	0.4925659
Portefeuille_11	5	0.210	0.4096732	0.6420115	0.4943810
Portefeuille_5	6	0.187	0.3968328	0.6613949	0.4905998
Portefeuille_4	7	0.188	0.3953597	0.6626022	0.4896542
Portefeuille_6	8	0.213	0.4087716	0.6364188	0.4917941
Portefeuille_2	9	0.211	0.4038526	0.6395070	0.4897669
Portefeuille_3	10	0.228	0.4108505	0.6219353	0.4890941
Nombre_etages	11	0.200	0.3986405	0.6470717	0.4879692
Portefeuille_12	12	0.199	0.4016948	0.6535500	0.4921385
Taux_engt_PE	13	0.216	0.4104996	0.6393350	0.4939203
Portefeuille_10	14	0.205	0.4048778	0.6502051	0.4935979
Taux_engt_contenu	15	0.227	0.4175030	0.6254789	0.4936459
Portefeuille_9	16	0.213	0.3927243	0.6323243	0.4788488
Impermeabilite_sol	17	0.161	0.3731855	0.7027727	0.4814545
Distance_riviere	18	0.199	0.3771107	0.6411880	0.4699266
Difference_altitude_riviere_large	19	0.170	0.3200746	0.5766670	0.4068950

Tableau 54 - Algorithme de Recursive Feature Elimination appliqué sur le XGBoost - Exposed_fluvial

L'algorithme *GridSearch* nous permet de faire passer le score F1 au-dessus de la barre des 50% en sélectionnant un nombre d'étapes de *boosting* de 100, une profondeur maximale d'arbre égale à 6 et un taux d'apprentissage eta de 0.15. Le seuil permettant d'obtenir le meilleur score F1 de validation croisée pour ce jeu de paramètres est de 0.248.

nrounds	max_depth	eta	seuil_opti	precision_cv	rappel_cv	f_score_cv
100	6	0.15	0.248	0.4271718	0.6252551	0.5009043
100	6	0.30	0.210	0.4096732	0.6420115	0.4943810
300	6	0.15	0.203	0.4068541	0.6358793	0.4902328
100	10	0.15	0.182	0.4000815	0.6363447	0.4854183
300	10	0.15	0.116	0.3790701	0.6574293	0.4749863
300	6	0.30	0.161	0.3841109	0.6289961	0.4712979
100	10	0.30	0.144	0.3826087	0.6339752	0.4709741
300	10	0.30	0.086	0.3752824	0.6349843	0.4653891

Tableau 55 - Recherche des meilleurs paramètres pour le XGBoost – Exposed_fluvial

On sélectionne donc finalement ces paramètres permettant d'obtenir un score F1 sur la base de test de 58% avec une précision de 52% et un rappel de 67%.

Modèle	Seuil_opti_cv	Cross_val_precision	Cross_val_rappel	Cross_val_f_score	TestSet_precision	TestSet_rappel	TestSet_f_score
Gradient Boosting	0.248	0.4271718	0.6252551	0.5009043	0.515733	0.6658199	0.581244

Tableau 56 - Résultats des indicateurs de performance sur le modèle de XGBoost – Exposed_fluvial

c. Résultats du *Random Forest*

L'algorithme de sélection de variables sur la forêt aléatoire nous indique qu'il faut conserver l'ensemble des variables étant donné que le score F1 se dégrade au fur et à mesure à chaque nouvelle suppression de variables. Notons un apport plus important des variables distance à la rivière, différence d'altitude à la rivière large et distance à la rivière large (qui est la dernière variable conservée par l'algorithme) pour lesquelles leur suppression a un impact significatif sur la diminution de la performance.

Var_remove	Step	Meilleur_seuil	precision_cv	rappel_cv	f_score_cv
Aucune	1	0.225	0.4057949	0.6387628	0.4897248
Nombre_etages	2	0.228	0.3980998	0.6294160	0.4812753
Taux_engt_PE	3	0.218	0.3878055	0.6388071	0.4759998
Portefeuille	4	0.206	0.3756830	0.6443818	0.4677565
Taux_engt_contenu	5	0.158	0.3455659	0.6948325	0.4553697
Impermeabilite_sol	6	0.195	0.3526770	0.6276738	0.4452312
Distance_riviere	7	0.184	0.3239666	0.5978113	0.4133693
Difference_altitude_riviere_large	8	0.041	0.2014390	0.5168732	0.2869801

Tableau 57 - Algorithme de Recursive Feature Elimination appliqué sur le Random Forest - Exposed_fluvial

De la même manière que pour le boosting de gradient, l'optimisation des paramètres permet de faire passer l'indicateur F1 au-dessus des 50%. Pour obtenir ce score optimal, il faut initialiser le nombre de variables testées à chaque division avec la valeur renseignée par défaut, la profondeur d'arbre à 12 et enfin le nombre d'arbres considérés à 100.

mtry	max.depth	num.trees	seuil_opti	precision_cv	rappel_cv	f_score_cv
0	12	100	0.234	0.4282011	0.6367665	0.5057002
4	8	100	0.217	0.4133437	0.6707143	0.5044686
0	8	100	0.212	0.4254148	0.6375195	0.5039485
0	16	100	0.240	0.4202912	0.6283868	0.4975689
0	0	100	0.229	0.4062506	0.6370169	0.4901549
4	12	100	0.238	0.4081441	0.6273177	0.4872603
4	16	100	0.227	0.3887766	0.6205153	0.4712938
4	0	100	0.217	0.3769671	0.6253740	0.4638859

Tableau 58 - Recherche des meilleurs paramètres pour le Random Forest – Exposed_fluvial

On obtient finalement les résultats suivants, avec une précision sur la base de test de 52%, un rappel à 67% et un score F1 de 59%.

Modèle	Seuil_opti_cv	Cross_val_precision	Cross_val_rappel	Cross_val_f_score	TestSet_precision	TestSet_rappel	TestSet_f_score
Random Forest	0.234	0.4282011	0.6367665	0.5057002	0.5242377	0.6664884	0.586866

Tableau 59 - Résultats des indicateurs de performance sur le modèle de Random Forest – Exposed_fluvial

Annexe 7 – Résultats détaillés de modélisation pour la partie régression fluviale

Cette annexe vise à présenter le détail des résultats pour les différents modèles, ainsi que le processus de choix des variables et des paramètres pour les modèles de régression qui permettent de prédire le taux de destruction annuel moyen pour le risque fluvial.

a. Résultats du GLM Gamma

Pour cette partie, on essayera les mêmes modèles que pour la prédiction du taux de destruction pluvial. On commence donc par le GLM Gamma avec une fonction de lien log et pour lequel il apparait que l'ensemble des variables est significatif au seuil de 5% selon le test de Student à l'exception de la variable concernant la différence d'altitude à la rivière large, ainsi que la modalité liée au portefeuille 11.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.4748  -2.3870  -1.2927   0.1328  13.1876

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     -5.202e+00  1.030e-01 -50.494  < 2e-16 ***
Nombre_etages                   -2.170e-01  5.384e-02  -4.030 5.62e-05 ***
Distance_riviere                -5.369e-04  3.560e-05 -15.083  < 2e-16 ***
Distance_riviere_large          -2.179e-06  1.062e-06  -2.052  0.04020 *
Difference_altitude_riviere_large -7.305e-04  4.115e-04  -1.775  0.07592 .
Impermeabilite_sol              -1.886e-02  7.992e-04 -23.600  < 2e-16 ***
Taux_engt_contenu                8.068e-01  7.074e-02  11.405  < 2e-16 ***
Taux_engt_PE                    -6.861e-01  3.294e-01  -2.083  0.03731 *
Portefeuille_1                  -1.336e+00  3.178e-01  -4.204 2.64e-05 ***
Portefeuille_2                  -1.059e+00  1.482e-01  -7.148 9.26e-13 ***
Portefeuille_3                  -1.040e+00  2.016e-01  -5.159 2.52e-07 ***
Portefeuille_4                  -1.918e+00  2.014e-01  -9.523  < 2e-16 ***
Portefeuille_5                  -4.899e-01  2.449e-01  -2.000  0.04550 *
Portefeuille_6                  -2.159e+00  2.662e-01  -8.108 5.62e-16 ***
Portefeuille_7                  -1.082e+00  3.753e-01  -2.884  0.00394 **
Portefeuille_8                  -6.393e-01  3.150e-01  -2.030  0.04242 *
Portefeuille_9                  -1.685e+00  1.085e-01 -15.530  < 2e-16 ***
Portefeuille_10                 -6.104e-01  1.000e-01  -6.103 1.07e-09 ***
Portefeuille_11                  1.260e+00  1.060e+00   1.189  0.23459
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for Gamma family taken to be 6.683443)

    Null deviance: 73772  on 12486  degrees of freedom
Residual deviance: 56120  on 12468  degrees of freedom
AIC: -189657

Number of Fisher Scoring iterations: 25
```

Tableau 60 - Résultats du GLM Gamma sur les taux de destruction du risque fluvial

La sélection de variables utilisant l'AIC indique que l'on peut conserver l'ensemble des variables étant donné que dès la première étape de l'algorithme, l'AIC augmente, peu importe la variable qui est enlevée. Notons tout de même l'importance de la variable distance à la rivière, portefeuille ainsi que celle liée à l'imperméabilité du sol qui ont le plus fort impact sur la diminution du pseudo R2 lorsqu'elles sont retirées. Le pseudo R2 obtenu avec l'ensemble des variables étant de 24%.

Variable_enlevee	Deviance	Pseudo_R2	AIC
Aucune	56120.44	0.2392725	-189657.1
Distance_riviere_large	56133.93	0.2390896	-189654.8
Taux_engt_PE	56144.64	0.2389445	-189651.3
Nombre_etages	56196.45	0.2382422	-189634.7
Difference_altitude_riviere_large	56334.27	0.2363741	-189590.4
Taux_engt_contenu	56918.35	0.2284566	-189403.7
Distance_riviere	58468.41	0.2074451	-188916.0
Portefeuille	58755.19	0.2035578	-188846.9
Impermeabilite_sol	60386.38	0.1814466	-188327.3

Tableau 61 - Première étape de la procédure backward AIC pour le modèle GLM Gamma - taux de destruction fluvial

On obtient pour finir les indicateurs de performance ci-dessous, notons que le coefficient de Spearman est meilleur que pour le risque pluvial, vu qu'il atteint quasiment les 50%. Il apparait également que l'on semble surestimer encore une fois la charge totale de près de 50% en moyenne sur les dix itérations de validation croisée.

Model	Cross_val_RMSE	Cross_val_Total_AAL_diff	Cross_val_Spearman	TestSet_RMSE	TestSet_Total_AAL_diff	TestSet_Spearman
Gamma	0.001198327	0.485765	0.4917637	0.001198234	0.5183526	0.4978091

Tableau 62 - Résultats des indicateurs de performance sur le modèle GLM Gamma optimal - taux de destruction fluvial

b. Résultats de la régression Bêta

En appliquant une régression Bêta à nos taux de destruction il apparait que la variable nombre d'étages et que les portefeuilles 1,2,3,5,7 et 8 ne sont pas significatifs au seuil de 5% selon le test de Wald.

```
Standardized weighted residuals 2:
    Min      1Q  Median      3Q     Max
-6.0403 -0.3849  0.1386  0.5302  3.7722

Coefficients (mean model with logit link):
                                  Estimate Std. Error  z value Pr(>|z|)
(Intercept)                     -6.778e+00  4.054e-02 -167.206  < 2e-16 ***
Nombre_etages                   -1.346e-02  2.035e-02   -0.662  0.50816
Distance_riviere                -2.260e-04  1.350e-05  -16.739  < 2e-16 ***
Distance_riviere_large          -5.009e-06  4.025e-07  -12.444  < 2e-16 ***
Difference_altitude_riviere_large -2.567e-03  5.790e-05  -44.338  < 2e-16 ***
Impermeabilite_sol              -2.786e-03  2.948e-04   -9.451  < 2e-16 ***
Taux_engt_contenu                1.590e-01  2.633e-02    6.037 1.57e-09 ***
Taux_engt_PE                    -2.442e-01  1.228e-01   -1.989  0.04674 *
Portefeuille_1                  -1.667e-01  1.184e-01   -1.408  0.15914
Portefeuille_2                   8.062e-02  5.494e-02    1.467  0.14231
Portefeuille_3                   2.469e-02  7.301e-02    0.338  0.73523
Portefeuille_4                  -1.586e-01  7.552e-02   -2.101  0.03567 *
Portefeuille_5                   4.291e-02  9.066e-02    0.473  0.63595
Portefeuille_6                  -3.077e-01  1.004e-01   -3.064  0.00219 **
Portefeuille_7                   1.886e-01  1.372e-01    1.374  0.16930
Portefeuille_8                  -1.237e-01  1.187e-01   -1.042  0.29745
Portefeuille_9                  -1.137e-01  4.036e-02   -2.818  0.00483 **
Portefeuille_10                  7.951e-02  3.711e-02    2.143  0.03215 *
Portefeuille_11                  1.899e+00  2.906e-01    6.535 6.37e-11 ***

Phi coefficients (precision model with identity link):
      Estimate Std. Error z value Pr(>|z|)
(phi)  320.743      6.511   49.27   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Type of estimator: ML (maximum likelihood)
Log-likelihood: 9.323e+04 on 20 Df
Pseudo R-squared: 0.06298
Number of iterations: 117 (BFGS) + 2 (Fisher scoring)
```

Tableau 63 - Résultats du GLM Bêta sur les taux de destruction du risque fluvial

Le résultat se confirme avec l'algorithme descendant vu que le retrait de la variable nombre d'étages de notre modèle permet de faire diminuer l'AIC. On décide donc de retirer cette variable pour ce modèle. Le processus s'arrête au bout de l'étape 1 étant donné qu'aucun autre retrait de variable ne permet de faire diminuer l'AIC par la suite.

Variable_enlevee	Etape	AIC
Toutes variables	0	-186426.1
Nombre_etages	1	-186427.7

Tableau 64 - Résumé de la procédure backward AIC pour la régression Bêta – taux de destruction risque fluvial

Si la corrélation de Spearman est globalement similaire à ce qu'on a pu trouver pour le GLM gamma, il apparait que la différence entre l'AAL totale observée et modélisée est près de 1,5 fois supérieure selon la validation croisée ce qui fait une différence beaucoup trop importante pour notre étude. De plus, on obtient également un RMSE supérieur à ce qui a été obtenu pour le GLM gamma.

Model	Cross_val_RMSE	Cross_val_Total_AAL_diff	Cross_val_Spearman	TestSet_RMSE	TestSet_Total_AAL_diff	TestSet_Spearman
Beta	0.001199956	1.513671	0.4930183	0.001222359	1.907303	0.5019801

Tableau 65 - Résultats des indicateurs de performance sur le modèle GLM Bêta optimal - taux de destruction fluvial

c. Résultats du XGBoost

Concernant le boosting de gradient, il apparait selon l'élimination récursive de variable que l'optimisation du RMSE est obtenu en retirant une majorité des variables et en conservant uniquement la variable distance à la rivière ainsi que distance à la rivière large, qui est la dernière variable restante à l'issue de l'algorithme. Cette sélection de variables permet d'améliorer significativement le RMSE, passant en effet de 0.001448 avec l'ensemble des variables à 0.001328 en ne sélectionnant que les deux variables en question.

Var_remove	Step	RMSE_cv	Total_AAL_Diff_cv	Spearman_score_cv
Aucune	1	0.001447619	48.0415 %	49.1698 %
Portefeuille_4	2	0.001448770	47.3222 %	49.1829 %
Portefeuille_6	3	0.001448991	55.4561 %	48.9892 %
Portefeuille_7	4	0.001448572	55.6948 %	49.0362 %
Portefeuille_8	5	0.001448572	55.6948 %	49.0362 %
Portefeuille_11	6	0.001437651	54.0387 %	49.1527 %
Portefeuille_9	7	0.001456750	48.9734 %	49.0551 %
Taux_engt_PE	8	0.001448950	42.7509 %	49.0383 %
Portefeuille_2	9	0.001442450	46.6471 %	49.0477 %
Portefeuille_10	10	0.001452496	49.9614 %	49.025 %
Nombre_etages	11	0.001455446	52.9145 %	49.0895 %
Portefeuille_3	12	0.001457869	55.7581 %	49.2236 %
Portefeuille_5	13	0.001461314	53.4539 %	49.1659 %
Portefeuille_1	14	0.001459305	58.8503 %	49.0293 %
Impermeabilite_sol	15	0.001514926	61.53 %	49.1005 %
Portefeuille_12	16	0.001465160	55.4473 %	49.0587 %
Taux_engt_contenu	17	0.001477894	66.4942 %	49.2407 %
Difference_altitude_riviere_large	18	0.001327823	67.5133 %	48.9772 %
Distance_riviere	19	0.001352423	48.5966 %	49.3088 %

Tableau 66 - Algorithme de Recursive Feature Elimination appliqué sur le XGBoost risque fluvial

En choisissant un nombre d'itérations de 50, une profondeur maximale d'arbre de 4 et un taux d'apprentissage de 0.3, cela permet d'améliorer davantage le RMSE. On décide donc de sélectionner ce jeu de paramètres malgré une légère dégradation au niveau du score d'AAL qui passe de 68 à 76%.

nrounds	max_depth	eta	RMSE_cv	Total_AAL_Diff_cv	Spearman_score_cv
50	4	0.3	0.001310219	0.7572280	0.4919843
200	10	0.1	0.001321782	0.6524816	0.4931629
50	8	0.3	0.001324566	0.6068095	0.4927733
200	8	0.3	0.001324566	0.6069269	0.4927733
50	6	0.3	0.001327823	0.6751327	0.4897715
200	6	0.3	0.001327861	0.6788827	0.4897685
200	6	0.1	0.001334756	0.6522429	0.4925288
200	10	0.3	0.001336872	0.5922395	0.4931450
50	10	0.3	0.001336872	0.5922520	0.4931450
200	4	0.3	0.001337127	0.7370698	0.4916131
200	8	0.1	0.001344660	0.6701709	0.4928262
200	4	0.1	0.001353513	0.9111827	0.4919312
50	4	0.1	0.001509973	8.7359672	0.4926171
50	6	0.1	0.001512108	8.5942333	0.4932573
50	8	0.1	0.001519000	8.5722884	0.4925130
50	10	0.1	0.001525827	8.5661900	0.4927393

Tableau 67 - Recherche des meilleurs paramètres pour le XGBoost risque fluvial

Cette fois-ci, il semble que l'on sous-estime la charge globale sur la base de test, avec en effet une charge annuelle moyenne inférieure de 40% à ce qui est calculé par notre modélisateur.

Modele	GB_RMSE_cv	GB_Total_AAL_Diff_cv	GB_Sperman_score_cv	GB_RMSE_test	GB_Total_AAL_diff_test	GB_Sperman_score_test
Gradient Boosting RFE	0.001310219	0.757228	0.4919843	0.001361898	-0.3838808	0.5031777

Tableau 68 - Résultats des indicateurs de performance sur le modèle XGBoost risque fluvial

d. Résultats du Random Forest

Pour finir, on essaye un modèle de forêt aléatoire. La racine de l'erreur quadratique moyenne optimale est de 0.00129 et est obtenue en conservant l'ensemble des variables. Cela permet également de minimiser le score d'AAL égal à environ 60%.

Var_remove	Step	RMSE_cv	Total_AAL_Diff_cv	Spearman_score_cv
Aucune	1	0.001292089	59.7664 %	49.3059 %
Nombre_etages	2	0.001304915	61.3986 %	49.325 %
Taux_engt_PE	3	0.001317169	61.0379 %	49.3272 %
Taux_engt_contenu	4	0.001332360	66.0232 %	49.3391 %
Impermeabilite_sol	5	0.001414059	72.8646 %	49.3032 %
Portefeuille	6	0.001380509	83.982 %	49.367 %
Distance_riviere	7	0.001435655	84.012 %	49.4208 %
Difference_altitude_riviere_large	8	0.001363618	78.9881 %	49.3632 %

Tableau 69 - Algorithme de Recursive Feature Elimination appliqué sur le Random Forest - taux de destruction fluvial

En optimisant par la suite les paramètres, on arrive à améliorer encore une fois le score de RMSE, qui passe ainsi de 0.00129 à 0.00125 avec une profondeur maximale d'arbre égale à 4 et un nombre d'arbres égal à 500.

mtry	max_depth	num.trees	RMSE_cv	Total_AAL_Diff_cv	Spearman_score_cv
2	4	500	0.001252056	0.8550581	0.4915053
2	4	100	0.001263595	0.8783447	0.4914833
2	6	500	0.001272668	0.6924776	0.4922117
2	6	100	0.001283136	0.7026327	0.4921539
2	8	500	0.001284377	0.6236337	0.4926249
2	12	500	0.001290725	0.5934129	0.4930578
2	0	500	0.001292089	0.5976645	0.4930587
2	8	100	0.001296739	0.6239267	0.4927249
2	12	100	0.001302491	0.5931257	0.4932568
2	0	100	0.001304084	0.6021999	0.4933620

Tableau 70 - Recherche de meilleurs paramètres pour le Random Forest - taux de destruction pluvial

Enfin comme pour le boosting de gradient ce modèle a tendance à sous-estimer la charge globale comme on peut le voir sur les résultats sur les données de test. Pour rappel, le score d'AAL par validation croisée est calculé en valeur absolue, tandis que celui sur la base de test est calculé sans. Il n'est donc pas incohérent de voir un score positif sur la validation croisée et négatif sur les données de test.

Model	Cross_val_RMSE	Cross_val_Total_AAL_diff	Cross_val_Spearman	TestSet_RMSE	TestSet_Total_AAL_diff	TestSet_Spearman
Random Forest	0.001252056	0.8550581	0.4915053	0.00116439	-0.6047074	0.5072335

Tableau 71 - Résultats des indicateurs de performance sur le modèle Random Forest - taux de destruction fluvial

Annexe 8 – Impact du changement climatique sur l'ensemble des indicateurs climatiques

On a présenté en partie V, l'évolution de quatre indicateurs de précipitations avec le changement climatique, à savoir les précipitations moyennes, le nombre de jours de pluie, le nombre de jours de forte pluie, ainsi que la variable « précipitations extrêmes » représentant le quantile à 99% des précipitations quotidiennes. On s'intéressera ci-dessous à l'évolution des autres indicateurs, selon la médiane des modèles DRIAS-2020.

MEDIANE DES MODELES	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100
	Auvergne-Rhône-Alpes	4%	6%	5%	4%	6%	8%	5%	7%	12%
	Bourgogne-Franche-Comté	5%	6%	6%	5%	7%	9%	7%	9%	15%
	Bretagne	4%	4%	3%	4%	6%	7%	5%	8%	14%
	Centre-Val de Loire	4%	4%	4%	4%	6%	8%	5%	8%	14%
	Corse	4%	5%	6%	4%	6%	8%	4%	7%	8%
	Grand Est	5%	5%	5%	5%	7%	9%	7%	10%	16%
	Hauts-de-France	5%	5%	5%	4%	7%	8%	5%	9%	17%
	Île-de-France	4%	4%	4%	5%	7%	8%	5%	9%	15%
	Normandie	4%	4%	5%	4%	6%	8%	5%	8%	14%
	Nouvelle-Aquitaine	4%	5%	5%	5%	6%	7%	3%	6%	11%
	Occitanie	4%	5%	6%	4%	4%	6%	4%	5%	8%
	Pays de la Loire	4%	5%	4%	5%	7%	9%	5%	8%	14%
	Provence-Alpes-Côte d'Azur	3%	5%	5%	3%	5%	8%	5%	6%	9%
	Total général	4%	**5%**	**5%**	**4%**	**6%**	**8%**	**5%**	**7%**	**12%**

Tableau 72 - Projections d'évolution des précipitations moyennes les jours pluvieux par région selon le jeu de données DRIAS-2020

L'évolution des précipitations moyennes les jours pluvieux est cohérente avec les premières conclusions déduites en partie V. Il apparait en effet que si le nombre de jours de pluie va certes diminuer, les précipitations observées durant ces jours de pluie seront plus intenses. Comme observé précédemment l'augmentation pour le scénario RCP2.6 ne devrait plus évoluer à partir de la période 2041-2070 avec une augmentation se stabilisant à +5%. Le constat est différent pour les deux autres scénarios et l'augmentation devrait atteindre les +8% à horizon 2100 pour le scénario RCP4.5 et +12% à horizon 2100 pour le scénario RCP8.5. Le sud de la France, avec les régions Corse, Occitanie et Provence-Alpes-Côte d'Azur, semble moins touché par ces évolutions et notamment pour le scénario RCP8.5 pour lequel les évolutions ne dépassent pas la barre des +10% d'ici 2100.

MEDIANE DES MODELES	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100
	Auvergne-Rhône-Alpes	4%	6%	5%	2%	2%	4%	3%	3%	-1%
	Bourgogne-Franche-Comté	6%	6%	5%	4%	5%	7%	6%	7%	8%
	Bretagne	3%	3%	3%	3%	1%	2%	2%	3%	5%
	Centre-Val de Loire	4%	4%	4%	3%	3%	4%	4%	4%	4%
	Corse	0%	4%	7%	-3%	-3%	-5%	-8%	-7%	-21%
	Grand Est	6%	6%	5%	4%	5%	8%	6%	8%	12%
	Hauts-de-France	6%	6%	5%	4%	4%	6%	4%	7%	12%
	Île-de-France	5%	5%	3%	3%	3%	4%	4%	4%	7%
	Normandie	5%	3%	4%	3%	3%	4%	4%	6%	9%
	Nouvelle-Aquitaine	3%	4%	5%	3%	1%	3%	1%	1%	-2%
	Occitanie	1%	4%	6%	0%	-1%	-1%	-1%	-2%	-10%
	Pays de la Loire	4%	4%	4%	4%	1%	3%	2%	3%	4%
	Provence-Alpes-Côte d'Azur	2%	7%	5%	-1%	-3%	-2%	-3%	-5%	-18%
	Total général	4%	**5%**	**5%**	**3%**	**2%**	**3%**	**2%**	**3%**	**1%**

Tableau 73 - Projections d'évolution des précipitations intenses (quantile à 90%) par région selon le jeu de données DRIAS-2020

L'augmentation du quantile à 90% est beaucoup moins nette que pour le quantile à 99%. L'augmentation sur l'ensemble du territoire sera plus importante pour le scénario RCP2.6 que pour le scénario RCP8.5, notamment dû au fait que le nombre de jours de pluie devrait davantage diminuer pour ce dernier scénario, le quantile à 90% est donc impacté par ce résultat. De plus, il apparait que pour les scénarios les plus pessimistes la variabilité d'évolution entre les régions est beaucoup plus importante, avec des évolutions allant de -21 à +12% d'ici 2100 pour le scénario le plus pessimiste.

MEDIANE DES MODELES	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100
	Auvergne-Rhône-Alpes	0%	0%	1%	1%	3%	3%	2%	3%	7%
	Bourgogne-Franche-Comté	0%	0%	1%	1%	2%	3%	2%	3%	7%
	Bretagne	1%	1%	1%	2%	4%	4%	3%	4%	9%
	Centre-Val de Loire	1%	1%	0%	1%	3%	3%	2%	3%	8%
	Corse	1%	1%	0%	2%	3%	4%	3%	4%	8%
	Grand Est	0%	0%	1%	1%	2%	2%	2%	3%	6%
	Hauts-de-France	0%	0%	0%	1%	2%	2%	1%	3%	6%
	Île-de-France	1%	1%	0%	1%	3%	3%	2%	3%	7%
	Normandie	1%	1%	0%	1%	3%	3%	2%	3%	7%
	Nouvelle-Aquitaine	1%	1%	0%	1%	3%	3%	2%	4%	9%
	Occitanie	1%	0%	0%	2%	2%	3%	2%	3%	8%
	Pays de la Loire	1%	1%	0%	1%	4%	4%	3%	4%	8%
	Provence-Alpes-Côte d'Azur	1%	0%	0%	1%	2%	2%	2%	3%	6%
	Total général	1%	1%	0%	1%	3%	3%	2%	3%	8%

Tableau 74 - Projections d'évolution de l'indicateur « fraction des précipitations quotidiennes intenses » par région selon le jeu de données DRIAS-2020

La fraction des précipitations quotidiennes intenses désigne la part que représentent les précipitations supérieures au quantile à 90% sur le total des précipitations. Cet indicateur est notamment particulièrement élevé dans le sud de la France. L'évolution semble uniforme entre les territoires et devrait augmenter de +3% à horizon 2100 pour le scénario RCP4.5 et de +8% pour le scénario RCP8.5, signe que les précipitations intenses devraient représenter une part de plus en plus importante sur le total des précipitations.

MEDIANE DES MODELES	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100
	Auvergne-Rhône-Alpes	2%	5%	2%	2%	0%	0%	1%	-1%	-3%
	Bourgogne-Franche-Comté	4%	8%	5%	3%	1%	2%	1%	2%	2%
	Bretagne	1%	1%	2%	0%	-2%	-1%	0%	0%	-3%
	Centre-Val de Loire	8%	9%	8%	6%	5%	4%	6%	6%	3%
	Corse	0%	3%	2%	-1%	-2%	-2%	-3%	-4%	-11%
	Grand Est	5%	9%	7%	3%	1%	4%	2%	4%	4%
	Hauts-de-France	8%	9%	5%	4%	3%	3%	4%	5%	4%
	Île-de-France	6%	8%	8%	6%	2%	0%	3%	3%	2%
	Normandie	4%	4%	5%	2%	1%	1%	2%	3%	0%
	Nouvelle-Aquitaine	0%	2%	2%	2%	1%	2%	2%	0%	-4%
	Occitanie	-1%	1%	2%	-1%	-2%	-1%	-1%	-2%	-8%
	Pays de la Loire	2%	3%	1%	2%	-1%	-1%	1%	1%	-2%
	Provence-Alpes-Côte d'Azur	2%	5%	2%	-1%	-3%	-2%	-2%	-4%	-10%
	Total général	3%	5%	4%	2%	0%	1%	1%	1%	-2%

Tableau 75 - Projections d'évolution du nombre maximum de jours pluvieux consécutifs par région selon le jeu de données DRIAS-2020

Le nombre maximum de jours pluvieux consécutifs semble globalement stagner avec le changement climatique avec une légère hausse de +4% pour le scénario RCP2.6 d'ici 2100, de +1% pour le RCP4.5 et une légère baisse de -2% pour le scénario RCP8.5.

MEDIANE DES MODELES	Régions	RCP2.6			RCP4.5			RCP8.5		
		2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100	2021-2050	2041-2070	2071-2100
	Auvergne-Rhône-Alpes	4%	6%	2%	5%	6%	6%	3%	6%	17%
	Bourgogne-Franche-Comté	5%	6%	1%	3%	4%	4%	1%	3%	12%
	Bretagne	4%	4%	-1%	4%	6%	14%	4%	12%	27%
	Centre-Val de Loire	6%	5%	3%	5%	6%	13%	5%	11%	26%
	Corse	3%	-1%	-1%	9%	9%	12%	4%	6%	26%
	Grand Est	5%	4%	1%	2%	5%	4%	1%	3%	9%
	Hauts-de-France	4%	3%	0%	2%	5%	4%	3%	5%	14%
	Île-de-France	3%	3%	2%	3%	5%	8%	6%	8%	20%
	Normandie	5%	3%	2%	4%	6%	12%	4%	10%	22%
	Nouvelle-Aquitaine	5%	4%	0%	6%	7%	12%	5%	11%	27%
	Occitanie	4%	4%	0%	7%	7%	11%	7%	11%	30%
	Pays de la Loire	5%	4%	1%	5%	8%	18%	6%	13%	30%
	Provence-Alpes-Côte d'Azur	4%	2%	-1%	5%	5%	6%	5%	5%	22%
	Total général	4%	4%	1%	5%	6%	9%	4%	8%	22%

Tableau 76 - Projections d'évolution des périodes de sécheresse (nombre maximum de jours sans pluie) par région selon le jeu de données DRIAS-2020

L'indicateur des périodes de sécheresse, calculé à partir du nombre maximum de jours sans pluie, devrait significativement augmenter avec le changement climatique. L'augmentation est particulièrement importante pour le scénario RCP8.5 entre la période 2041-2070 et 2071-2100 avec un passage de +8 à +22%. Pour le scénario RCP2.6, on devrait retrouver un retour à la normale à horizon 2100 avec un passage de +4% sur la période 2041-2070 à +1% sur la période 2071-2100.