Université Assane Seck de Ziguinchor
UFR Sciences et Technologies
Département de Physique

Mémoire de Master Physique et Applications

Spécialité : Sciences de l'Atmosphère et de l'Océan

Thème

Modélisation des précipitations par l'intelligence artificielle ou Apprentissage automatique en Casamance

Présenté par :

David SAGNA

Sous la direction de

Dr Samo DIATTA

Soutenu publiquement le 27 Mars 2021

Sous la supervision du Pr Moctar CAMARA

Devant le jury composé de :

M. Moctar CAMARA	Professeur Titulaire	Président	UASZ
M. Khadim DRAME	Maitre Assistant	Rapporteur	UASZ
M. Mamadou Lamine MBAYE	Maitre Assistant	Examinateur	UASZ
M. Joseph Sambasene DIATTA	Maitre Assistant	Examinateur	UASZ
M. Samo DIATTA	Maitre Assistant	Encadrant	UASZ

Dédicaces

Je dédie ce travail à toute ma famille particulièrement à mon père, ma mère, mes frères et soeurs,la famille Sambou de Lyndiane Ziguinchor ainsi qu'à mes amis.

Que DIEU vous garde et vous accorde sa bénédiction .

Remerciements

Je tiens à remercier les personnes impliquées de près ou de loin dans mon travail de recherche.

Tout d'abord, mon directeur de mémoire, ***Dr. Samo DIATTA*** *, pour avoir dirigé ce travail bien qu'il soit difficile d'exprimer en quelques mots ma reconnaissance envers lui. Sa patience, sa disponibilité et ses conseils m'ont permis de vivre une expérience stimulante, enrichie par son expertise et sa rigueur scientifiques qui influenceront longtemps mes projets professionnels.*

Merci également aux membres du jury avec à sa tête le président, pour m'avoir fait l'honneur d'accepter d'évaluer ce travail.

Je remercie les enseignants chercheurs du département de Physique et plus particulièrement les enseignants du LOSEC : Pr Bamol Ali Sow, Pr Moctar Camara, Dr Mamadou Lamine Mbaye, Dr Joseph Diatta, Dr Ababacar Ndiaye, pour cette formation pleine de qualités.

Je remercie l'ensemble des étudiants du laboratoire LOSEC en commençant par les Doctorants particulièrement à Serigne Mbacké COLY et Birane NDOM pour l'aide qu'ils m'ont apporté. Je terminerais par mes camarades de promo pour les nombreuses discussions et le soutient qu'ils m'ont apportés : Dioumacore FAYE,Jacques D. DIOUF, Assane NDIAYE, Adama THIANDIOUME, Amadou DIOUF et Fatou KHOULE .

Enfin, merci à ma famille et à Bintou qui m'ont soutenu durant mes études.

Résumé

La réduction d'échelle atténue considérablement les inconvénients de la simulation du climat régional par les modèles de circulation générale (MCG). Cependant, peu d'informations sont disponibles concernant la réduction d'échelle par des méthodes d'apprentissage automatique. De plus, la complexité est encore accrue par la rareté des données d'observations dans certaines régions telle que le sahel. Cette étude utilise les modèles des "K plus proches voisins" (KNN), le "support vector machine" (SVM), l' "extreme machine learning" (ELM) et le "poursuit preject regression" pour réduire l'échelle des précipitations en Casamance au sud du Sénégal, qui aussi est considérée comme une zone vulnérable aux changements climatiques. L'ensemble des données de réanalyse du " National Center for Environmental Prediction" et du" National Center for Atmospheric Research" NCEP/NCAR a été sélectionné comme prédicteur à partir des corrélations0 de Pearson. Les données mensuelles des précipitations pour les périodes 1950-1994 et 1994-2015 ont été utilisées pour l'étalonnage et la validation des modèles, respectivement. La performance des modèles a été évaluée à l'aide de diverses mesures statistiques, notamment l'erreur de biais (B), le coefficient de détermination (R^2) et l'erreur quadratique moyenne (RMSE). Les comparaisons des séries chronologiques mensuelles moyennes des précipitations observées et simulées ont montré une bonne concordance de l'évolution pendant les périodes d'étalonnage et de validation. Cependant, l'étude des variances montre une sous-estimation de la précipitation par les modèles d'apprentissage automatique. Le diagramme de Taylor nous a permis de comparer les modèles et de constater que le KNN donnait les meilleures représentations de la précipitation.

Mots clés : ***Réduction d'échelle, Apprentissage automatique, precipitations, Casamance***

Abstract :

Downscaling considerably reduces the drawbacks of regional climate simulation by general circulation models (GCMs). However, little information is available on downscaling by machine learning methods. Moreover, the complexity is further increased by the scarcity of data in some regions such as the Sahel. This study uses Nearer Neighbours (NNN), Support Vector Machine (SVM), Extreme Machine Learning (ELM) and continued preject regression to downscale rainfall in Casamance in southern Senegal, which is also considered a vulnerable area to climate change. The National Center for Environmental Prediction and National Center for Atmospheric Research NCEP/NCAR reanalysis datasets were used to select predictors in the Person correlation. Monthly precipitation data for the periods 1950-1994 and 1994-2015 were used for model calibration and validation, respectively. Model performance was evaluated using various statistical parameters including bias error (B), coefficient of determination (R^2) and root mean square error (RMSE). Comparisons of the mean monthly time series of observed and reduced precipitation showed good agreement during the calibration and validation periods, while the reduced models were found to underestimate the variance of precipitation in both periods. Other parameters such as the Taylor plot were equal and allowed the models to be compared and KNN was better.

Keywords : Downscaling, machine learning, precipitations, Casamance

Table des matières

Table des figures

Introduction

Les problèmes liés aux changements climatiques occupent une importante place parmi les préoccupations majeures de notre siècle. Plusieurs auteurs s'accordent à dire que les aspects liés à l'eau occuperont une place prépondérante parmi les impacts potentiels des changements climatiques [1, 2] . Il y' a donc un intérêt particulier des scientifiques à l'étude de la variabilité climatique et des ressources en eau.

L'Afrique de l'ouest est l'une de ces régions, soumise à un régime de mousson marqué par une saison sèche en hiver et des pluies en été. Depuis des dizaines d'années, ce régime des pluies de mousson subit d'importantes variations temporelles, comme en témoigne la longue période de sécheresse liée aux déficits pluviométriques qui a sévi de la fin des années 1960 au milieu des années 1990 [3]. Cette sécheresse, particulièrement marquée dans la région sahélienne, a accentué davantage la vulnérabilité des populations face à la disponibilité des ressources en eau. Récemment, cette région a également connu de fréquentes épisodes de précipitations intenses qui ont entrainé des inondations et des pertes de vies humaines à Dakar au Sénégal et dans d'autres pays en Afrique de l'ouest. Par exemple, entre le 16 et le 22 Août 2005, Dakar a enregistré 367 mm de pluie, soit plus de la moitié de la moyenne annuelle des précipitations totales. Le 26 Aout 2012, Dakar a reçu un quart des précipitations annuelles (168 mm) en une heure [4]. Ces précipitations ont eu des impacts psychosociaux et sanitaires considérables, qui sont autant de pertes éventuelles pour les secteurs vulnérables tels que l'agriculture, les infrastructures, le commerce, etc. La perception d'une augmentation des catastrophes liées au climat a fait craindre que la fréquence des précipitations extrêmes ne déclenche des événements plus extrêmes, caractérisés par leur rareté et leur grande ampleur [4].

Face à cette situation, des modèles de circulations générales (MCG) sont utilisés pour mieux comprendre les effets du changement climatique et de les prévenir dans le futur en simulant des processus physiques connus jusqu'à deux cent ans dans le futur. Ces modèles sont utilisés pour comprendre les changements climatiques à l'échelle mondiale et continentale. Cependant, il est bien connu que la résolution spatiale grossière (300 km) des MCG limite l'utilisation fiable de ces données dans la prise de décision et les études d'impact basées sur des modèles [5, 6]. Pour tenir compte de ces phénomènes, des techniques de réduction d'échelle (downscaling) sont appliqués ; elles permettent non seulement de décrire ces phénomènes mais également de fournir des projections climatiques à des échelles spatiales plus fines.Deux principales approches de réduction d'échelle sont généralement utilisées, à savoir dynamique et statistique. La réduction d'échelle dynamique fait référence à l'utilisation de modèles climatiques régionaux (MCR) qui utilisent les conditions aux limites latérales et à grande échelle des MCG pour produire des sorties à plus haute résolution [7, 8]) mais sur une partie du globe. Les méthodes de réduction d'échelle statistique sont quant à elles fondées sur certaines relations entre les variables atmosphériques à grande échelle (prédicteurs) et les variables climatiques locales (prédicants). Un des principaux avantages de ces méthodes est leur faible coût en terme de temps de calcul par rapport aux méthodes dynamiques, elles peuvent donc être facilement appliquées à différents simulations de MCG. L'échelle spatiale beaucoup plus fine permet de prendre en compte les

particularités physiographiques non représentées dans les MCGs et de générer une information climatique en un site spécifique, via l'utilisation des données observées pour calibrer le modèle statistique et qui incorporent directement ou indirectement ces effets locaux [9]. Cette dernière approche peut utiliser des techniques d'apprentissage machine ou " machine learning ". Les méthodes d'apprentissage machine, qui sont issues de l'intelligence artificielle, sont de plus en plus présentes dans les sciences de l'environnement.

Une évaluation de plusieurs méthodes de réduction d'échelle statistique qui utilisent des techniques de "machine learning" est présentée dans ce travail, afin d'améliorer la simulation des précipitations à l'échelle locale par rapport aux informations de grandes échelles fournies. Cette évaluation est réalisée dans la zone de la Casamance où la plupart des activités des populations dépendent de la précipitation.

Il s'agit dans ce travail de caractériser et d'analyser le climat sahélien, d'élaborer des modèles statistiques intelligents permettant d'estimer l'évolution des précipitations dans notre zone d'étude, d'identifier les prédicteurs potentiels pour la prévision et en fin d'évaluer la performance des modèles sur la zone de la Casamance à partir des observations et des séries de ré-analyses disponibles

Le chapitre 1 présente une synthèse des connaissances sur la climatologie ouest africaine, les mécanismes ou processus physiques caractérisant le climat de l'Afrique de l'ouest et sur l'apprentissage machine. Le chapitre 2 donne la présentation de la zone et la méthodologie utilisée dans ce travail. Le chapitre 3 présente les résultats ainsi que les analyses faites sur ces résultats. Enfin la dernière partie est réservée à la conclusion et aux perspectives.

Chapitre 1

Généralité sur la climatologie ouest africaine et sur l'apprentissage automatique

Dans ce chapitre, nous allons faire une description du climat ouest africaine avec un accent particulier sur les processus clés qui le régissent, ensuite nous donnerons une brève définition de apprentissage automatique et les différentes méthodes qui sont utilisées.

1.1 La climatologie ouest africaine

L'analyse des données climatiques montre qu'en Afrique de l'Ouest les dernières décennies ont été caractérisées par des évolutions très marquées, en particulier par des épisodes de sécheresse significative et des inondations. Les conséquences de cette sécheresse ont été importantes sur les ressources en eau, l'agriculture et l'élevage pastoral, qui sont des secteurs clés de l'économie de la région. L'Afrique de l'Ouest et en particulier le Sahel est reconnu comme une région très vulnérable aux changements climatiques. Face à ce constat, il est nécessaire de s'intéresser à la climatologie et à l'évaluation des impacts du changement climatique dans cette région à risque . Cependant, pour comprendre la climatologie, il est important de connaitre la circulation atmosphérique dans cette zone.

1.1.1 La circulation atmosphérique générale

La circulation générale est le mouvement à l'échelle planétaire de la couche d'air entourant la terre qui redistribue la chaleur provenant du soleil en conjonction avec la circulation océanique. C'est l'ensemble des grands mouvements horizontaux et verticaux de l'atmosphère sur toute l'étendue du globe (Figure 1.1) [10].
Le moteur principal des mouvements atmosphériques est le soleil. Celui-ci réchauffe la surface de la Terre, qui réchauffe à son tour l'air ambiant. Des mouvements ascendants se créent, mais en s'élevant, l'air se refroidit, environ 1°C tous les $100m$ dans la troposphère, ce qui le rend négativement instable en atteignant la tropopause. L'air redescend alors vers le sol. Cette circulation constitue un courant de convection. De telles boucles de circulation portent le nom de cellule. Les différentes cellules sont disposées en bandes selon les latitudes : c'est une organisation zonale. Le modèle de circulation générale proposé comporte trois cellules de convection dans chaque hémisphère : une cellule équatoriale dans le sens direct dite cellule de Hadley [11], une cellule à circulation inverse de la précédente dite cellule de Ferrel et une cellule polaire à nouveau à circulation directe. L'Afrique de l'ouest étant située dans la région équatoriale, rassemble les cellules de Hadley. L'air est surchauffé par le soleil et est allégé puis

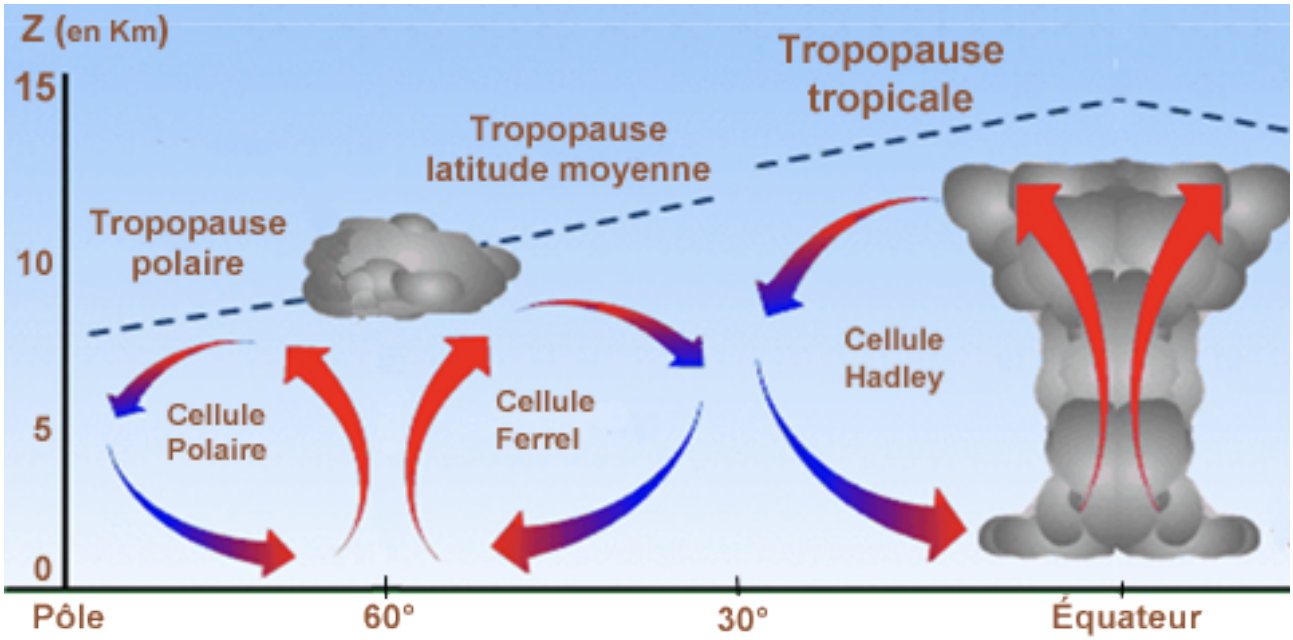


FIGURE 1.1 – Circulation à grande échelle de l'atmosphère, source : https://www.lavionnaire.fr/MeteoCirculation.php

il s'élève vers le haut de la troposphère et aspire l'air situé tout autour.Il engendre ainsi des vents qui convergent vers l'équateur. Sous l'influence de la force de Coriolis, l'air venant du nord est dévié vers la droite, celui venant du sud l'est vers la gauche (ce sont les alizés). Le courant ascendant des alizés se charge en humidité lors de son passage au-dessus des océans. En traversant la troposphère il se refroidit et s'assèche par condensation dans les hautes altitudes, et il perd progressivement de la vitesse. Il ne parvient pas à dépasser l'altitude de la tropopause qui est très stable, mais son débit massique doit être conservé. Ceci n'est possible que si sa trajectoire se courbe sous la forme de vents horizontaux orientés, soit vers le nord, soit vers le sud, selon l'hémisphère, formant ainsi deux cellules convectives. Ces courants ne restent pas dans le plan zonal, car bien avant qu'ils atteignent les pôles, la force de Coriolis a pour effet de dévier leurs trajectoires, systématiquement vers l'est, dans l'hémisphère nord comme dans l'hémisphère sud. Cette pseudo-force, les empêche donc de demeurer dans les plans méridiens et impose une circulation atmosphérique en hélice au sein de cette cellule de Hadley. L'influence de la force de Coriolis limite ainsi l'étendue de cette cellule de Hadley à des latitudes voisines de plus ou moins 30° C [12, 13].

1.1.2 Quelques éléments du système de mousson ouest-africaine

Le climat de l'Afrique de l'ouest est régi par la mousson ouest africaine [14, 15]. Les zones de mousson sont caractérisées par l'opposition de part et d'autre de l'équateur d'une masse continentale et d'une masse océanique, créant ainsi des différences importantes de température et de pression, le vent se dirigeant alors des hautes pressions situées au-dessus de l'océan, et transportant donc des masses d'air chargées en humidité, vers les basses pressions situées au-dessus du continent, venant ainsi alimenter en énergie la Zone de Convergence Inter-Tropicale (ZCIT). Ces systèmes de mousson correspondent donc régionalement à une intensification de la circulation de Hadley. La circulation du système de mousson africain s'organise donc par les anticyclones de Sainte-Hélène et des Açores, situés respectivement sur l'Atlantique tropical sud et nord, et la dépression thermique saharienne que l'on voit centrée vers 20° N sur l'Afrique de l'Ouest, soit environ 10° plus au nord que la ZCIT [16]. Cela conduit à la mise en place de la mousson d'été en Afrique de l'Ouest, par le développement de vents de sud-est humides

issus de l'anticyclone de Sainte-Hélène qui tournent au sud-ouest en passant par l'équateur sous l'effet du changement de sens de la force de Coriolis et de vents de nord-est secs (harmattan) venant du Sahara. Le lieu des pressions minimales dans la dépression thermique continentale représente ainsi la confluence de ces vents de sud-ouest et de nord-est dans les basses couches et est appelé Front intertropical (FIT) [12, 14, 17] .
Le système de mousson ouest africaine est régulé par plusieurs flux [18] qui sont présentés à la Figure 1.2 :

Les fluctuations : Ce sont les fluctuations de la mousson (sud-ouest) chargées d'humidité et l'harmattan (nord-est) vent chaud et sec, provenant du Sahara et correspondant aux alizés de l'hémisphère Nord. Ces deux fluctuations aux directions opposées se rencontrent et convergent ensemble dans les basses couches de l'atmosphère.

Le jet d'est africain (JEA) : Situé entre 600 et 700 hPa, est généré par le gradient thermique, dirigé vers le nord du contient. Il joue un rôle important dans le climat sahélien. Le JEA organise le mouvement ascendant de l'air des systèmes convectifs les plus développés [18] .

Le jet d'est tropical (JET) : Son origine est liée à l'établissement de la mousson indienne notamment aux contrastes thermiques existant en été entre les hauts plateaux du Tibet et les régions océaniques dans le sud-est asiatique. Le JET est observé en Afrique de l'ouest entre 100 et 200 hPa autour de 10◦ N.

Le jet sub tropical JOST : Aux mêmes niveaux de pression que le JET, mais autour de 35◦ N en été, on trouve le JOST, soufflant vers l'Ouest et dont certaines instabilités peuvent conduire à des intrusions d'air sec dans la moyenne troposphère au niveau de la ZCIT, qui coïncident avec de la convection humide organisée à grande échelle [19] .

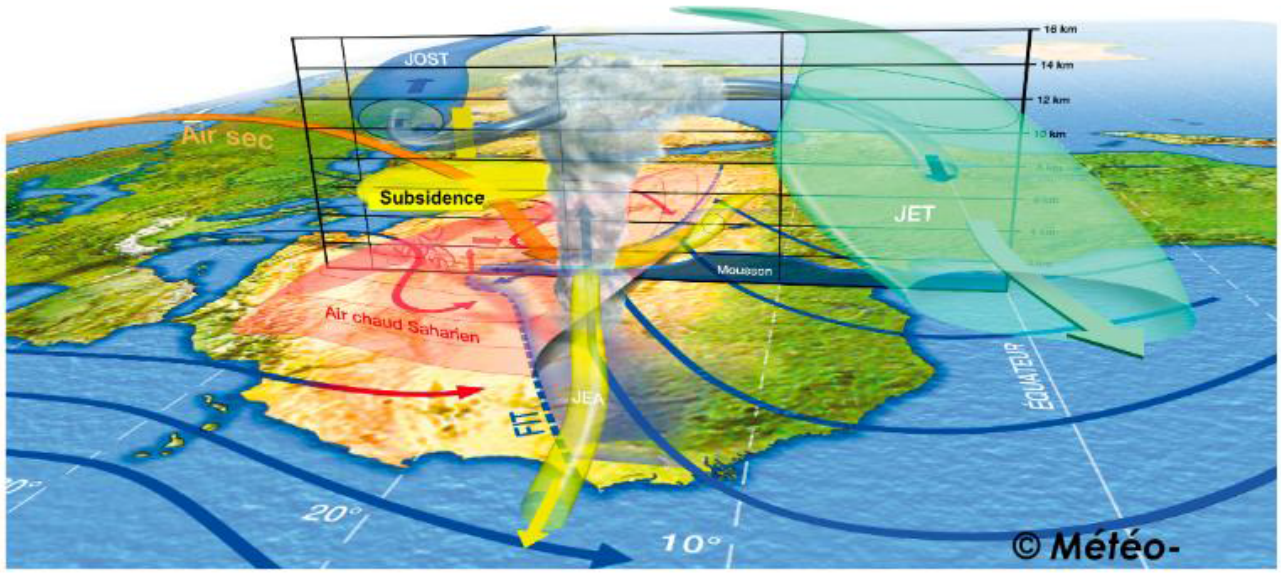


FIGURE 1.2 – Schéma tridimensionnel de la dynamique atmosphérique de la mousson ouest Africaine (Tirée de Lafore et al 2010)

Le régime pluviométrique de l'Afrique de l'ouest est lié au mouvement saisonnier de la zone de convergence intertropicale (ZCIT). La zone semi-aride, qui comprend essentiellement la bande sahélienne et sahélo-saharienne, est marquée par une seule saison des pluies). Le Sahel reçoit la plus grande partie de ses précipitations entre juillet et septembre. Plus au sud, l'alternance de deux saisons pluvieuses et deux saisons sèches marque le climat des pays du golfe de Guinée.

1.1.3 Changement climatique et impacts en Afrique de l'ouest

Le changement climatique désigne l'ensemble des variations des caractéristiques climatiques en un endroit donné, au cours du temps : réchauffement ou refroidissement. Certaines formes de pollution de l'air, résultant d'activités humaines, menacent de modifier sensiblement le climat, dans le sens d'un réchauffement global. Ce phénomène peut entraîner des dommages importants. A l'échelle mondiale, on note une hausse des températures moyennes de l'atmosphère et de l'océan, une fonte massive de la neige et de la glace et une élévation du niveau de la mer [20]. Notre sous-région est exposée aux changements suivants : une augmentation des températures qui atteindra entre 3^0C et 6^0C [17,21] d'ici à la fin du siècle et une irrégularité des précipitations, avec possiblement un retard du début de la saison des pluies ; une augmentation des phénomènes météorologiques extrêmes (canicules, averses orageuses, vents violents) ; une élévation du niveau marin. Si les prévisions sont relativement nettes pour les pays situés au nord de la Gambie, elles sont moins précises plus au sud en ce qui concerne les températures et les précipitations. Le Sahel et l'Afrique de l'ouest sont considérés dans leur ensemble comme des régions de forte sensibilité au changement climatique. En effet la persistance des sécheresses à partir des années 1970 entraînant des déficits pluviométriques assez importants et une évolution des isohyètes vers le sud, ce qui a provoqué d'importantes conséquences dans cette zone ou l'économique dépend fortement de l'agriculture et de l'élevage. Un tiers des populations en Afrique vit dans des zones prédisposés à la sécheresse et sont vulnérables aux impacts des sécheresses [22]. Par exemple, plusieurs millions de personnes souffrent régulièrement des impacts liés à la sécheresse et aux inondations. Ces impacts sont souvent encore aggravés par des problèmes de santé, en particulier la diarrhée, le choléra et la malaria [23]. Les pertes économiques aux sécheresses des années 80 ont été estimées à plusieurs centaines de millions de dollars des États-Unis [24]. En Afrique de l'ouest, les sécheresses ont principalement affecté le sahel, en particulier depuis la fin des années 60 [23, 25]. Les conséquences décrites dans cette partie montrent à quel point, il est important pour la communauté scientifique de comprendre les effets du changement climatique de la sous-région pour pouvoir prévenir d'éventuelles menaces dans le futur. L'évaluation de la vulnérabilité au changement climatique nécessite des informations sur les projections de celui-ci à l'échelle locale ou régionale. Les MCG ayant des projections assez grossières, les techniques de réduction d'échelles sont appliquées pour fournir des projections climatiques à des échelles spatiales plus fines. Dans le cadre de ce travail, l'approche statistique, qui utilise les méthodes d'apprentissage automatique, est utilisée.

1.2 Généralités sur l'apprentissage automatique

L'apprentissage automatique fait partie de ce qui est communément appelé intelligence artificielle (IA) dont les définitions sont très diverses. Cependant on pourrait dire que l' IA est un ensemble de techniques permettant à des machines d'accomplir des tâches et de résoudre des problèmes normalement réservés aux humains et à certains animaux. Les tâches peuvent être simples comme la reconnaissance et la localisation d'objet dans une image, elles requièrent parfois de la planification complexe comme jouer aux échecs, et en fin, les plus compliquées requièrent beaucoup de connaissances et de bon sens commun comme la traduction de texte [26].

Depuis quelques années, on associe presque toujours l'intelligence aux capacités d'apprentissage. C'est grâce à l'apprentissage qu'un système intelligent capable d'exécuter une tâche peut améliorer ses performances avec l'expérience, apprendre à exécuter de nouvelles tâches et acquérir de nouvelles compétences.
Benureau [27] propose cette définition :" *L'apprentissage est une modification d'un comportement sur la base d'une expérience* ".

Dans le cas d'un programme informatique, qui est celui qui nous intéresse dans ce travail, on parle d'apprentissage automatique, ou " machine learning " [27]. On en trouve une première définition dès 1959, grâce à Arthur Samuel [28], l'un des pionniers de l'intelligence artificielle, qui définit le " machine learning " comme le champ d'étude visant à donner la capacité à une machine d'apprendre sans être explicitement programmée (). On peut ainsi opposer un programme classique, qui utilise une procédure et les données qu'il reçoit en entrée pour produire en sortie des réponses, à un programme d'apprentissage automatique, qui utilise les données et les réponses afin de produire la procédure qui permet d'obtenir les secondes à partir des premières.

En 1997 , Tom Mitchell , de l'université de Carnegie Mellon, propose une définition beaucoup plus précise : *" A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E ".*

Ici, l'expérience ce sont les données qui sont fournies à l'ordinateur et qui vont être traiter par des méthodes statistiques (ce sont les taches) puis la performance de ces méthodes est évaluée. Grâce à cette évaluation, les taches sont améliorées (c'est l'optimisation). Le " machine learning" sert à faire des prévisions, de la classification et de la segmentation automatiques en exploitant des données en général multidimensionnelles, il relève d'une approche probabiliste. Les outils du " machine learning " servent à exploiter les gros volumes de données des entreprises, autrement dit le " big data ". Les algorithmes ne sont pas tous destinés aux mêmes usages. On les classe usuellement selon deux composantes :

Le mode d'apprentissage : on distingue les algorithmes supervisés des algorithmes non supervisés ;

Le type de problème à traiter : on distingue les algorithmes de régression de ceux de classification.

1.2.1 Méthodes d'apprentissages automatiques

Les méthodes d'apprentissage automatique ou " machine learning" [29, 30]les plus adoptées sont l'apprentissage supervisé et l'apprentissage non supervisé. La différence entre ces deux algorithmes est fondamentale.

— Les algorithmes supervisés extraient de la connaissance à partir d'un ensemble de données contenant des couples entrée-sortie. Ces couples sont déjà « connus », dans le sens où les sorties sont définies à priori. La valeur de sortie peut être une indication fournie par un expert : par exemple, des valeurs de vérité de type OUI/NON. Ces algorithmes cherchent à définir une représentation compacte des associations entrée-sortie, par l'intermédiaire d'une fonction de prédiction [31].

— A contrario, les algorithmes non supervisés n'intègrent pas la notion d'entrée-sortie. Toutes les données sont équivalentes (on pourrait dire qu'il n'y a que des entrées). Dans ce cas, les algorithmes cherchent à organiser les données en groupes. Chaque groupe doit comprendre des données similaires et les données différentes doivent se retrouver dans des groupes distincts. Dans ce cas, l'apprentissage ne se fait plus à partir d'une indication qui peut être préalablement fournie par un expert, mais uniquement à partir des fluctuations observables dans les données.

1.2.2 Différentes approches d'apprentissage automatique

Les approches de l'apprentissage automatique sont continuellement développées. Nous présentons ici les grandes familles de techniques , sachant que certaines des méthodes utilisées dans le cadre de ce travail seront vues plus en détails dans la partie méthodologie.

1.2.2.1 Arbre de décision :

Les arbres de décision sont utilisés pour représenter visuellement les décisions et montrer ou éclairer la prise de décision. Dans le cas du " machine learning " ces arbres sont utilisés comme modèles prédictifs. L'objectif est de créer un modèle qui prédira la valeur d'une cible en fonction de variables d'entrée. Dans le modèle, les attributs des données qui sont déterminés par l'observation sont représentés par les branches, tandis que les conclusions sur la valeur cible des données sont représentées dans les feuilles. Lors de l'apprentissage d'un arbre, les données sources sont divisées en sous-ensembles en fonction d'un test de valeur d'attribut, qui est répété récursivement sur chacun des sous-ensembles dérivés. Une fois que le sous-ensemble d'un nœud a la valeur équivalente à sa valeur cible, le processus de récursions sera terminé.

1.2.2.2 " Deep learning " :

Le " deep learning " ou apprentissage profond, permet d'aller plus loin que le " machine learning " pour reconnaître des objets complexes comme les images, l'écriture manuscrite, la parole et le langage [32]. Le " deep learning " exploite des réseaux de neurones multicouches, sachant qu'il en existe de très nombreuses variantes. Le " deep learning " permet aussi de générer des contenus ou d'améliorer des contenus existants, comme pour colorier automatiquement des images en noir et blanc. Il est profond parce qu'il exploite des réseaux de neurones avec de nombreuses couches de filtres. Par contre, le " deep learning " n'est pas dédié exclusivement au traitement de l'image et du langage. Il peut servir dans d'autres environnements complexes.

Dans ce chapitre, il s'est agi de définir, d'expliciter la climatologie ouest africaine et l'apprentissage automatique en générale et présenter les concepts clés faisant objet de cette étude, notamment : la circulation atmosphérique, la mousson, etc . Il s'est agi aussi, bien évidemment, de présenter quelles que approches de l'apprentissage automatique en vue d'asseoir une partie de la théorie en rapport avec notre thème de recherche.

Chapitre 2

Domaine d'étude, données et méthodologie utilisée

Ce chapitre est consacré à la présentation du domaine d'étude et de la description des différentes données et méthodes utilisées au cours de ce travail.

2.1 Domaine d'étude

Cette présente étude est menée sur la zone de la Casamance localisée au Sud du Sénégal. Le Sénégal est situé dans la boite comprise entre 12°8 Nord et 16°41 Nord de latitude, 11°21 Ouest et 17°32 Ouest de longitude dans la zone intertropicale (Figure 2.1). Il est limité par l'océan Atlantique à l'ouest, la Mauritanie au nord, à l'est par le Mali, la Guinée et la Guinée-Bissau au sud. La Gambie est quasi enclavée dans le Sénégal et pénètre jusqu'à plus de $300km$ à l'intérieur des terres. Le climat de la zone est tropical de type soudano-guinéen et sec avec deux saisons : une saison sèche et une saison humide qui dure environ 4 mois avec un gradient de pluie plus intense au sud du pays (en Casamance). La Casamance a une superficie d'environ 28.300 km^2, soit $\frac{1}{7}^{\text{ème}}$ du territoire national et contient les villes urbaines comme Ziguinchor, Sédhiou et Kolda . Le fleuve qui lui donne son nom coule sur $300km$ de longueur et prend sa source à l'est de Kolda près du Fafakourou en haute Casamance [33].
La Casamance est limitée à l'ouest par l'océan Atlantique, à l'est par le Koulountou, affluent du fleuve Gambie, au Nord par la Gambie, au Sud par la Guinée-Bissau et la Guinée. Elle est divisée en trois zones géographiques : la basse Casamance qui comprend la région de Ziguinchor qui s'étend de l'océan atlantique au fleuve Soungrougou, la moyenne Casamance qui comprend la région de Sédhiou, celle-ci est située entre la basse et la haute Casamance et en fin la haute Casamance, elle comprend la région de Kolda. La pluviométrie de la zone diminue en allant vers la haute Casamance. Le climat de la région est de type soudano guinéen recevant des précipitations qui s'étalent de Juin en Octobre avec une intensité maximale en août et septembre, et une saison sèche qui couvre la période de Novembre à Mai. Les précipitations moyennes varient de 700 à $1300mm$. Les températures moyennes mensuelles les plus basses sont enregistrées entre décembre et janvier et varient entre 25 à 30°C, les plus élevées sont notées entre Mars et Septembre avec des variations de 30 à 40°C.

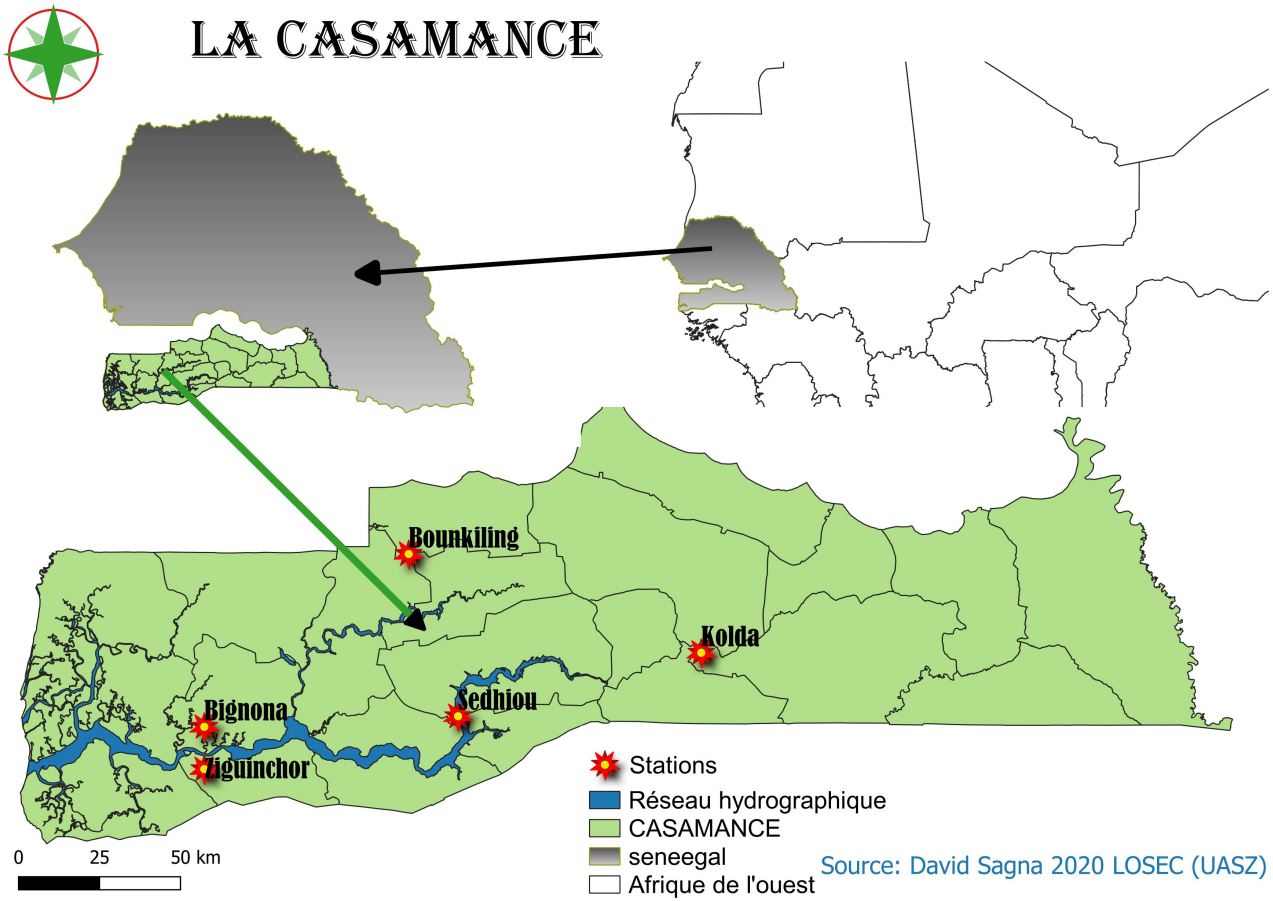


FIGURE 2.1 – Domaine d'étude

2.2 Données utilisées

Des séries chronologiques de précipitations mensuelles allant de 1950 à 2015 de cinq stations (Tableau 2.1) localisées dans la zone de la Casamance obtenues de l'Agence Nationale de l'Aviation Civile et de la Météorologie (ANACIM) sont utilisées. Des données de réanalyses du " National Center of Environmental Prediction/National Center for Atmospheric reseach" (NCEP/NCAR) représentant de résolution 2.2∗2.5 des paramètres météorologiques de grandes échelles sont également extraits sur ces stations. Ces données sont disponibles gratuitement en ligne dans les bases de données .

Station	Longitude	latitude
Ziguichor	16.27° W	12.55° N
Bignona	16.27° W	12.67° N
Kolda	14.96° W	12.88° N
Bounkiling	15.69° W	13.16° N
Sédhiou	15.55° W	12.70° N

TABLE 2.1 – Stations pluviométriques de la Casamance

2.3 Méthodologie utilisée

2.3.1 Techniques d'apprentissage machines utilisées

Dans cette étude, pour chaque station, des modèles de réduction d'échelle ont été appliqués en utilisant quatre techniques d'apprentissage machine, à savoir le machine à vecteur de soutien ou "support vector machine" (SVM), l'extrême machine learning (ELM), le k-plus proche voisin ou "K Nearest Neighbor" (KNN) et la régression poursuit par projection ou "Projection pursuit regression" (PPR). En effet plusieurs études sur la réduction d'échelle des précipitations suggèrent d'appliquer différentes méthodes plutôt qu'une seule pour obtenir des résultats robustes [34]. Nous verrons plus en détails ces différentes techniques dans les paragraphes qui suivent.

2.3.1.1 K Plus Proches Voisins ou "K Nearest Neighbor"

Le K Plus Proches Voisins ou "K Nearest Neighbor" (KNN) en Anglais, est un algorithme d'apprentissage automatique supervisé qui place un nouveau point de données dans la classe cible, en fonction des caractéristiques de ses points de données voisins. Il peut être utilisé aussi bien pour la régression que pour la classification [35,36]. Son fonctionnement peut être assimilé à l'analogie suivante "dis moi qui sont tes voisins, je te dirais qui tu es...".
L'algorithme KNN présente les caractéristiques suivantes :

- C'est un algorithme d'apprentissage supervisé qui utilise un ensemble de données d'entrées étiquetées pour prédire la sortie des points de données ;
- C'est l'un des algorithmes d'apprentissage automatique les plus simples et il peut être facilement implémenté pour un ensemble varié de problèmes ;
- Il est principalement basé sur la similitude des fonctionnalités. KNN vérifie à quel point un point de données est similaire à son voisin et classe le point de données dans la classe à laquelle il est le plus similaire ;
- Contrairement à la plupart des algorithmes, KNN est un modèle non paramétrique, ce qui signifie qu'il ne fait aucune hypothèse sur l'ensemble de données. Cela rend l'algorithme plus efficace car il peut gérer des données réalistes,
- C'est un algorithme paresseux, cela signifie qu'il mémorise l'ensemble de données d'apprentissage au lieu d'apprendre une fonction discriminante à partir des données d'apprentissage.

Sur la Figure 2.2 , nous avons deux classes de données, à savoir la classe A (carrés) et la classe B (triangles). L'énoncé du problème consiste à affecter le nouveau point de données d'entrée à l'une des deux classes à l'aide de l'algorithme KNN. Gangopadhyay et al. 2005 [37] ont utilisé le modèle KNN pour réduire la température et les précipitations à l'échelle locale aux États-Unis. La première étape de l'algorithme KNN consiste à définir la valeur de «K» qui représente le nombre de voisins les plus proches et d'où le nom K-plus proches voisins (noté KNN en Anglais).

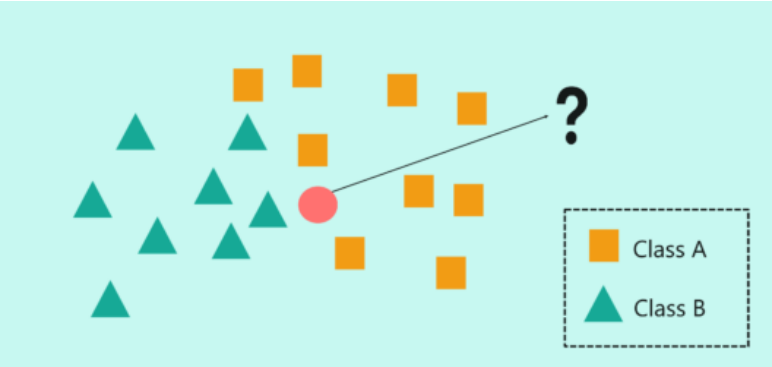


FIGURE 2.2 – Ensemble de données séparées en deux classes, classe A et B, source : https://www.edureka.co/blog/support-vector-machine-in-r

Dans l'image ci-dessus, on a défini la valeur de K comme étant égale à 3. Cela signifie que l'algorithme considérera les trois voisins les plus proches du nouveau point de données afin de décider de la classe de ce nouveau point de données. La proximité entre les points de données est calculée en utilisant des mesures telles que la distance euclidienne.
À K = 3, les voisins comprennent deux carrés et 1 triangle. Donc, si on devait classer le nouveau point de données en fonction de K = 3, il serait alors attribué à la classe A (carrés) (Figure 2.3).

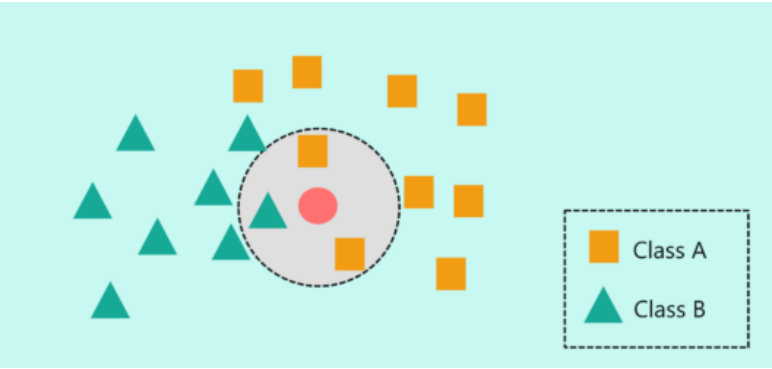


FIGURE 2.3 – Méthode K-plus proches voisins avec K=3, Source : https://www.edureka.co/blog/support-vector-machine-in-r/

KNN utilise des mesures simples telles que la distance euclidienne pour résoudre des problèmes complexes, c'est l'une des raisons pour lesquelles KNN est couramment utilisé.

2.3.1.2 Support vecteur machine

Le machine à vecteur de soutien ou "Support vector machine" (SVM) en Anglais [38] peut être représenté comme un réseau de neurones à deux couches [39] qui peut être utilisé pour la régression linéaire et non linéaire. Le principe du SVM est basé sur la théorie de l'apprentissage statistique et de la méthode de minimisation de risque structurel. C'est de trouver l'hyperplan de séparation optimal qui sépare une classe de données à une autre (maximisation de la marge et minimisation de l'erreur) [40].

Considérons l'exemple suivant (Figure 2.4). On se place dans le plan, et l'on dispose de deux classes : les ronds rouges et les ronds bleus, chacune occupant une région différente du plan. Cependant, la frontière entre ces deux régions n'est pas connue. Ce que l'on veut, c'est que quand on lui présentera un nouveau point dont on ne connaît que la position dans le plan, l'algorithme de classification sera capable de prédire si ce nouveau point est un rond rouge

ou un rond bleu. Pour cela, il faut être capable de trouver la frontière entre les différentes catégories. Si on connaît la frontière, savoir de quel côté de la frontière appartient le point, et donc à quelle classe il appartient.

Pour que le SVM puisse trouver cette frontière, il est nécessaire de lui donner des données d'entraînement. En l'occurrence, on donne au SVM un ensemble de points, dont on sait déjà si ce sont des rond rouges ou des ronds bleus, comme dans la Figure 1. A partir de ces données, le SVM va estimer l'emplacement le plus plausible de la frontière : c'est la période d'entraînement, nécessaire à tout algorithme d'apprentissage automatique.

Une fois la phase d'entraînement terminée, le SVM a ainsi trouvé, à partir de données d'entraînement, l'emplacement supposé de la frontière. En quelque sorte, il a «appris» l'emplacement de la frontière grâce aux données d'entraînement. Le SVM est maintenant capable de prédire à quelle catégorie appartient une entrée qu'il n'avait jamais vue avant, et sans intervention humaine : c'est là tout l'intérêt de l'apprentissage automatique supervisé.

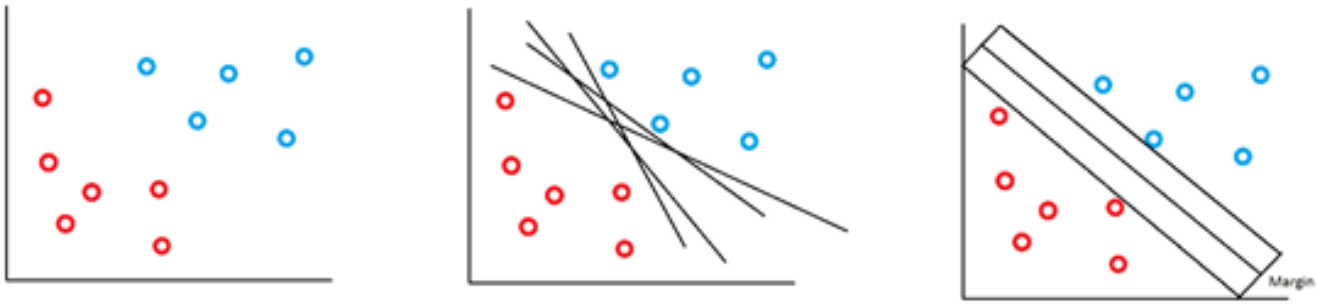


FIGURE 2.4 – ensemble de données linéairement séparables,
source : https://www.sciencedirect.com/science/article/abs/pii/S0023643816302328

Comme vous pouvez le constater dans la figure ci-dessus, pour notre problème, le SVM a choisi une ligne droite comme frontière. C'est parce que, le SVM est un classificateur linéaire. Quand on a un ensemble de données d'entraînement, il existe plusieurs lignes droites qui peuvent séparer nos catégories. La plupart du temps, il y en a une infinité [41]. Alors, laquelle choisir ? Un SVM dans notre exemple va placer la frontière aussi loin que possible des ronds bleus, mais également aussi loin que possible des ronds rouges. Comme on le voit dans la Figure 2.4, c'est bien la frontière la plus éloignée de tous les points d'entraînement qui est optimale, on dit qu'elle a la meilleure capacité de généralisation. Ainsi, le but d'un SVM est de trouver cette frontière optimale, en maximisant la distance entre les points d'entraînement et la frontière. Les points d'entraînement les plus proches de la frontière sont appelés vecteurs support, et c'est d'eux que les SVM tirent leur noms.

Dans le cas ou les données ne sont pas linéairement séparables (Figure 2.5), les SVM utilisent une méthode appelée kernel trick, ou astuce du noyau en français qui consiste à faire une transformation. De façon plus générale que dans les exemples donnés précédemment, les SVM ne se bornent pas à séparer des points dans le plan. Ils peuvent, en fait, séparer des points dans un espace de dimension quelconque. Fondamentalement, un SVM cherchera simplement à trouver un hyperplan qui sépare les deux classes de notre problème.

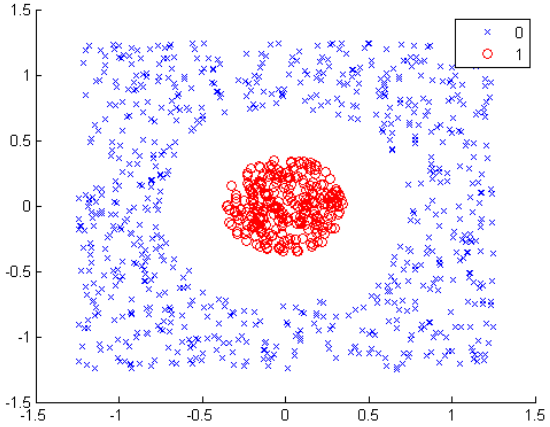


FIGURE 2.5 – Ensembles de données non-linairement séparables, source : https://quantdare.com/svm-versus-a-monkey/

Dans un espace vectoriel de dimension finie n, un hyperplan est un sous-espace vectoriel de dimension $n-1$. Ainsi, dans un espace de dimension 2 un hyperplan sera une droite, dans un espace de dimension 3 un hyperplan sera un plan, etc. Soit un espace vectoriel E de dimension n. L'équation caractéristique d'un hyperplan est de la forme :

$$w_1x_1 + w_2x_2 + \cdots + w_nx_n = 0 \tag{2.1}$$

où $w_1, \ldots, w_n$ sont des scalaires. Par définition, tout vecteur $x = (x_1 \cdots x_n) \in E$ vérifiant l'équation appartient à l'hyperplan. Par exemple, en dimension 2, $ax + by = 0$ est l'équation caractéristique d'une droit vectorielle (qui passe par l'origine). C'est pourquoi on utilise un hyperplan affine qui n'est pas obligé de passer par l'origine.

Ainsi dans un espace de dimension $\mathbb{R}^n$, un SVM calculera un hyperplan d'équation :

$$w_1x_1 + w_2x_2 + \cdots + w_nx_n = 0 \tag{2.2}$$

ainsi qu'un scalaire b. Le vecteur $w = (w_1 \cdots w_n)$ est appelé **vecteur poids**, le scalaire b est appelé **biais**. Une fois l'entraînement terminé, pour classer une nouvelle entrée $x = (a_1 \cdots a_n) \in \mathbb{R}^n$, le SVM regardera le signe de l'équation de l'hyperplan :

$$h(x) = w_1x_1 + w_2x_2 + \cdots + w_nx_n = \sum_{i=1}^{n} w_i.a_i + b = w^T \cdot x + b \tag{2.3}$$

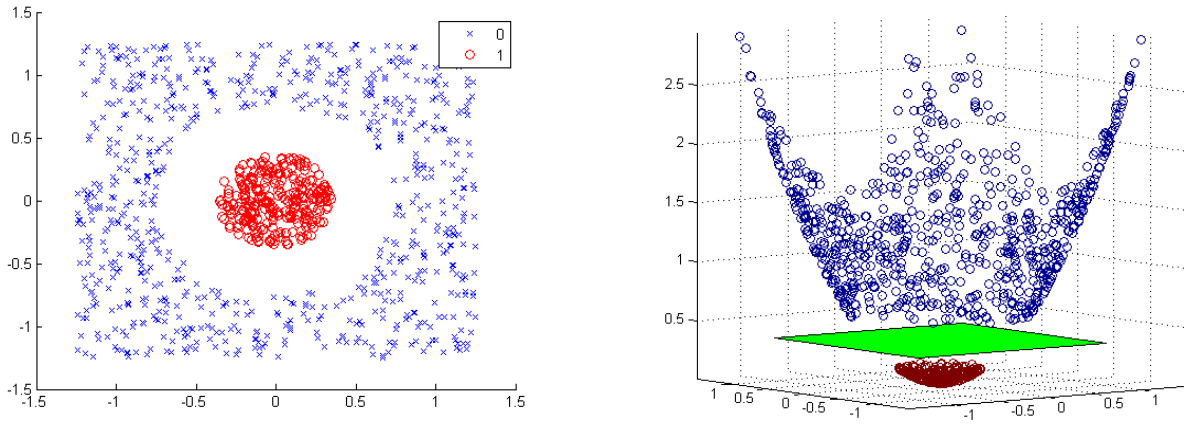
Par exemple, dans le cadre de la réduction d'échelle statistique, un ensemble de données prédicteurs à deux classes pouvant être séparées linéairement, pourrait être classé dans deux classes en satisfaisant les équations suivants :

$$\begin{cases} h(x) \geq 0 & \Rightarrow x \in \text{classe 1} \\ h(x) < 0 & \Rightarrow x \in \text{classe 2} \end{cases} \tag{2.4}$$

Comme on l'a vu précédemment, on choisira l'hyplerplan qui maximise la marge, c'est-à-dire la distance minimale entre les vecteurs d'entraînement et l'hyperplan. De tels vecteurs situés à la distance minimale sont appelés vecteurs supports. C'est de là que vient le nom des SVM. En effet, on peut prouver que l'hyperplan donné par un SVM ne dépend que des vecteurs supports, c'est donc tout naturellement qu'on a appelé cette structure les Support Vectors Machines,

c'est-à-dire les machines à vecteurs support.

Revenons à notre cas ou les données ne sont pas linéairement séparables. Des données sont non linéairement séparables quand il n'existe pas d'hyperplan capable de séparer correctement les deux classes. De façon générale, il est courant de ne pas pouvoir séparer les données parce que l'espace est de trop petite dimension. Si l'on arrivait à transposer les données dans un espace de plus grande dimension, on arriverait peut-être à trouver un hyperplan séparateur (Figure 2.6).



(a) (b)

FIGURE 2.6 – Méthode SVM, cas des données non linéairement séparable, transposé des données de l'espace de description (a) vers l'espace de redescription (b), source : https://quantdare.com/svm-versus-a-monkey/

Plus formellement, l'idée de cette redescription du problème est de considérer que l'espace actuel, appelé espace de description, est de dimension trop petite ; alors que si on plongeait les données dans un espace de dimension supérieure, appelé espace de redescription, les données seraient linéairement séparables. Les points sont redistribués depuis l'espace de description vers l'espace de redescription, à l'aide d'une fonction φ, nécessairement non-linéaire.
On définit l'opération de redescription des points x de E vers E' par :

$$\begin{aligned} \varphi : \quad & E \longrightarrow E' \\ & x \longmapsto \varphi(x) \end{aligned}$$

Dans ce nouvel espace E', on va entrainer le SVM, comme nous l'aurions fait dans l'espace E. Si les données y sont linéairement séparables, c'est gagné ! Par la suite, si l'on veut classer x, il suffira de classer $\varphi(x)$: on obtient un SVM fonctionnel.
Mais plus la dimension augmente plus le calcul devient long. Quand on pose le problème de l'optimisation quadratique dans l'espace E', on s'aperçoit que les seules apparitions sont de la formes : $\varphi(x_i)^T.\varphi(x_j)$. Par conséquent, il n'y a pas besoin de connaître expressément E', ni même φ, il suffit de connaître toutes les valeurs $\varphi(x_i)^T.\varphi(x_j)$.

On appelle donc fonction noyau la fonction $K : E \to E' \to \mathbb{R}$ défini par de la façon suivante $K(x, x') = \varphi(x_i)^T.\varphi(x_j)$. A ce moment, le calcul de l'hyperplan séparateur dans E' ne nécessite ni la connaissance de E', ni de φ, mais seulement de K. Grâce au théorème de Mercer, on sait qu'une condition suffisante pour qu'une fonction K soit une fonction noyau est que K soit continue, symétrique et semi-définie positive. Ainsi, il est possible d'utiliser n'importe quelle fonction noyau afin de réaliser une redescription dans un espace de dimension supérieure. La

fonction noyau étant définie sur l'espace de description E (et non sur l'espace de redescription E', de plus grande dimension), les calculs sont beaucoup plus rapides. Les noyaux les plus utilisées sont : le noyaux polynomial,le noyau gaussien, le noyau rationnel et le noyau laplacien. Voici un lien pour plus de détails.

2.3.1.3 Extrem Learning Machine (ELM)

L'extrême " machine learning " (ELM) est un nouvel algorithme d'apprentissage pour les réseaux de neurones à action directe à couche cachée unique [42, 43] . L'essence d'ELM est que les paramètres d'apprentissage des nœuds cachés, y compris les poids d'entrée et les biais, sont assignés de manière aléatoire et n'ont pas besoin d'être réglés tandis que les poids de sortie peuvent être déterminés analytiquement par la simple opération inverse généralisée. Le seul paramètre à définir est le nombre de nœuds masqués. Par rapport aux autres algorithmes d'apprentissage, ELM offre une vitesse d'apprentissage extrêmement rapide, de meilleures performances de généralisation et une intervention humaine minimale. ELM a été appliquée avec succès à de nombreuses applications, telles que les problèmes de classification et de régression. Cependant, un des problèmes de cette méthode ELM est que la classification des données peut ne pas être optimale pour l'apprentissage, les paramètres des nœuds cachés sont attribués de manière aléatoire alors qu'ils restent inchangés pendant la phase de formation. Ainsi, certains échantillons peuvent être mal classés par l'ELM, en particulier ceux qui sont proches de la limite de la classification. On constate également que L'ELM a tendance à exiger plus de neurones cachés que les neurones conventionnels des algorithmes basés sur le réglage dans de nombreux cas. Pour surmonter les lacunes susmentionnées de l'ELM, certains chercheurs ont proposé plusieurs variantes de l'ELM.
Le plus simple algorithme d'apprentissage ELM a un modèle de la forme :

$$\mathbf{Y} = W_2\sigma(W_1 x) \tag{2.5}$$

où W_1 est la matrice des pondérations d'entrées à couche cachée, σ est la fonction d'activation et W_2 est la matrice des pondérations de sortie à couche cachée.

2.3.1.4 Projection pursuit regression

Le Projection pursuit regression (PPR) ou régression poursuite par projection est proposée par [44] . Il a la forme d'un modèle additive, Ce modèle adapte les modèles additifs en ce qu'il projette d'abord la matrice de données des variables explicatives dans la direction optimale avant d'appliquer des fonctions de lissage à ces variables explicatives.

Supposons $X^T = (X_1, X_2, X_3, ..., X_p)$ est un vecteur avec p variables. Y est la variable de réponse correspondante.
$w_m, m = 1, 2, ..., M$ est un vecteur de paramètre avec p éléments.

$$f(X) = \sum_{m=1}^{M} Gm(w_m^T X) \tag{2.6}$$

La nouvelle fonction $V_m = w_m^T X$ est une combinaison linéaire de variable d'entrée X. Le modèle additif est basé sur la nouvelle fonction. Ici w_m est un vecteur unitaire, et la nouvelle fonction V_m est en fait la projection de X sur w_m. Il projette l'espace $p - dimensionnel$ des variables indépendantes sur le nouvel espace $M - dimensionnel$ des caractéristiques. Cette méthode est similaire à l'analyse en composantes principales, sauf que la composante principale est la projection orthogonale, mais elle n'est pas nécessairement orthogonale ici.

Il s'agit essentiellement d'effectuer une transformation non linéaire de la combinaison linéaire. Vous pouvez utiliser cette méthode pour saisir les différentes relations.

Le PPR était une idée nouvelle qui a conduit au début du modèle de réseau de neurones. L'idée technique de base derrière l'apprentissage approfondi existe depuis des décennies. Cependant, pourquoi a-t-elle pris son essor ces dernières années ?
Voici les principaux moteurs de cette montée en puissance :

Premièrement, grâce à l'explosion des données de ces dernières années le " big data ". Les algorithmes d'apprentissage traditionnels, comme la régression logistique, la machine à vecteurs de soutien, la forêt aléatoire, ne peuvent pas tirer efficacement parti d'une telle quantité de données.

Deuxièmement, la puissance de calcul croissante nous permet d'entraîner un grand réseau neuronal sur un CPU ou un GPU à l'aide du " big data ". L'échelle des données et la capacité de calcul permettent de réaliser de nombreux progrès, mais l'innovation algorithmique est également un moteur important. Nombre de ces innovations visent à accélérer l'optimisation des réseaux de neurones. L'un des exemples est l'utilisation de ReLU comme fonction d'activation de la couche intermédiaire au lieu de la fonction sigmoïde précédente. Ce changement a rendu le processus d'optimisation beaucoup plus rapide car la fonction sigmoïde précédente souffre d'un gradient de fuite.

2.3.1.5 Implémentation des modèles

La modélisation a été entièrement faite avec le logiciel **R**. Ce dernier est essentiellement un langage de programmation statistique à source ouverte, utilisé principalement dans le domaine de la science des données. Dans cette étude, nous avons utiliser le " **paquet Caret**". Le paquet Caret est également connu sous le nom de "Formation à la classification et à la régression " ; il possède plusieurs fonctions qui aident à construire des modèles prédictifs. Il contient des outils pour le fractionnement des données, le pré-traitement, la sélection des caractéristiques, le réglage, les algorithmes d'apprentissage non supervisés, etc. Le package caret est très utile car il nous fournit un accès direct à diverses fonctions pour entraîner notre modèle avec divers algorithmes d'apprentissage automatique tels que KNN, SVM, arbre de décision, régression linéaire, etc.

Lorsque vous avez un nouvel ensemble de données, il est judicieux de visualiser les données à l'aide d'un certain nombre de techniques graphiques différentes afin de les examiner sous différents angles. La même idée s'applique à la sélection du modèle. Vous devez utiliser différentes façons d'examiner la précision estimée de vos algorithmes d'apprentissage automatique afin de choisir le ou les deux meilleurs.

Une fois les données importées dans R, ils sont séparés en données d'entrainement et de test. Ensuite, avant d'entraîner nos modèles, on implémente d'abord la méthode "**trainControl()**". Elle contrôle tous les paramètres généraux de calcul qui nous permettent d'utiliser la fonction "**train ()**" fournie par le paquet caret. Cette fonction nous permet d'entraîner les données.

La méthode "**trainControl ()**" ici, prend 3 paramètres.

1. Le paramètre " method " définit la méthode de ré-échantillonnage dans cette étude, nous utiliserons la méthode répétée ou la méthode de validation croisée répétée.

2. Le paramètre suivant est le " number ", il contient essentiellement le nombre d'itérations de rééchantillonnage.
3. Le paramètre " repeats " contient les ensembles à calculer pour notre validation croisée répétée.

Nous utiliserons une validation croisée répétée avec 10 plis et 10 répétitions, une configuration standard commune pour comparer les modèles.

Une fois les modèles entraînés, ils sont ajoutés à une liste et une fonction "**resamples ()**" est appliquée sur la liste pour le rééchantionnage. Cette fonction vérifie que les modèles sont comparables et qu'ils ont utilisés le même schéma de formation (configuration **trainControl**). Cet objet contient les métriques d'évaluation pour chaque pli et répétition pour chaque algorithme à évaluer.

2.3.2 Sélection des prédicteurs et développement de modèle de réductions d'échelle

Un ensemble de prédicteurs probables communs à toutes les stations a été choisie sur la base des études antérieures sur la réduction d'échelle des paramètres tels que les précipitations, l'évaporation, les températures, etc. Ces prédicteurs probables sont résumés sur le Tableau 2.2. Les prédicteurs les plus probables sont les prédicteurs potentiels.

Prédicteur probable	Description	Hauteurs (hPa)
air	Température de l'air	1000
spl	Pression	1000
pr_wtr	Teneur en eaux appréciable	1000
hgt	Hauteur du geo-potentiel	850
		500
		250
		200
omega	Vent zonal	850
		500
		250
		200
rhum	Humidité relatif	1000
		850
		500
shum	Humidité spécifique	850
		500

TABLE 2.2 – Ensemble de prédicteurs probables

En général, la sélection des prédicteurs potentiels dépend du prédictant à estimer et de la région étudiée [7, 45, 46]. Tous les types de prédicteurs peuvent être utilisés pour la réduction d'échelle à condition qu'ils présentent une corrélation acceptable avec les prédictants [47–49]. Toutefois, l'utilisation d'un grand nombre de prédicteurs peut ne pas aboutir à de meilleures relations qu'un prédicteur défini avec une taille parcimonieuse. Il existe plusieurs façons de sélectionner les prédicteurs si un grand nombre d'entre eux est disponible [40]. Dans cette étude, la méthode de la corrélation de Pearson a été appliquée sur l'ensemble des données de

ré-analyses afin d'identifier les prédicteurs potentiels les plus significatifs.

Pour la mise au point du modèle de prédiction, la série temporelle de nos données a été divisé en deux ensembles de telle sorte que les $\frac{2}{3}$ des données (période $1950 - 1993$) soient réservées à la période d'entrainement du modèle et le reste (période 1994-2015) à la période de test.

2.3.3 Paramètres statistiques

Les paramètres utilisés pour évaluer la performance des modèles choisis pour notre étude, mais également pour leur comparaison sont :

— **Le coefficient de corrélation** (R^2) qui permet d'évaluer l'intensité de la relation statistique entre les prédicteurs et le prédicant, soit la capacité du modèle à reproduire la variabilité de la précipitation à l'échelle locale en se basant sur des variables météorologiques de grandes échelles choisies. Mais également permet de traduire la relation entre la variable observée et celle simulée par le modèle. R^2 est compris entre 0 et 1, 1 signifiant une relation parfaite avec les fluctuations de la variable considérée, et 0 un lien nul entre la variable simulée et la variable observée.

— Le $RMSE$ **(Root Mean Square Error)** est calculée à partir des valeurs d'observations et ensuite moyennées pour toutes les simulations faites avec les différents modèles. il mesure la différence entre la simulation et les observations.

— Le biais (B) ou l'erreur du biais qui mesure la tendance moyenne des valeurs simulées à être plus petites ou plus grandes que celle des observations. La valeur optimale du biais est 0.0 avec une faible amplitude indiquant une simulation parfaite du modèle. Des valeurs positives indiquent une surestimation, alors que les valeurs négatives indiquent une sous-estimation.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}[y_i(i) - x_i(i)]^2}{\sum_{i=1}^{n}[x_i(i) - X_i]^2} \tag{2.7}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[y_i(i) - x_i(i)]^2} \tag{2.8}$$

$$B = \frac{\sum_{i=1}^{n}(y_i(i) - x_i(i))}{\sum_{i=1}^{n} x_i(i)} \tag{2.9}$$

Où n est le nombre d'observations, x les variables d'observations et y les variables simulées.

Chapitre 3

Résultats et Discussions

Dans cette partie du travail, les résultats obtenus sont présentés et les performances des modèles seront analysées et discutées.

3.1 Selection des prédicteurs explicatifs

Dans cette étude, les informations requises, telles que les données de réanalyse mensuelles, ont été obtenues auprès du NCEP/NCAR et les données de précipitations des stations sélectionnées ont été utilisées pendant les périodes d'entrainement et de test des outils de réduction d'échelle. L'ensemble des données de réanalyse du NCEP/NCAR est issu d'un modèle de MCG [50]. Les modèles de réduction d'échelle ont été développés en utilisant les modèles KNN, ELM, SVM et PPR. Afin d'obtenir les prédicteurs probables à partir des paramètres climatiques, l'analyse de la corrélation de Pearson [51] a été utilisé [40, 48, 52]. Les variables atmosphériques optimales finales à grande échelle de l'ensemble des données du NCEP/NCAR présentant des corrélations statistiquement significatives avec un niveau de confiance d'au moins 60% ont été choisis comme prédicteurs potentiels (Figure 3.1).

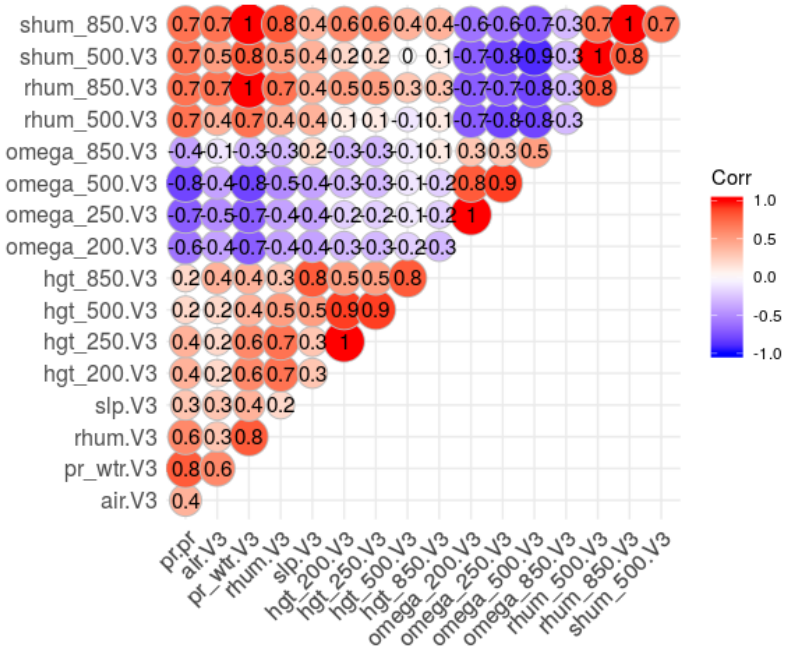


FIGURE 3.1 – Corrélation de Pearson (exemple de la station de Ziguinchor)

La matrice de la Figure 3.1 révèle que l'humidité spécifique (shum) aux niveaux 850 et 500 hPa ont une corrélation positive avec l'humidité relatif (rhum) aux même niveaux et à 1000 hPa. On a aussi la teneur en eau précipitable (pr_wtr) qui a aussi une corrélation positive à 1000 hPa. Au contraire, les valeurs d'omega aux niveaux 200, 250, et 500 hPa ont une corrélation négative par rapport aux variable citées ci-dessus. La précipitation observée dans toutes les stations est beaucoup plus corrélées avec les paramètres atmosphériques à un niveau élevé qu'en surface sauf pour pr_wrt.
La température de l'air a une corrélation en-dessous du niveau de confiance, mais dans la littérature ce paramètre est toujours utilisée pour la réduction d'échelle des précipitations. C'est pour cette raison qu'il est aussi utilisé dans cette étude. Les prédicteurs utilisés sont présentés dans le Tableau 3.1.

TABLE 3.1 – Prédicteur du NCEP reanalysis

Predicteurs	Hauteurs hPa
shum	850
	500
rhum	1000
	850
	500
omega	500
	250
	200
pr_wrt	1000
air	1000

3.2 Evaluation des modèles au niveau des différentes stations

Après la sélection des prédicteurs, les modèles sont appliqués à chaque station pour modéliser les précipitations. Les modèles sont appliqués aux données des stations de Ziguinchor, Bignona, Kolda, Sedhiou et Bounkiling en utilisant le logiciel **R**. Pour ces stations, les données ont été partitionnées en données d'entrainement$(1950 - 1993)$ et en données de test$(1994 - 2015)$ sauf pour la station de Bounkiling qui est une station particulière (voir sous-section 3.2.5). La période $1950 - 1993$ a été utilisée comme ligne de base car elle est d'une durée suffisante pour établir une climatologie fiable, mais pas trop longue, ni trop contemporaine pour inclure un signal de changement global fort [53]. La performance des modèles pendant l'entrainement et le test a été évaluée en utilisant l'erreur quadratique moyenne ($RMSE$), le biais (B) et le coefficient de détermination (R^2). L'évaluation se fait par station afin de voir quelle est le modèle le plus performant dans chaque station.

3.2.1 Station de Ziguinchor

Les performances des modèles à la station de Ziguinchor ont été évaluées en commençant par la comparaison des séries temporelles des précipitations. Les précipitations observées et simulées par les modèles ont été converties en séries chronologiques mensuelles et présentées sur la Figure 3.2 à des fins de comparaison sur les deux périodes : période d'entrainement et période de test.

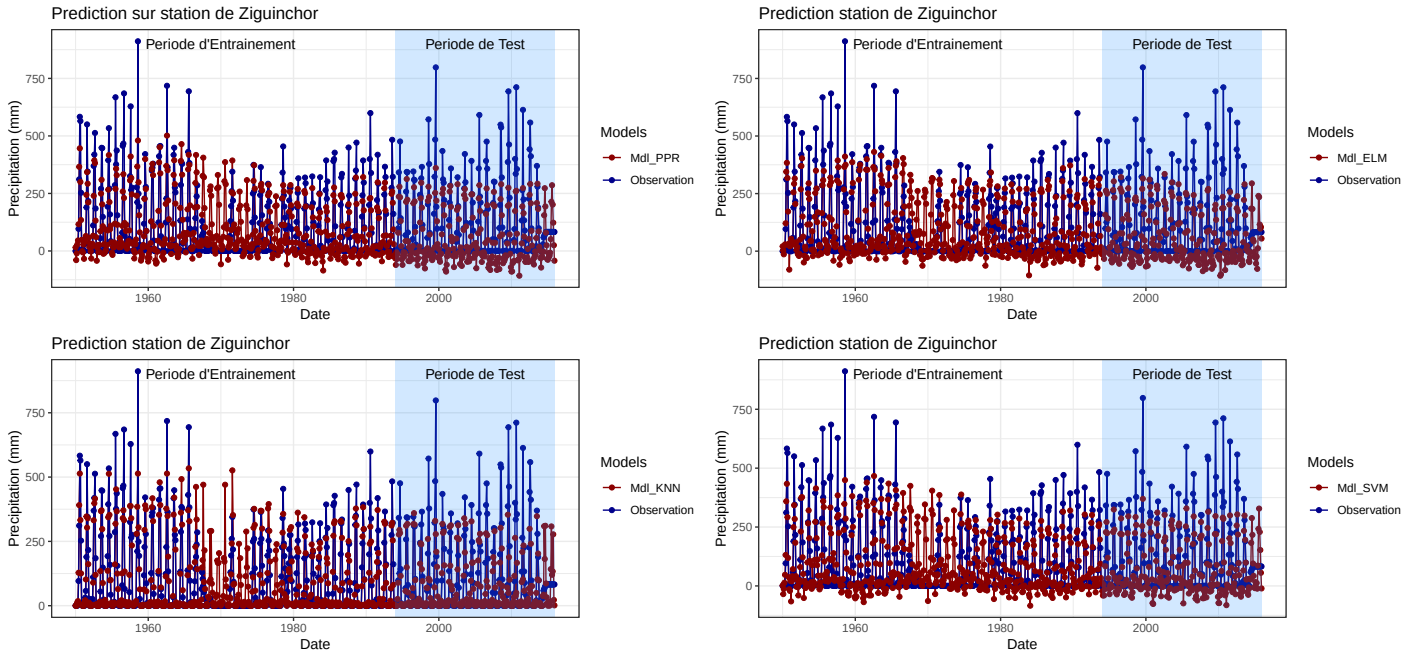


FIGURE 3.2 – Série chronologique des précipitations observées et simulées au niveau de la station de Ziguinchor

Nous constatons que les précipitations mensuelles simulées suivent une évolution similaire à celles des précipitations observées. Par contre, certains mois présentent des valeurs de précipitations très élevées, qui ont été sous-estimées par les modèles. L'apparition d'événements extrêmes est un phénomène courant dans l'hydrologie des précipitations, qui souvent ne peut être prévu par les prédicteurs du NCEP. Tripathi [39] a signalé que les modèles de réduction d'échelles ne parviennent pas à reproduire correctement les précipitations extrêmes. En général, la variance des précipitations observées est beaucoup plus importante que la variance des variables atmosphériques obtenues à partir des réanalyse ou des MCG et donc les modèles de réduction d'échelle ne parviennent pas à saisir toute la gamme de variance des précipitations [54]. Cependant, ils peuvent réussir à capturer la moyenne ou les faibles précipitations. Nous avons observé que les modèles utilisés dans l'étude reproduisent avec précision la moyenne et un peu moins les faibles précipitations. On voit aussi que le modèle KNN s'accorde mieux avec les observations que les autres modèles surtout pour les faibles précipitations. Ces résultats confirme ceux de Ahmed et al., 2015 [55] qui ont montré que les modèles reproduisent mieux la précipitation moyenne que la variance de la précipitation.

Les performances des modèles ont également été évaluées à l'aide d'approches statistiques standard, à savoir le biais (B), l'erreur quadratique moyenne (RMSE) et le coefficient de détermination (R^2) pendant les phases l'étalonnage et la validation des modèles. Les résultats obtenus pour la station de Ziguinchor sont présentés dans le Tableau 3.2 .

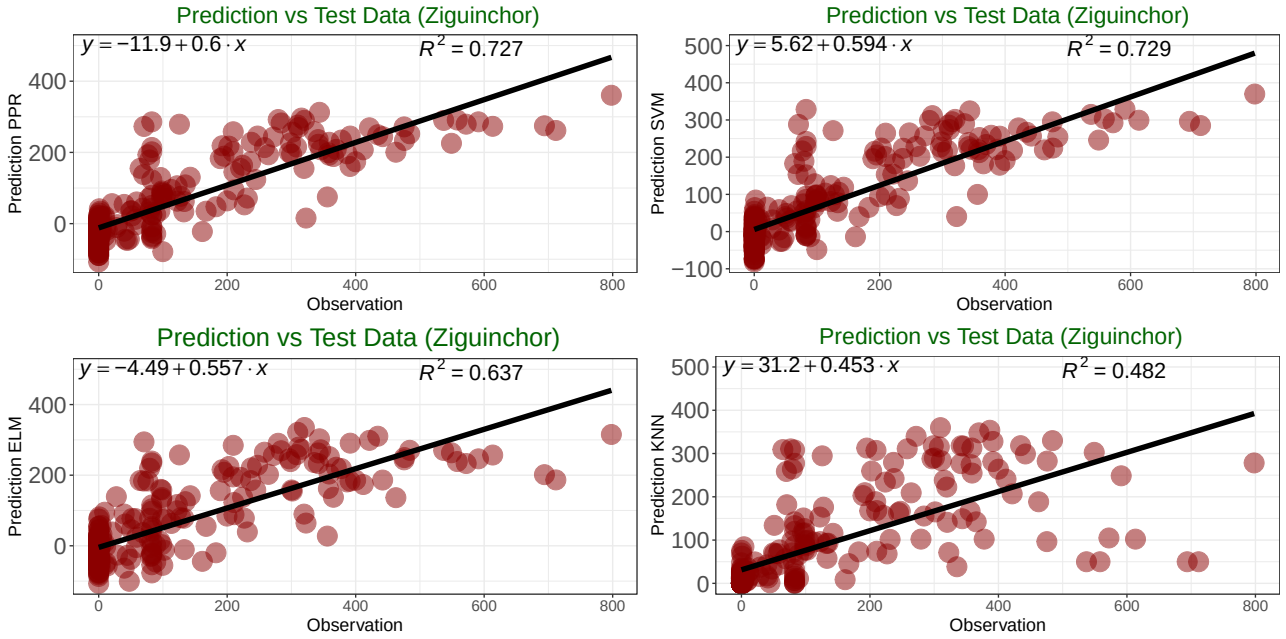
Le RMSE explique la différence entre les précipitations observées et celles qui sont simulées, et fournit donc la répartition de l'erreur ou la performance du modèle. On peut voir dans le Tableau 3.2 que les différents modèles ont des valeurs efficaces différentes pendant l'étalonnage et la validation. En effet pendant l'étalonnage des modèles, les valeurs varient entre $92.208mm$ pour PPR et $79.786mm$ pour KNN alors que pendant la validation, la plupart des modèles franchissent la barre de $100mm$ avec une valeur maximale de $120.201mm$ pour le modèle KNN. On remarque une certaine incohérence sur les valeur de RMSE surtout pour le modèle KNN où on note une grande différence entre l'étalonnage ($79.786mm$) et la validation ($120.208mm$).

TABLE 3.2 – Mesures de performances station de Ziguinchor

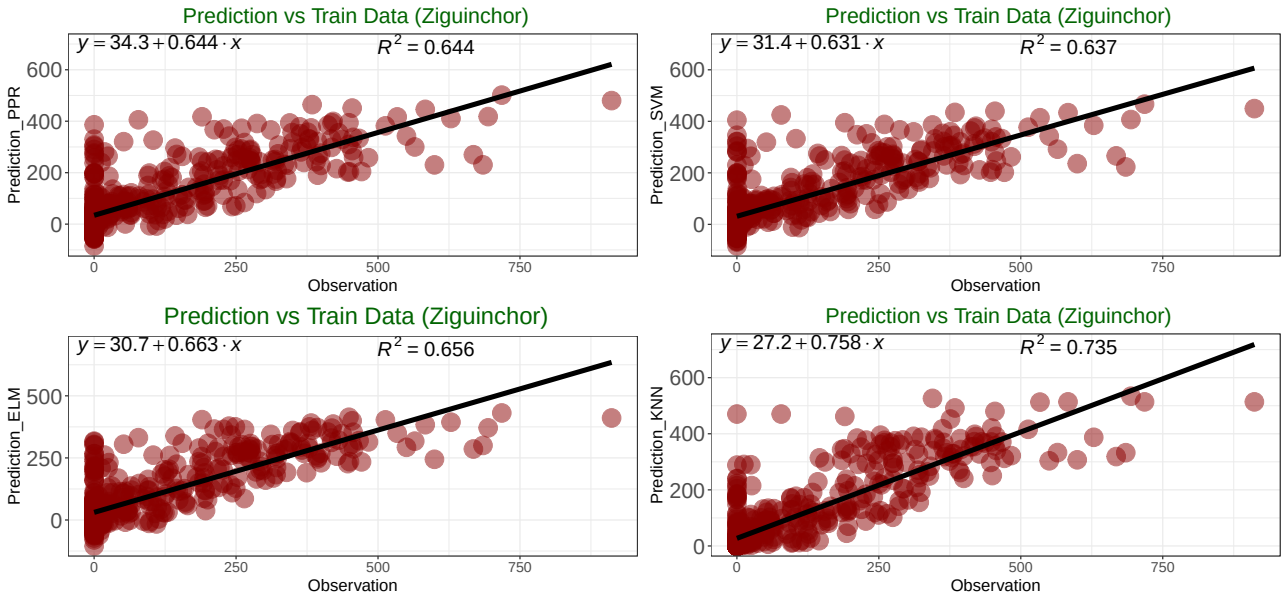
Modèles	Période entrainement			Période test		
	RMSE (mm)	B (mm)	R^2	RMSE (mm)	B (mm)	R^2
PPR	92.208	1.325	0.644	104.179	-55.766	0.727
KNN	79.786	3.884	0.735	120.201	-28.832	0.482
SVM	93.157	-4.112	0.637	96.271	-38.917	0.729
ELM	89.498	0.792	0.656	103.462	-17.029	0.637

L'erreur de biais moyenne (B) mesure la différence moyenne entre deux ensembles de données et, par conséquent, fournit une mesure utile du degré de sous-estimation ou de sur-estimation par les modèles. Le Tableau 3.2 présente l'erreur moyenne de biais dans les précipitations simulées lors de l'étalonnage et de la validation. Il ressort de ce tableau que le biais varie considérablement selon les modèles. Les précipitations se sont avérées légèrement surestimées pour la plupart des modèles lors de l'étalonnage sauf pour le modèle SVM qui quant-a-lui les sous-estimes. Les précipitations sont sous-estimées que lors de la validation des modèles et cela pour tous les modèles. Le modèle PPR présente le biais négatif le plus élevé et donc sous-estime les précipitations. Nous avons observé un comportement différent des quatre modèles dans la reproduction des précipitations pendant la phase d'étalonnage et la validation des modèles. Le signe des biais varie de positif à négatif pendant la calibration et la validation pour la plupart des modèles.

Les diagrammes de dispersions des précipitations observées et simulées par les différents modèles pendant l'étalonnage et la validation sont présentés sur la Figure 3.3 . On peut voir sur cette figure que pratiquement tous les modèles sous-estiment les valeurs maximales des précipitations observées dans la station. Cependant, nous avons constaté que les modèles reproduisaient efficacement les faibles valeurs de précipitations. Les valeurs du coefficient de détermination sont supérieures à 0.6 pour presque tous les modèles pendant l'étalonnage avec une valeur maximale pour le modèle KNN de 0.735. Cependant, la performance des modèles pendant la validation est moins précise que celle de l'étalonnage des modèles surtout pour le modèle KNN où la valeur du coefficient de détermination est en-dessous de 0.5 pendant la validation du modèle. Cela peut être expliqué par le grand nombre de donnés manquantes pendant la période de validation. Des résultats similaires ont été obtenus par Pervez [56] lors de la simulation des précipitations dans le bassin du Gange-Brahmapoutre, où le modèle utilisé s'est avéré plus performant lors de l'étalonnage du modèle dans la plupart des stations.



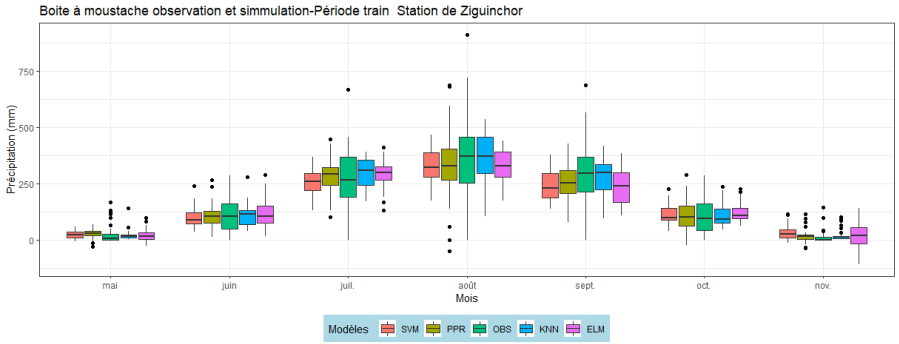
(a) Période de validation



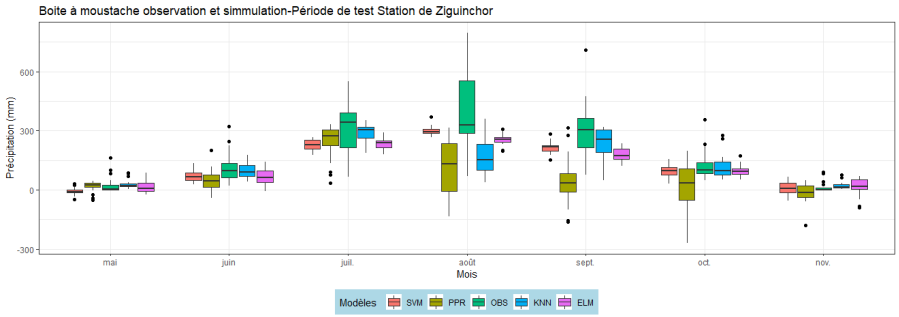
(b) Période d'étalonnage

FIGURE 3.3 – Diagrammes de dispersion des précipitations observées et réduites pendant la période d'étalonnage en (b) et de validation en (a) à la station de Ziguinchor

Il apparait donc claire que ces indices de performance ne sont pas suffisantes pour comparer la pertinence des modèles à la station de Ziguinchor. Donc, nous avons utilisé d'autres outils supplémentaires pour faire des comparaisons supplémentaires. La Figure 3.4 représente les boites à moustaches des précipitations observées et simulées pendant la période d'étalonnage et de validation pour les mois de Mai à Novembre. Cette figure nous montre que les modèles sous-estiment les observations tant pendant la période d'étalonnage que pour la période de validation surtout pour les mois de juillet à septembre avec une forte sous-estimation en Août. Nous observons une fois que les modèles ne parviennent pas à représenter les fortes précipitations. Cependant, nous constatons que le modèle KNN simule mieux les observations que les autres modèles surtout pour les fortes valeurs de précipitations. De plus, on observe moins de valeurs aberrantes pour KNN, cela confirme la performance du modèle au niveau de la station.



(a) Période d'étalonnage



(b) Période de validation

FIGURE 3.4 – Boite à moustache des précipitations mensuelles observées et simulées à la station de Ziguinchor pendant l'étalonnage (a) et la validation (b)

Une autre méthode pour comparer la précision estimée des modèles construits, consiste à un tracé de points (Figure 3.5). Ce sont des graphiques utiles car ils montrent à la fois la précision moyenne estimée et l'intervalle de confiance à 95% (par exemple, l'intervalle dans lequel se situent 95% des données observées). Notons que les points sont classés de la précision moyenne la plus faible à la plus élevée. Il apparait donc que le modèle KNN est plus performant avec un RMsE et un MAE(mean absolute error) plus petits que pour les autres modèles.

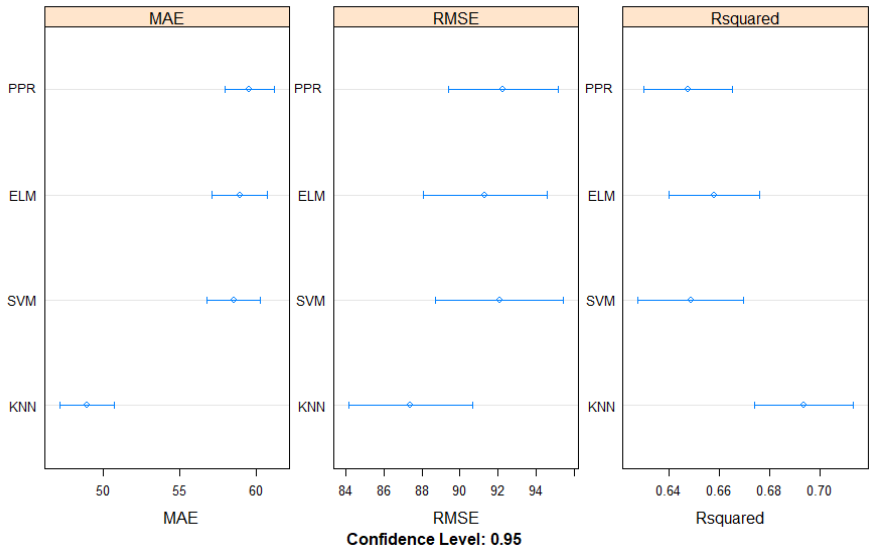


FIGURE 3.5 – Comparaison des algorithmes d'apprentissage automatique dans les tracés de points station à la de Ziguinchor

Un dernier outil de mesure de performance utilisé dans ce travail est le tracé du diagramme de Taylor (Figure 3.6), Le diagramme de Taylor fournit un moyen de résumer graphiquement le degré de correspondance entre un ensemble de modèles et l'observation [57]. La similarité entre deux modèles est quantifiée en termes de corrélation **r**, de différence moyenne quadratique centrée **RMS** et d'amplitude de leurs variations (représentées par leurs écarts types). Ce diagramme est particulièrement utile pour évaluer de multiples aspects de modèles complexes ou pour mesurer la performance relative de nombreux modèles différents. Les modèles simulés qui concordent bien avec les observations se situent le plus près du point marqué "**observed**" sur l'axe des abscisses. Ces modèles auront une corrélation relativement élevée et de faibles erreurs RMS. Sur la Figure 3.6, le modèle KNN se démarque complètement des autres avec une corrélation **r** plus élevée et un **RMS** plus faible.

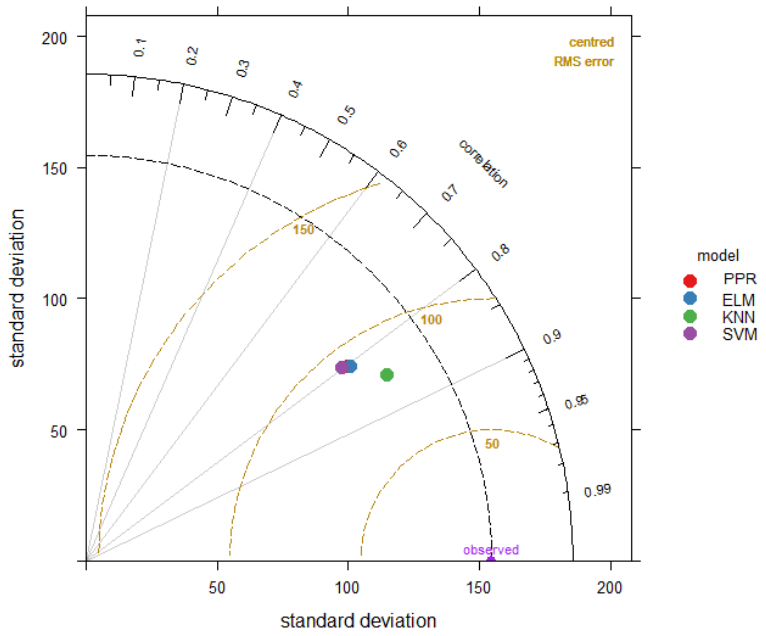


FIGURE 3.6 – Diagramme de Taylor des précipitations mensuelles station de Ziguinchor

En somme, l'analyse de ces derniers résultats nous permet de voir que le modèle KNN est beaucoup plus performant que les autres au niveau de la station de Ziguinchor. L'évaluation au niveau des autres stations se fera en s'appuyant sur l'étude faite à la station de ziguinchor.

3.2.2 Station de Bignona

A l'image de la station de Ziguinchor, l'évaluation des performances des modèles à la station de Bignona commence par la comparaison des séries temporelles. La Figure 3.7 représente la série chronologique des précipitations observées et simulées à la station de Bignona. On peut observer sur cette série un manque de données surtout pendant la période de test. Nous constatons que comme à la station de Ziguinchor, les modèles ont une évolution similaire à celles des observations et parviennent à simuler avec précision les moyennes et un peu moins les faibles précipitations. Cependant, les modèles sous-estiment les précipitations extrêmes tout comme à la station de Ziguinchor. Nous observons aussi que le modèle des KNN simule mieux les faibles précipitations.

Les critères de performance ont étaient évalués et présentés dans Tableau 3.3. On peut voir dans ce tableau que les RSME des modèles sont sous la barre de $90mm$ pendant l'étalonnage avec le

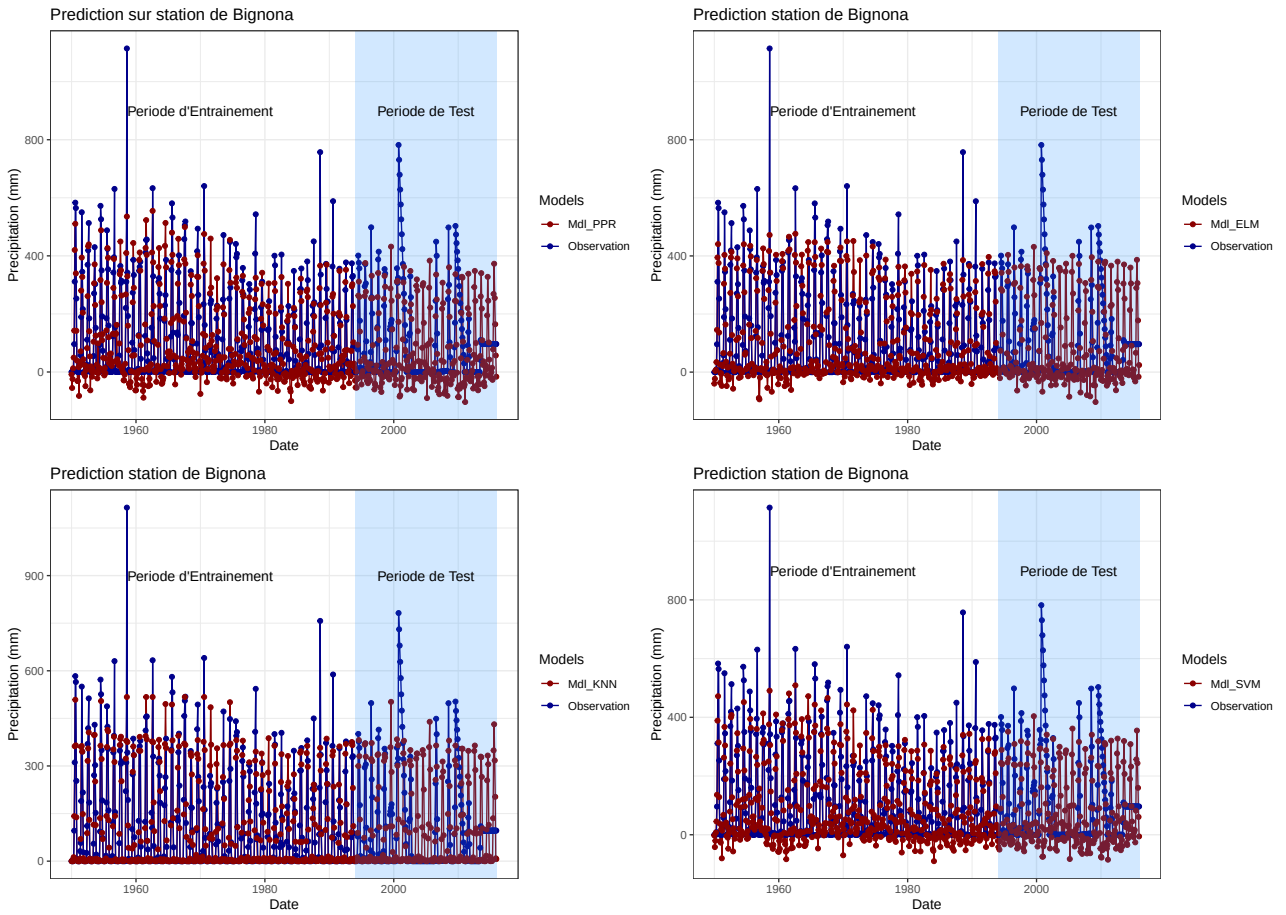


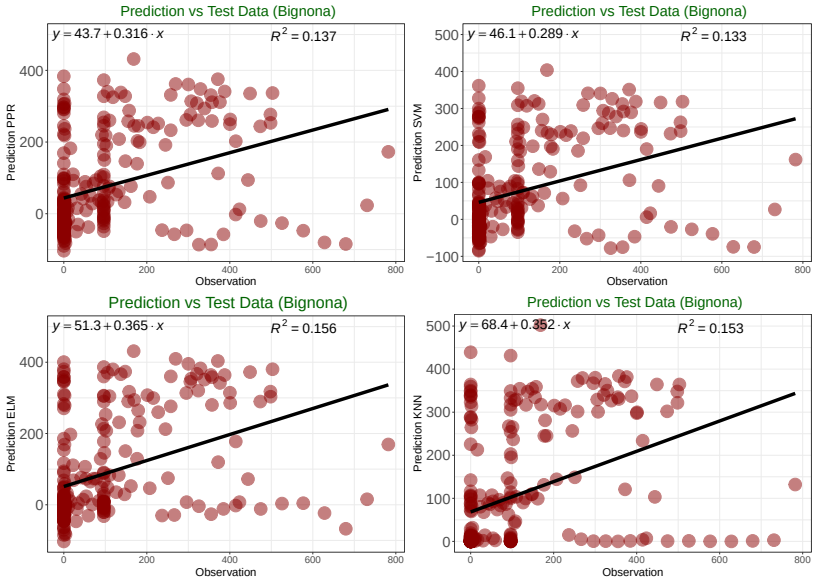
FIGURE 3.7 – Série chronologique des précipitations observées et simulées au niveau de la station de Bignona

modèle PPR qui enregistre la plus faible valeur 65.52mm suivie du modèle KNN. Pendant la période de validation, on constate que cette valeur de RMSE atteint pratiquement les 160mm, ce qui montre une certaine incohérence sur son comportement entre l'étalonnage et la validation. Cette incohérence s'explique bien par le manque important de donnée pendant la période de validation.

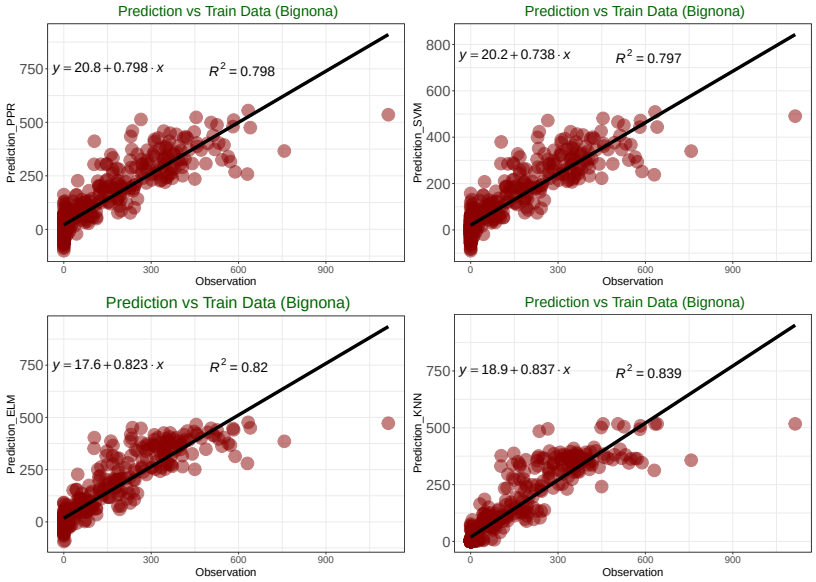
TABLE 3.3 – Mesures de performances station de Bignona

Modèles	Période entrainement			Période test		
	RMSE (mm)	B (mm)	R^2	RMSE (mm)	B (mm)	R^2
PPR	65.520	1.001	0.798	158.151	-7.082	0.137
KNN	79.786	2.607	0.839	161.634	5.434	0.153
SVM	93.157	6.798	0.797	158.925	-22.977	0.133
ELM	89.498	0.882	0.820	161.748	-1.228	0.156

L'analyse du biais montre une légère surestimation des observations par presque tout les modèles avec une valeur maximale du modèle SVM pendant la phase d'étalonnage. Pendant la validation, presque tous les modèles sous-estiment les observations sauf le modèle KNN dont la valeur est positive. Le modèle KNN montre un biais assez proche entre l'étalonnage et la validation.



(a) Période de validation

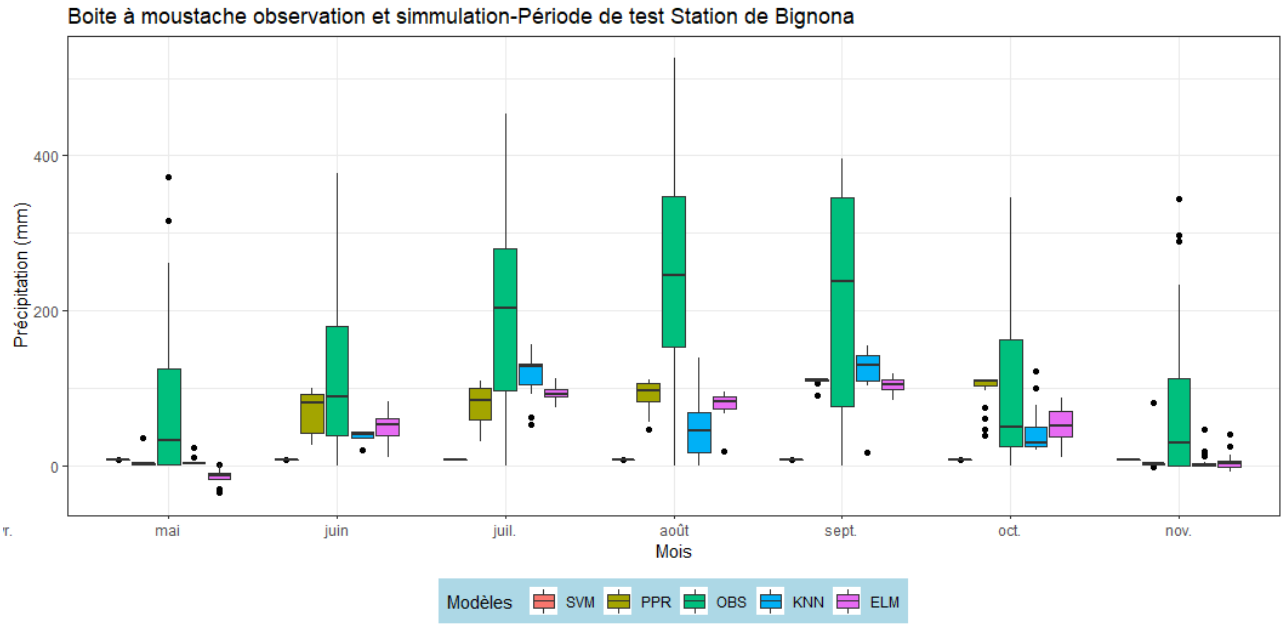


(b) Période d'étalonnage

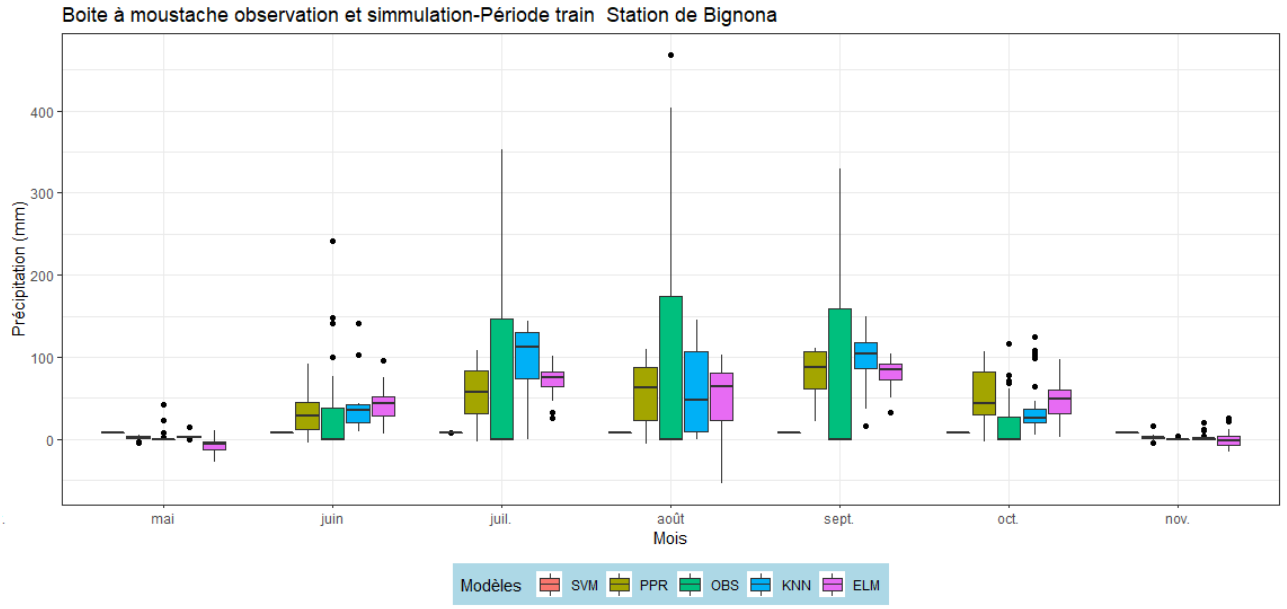
FIGURE 3.8 – Diagrammes de dispersion des précipitations observées et simulées pendant la période d'étalonnage en (b) et de validation en (a) à la station de Bignona

Les diagrammes de dispersions des précipitations observées et simulées à la station de Bignona sont représentés sur la Figure 3.8 . On constate une sous-estimation des précipitations extrêmes par les modèles pendant l'étalonnage et la validation. Nous notons aussi une très grande dispersion des données pendant la phase de validation avec des coefficients de détermination qui sont en-dessous de 0.2 . Cela peut être expliqué par le grand nombre de données manquantes pendant cette période. Cependant, on note une bonne corrélation des modèles pendant la période d'étalonnage avec des coefficients de corrélations qui dépassent 0.79 pour tous les modèles. La valeur maximale a été obtenue par le modèle KNN et est égale à 0.839 suivi du modèle PPR avec une valeur de 0.82.

Contrairement à la station de Ziguinchor, les résultats présentés dans la Figure 3.8 nous permettent de voir que les modèles KNN et PPR simulent mieux les précipitations observées que les autres modèles.



(a) Période de validation



(b) Période d'étalonnage

FIGURE 3.9 – Boite à moustache des précipitations mensuelles observées et réduites à la station de Bignona pendant l'étalonnage (b) et la validation (a)

Ceci peut être vérifié sur la Figure 3.9 qui représente les boites à moustaches des précipitations observées et simulées entre les mois de Mai et Septembre à la station de Bignona. Nous notons sur cette figure une forte sous-estimation des observations par les modèles pendant la validation. Nous remarquons également que les modèles ont du mal à bien simuler les observations surtout pour les mois les plus pluvieuses (Juillet à Septembre), seule les modèle KNN et PPR se distinguent dans cette phase. Cependant, pendant l'étalonnage, on voit que presque tous les modèles parviennent à mieux simuler les précipitations avec KNN et PPR qui présentent une meilleure dispersion des données par rapport à ELM surtout au mois d'Août . Le modèle SVM quant-à lui simulem très faiblement les observations.

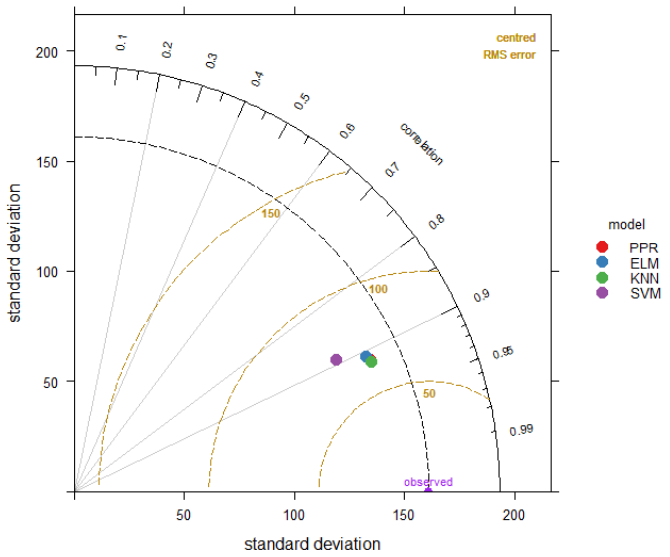


FIGURE 3.10 – Diagramme de Taylor des précipitations mensuelles station de Bignona

Le diagramme de Taylor pour la station de Bignona est présenté à la Figure 3.11. Il sort de cette figure que les modèles KNN , PPR et ELM ont des coefficients de corrélations supérieurs à 0.9 , la valeurs des RMS pour les différents modèles est inférieur à $75mm$ et la valeur de l'écart-type des modèles est inférieur à celle des observations.

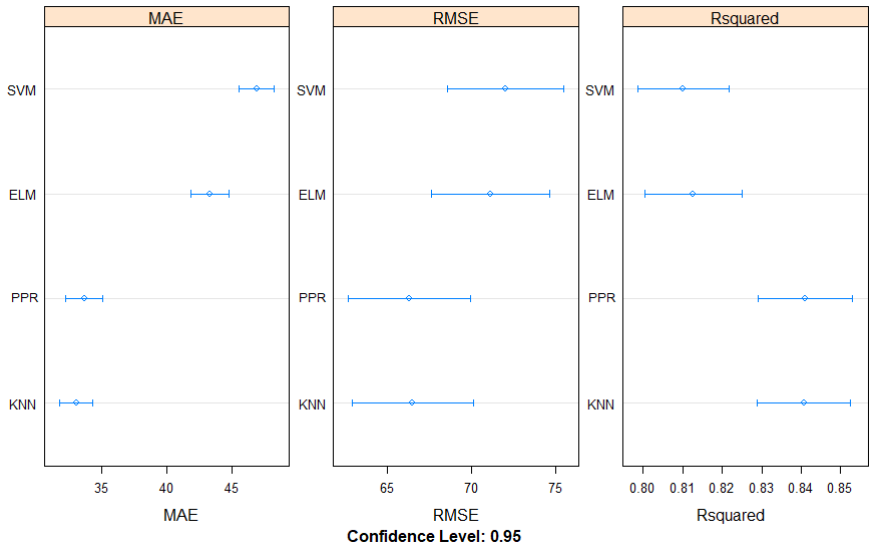


FIGURE 3.11 – Comparaison des algorithmes d'apprentissage automatique dans les tracés de points station de Bignona

Ces résultats nous montrent bien que les modèles KNN, PPR et ELM sont beaucoup plus performants que le SVM au niveaux de cette station.
La Figure 3.11 représente le tracée de points qui permet de comparer la précision estimée des modèles. Nous constatons sur cette figure que deux modèles se distinguent des autres (KNN, PPR). Cependant le modèle KNN reste meilleur. Tout comme la station de Ziguinchor, on peut dire aussi que le modèle KNN simule mieux les précipitations à la station de Bignona.

3.2.3 Station de Sédhiou

La présentation de la série chronologique des précipitations mensuelles observées et simulées est représente sur la Figure 3.12. Comme pour les autres stations, il apparait que les modèles ont une évolution similaire à celles des précipitations observées et on note également une sous-estimation de celles-ci pour les fortes et faibles précipitations. Cependant le modèle KNN arrive à simuler efficacement les faibles et moyennes précipitations.

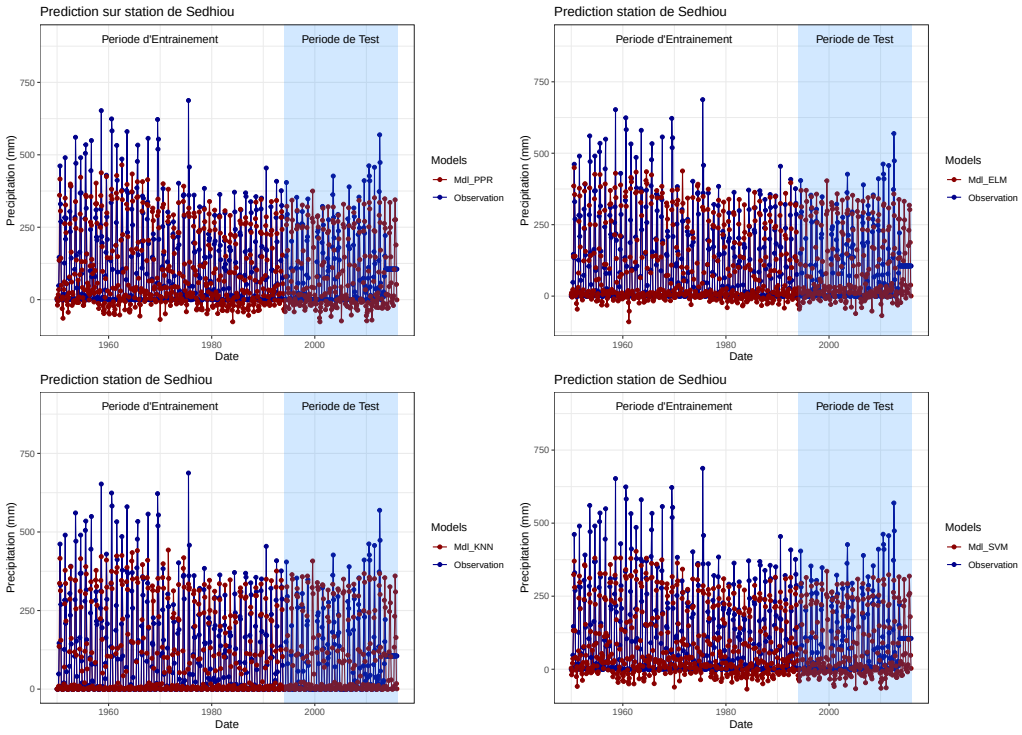


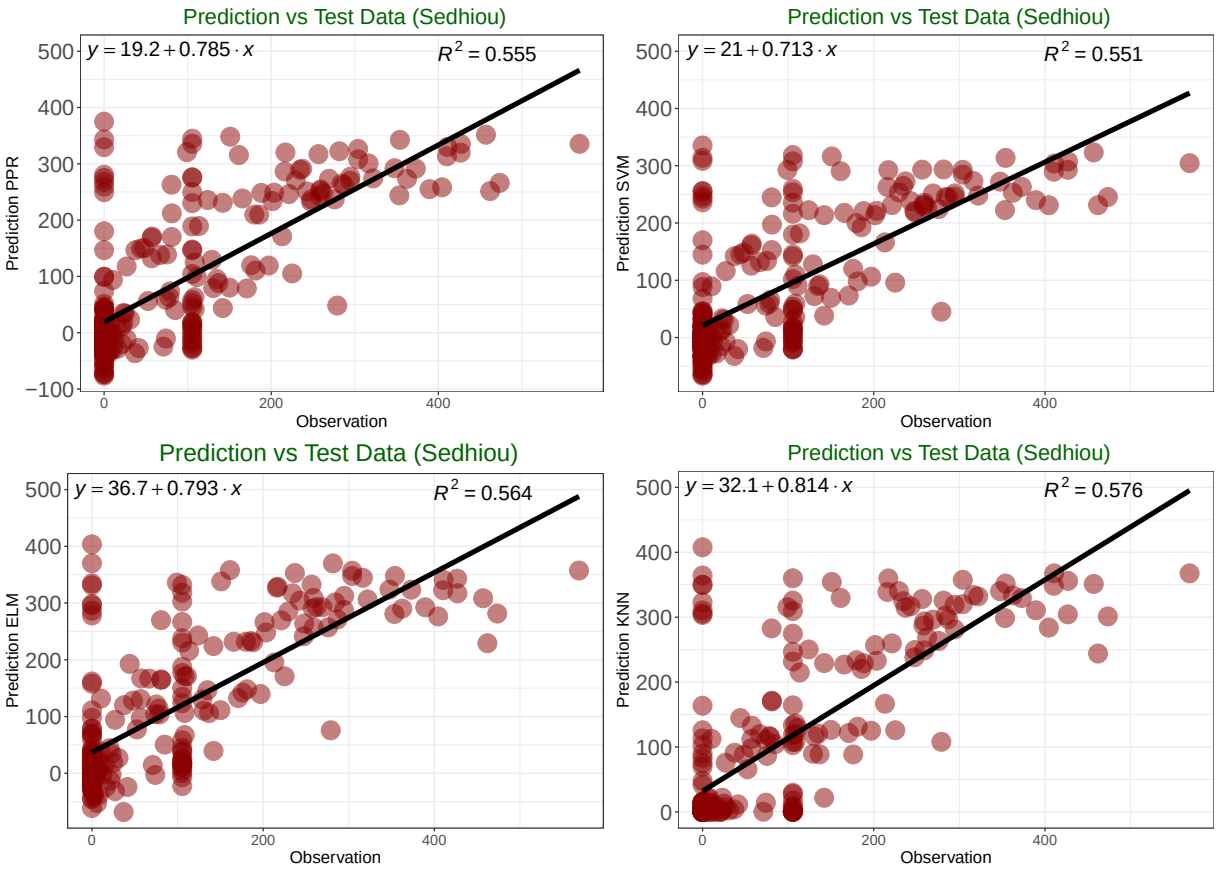
FIGURE 3.12 – Série chronologique des précipitations observées et simulées au niveau de la station de Sédhiou

Le Tableau 3.4 donne l'évaluation des performances des modèles à travers les critères R^2, $RMSE$ et du Biais. Les résultats montrent que les modèles KNN et PPR présentent les meilleurs résultats par rapport aux modèles comparatifs pendant les périodes d'étalonnage et de validation. Selon les valeurs moyennes de R^2, $RMSE$ et du Biais, le modèle KNN a obtenu respectivement $62.489mm$, 0.826 et $0.185mm$ pendant la période d'étalonnage et $84.707mm$, 0.576 et $16.142mm$ pendant la période de validation.

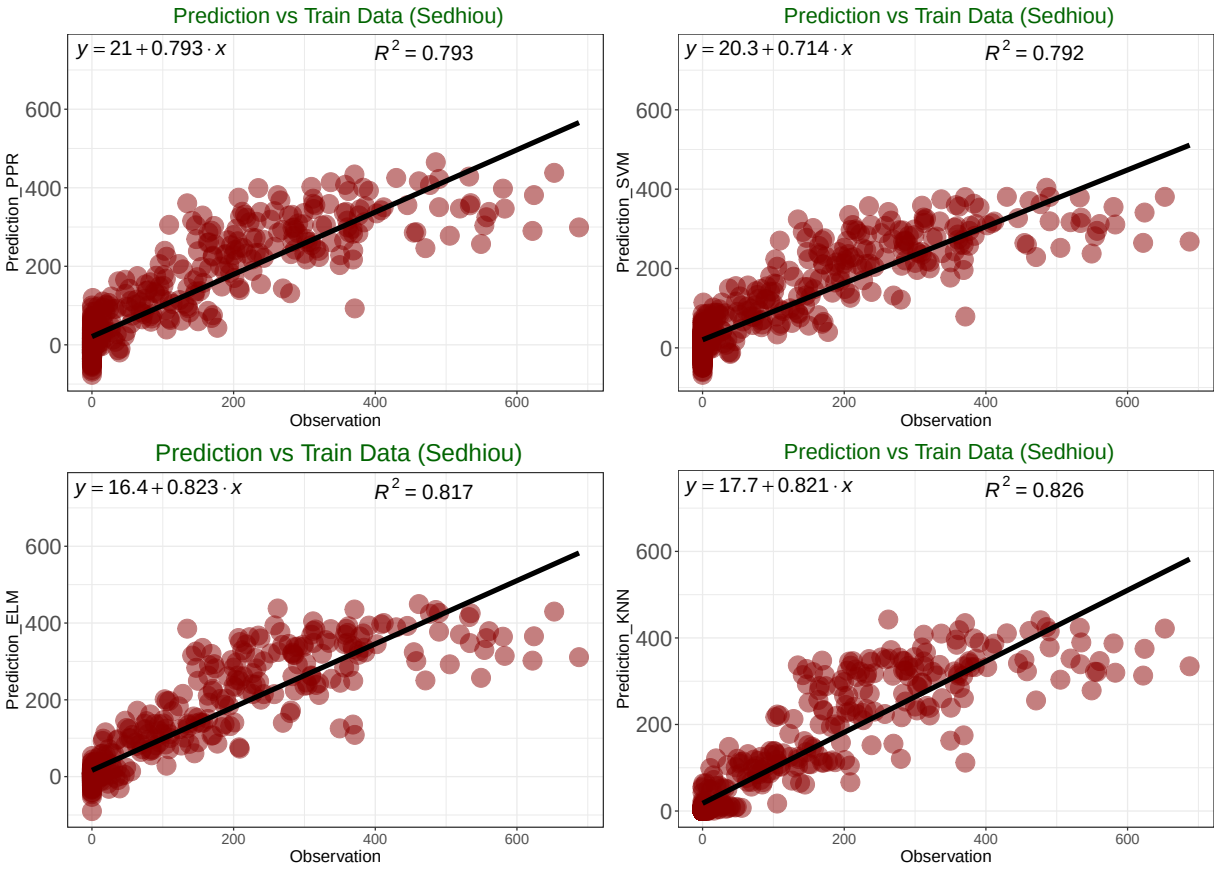
TABLE 3.4 – Mesures de performance station de Sedhiou

Modèles	Période entrainement			Période test		
	RMSE (mm)	B (mm)	R^2	RMSE (mm)	B (mm)	R^2
PPR	62.509	-0.0001	0.793	88.124	18.823	0.555
KNN	62.489	0.185	0.826	84.707	16.142	0.576
SVM	70.301	-8.661	0.792	87.249	-2.8162	0.551
ELM	64.181	-0.074	0.817	93.757	13.145	0.564

Tous ces critères sont supérieurs à ceux des modèles comparatifs sauf pour le biais pendant la validation. On remarque que SVM enregistre le plus faible biais $-2.8162mm$ au moment où KNN à un biais de $-16.142mm$, mais cela n'entache en rien de la pertinence du modèle KNN.



(a) Période de validation



(b) Période d'étalonnage

FIGURE 3.13 – Diagrammes de dispersion des précipitations observées et réduites pendant la période d'étalonnage en (b) et de validation en (a) à la station de Sédhiou

On peut également observer dans les Figure 3.13 et Figure 3.14 que le modèle KNN montre une plus grande précision et une plus grande stabilité que les autres modèles dans la station. Ces résultats nous permettent de dire que le modèle KNN obtient les meilleurs résultats à la station de Sédhiou comme au niveau des autres stations.

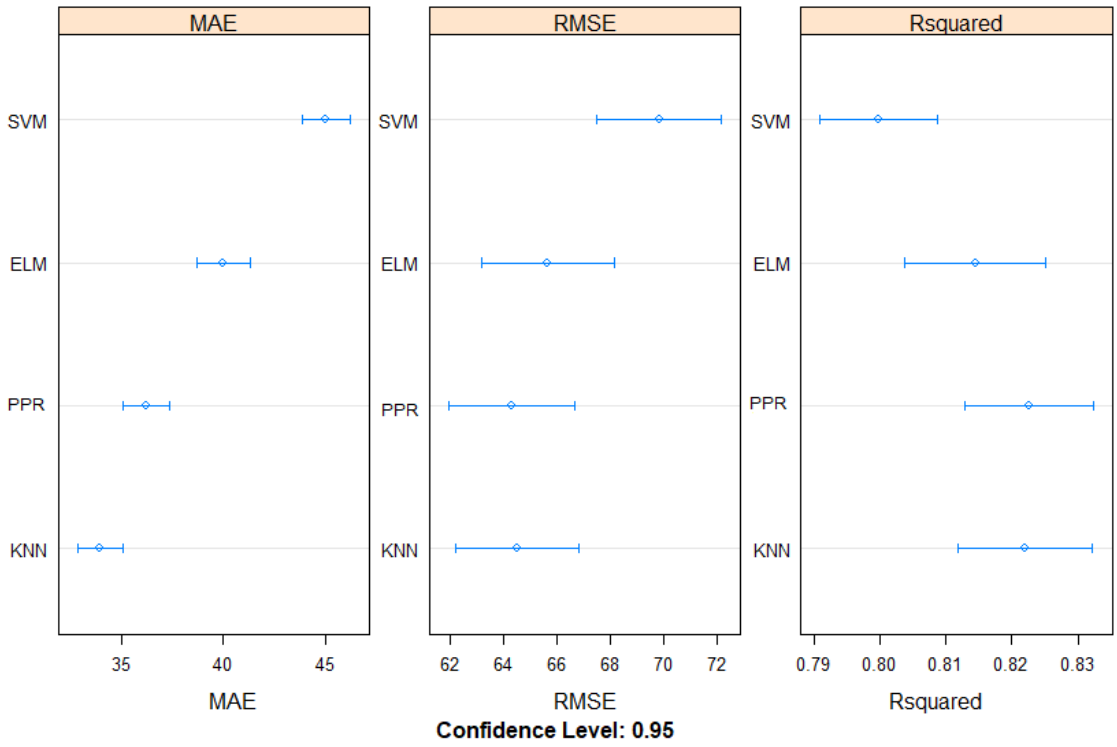


FIGURE 3.14 – Comparaison des algorithmes d'apprentissage automatique dans les tracés de points station de Sédhiou

3.2.4 Station de Kolda

La Figure 3.15 représente la série chronologique des précipitations observées et simulées pour chaque modèle pris individuellement à la station de Kolda. Tout comme dans les stations précédentes, on note que les modèles ont une évolution qui est similaire à celles des précipitations observées. Les modèles sous-estiment les précipitations de point. La figure nous montre aussi que le KNN simule mieux les faibles précipitations que les autres modèles.

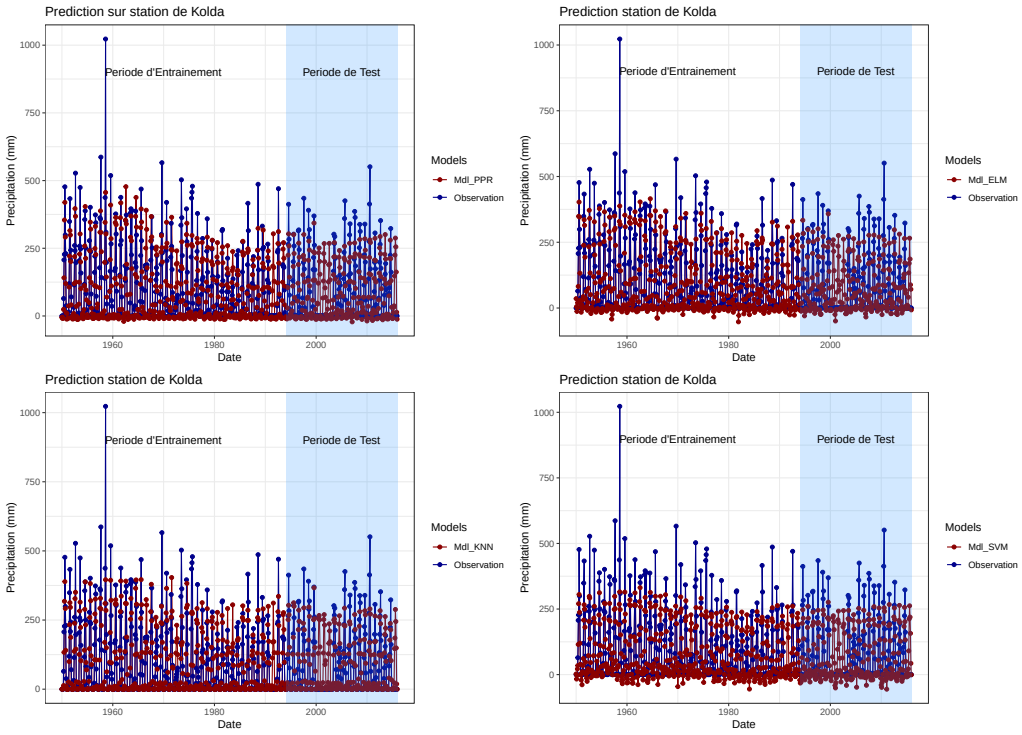


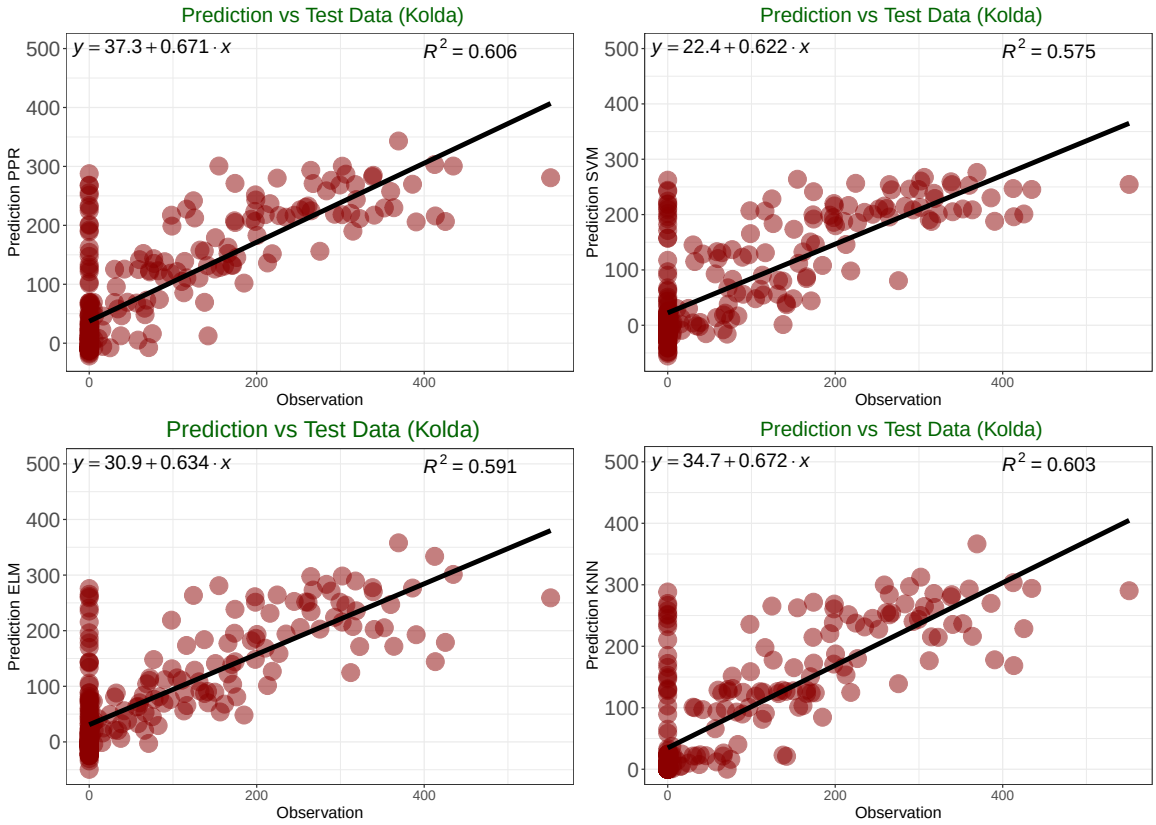
FIGURE 3.15 – Série chronologique des précipitations observées et simulées au niveau de la station de Kolda

L'évaluation des performances à travers les critères R^2, $RMSE$ et du Biais est résumée dans le Tableau 3.5. Il sort de ce tableau de faibles valeurs de RMSE pendant l'étalonnage et la validation pour tous les modèles. La valeur maximale est obtenue par le SVM, $77.279mm$ pendant la validation. En terme de biais, nous notons une légère surestimation des observations par presque tous les modèles pendant l'étalonnage et la validation. Seul le modèle SVM, sous-estime les observations pendant les deux phases. Cependant, on peut voir que KNN , PPR et ELM ont obtenu les meilleurs résultats par rapport au modèle SVM.

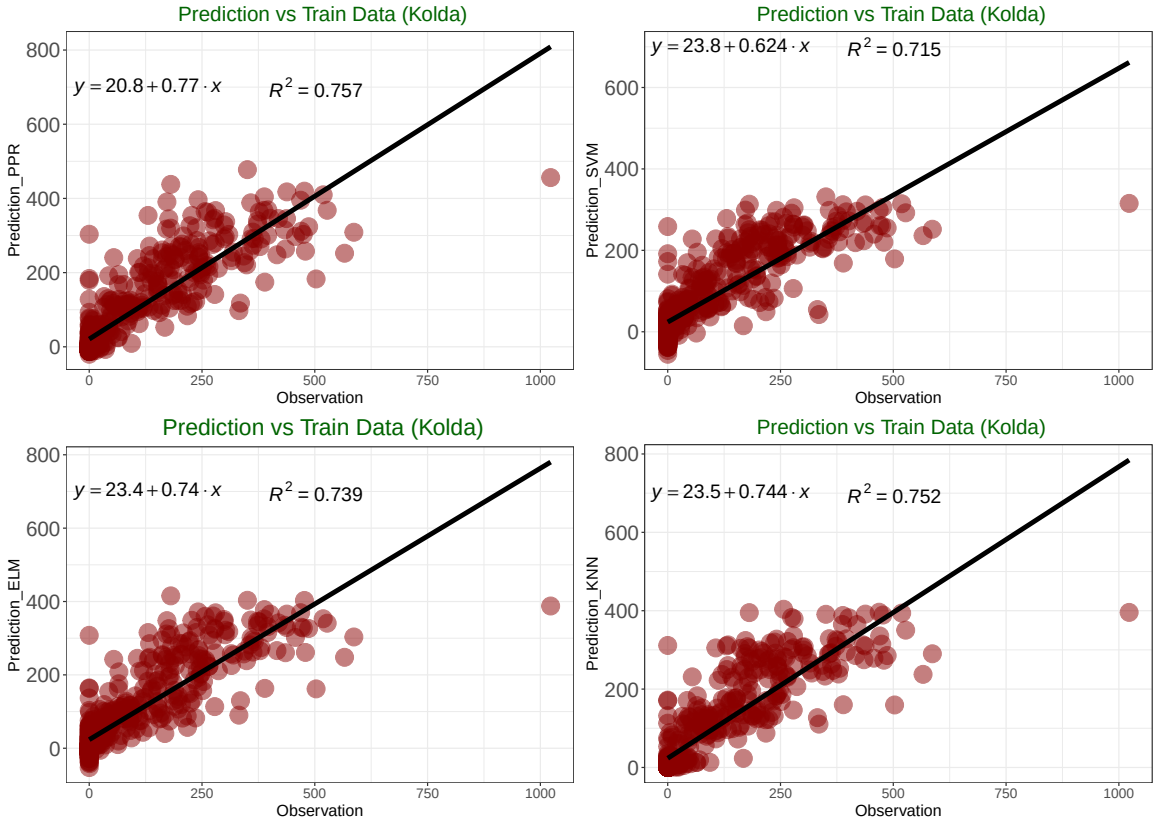
TABLE 3.5 – Mesures de performances station de Kolda

Modèles	Période entrainement			Période test		
	RMSE (mm)	B (mm)	R^2	RMSE (mm)	B (mm)	R^2
PPR	68.810	0.0003	0.757	75.078	6.068	0.606
KNN	67.584	0.752	0.752	75.799	8.726	0.603
SVM	74.601	-10.236	0.715	77.279	-6.350	0.575
ELM	70.411	0.161	0.739	76.933	5.904	0.591

L'analyse des diagrammes de dispersions (Figure 3.16) nous montre une bonne corrélation des modèles par rapport aux observations et aussi une sous-estimation de ceux-ci des fortes précipitations. La valeur du coefficient de détermination pendant l'étalonnage est supérieur à 0.7 et à 0.5 pendant la validation pour toutes les méthodes. Cependant on note que les modèles des KNN et PPR enregistrent les plus fortes valeurs de R^2 .



(a) Période de validation



(b) Période d'étalonnage

FIGURE 3.16 – Diagrammes de dispersion des précipitations observées et simulées pendant la période d'étalonnage en (b) et de validation en (a) à la station de Kolda

Nous remarquons également qu'à la station de Kolda, trois modèles se distinguent à savoir

le KNN, PPR et ELM. Pour plus de précision, le diagramme de Taylor est présenté sur la Figure 3.17, on constate sur cette figure que les trois modèles ont pratiquement les mêmes caractéristiques en terme de coefficient de corrélation, d'écart-type et RMS. Cependant, la Figure 3.18 sur le tracé de points nous permet de trancher sur le meilleur modèle parmi les trois. On peut voir sur cette figure que le modèle KNN est meilleur pour la station de Kolda comme c'est le cas pour les stations précédentes.

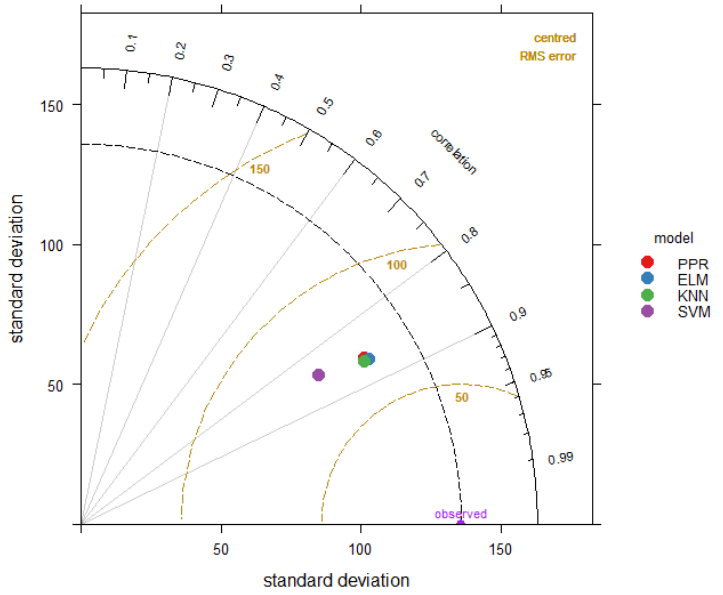


FIGURE 3.17 – Diagramme de Taylor des précipitations mensuelles station de Kolda

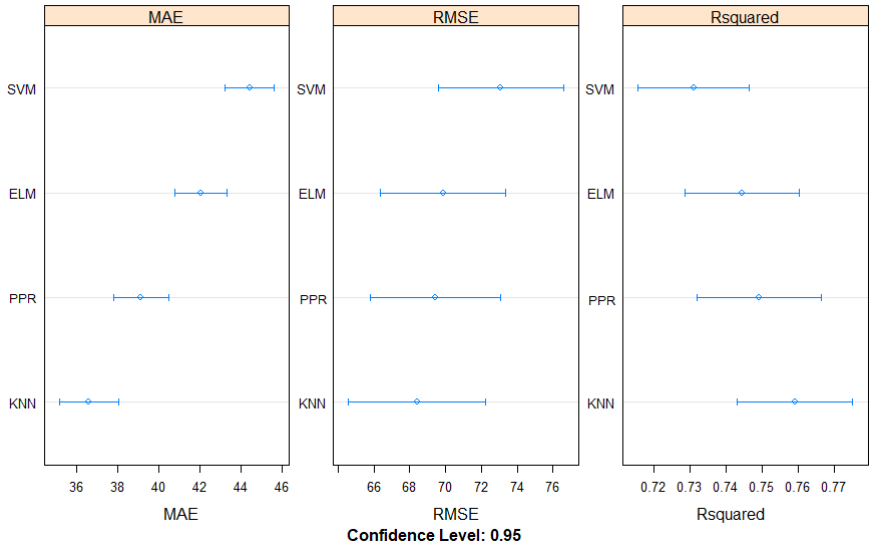


FIGURE 3.18 – Comparaison des algorithmes d'apprentissage automatique dans les tracés de points station de Kolda

3.2.5 Station de Bounkiling

La station de Bounkiling est une station particulière. La collecte des données de précipitations observées dans cette zone a commencé à partir de 1980 . Le jeu de données s'étend donc sur la période de $(1980 - 2015)$ donc sur 36 ans au lieu de 66 ans comme dans les autres stations.

Les périodes d'étalonnage et de validation pour cette station sont respectivement $(1980-2006)$ et $(2007-2015)$ moins que dans les autres stations. Comme dans les autres stations, la série chronologique des précipitations observées est présentée sur la Figure 3.19. Il sort de cette figure le manque de données sur une partie de la période d'entrainement et de validation. Nous notons aussi sur cette figure que les précipitations simulées sont bien corrélées aux observations pour les moyennes et faibles précipitations. les fortes valeurs de précipitation sont sous-estimées par les modèles.

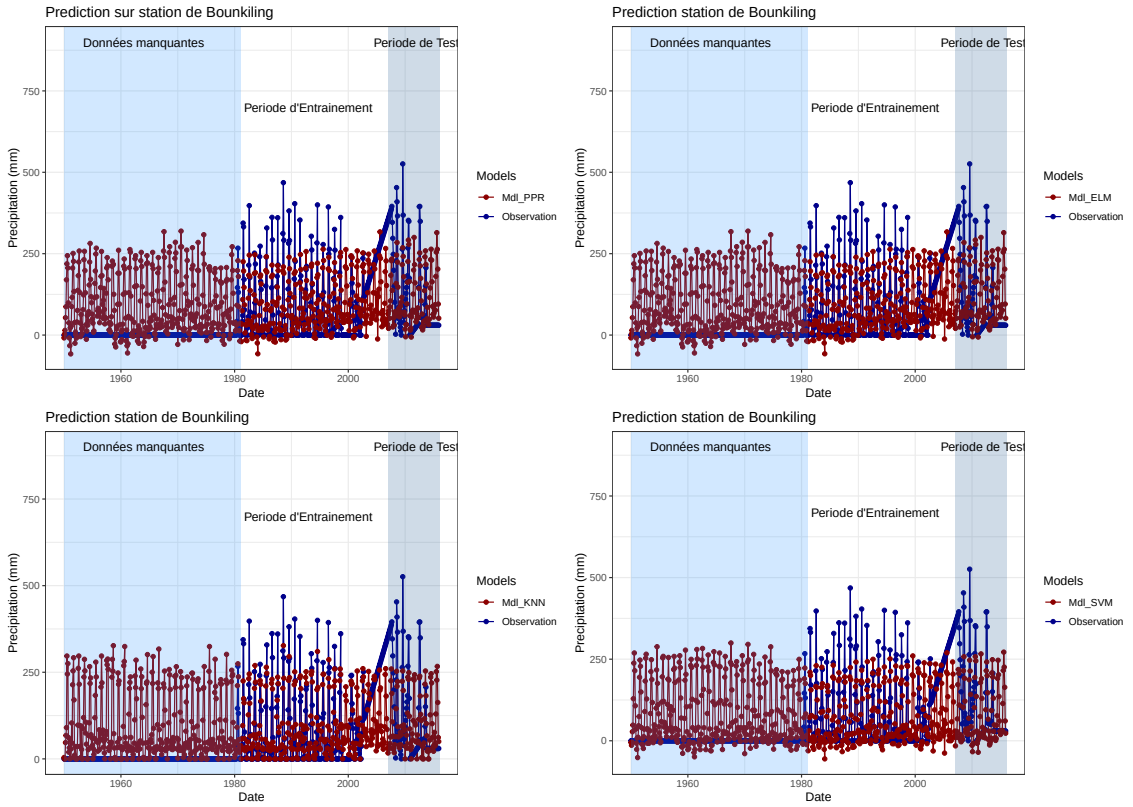


FIGURE 3.19 – Série chronologique des précipitations observées et simulées au niveau de la station de Bounkiling

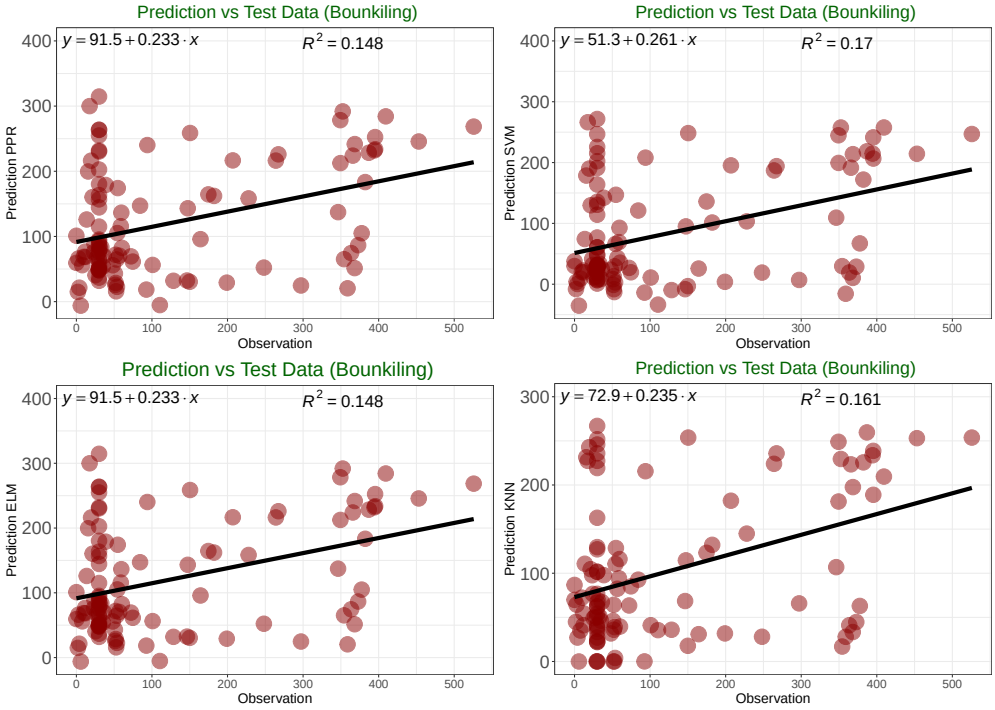
Les performances en terme de R^2, $RMSE$ et du Biais sont résumées dans le Tableau 3.6. On peut voir que ces critères changent fortement de l'étalonnage à la validation. Le RMSE garde une valeur inférieur à $100mm$ pendant l'étalonnage alors que pendant la validation, cette valeur atteint presque $130mm$ pour tous les modèles. On note également un biais qui varie fortement entre les deux périodes, soit il varie de positif à négatif (PPR et ELM), soit on a une différence considérable entre les biais (KNN et SVM).

TABLE 3.6 – Mesures de performance station de Bounkiling

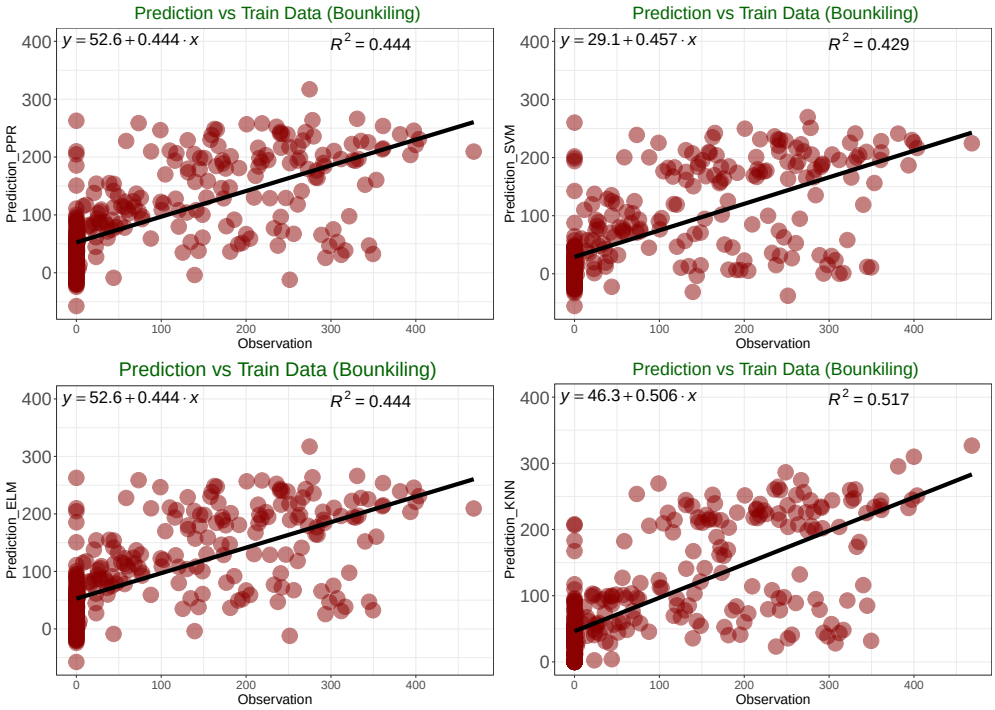
Modèles	Période entrainement			Période test		
	RMSE (mm)	B (mm)	R^2	RMSE (mm)	B (mm)	R^2
PPR	95.625	0.00002	0.444	136.562	-2.748	0.148
KNN	87.956	-0.463	0.517	129.298	-40.330	0.161
SVM	93.416	-22.265	0.429	136.568	-39.450	0.170
ELM	89.414	0.00019	0.444	132.383	-2.749	0.148

Les diagrammes de dispersions des précipitations observées et simulées pendant les phases d'étalonnage et de validation sont présentés sur la Figure 3.20. Il sort de cette figure une grande

dispersion des données pendant la validation qui est illustrée par les faibles valeurs du coefficient de détermination qui ne dépasse pas 0.2 pendant cette phase. Cependant, pendant l'étalonnage, on note une meilleure dispersion des précipitations avec le coefficient de détermination qui dépasse 0.4. La valeur maximale (0.517) est obtenue par le modèle KNN pendant l'étalonnage .



(a) Période de validation



(b) Période d'étalonnage

FIGURE 3.20 – Diagrammes de dispersion des précipitations observées et réduites pendant la période d'étalonnage en (b) et de validation en (a) à la station de Bounkiling

Les informations obtenues des critères de performances ne sont pas suffisantes et ne nous

permettent pas de faire une comparaison sur la performance des modèles. En effet, la station de Bounkiling nous permet de voir quelques limites de nos modèles. Les modèles statistiques ont besoin de suffisamment de données pendant l'entrainement des modèles pour que ceux-ci puissent donner de très bonnes estimations de la réponse. Les données au niveau de la station de Bounkiling ne sont pas suffisantes pour avoir une bonne estimation de la réponse.

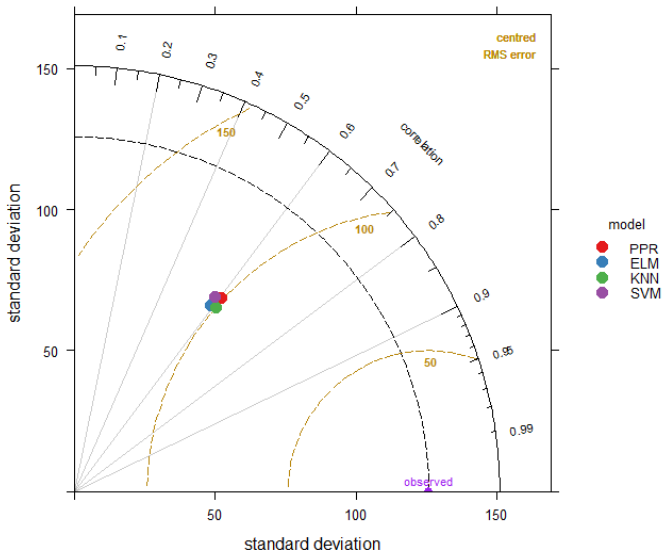
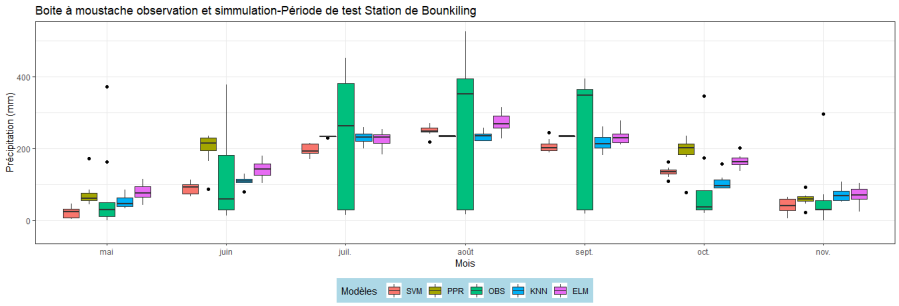
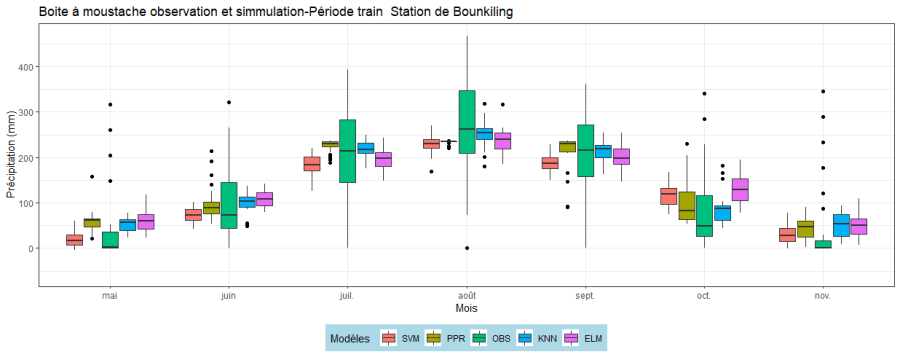


FIGURE 3.21 – Diagramme de Taylor des précipitations mensuelles station de Bounkiling



(a) Période de validation



(b) Période d'étalonnage

FIGURE 3.22 – Boite à moustaches des précipitations mensuelles observées et simulées à la station de Bounkiling pendant l'étalonnage (b) et la validation (a)

Les Figure 3.21 et Figure 3.22 représentent respectivement le diagramme de Taylor et les boites à moustaches pendant l'étalonnage et la validation des précipitations observées et simulées entre

les mois de Mai à Novembre. Sur le diagramme, on remarque que les modèles ont pratiquement des critères similaires avec un RMS qui tourne au tour de $100mm$ et un coefficient de corrélation qui tourne autour de 0.6. Les boites à moustaches nous montrent une forte sous-estimation des précipitations par les modèles pendant l'étalonnage et la validation et aussi pendant les mois les plus pluvieux. Cependant, les résultats présentés à la Figure 3.23 nous permettent de voir que le modèle SVM est beaucoup plus stable que les autres modèles à la station de Bounkiling. Des résultats ont était obtenus par Devak et Dhanya [58] qui ont montré que le modèle SVM était plus performant quant on lui présente moins de données d'entrainement.

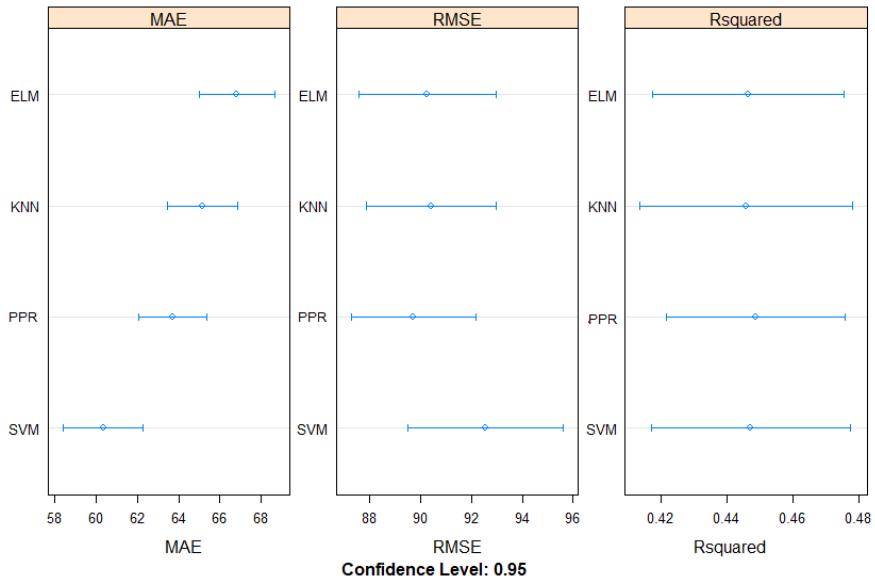


FIGURE 3.23 – Comparaison des algorithmes d'apprentissage automatique dans les tracés de points station de Bounkiling

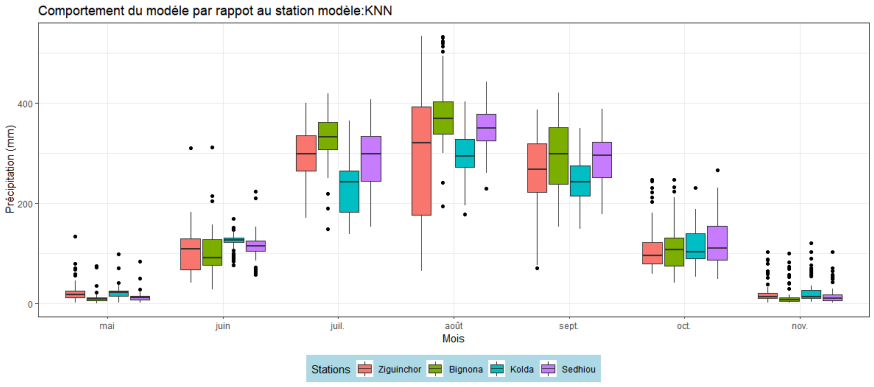
D'une manière globale, on peut dire dans cette étude que les modèles KNN et PPR sont beaucoup plus performants dans notre zone d'étude. Cependant, quel est l'impact réel de chacun au niveau de chaque station ?

3.3 Comportement de chaque modèle dans les différentes stations de la zone d'étude

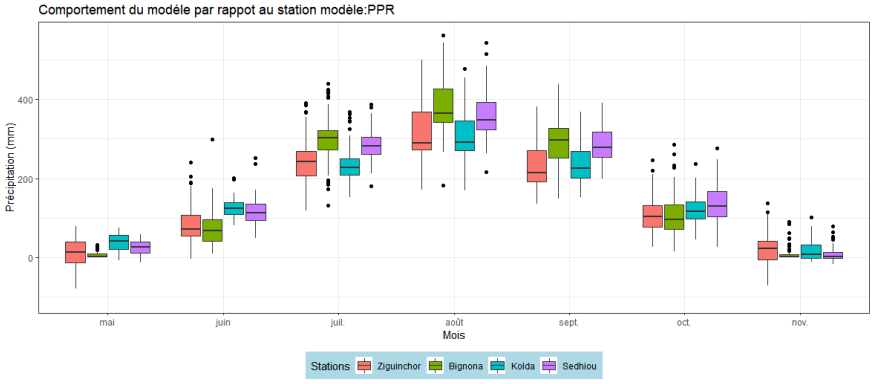
L'étude de chaque modèle au niveau des différentes stations est importante car elle nous permet de voir le comportement du même modèle au niveau de chaque station. L'étude a été faite sur quatre stations à savoir Ziguinchor, Bignona, Sédhiou et Kolda. La station de Bounkiling n'en fait pas parti car n'ayant pas la même période d'observation que les autres stations. La Figure 3.24 représente les boites à moustaches des précipitations simulées entre Mai et Septembre de chaque modèle au niveau des différentes stations. Il sorte de cette figure que pratiquement tous les modèles simulent assez bien les précipitations sur toutes les stations de la zone surtout pour les stations de Zguinchor et Bignona où on note une plus importante distribution des données simulées surtout pendant les mois de Juillet à Septembre . Cependant, on peut voir que deux modèles sont plus stables que les autres dans la zone, il s'agit du KNN et PPR.

Dans l'ensemble, on peut conclure que le modèle KNN est le meilleur car il a fourni les meilleurs résultats pour la prévision des précipitations mensuelles dans la zone d'étude. Il est suivi par

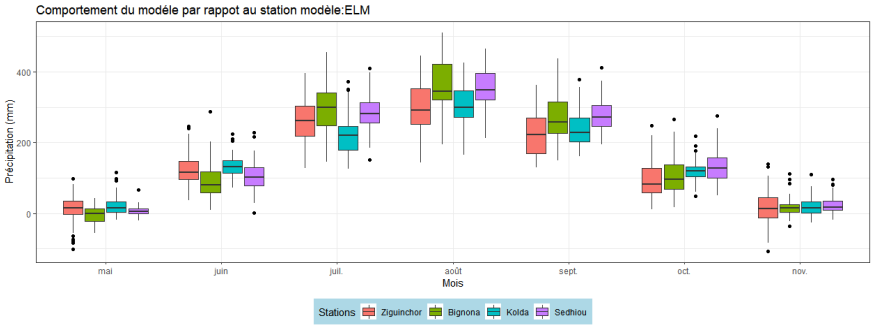
le modèle PPR.



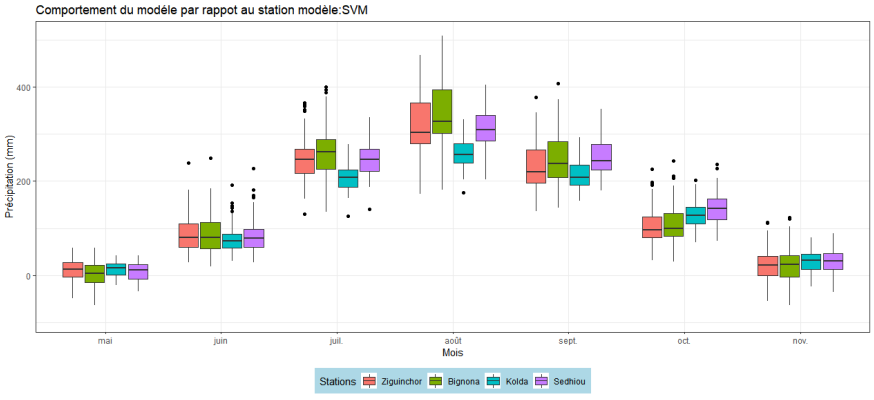
(a) Modèle KNN



(b) Modèle PPR



(c) Modèle ELM



(d) Modèle Modèle SVM

FIGURE 3.24 – Boite à moustache des précipitations réduites par chaque modèles au niveau des stations de Ziguinchor, Bignonan Kolda et Sédhiou.

Sur la base d'études antérieurs, Devak et Dhanya [58] ont appliqué les modèles KNN et SVM pour modéliser les précipitations sur la base des paramètres climatiques à l'échelle grossière. Les résultats de cette étude ont montré une meilleur performance du modèle KNN par rapport au SVM dans la réduction d'échelle des précipitations mensuelles. Dans une autre étude, Giogri et al., 2001 [48] a appliqué la régression polynomiale locale, la régression linéaire multiple et le réseau neuronal artificiel pour prédire les précipitations dans le bassin versant du réservoir d'Idukky au Kerala, en Inde. Comme le montre l'étude, la régression polynomiale locale a offert une meilleure performance dans la prévision des précipitations dans ce bassin.

Conclusion et Perspectives

L'objectif de notre étude portait sur la modélisation des précipitations en Casamance en utilisant des méthodes de "machine learning". Il s'agit pour nous d'étudier la performance de quatre modèles (KNN, PPR, SVM, et ELM) pour l'estimation de la précipitation en Casamance à partir d'un ensemble de prédicteurs.

Les résultats obtenus dans cette étude sont présentés comme suit :

D'abord, les variables climatiques de grande échelle les plus explicatives pour décrire l'évolution mensuelle des précipitations aux stations météorologiques de la Casamance sont sélectionnées à partir des données de ré-analyses NCEP/NCAR. La sélection de ces variables appelées prédicteurs est basée sur l'analyse des valeurs du coefficient de corrélation de Pearson entre les précipitations obtenues sur les stations météorologique de l'ANACIM et les variables météorologiques de NCEP/NCAR.

Plusieurs outils statistiques ont été utilisés pour déterminer la performance des modèles dans la région étudiée. Il s'agit des mesures statistiques tels que le bais, RMSE et R^2, de diagrammes en boite et de dispersions. Le diagramme de Taylor à été aussi utilisé pour examiner quantitativement comment les simulations des modèles et les données d'observation sont carrelées entre elles.

La comparaison des résultats a montré que le modèle KNN a obtenu de meilleurs résultats comparé aux autres modèles. Toutefois, nous avons constaté que les modèles avaient une évolution similaire à celle des précipitations observées. Cependant, ils étaient moins habiles à reproduire les précipitations mensuelles extrêmes. Seul le modèle des KNN est parvenu à reproduire efficacement les faibles précipitations dans la zone. Par conséquent, l'applicabilité du modèle dépend parfois aussi de l'emplacement du site. Ces modèles peuvent également être utilisés pour simuler d'autres paramètres tels que la température maximale, l'humidité spécifique, la température minimale, etc., ce qui aide à évaluer les changements climatiques en fonction du temps et du lieu. Le modèle des KNN a obtenu les meilleur résultats dans l'ensemble de notre zone d'étude suivi du modèle PPR. Cependant au niveau de la station de Bounkiling c'est SVM qui présente les meilleurs résultats.

En outre, les périodes d'entrainement et de validation ont été formées en utilisant les premiers 2/3 et les 1/3 restant des données, respectivement. Cette approche de calibrage et de validation des modèles est souvent insuffisante pour représenter les grandes variations des précipitations dans ces régions. La sélection aléatoire des données peut être utilisée pour améliorer l'efficacité des modèles de réduction d'échelle. De plus, les paramètres de surface, comme l'altitude, la pente, la végétation, etc., influencent également la distribution de ces éléments météorologiques, en particulier en terrain complexe [59], car la faible résolution des MCG empêche l'étude efficace du changement climatique à cette échelle. Il serait donc intéressant de poursuivre les recherches en considérant ces paramètres comme faisant parti des données d'entrées du modèle.

Cette étude se concentre sur les stations locales actuelles, il serait particulièrement intéressant d'étendre les méthodes utilisées aux zones sans stations en utilisant des données satellitaires à haute résolution comme CHIRPS. Il est certain que des données à plus haute résolution temporelle, par exemple les données quotidiennes, seraient d'un grand intérêt pour des résultats plus approfondis. En outre, l'application sur d'autres ensembles de données de réanalyse basés sur différentes représentations de la surface terrestre pourrait également être utile pour valider les algorithmes.
Dans l'ensemble, cette recherche a proposé l'application de méthodologies informatiques douces qui sont constructives dans l'évaluation des impacts du changement climatique à l'échelle de la zone. Ainsi, les approches proposées peuvent être utilisées pour générer des paramètres d'entrée plus précis qui sont essentiels dans la gestion des ressources en eau et la planification pour résoudre les problèmes connexes .

Bibliographie

[1] Peter H Gleick. Methods for evaluating the regional hydrologic impacts of global climatic changes. *Journal of hydrology*, 88(1-2) :97–116, 1986.

[2] Mike Hulme. Recent climatic change in the world's drylands. *Geophysical Research Letters*, 23(1) :61–64, 1996.

[3] Romain Roehrig, Dominique Bouniol, Francoise Guichard, Fré déric Hourdin, and Jean Luc Redelsperger. The present and future of the west african monsoon : A process-oriented assessment of CMIP5 simulations along the AMMA transect. *Journal of Climate*, 26(17) :6471–6505, 2013.

[4] Seidou Sarr MA, Tramblay O, and Yves El Adlouni. Comparison of downscaling methods for mean and extreme precipitation in Senegal. *Journal of Hydrology : Regional Studies*, 4 :369–385, 2015.

[5] Lu Gao, Karsten Schulz, and Matthias Bernhardt. Statistical downscaling of ERA-interim forecast precipitation data in complex terrain using lasso algorithm. *Advances in Meteorology*, 2014 :16–21, 2014.

[6] Thomas Lafon, Simon Dadson, Gwen Buys, and Christel Prudhomme. Bias correction of daily precipitation simulated by a regional climate model : A comparison of methods. *International Journal of Climatology*, 33(6) :1367–1381, 2013.

[7] Claudia Fowler, Hayley J and Blenkinsop, Stephen and Tebaldi. Linking climate change modelling to impacts studies : recent advances in downscaling techniques for hydrological modelling. *International Journal of Climatology : A Journal of the Royal Meteorological Society*, 27(12) :1547–1578, 2007.

[8] L. Ruby Leung, Linda O. Mearns, Filippo Giorgi, and Robert L. Wilby. Regional climate research. *Bulletin of the American Meteorological Society*, 84(1) :89–95, 2003.

[9] Sahar Hadi, Shamsuddin Shahid, and Eun Sung Chung. A Hybrid Model for Statistical Downscaling of Daily Rainfall. In *Procedia Engineering*, volume 154, pages 1424–1430. The Author(s), 2016.

[10] Ouedraogo. Contribution à l ' étude de l'impact de la variabilité climatique sur les ressources en eau en Afrique de l'ouest. Analyse des conséquences d' une sécheresse persistante : normes hydrologiques et modélisation régionale . *Thèse de doctorat à l'université de Montpellier II de France*, pages 19–59, 2001.

[11] George Hadley. VI. Concerning the cause of the general trade-winds. *Philosophical Transactions of the Royal Society of London*, 39(437) :58–62, 1735.

[12] Justine Ringard. Etude rétrospective et prospective des vagues de chaleur en Afrique de l ' Ouest. (February), 2013.

[13] Adama Traore. Evolution des Précipitations au Sahel sur la période 1950-2000 : Extraction de Scénarios intra-saison. Technical report, 2014.

[14] Serge Janicot, Jean-Luc Redelsperger, and Thierry Lebel. La mousson ouest-africaine : introduction à quelques contributions du programme d'étude multidisciplinaire AMMA. *La Météorologie*, 8(Special-AMMA) :2, 2012.

[15] Théo Vischel, Geremy Panthou, Gillaume Quantin, Aurélien Rossi, and Maxime Martinet. L e retour d ' une période humide au Sahel ? Observations et perspectives. pages 43–61, 2015.

[16] Sira Diouf. Évolution spatio-temporelle des vagues de chaleur en Afrique de l'Ouest et risques sanitaires associés. Technical report, Université Assane Seck de Ziguinchor, 2018.

[17] Vincent Fontaine, Bernard and Janicot, Serge and Moron. Rainfall anomaly patterns and wind field signals over West Africa in August (1958–1989). *Journal of Climate*, 8(6) :1503–1510, 1995.

[18] Samuel Louvet. Modulations intrasaisonnières de la Mousson d'Afrique de l'Ouest et impacts sur les vecteurs du paludisme à NDIOP (Sénégal) : Diagnostique et prévisibilité. *Climatologie, Université de Bourgone*, page 246, 2008.

[19] Jea Roca, Rèmy and Lafore, Jean-Philippe and Piriou, Catherine and Redelsperger. Extratropical dry-air intrusions into the West African monsoon midtroposphere : An important factor for the convective activity over the Sahel. *Journal of the Atmospheric Sciences*, 62(2) :390–407, 2005.

[20] Dr. Jean-Marie Barrat. Changements Climatiques en Afrique de l'Ouest et Conséquences sur les Eaux Souterraines. Technical Report 13, 2012.

[21] Nico Rozemeijer. External Review of IUCN 2007 Report on Linking Conservation to Livelihoods in Africa (Objective 2). *Review Literature And Arts Of The Americas*, 1(March), 2008.

[22] Laurent Mermet. Quelle unité territoriale pour la gestion durable de la ressource en eau ? *Annales des mines*, pages 67–80, 2001.

[23] Daouda Zan Diarra. Impacts des changements climatiques en Afrique de l'Ouest. page 35, 2005.

[24] Aondover Tarhule and Peter J. Lamb. Climate research and seasonal forecasting for West Africans. *Bulletin of the American Meteorological Society*, 84(12) :1741–1759, 2003.

[25] Yves Richard, Nicolas Fauchereau, Isabelle Poccard, Mathieu Rouault, and Sylwia Trzaska. 20th century droughts in Southern Africa : Spatial and temporal variability, teleconnections with oceanic and atmospheric conditions. *International Journal of Climatology*, 21(7) :873–885, 2001.

[26] CÉDRIC BEAULAC. Intelligence Artificielle Avec Apprentissage Automatique Pour L ' Estimation De La Position D ' Un Agent Mobile En Utilisant Les Modèles De Markov Cachés Par Cédric Beaulac Novembre 2015. page 116, 2015.

[27] Fabien Benureau. *Self Exploration of Sensorimotor Spaces in Robots* . PhD thesis, 2016.

[28] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, 44(1-2) :207–219, 2000.

[29] Thomas Vandal, Evan Kodra, and Auroop R. Ganguly. Intercomparison of machine learning methods for statistical downscaling : the case of daily and extreme precipitation. *Theoretical and Applied Climatology*, 137(1-2) :557–570, 2019.

[30] D. A. Sachindra, K. Ahmed, Md Mamunur Rashid, S. Shahid, and B. J.C. Perera. Statistical downscaling of precipitation using machine learning techniques. *Atmospheric Research*, 212(May) :240–258, 2018.

[31] Adjon Kouassi, Paul Assamoi, Sylvain Bigot, Adama Diawara, Guy Schayes, Fidèle Yoroba, and Benjamin Kouassi. Étude du climat Ouest-Africain à l'aide du modèle atmosphérique régional M.A.R. *Climatologie*, 7 :39–55, 2010.

[32] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553) :436–444, 2015.

[33] Cheikh Waly Diedhiou. Analyse pluviométrique au Sénégal, récentes modifications, influence sur le régime hydrologique : focus sur la Casamance. Technical report, Université Assane Seck de Ziguinchor, 2017.

[34] Chong-Yu Wetterhall, Fredrik and Bardossy, Andras and Chen, Deliang and Halldin, Sven and Xu. Daily precipitation-downscaling techniques in three Chinese regions. *Water Resources Research*, 42(11), 2006.

[35] Deepashree Raje and P. P. Mujumdar. A comparison of three methods for downscaling daily precipitation in the Punjab region. *Hydrological Processes*, 25(23) :3575–3589, 2011.

[36] N S Altman. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. 46(3) :175–185, 1992.

[37] Subhrendu Gangopadhyay, Martyn Clark, and Balaji Rajagopalan. Statistical downscaling using K-nearest neighbors. *Water Resources Research*, 41(2) :1–23, 2005.

[38] Vladimir Vapnik. *The support vector method of function estimation*. Springer, 1998.

[39] Ravi S Tripathi, Shivam and Srinivas, VV and Nanjundiah. Downscaling of precipitation for climate change scenarios : a support vector machine approach. *Journal of hydrology*, 330(3-4) :621–640, 2006.

[40] D. A. Sachindra, F. Huang, A. Barton, and B. J.C. Perera. Least square support vector and multi-linear regression for statistically downscaling general circulation model outputs to catchment streamflows. *International Journal of Climatology*, 33(5) :1087–1106, 2013.

[41] Ronei Jesus Bona, Evandro and Marquetti, Izabele and Link, Jade Varaschim and Makimori, Gustavo Yasuo Figueiredo and da Costa Arca, Vinicius and Lemes, André Luis Guimares and Ferreira, Juliana Mendes Garcia and dos Santos Scholz, Maria Brigida and Valderrama, Patric. Support vector machines in tandem with infrared spectroscopy for geographical classification of green arabica coffee. *LWT-Food Science and Technology*, 76 :330–336, 2017.

[42] Guang Bin Huang, Qin Yu Zhu, and Chee Kheong Siew. Extreme learning machine : Theory and applications. *Neurocomputing*, 70(1-3) :489–501, 2006.

[43] Gao Huang, Guang Bin Huang, Shiji Song, and Keyou You. Trends in extreme learning machines : A review. *Neural Networks*, 61 :32–48, 2015.

[44] Jerome H. Friedman and Werner Stuetzle. Projection pursuit regression. *Journal of the American Statistical Association*, 76(376) :817–823, 1981.

[45] B. C. Hewitson and R. G. Crane. Climate downscaling : Techniques and application. *Climate Research*, 7(2) :85–95, 1996.

[46] D. Maraun, F. Wetterhall, A. M. Ireson, R. E. Chandler, E. J. Kendon, M. Widmann, S. Brienen, H. W. Rust, T. Sauter, M. Themel, V. K.C. Venema, K. P. Chun, C. M. Goodess, R. G. Jones, C. Onof, M. Vrac, and I. Thiele-Eich. Precipitation downscaling under climate change : Recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3) :1–34, 2010.

[47] Keisuke Wilby, Robert L and Hassan, Hany and Hanaki. Statistical downscaling of hydrometeorological variables using general circulation model output. *Journal of Hydrology*, 205(1-2) :1–19, 1998.

[48] F Giorgi, B Hewitson, J Christensen, M Hulme, H Von Storch, P Whetton, R Jones, L Mearns, and C Fu. Regional Climate Information Evaluation and Projections. *Climate Change 2001 : The Scientific Basis. Contribution of Working Group I to the Third Assessment Report of the Intergovernmental Panel on Climate Change*, 2001.

[49] Eduardo Zorita and Hans Von Storch. The analog method as a simple statistical downscaling technique : Comparison with more complicated methods. *Journal of Climate*, 12(8 PART 2) :2474–2489, 1999.

[50] John and others Kalnay, Eugenia and Kanamitsu, Masao and Kistler, Robert and Collins, William and Deaven, Dennis and Gandin, Lev and Iredell, Mark and Saha, Suranjana and White, Glenn and Woollen. The NCEP/NCAR 40-year reanalysis project. *Bulletin of the American meteorological Society*, 77(3) :437–472, 1996.

[51] Karl Pearson. *Correlation coefficient*. Royal Society Proceedings, 1895.

[52] D Anandhi, Aavudai and Srinivas, VV and Nanjundiah, Ravi S and Nagesh Kumar. Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. *International Journal of Climatology : A Journal of the Royal Meteorological Society*, 28(3) :401–420, 2008.

[53] R L Wilby, S P Charles, E Zorita, B Timbal, P Whetton, and L O Mearns. Guidelines for Use of Climate Scenarios Developed from Statistical Downscaling Methods. *Analysis*, 27(August) :1–27, 2004.

[54] Geoge H Wilby, Robert L and Hay, Larren E and Leavesley. A comparison of downscaled and raw GCM output : implications for climate change scenarios in the San Juan River basin, Colorado. *Journal of Hydrology*, 225(1-2) :67–91, 1999.

[55] Kamal Ahmed, Shamsuddin Shahid, Sobri Bin Haroon, and Xiao Jun Wang. Multilayer perceptron neural network for downscaling rainfall in arid region : A case study of Baluchistan, Pakistan. *Journal of Earth System Science*, 124(6) :1325–1341, 2015.

[56] Geoffrey M Pervez, Md Shahriar and Henebry. Projections of the Ganges–Brahmaputra precipitation—Downscaled from GCM predictors. *Journal of Hydrology*, 517 :120–134, 2014.

[57] Karl E Taylor. in a Single Diagram. *Journal of Geophysical Research*, 106(D7) :7183–7192, 2001.

[58] CT Devak, Manjula and Dhanya. Downscaling of precipitation in Mahanadi Basin, India using support vector machine, K-nearest neighbour and hybrid of support vector machine with K-nearest neighbour. In *Geostatistical and geospatial approaches for the characterization of natural resources in the environment*, pages 657–663. Springer, 2016.

[59] Zachary A. Holden, John T. Abatzoglou, Charles H. Luce, and L. Scott Baggett. Empirical downscaling of daily minimum air temperature at very fine resolutions in complex terrain. *Agricultural and Forest Meteorology*, 151(8) :1066–1073, 2011.