M1 SOAC
UE: Stage Master 1

Classification des trajectoires automatiques de prévision cyclonique fournies en post-traitement de la prévision d'ensemble IFS pour les cyclones tropicaux

Réalisé par : **Mike PAYET, Étudiant en M1 Sciences de l'Océan, de l'Atmosphère et du Climat**

Sous la tutelle de : **Quoc-Phi DUONG, Ingénieur des travaux de Météorologie**

Et sous la tutelle universitaire du : **Pr. Patrick RAIROUX**

Du 16 mai au 15 juillet 2022

À Météo-France / DIROI LACy
Laboratoire de l'Atmosphère et des Cyclones

REMERCIEMENTS

Tout d'abord, je tiens à adresser mes remerciements à la personne sans qui je n'aurai pas pu passer un stage aussi stimulant : Quoc-Phi DUONG. Merci d'avoir retenu ma candidature et de m'avoir permis de travailler sur un sujet aussi passionnant. J'ai beaucoup apprécié ta pédagogie et notamment le cadre que tu as su mettre. J'ai eu le sentiment d'évoluer en autonomie tout en ayant un fil rouge sur lequel m'accrocher.

J'aimerais également remercier toute l'équipe Cyclone du LACy. Merci à Sylvie Malardel de m'avoir accepté dans son équipe. L'accueil, le cadre et l'ambiance y étaient juste fabuleux. De par les différentes conversations et présentations, j'ai beaucoup appris sur différents sujets, tous aussi passionnants les uns que les autres. Merci aussi à mes collègues de raquettes, Rémi, Soline et Andréa pour ces chouettes moments d'échanges après le travail.

Enfin, j'aimerais remercier le Pr. Patrick Rairoux d'avoir accepté d'être mon tuteur universitaire. Merci de prendre le temps de lire mes minutes de stage.

Table des matières

1 Introduction

Selon le dernier rapport du GIEC [1], les événements extrêmes vont à l'avenir se multiplier et s'intensifier sur l'ensemble du globe. Il en va de même pour les cyclones tropicaux qui, en l'espace de 32 ans, ont augmenté de 8% leur propension à atteindre les stades les plus intenses. [2] Au cours des 50 dernières années, lorsque ces derniers ont traversé une zone habitée par l'homme, on estime qu'ils ont tué près de 800 000 personnes et provoqué de lourds dégâts matériels approximés à 2 000 milliards d'euros. Pendant la saison cyclonique (novembre-avril), le Sud-Ouest de l'Océan Indien est une zone réunissant toutes les conditions à l'apparition de ces systèmes dépressionnaires tropicaux (SDT). La vigilance face à ce phénomène est donc essentielle dans cette zone et il convient d'étendre rapidement notre connaissance sur l'endroit où il fait son apparition (cyclogenèse) ainsi que sur sa trajectoire. Cela permettrait d'avertir au mieux les éventuels territoires touchés pour qu'ils se préparent à la catastrophe et limitent au maximum les dommages. De nombreux outils en instrumentation et modélisation (satellites, radar météo …) sont donc mis en place pour permettre aux météorologues d'observer et de prédire l'évolution des cyclones tropicaux. L'ECMWF (Centre Européen pour les prévisions Météorologiques à moyen terme) est un organisme qui a par exemple mis en place un modèle de prévision numérique ensembliste appelé IFS (Integrating Forecasting System) permettant de prédire l'apparition et les trajectoires des cyclones sur une échelle de temps donné. La quantité d'informations importantes contenue dans ces prévisions reste difficilement interprétable en mode opérationnel où le temps disponible est fortement contraint. Parallèlement, l'émergence de nouveaux outils statistiques dits "d'apprentissage machine" ont rendu possibles la résolution de nombreux problèmes scientifiques [3].

L'objectif de ce stage est donc de s'appuyer sur les prédictions faites par le modèle IFS afin de définir, par le biais de l'apprentissage machine, une méthode capable de classifier les potentielles zones de cyclogenèses ainsi que les trajectoires de ces cyclones sur la période du 10 janvier au 7 février 2022.

2 Contexte Scientifique

2.1 Fonctionnement d'un cyclone tropical

2.1.1 Formation : Cyclogenèse

Pendant la saison cyclonique (Novembre-Avril), le Sud-Ouest de l'Océan Indien, de par les conditions météorologiques présentes, est une zone propice à la cyclogenèse. En effet, pendant l'été austral, la température océanique excède les 26°C sur une profondeur d'une dizaine de mètres. Le cyclone y tire son énergie et sa matière pour son développement et son entretien car il lui faut une énorme quantité d'eau pour qu'il puisse se créer. Il faut que cela soit accompagné de mouvements verticaux importants (atmosphère instable) avec une humidité relative élevée dans les couches moyennes (présence d'amas nuageux). Une dépression initiale doit déjà exister et le vent doit se renforcer sur une ou plusieurs faces afin d'accentuer le mouvement tourbillonnaire. Bien que ces conditions soient présentes, cela ne garantit pas systématiquement la cyclogenèse. [4]

2.1.2 Vie du cyclone et trajectoire

Avant d'atteindre le stade de cyclone tropical, un système dépressionnaire tropical passe par différents stades d'intensité. Le Sud-Ouest de l'Océan Indien répertorie ces systèmes en s'appuyant sur une échelle faite par l'Organisation Mondiale de Météorologie (OMM) (Annexe 1). Celle-ci se distingue de celle de Saffir-Simpson utilisée dans l'Atlantique Nord pour les ouragans.

Un cyclone tropical se déplace en fonction de la circulation des vents en basses couches et moyennes couches. Dans le Sud-Ouest de l'Océan Indien, 3 types de trajectoires prédominent, celles de type parabolique (Firinga en orange sur la figure 1) débutant dans le Nord-Est des Mascareignes pour se terminer vers le Sud-Est. Celles de type zonal allant d'Est en Ouest (Bonita et Eline sur la figure 1). Et enfin, celles de type méridional allant du Nord au Sud (Anacelle sur la figure 1).

Bien que les cyclones puissent suivre un certain motif, les trajectoires sont parfois erratiques comme on peut le voir avec le cyclone Hyacinthe (en pointillé sur la figure 1) qui réalise des boucles lors de sa traversée.

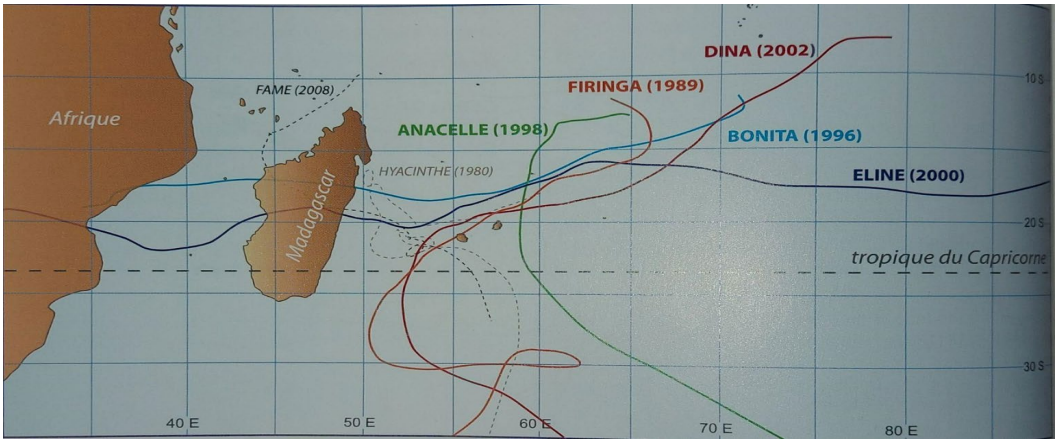


Figure 1 : *Différents types de trajectoire des cyclones dans le Sud-Ouest de l'Océan Indien*

Un cyclone arrive en fin de vie lorsqu'il touche la terre ou une surface d'eau froide. Étant privé de son énergie, il s'affaiblit au fil du temps puis se désagrège.

Les dommages que peuvent causer un cyclone sont si importants que les météorologues développent davantage d'outils capable de prédire l'apparition et la trajectoire de ces phénomènes. Le modèle IFS en fait partie.

2.2 Le modèle IFS

Pour effectuer la prévision à moyen terme des cyclones tropicaux, le Centre Météorologique Régional Spécialisé (CMRS) de La Réunion se base en grande partie sur la prévision d'ensemble basée sur le modèle IFS (Integrating Forecasting System). Ce modèle de prévision numérique du temps global a été développé par le Centre Européen pour les prévisions météorologiques à moyen terme (ECMWF). Nous utilisons plus spécifiquement le modèle ENS (Ensembles-Atmospheric Model) capable de prédire les cyclogenèses et trajectoires de celles-

ci sur les 4 prochaines semaines après son exécution. [5] L'annexe 3 fournit davantage de détails sur ce modèle.

Il existe deux types de prévision : celle déterministe et celle ensembliste. La première base ses calculs sur un état donné de l'atmosphère alors que pour la seconde plusieurs prédictions sont réalisées à l'aide de conditions initiales, états de l'atmosphère et paramétrisations différentes mais plausibles. De ce fait, elle génère un échantillon représentatif des états futurs possibles d'un système dynamique atmosphérique très sensible à la moindre fluctuation. Contrairement à la prévision déterministe, la prévision d'ensemble produit une information plus précise et riche mais cependant souvent difficile à comprendre. Elle est plus souvent utilisée à des échéances allant au-delà de 3 jours où les erreurs de prévisions météorologiques s'amplifient à cause du caractère chaotique de l'atmosphère.

Le modèle IFS est lui un modèle de prévision ensembliste. Les calculs ont été faits sur 51 états et paramétrisations différentes. On définira par la suite chaque état et paramétrisation comme un membre. Pour une même prévision, un membre est capable de prédire plusieurs SDT.

Sur la figure 2, chaque point bleu foncé correspond à une cyclogenèse prévue par un membre au cours de la période de prévision. Chacun d'entre eux est associé à une courbe tracée en bleu cyan représentant la trajectoire du cyclone. On peut voir sur la figure 2 que le modèle produit une quantité importante d'informations parfois difficile à appréhender par un prévisionniste. On n'est par exemple pas à même de voir à quelles dates apparaissent ces points. De plus, il est parfois difficile de distinguer une trajectoire prédominante sur cette carte tant les courbes se mêlent. Bien qu'il y ait 51 membres, notons qu'on trouve généralement plus de 51 points sur une carte de prévision car un membre peut prédire plusieurs SDT.

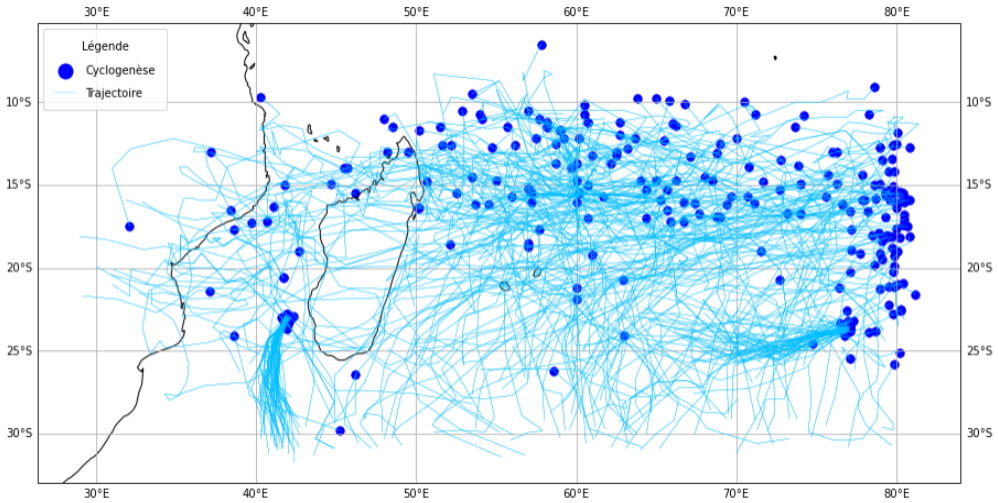


Figure 2 : *Prévision d'ensemble des 51 membres du modèle IFS du 07/02/2022 pour les 4 prochaines semaines*

On peut voir sur la figure 2 que certaines zones présentent une plus grande densité de points. La classification (ou clustering) par le biais de l'apprentissage machine (Machine Learning) semble être une option viable pour faire apparaître les zones de cyclogenèse les plus probables.

2.3 L'apprentissage machine : Méthode de classification

L'apprentissage machine (ou Machine Learning) est une méthode statistique pour construire un modèle de prédiction par apprentissage sur un jeu de données sans connaissance experte.

On distingue deux types de Machine Learning, celui supervisé et celui non supervisé. L'apprentissage supervisé consiste à montrer à la machine des exemples de ce qu'elle doit apprendre alors que l'apprentissage non-supervisé consiste à dire à la machine d'analyser la structure des données (par elle-même) pour délivrer une information précise. Vous trouverez en Annexe 2 un exemple décrivant la distinction entre ces deux types d'apprentissage machine.

L'apprentissage non-supervisé est capable de regrouper, selon un critère de similarité, une grande quantité de données en plusieurs sous-ensembles, c'est de la classification ou clustering en anglais. Chaque sous-ensemble est appelé cluster.

Pour notre étude, on s'appuiera sur l'apprentissage non-supervisé pour faire du clustering des différents SDT. Il existe des dizaines de méthodes de classification ayant des usages propres à chaque situation. On se concentre ici sur trois méthodes de classification parmi les plus couramment utilisées : DB-scan, la classification hiérarchique (Agglomerative Clustering) et enfin K-Means.

L'observation de ces méthodes en fonction de différentes situations (figure 3) montre immédiatement la différence de pertinence des clusters crées en fonction de la distribution des données étudiées.

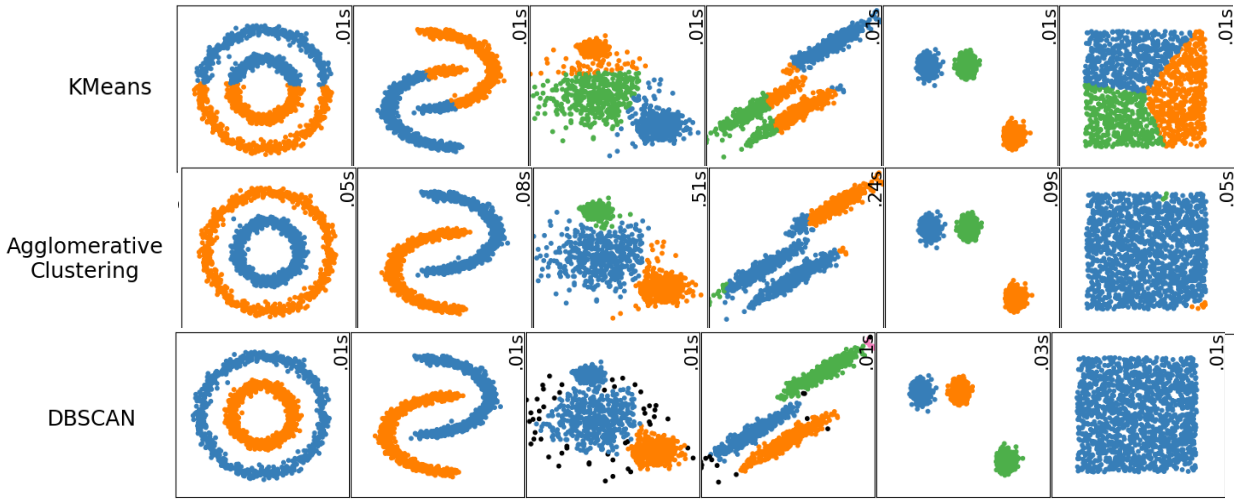


Figure 3 : *Comportement de différentes méthodes de classification en fonction de différentes situations [6]*

Chacune de ses méthodes mesure différemment la similarité entre deux individus (ou points sur la figure 2). Les points les plus proches seront classés ensemble (ici par couleur sur la figure 2).

Ce sont donc des méthodes ayant des usages différents de par leurs différences pour calculer la distance entre les points.

Le tableau 1 compare les caractéristiques des différentes méthodes de clustering pour voir théoriquement laquelle serait la plus appropriée à notre étude. [6]

	K-Means	**DB Scan**	**Agglomerative Clustering**
Type de méthode	Centroïde : chaque cluster représente un vecteur moyen	Densité : les clusters sont réprésentés comme des régions denses connectées	Connectivité : formation de clusters sur l'idée que les objets sont plus liés aux objets proches qu'aux objets éloignés
Usage	Usage général, cluster uniforme, géométrie plate, peu de clusters, cluster convexe et isotropique	Géométrie non-plate, taille inégale des grappes, cluster non convexe	Nombreux clusters clusters, distance non euclidiennte
Métrique utilisé	Distance entre les points	Distance entre les points les plus proches	Toute distance par paire
Fonction de distance	Distance euclidienne	Distance euclidienne	Clustering hiérarchique[1]
Comportement aux valeurs aberrantes	Sensible	Robuste (exclue des clusters les valeurs trop éloignées)	Sensible
Nombre de cluster à prédéfinir	Oui	Non	Oui
Hyper paramètres	N_clusters	Epsilon ; min_samples ;	N_clusters ; linkage

Tableau 1 : *Comparaison des méthodes de clustering K-Means, DB Scan, Agglomerative clustering*

Si on s'intéresse exclusivement à la cyclogenèse, on cherche une méthode capable d'être discriminante et de cibler uniquement les zones où il y a une potentielle apparition de cyclone. Les clusters de cyclogenèse ne seront pas nécessairement convexe et isotropique, cela exclut donc K-Means comme une méthode valable pour l'étude de la cyclogenèse. De plus, la méthode Agglomerative Clustering est sensible aux valeurs aberrantes et aura tendance à former des clusters même avec celles-ci. Cela peut induire des imprécisions sur la zone d'apparition de cyclone. La méthode DB-Scan semble dont la plus intéressante car elle serait plus discriminante et diminuerait ainsi la zone d'incertitude pour le prévisionniste.

Maintenant, si on s'intéresse aux trajectoires des cyclones, on cherche une méthode capable de distinguer les différentes trajectoires prédominantes parmi l'ensemble des trajectoires des cyclogenèse. La méthode K-Means, de par le fait qu'elle moyenne les clusters risque de ne pas discriminer les différentes trajectoires entre-elles. DB-Scan et Agglomerative Clustering semblent être de bonnes méthodes pour cette étude. En effet, DB Scan, de par sa nature discriminante, peut se passer des trajectoires qui dérivent de celles qui prédominent. La

[1] Chaque observation commence dans son propre cluster, et les clusters sont successivement fusionnés ensemble.

méthode Agglomerative Clustering semble être aussi utile pour classifier les trajectoires car au début elle considère tous les points comme un seul et même cluster puis forme plusieurs clusters au fur et à mesure qu'elle trouve des différences.

Notons néanmoins que les méthodes Agglomerative Clustering et K-Means nécessitent de connaître à l'avance le nombre de cluster que l'on souhaite réaliser. C'est un paramètre important qui doit être optimisé. La méthode DB-Scan ne nécessite pas de préciser en entrée le nombre de clusters que l'on souhaite avoir, cependant, il faut bien choisir son epsilon[2] et son min_samples[3]

3 Méthodologie

Après avoir compris l'intérêt de la classification pour traiter toutes les données fournies par le modèle IFS. L'objectif à présent est d'évaluer graphiquement et en se basant sur les cyclones apparus durant la saison 2022, quelle méthode de classification, parmi K-Means DB-Scan et Agglomerative Clustering, est la plus à même de trouver les zones de cyclogenèse ainsi que les meilleures trajectoires comme le ferait un prévisionniste dans son analyse. Pour ce faire, nous utilisons le langage de programmation Python.

Au cours de ce rapport, je mentionnerai des cyclones qui sont apparus sur cette période. Vous trouverez en annexe 4 un tableau indiquant les informations suivantes sur ces cyclones : nom du cyclone, latitude et longitude de la cyclogenèse, date de la cyclogenèse, résumé de la trajectoire. Ce tableau permettra au lecteur de relater sur la concordance des résultats de notre modèle avec ce qui s'est passé réellement.

Présentons brièvement le jeu de données nous ayant permis de réaliser cette étude. Tous les codes sont disponibles en Annexe 5 sur le lien GitHub.

3.1 Création et présentation des données

Les données sur lesquelles nous avons travaillées sont sur la période du 10 janvier au 7 février 2022. Les prévisions sont faites tous les 7 jours. Nous avons donc obtenu des données pour les 10, 17, 24 et 31 janvier 2022 et pour le 7 février 2022.

Pour chacune de ces dates, chaque membre prévoit un certain nombre de SDT et chaque SDT est décrit par plusieurs variables dans un fichier texte. Les variables sont les suivantes : la date, la latitude, la longitude, la vitesse maximale du vent (en nœud) et la pression (hPa) pour chaque échéance de prévision. Le système est décrit toutes les 12 heures et ça permet finalement d'avoir un fichier texte décrivant la trajectoire prévue de ce dernier.

Étant donné que l'on s'intéresse également à la cyclogenèse et pas qu'à la trajectoire, il nous fallait extraire des nouvelles variables, décrites dans le tableau 2, pour obtenir la date de cyclogenèse ainsi que les coordonnées de départ. Un code python nous a permis d'implémenter ces paramètres dans notre jeu de données et ainsi mieux décrire la cyclogenèse d'un SDT.

[2] Distance maximale entre deux échantillons pour que l'un soit considéré comme dans le voisinage
[3] Nombre minimum de points pour former un cluster

Enfin, nous avons écrit une boucle sur Python permettant de lire l'ensemble des fichiers textes contenant les différentes trajectoires de chaque système de chaque membre et nous les avons affichés sur une carte. C'est ce que vous pouvez par exemple voir sur la figure 2.

À la fin, nous avions un jeu de données contenant les variables suivantes :

Variable	Numéro du membre	Numéro SDT dans le membre	Date de cyclogenèse	Latitude cyclogenèse	Longitude cyclogenèse	Durée de vie (en jours)	Pression minimale (en hPa)	Vitesse max vent (en nœud)
Appellation sur Python	Member_number	TC_number_in_member	TC_start_date	TC_start_lat	TC_start_lon	TC_duration	TC_min_PMIN	TC_max_VMAX

Tableau 2 : *Descriptif des variables dans notre jeu de données avec leur appellation sur le code Python*

Ce sont ces variables que nous utiliserons pour entraîner notre modèle de classification. Précisons qu'en fonction des dates, la taille du jeu de données peut varier du fait que le nombre de SDT prédit par chaque membre varie.

Maintenant que nous avons notre jeu de données complets, passons à l'utilisation du modèle de classification sur la cyclogenèse.

3.2 Classification pour la cyclogenèse

Dans cette partie, par soucis de clarté pour le lecteur, les exemples suivants s'appuieront exclusivement sur les prévisions faites le 7 février 2022. L'analyse a cependant été faite pour toutes les dates.

3.2.1 Filtrage des données

L'étape de filtrage des données est importante car elle permet de faire abstraction des données dont nous n'avons pas usage et qui, in fine, perturbent nos résultats.

Chaque prévision prévoit l'apparition des différents SDT sur les 4 prochaines semaines et nous avons des données tous les 7 jours. La dispersion des scénarios possible est si large qu'elle n'est pas réellement exploitable. L'usage d'un filtre permet de se concentrer sur les quinze premiers jours de prévision.

De plus, le modèle crée beaucoup de SDT dont la durée de vie est inférieure à 3 jours sans réalité physique et qui s'apparentent plutôt à du bruit. Nous filtrons donc ces données.

Le fait que nous nous intéressons, pour l'instant, uniquement aux cyclogenèses fait que nous ne souhaitons pas voir apparaître les cyclones qui sont déjà apparus. On peut voir par exemple sur la figure 2, à la longitude 42 et à la latitude 23, un amas de points ayant des trajectoires similaires. Cela correspond à un SDT qui a déjà effectué sa cyclogenèse. Nous souhaitons filtrer ce genre de SDT pour ne pas perturber nos résultats.

Finalement, grâce à cette étape de filtrage, on passe de la figure 2 à la figure 3 ci-dessous où on obtient une carte plus lisible indiquant les différentes zones de cyclogenèses. Le code ayant permis de faire ce filtrage est disponible en Annexe 6.

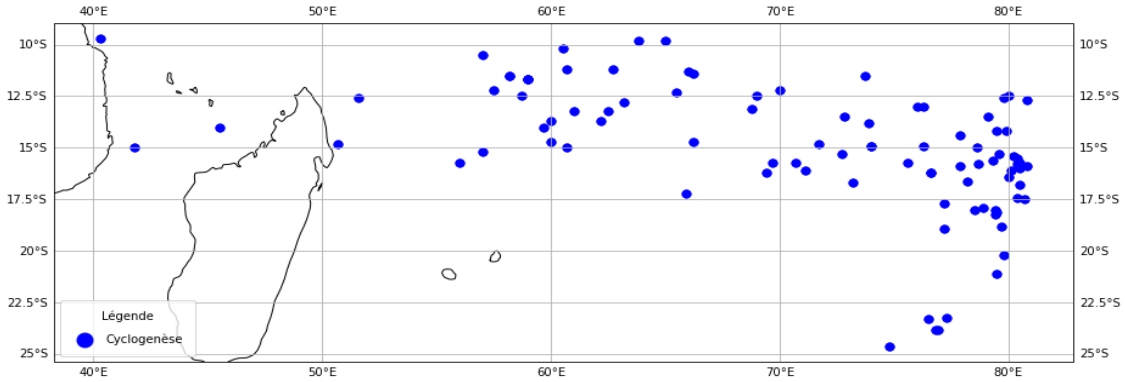
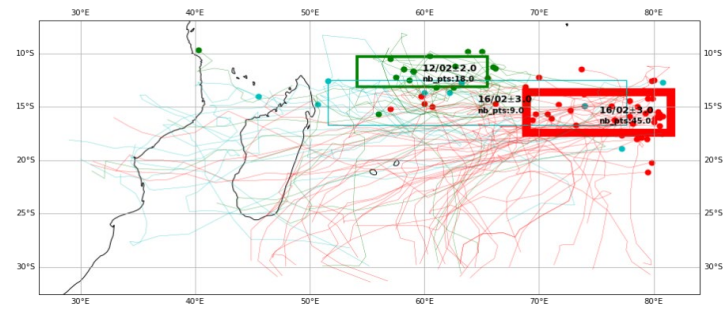
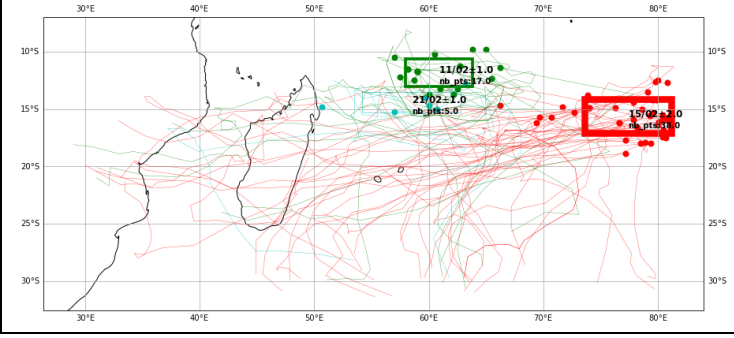


Figure 4 : *Prévision d'ensemble des 51 membres du modèle IFS du 07/02/2022 pour les 2 prochaines semaines (après filtrage)*

3.2.2 Vérification de l'analyse théorique et amélioration de la lisibilité des cartes

Nous avons à présent une carte de prévision des SDT filtrés, on peut appliquer nos différentes méthodes de clustering pour la prévision du 7 février. Nous cherchons ici à vérifier graphiquement que la méthode DB Scan est plus efficace que les méthodes K-Means et Agglomerative Clustering comme on a pu le voir dans la partie 2.3.

K-Means	
DB Scan	

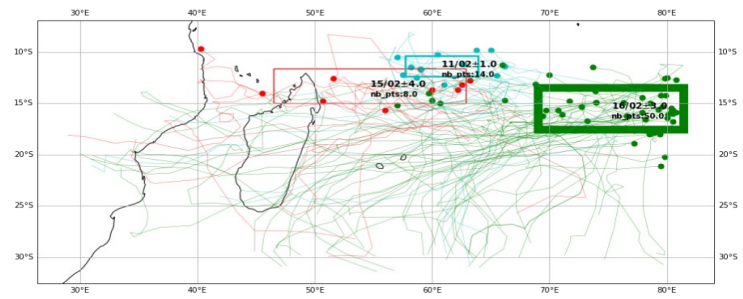
Agglomerative Clustering	

Tableau 3 : *Comparaison des méthodes de clustering DB Scan, K-Means et Agglomerative Clustering pour la prévision du 7 février 2022.*

Notons que nous avons pris soin d'afficher les dates de cyclogenèse de nos différents clusters ainsi que leurs incertitudes. En effet, sans cet affichage, il est difficile à l'œil nu de comprendre que deux clusters, bien qu'ils soient géographiquement proches, ne sont pas regroupés. On peut voir par exemple sur le tableau 3 que pour la méthode DB SCAN, bien que les clusters vert et bleu soient proches, ces derniers n'apparaissent pas à la même date (respectivement 11/02 et 21/02). Le modèle discrimine les deux clusters par leurs dates de cyclogenèse et c'est pourquoi ils ne les regroupent pas.

Nous avons également affiché le nombre de points présents dans chaque cluster afin de donner au lecteur une estimation de la probabilité d'un scénario. Pour la méthode DB Scan, on peut voir que le scénario en rouge est prédit par plus de membre que celui en bleu par exemple.

Les cadres sur les différentes cartes sont tracés à partir des écarts-types de la longitude et de la latitude. Cela permet de visualiser où il y a plus de chances que la cyclogenèse se déroule. L'épaisseur du cadre est proportionnelle au nombre de points dans le cluster. Plus il y a de points et plus le cadre est épais.

On remarque que les 3 méthodes sont capables d'identifier des SDT qui sont apparus réellement : Dumako (12 février) et Emnati (15 février)[4]. Cependant, on constate que K-Means et Agglomerative Clustering créent des cadres très grand de par sa faible robustesse aux valeurs aberrantes alors que DB Scan cible mieux les zones de cyclogenèse. De par sa meilleure précision, on gardera DB Scan pour classifier les cyclogenèses.

Il faut maintenant optimiser les features et les hyperparamètres pour notre modèle de clustering.

3.2.3 Optimisation des features et des hyperparamètres

Les hyperparamètres sont des paramètres réglables qui permettent de contrôler le processus d'entraînement du modèle. Les features sont les différentes variables que l'on utilise en entrée du modèle. L'optimisation de ces derniers consiste à rechercher la configuration pour laquelle on obtient les meilleurs résultats.

Optimisons tout d'abord les features. Nous avons les 8 paramètres du tableau 2 sur lesquels on peut entraîner notre modèle. On peut exclure les variables « Member_number » et « TC_number_in_member » car ces derniers ne jouent aucun rôle dans une cyclogenèse. Pour

[4] Voir Annexe 4 pour plus d'informations sur ces cyclones

vérifier quelles sont les meilleures variables à garder parmi les 6, nous faisons une boucle avec plusieurs combinaisons pour voir lesquelles donnent, graphiquement, de meilleurs résultats. Nous devons néanmoins forcément avoir les variables suivantes « TC_start_date », « TC_start_lat » et « TC_start_lon » car celles-ci sont primordiales pour caractériser une cyclogenèse. Le lecteur trouvera en annexe 7 les comparaisons des différentes cartes pour différentes prévisions (24 et 31 janvier, 7 février) et cela pour 4 combinaisons de variables différentes. L'analyse graphique montre que la combinaison « TC_start_date », « TC_start_lat », « TC_start_lon » et « TC_duration » donne les résultats les plus probants.

Avec ces features, on va pouvoir à présent optimiser les hyperparamètres. La méthode DB Scan contient 2 paramètres cruciaux : epsilon et le nombre minimal d'échantillon. La valeur de l'epsilon correspond à la distance maximale entre deux échantillons pour que l'un soit considéré comme dans le voisinage et le nombre minimal d'échantillon correspond au nombre minimal de points que le cluster doit avoir pour qu'il soit considéré comme tel. Comme pour les features nous réalisons une boucle afin de voir quelles sont les valeurs donnant les meilleurs résultats. Le lecteur trouvera en annexe 8 et 9 les comparaisons des différentes cartes pour différentes valeurs de « eps » et de « val_min_sample ». L'analyse graphique suggère que pour une valeur de 0,9 pour epsilon et une valeur de 5 pour le nombre minimal d'échantillon, nous obtenons les meilleurs résultats.

3.3 Classification pour les trajectoires

Nous souhaitons maintenant classifier les trajectoires de nos différents SDT. Pour pouvoir utiliser nos méthodes de classification, chaque trajectoire doive être décrite par un nombre fixe de features, liés à davantage de coordonnées spatiales, que nous obtenons à partir d'une méthode de prétraitement décrite ci-dessous.

3.3.1 Méthode de l'Equal Division

L'Equal Division est une méthode consistant à diviser une trajectoire en un certain nombre de points avec le même pas de temps entre les points. Chaque point aura des coordonnées spatiales (longitude et latitude) que nous pouvons utiliser comme paramètre pour notre méthode de classification. Si nous divisons, par exemple, la droite en 5 points, nous aurons 10 paramètres. L'annexe 10 détaille le procédé d'Equal Division pour une seule trajectoire. Nous avons fait ceci pour chaque trajectoire.

Précisons qu'une régression polynomiale d'ordre 2 a été faite, car elle est plus à même de représenter des trajectoires zonales et paraboliques.

3.3.1 Filtrage et optimisation

Nous avons gardé les mêmes filtres que pour la cyclogenèse, excepté le fait que nous formerons des clusters de nos trajectoires à partir des clusters faits lors de la classification de la cyclogenèse. Concernant les features, l'Equal Division nous a permis d'obtenir 5 features relatifs à la latitude et 5 features relatifs à la longitude. L'optimisation ici n'est pas nécessaire tant ses paramètres sont essentiels au fonctionnement du modèle.

Pour faciliter la lisibilité pour le lecteur, lorsque nous faisons des clusters de trajectoires, nous traçons la moyenne des trajectoires de chaque cluster. Autour de celle-ci, nous traçons une enveloppe gardant uniquement les points se trouvant à une distance ne dépassant pas 1,5 fois la moyenne. De cette manière, l'enveloppe nous préserve des points que l'on peut considérer comme aberrants. Des exemples sont visibles dans la partie 4.2.

3.3.3 Choix de la méthode de classification

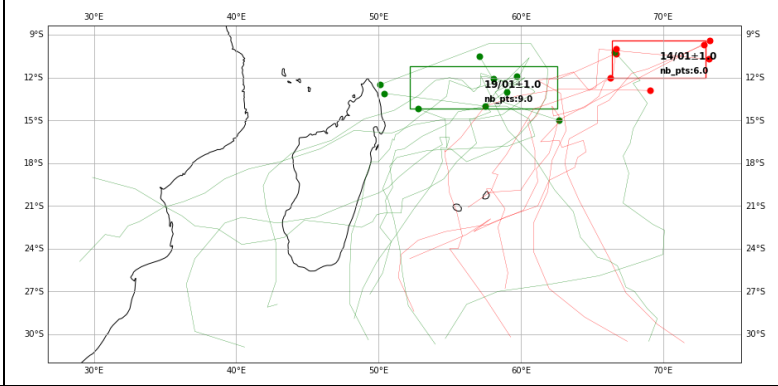
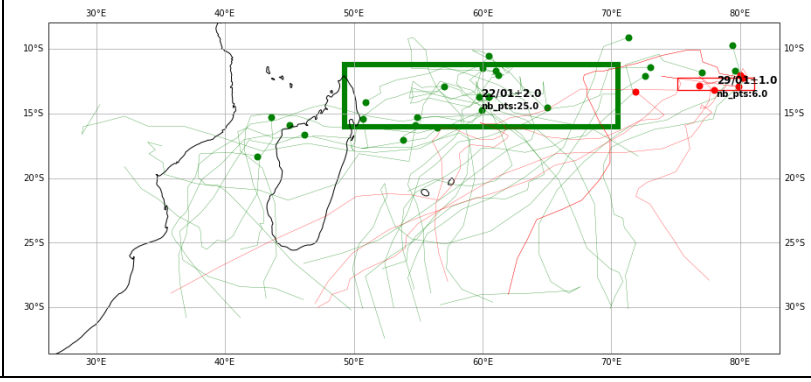
Parmi les 3 méthodes de classification, la méthode Agglomerative Clustering est la meilleure. On constate que de par ses propriétés se basant sur la densité des points, DB Scan n'arrive pas du tout à former plusieurs clusters. La méthode K-Means, bien qu'elle soit meilleure que la méthode DB-Scan, ne prend pas en considération le paramètre sur la durée de vie du cyclone. Comme vu dans la partie 2.3, la présence de nombreux features (11) et la non-uniformité des clusters viennent contraindre le fonctionnement de cette méthode, et l'empêche a fortiori de distinguer clairement les trajectoires. L'Annexe 11 vous montre un exemple de ces trois méthodes pour la prévision du 31 janvier.

4 Résultats

Le code final obtenant les résultats suivants est sur le lien GitHub disponible en Annexe 5. Le code est commenté permettant au lecteur de comprendre les différentes étapes.

4.1 Cyclogenèse

Le tableau 4 ci-dessous, nous montre les résultats que l'on obtient après avoir optimisé notre méthode DB Scan pour la classification des cyclogenèses.

10 janvier	
17 janvier	

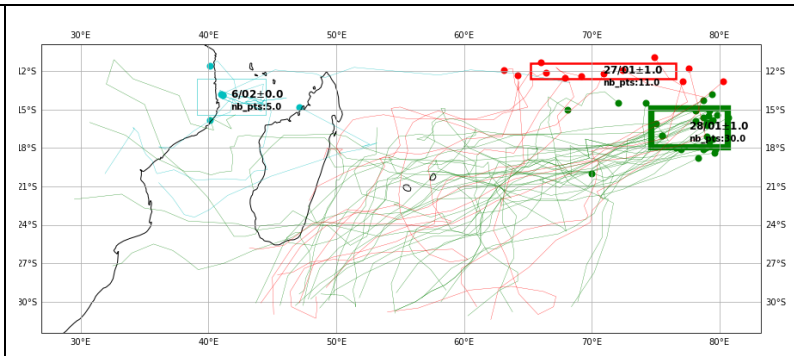
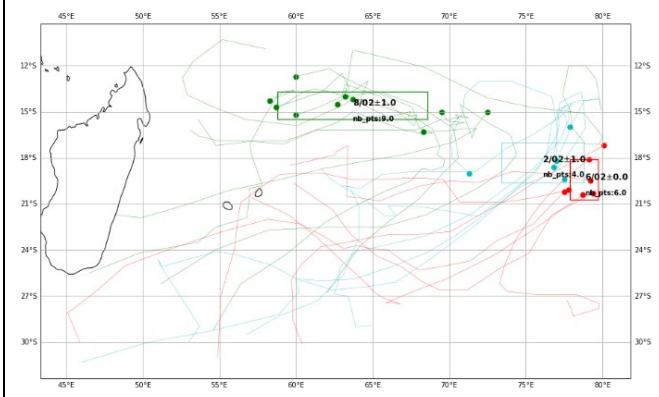
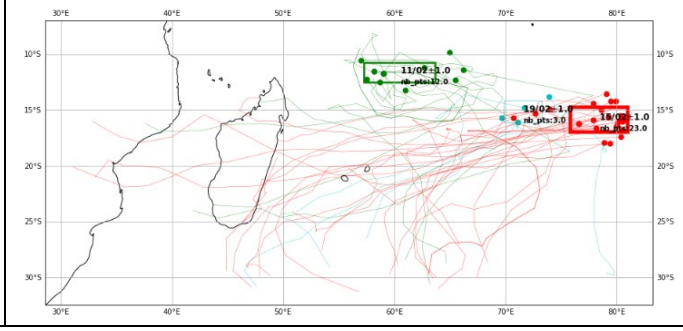
24 janvier	
31 janvier	
7 février	

Tableau 4 : *Résultats obtenus par la méthode DB Scan pour les dates du 10,17,24,31 janvier et 7 février*

Pour la prévision du 10 janvier, notre modèle est capable de prédire un système apparaissant le 19 ± 1 janvier (en vert). Cela correspond au cyclone Ana qui est réellement apparu le 20 janvier. Néanmoins le nombre de membre prédisant ce système est très faible. Il prévoit aussi un autre scénario (en rouge) mais celui-ci n'a pas réellement abouti.

Le 17 janvier, notre modèle a du mal à détecter précisément Ana. L'incertitude sur sa position est importante. Il n'est pour l'instant pas capable de détecter Batsirai qui apparaît le 28 janvier.

Le 24 janvier, notre modèle détecte deux scénarios probables (en vert et rouge) qui semblent être pour Batsirai. Il est intéressant de voir que notre modèle discrimine les 2 scénarios quand ceux-ci ont un jour d'écart seulement pour la date de cyclogenèse. Cela indique qu'à cette échéance, la zone de cyclogenèse de Batsirai est encore incertaine avec deux zones de cyclogenèse préférentielles qui se distinguent.

Le 31 janvier, notre modèle rencontre des difficultés à détecter Cliff. Ce cyclone est apparu le 2 février à l'endroit où les clusters bleu et rouge se situent. Ces derniers sont prédits par peu de

membres. Néanmoins de par leur écart de date de cyclogenèse, le modèle ne les a pas regroupés. L'erreur vient probablement de la qualité des données IFS.

Le 7 février, le modèle a été capable de prédire avec précision les cyclones Dumako et Emnati (respectivement vert et rouge). Ces derniers sont prédits par de nombreux membres.

Finalement, notre modèle permettant de classifier les cyclogenèses est capable de détecter des cyclones tropicaux qui sont réellement apparus. La date de cyclogenèse est une variable importante pour notre modèle car on peut voir que pour les prévisions faites les 24 janvier et 7 février, il fait des distinctions par rapport à la date de cyclogenèse bien que les points soient proches spatialement. Le modèle crée, tout de même, des clusters de scénarios qui ne se sont pas passés mais ces derniers sont en moyenne prédit par peu de membres. Cependant, on constate qu'il est limité à une échéance temporelle ne dépassant pas une dizaine de jours. De plus, dans certaines situations, notre modèle perd de ses propriétés discriminantes (ex le 17 janvier) mais cela est sûrement dû à la qualité des prévisions faites par l'IFS.

4.2 Trajectoires des cyclones tropicaux

Dans cette partie, nous nous intéresserons uniquement aux scénarios susceptibles de représenter un cyclone qui est réellement apparu. En effet, s'intéresser aux trajectoires d'un scénario qui n'est pas apparu pendant la saison cyclonique ne nous permettra pas de savoir si notre méthode de classification de trajectoires fonctionne. Pour chacune des prévisions, nous appuierons donc notre étude uniquement sur les clusters ayant trouvé les bonnes zones de cyclogenèse

Sur la figure 5 ci-dessous, en comparant avec la courbe noire, on peut constater que la trajectoire vraie colle avec le scénario bleu mais que celle-ci diverge pour aller plus au nord des côtes africaines.

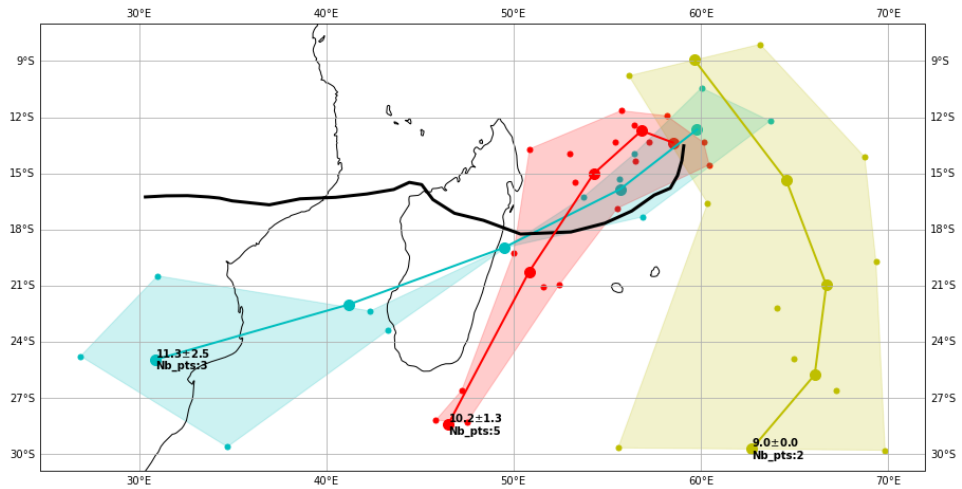


Figure 5 : *Différentes trajectoires envisagées par le modèle pour la prévision du 10 janvier (cluster vert du tableau 4) ; la courbe noire est la vraie trajectoire d'Ana*

Sur la figure 6 ci-dessous, la trajectoire vraie est bien incluse dans l'enveloppe d'un scénario (en jaune) relativement probable bien que minoritaire.

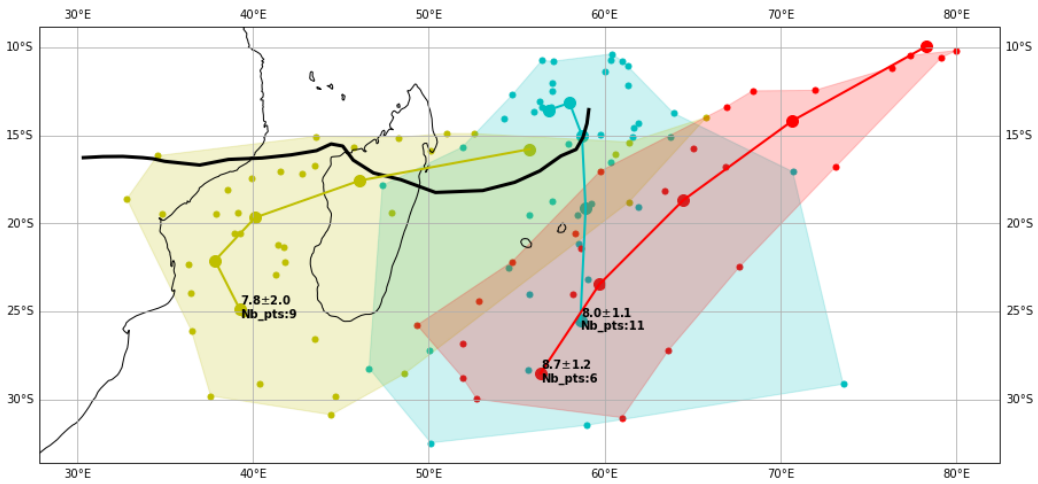


Figure 6 : *Différentes trajectoires envisagées par le modèle pour la prévision du 17 janvier (cluster vert dans le tableau 4) ; la courbe noire est la vraie trajectoire de Ana*

Sur la figure 7 ci-dessous, il est intéressant de voir qu'au début, le scénario bleu est celui qui colle à la réalité puis qu'au fil du temps, il cède sa place au scénario jaune. On peut remarquer qu'encore une fois la trajectoire vraie n'est pas incluse dans le scénario le plus probable.

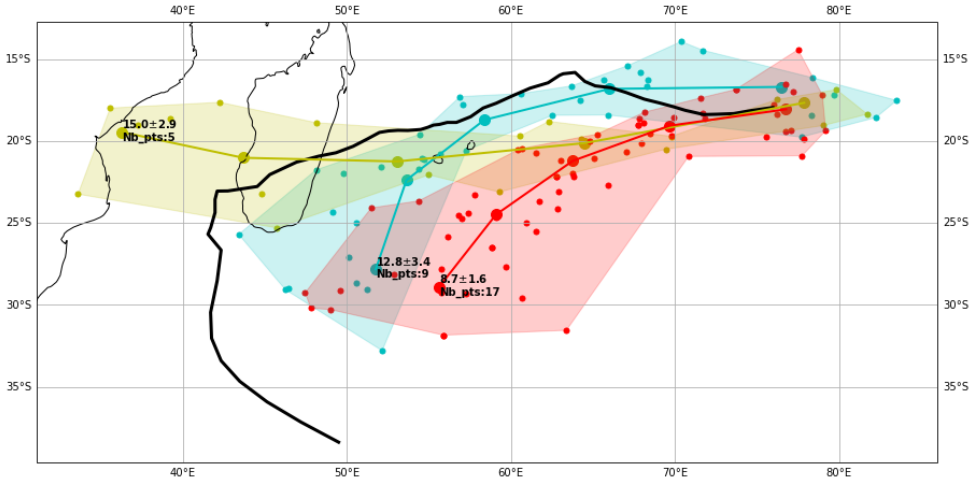


Figure 7 : *Différentes trajectoires envisagées par le modèle pour la prévision du 24 janvier (cluster vert du tableau 4) ; la courbe noire est la vraie trajectoire de Batsirai*

Les résultats pour la prévision du 31 janvier sont en Annexe 12 car le scénario ne contient pas assez de points pour tracer une enveloppe.

Sur la figure 8 ci-dessous, on remarque que Dumako (figure 8-a), que cela soit sa cyclogenèse ou sa trajectoire, est bien prédit par le modèle. Ce scénario est prédit par plusieurs membres. Concernant Emnati, il suit bien l'un des scénarios mais c'est celui le moins prédit parmi les membres

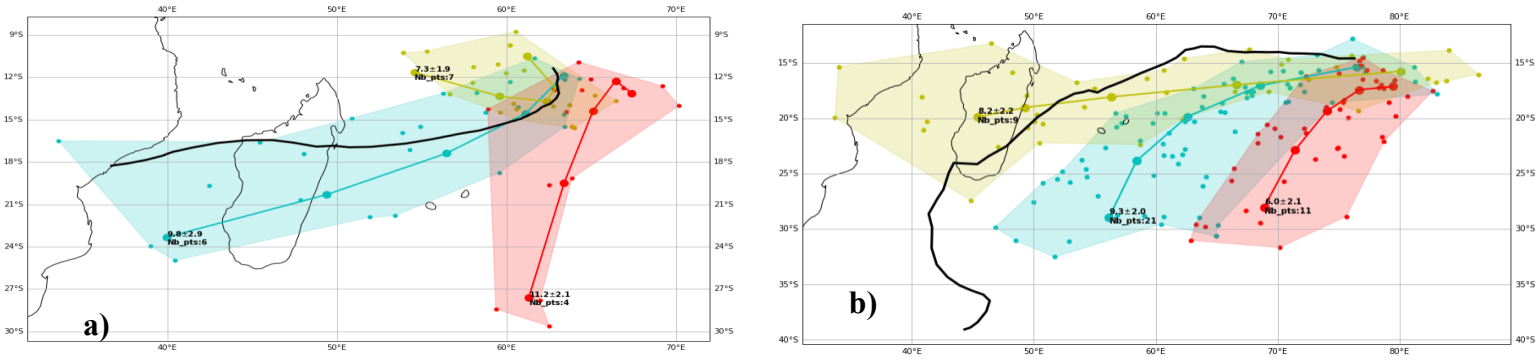


Figure 8 : *Différentes trajectoires envisagées par le modèle pour la prévision du 7 février ; a) Scénario pour Dumako (cluster vert du tableau 4) ; b) Scénario pour Emnati (cluster rouge du tableau 4) ; la courbe noire est la vraie trajectoire de Dumako (a) et de Emnati (b)*

Il est intéressant également de regarder si notre modèle fonctionne pour les cyclones déjà apparus. Prenons par exemple sur la figure 10 ci-dessous, la prévision du 31 janvier où l'on peut voir Batsirai déjà en activité. On voit que le scénario en jaune colle bien avec la trajectoire réelle. C'est d'ailleurs celui prédit par le plus de membre (25 /51 soit ~50%). Précisons que pour obtenir cette figure, il faut enlever le filtre sur les cyclones déjà existant.

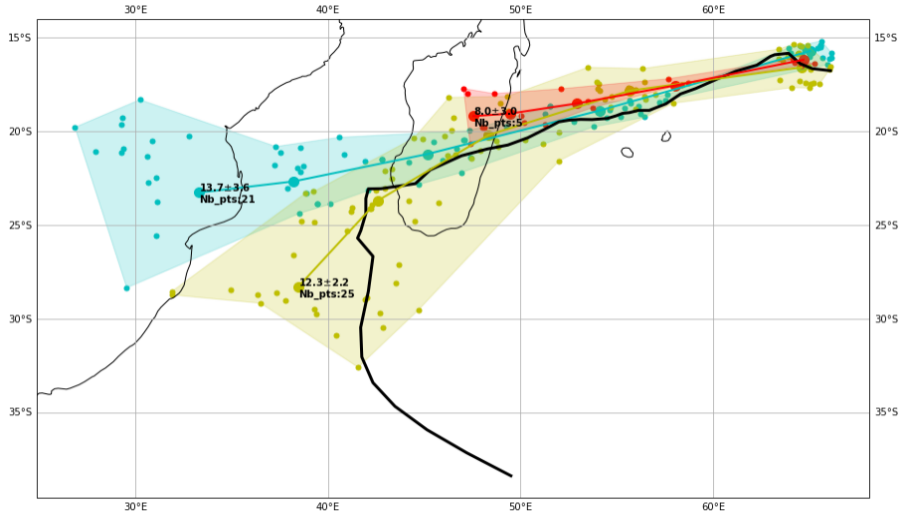


Figure 10 : *Différentes trajectoires prédites par le modèle pour la prévision du 31 janvier ; la courbe noire est la trajectoire réelle de Batsirai.*

Finalement, il est important de faire la distinction entre la qualité des prévisions provenant du modèle IFS et la pertinence de notre méthode de classification. La cyclogenèse et trajectoire de plusieurs SDT de la saison ont été prévues très tardivement par la prévision d'ensemble elle-même. Il n'y avait pas donc grand-chose que notre méthode puisse faire dans ce cas. Cela montre qu'avoir seulement 51 membres ne permet pas d'identifier de façon exhaustive l'ensemble des évolutions des SDT potentiels sur le bassin. Lorsque le signal d'un SDT, que ce soit en cyclogenèse ou trajectoire, était présent dans la prévision d'ensemble, celui-ci était presque systématiquement inclus dans un des clusters proposés par la méthode, même si celui-ci était parfois minoritaire. Lorsqu'au moment de la prévision, le cyclone est déjà existant, il y a moins d'incertitude sur la trajectoire, et la capacité de prédiction du modèle est meilleure.

5 Conclusion et perspectives

Ce stage a consisté à développer, par le biais de l'apprentissage machine, une méthode de classification capable de cibler les différentes zones de cyclogenèse ainsi que leurs trajectoires à partir des prévisions faites par le modèle IFS pendant les mois de janvier et février 2022. La prévision d'ensemble IFS est dense en informations et difficilement déchiffrable à l'œil nu mais notre méthode se trouve être pertinente pour extraire l'information contenue dans celle-ci. L'enjeu est d'autant plus important car le Centre Européen a pour projet de faire passer le nombre de membres dans le modèle IFS de 51 à 100. L'usage d'outils statistiques qui fonctionnent est donc essentiel. Lorsque nous nous sommes intéressés à la classification de la cyclogenèse, la méthode DB Scan était la plus efficace de par sa capacité à cibler précisément les zones de cyclogenèse sans tenir compte des valeurs aberrantes. Ses propriétés discriminantes en font néanmoins une méthode diminuant le nombre de prédiction faite par les différents membres. De plus, elle a du mal à prédire des SDT apparaissant au-delà d'une dizaine de jours. L'étude de la classification des trajectoires nous a permis, quant à elle, de montrer que la méthode Agglomerative Clustering était la plus appropriée. Néanmoins, l'information correcte de la trajectoire du cyclone n'est pas toujours contenue dans la prévision d'ensemble du modèle IFS. Cela influe sur la qualité de notre méthode de classification de trajectoire qui dépend intrinsèquement des données fournies par le Centre Européen.

De futurs améliorations peuvent être réalisées, elles concernent essentiellement la classification des trajectoires. Dans un premier temps, il serait intéressant d'automatiser le choix du nombre de cluster lorsque l'on s'intéresse à la classification des trajectoires. Cela éviterait à l'utilisateur du modèle de lancer une première fois le modèle pour qu'il estime lui-même le nombre de trajectoire présent. Nous pourrions également automatiser le choix du nombre de fois dont on divise la trajectoire lors de l'Equal Division car en fonction de la forme de la courbe et de sa longueur, la division systématiquement par 5 peut influencer nos variables sur lesquelles notre modèle apprend. Une perspective envisageable serait de lier les scénarios à des contextes synoptiques (position de la Zone de Convergence Intertropicale par exemple) qui ont une influence directe sur la cyclogenèse ainsi que sur les trajectoires des cyclones. Un autre axe de travail serait d'ajuster la méthode selon l'échéance de prévisions pour essayer de prédire des cyclogenèses au-delà de 10 jours par exemple.

Bibliographie

[1] Collins M., M. Sutherland, L. Bouwer, S.-M. Cheong, T. Frölicher, H. Jacot Des Combes, M. Koll Roxy, I. Losada, K. McInnes, B. Ratter, E. Rivera-Arriaga, R.D. Susanto, D. Swingedouw, and L. Tibig, 2019: Extremes, Abrupt Changes and Managing Risk. In: IPCC Special Report on the Ocean and Cryosphere in a Changing Climate [H.-O. Pörtner, D.C. Roberts, V. Masson-Delmotte, P. Zhai, M. Tignor, E. Poloczanska, K. Mintenbeck, A. Alegría, M. Nicolai, A. Okem, J. Petzold, B. Rama, N.M. Weyer (eds.)]. In press.

[2] Klotzbach, P. J., Wood, K. M., Schreck III, C. J., Bowen, S. G., Patricola, C. M., & Bell, M. M. (2022). Trends in global tropical cyclone activity: 1990–2021. Geophysical Research Letters, 49, e2021GL095774. https://doi.org/10.1029/2021GL095774Received 26 AUG

[3] Miller T., Macrae J., (2018), Machine Learning for Environmental Toxicology: a call for integration and innovation. Environmental Science and Technology

[4] G. Jumeaux, H. Quetelard, D. Roy (2011) Atlas Climatique de la Réunion, Sainte-Clotilde, Météo-France

[5] European Centre for Medium-Range Weather Forecasts, Reading in United Kingdom, https://www.ecmwf.int/en/forecasts/documentation-and-support

[6] Scikit-Learn developers (BSD License), 2007-2022, https://scikit-learn.org/stable/modules/clustering.html

Annexes

Annexe 1 : Terminologie utilisée dans le Sud-Ouest de l'Océan Indien pour classer les différents états d'un système dépressionnaire tropical

Terminologie	Vent maximal moyen sur 10 minutes	Pression minimale centrale
Dépression tropicale	51 à 62 km/h	999 à 995 hPa
Tempête tropicale modérée	63 à 88 km/h	995 à 985 hPa
Forte tempête tropicale	89 à 117 km/h	985 à 970 hPa
Cyclone tropical	118 à 165 km/h	970 à 941 hPa
Cyclone tropical intense	166 à 212 km/h	941 à 910 hPa
Cyclone tropical très intense	> 212 km/h	< à 910 hPa

Annexe 2 : Explication des différences entre l'apprentissage supervisé et l'apprentissage non-supervisé

Apprentissage supervisé	Apprentissage non-supervisé
 X : forme des animaux ; Y : Nom des animaux La machine apprend à partir de X qui est associé à Y ainsi elle va être capable de donner le nom de l'animal en fonction de sa forme.	 Classification en fonction des formes  La machine analyse la structure des données X pour apprendre elle-même à faire des clusters en fonction des formes.

Annexe 3 : Précisions sur le modèle IFS – ENS

	Prévision	Nombre de membres	Résolution horizontale	Nombre de niveau vertical	Pression en haut de la dernière couche	Perturbation du modèle
ENS Ensemble-Atmospheric Model	0-15 jours	51	~18 km Interpolé à 0,2°	137	0,01 hPa	Oui

Annexe 4 : Tableau indiquant la date et la zone d'apparition des différents cyclones tropicaux ainsi qu'un résumé de sa trajectoire

	Date d'apparition	Coordonnées de la cyclogenèse	Résumé trajectoire
Ana	20 janvier	13°lat / 59°lon	Apparaît à mi-distance entre les Seychelles et la Réunion puis le 22 janvier il est entre la Réunion et Madagascar et se désagrège le 26 janvier sur la côte africaine.
Batsirai	26-27 janvier	9,5° lat / 89,5° lon	Apparaît dans l'est de l'Océan Indien puis traverse l'océan d'est en ouest. Il est proche des côtes réunionnaises le 2 février. Il touche les côtes malgaches le 5 février et se désagrège dans le sud le 10 février.
Cliff	2 février	18° lat / 81,1° lon	Apparaît à l'est de la Réunion pour avoir une trajectoire zonale et s'éteint à mi-distance entre la Réunion et les côtes malgaches.
Dumako	12 février	11,4° lat / 62,8° lon	Apparaît au nord-est de la Réunion et se dirige d'est en ouest jusqu'au Mozambique où il se désagrège.
Emnati	15 février	14,58° lat / 76,3° lon	Apparaît à l'est de la Réunion pour avoir une trajectoire parabolique où il touche les côtes malgaches puis s'éteint.
Gombe	8 mars	15° lat / 48° lon	Apparaît à une dizaine de degrés au nord de la Réunion pour avoir une trajectoire zonale et se désagréger sur les côtes Africaines.

Annexe 5 : Lien du GitHub et du fichier drive

Le fichier drive vous mène vers le dossier 'Données' dans lequel vous trouverez :

- Des fichiers 'csv' qui nous donnent les vraies trajectoires des cyclones entre le 10 janvier et le 7 février.
- Un fichier 'M1SOAC_Code_Rapport ' qui permet d'obtenir les résultats de la partie 4. Seul la date devrait être modifié par le lecteur pour obtenir les mêmes résultats
- 5 dossiers dont l'intitulé dépend de la date de prévision. Notre programme 'M1SOAC_Code_Rapport' les lit automatiquement. Pour une bonne exécution, veuillez à bien placer ce dernier dans le même dossier que l'ensemble des 5 dossiers. Tous ces fichiers doivent être téléchargés pour que le code fonctionne pour chaque date.

Lien du drive :
https://drive.google.com/drive/folders/1aXH9OgtCYyCBSa53pGs9sitmot9B07fg?usp=sharing

Le lien GitHub vous mène vers mon répertoire contenant le code Jupyter Notebook 'M1SOAC_Code_Rapport'. Ce code permet de lire les données et d'afficher les résultats.

Le lien GitHub :

https://github.com/Gleam974/Code_stage_M1

Si le lecteur n'arrive pas à télécharger les dossiers à partir du lien drive, voici un lien WeTransfer contenant la même chose que le lien drive. Celui-ci expire le 6 septembre 2022.

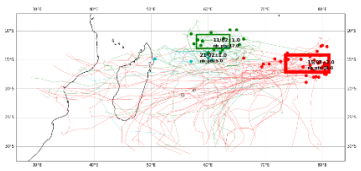
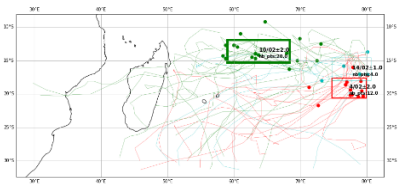
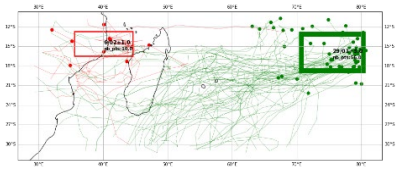
Le lien WeTransfer :

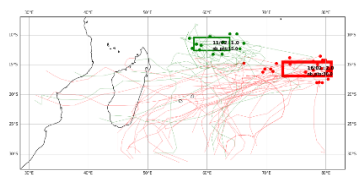
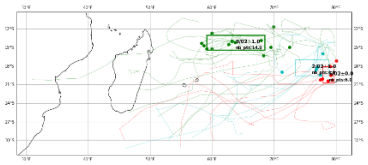
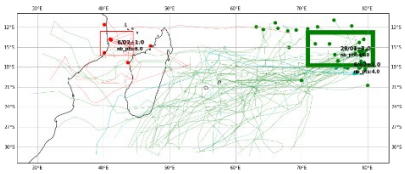
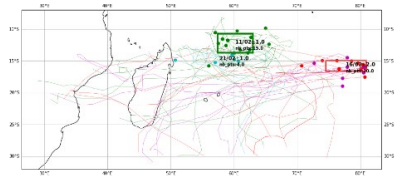
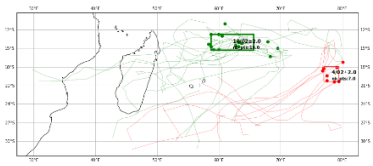
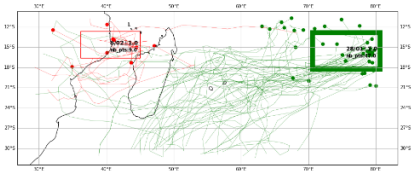
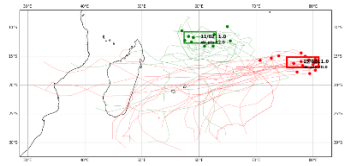
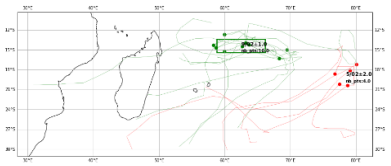
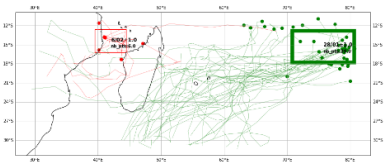
https://wetransfer.com/downloads/a7e448685cbba5e2cad5bfb4104fcdcd20220830124738/f41690faaef4bd04e865837c21da54a020220830124756/c13af9

Annexe 6 : Code ayant permis de filtrer les données pour la cyclogenèse

```
#suppression des sytèmes dont la durée de vie est inférieure à 3 jours
df_ens_filtered = df_ens[df_ens['TC_duration'] > 3 ]
#suppresion des système ayant déja démarré. Cyclogenèse donc TC_Start_Date =/= 0
df_ens_filtered = df_ens_filtered[df_ens_filtered['TC_start_date'] >1]
#suppresion des système démarrant au delà de 15 jours.
df_ens_filtered = df_ens_filtered[df_ens_filtered['TC_start_date'] <15]
```

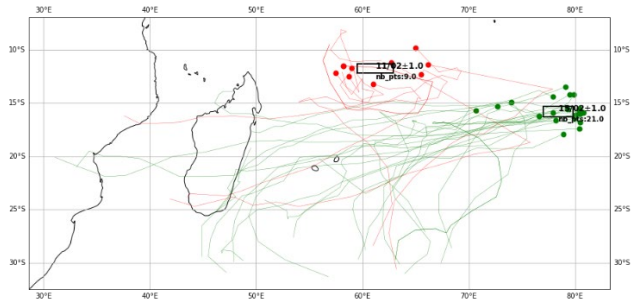
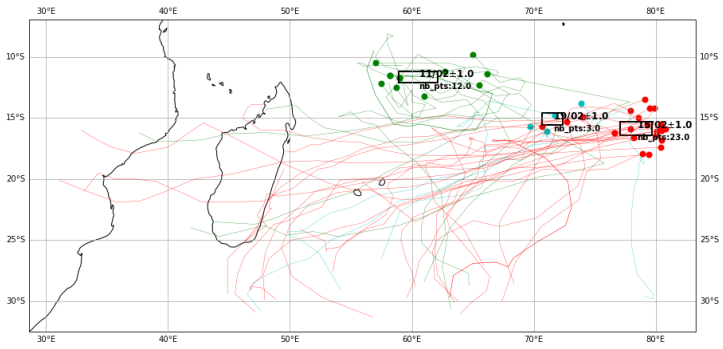
Annexe 7 : Test de classification de cyclogenèse pour différents features

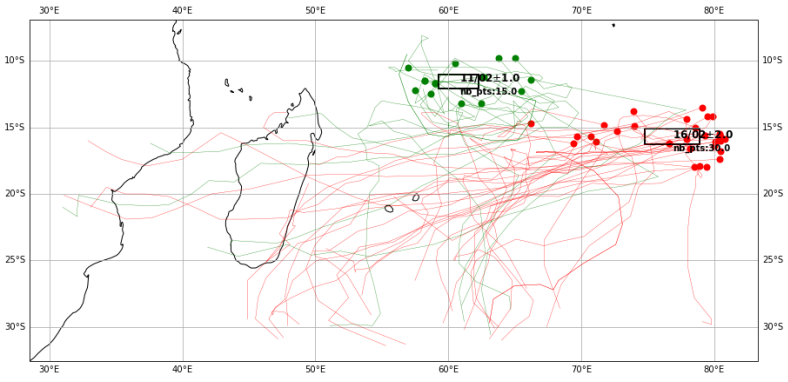
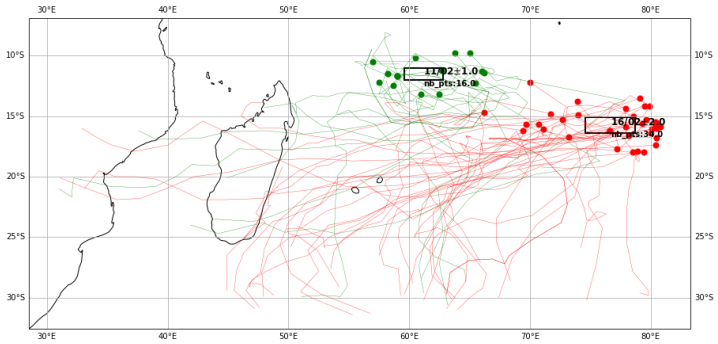
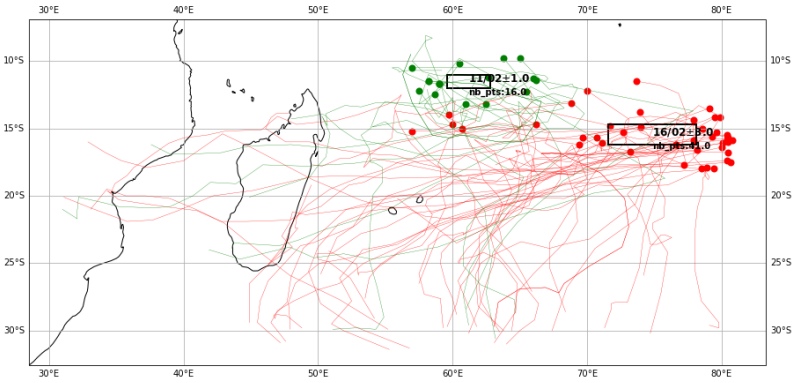
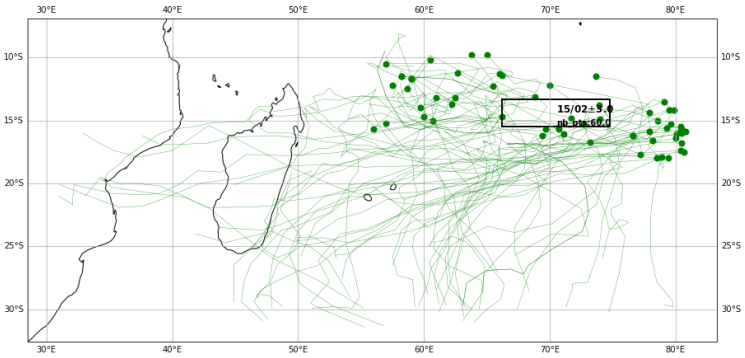
Features	Prévision du 7 février 2022	Prévision du 31 janvier	Prévision du 24 janvier
Combinaison 1: TC_start_date ; TC_start_lon ; TC_start_lat			

Combinaison 2: TC_start_date ; TC_start_lon ; TC_start_lat ; TC_duration			
Combinaison 3: TC_start_date ; TC_start_lon ; TC_start_lat ; TC_min_PMIN ; TC_max_VMAX			
Combinaison 4: TC_start_date ; TC_start_lon ; TC_start_lat ; TC_duration ; TC_min_PMIN ; TC_max_VMAX			

On remarque que la combinaison 2 est la plus viable. Elle est parmi les quatre combinaisons, celle qui regroupe le mieux les critères de discrimination, précision et inclusion de nombreux membres.

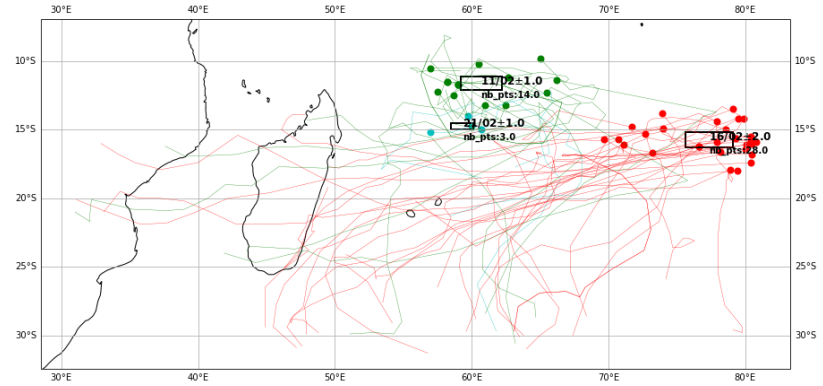
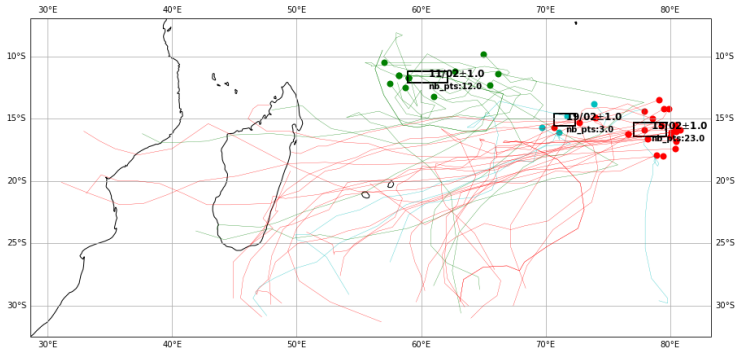
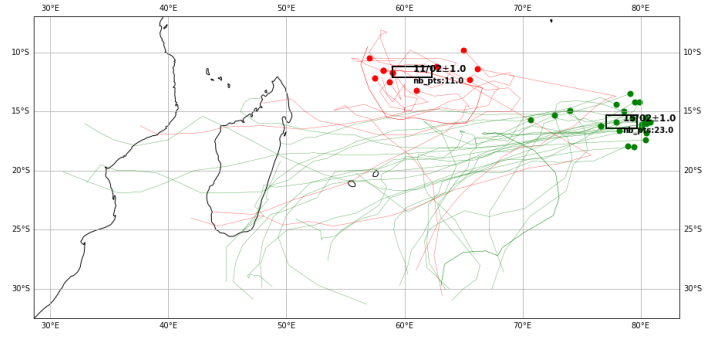
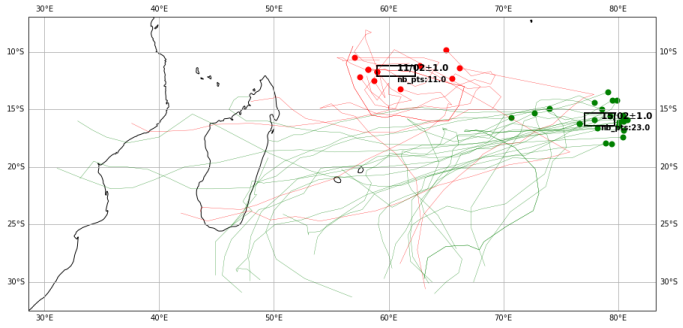
Annexe 8 : Test de classification de cyclogenèse pour différents epsilons

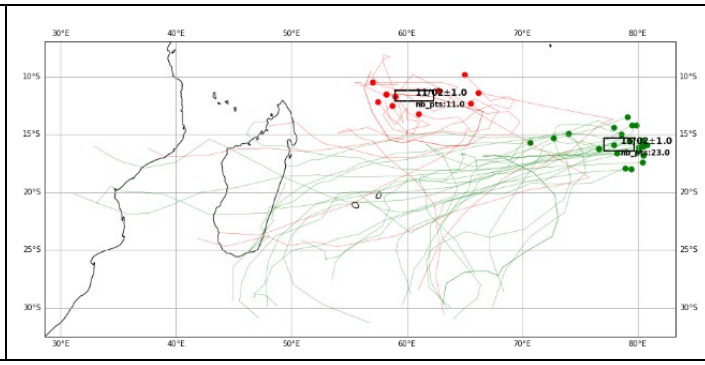
Valeur d'epsilon	Prévision du 7 février
0,8	
0,9	

1	
1,1	
1,2	
1,3	

On constate que pour epsilon égal à 0,9, le modèle discrimine suffisamment par rapport aux dates de cyclogenèse alors qu'en dessous ou au-dessus de cette valeur, le modèle n'est pas assez discriminant.

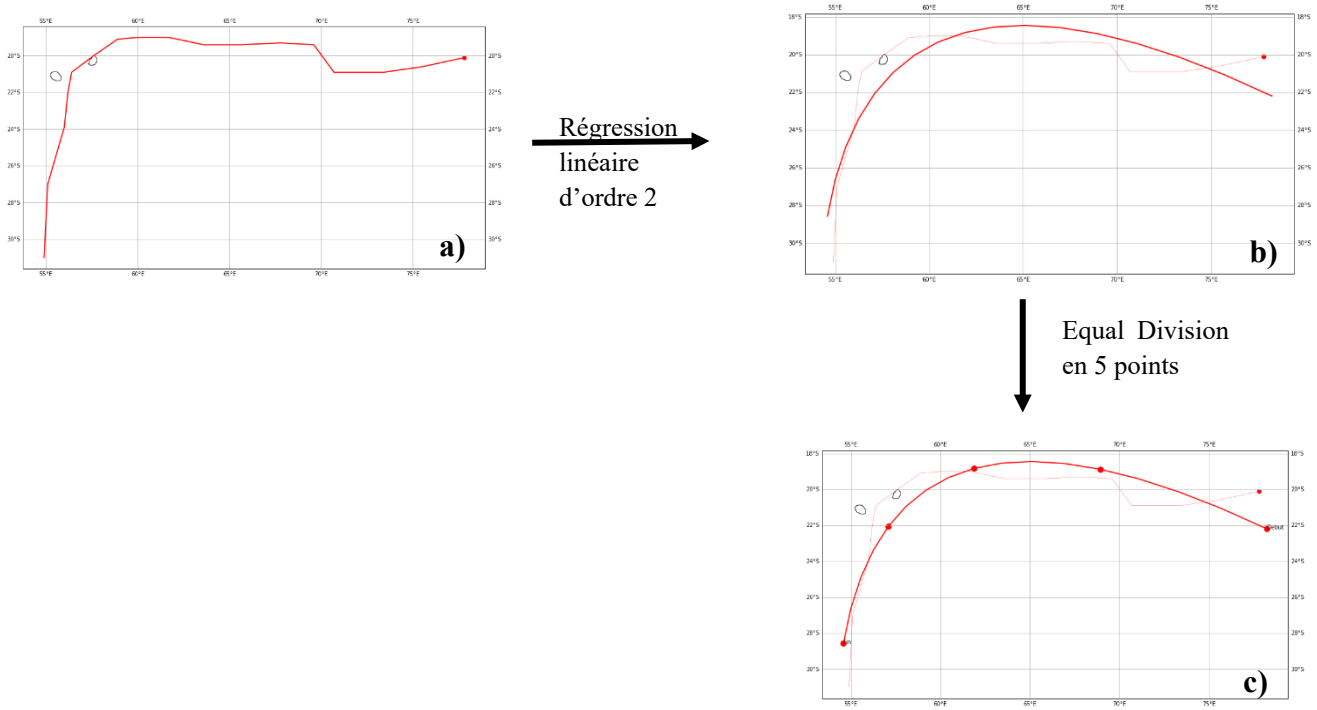
Annexe 9 : Test de classification de cyclogenèse pour différents nombre minimal d'échantillons

Valeur minimal de l'échantillon	Prévision du 7 février
4	
5	
6	
7	

8	

On constate que pour une valeur minimale d'échantillon égal à 5, le modèle discrimine suffisamment par rapport aux dates de cyclogenèse alors qu'au-dessus de cette valeur, le modèle n'est pas assez discriminant. La valeur 4 pourrait éventuellement fonctionner, mais elle aussi n'est pas suffisamment discriminante.

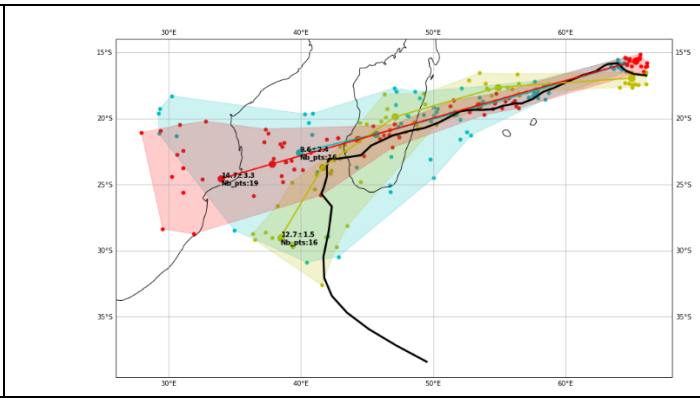
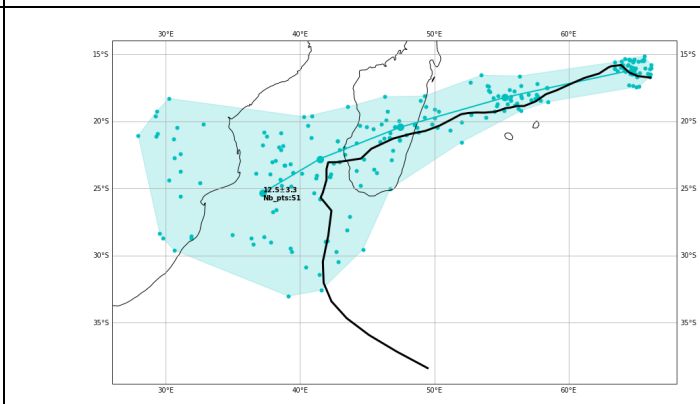
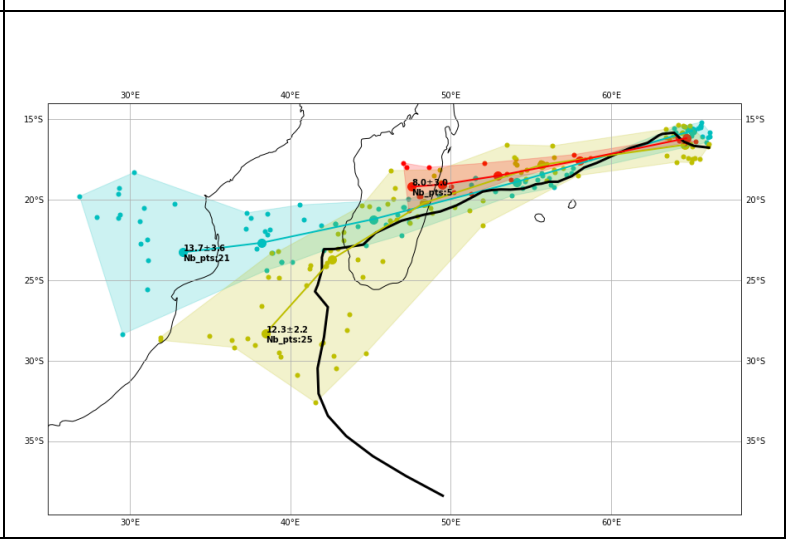
Annexe 10 : Différentes étapes de l'Equal Division



a) Une trajectoire d'un SDT prédite par un membre ; b) Régression linéaire d'ordre 2 d'une trajectoire d'un SDT ; c) Division en 5 points de notre régression linéaire

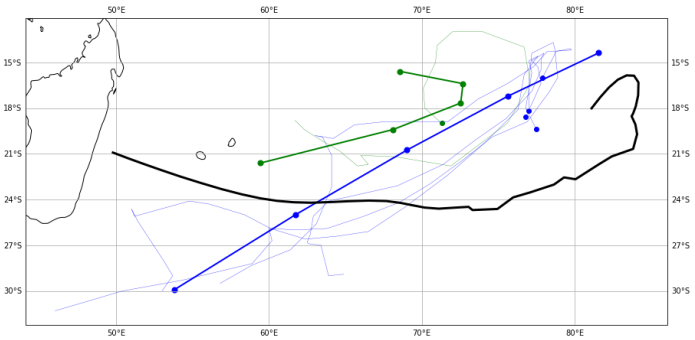
Annexe 11 : Choix de la méthode de classification pour les trajectoires

Méthode de classification	Prévision du 31 janvier ; Ici sur un cyclone déjà existant Batsirai

K-Means	
DB Scan	
Agglomerative Clustering	

On remarque que la méthode DB Scan n'arrive pas à créer plusieurs clusters. La méthode K-Means, elle, forme des clusters qui se mélangent entre eux et distingue difficilement la durée des trajectoires. Ce n'est donc pas intéressant pour un prévisionniste. La meilleure méthode est l'Agglomerative Clustering qui arrive à bien séparer les trajectoires et de prendre en considération la durée de celle-ci.

Annexe 12 : Résultats classification de trajectoire pour la prévision du 31 janvier (Scénario du cyclone Cliff)



Bien qu'il n'y ait pas assez de points pour faire une enveloppe, le modèle semble prédire l'orientation de la trajectoire mais avec si peu de points (4), qu'il peut être difficile pour un prévisionniste de se fier à ce scénario.