ENSTA Paris - Institut Polytechnique de Paris

Laboratoire de Météorologie Dynamique - IPSL

Projet de Recherche (PRe)

Spécialité : *Intelligence Artificielle*
Année scolaire : *2025*

Caractérisation des précipitations Antarctiques : étude de cas, méthodes statistiques de croisement de données de différents instruments de télédétection depuis la surface

DOCUMENT NON CONFIDENTIEL Libre Accès

Auteur : Maxime COSTE
Promotion : 2026
Tuteur ENSTA Paris : Laure GIOVANGIGLI
Tuteur organisme d'accueil : Christophe GENTHON
Stage effectué du 28/05/2025 au 22/08/2025
Organisme d'accueil : Laboratoire de Météorologie Dynamique
Adresse : 4 Place Jussieu, 75005 Paris

Attestation de confidentialité / non-confidentialité

A remplir par l'entreprise d'accueil et à remettre à la bibliothèque ENSTA

Rapport de Projet de recherche (PRe)

Par la présente, je soussigné Monsieur[1] GENTHON, Christophe

Employé(e) en tant que directeur de recherches

dans la société CNRS, Laboratoire de Météorologie Dynamique, 4 Place Jussieu, 75252 Paris Cedex 05

Atteste sur l'honneur que LES DONNÉES contenues dans le rapport de Monsieur COSTE Maxime **SONT:**

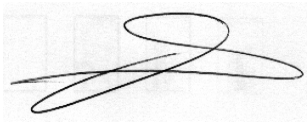
X NON CONFIDENTIELLES

X L'entreprise autorise une **mise en ligne du rapport de stage.** Dans ce cas, seules les métadonnées (titre, résumé mots-clés, entreprise, noms de l'élève et des encadrants...) sont accessibles à tous. Le rapport lui-même n'est accessible qu'aux personnes de la communauté ENSTA. L'étudiant doit déposer son rapport sur le site des archives institutionnelles de l'école BibNum en paramétrant « Accès libre » **- https://bibnum.ensta.fr.**

Fait à Paris

Le 19 Août 2025

Signature et cachet



Cette attestation doit être déposée sur l'espace dédié dans le moodle IP Paris : https://moodle.ip-paris.fr/mod/assign/view.php?id=162720
Si le rapport est confidentiel, celui-ci ne doit pas être transmis à la bibliothèque de l'ENSTA et sera archivé par l'entreprise d'accueil.

1 Rayer la mention inutile.

Résumé

Le stage que j'ai effectué dans le Laboratoire de Météorologie Dynamique (LMD) se décompose en 3 parties. Elles correspondent aux 3 programmes informatiques que j'ai développés en Python dans le but d'optimiser l'observation du climat en Antarctique. Ils se basent sur les données de 2 instruments de mesure installés dans la station Dumont D'Urville en terre d'Alédie : le Micro Rain Radar de la société METEK (noté MRR dans la suite du rapport), qui mesure la quantité de précipitation dans l'atmosphère à une altitude et un moment donnés, et le Célomètre de la société VAISALA (noté CL31 dans la suite du rapport) dont la fonction première est de détecter la présence et mesurer l'altitude de la base des nuages.
Le premier programme a pour objectif de rechercher un évènement de précipitation intense, selon des critères présentant un intérêt d'étude pour les chercheurs du projet AWACA (Atmospheric WAter Cycle over Antarctica, projet ERC Synergie porté par Christophe GENTHON), en se basant sur les données du MRR. Il ne se base pas sur le développement d'un algorithme de machine learning, contrairement aux 2 programmes suivants.
La deuxième partie de ce stage consiste, en se basant sur les données du CL31, à détecter la présence de précipitation dans l'atmosphère, alors que cet instrument n'est pas initialement prévu à cet effet.
Pour finir, le dernier programme a pour ambition d'estimer la quantité de précipitation à partir de cet instrument. Ces 2 algorithmes travaillent en synergie car le premier va détecter la présence de précipitation et le second va en estimer la quantité.
L'objectif de cette étude, est de mesurer les précipitations en Antarctique à moindres coût et d'étendre les mesures de précipitations par le MRR en dessous de 300m d'altitude, zone aveugle pour le radar mais pas pour le CL31.

Mots-clés

Climat, Antarctique, Célomètre, Machine Learning, Random Forest, GradientBoosting, Réseau de neurones, Précipitations, estimations, quantifications, prédictions

Abstract

The internship I completed at the Dynamic Meteorology Laboratory (LMD) is divided into three parts. These correspond to the three computer programs I developed to optimize climate observation in Antarctica. They rely on data from two measurement instruments installed at the Dumont D'Urville station in Adélie Land : the Micro Rain Radar from the METEK company (referred to as MRR in the rest of the report), which measures the amount of precipitation in the atmosphere at a given altitude and time, and the Célometer from the VAISALA company (referred to as CL31 in the rest of the report), whose primary function is to detect the presence and measure the altitude of the cloud base height.

The first program aims to identify intense precipitation events based on criteria of interest for researchers in the AWACA project (Atmospheric WAter Cycle over Antarctica, an ERC Synergy project led by Christophe GENTHON), using MRR data. Unlike the two following programs, it does not rely on a machine learning algorithm.

The second part of this internship involves using CL31 data to detect the presence of precipitation in the atmosphere, even though this instrument is not initially designed for this purpose. Finally, the third program aims to estimate the amount of precipitation using this instrument. These two algorithms work in synergy, as the first detects the presence of precipitation, and the second estimates its quantity.

The objective of this study is to measure precipitation in Antarctica at a lower cost and to extend precipitation measurements by the MRR below 300 meters in altitude, a blind zone for the radar but not for the CL31.

Keywords

Climate, Antarctica, Célometer, Machine Learning, Random Forest, Gradient Boosting, Neural Network, Precipitation, Estimations, Quantifications, Predictions

Remerciements

J'aimerais remercier Christophe GENTHON, mon maître de stage, pour la confiance qu'il m'a accordée. Il m'a laissé travailler en autonomie et proposer mes propres idées ce qui fût une façon très constructive de fonctionner pour moi. J'aimerais aussi remercier l'équipe AWACA devant laquelle j'ai eu la chance de faire une présentation de mon travail, merci pour l'intérêt qu'ils y ont porté et leur bienveillance. Je voudrais également remercie Valentin WIENER, avec qui j'ai pu discuter quelques fois et qui m'a donné des conseils et des indications pendant ce stage. Je voudrais aussi remercier Jean-Louis DUFRESNE pour les données du Pluviomètre qu'il a traitées afin que je puisse m'en servir. malheureusement, le manque de temps à fait que je n'ai finalement pas pu exploiter durant ce stage. Merci à l'équipe du SIRTA, notamment Marc-Antoine DROUIN qui m'a aidé à plusieurs reprises avec les données du célomètre.

Table des matières

Chapitre 1

Introduction : Présentation du contexte météorologique et climatique en Antarctique, des appareils de mesures et de leurs données

1.1 Station Dumont D'Urville, Antarctique et climat

1.1.1 Présentation de la station

Les differents appareils de mesures dont les données ont été utilisées durant ce stage ont été installé à la station Dumont D'Urville en terre d'Adélie au niveau de la côte Antarctique (Figure 1.1). Lors d'une expedition en 2015 attaché au programme APRES3 le MRR a été installé et calibré. Le CL31 lui, n'est déployé que depuis 2018.

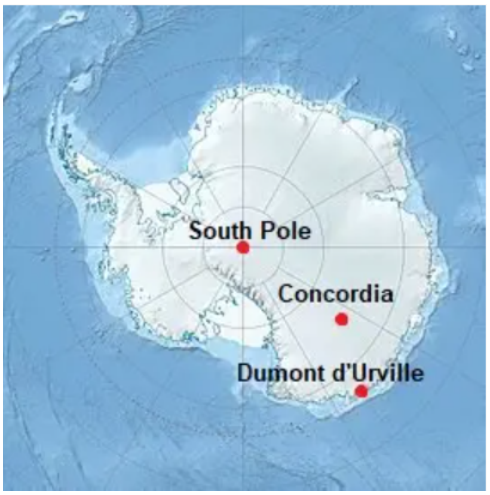


FIGURE 1.1 – Carte de l'Antarctique figurant la position de la station Dumont D'Urville

1.1.2 Climat et phénomènes météorologiques

L'Antarctique représente un endroit unique et assez mal connu sur Terre. Le climat y présente des caractéristiques particulières en raison d'une faible température et d'un climat sec. Des phénomènes comme de la neige soufflé ou des vents catabatiques s'y produisent faisant de l'Antarctique une zone particulère. De plus, les conditions météorologiques difficiles rendent l'accès compliqué et l'installation de stations météo difficile et coûteuse. La mesure des précipitations solides est complexe, elle l'est encore plus en Antarctique.

1.2 Appareils de mesure : Le Célomètre et le Micro Rain Radar

Célomètre

Le CL31 est un appareil de télédetection active de type LIDAR qui émet vers le haut, un rayonnement de longueur d'onde $\lambda = 910\,\text{nm}$ à travers l'atmosphère, puis collecte le signal rétrodiffusé dans sa direction par les différents constituants de l'atmosphère (molécule d'air, d'eau solide et liquide, aérosols ...)(Figure 1.2). Ensuite grâce à un algorithme d'analyse du signal fournis par VAISALA, l'entreprise qui fabrique le célomètre CL31 utilisé lors de ce stage, l'appareil est capable de fournir une mesure de l'altitude de la base des nuages et de la visibilité verticale. Le CL31 collecte des données toutes les 30 secondes, depuis la surface jusqu'à 7600 mètres avec une résolution verticale de 10 mètres. En théorie, il est capable de détecter jusqu'à 3 couches de nuages. En général seule la couche la plus basse est correctement renseignée, l'information des deuxième et troisième couches n'a pas été utilisée lors de ce stage.

FIGURE 1.2 – Célomètre CL31 VAISALA installé à Dumont D'Urville

1.2.1 Données du Célomètre

Le signal 2D (temps,altitude, et intensité en couleur)(Figure 1.3) rétrodiffusé est riche en information, la résolution temporelle de l'appareil est de 30 secondes, et la résolution verticale est de 10 mètres. L'exemple détaillé figure 1.3 est pour la journée du 4 décembre 2022. On y voit une zone d'apparence hachurée entre 10h et 18h, et une base nuageuse entre 1500 et 2000 mètres sur l'ensemble de la journée. Chaque point en rose fuchsia représente la base d'un nuage, calculée par le célomètre à partir d'un algorithme sur

l'intensité de la rétrodiffusion. La zone de hachures, nous allons le voir dans les chapitres 3 et 4 peut-être relié à la présence de précipitations sous un nuage.

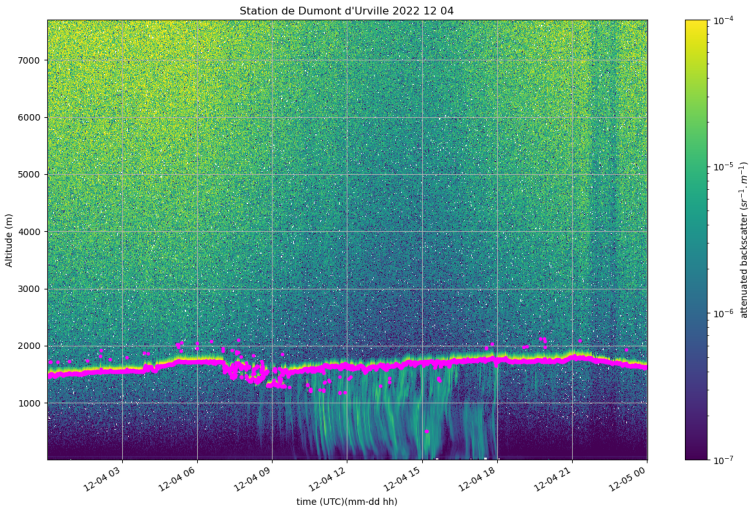


FIGURE 1.3 – Intensité du signal rétrodiffusé du célomètre.Journée du 04/12/2022. Les points fushia indiquent les bases des nuages. Noter la zone d'apparence hachurée sous les nuages interprétée comme la présence de précipitation.

1.2.2 MRR

Un MRR (Micro Rain Radar, voire figure 1.4) est un autre instrument de télédétection active permettant notamment d'estimer la distribution en taille et en densité des gouttelettes d'eau liquide présentes dans l'atmosphère, ainsi que la vitesse de chute. Le traitement des données de celui de Dumont d'Urville, de marque METEK à 24GHz, a dû être adapté afin d'observer des précipitations solides. En effet, la mesure du MRR pour la densité de précipitations se base en partie sur l'effet Doppler, or les conditions météorologiques en Antarctique font qu'il y a principalement des particules d'eau solide, qui ont une vitesse de chute plus faible que celle des particules d'eau liquide.

FIGURE 1.4 – Micro Rain Radar (MRR) METEK installé à Dumont D'Urville

1.2.3 Données du MRR

L'intensité des précipitations neigeuses S du MRR est reliée au profil de réflectivité Z par la relation :

$$Z = A \times S^B$$

où les paramètres A et B sont ajustés empiriquement par croisement avec les données d'autres appareils qui n'ont pas besoin de calibration. Dans notre cas, les facteurs A et B sont ajustés à respectivement 43.3 et 0.88 (Wiener et al. 2025). En pratique nous allons nous intéresser uniquement au 'SnowfallRate' c'est-à-dire l'intensité des précipitations mesurées et calculées d'après la relation ci-dessus par le MRR. Le MRR à une résolution temporelle de 1 minute et une résolution verticale de 100 mètres. Les données s'étendent de 300 à 3100 mètres, la zone en dessous de 300 mètres est une 'blind zone' pour le MRR à cause des interaction d'onde avec la surface. Un exemple des données du MRR est fourni Figure 1.5.

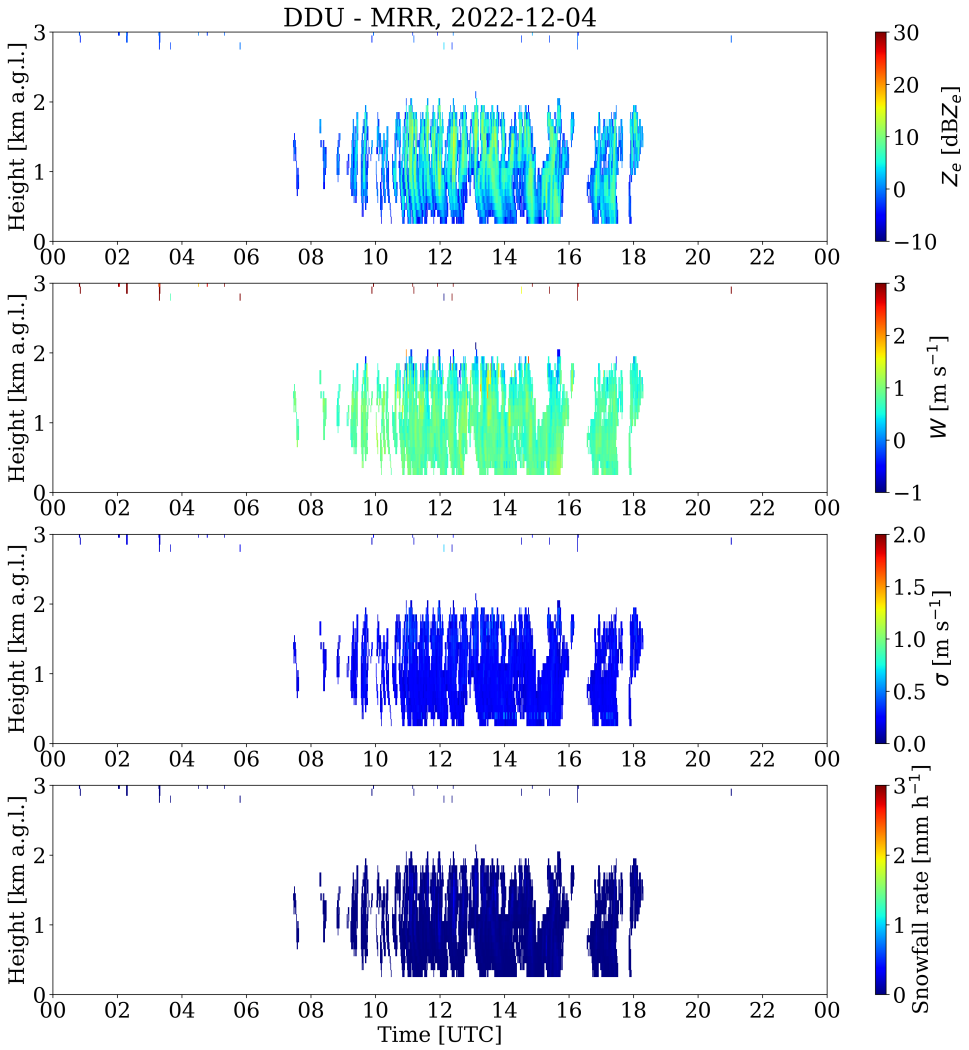


FIGURE 1.5 – Données du MRR sur 4 graphiques, de haut en bas : la réfléctivité Ze, la vitesse de chute W, la vitesse horizontal σ, l'intensité des précipitations. Journée du 04/12/2022

Chapitre 2

Détéction d'un évènement de précipitation intense

2.1 Présentation de l'évènement et motivations

2.1.1 Présentation

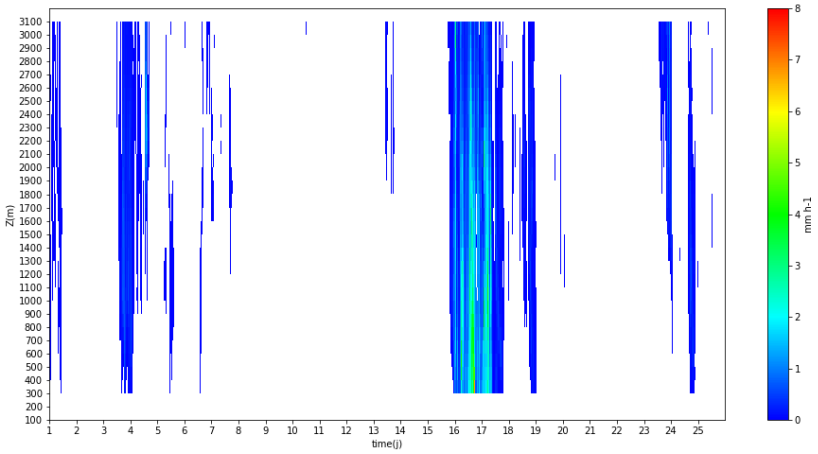


FIGURE 2.1 – Profil de la quantité de précipitation en fonction de l'altitude extrait du MRR(Micro Rain Radar), mois de Février 2025

La Figure 2.1 est un profil de précipitation issue du MRR sur le mois de février 2025. L'évènement d'intérêt se produit entre le 16 et le 18 février. Cet évènement a été observé par une variété d'instruments déployés par le projet AWACA au cours de la dernière saison australe. Je suis missionné pour évaluer grâce aux longues séries de données du MRR et du CL31 déployés antérieurement la représentativité statistique dans le temps de cet évènement.

2.1.2 Extraction des caractéristiques

Dans le but d'étudier la représentativité de l'évènement de février 2025 j'ai dû définir les caractéristiques les plus pertinences et représentatives de ce dernier. J'en ai sélectionné 4 principales.

Durée de l'évènement

La durée de l'évènement est le premier critère de filtrage. Comme on peut le constater(Figure 2.1) la durée est d'environ 2 jours (du 16 au 18). Cependant dans l'idée de collecter un maximum d'échantillons le seuil est volontairement élargi et donc la première condition est une durée minimale de 0.75 jours.

Proportion de précipitation

La deuxième condition est une proportion de précipitation supérieure à 75% dans la colonne et sur la durée de l'évènement. En effet on peut observer des 'trous' dans la colonne de précipitation qui va de 300 à 3100 mètres (limite physique du MRR) L'évènement représente en rectangle dont la base est égale à la durée de l'évènement et la hauteur est égale à la portée maximale moins la portée minimale,c'est-à-dire, 2900 mètres répartis sur 29 valeurs. Ainsi il faut qu'une proportion supérieure à 75% de ce rectangle présente des valeurs de précipitation. On détecte la précipitation si le MRR renvoie une valeur positive.

Valeur du maximum

La troisième condition est celle d'une valeur maximale enregistrée supérieure à 1 $mm.h^{-1}$, sachant que dans l'exemple donné la valeur maximale est $7\ mm.h^{-1}$

Précipitation avant et après l'évènement

On observe que l'évènement est séparé par deux périodes de faible précipitation avant et après. D'où le dernier critère qui est dans les 2 jours avant et après l'évènement la présence de précipitation, en suivant la même logique dans le calcul que pour le critère de 75 % de présence pendant l'évènement, doit être inférieure à 25%. C'est-à-dire, que dans les 2 rectangles de base 2 jours et de hauteur, la hauteur maximale capté par le MRR moins la hauteur minimale, il faut qu'il y ait moins de 25% de précipitation détectée par le MRR.

2.2 Résultats et Analyses

2.2.1 Résultats

Les résultats de l'algorithme qui a traité les données du MRR sur 8 ans de 2016 à 2024 sont présentés dans le tableau si après (Figure 2.2). Les R signifient qu'il n'y a pas eu d'évènement ce mois. Si un évènement est détecté le jour du début et le jour de fin sont inscrits : jour du début/ jour de fin.

	2016	2017	2018	2019	2020	2021	2022	2023	2024
Janvier	R	27/28	R	R	R	21/22 27/29	R	R	R
Février	16/18 19/21	R	10/12	R	R	12/14	6/8 16/17	13/14 22/24	20/21
Mars	R	R	10/11	24/25	R	R	R	4/6	4/5
Avril	R	R	R	7/9	R	R	6/7	24/25	R
Mai	1/2 10/11	R	4/5	6/8	6/7	R	22/24	10/11	20/21
Juin	9/10	25/26	R	R	R	13/14	R	R	R
Juillet	11/13	R	4/6	11/13	absent	R	13/14	R	R
Août	R	R	R	R	12/13	R	R	R	R
Septembre	R	2/4	R	19/20	R	13/14	R	22/23	11/12
Octobre	R	R	R	11/13	R	20/22	5/6	R	17/18
Novembre	R	R	R	9/10	22/23	R	5/6	19/20 23/24	R
Décembre	R	R	R	R	R	R	R	R	9/10 16/17

FIGURE 2.2 – Tableau de la représentativité de l'évènement de précipitation intense de 2016 à 2024

La question sous-jacente était de savoir si l'évènement était représentatif, on peut déjà affirmer qu'il ne s'agissait pas d'un évènement isolé. La moyenne annuelle est de 5,7 évènements/ans.
Le mois avec le plus grand nombre d'occurrences de cet évènement est février avec 9 détections en 9 ans et le pire mois est août avec seulement 1 détection en 9 ans.

2.2.2 Conclusion

Selon les critères que je sélectionne pour caractériser un évènement de précipitation observée par les instruments du programme AWACA en février 2025, on constate que ce type d'évènement s'est produit à différentes reprises, au cours des années 2016 à 2024. Il n'est pas très fréquent mais il caractérise de façon assez uniforme les différentes périodes de l'année.

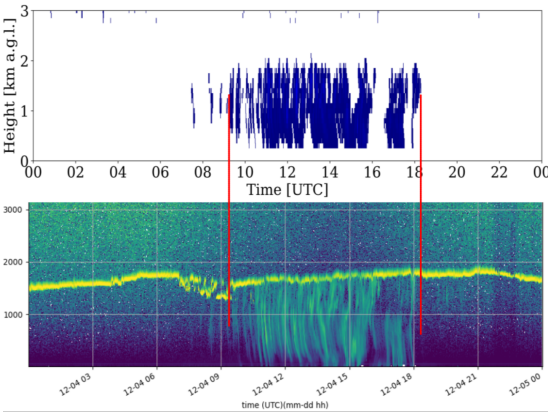
Chapitre 3

Programme de détection de précipitation dans les données du célomètre

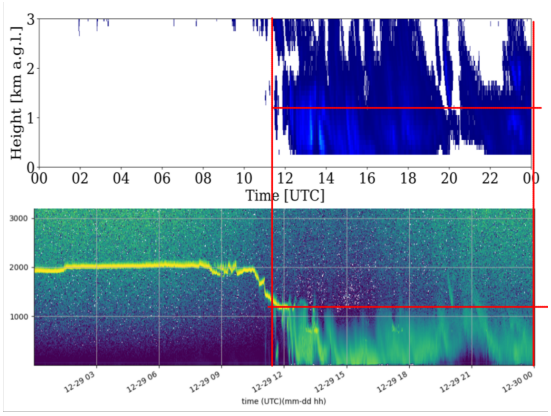
3.1 Données et traitement

3.1.1 Correspondance entre les données du CL31 et celles du MRR, limites du célomètre

Comme vu dans le chapitre 1, les diagrammes d'intensité du signal rétrodiffusé du CL31 peuvent présenter des zones d'apparence hachurée. En croisant ces observations avec celles du MRR nous avons décerné une correspondance entre ces zones et la présence de précipitations au même endroit et au même moment. Quelques exemples illustratifs issus du mois de décembre 2022 sont présents par Figure 3.1.



(a) Correspondance données du CL31 et du MRR. Journée du 04/12/2022



(b) Illustration des limites du CL31 au delà de 1000 mètres. Journée du 29/12/2022

FIGURE 3.1 – Exemples illustratifs de la correspondance des données du CL31 (en bas) et du MRR (en haut) sur des évènements de précipitation avec illustration de la limite de visibilité du CL31 au delà de 1000 mètres environ

Comme on peut l'observer sur la figure 3.1(b) le CL31 possède une limitation au niveau de la hauteur maximale où il peut détecter les précipitations qui sont aux alentours de 1000 mètres. On observe qu'après 12h, le MRR mesure des précipitations au-delà de 1000

mètres (quasiment jusqu'à 3000 mètres, sa portée maximale) cependant la zone hachurée du CL31 s'arrête, pour le même évènement de précipitation, à environ 1000 mètres. Cela vient du fait que le signal rétrodiffusé devient trop atténué en traversant une zone trop épaisse de précipitation. Le faisceau laser est diffusée dans l'atmosphère et une partie de plus en plus faible est captée par le CL31. Ce phénomène est accentué par la présence de particules d'eau solides (et liquides) dans l'atmosphère, c'est-à-dire, de la précipitation. C'est une limite qui doit être prise en compte lors de l'entraînement des algorithmes qui seront présentés dans les sections suivantes.
Pour nuancer encore et préciser le propos sur la précision, un autre exemple qui démontre bien les limitations physiques et donc de performance des algorithmes qui vont suivre, est présenté Figure 3.2.

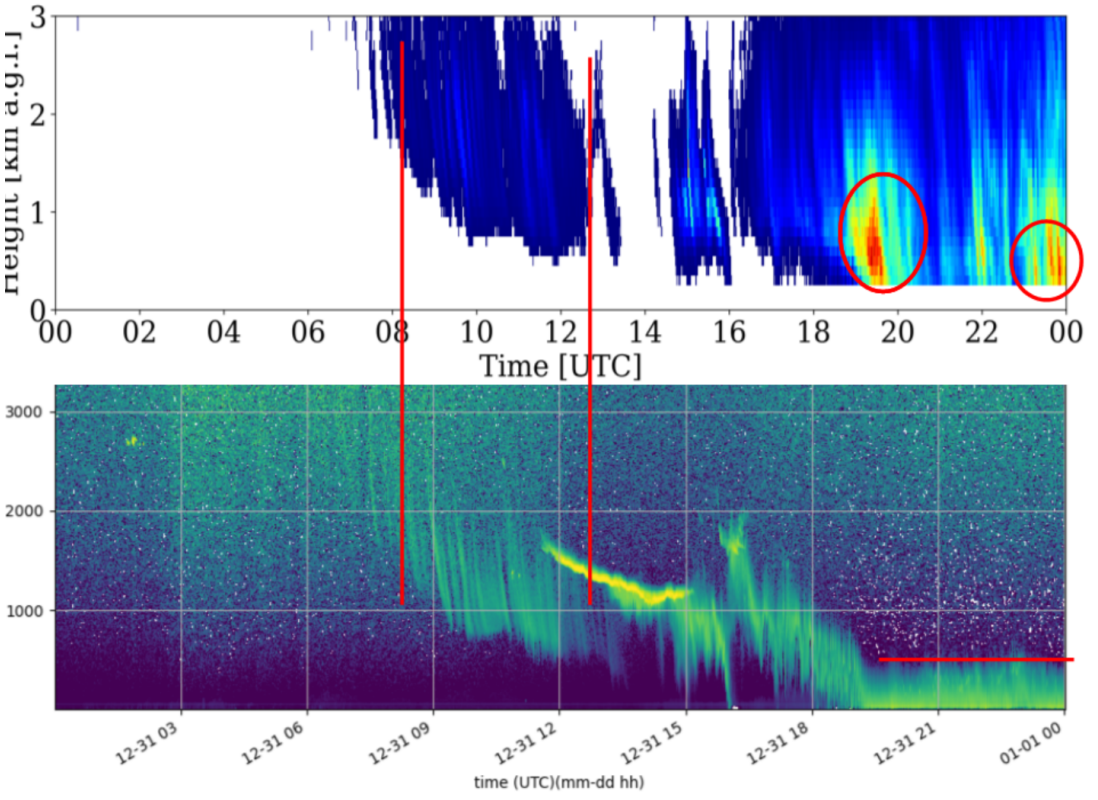


FIGURE 3.2 – Limite de la performance du CL31 illustrée. Journée du 31/12/2022

L'évènement de précipitation entre 8h et 12h30 environ montre bien que le CL31 peut 'voir' au-delà de 1000 mètres dans l'absolu. Le problème semble survenir lorsque l'épaisseur de la précipitation dépasse 1000 mètres. On le voit clairement ici le signal rétrodiffusé caractéristique de la précipitation est entre 800 mètres et 1800 - 1900 mètres, tandis que la précipitation détectée par le MRR commence bien à la même altitude mais s'étend jusqu'à 3000 mètres (portée maximale du MRR). Ainsi ce qu'il faut prendre en compte c'est bien une épaisseur de 1000 mètres de précipitation mais sans réstrictions apparentes sur l'altitude de début de cette épaisseur.

Nuance sur la limite

Il y a cependant une nuance à apporter, comme on pouvait l'imaginer l'épaisseur de la couche de précipitation détectable par le CL31 est reliée à l'intensité de la précipitation. Ainsi en prenant l'évènement entre 18h et 00h du 31/12/2022, toujours sur la figure 3.2. On voit des zones de précipitation particulièrement intenses jusqu'à 3 $\mathrm{mm\,h^{-1}}$ et sur cet évènement le CL31 ne voit pas au-delà de 500 mètres et donc une grande partie de la précipitation reste indétectable. Pour la suite nous prendrons quand même une épaisseur limite de 1000 mètres dans les données d'entraînement car les épisodes de précipitation intense sont rares.

3.1.2 Traitement

Le traitement des données se fait en plusieurs étapes.

Remise à l'échelle

Le CL31 et le MRR ont des caractéristiques différentes, notamment la résolution verticale et temporelle et les portées minimales et maximales. Pour que les données soient cohérentes entre les 2 appareils il faut uniformiser les résolutions.

Table 3.1 – Caractéristiques techniques du CL31 et du MRR

Caractéristique	**CL31**	**MRR**
Résolution verticale	10 m	100 m
Résolution temporelle	30 s	1 min
Portée maximale	7600 m	3100 m
Portée minimale	0 m	300 m

Il a donc fallu moyenner les données du CL31 sur 2 périodes de temps pour obtenir une résolution de 1 min, identique au MRR. Les données au-dessus de 3100 mètres ont aussi dû être retirées du jeu de données et les données sous 300 mètres également.
Pour ce qui est de la résolution verticale, il n'y a pas eu de moyenne. Chaque point du MRR, qui correspond donc à une valeur de précipitation sur une hauteur de 100 mètres, est en correspondance avec un vecteur de taille (1x10) provenant du CL31. La correspondance est conservée puisque la résolution du célomètre est de 10 mètres et donc chaque vecteur couvre bien une hauteur de 100 mètres correspondant à la valeur de précipitation donnée par le MRR. Le première indice du vecteur correspond à l'altitude de la mesure du MRR.

Equilibrage des données

Pour que le modèle apprenne correctement à reconnaître les évènements neigeux il faut que les données d'entraînement soient équilibrées. Le problème c'est qu'en pratique il y a beaucoup plus d'échantillons temporels pour lesquels il n'y a pas de précipitations. Sur l'ensemble du mois de décembre 2022 qui est le mois d'entraînement la proportion de précipitation est de 9%, sur 1 009 863 points seulement 88 187 sont des cas avec précipitation. On se ramène donc à un jeu de données contenant 176 372 échantillons mais avec une répartition à 50/50 des cas positifs et négatifs en supprimant 833 491 échantillons qui ne présentent pas de précipitations.

Application d'un filtre passe bas

Un filtre passe bas a été appliqué aux données du CL31 (Rocadenbosh et al 2020). Le filtrage est réalisé grâce à un filtre numérique FIR d'ordre 50 et avec une fréquence de coupure de 100 Hz et une fenêtre de Hamming dans le but d'obtenir une forte réduction du bruit contenu dans le signal de départ.

Elimination des outliers

Pour limiter l'influence des valeurs extrêmes dans les mesures du célomètre, un autre filtrage basé sur l'*écart interquartile* (Interquartile Range, IQR) a été appliqué. Le principe est le suivant : On calcule d'abord le premier quartile (Q_1, 25^e percentile) et le troisième quartile (Q_3, 75^e percentile) des données du célomètre. Puis, l'écart interquartile est défini comme suit :

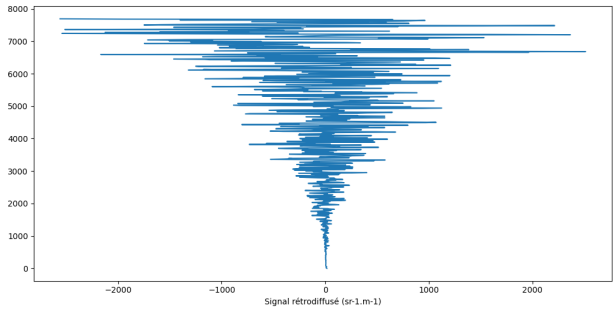
$$\text{IQR} = Q_3 - Q_1$$

Les bornes de détection des valeurs aberrantes sont alors :

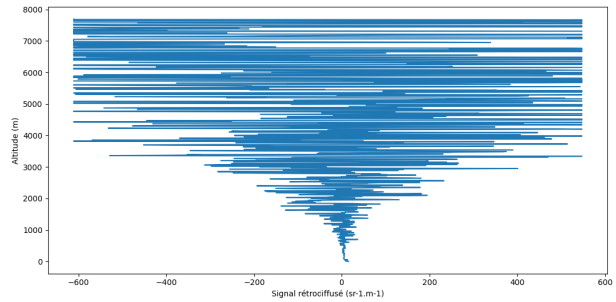
$$\text{Borne inférieure} = Q_1 - 1.5 \times \text{IQR}$$

$$\text{Borne supérieure} = Q_3 + 1.5 \times \text{IQR}$$

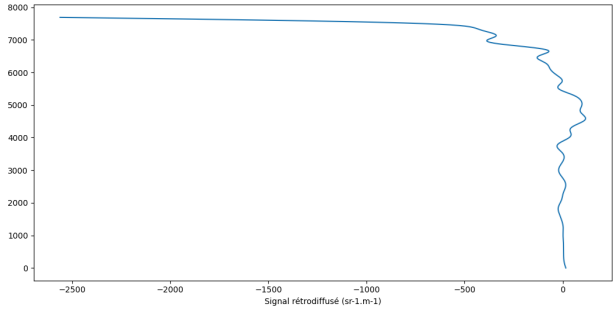
Toutes les valeurs issues du célomètre qui ne sont pas comprises dans cet intervalle sont remplacées par la valeur limite correspondante (*clipping*), le but est de réduire l'impact des mesures aberrantes en conservant l'allure générale des données. La figure ci-dessous présente un profil du CL31 issue du 29/04/2024 et montre visuellement l'effet des différents filtres sur le signal rétrodiffusé.



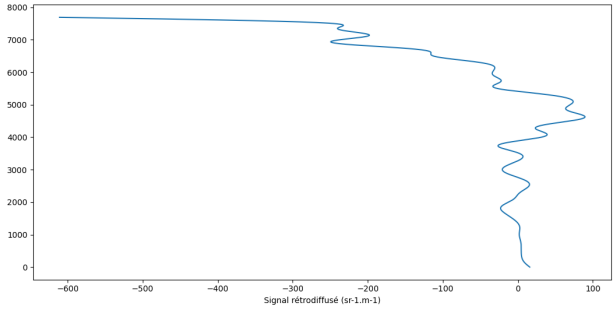
(a) Signal brut



(b) Signal sans les valeurs aberrantes



(c) Signal filtré par le filtre passe bas



(d) Signal sans valeurs aberrantes et filtré

FIGURE 3.3 – Signal issue du CL31 avec différents traitements. Journée du 29/04/2024

La figure 3.3 représente un même profil du CL31 issu de la journée du 29/04/2024 sur lequel est appliqué différentes combinaisons de filtre. On voit que le signal de base est extrêmement bruité, d'où la nécessité d'utiliser un filtre passe bas pour débruiter et accéder à l'information pertinente du signal. Confiner les données dans l'intervalle déterminé par la méthode IQR est surtout utile pour les hautes altitudes où le signal prend les valeurs les plus extrêmes.

3.2 Algorithmes utilisés

Trois algorithmes de machine learning ont été utilisés lors de ce stage. Les deux premiers sont dans ce chapitre et le dernier qui est un réseau de neurones est utilisé dans le chapitre suivant (Yong-Hyu et al 2020).

3.2.1 Random Forest

Le *Random Forest* est une méthode d'*ensemble learning* utilisée en classification et en régression. Elle consiste à entraîner plusieurs arbres de décision indépendants, puis à combiner leurs prédictions.

Le *Random Forest*, repose sur la construction d'arbres selon l'algorithme CART (*Classification and Regression Tree*) en utilisant deux techniques :
(i) Un choix aléatoire des variables pour réduire la corrélation entre les arbres,
(ii) l'apprentissage de chaque arbre sur un échantillon tiré avec remise (*bagging*).

Deux paramètres influencent principalement les performances : le nombre d'arbres et la profondeur (= le nombre de noeuds entre la racine et une feuille) maximale autorisée. Un nombre élevé d'arbres améliore la généralisation mais augmente le temps d'entraînement, tandis qu'une profondeur trop faible peut provoquer : un *underfitting*, c'est-à-dire, un sous-apprentissage, le modèle n'a pas bien appris les relations importantes. Une profondeur trop grande un *overfitting*, c'est-à-dire, un sur-apprentissage, le modèle s'adapte aux données d'entraînement et n'est donc pas capable de généraliser.

Dans le cadre de ce stage, le nombre d'arbres a été fixé à 500 et aucune limite n'a été imposée à la profondeur.

3.2.2 Gradient Boosting

Le Gradient Boosting est une méthode d'ensemble utilisée en régression et classification. Elle construit plusieurs arbres de décision de manière séquentielle, chaque arbre corrigeant les erreurs du précédent. Les principaux paramètres sont le nombre d'arbres, la profondeur maximale des arbres et le taux d'apprentissage (learning rate). Un nombre trop élevé d'arbres ou une profondeur trop grande peuvent entraîner un sur-apprentissage (overfitting). Les deux méthodes ainsi que leurs résultats sont similaires.

3.3 Données entraînement

3.4 Résultats et hyperparamètres

3.4.1 Choix des données d'entraînement

Dans cette section nous explorons les résultats et les performances que les algorithmes présentés ont obtenus. Sont présentes dans le tableau ci-dessous l'importance de chaque caractéristique par rapport à leur pertinence dans l'apprentissage du modèle.

Importance d'une variable Dans un modèle de machine learning, l'importance d'une variable (Caractéristique) permet de quantifier sa contribution à la prédiction. L'importance d'une caractéristique est calculée en moyenne sur tous les arbres de la forêt : chaque

fois qu'une variable est utilisée pour scinder un noeud, on mesure la diminution de l'impureté qu'elle apporte. La somme de ces diminutions sur tous les arbres est normalisée pour donner un score relatif, qui indique quelles variables sont les plus pertinentes dans la prédiction.

TABLE 3.2 – Importance des caractéristiques retenues pour l'entraînement

Caractéristique	**Importance (%)**	**Signification**
Somme des carrés	27.60	Racine carrée de la somme des carrés des valeurs du profil
Symétrie	1.48	Coefficient d'asymétrie : mesure de la dissymétrie d'une distribution. Valeur $= 0$: distribution symétrique (ex. loi normale centrée), < 0 : plus de valeurs extrêmes petites, > 0 : plus de valeurs extrêmes grandes.
Autocorrélation	12.71	Autocorrélation au pas 1
Slope	2.60	Pente de la tendance linéaire globale du profil
Minimum	31.23	Valeur minimale du profil
Maximum	20.77	Valeur maximale du profil
Gradient moyen	3.65	Moyenne du gradient du profil

On voit que globalement chaque caractéristique apporte une information pertinente pour la détéction de la précipitation. Il est a noter que la variable *Symétrie* présente une importance de moins de 2% , ce qui relativement faible par rapport aux autres, mais le choix a été fait de ne supprimer que les caractéristiques qui présentent une importance $<$ 1% . A noter que l'importance à été calculé à partir du modèle entraîné par les caractéristiques uniquement(c'est-à-dire avec les variables du tableau sans entraîner avec le signal en lui-même). Nous verrons les résultats du modèle RandomForest en fonction du type de donnée que l'on donne au modèle. Les résultats en fontion du type de données seront présentés et discutés dans la section *Resultats* de ce chapitre.

3.4.2 Hyperparamètres

Les hyperparamètres retenus pour ce stage sont présenté dans la table 3.3.

TABLE 3.3 – Hyperparamètre des algorithmes RandomForest et GradientBoosting

Hyperparamètre	**RandomForest**	**GradientBoosting**
n_estimators	500	500
Learning Rate	None	0.1
Min_samples_leaf	7	None
Max depth	None	3

— **n_estimators** : nombre d'arbres de l'ensemble, fixé à 500 pour garantir la stabilité des prédictions.
— **min_samples_leaf** : taille minimale d'une feuille ; pour Random Forest (7) afin de limiter la complexité des arbres et réduire le surapprentissage.
— **max_depth** : profondeur maximale d'un arbre ; pour Gradient Boosting (3) pour que les arbres restent faibles, favorisant la régularisation.
— **learning_rate** : influence de chaque arbre dans le boosting (0.1), permettant une convergence progressive et une meilleure généralisation.

Le choix des hyperparamètres ne suit pas une méthodologie scientifique très poussée, l'idée avec *max_deph*, *learning rate* et *min_samples_leaf* est de réduire le surapprentissage du modèle. Le but est d'équilibrer la complexité du modèle avec *n_estimators* et sa stabilité avec les autres paramètres.

3.4.3 Coefficient de mesure de performance

La question se pose de savoir comment on mesure la performance du modèle une fois entraîné. Il existe plusieurs coefficients très largement utilisés et répendus qui sont : accuracy, précision, recall et F1-score. Dont nous allons détailler la signification dans ci-dessous.

Dans une tâche de classification binaire, on définit quatre grandeurs de base à partir de la matrice de confusion :

— TP (True Positives) : nombre d'exemples positifs correctement prédits comme positifs,
— TN (True Negatives) : nombre d'exemples négatifs correctement prédits comme négatifs,
— FP (False Positives) : nombre d'exemples négatifs incorrectement prédits comme positifs,
— FN (False Negatives) : nombre d'exemples positifs incorrectement prédits comme négatifs.

À partir de ces quantités, on définit les coefficients de performance :

Précision (Precision).

$$\text{Précision} = \frac{TP}{TP + FP}$$

Elle mesure la proportion de prédictions positives qui sont effectivement correctes.

Rappel (Recall).

$$\text{Recall} = \frac{TP}{TP + FN}$$

Il indique la proportion d'exemples positifs correctement identifiés par le modèle.

Exactitude (Accuracy).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

C'est la proportion totale de prédictions correctes (positives et négatives) par rapport au nombre total d'exemples.

F1-score.

$$F_1 = 2 \cdot \frac{\text{Précision} \cdot \text{Recall}}{\text{Précision} + \text{Recall}}$$

Il s'agit de la moyenne harmonique entre la précision et le rappel, utile lorsque l'on souhaite un compromis entre ces deux métriques.

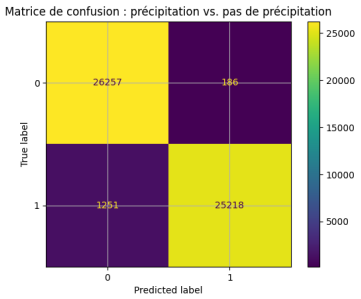
3.4.4 Resultats

RandomForest

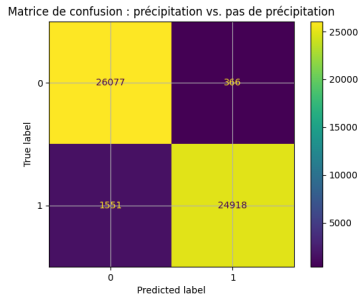
TABLE 3.4 – Résultats du RandomForest selon le type de données d'entraînement

Coefficient de performance	Signal filtré	Caractéristiques	Signal filtré + Caractéristiques
F1-score classe 0	97 %	96 %	97 %
F1-score classe 1	97 %	96 %	97 %
Précision classe 0	95 %	94 %	95 %
Précision classe 1	99 %	99 %	99 %
Recall classe 0	99 %	99 %	99 %
Recall classe 1	95 %	94 %	95 %
Accuracy	97 %	96 %	97 %

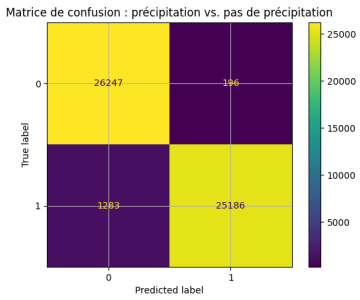
La classe 0 indique qu'il n'y a pas de précipitations et la classe 1 indique la présence de précipitations indépendamment de sa quantité. On voit à travers ces résultats que les meilleurs performances sont obtenus quand le modèle prend en donnée d'entrée le signal filtré concaténé avec les caractéristiques comme décris dans les sections précédentes. Il est à signaler que les résultats sont très proches les uns des autres, on parle maximum 5 points de différence ce qui est peu. Le but étant de coder le modèle le plus performant possible il était cependant intéressant de comparer les résultats en fonction de ce qu'on entraîne. Les matrices de confusion sont disponibles figure 3.4.



(a) Caractéristique et Signal



(b) Caractéristiques seulement



(c) Signal seulement

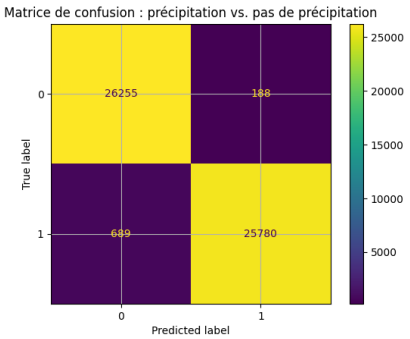
FIGURE 3.4 – Matrices de confusion issues de l'algorithme de RandomForest en fonction des données d'entrée : Caractéristiques de la table 3.2, Signal, ou les deux concaténés

GradientBoosting

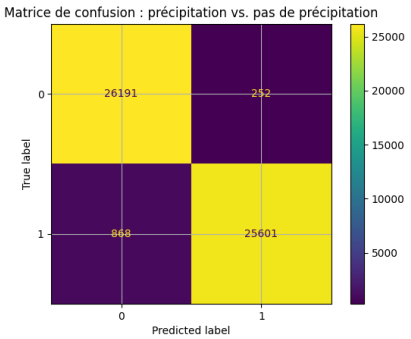
TABLE 3.5 – Résultats du GradientBoosting selon le type de données d'entraînement

Coefficient de performance	**Signal filtré**	**Caractéristiques**	**Signal filtré + Caractéristiques**
F1-score classe 0	98 %	98 %	98 %
F1-score classe 1	98 %	98 %	98 %
Précision classe 0	97 %	97 %	97 %
Précision classe 1	99 %	99 %	99 %
Recall classe 0	99 %	99 %	99 %
Recall classe 1	97 %	97 %	97 %
Accuracy	98 %	98 %	98 %

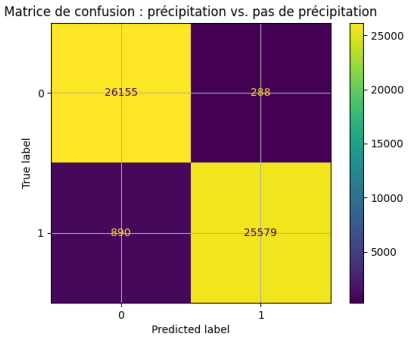
Les résultats de la Table soulignent le fait que les deux algorithmes ont des performances très similaire. On remarque cependant que l'algorithme de GradientBoosting a des résultats un peu meilleurs. La raison doit probablement venir des hyperparamètres qui doivent être plus optimisés pour cet algorithme. Ce qui est significatif c'est le traitement des données en amont, c'est-à-dire, l'application du filtre passe bas et la coupure des valeurs aberrantes. Les bons résultats montrent que les modèles ont correctement appris les relations importantes pour détecter les précipitations sans sur-apprendre. Les matrices de confusion sont disponibles figure 3.4.



(a) Caractéristique et Signal



(b) Caractéristiques seulement



(c) Signal seulement

FIGURE 3.5 – Matrices de confusion issues de l'algorithme de GradientBoosting en fonction des données d'entrée : Caractéristiques de la table 3.2, Signal, ou les deux concaténés

3.4.5 Test sur un mois inconnu

Pour tester la robustesse de ces algorithmes de détection de précipitation on les teste sur un mois inédit. Dans cette section nous allons donc tester l'algorithme de Gradient-Boosting en utilisant le signal du CL31 concatété avec les caractéristiques de la Table 3.2, car c'est cette configuration qui a donnée les meilleurs résultats comme le montre les matrices de confusions et les Table 3.4 et 3.3. Tous les traitements décrits dans ce chapitre,c'est-à-dire, l'application d'un filtre passe bas FIR a la fréquence de coupure de 100Hz, la suppression des valeurs aberrantes, et le rééchantillonage pour avoir une proportion égale de cas avec et sans précipitations, ont été appliqués sur ce mois de test qui est décembre 2024. Un total de 68502 échatillons ont donc été prédits par le modèle de GradientBoosting qui a été lui entraîné comme dans la section précédente, sur les données du moi de décembre 2022. Les résultats sont présentés dans la table 3.6 et la matrice de confusion figure 3.5.

TABLE 3.6 – Résultats des prédictions de l'algorithme de GradientBoosting sur un mois inédit (Décembre 2024) avec comme données d'entrée la concaténation entre le signal et les caractéristiques de la Table 3.2

Coefficient de performance	**Valeurs**
F1-score classe 0	96 %
F1-score classe 1	96 %
Précision classe 0	93 %
Précision classe 1	99 %
Recall classe 0	99 %
Recall classe 1	93 %
Accuracy	96 %

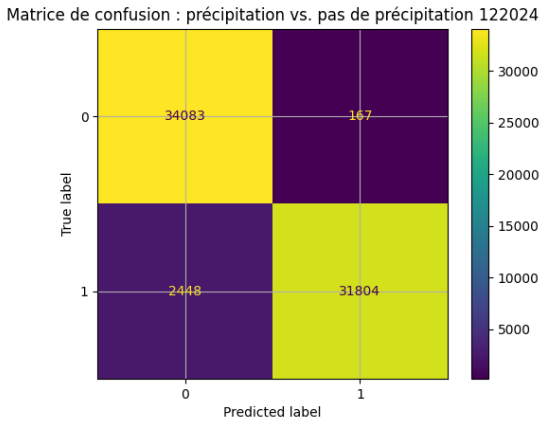


FIGURE 3.6 – Matrice de confusion des prédictions du mois de décembre 2024 en utilisant l'algorithme de GradientBoosting entraîné sur les données du mois de décembre 2022, avec pour données d'entrée la concaténation du signal et les caractéristiques de la table 3.2

Globalement on observe une légère baisse de performance du modèle, la classe 1, c'est-à-dire, la classe des précipitations est moins bien prédite que la classe 0. La baisse de la précision de la prédiction est a relativiser, on garde une accuracy de 96%. Le fait que

les données soient "inédites" pour le modèle car elles proviennent d'un autre mois que le mois d'entraînement et que les performances restent très élevées, montre un très léger sur-apprentissage du modèle mais qui globalement arrive à assez bien généraliser d'un mois à l'autre.

3.5 Conclusion

L'utilisation de modèle de machine learning pour la classification binaire : précipitation ou absence de précipitation, est une réussite. Les résultats obtenus sont proches de 100 % de précision ce qui est très satisfaisant. La modification des hyperparamètres n'a pas eu qu'une importance assez faible empiriquement. Les différences de performance entre GradientBoosting et RandomForest viennent certainement d'une meilleure optimisation de ces hyperparamètres dans un cas plus que dans l'autre.

Chapitre 4

Estimation de la quantité de précipitation

4.1 Données et Traitement

Dans le cadre de ce dernier chapitre l'ambition est plus grande : entraîner un algorithme de machine learning pour estimer la quantité de précipitation en se basant sur les données du CL31(qui, on le rappelle, n'est pas conçu pour cette fonction).
Pour cela nous utilisons un jeu de données issu de l'ensemble de l'année 2024. Cependant une proportion non négligeable de jours était corrompue, ainsi par soucis de précision et bonne correspondance entre les données du CL31 et du MRR, chaque jour avec un problème (ex : manque d'une donnée) a été supprimé du jeu de donnée initial.
Le tableau ci-dessous regroupe tous les jours où les fichiers contenants les données du CL31 ou du MRR ont eu un problème et ont dû être supprimer du jeu de donnée d'entraînement.

TABLE 4.1 – Tableau des jours supprimés de l'année 2024 dans les données d'entraînement

Mois	**Numéro des jours supprimés**
Janvier	04 / 08 / 09 / 10 / 06 / 16 / 27 / 28
Février	06 / 14 / 17 / 23 / 24 / 27
Mars	09 / 19 / 20 / 30 / 31
Avril	09 / 15 / 20 / 30
Mai	11 / 14 / 21 / 22
Juin	01 / 10 / 11 / 12 / 22
Juillet	01 / 02 / 08 / 13 / 14 / 24
Août	04 / 15 / 26 / 31
Septembre	06 / 16 / 26 / 28
Octobre	06 / 17 / 25 / 27 / 28 / 29 / 30 / 31
Novembre	01 / 02 / 03 / 06 / 16 / 18 / 19 / 20 / 27
Décembre	08 / 17 / 18 / 19 / 29 / 30 / 31

4.1.1 Traitement

Pour ce qui est du traitement des données, le filtre passe bas, l'élimination des outliers est identique au chapitre précédent.

La différence avec le chapitre précédent est que les données d'entraînement sont issues de la première couche visible du MRR seulement. Nous avons donc un seul vecteur par pas de temps issu du CL31 qui représente l'altitude entre 300 m et 400 m. Cela a été fait dans le but d'avoir des données d'entrée les plus précises possible car l'estimation est beaucoup plus sensible que la classification. Pour eviter le bruit et les perturbations liés aux plus hautes altitudes ce choix a donc été fait.

Suppression des valeurs nulles

Dans ce chapitre les valeurs nulles de précipitation des données d'entraînement ont été supprimées pour permettre au modèle de mieux apprendre à estimer les précipitations. Le but est d'apprendre au réseau de neurones à éstimer la quantité de précipitation une fois que l'algorithme précédent la détecte. Cette dynamique permet au réseau de neurones de se focaliser sur les données qui représentent une valeure positive de précipitation et donc d'améliorer les performances de ce dernier.

Finalement

En prenant en compte la suppression des valeurs nulles et les données manquantes et le fait que l'entraînement ne se fait que sur la première altitude du MRR, on a un total de 51 921 données pour entraîner le réseau de neurones.

4.2 Algorithme utilisé

4.2.1 Artificial Neural Network

L'algorithme que nous tentons de mettre en place dans ce chapitre pour estimer la quantité de précipitation dans l'atmosphere est une tâche plus complexe que la classification binaire du chapitre précédent. Pour ce faire, un Réseau de Neurones Artificiels a été utilisé dans ce cadre.

Fonctionnement et Principe

Un *réseau de neurones artificiels* (Artificial Neural Network, ANN) est un modèle d'apprentissage inspiré de l'organisation biologique du cerveau. Il est constitué de couches de *neurones artificiels*, connectés entre eux par des pondérations ajustables appelées poids synaptiques. Le réseau se compose généralement d'une couche d'entrée, d'un ensemble de couches cachées et d'une couche de sortie. L'apprentissage consiste à optimiser les poids de manière à minimiser une fonction de coût mesurant l'écart entre les prédictions du modèle et les données observées. Cette architecture permet de modéliser des relations complexes et non linéaires, et trouve ainsi des applications dans des domaines variés tels que la classification, la régression ou encore la reconnaissance de motifs.

4.2.2 Entraînement du modèle de réseau de neurones

Le modèle a été entraîné à l'aide de la librairie `FastAI` via la fonction `tabular_learner` du language Python. Nous avons retenu une architecture de réseau de neurones entièrement connecté composée de trois couches cachées contenant respectivement 500, 300 et 100 neurones. Cette profondeur permet de capturer des relations non linéaires complexes entre les variables explicatives tout en conservant un temps d'entraînement raisonnable. Les fonctions d'activation standards de type ReLU ont été utilisées afin de garantir une propagation efficace du gradient et de limiter les problèmes liés au phénomène de *vanishing gradient*.

Transformation de la variable cible. La variable cible y présente une distribution très déséquilibrée avec de nombreuses petites valeurs et quelques valeurs extrêmes. Pour stabiliser la variance et réduire l'influence des outliers, nous avons appliqué une transformation logarithmique $\log(1+y)$. Cette transformation est adaptée dans le cas de variables strictement positives. Afin de contraindre les prédictions à rester dans un domaine réaliste, la sortie du modèle est bornée par l'intervalle $[0, \max(\log(1+y)) + 1]$, spécifié via l'argument `y_range`.

Fonction de coût. Nous avons utilisé la *Huber loss* avec un paramètre $\delta = 0.1$. Cette fonction combine les avantages de l'erreur quadratique moyenne (MSE) et de l'erreur absolue moyenne (MAE) : elle se comporte comme la MSE pour les petites erreurs (stabilité et différentiabilité) et comme la MAE pour les grandes erreurs, ce qui la rend plus robuste aux valeurs aberrantes. Ainsi, elle constitue un compromis efficace dans le contexte de données présentant des outliers.

Stratégie d'optimisation. L'entraînement a été réalisé sur 500 époques en utilisant la politique de taux d'apprentissage *One Cycle*. Cette stratégie fait varier le taux d'apprentissage de façon non monotone, en commençant par une phase d'exploration avec des taux relativement élevés, puis en diminuant vers de petits taux favorisant la convergence. L'intervalle des taux d'apprentissage a été fixé à $[10^{-5}, 10^{-3}]$. Cette approche a montré de bonnes propriétés empiriques en termes de vitesse de convergence et de généralisation.

Régularisation. Pour éviter le surapprentissage, nous avons utilisé un mécanisme d'arrêt prématuré (*Early Stopping*). Si la perte de validation ne s'améliore pas pendant 50 époques consécutives, l'entraînement est automatiquement interrompu. Ce mécanisme permet d'éviter que le modèle ne s'adapte trop spécifiquement aux données d'entraînement.

Métriques de suivi. Deux métriques de performance ont été suivies durant l'entraînement :

- la racine de l'erreur quadratique moyenne (RMSE), qui quantifie l'erreur absolue de prédiction dans l'unité d'origine et permet une interprétation directe,
- le coefficient de détermination R^2, qui mesure la proportion de variance expliquée par le modèle et reflète sa capacité explicative.

Ces deux métriques sont complémentaires : la RMSE donne une idée de l'erreur en valeur absolue, tandis que le R^2 renseigne sur la qualité globale de l'ajustement.

4.3 Entraînement du modèle et résultats

Les résultats obtenus par le réseau de neurones ne sont pas aussi remarquables et montrent une performance inférieure à la simple classification binaire, ce qui est logique compte tenu de la tâche demandé qui est bien plus complexe.
Empiriquement, il a été observé que la dépendance des résultats à la valeur de la fréquence de coupure du filtre FIR passe bas (dont le but premier est de débruiter le signal) est grande.

4.3.1 Choix des données d'entraînement

Comme dans le chapitre précédent, la performance du modèle varie en fonction du choix des données d'entrée. Nous pouvons entraîner sur le signal (filtré), les caractéristiques, ou les deux.

TABLE 4.2 – Importance des caractéristiques retenues pour l'entraînement, cutoff = 1e-6

Caractéristique	**Importance (%)**	**Signification**
Somme des carrés	3.26%	Racine carrée de la somme des carrés des valeurs du profil
Symétrie	16.86%	Coefficient d'asymétrie : mesure de la dissymétrie d'une distribution. Valeur = 0 : distribution symétrique (ex. loi normale centrée), < 0 : plus de valeurs extrêmes petites, > 0 : plus de valeurs extrêmes grandes.
Autocorrélation	12.18%	Autocorrélation au pas 1
Slope	7.79%	Pente de la tendance linéaire globale du profil
Minimum	32.12%	Valeur minimale du profil
Maximum	3.85%	Valeur maximale du profil
Gradient moyen	9.29%	Moyenne du gradient du profil
Heure de la journée	14.65%	Heure de la journée à laquelle la mesure a été faite

La table 4.2 présente l'importance de chaque caractéristique pour le modèle, on remarquera l'ajout de l'heure de la journée à laquelle la mesure a été faite. Cet ajout s'est avéré pertinent puisque l'importance de cette caractéristique est de presque 15 % ce qui est en fait la 3° la plus importante. On remarquera également toujours la pertinence excessive du minimum du profil qui contribue seul à presque un tiers de la prédiction, cette caractéristique joue donc un rôle clé dans l'estimation de la quantité de précipitation.

4.3.2 Présentation des résultats

Les résultats du modèle dépendent grandement du choix de la valeur du cutoff, c'est-à-dire, la fréquence de coupure du filtre passe bas FIR utilisé.

Fc	**Train/Test Split**			**202406**		
	R^2	RMSE	MAE	R^2	RMSE	MAE
10Hz	0.6227	0.4090	0.1646	0.3798	0.7081	0.2791
100Hz	0.6973	0.3664	0.1523	0.5805	0.5824	0.2560
1000Hz	0.6290	0.4056	0.1571	0.4332	0.6769	0.2773

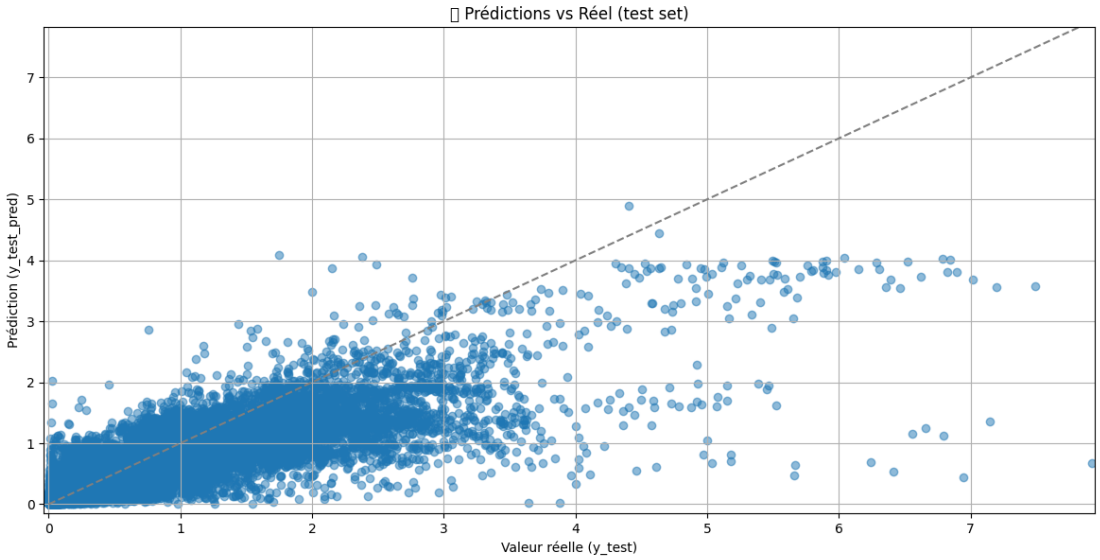
TABLE 4.3 – Performances du modèle selon la fréquence de coupure 'Fc'.

On voit grâce aux données de la Table 4.3, que les meilleurs résultats sont atteint avec une fréquence de coupure du filtre passe bas Fc = 100 Hz. Il doit certainement exister un optimum mais les temps de calcul pour chaque entraînement du modèle étant relativement long, il est complexe de tester un grand nombre de fréquence de coupure différentes.

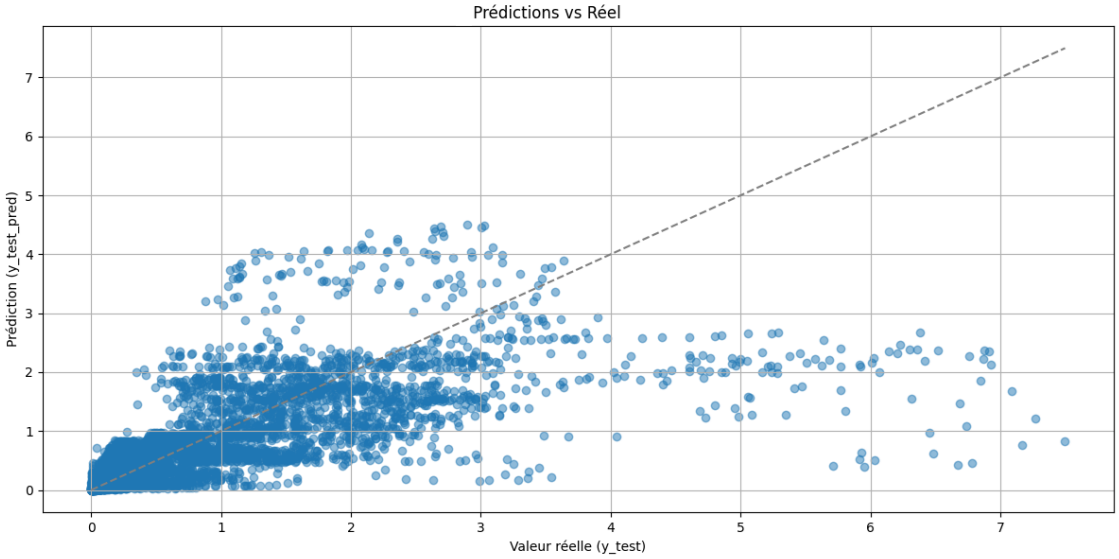
On voit néamoins que l'entraînement du modèle a assez bien fonctionné puisque sur les données de test, c'est-à-dire, les données issues de la première altitude mesurée par le MRR (à 300 mètres) mais qui ont été séparées du jeu de données d'entraînement à raison d'un ratio 80/20. Le R^2 est de presque 0.70, autrement dit le modèle explique près de 70% de la variance des résulats. Tandis que sur le mois de test, qui est le mois de Juin 2024 mais dont les données sont issues de la deuxième altitude mesurée par le MRR (à 400 mètres), on a quand même un R^2 de 0.58, donc près d'un tier de la variance expliquée.

Pour les données de test RMSE = 0.3664 et MAE = 0.1523 donc en moyenne on a un écart entre la prédiction et la valeur réelle de 0.15 et une erreur quadratique moyenne qui reste relativement faible.

Pour le mois de Juin 2024 qui est plus hors-échantillon RMSE = 0.5824 et MAE = 0.2560 on à une augmentation assez significative des erreurs par rapport au split classique. Les prédictions sont donc moins bonnes, mais restent acceptables.



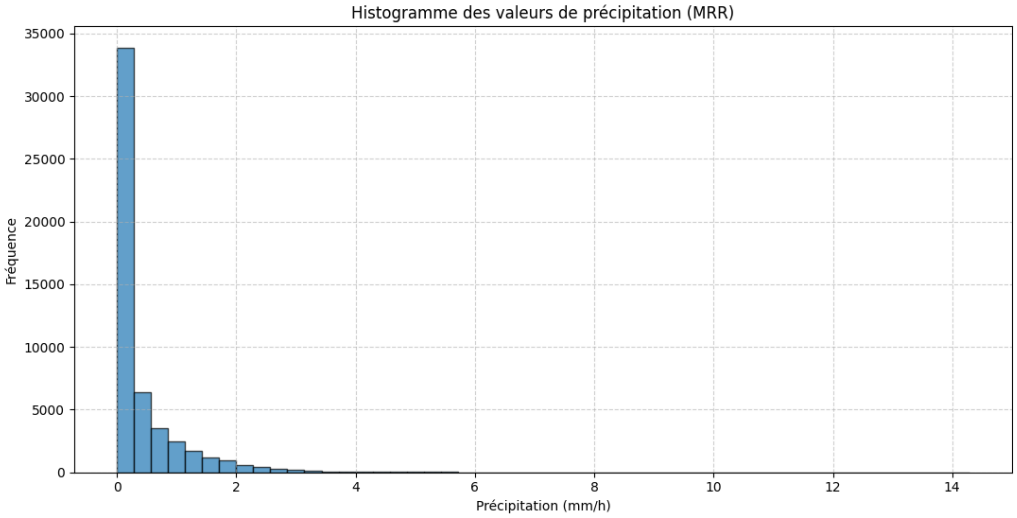
(a) Prédiction sur le split Train/Test



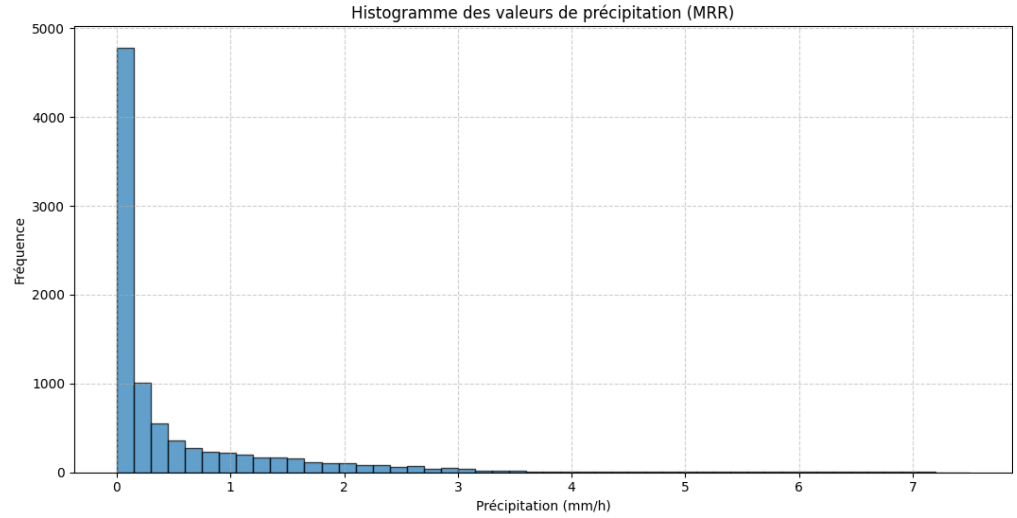
(b) Prédiction mois Juin 2024 inconnue

FIGURE 4.1 – Valeur réelle en fonction de la prédiction sur le split Train/Test et sur les données 'inconnues' du mois de Juin 2024

La figure 4.1 présente les résultats de la prédiction du modèle en fonction de la valeur réelle. Si le modèle était parfait on aurait une droite, on observe que ce n'est pas tout à fait le cas. Les erreurs sont d'autant plus importante que les valeurs à prédire sont grandes, ce qui est peut s'expliquer si on regarde la distribution des quantité de précipitations Figure 4.2.



(a) Année 2024 complète



(b) Juin 2024

FIGURE 4.2 – Histogramme des quantités de précipitations mesurées par le MRR sur l'année 2024 et sur spécifiquement sur Juin 2024

On voit clairement une sur-abondance des faibles valeurs de précipitation Figure 4.2 ce qui oriente le modèle à apprendre mieux sur ces valeurs. Cette forte asymétrie dans la distribution des quantités de précipitations est un des axes d'amélioration du modèle. Une idée pourrait être de suréchantilloner ces fortes valeurs pour avoir une distribution plus homogène.

Chapitre 5

Conclusion

Les différents résultats des programmes développés pendant ce stage sont intéressants scientifiquement. La détection de l'évènement de précipitation particulier du Chapitre 1, va servir aux équipes d'AWACA. Le fait que cet évènement ne soit pas un cas isolé et qu'on en observe plusieurs occurences, grâce à la longueur de la série d'observations des appareils de mesure déployés à Dumont D'Urville, en fait un cas d'étude intéressant pour les scientifiques du projet AWACA.

Concernant l'utilisation d'algorithmes de machine learning dans la détection et l'estimation de la quantité de précipitation en Antarctique, les résultats sont plus qu'encourageants. En effet, l'état de l'art ne montrait que peu d'utilisations des données du CL31 pour détecter et quantifier des précipitations. Il est intéressant en particulier pour l'Antarctique car le déploiement d'un appareil de mesure y est plus compliqué et coûteux qu'ailleurs. Le manque de temps m'a malheureusement empêché d'utiliser mes résultats pour compléter les données du MRR dans sa zone aveugle, en dessous de 300 mètres sur toute la durée de la série temporelle disponible (depuis janvier 2016). Il resterait à faire cette étude en prenant en compte les phénomènes de surfaces qui peuvent se produire en Antarctique comme les vents catabatiques, le brouillard ou la neige soufflé qui pourraient affecter les résultats du modèle.
Les scripts python et l'ensemble du code source utilisé pendant ce stage sont disponibles en ligne en cliquant : Codesource
Assurant le transfert d'information en vue de la poursuite de ces travaux.

Bibliographie

— Yong-Hyu, K., Seung-Hyun, M., &Yourim, Y. *Detection of Precipitation and Fog Using Machine Learning on Backscatter Data from Lidar Ceilometer.* Applied Sciences, 16 Septembre 2020

— Rocadenbosh, F., Barragàn, R., Frasier, S., Waldinger, J., Turner, D., Tanamachi, R., &Dawson, D. *Ceilometer-Based Rain-Rate Estimation : A Case-Study Comparison With S-Band Radar and Disdrometer Retrievals in the Context of VORTEX-SE*, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, VOL. 58, NO. 12, Décembre 2020

— Salman, H A., Kalakech, A., &Steiti, A., *Random Forest Algorithm Overview*, Babylonian Journal of Machine Learning, Vol.2024, p. 69–79, Juin 2024

— Wiener, V., Madeleine J-B., &Genthon, C., *Nuages et Précipitations sur la côte AntarctiqueObservations in situ et Modélisation*, Science Sorbonne Université & Laboratoire de Météorologie Dynamique, Juin 2020

— Roussel, M-L., *Les précipitations au dessus de la calotte Antarctique : une approche conjointe observations et modélisation*, Océan, Atmosphère. Institut Polytechnique de Paris, 2023. Français. NNT : 2023IPPAX061. tel-04540632

— Charrel, J., Genthon, C., &Dufresne J-L., textitÉtude des nuages à Dumont d'Urville en Antarctique et impact sur le rayonnement en surface, Ecole normale supérieure de Paris département des géosciences & Laboratoire de Météorologie Dynamique, Juillet 2022.

— VAISALA, *CL31 User's Guide*, `https://www.manualslib.com/manual/1226537/Vaisala-Cl31.html`

— FastAI documentation, `https://docs.fast.ai/`

— SCIKIT-LEARN `https://scikit-learn.org/stable/user_guide.html`