Rapport d’étude

Analyse des mesures atmosphériques et production de statistiques

Lot 5 - AO5 Bretagne Sud - Bouée Nord

Version 1.0 du 28/02/2022

Client Direction Générale de l’Énergie et du Climat du Ministère de la Transition Écologique	@guillaume.rouviere@developpement-durable.gouv.fr 01 40 81 20 98	MINISTÈRE DE LA TRANSITION ÉCOLOGIQUE Liberté Égalité Fraternité

Chargé d’affaires Météo-France D2C Toulouse	@ +33 5 61 07 86 85	METEO FRANCE
Chef de projet Météo-France DSM/CS/DC	@Raphael.Legrand@meteo.fr +33 5 61 07 86 11	

- page laissée intentionnellement vide -

Documents de référence antérieurs

	Intitulé	Référence	Date	Version
DR1	Acquisition et suivi des mesures sur le site AO5 Bretagne – Bouée Nord	AO5_BretagneSud_Lot4_Rapport-final_V2_20220126.odt	26/01/2022	1

Évolutions successives

Référence	Date	Version	Évolution
DGEC_AO5_BretagneSud_Bouée_Nord_Lot5_Rapport_V1_20220228	02/02/2022	V1	Création

Signatures

	Nom		Service		Signature
Rédacteur(s)	Daka KEITA		DSM/CS/DC		
Relecteur(s)	Julie Capo	Valentine Chatel	DSM/CS/ Energie	DSM/CS/ DC	
	Raphaël Legrand		DSM/CS/DC/D		
Approbateur(s)	Christophe Jacolin		D2C/DV/PRO		

RÉSUMÉ

L'objectif de ce lot 5 est d'**analyser les mesures atmosphériques** notamment la force et la direction du vent, d'en **produire des statistiques** associées et d'**étendre les séries d'observations horaires** dans le cadre d'une problématique d'éolienne off-shore.
Ce rapport concerne l'étude d'une campagne de mesure réalisée entre le 08/07/2020 01H TU et le 07/10/2021 23H TU. Pendant cette période, il n'y a pas eu de mesure au mois de février 2020 et globalement peu au 1er trimestre 2021. De plus, la campagne s'étendant sur deux années différentes (2020 et 2021), 4 mois (juillet, août, septembre et octobre) présentent simultanément des mesures pour les deux années. Cela a ainsi affecté la restitution du cycle saisonnier que nous utilisons comme variable explicative. La variable qui prend en compte le cycle saisonnier a donc été adapté en faisant un regroupement de certains mois dans un mois virtuel afin de le garder dans l'étude.

L'**extension de série est réalisée sur l'observation horaire du vent à 100 m**, une hauteur qui revêt un caractère important pour la production éolienne. La **période reconstituée est d'environ 21 ans** en utilisant les données horaires du **modèle météorologique AROME** (à la résolution horizontale de 2,5 km) sur la période 2000-2020, permettant de disposer d'un jeu de données dense sur la zone étudiée.

Une étude complète réalisée sur deux sites de test est présentée en annexe. Elle détaille les différentes étapes ainsi que la méthodologie adoptée pour le choix du modèle statistique d'extension des séries horaires.

Dans cette étude préliminaire, plusieurs modèles statistiques (notamment arbre binaire de décision, modèle linéaire général, modèle linéaire avec anamorphose, forêt aléatoire et réseau de neurones) ont été testés afin de **choisir le modèle capable de restituer au mieux les caractéristiques du vent horaire observé**.
Les scores de qualité suivant sont utilisés pour la recherche du meilleur modèle : le biais, l'écart-type (ECT), l'erreur absolue moyenne (MAE), l'erreur quadratique moyenne (RMSE) et les scores Pierce Skill Score (PSS). En plus de ces scores, la courbe de fiabilité (QQ-Plot), le critère BIC (pour les modèles linéaires), la restitution des cycles diurne et annuel du vent, ainsi que des roses de vent sont utilisées comme indicateurs de qualité.
Après analyse des différents indicateurs, **le modèle de forêt aléatoire s'est révélé comme le meilleur des modèles statistiques que nous avons testés**. Nous avons donc décidé de l'utiliser pour réaliser l'extension des séries horaires dans le cadre des études du lot 5.

Pour vérification, les résultats de l'extension ont été confrontés aux données brutes du modèle AROME sur la période d'observation. Selon les scores mentionnés plus haut, calculées pour la force du vent, la direction du vent et la puissance produite par une éolienne, il est confirmé que l'extension statistique est de meilleur qualité que le modèle brut pour étendre la série de mesure du lidar Nord de la zone de l'AO5 Bretagne sud. A noter que la série brute du modèle AROME sera également mis à disposition des porteurs de projets.

Table des matières

Liste des illustrations

Liste des tableaux

1 Contexte de l'étude

Cette étude s'intéresse à l'analyse des mesures atmosphériques et à la production de statistiques sur le paramètre vent à 100 m.

Le domaine couvert par cette étude est la zone de l'AO5 Bretagne-Sud, zone sur laquelle un projet d'implantation d'éoliennes off-shore est prévuCe rapport traite de la méthode d'extension de la série d'observations horaires et des principaux résultats de la campagne de mesure réalisée entre le 08/07/2020 01H TU et le 07/10/2021 23H TU pour le LiDAR situé dans la partie nord de cette zone de l'AO5 Bretagne-Sud .

La période reconstituée est de presque 22 années (01/01/2000 – 07/07/2020). Cette période correspondà la profondeur de la base climatologique AROME 2,5 km utilisée.

1.1 Livrables

Type de données	Paramètres	Caractéristiques	Format
Série horaire complète à 100 mètres sur la période de la campagne de mesures	Vitesse et direction du vent horizontal	- Période : campagne de mesures - Hauteur : 100 m - Fréquence temporelle : horaire	CSV (avec en entête les métadonnées associées aux données) (date/valeur)
Statistiques de la série horaire complète à 100 mètres sur la période de campagne de mesures	- Vent moyen horizontal - distribution statistique associée (quantiles de force de vent) - variation diurne horaire de la force du vent sur la période - roses de vent horizontal fréquentielles - gradients verticaux de vent	Hauteur : 100 m	CSV (avec en entête les métadonnées associées aux données) (date/valeur)
Chronique longue durée	Vitesse et direction du vent horizontale	- Période : à minima 20 ans - Hauteur : 100 mètres - Fréquence temporelle : horaire	CSV (avec en entête les métadonnées associées aux données) (date/valeur)
Rapport de fin de campagne		Extension de la série d'observations horaires	.PDF

À noter que « la série horaire complète à 100 mètres sur la période de la campagne de mesures » ainsi que les « statistiques de la série horaire complète à 100 mètres sur la période de campagne de mesures » ont déjà été transmises lors de la livraison finale du lot 4 de ce présent AO.

2 Méthodologie d’extension de la série temporelle d’observations horaires

2.1 Méthodologie

Pour étendre la série temporelle d’observation de force de vent sur une période plus longue, on procède de la manière suivante :

- tout d’abord, on quantifie la différence entre la force du vent issue de la série temporelle LiDAR et la force du vent issue du modèle AROME au point de grille le plus proche du LiDAR choisi comme référence ;
- ensuite on explique cette différence à l’aide d’autres paramètres atmosphériques provenant du modèle AROME. Un lien statistique est alors établi entre la force du vent du LiDAR, la force du vent du modèle AROME et d’autres paramètres météorologiques explicatifs du modèle AROME ;
- le lien statistique est validé à l’aide des différents scores statistiques puis appliqué sur une série temporelle passée pour reconstituer la force du vent.

Le processus est le même pour la direction du vent que l’on décompose en deux paramètres – le vent zonal U et méridien V – sur lesquels le processus est appliqué.

Le schéma de l’illustration 2.1 présente la méthodologie complète adoptée pour la réalisation de l’extension de série.

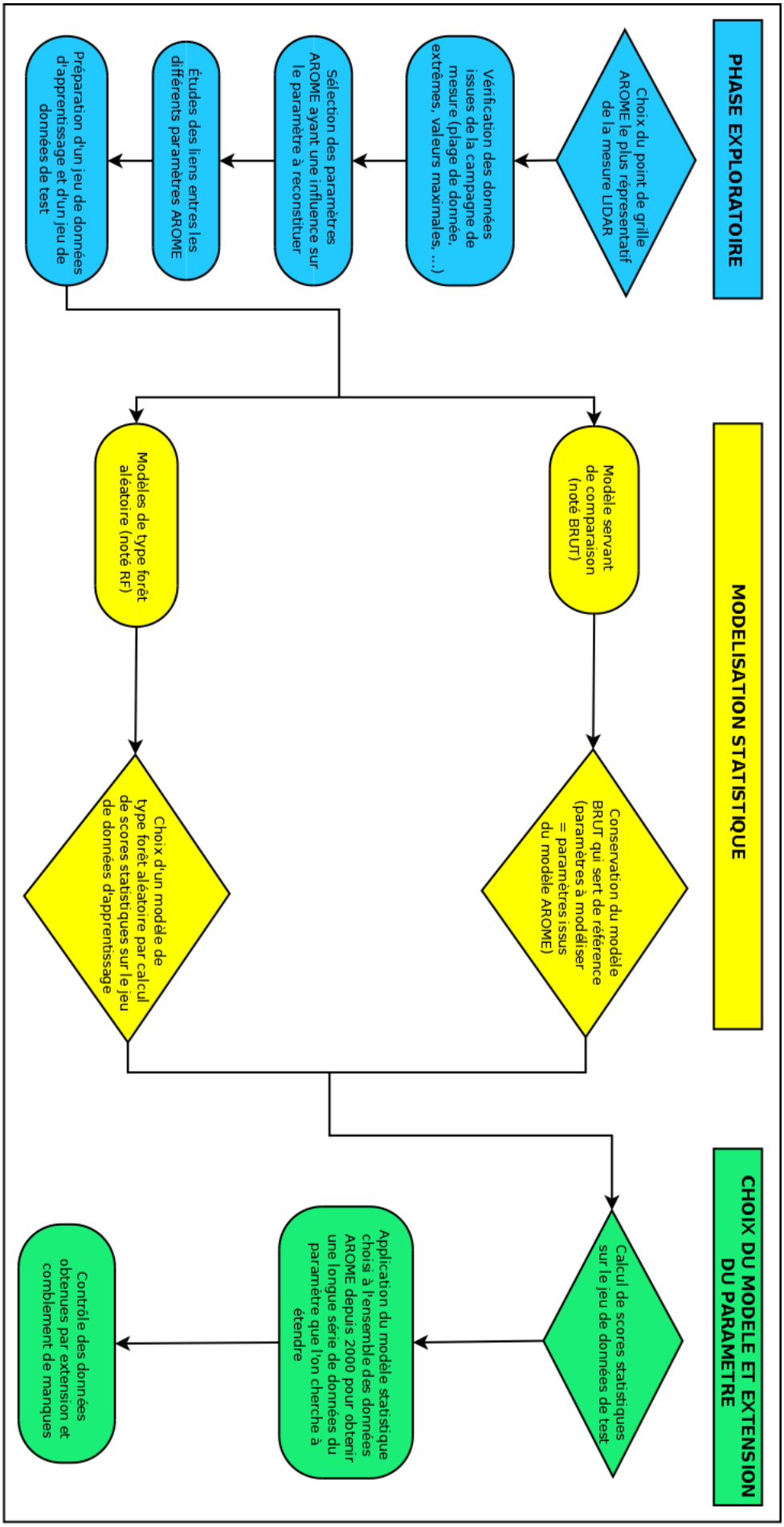

Illustration 2.1: Schéma récapitulatif de la méthodologie employée

2.2 Phase exploratoire des données d'entrée

La modélisation statistique est précédée d'une phase d'exploration des données dans laquelle on analyse les liens entre les variables afin d'aboutir à un sous-ensemble de données pertinentes pour la mise en place des modèles statistiques.

Pour les travaux de ce rapport, il s'agit notamment :

- d'identifier le point de grille AROME de référence dont les caractéristiques du vent (force et direction) se rapproche au mieux de ceux du point d'observation,
- d'identifier les paramètres pouvant expliquer les variations de la force et de la direction du vent au cours du temps, pour ensuite les intégrer dans les modèles statiques.

À l'issue de cette analyse, on dispose d'une sélection des variables (dites variables explicatives), qui sont utilisées par la suite pour ajuster les modèles statistiques aux données d'observations.

2.2.1 Identification du point AROME de référence

Le choix du point AROME de référence s'effectue en s'appuyant sur les indicateurs de qualité (B95+) des roses de vent. Ces indicateurs de qualité concernent notamment quatre critères notés C1, C2, C3 et C4 présentés ci-dessous. Ils permettent de comparer deux roses des vents entre-elles sur des critères comme la distribution de la force du vent par classe ou la direction du vent par secteur.

Soit f_0^i la rose des vents climatologiques de référence et f_m^i la rose des vents du modèle AROME. Ces roses fréquentielles sont de 18 secteurs et 4 classes de vent ([0 ; 3[,[3 ; 9[,[9 ; 12[et >12 m/s).

- **Le critère C1** est le critère principal où la comparaison se fait sur l'ensemble des classes. Il permet d'étudier la qualité de la modélisation des fréquences de **force et direction du vent simultanément :**

$$C1 = 100 - 0.5 \times \sum_{i=1}^{18\times 4} |f_0^i - f_m^i|$$

- **Le critère C2** est celui pour lequel les classes de vitesse sont regroupées et les vents calmes (<3 m/s, pour lesquels on sait qu'il y a de grosses incertitudes instrumentales sur la direction) ne sont pas pris en compte. Le critère C2 permet d'étudier la qualité de la modélisation des fréquences de **direction du vent :**

$$C2 = 100 - 0.5 \times \sum_{i=1}^{18} |\sum_{cl=1}^{3} f_0^i - \sum_{cl=1}^{3} f_m^i|$$

- **Le critère C3** est celui pour lequel les classes de direction sont regroupées. I permet d'étudier la qualité de la modélisation des fréquences de **force du vent** :

$$C3 = 100 - 0.5 \times \sum_{i=1}^{4} |\sum_{secteur=1}^{18} f_0^i - \sum_{secteur=1}^{18} f_m^i|$$

Ces critères de qualité (C1, C2 et C3) varient entre 0 et 100, 100 étant le meilleur score. Ils ont pour principal avantage leur simplicité d'interprétation, en particulier si le critère global est mauvais, on peut rapidement détecter si l'erreur vient d'un biais en direction ou d'un biais en force du vent. Par contre, ils présentent le défaut d'être fortement dépendant du nombre de classes : plus on a de classes, plus le critère est sévère. Dans notre cas, nous travaillons sur 18 classes en direction (tous les 20 degrés), et 4 classes en force du vent. Ceci rend le critère de direction très exigeant : un biais de 20° (possible lorsque la résolution horizontale du modèle n'est pas suffisante pour représenter correctement le relief par exemple) dans la direction du vent modèle, donnera un critère de direction (C2) nulle.

- Enfin le dernier **critère C4** est la corrélation circulaire. Il est appliqué uniquement sur la **direction du vent**. Le coefficient de corrélation est calculé comme la corrélation de Pearson pour deux variables linéaires X et Y . Dans la formule de calcul, $(x_i-\overline{x})$ et $(y_i-\overline{y})$ sont remplacés par $\sin(x_i-\overline{x})$ et $\sin(y_i-\overline{y})$ car les deux variables X et Y sont circulaires, où :
  - x_i et y_i sont les valeurs de directions pour X et Y .
  - $\overline{x}$ et $\overline{y}$ sont les directions moyennes des échantillons

$$C4=\frac{\sum_{i=1}^{n}\sin(x_i-\overline{x})\times\sin(y_i-\overline{y})}{\sqrt{\sum_{i=1}\sin^2(x_i-\overline{x})}\times\sqrt{\sum_{i=1}\sin^2(y_i-\overline{y})}}$$

2.2.2 Identification des variables explicatives

Le processus détaillé de la sélection des variables explicatives est décrit dans la section 7.4 de l'annexe 2.

Le choix des variables explicatives est guidé par la connaissance météorologique de l'origine du vent. Les variables ci-dessous ont été pré-sélectionnées pour participer à la modélisation statistique. Celles présentant une trop forte covariance entre elles ont été séparées et intégrées dans une analyse en composante principale (ACP). **Le détail de la sélection des variables explicatives est présenté dans la section 3.1.2 pour la force du vent et la section 3.1.3 pour la direction du vent.**

- Prise en compte du paramètre à prédire :
  - Pour la force du vent :
    - **FF** : la force du vent AROME aux niveaux 10, 50, 100, 250 et 500 m – variable quantitative notée FFARO_10, FFARO_50, FFARO_100, FFARO_250 et FFARO_500 dans la suite du rapport**,**
      et
    - **SECTEUR :** la direction du vent AROME à 100 m par secteur de 20° – variable catégorielle à 18 facteurs – notée SECTEURARO_100 dans la suite du rapport,
  - Pour la direction du vent :
    - **U :** le vent zonal AROME aux niveaux 10, 100, 250 et 500 m – variable quantitative notée UARO_10, UARO_100, UARO_250 et UARO_500 dans la suite du rapport**, et**
    - **V** : le vent méridien AROME aux niveaux 10, 100, 250 et 500 m – variable quantitative notée VARO_10, VARO_100, VARO_250 et VARO_500 dans la suite du rapport**,**
- Prise en compte de la situation météorologique générale :
  - **PMER** : la pression AROME réduite au niveau de la mer – variable quantitative notée PMERARO dans la suite du rapport,
  - **TPW850**: la température pseudo-adiabatique potentielle du thermomètre mouillé AROME à 850 hPa. Cette variable caractérise une masse d'air. C'est une variable quantitative – notée TPWARO_850 dans la suite du rapport,
  - **HU2M :** l'humidité AROME relative à 2 m – variable quantitative notée HUARO_2 dans la suite du rapport,
- Prise en compte d'éléments de stabilité de l'atmosphère :

- **T :** la température AROME à 2, 50, 100, 250 et 500 m – variable quantitative notée TARO_2, TARO_50, TARO_100, TARO_250, TARO_500 dans la suite du rapport.
- **TKE et SQRTTKE** : la turbulence AROME et sa racine carrée à 10, 50, 100 et 160 m. On exploite la racine carrée de cette variable, de manière à privilégier un lien linéaire avec la force du vent de l'observation. Ce sont des variables quantitatives – noté TKEARO_10, TKEARO_50, TKEARO_100, TKEARO_160, SQRTTKEARO_10, SQRTTKEARO_50, SQRTTKEARO_100 et SQRTTKEARO_160 dans la suite du rapport,

- Prise en compte des cycles diurne et saisonnier du vent :
  - cycle diurne **HH** : variable catégorielle à plusieurs facteurs, fonction de la variable d'intérêt (FF ou DD), représentant globalement l'heure de la journée – notée respectivement **RegHH2_FF et RegHH_DD** dans la suite du rapport,
  - cycle saisonnier **MM** : variable catégorielle à plusieurs facteurs, fonction de la variable d'intérêt (FF ou DD), représentant globalement le mois de l'année – notée respectivement **RegMM2 et RegMM** dans la suite du rapport. Cette variable est notamment sujette au taux de disponibilité de la mesure LiDAR au long de la campagne de mesure.

2.3 Modélisation statistique par forêt aléatoire

La modélisation statistique permet d'établir un lien statistique robuste entre les variables explicatives X^i et les observations réelles de la force du vent à 100 m correspondant à la **variable** Y **à expliquer** :

$$Y = f\left(X^1, \ldots, X^i, \ldots, X^n\right) \quad .$$

Le type des variables statistiques diffère selon l'espace dans lequel elles prennent leurs valeurs.

Elles peuvent être qualitatives à valeurs dans un ensemble de cardinal fini ou quantitatives à valeurs réelles voire fonctionnelles (à valeurs dans un espace de dimension infinie). Certaines méthodes de modélisation s'adaptent à tout type de variables explicatives tandis que d'autres sont spécialisées.

Si la variable Y à expliquer est qualitative, on parle de discrimination, classement ou reconnaissance de forme tandis que si la variable Y est quantitative comme pour cette étude, on parle d'un problème de régression.

Certaines méthodes sont dites spécifiques (régression linéaire, analyse discriminante), c'est-à-dire qu'elles sont soumises à des hypothèses de travail qu'il s'agit de vérifier, tandis que d'autres s'utilisent sans hypothèses contraignantes (arbres de décision, forêt aléatoire, réseau de neurone, etc).

Compte tenu du problème à résoudre (estimation de la force du vent à 100 m) et de l'expérience de Météo-France en matière de famille de modèles pouvant répondre au besoin, nous avons pré-sélectionné les types de modèles suivants : **arbre binaire de décision, modèle linéaire général, forêt aléatoire, modèle linéaire avec anamorphose et réseau de neurone** (descriptions faites en annexe au paragraphe 6)**.** Suite à l'étude approfondie de l'ensemble de ces méthodes statistiques présentées en annexe au paragraphe 7, **c'est le modèle de forêt aléatoire qui a été sélectionné pour réaliser les extensions des séries horaires dont celle de l'AO5 Bretagne-Sud.**

2.3.1 Description du modèle de forêt aléatoire

La méthode consiste à utiliser le hasard pour améliorer les performances des algorithmes ayant une faible capacité de classification. Une part d'aléatoire est ajoutée au cours de la construction des différents arbres qui seront alors agrégés ensembles pour former une forêt.

La forêt aléatoire se construit en concevant un arbre sur un sous-échantillon tiré aléatoirement (ou échantillon « out-of-bag »). Ensuite, pour chacun des arbres à construire, un sous-ensemble de $q \leq P$ variables explicatives est sélectionné aléatoirement et servira à leur élaboration respective.

L'objectif de cette approche est de rendre les arbres construits plus indépendants entre eux, ce qui offre de meilleures performances lors de l'agrégation en forêt. L'approche possède l'avantage d'être très fructueuse en grande dimension et d'être simple à mettre en œuvre. L'utilisation de la forêt aléatoire permet également de s'affranchir de toute phase d'élagage et de tout problème lié à la multicolinéarité des variables.

Comme pour tout modèle construit par agrégation, il n'y a pas d'interprétation directe. Néanmoins des informations pertinentes sont obtenues par le calcul et la représentation graphique d'indices proportionnels à l'importance de chaque variable dans le modèle agrégé et donc de sa participation à la régression ou à la discrimination.

Deux critères sont ainsi proposés pour évaluer l'importance de la j-ème variable.

- Le premier (Mean Decrease Accuracy) repose sur une permutation aléatoire des valeurs de cette variable. Plus la qualité, de la prévision est dégradée par la permutation des valeurs de cette variable, plus celle-ci est importante. Une fois le b-ème arbre construit, l'échantillon out-of-bag est prédit par cet arbre et l'erreur estimée enregistrée. Les valeurs de la j-ème variable sont aléatoirement permutées dans l'échantillon out-of-bag et l'erreur à nouveau calculée. La décroissance de la qualité de prévision est moyennée sur tous les arbres et utilisée pour évaluer l'importance de la variable j dans la forêt. Il s'agit donc d'une mesure globale mais indirecte de l'influence d'une variable sur la qualité des prévisions.
- Le deuxième (Mean Decrease Gini) est local et basé sur la décroissance d'entropie ou encore la décroissance de l'hétérogénéité définie à partir du critère de Gini. L'importance d'une variable est alors une somme pondérée des décroissances d'hétérogénéité induites lorsqu'elle est utilisée pour définir la division associée à un nœud.

Le paramètre le plus important à fixer est celui du nombre d'arbres à générer.

Référence : *Robin Genuer, Jean-Michel Poggi. Arbres CART et Forêts aléatoires, Importance et sélection devariables. 2017. hal-01387654v2*

2.3.2 Les étapes de l'apprentissage

Les traitements s'enchaînent selon le schéma suivant :

1. Extraction des données.
2. Exploration des données pour la détection de valeurs aberrantes ou seulement atypiques, d'incohérences, pour l'étude des distributions des structures de corrélation, recherche de typologies, pour des transformations des données…
3. Partition de l'échantillon en fonction de sa taille et des techniques qui seront utilisées pour estimer une erreur de prévision en vue des étapes de choix de modèle, puis de choix et certification de méthode.
4. Pour la forêt aléatoire :
    - estimer ce modèle pour une valeur donnée du paramètre de complexité, le nombre d'arbres.
    - optimiser le nombre d'arbre en fonction de la technique d'estimation de l'erreur retenue : validation croisée (notre cas), approximation par pénalisation de l'erreur d'ajustement (critères de BIC).
5. Ré-estimation du modèle de forêt aléatoire avec sa complexité optimisée à l'étape précédente sur l'ensemble des données.
6. Comparaison du modèle obtenu avec le modèle AROME sur l'ensemble des données.

2.3.3 Choix des échantillons d'apprentissage et de test

Pour la définition du modèle statistique, nous devons séparer la série d'observation initiale en deux échantillons :

- l'échantillon d'apprentissage (on prend 70 % de la série initiale) pour ajuster les modèles statistiques,
- l'échantillon de test (on prend les 30 % restant) pour vérifier la robustesse du modèle.

L'échantillon initial étant petit, nous avons mis en place une stratégie nous permettant de séparer de façon itérative (20 itérations) la série d'observation initiale en deux échantillons (apprentissage et test). Les données seront alors tour à tour dans l'échantillon d'apprentissage et dans l'échantillon de test.

2.3.4 Estimation de l'erreur des modèles statistiques sur ces échantillons

Dans le but d'établir un modèle parcimonieux, on cherche à trouver un compromis entre complexité du modèle et erreur d'estimation. En effet, plus un modèle est complexe (intégrant plus de paramètres) plus il est capable de s'ajuster aux données d'apprentissage, engendrant ainsi une erreur d'ajustement moindre. Cependant, un tel modèle peut se révéler défaillant lorsqu'il sera appliqué à des données qui n'ont pas participé à son estimation.

Pour minimiser ce risque de surajustement, nous nous sommes fixés la règle de prendre le nombre minimum d'arbre à score équivalent.

Un ensemble de critères définis comme suit est utilisé pour mesurer la qualité d'estimation de notre erreur de modélisation :

- **RMSE**: l'erreur quadratique moyenne mesure la distance entre le modèle et la référence. Plus il est proche de 0, plus le modèle est proche de la référence.

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i-O_i)^2}$$

  Ce score est également calculé par classes de direction de vent de 20°.

- **BIAIS** : il caractérise l'erreur systématique du modèle ; plus il est proche de 0, plus le modèle est proche en moyenne des observations. Un BIAIS positif (négatif), signifie que le modèle surestime (sous-estime) le paramètre considéré.

$$BIAIS=\frac{1}{N}\sum_{i=1}^{N}(P_i-O_i)$$

- **ECT** : l'écart type donne la précision du modèle. Plus il est proche de 0, meilleur est le modèle.

$$ECT=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(P_i-O_i)^2-\left(\frac{1}{N}\sum_{i=1}^{N}(P_i-O_i)\right)^2}$$

  à noter que RMSE, BIAIS et ECT sont liés par la relation suivante: $RMSE^2=BIAIS^2+ECT^2$

- **MAE** : L'erreur absolue moyenne mesure l'ampleur moyenne des erreurs dans un ensemble de prédictions, sans tenir compte de leur direction. Plus il est proche de 0, meilleur est le modèle.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |P_i - O_i|$$

- **PSS** : Peirce Skill Score

  Soit un événement E (par exemple FF > 5 m/s), un échantillon d'observations et un modèle offrant la possibilité de prévoir cet événement, il est possible de construire la table de contingence suivante :

		observations	
		E est vrai sur l'observation	E est faux sur l'observation
prévisions	E est vrai sur la prévision	$n_{11}(s)$	$n_{10}(s)$
	E est faux sur la prévision	$n_{01}(s)$	$n_{00}(s)$

  Taux de bonnes prévisions: **H** = $n_{11}(s)/(n_{11}(s)+ n_{01}(s))$

  Taux de fausses alertes: **FA** = $n_{10}(s)/(n_{10}(s)+ n_{00}(s))$

  PSS = H – FA

  Le score PSS est compris entre -1 et 1. Si ce score est supérieur à 0, le taux de bonnes prévisions est supérieur à celui des fausses alertes. Plus il est proche de 1, meilleur est le modèle. Il est particulièrement adapté, pour évaluer la capacité du modèle à prévoir les valeurs extrêmes.

Ces critères lorsqu'ils sont appliqués à l'échantillon qui a été utilisé pour construire le modèle statistique (échantillon d'apprentissage) ne peuvent être qu'une estimation biaisée, car trop optimiste, de l'erreur de prévision puisqu'elle est liée aux données qui ont servi à l'ajustement du modèle.

Il est alors possible d'effectuer une validation dite croisée dont l'idée est d'itérer l'estimation de l'erreur sur plusieurs échantillons de validation puis d'en calculer la moyenne. Cette méthode a l'intérêt de réduire la variance des erreurs d'échantillonnages et ainsi améliorer la précision des scores lorsque la taille de l'échantillon initial est trop réduite pour en extraire des échantillons de validation et test de taille suffisante.

```
ALGORITHME DE LA VALIDATION CROISÉE :
1: Découper aléatoirement l'échantillon d'apprentissage en K parts (K-
fold) de tailles approximativement égales selon une loi uniforme ;
2: pour k allant de 1 à K do
      3: mettre de côté l'une des parties,
      4: estimer le modèle sur les K- 1 parties restantes,
      5: calculer l'erreur sur chacune des observations qui n'ont pas
      participé à l'estimation
6: fin boucle
7: moyenner toutes ces erreurs pour aboutir à l'estimation par validation
croisée.
```

En plus, du calcul de la moyenne des scores précédents sur les 20 échantillons tests, deux scores supplémentaires sont réalisées. Ils sont calculés sur la série contenant les prédictions des 20 échantillons de test concaténés, noté « échantillon test concaténé » par la suite :

- Corrélation : score permettant de tester l'association entre des échantillons appariés, en utilisant le coefficient de corrélation de Spearman.
- La courbe de fiabilité (**QQ-Plot**) : technique graphique qui vise à comparer deux fonctions de répartition en traçant les quantiles de l'une en fonction de ceux de l'autre (par exemple, le FF observé versus le FF estimé par le modèle statistique). Il peut être utilisé afin de déterminer si deux séries de données obéissent à la même loi de probabilité, pour comparer les fonctions de répartition

entre elles ou pour vérifier qu'un ensemble de données empiriques suit une certaine loi de probabilité théorique. La distribution générant le QQ-plot le plus proche d'une ligne droite de pente 1 peut alors être considérée comme le meilleur choix.

2.4 Qualification de l'extension par forêt aléatoire

Une fois le modèle statistique choisi grâce aux scores ci-dessus, l'apprentissage est à nouveau réalisé sur l'ensemble de la période d'observation. Les prédictions du modèle sur plus de 20 ans sont utilisés pour étendre la série horaire d'observation.

La qualité de cette extension est enfin comparée au modèle AROME notamment pour vérifier l'amélioration de la restitution des cycles diurne et saisonnier et des roses de vent (dont B95+).

La combinaison pour FF et DD du modèle de forêt aléatoire et d'AROME est aussi étudiée.

Enfin, le lien entre vitesse de vent et puissance produite par une éolienne étant non-linéaire, nous réalisons des scores (Biais, RMSE, PSS…) en exploitant une courbe de charge théorique pour la conversion vent/puissance. La courbe de fiabilité (**QQ-Plot**) est également considérée ici.

Afin de calculer la puissance théorique, nous utilisons la courbe de charge présentée en illustration 2.2. La puissance suit les formules du tableau 2.1. Pour plus de détail sur les caractéristiques de l'éolienne choisie, le lecteur pourra se référer au lien suivant https://github.com/IEAWindTask37/IEA-10.0-198-RWT.

Tableau 2.1: Caractéristiques de la courbe de charge théorique utilisée.

Force du vent (m/s)	FF < 4	4≤ FF ≤ 10,7	10,7 < FF ≤ 25	>25
Puissance (MW)	0	*P(FF)*	10	0

où $$P(FF) = a_3 \times FF^3 + a_2 \times FF^2 + a_1 \times FF + a_0$$

avec a_3 , a_2 , a_1 et a_0 valant respectivement 0.009808327 , -0.04421503 , 0.36130470 , et -0.9848518 déterminés à partir de l'ajustement de la courbe de charge théorique de l'illustration 2.2.

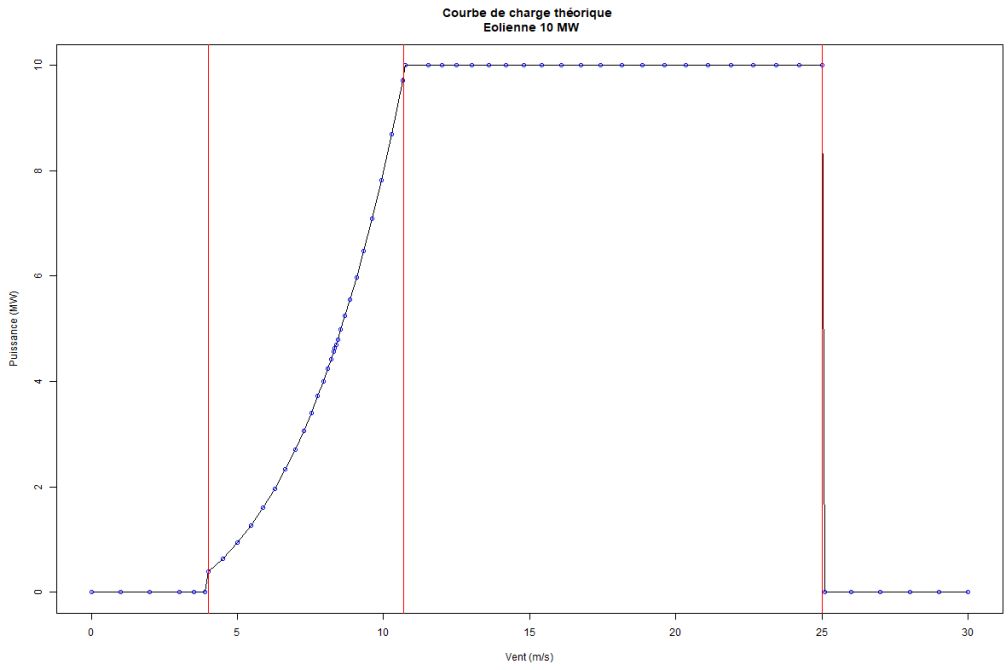

Illustration 2.2: Courbe de charge théorique d'une éolienne de 10MW

3 Principaux résultats sur la modélisation de FF et DD à 100 m

3.1 Phase exploratoire

3.1.1 Choix du point AROME de référence

Les variables explicatives des modèles statistiques sont issues des données du point de grille du modèle AROME servant de référence. Ce point de référence est choisi à travers la capacité de sa rose des vents à être le plus proche possible de la rose des vents du point d'observation.

L'illustration 3.1 présente les roses des vents du point d'observation et des quatre points AROME voisins.

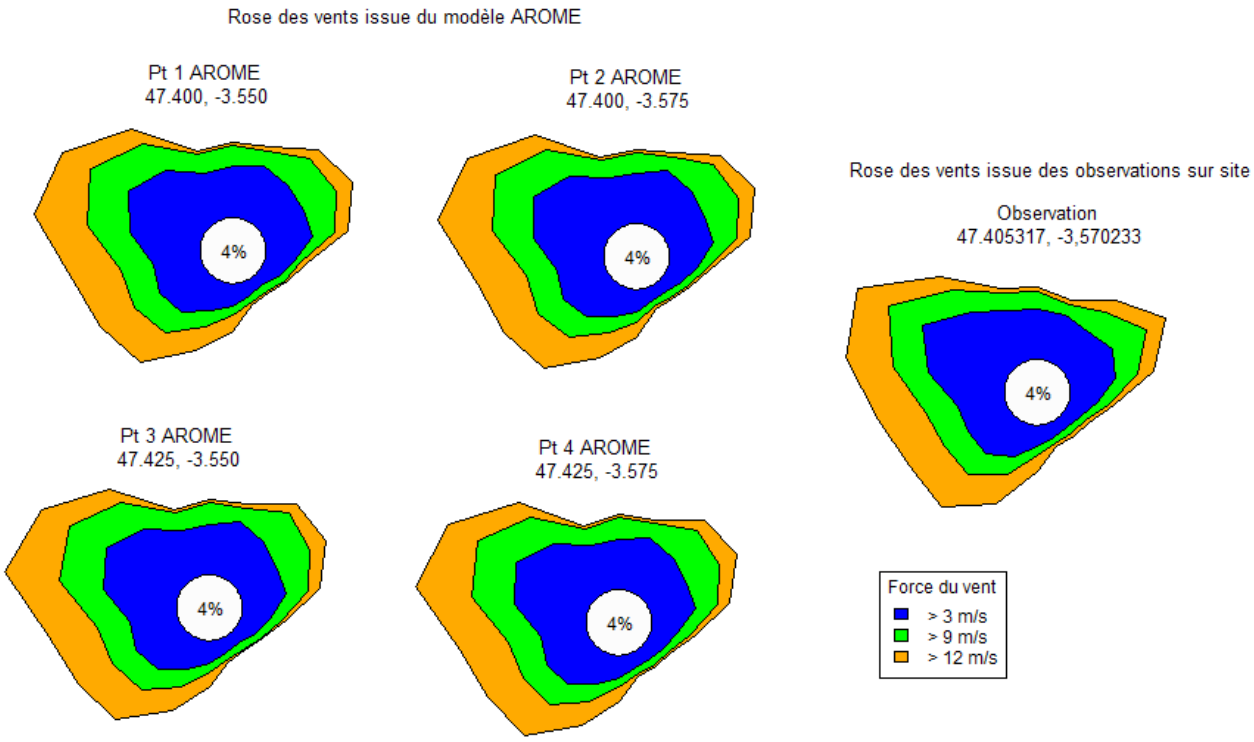

Illustration 3.1: AO5 Bretagne-Sud Bouée Nord – Choix du point AROME de référence - Roses des vents du point d'observation et des quatre points AROME voisins.

Visuellement les roses des quatre points AROME sont très similaires. Il est très difficile de distinguer celui qui se rapproche le plus de la rose des vents de l'observation. C'est donc les scores de qualité B95+ (présentés au paragraphe 2.2.1) qui ont été décisifs pour choisir le point de référence (tableau 3.1).

Tableau 3.1: AO5 Bretagne-Sud Bouée Nord – Choix du point AROME de référence – Scores B95+ des quatre points AROME voisins de l'observation

Score	Latitude	Longitude	C1 (DD&FF)	C2 (DD)	C3 (FF)	C4 (Corrélation circulaire)
Point 1	47.400	-3.550	92.17439	94.71302	96.26380	0.9850581
Point 2	**47.400**	**-3.575**	**92.94150**	**94.76821**	**97.31236**	**0.9853091**
Point 3	47.425	-3.550	92.20751	94.54746	96.36865	0.9833867
Point 4	47.425	-3.575	92.61589	94.63576	96.58940	0.983374

Les scores sont très proches les uns des autres pour chacun des quatre points, mais le point 2 à des scores C1, C2, C3 et C4 légèrement meilleurs que ceux des autres points. Par conséquent, **c'est le point 2, situé au Nord-Est du point d'observation, qui a été choisi comme point AROME de référence.**

3.1.2 Sélection des variables explicatives de la force du vent

Pour mieux modéliser les valeurs extrêmes (retour d'expérience des tests d'extension de série d'observations précédents), nous avons cherché à **ajuster** des modèles statistiques, non pas sur la force FF du vent observé, mais sur **l'écart de cette force du vent avec la force du vent AROME**. Ainsi, la variable à prédire est définie par :

$$Y = FF_{obs} - FF_{AROME} = FFmFFARO$$

L'ensemble des variables identifiées au paragraphe 2.2.2 est analysé. On examine en effet les liens entre la variable à expliquer Y et les variables explicatives X^i, ou entre les variables explicatives. Cette analyse permet de s'assurer qu'il y ait assez de corrélation entre chaque X^i et Y, et éviter toute corrélation forte entre variables X^i.

Certaines variables explicatives sont fortement corrélées entre-elles et notamment au sein des nombreux niveaux disponibles pour les paramètres vent, température et turbulence. Nous avons donc choisi de garder un ou deux niveaux par paramètres avec des coefficients de corrélations avec FF élevés mais inférieurs à 0,96 entre-eux. Les autres niveaux sont utilisés dans l'ACP.

L'ACP a permis de sélectionner 5 nouvelles variables explicatives (notées respectivement **PC1_FF, PC2_FF, PC3_FF, PC4_FF et PC5_FF**) combinant les paramètres FFARO_10, FFARO_50, FFARO_250, TARO_2, TARO_50, TARO_250, TARO_500, TKEARO_50, TKEARO_100, TKEARO_160, SQRTTKEARO_50, SQRTTKEARO_100 et SQRTTKEARO_160.

L'illustration 3.2 présente la corrélation entre les variables qui ont été définitivement sélectionnées pour la mise en place du modèle statistique, et leurs corrélations avec la variable à prédire FFmFFARO.

Le graphique des corrélations montre que la variable à expliquer est faiblement corrélée linéairement avec les **14 variables explicatives choisies** auxquelles s'ajoutent les variables liées aux cycles de FF qui sont détaillées ci-dessous.

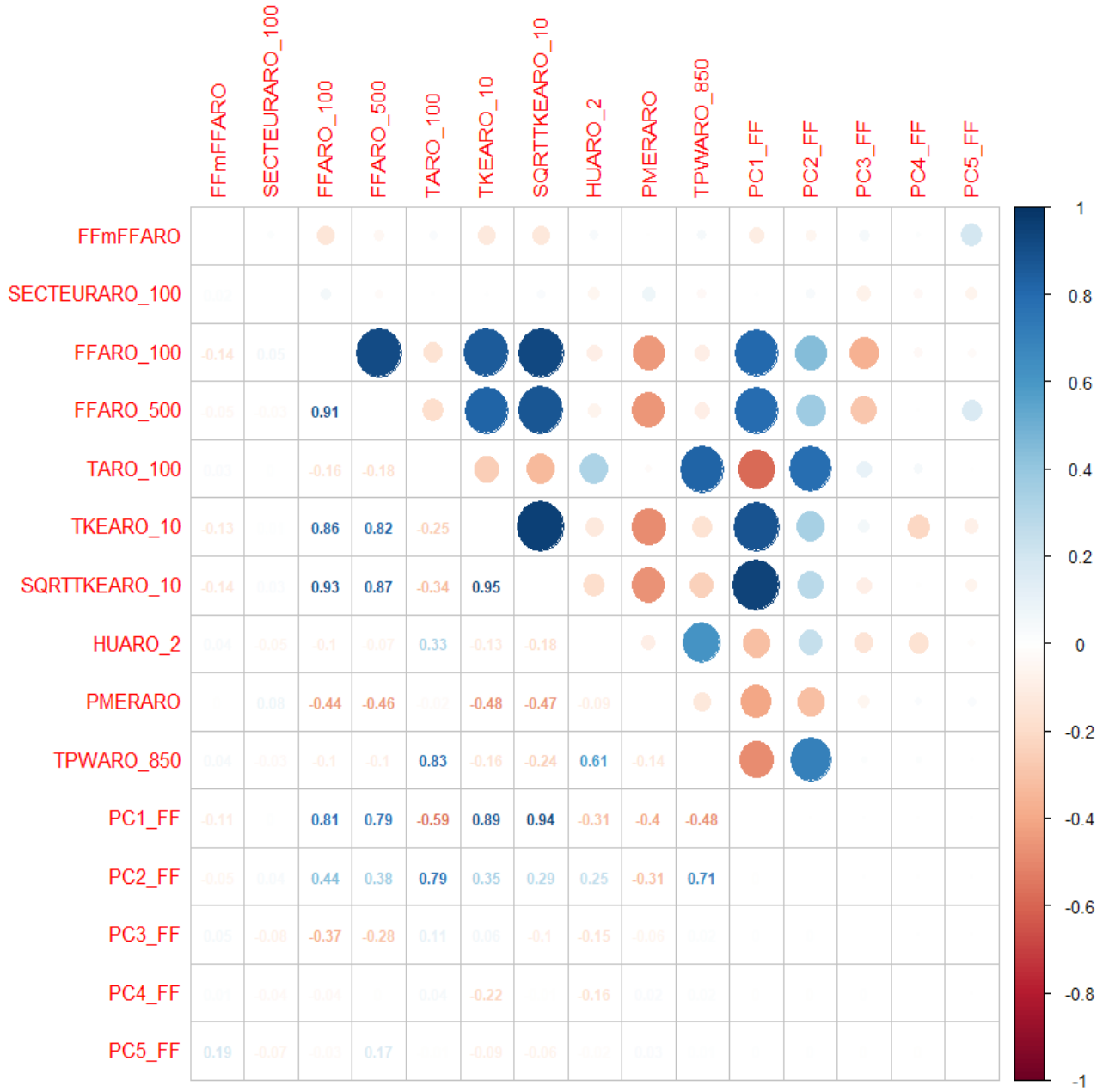

Illustration 3.2: AO5 Bretagne-Sud Bouée Nord – FF – Corrélation entre la variable à expliquer FFmFFARO et les variables explicatives. En haut : sous forme de bulles. En bas : chiffré (0 : absence de corrélation, 1 ou -1 : forte corrélation)

En effet, une dernière étape consiste en la définition des variables explicatives des cycles saisonnier et diurne pour la force du vent. Ces cycles sont déterminés à partir de la distribution du vent (illustration 3.3) en fonction des mois pour le cycle saisonnier (RegMM2) et des heures pour le cycle diurne (RegHH2_FF). Nous tenons également compte du taux de présence de données d'observation au cours de la campagne de mesure (tableau 3.2) pour ajuster la variable RegMM2.

Tableau 3.2: AO5 Bretagne-Sud Bouée Nord – FF et DD – Tableau de présence de mesure pour chaque mois de la campagne de mesure LiDAR de juillet 2020 à octobre 2021

	Janv.	Fév.	Mars	Avril	Mai	Juin	Juillet	Août	Sept.	Oct.	Nov.	Déc.	Année
2020	0	0	0	0	0	0	551	678	657	724	669	711	3990
2021	442	0	510	663	688	617	679	685	632	154	0	0	5070
Somme	**442**	**0**	**510**	**663**	**688**	**617**	**1230**	**1363**	**1289**	**878**	**669**	**711**	**9060**

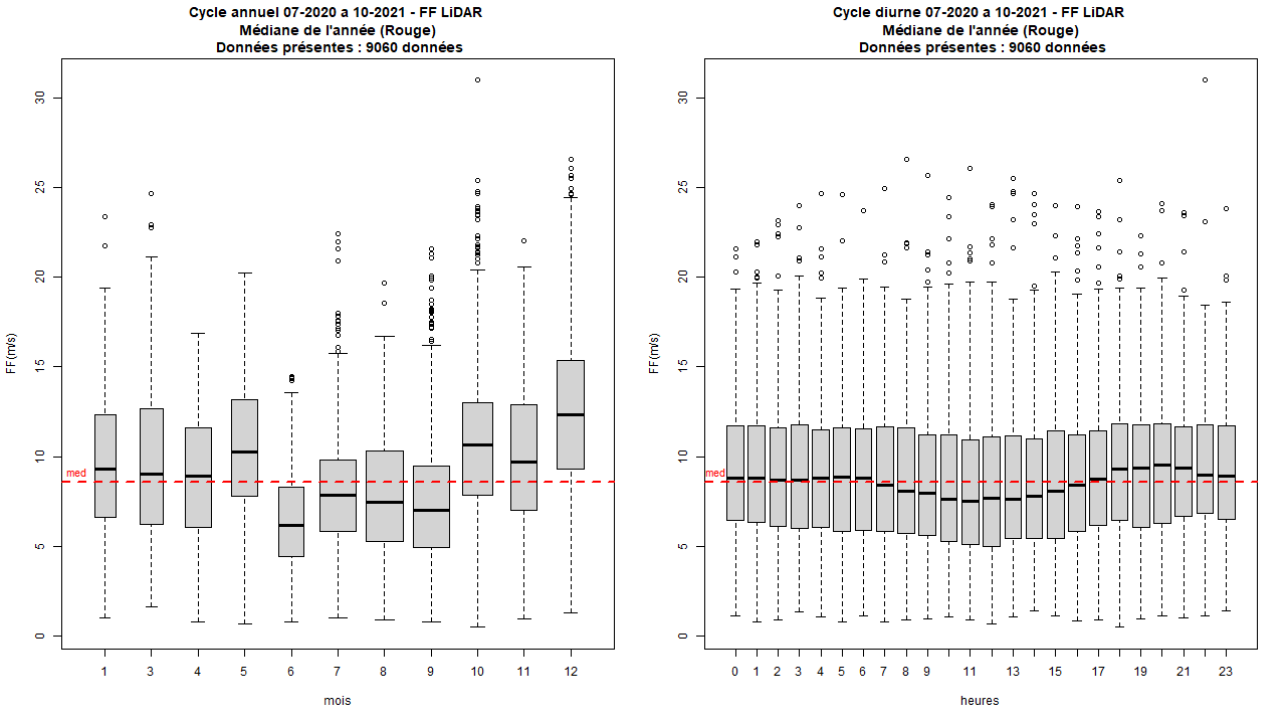

Illustration 3.3: AO5 Bretagne-Sud Bouée Nord – FF – Distribution du vent permettant de déterminer le cycle saisonnier (gauche) et le cycle diurne (droite)

Pour le cycle saisonnier, la distribution de FF le long des mois laisse apercevoir 3 groupements de mois :

- janvier, février, mars, avril et mai qui ont leurs médianes juste au-dessus de la médiane de l'année (c'est-à-dire celle de l'intégralité des données, en trait rouge sur les graphiques). Pour rappel, il n'y a pas eu de mesure au mois de février.
- juin, juillet, août et septembre qui ont leurs médianes en dessous de la médiane de l'année. Cependant, pour équilibrer le nombre de donnée au sein de chaque groupement, nous avons décidé de scinder ce groupement en deux sous-groupes : de juin à juillet, puis d'août à septembre.
- Et enfin octobre, novembre, décembre pour lesquels les vents sont globalement plus forts avec des valeurs extrêmes du fait des tempêtes hivernales.

Pour le cycle diurne, la distribution de FF fait ressortir trois catégories de groupement d'heures liés aux effets de brise :

- les heures qui ont leurs médianes très proches de celle de l'année : il s'agit notamment de 22H, 23H, puis de 00H jusqu'à 05H,
- les heures dont la médiane décroît progressivement par rapport à celles du premier groupement : il s'agit notamment de 06H jusqu'à 12H,
- et les heures dont la médiane croit progressivement par rapport au second groupement : il s'agit notamment de 13H jusqu'à 21H.

Ainsi, l'analyse de la distribution de FF et du taux de présence de données d'observations a permis de choisir 4 plages ([Janvier–Mai], [Juin–Juillet], [Août–Septembre], [Octobre–Décembre]) pour la variable RegMM2 représentant le cycle saisonnier, et 3 plages ([6h–12h], [13h–21h], [22h–5h]) pour la variable RegHH2_FF qui représente le cycle diurne.

3.1.3 Sélection des variables explicatives de la direction du vent

Comme pour la force du vent, le modèle statistique est ajusté sur l'écart de la direction du vent observé avec la direction du vent AROME. La direction du vent étant une variable circulaire, nous ne pouvons pas l'étudier directement. Nous devons donc élaborer un modèle statistique pour chacune des composantes U et V :

$$Y = U_{OBS} - U_{AROME} = UmUARO \quad \text{et} \quad Y = V_{OBS} - V_{AROME} = VmVARO \quad .$$

De la même façon que pour la force du vent, il s'agit d'examiner les liens entre la variable à expliquer Y et les variables explicatives X^i .

En suivant la même méthode que précédemment (paragraphe 3.1.2), nous avons choisi de garder moins de deux niveaux par paramètres (avec des coefficients de corrélations inférieurs à 0,96), les autres étant dans l'ACP.

L'ACP a permis de sélectionner 5 nouvelles variables explicatives (notées respectivement **PC1_DD, PC2_DD, PC3_DD, PC4_DD et PC5_DD**) combinant cette fois les variables UARO_10, UARO_250, VARO_10, VARO_250, TARO_2, TARO_50, TARO_250, TARO_500, TKEARO_10, TKEARO_50, TKEARO_160, SQRTTKEARO_10, SQRTTKEARO_50 et SQRTTKEARO_160.

L'illustration 3.4 présente la corrélation avec la variable à prédire et entre les variables qui ont été définitivement sélectionnées pour la mise en place de chaque modèle statistique en regroupant UmUARO et VmVARO sur le même graphique.

Ce graphique montre que les variables à expliquer $UmUARO$ et $VmVARO$ sont faiblement corrélées linéairement avec les **15 variables explicatives choisies** auxquelles s'ajoutent les variables liées aux cycles de DD qui sont détaillées ci-dessous.

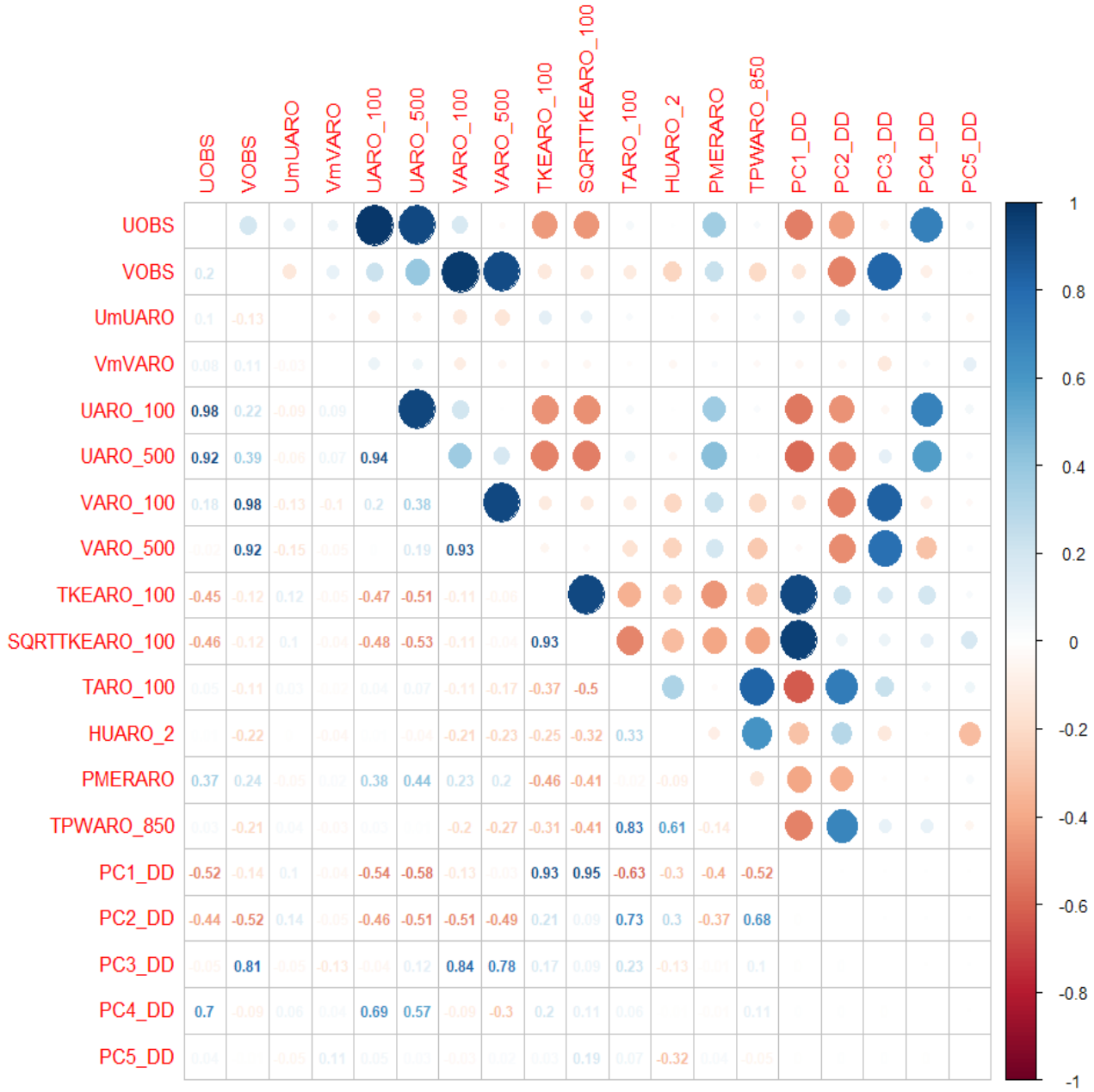

Illustration 3.4: AO5 Bretagne-Sud Bouée Nord – DD – Corrélation entre les variables à expliquer UmUARO et VmVARO et les variables explicatives. En haut : sous forme de bulles. En bas : chiffré (0 : absence de corrélation, 1 ou -1 : forte corrélation)

La dernière étape consiste en la définition des variables explicatives des cycles saisonnier et diurne pour la direction du vent. Comme pour FF, les cycles de DD (RegMM et RegHH_DD) sont déterminés à partir du taux de présence de données d'observation (tableau 3.2) et de la distribution du vent le long des mois pour le cycle saisonnier, et tout au long des heures pour le cycle diurne (illustration 3.5).

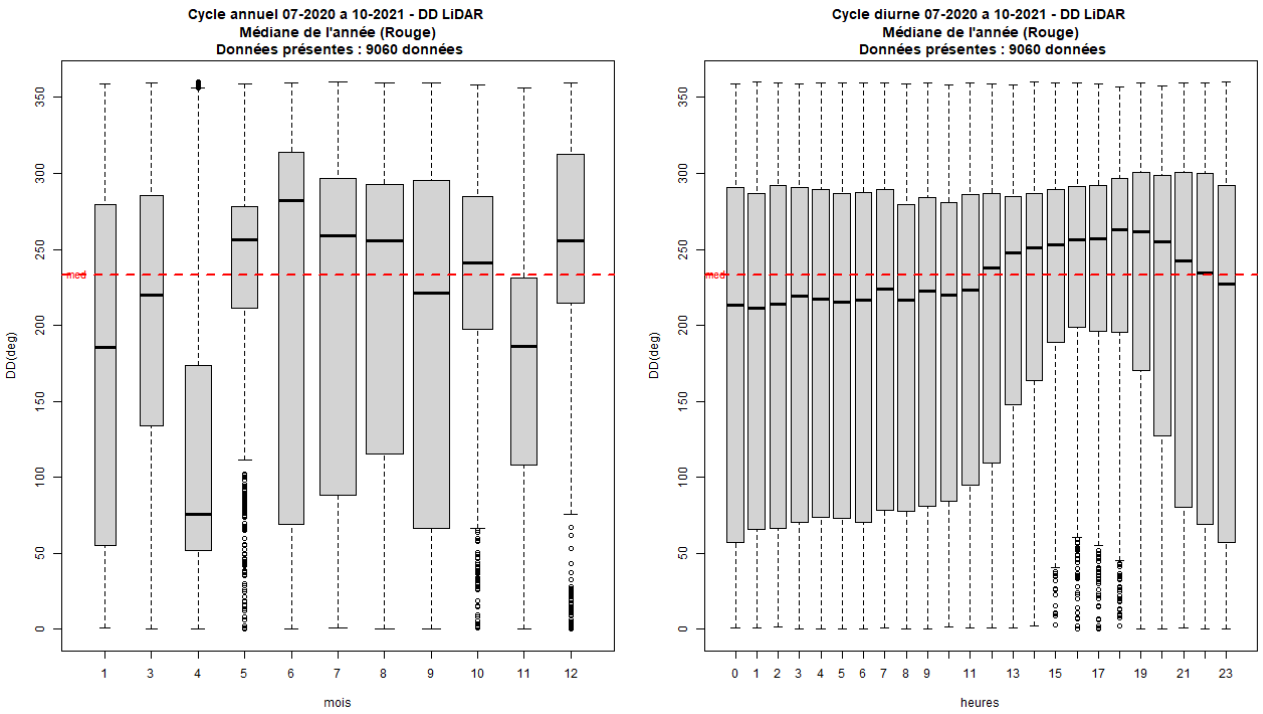

Illustration 3.5: AO5 Bretagne-Sud Bouée Nord – DD – Distribution du vent permettant de déterminer le cycle saisonnier (graphique à gauche) et cycle diurne (graphique à droite)

Pour le cycle saisonnier, la distribution de DD le long des mois ne permet pas de distinguer un cycle particulier. Nous avons donc essayé d'avoir un nombre équilibré de données pour chaque mois, ce qui nous a amené à regrouper les mois de janvier, février et mars en un seul mois virtuel, puis de considérer chaque mois comme un seul groupe.

Pour le cycle diurne, la distribution de DD fait ressortir deux catégories de groupement d'heures :

- les heures qui ont leurs médianes très proches ou en dessous de celle de l'année : il s'agit notamment de 21H, jusqu'à 12H,
- et les heures dont la médiane très au-dessus de celle de l'année : il s'agit notamment de 13H jusqu'à 20H.

Ainsi, l'analyse de la distribution de DD et de la présence de données à permis de construire la variable RegMM représentant le cycle saisonnier en mois de 10 facteurs (9 mois d'avril à décembre, et 1 mois virtuel composé de l'association des trois premiers mois de l'année, c'est-à-dire janvier, février et mars) ; et 2 plages (13h–20h], [21h–12h]) pour la variable RegHH_DD qui représente le cycle diurne.

3.1.4 Choix des échantillons de test et d'apprentissage

Le paragraphe 2.3.3 présente la manière dont sont construits les échantillons de test et d'apprentissage.

Dans notre étude, nous adaptions cette méthode pour construire des échantillons de test et d'apprentissage les plus représentatifs de la mesure du LIDAR Nord et des caractéristiques de la série de mesure.

Nous avons créé une variable année-mois noté AAAAMM avec comme donnée l'année et le mois correspondant à la date. De 07/2020 à 10/2021 la variable AAAAMM est une variable à 14 facteurs car nous avons regroupé les mois de septembre et novembre 2021 en un seul année-mois (fin de la campagne de mesure le 07/10/2021 à 23h TU).

Ainsi, pour chacune des 20 itérations de l'échantillonnage, 70 % des données correspondant à chaque année-mois est sélectionné aléatoirement pour constituer l'échantillon d'apprentissage, et les 30 % restant pour constituer l'échantillon de test.

3.2 Modélisation de la force du vent à 100 m

La force du vent FF a été estimée avec le modèle de forêt aléatoire.

On rappelle que la variable à expliquer est : $Y = FF_{OBS} - FF_{AROME} = FFmFFARO$.

Les variables explicatives qui ont été sélectionnées sont : RegMM2, RegHH2_FF, SECTEURARO_100, FFARO_100, FFARO_500, TARO_100, TKEARO_10, SQRTTKEARO_10, HUARO_2, PMERARO, TPWARO_850, PC1_FF, PC2_FF, PC3_FF, PC4_FF, PC5_FF.

Nous agissons sur le nombre d'arbres dans la forêt aléatoire pour trouver le nombre d'arbre optimal en termes de qualité d'ajustement sur l'échantillon d'apprentissage, tout en restant un modèle parcimonieux.

Les figures suivantes (illustrations 3.6, 3.7, 3.8 et 3.9) présentent les scores de validation croisée (apprentissage puis test) des modèles avec les différents niveaux d'arbres.

Les modèles RF_20, RF_50, RF_100, RF_200 et RF_500 correspondent respectivement aux modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres.

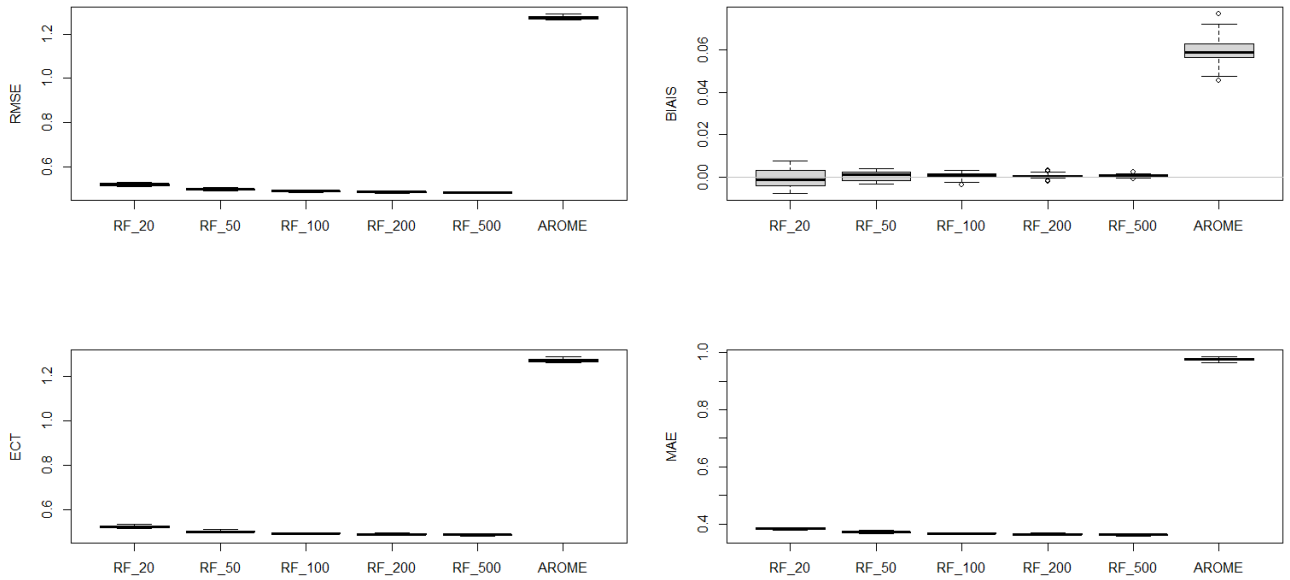

*Illustration 3.6: AO5 Bretagne-Sud Bouée Nord – FF – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'échantillon d'**apprentissage***

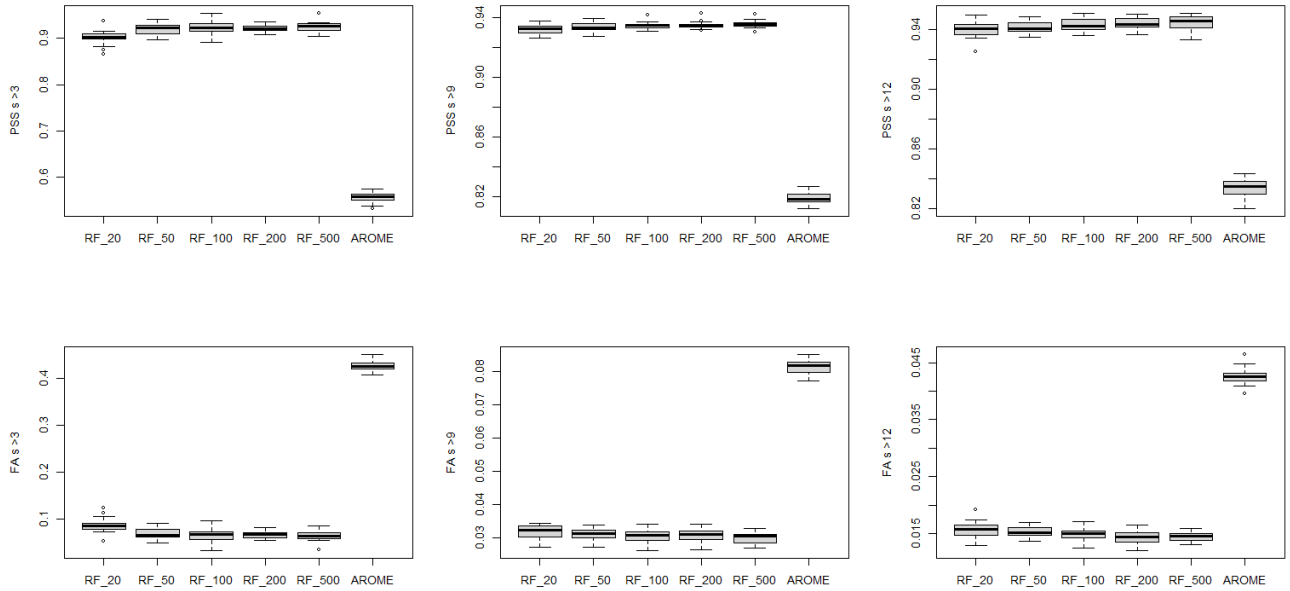

*Illustration 3.7: AO5 Bretagne-Sud Bouée Nord – FF – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot PSS (première ligne PSS FF > 3 m/s, PSS FF > 9 m/s et PSS FF > 12 m/s) et FA (deuxième ligne FA FF > 3 m/s, FA FF > 9 m/s et FA FF > 12 m/s) pour l'échantillon d'**apprentissage.***

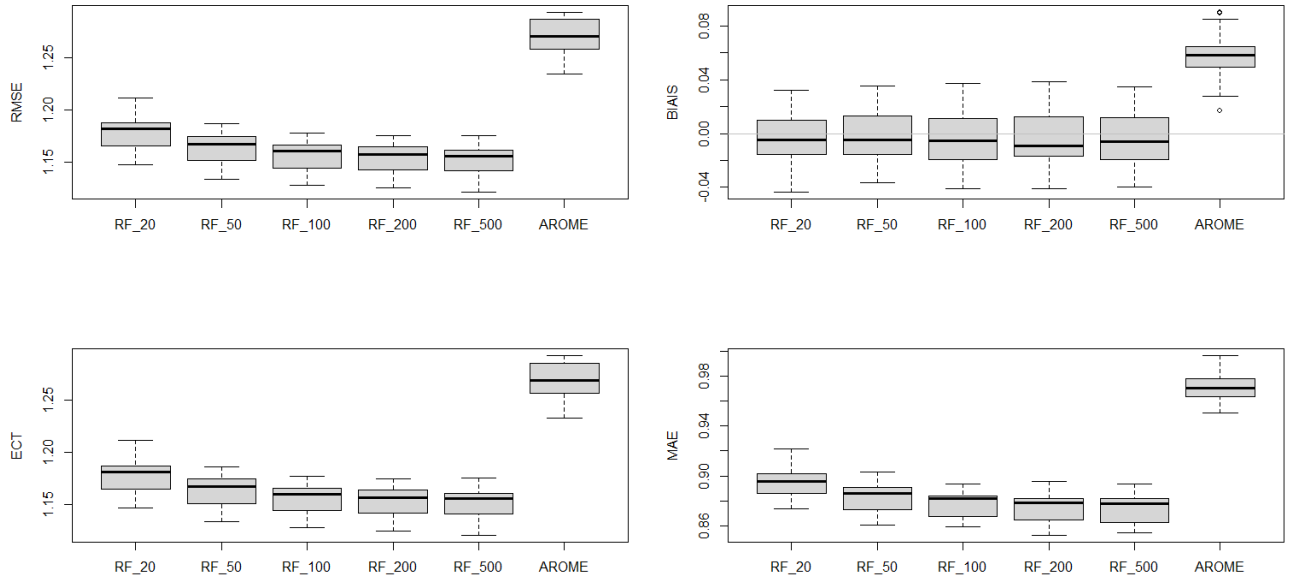

*Illustration 3.8: AO5 Bretagne-Sud Bouée Nord – FF – Modèles de forêt aléatoire - Box-plot RMSE, ECT, BIAIS et MAE pour l'échantillon de **test.***

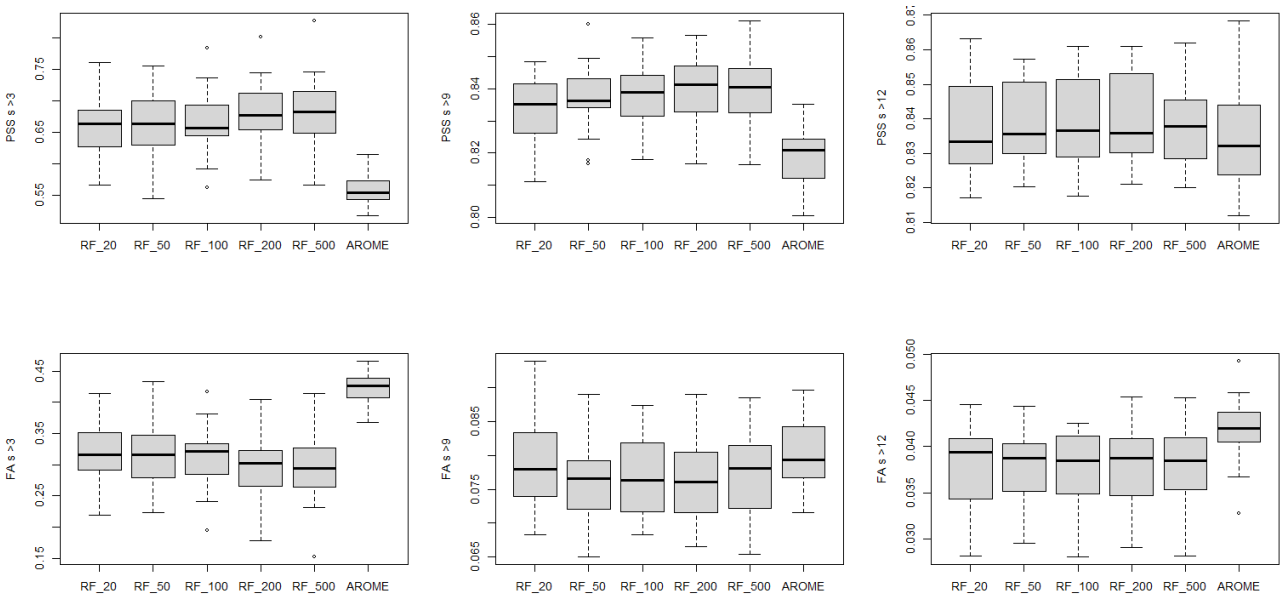

*Illustration 3.9: AO5 Bretagne-Sud Bouée Nord – FF – Modèles de forêt aléatoire - Box-plot PSS et FA pour l'échantillon de **test.***

Les tableaux 3.3 et 3.4 présentent le récapitulatif des scores de validation croisée (scores moyens des 20 échantillonnages) sur l'échantillon d'apprentissage puis de test.

*Tableau 3.3: AO5 Bretagne-Sud Bouée Nord – **FF** – Modèles de forêt aléatoire – Scores de validation croisée sur l'**échantillon d'apprentissage** (en vert, le modèle choisi)*

Modèle	RMSE	ECT	BIAIS	MAE	PSS FF>3 m/s	FA FF>3 m/s	PSS FF>9 m/s	FA FF>9 m/s	PSS FF>12m/s	FA FF>12m/s
RF_20	0.522	0.522	0	0.384	0.901	0.087	0.932	0.032	0.94	0.016
RF_50	0.501	0.501	0.001	0.371	0.92	0.069	0.933	0.031	0.942	0.015
RF_100	0.492	0.492	0.001	0.366	0.923	0.066	0.935	0.03	0.943	0.015
RF_200	**0.488**	**0.488**	**0.001**	**0.364**	**0.922**	**0.066**	**0.935**	**0.031**	**0.944**	**0.014**
RF_500	0.485	0.485	0.001	0.362	0.924	0.064	0.936	0.03	0.945	0.014
AROME	1.274	1.272	0.059	0.977	0.556	0.427	0.819	0.081	0.834	0.043

*Tableau 3.4: AO5 Bretagne-Sud Bouée Nord – **FF** – Modèles de forêt aléatoire – Scores de validation croisée sur l'**échantillon de test** (en vert, le modèle choisi)*

Modèle	RMSE	ECT	BIAIS	MAE	PSS FF>3 m/s	FA FF>3 m/s	PSS FF>9 m/s	FA FF>9 m/s	PSS FF>12m/s	FA FF>12m/s
RF_20	1.178	1.178	-0.005	0.895	0.659	0.319	0.834	0.078	0.837	0.038
RF_50	1.163	1.163	-0.002	0.882	0.663	0.315	0.837	0.076	0.839	0.038
RF_100	1.156	1.156	-0.003	0.877	0.668	0.31	0.837	0.077	0.838	0.038
RF_200	**1.153**	**1.153**	**-0.003**	**0.875**	**0.681**	**0.297**	**0.839**	**0.076**	**0.839**	**0.038**
RF_500	1.152	1.152	-0.003	0.874	0.682	0.295	0.839	0.077	0.839	0.038
AROME	1.27	1.268	0.057	0.971	0.56	0.422	0.818	0.08	0.834	0.042

Les scores des modèles RF_100, RF_200 et RF_500 sont très proches (aussi bien pour l'apprentissage que pour le test). Ils sont meilleurs que ceux des 2 autres modèles (RF_20 et RF_50 qui ont un nombre inférieur d'arbre). Cependant, nous avons constaté que l'évolution des scores est très minime autour du modèle de forêt à 200 arbres, RF_200. **Nous avons donc décidé de choisir le modèle RF_200 pour la réalisation de l'estimation de la force du vent (reconstitution de FF avec RF_200)**.

Pour ce modèle, l'importance des variables est donnée par le tableau 3.5. C'est la moyenne de l'importance des variables pour chacun des 20 échantillonnages décrites dans la section 2.3.2.

Tableau 3.5: AO5 Bretagne-Sud bouée Nord – ***FF –***
Importance des variables explicatives *pour le modèle de forêt aléatoire avec 200 arbres*

Variable	RegMM 2	RegHH 2_FF	TKEARO_10	SQRTTKEARO_10	PC2_FF	PC1_FF	TPWARO_850	PC4_FF	FFARO_100	TARO_100	FFARO_500	PC3_FF	HUARO_2	PMERARO	SECTEURARO_100	PC5_FF
Importance	174.54	179.73	539.75	542.45	584.07	594.77	601.13	603.38	606.8	616.17	633.5	659.26	676.53	807.76	934.44	1000.26

La composante principale 5 de l'ACP, la variable SECTEURARO_100 ainsi que PMERARO ont une importance plus élevée que les autres variables explicatives pour ce modèle. À l'exception des deux variables de prise en compte des cycles annuel et diurne, le reste des variables ont un poids similaire et important pour le modèle de forêt aléatoire.

On note déjà sur les scores de qualité de ces échantillons, l'apport de la forêt aléatoire par rapport à AROME. Les scores de corrélation sur l'échantillon de test concaténé vont également dans ce sens :

- pour FF issue de RF_200, la corrélation sur l'échantillon de test concaténé est de 0.959,
- pour FF issue de AROME, la corrélation sur l'échantillon de test concaténé est de 0.951.

3.3 Modélisation de la direction du vent à 100 m

La méthodologie reste la même que celle employée précédemment pour la force du vent à 100 m.

La différence majeure se situe dans la caractéristique de la variable Y à expliquer (DD à 100 m), qui est une variable circulaire (modulo 360°). Pour tenir compte de cette caractéristique, la régression est effectuée selon les deux composantes du vent : zonale U et méridienne V puis la direction du vent DD est estimée à partir des estimations de U et V.

Le passage de (DD, FF) à (U, V) s'effectue de la façon suivante :

$$U = \sin\left(DD \times \frac{\pi}{180}\right) \times FF$$
$$V = \cos\left(DD \times \frac{\pi}{180}\right) \times FF$$

3.3.1 La modélisation de U et V à 100 m

Chacune des composantes U et V du vent a été estimée avec le modèle de forêt aléatoire.

Les variables à expliquer sont : $Y_1 = U_{OBS} - U_{AROME} = UmUARO$ et $Y_2 = V_{OBS} - V_{AROME} = VmVARO$.

Les variables explicatives choisies sont : RegMM, RegHH_DD, UARO_100, UARO_500, VARO_100, VARO_500, TKEARO_100, SQRTTKEARO_100, TARO_100, HUARO_2, PMERARO, TPWARO_850, PC1_DD, PC2_DD, PC3_DD, PC4_DD, PC5_DD.

Comme pour FF, les scores RMSE, ECT, MAE, BIAIS, PSS et FA ont permis de sélectionner le modèle de **forêt aléatoire avec 200 arbres** (RF_200) comme meilleur modèle statistique pour la prédiction de U et de V, comme le montrent les graphiques et tableaux de ces scores : illustrations 3.10 à 3.13, tableaux 3.6 et 3.7 pour la composante U et illustrations 3.14 à 3.17, tableaux 3.8 et 3.9 pour la composante V.

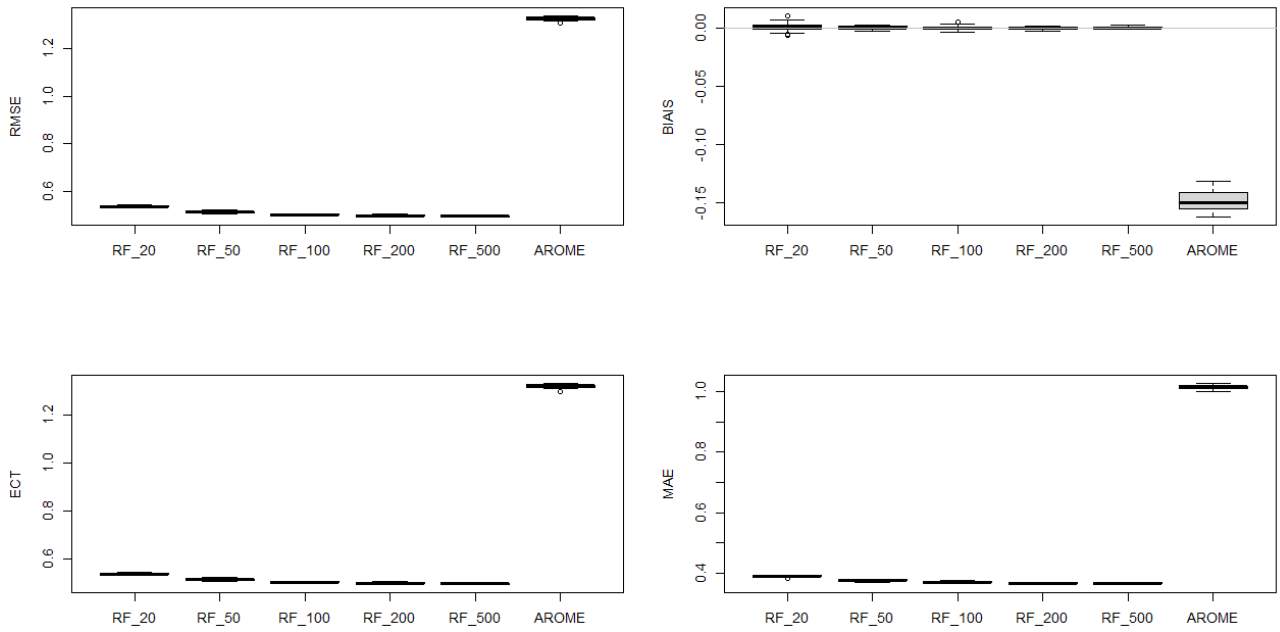

*Illustration 3.10: AO5 Bretagne-Sud Bouée Nord – **U** – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'**échantillon d'apprentissage***

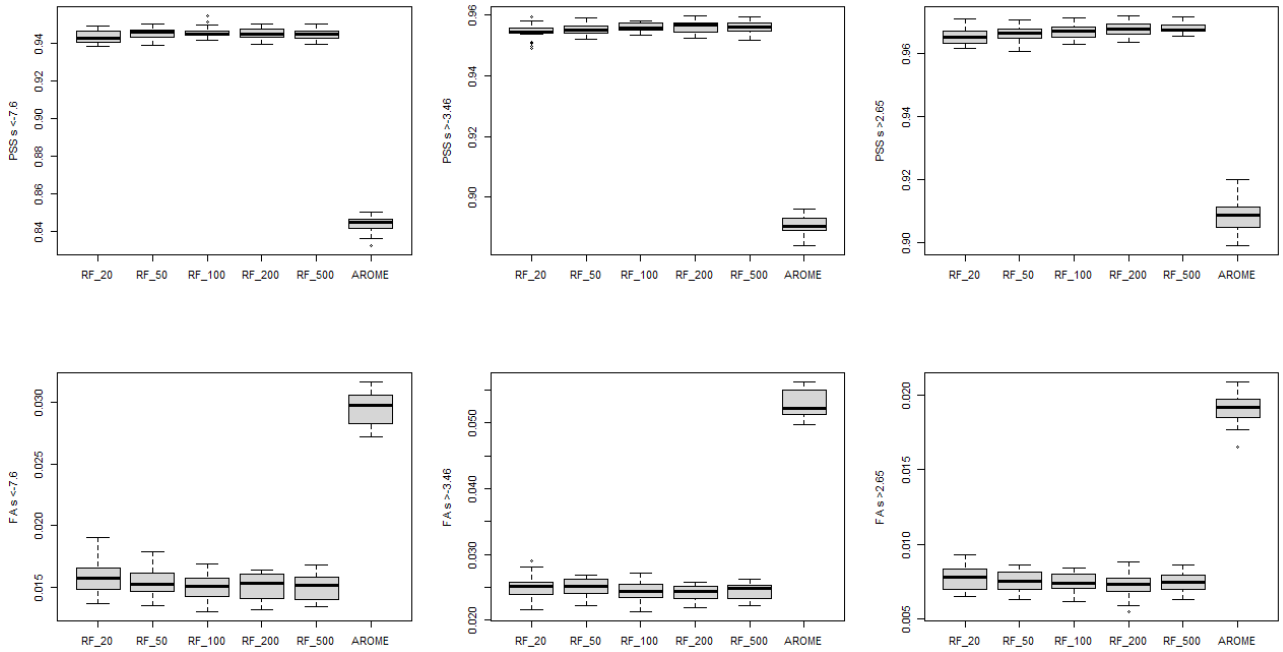

*Illustration 3.11: AO5 Bretagne-Sud Bouée Nord – **U** – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot PSS (première ligne PSS U < -7.6 m/s, PSS U > -3.46 m/s et PSS U > 2.65 m/s) et FA (deuxième ligne FA U < -7.6 m/s, FA U > -3.46 m/s et FA U > 2.65 m/s) pour l'**échantillon d'apprentissage.***

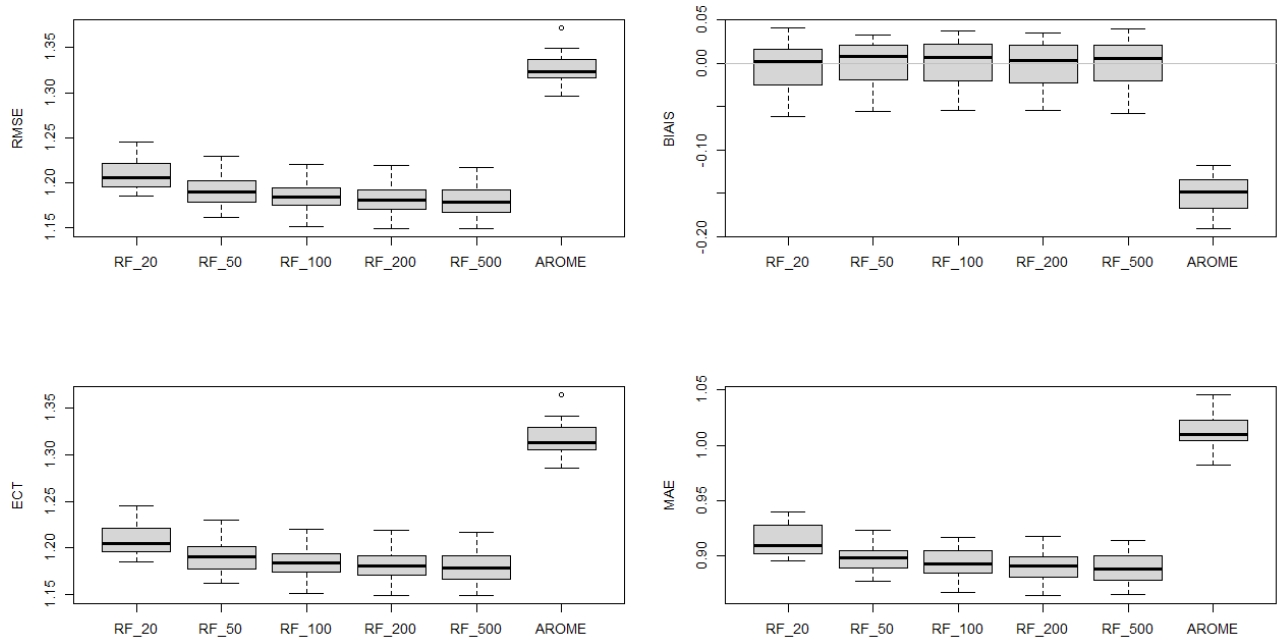

*Illustration 3.12: AO5 Bretagne-Sud Bouée Nord – **U** – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'**échantillon de test.***

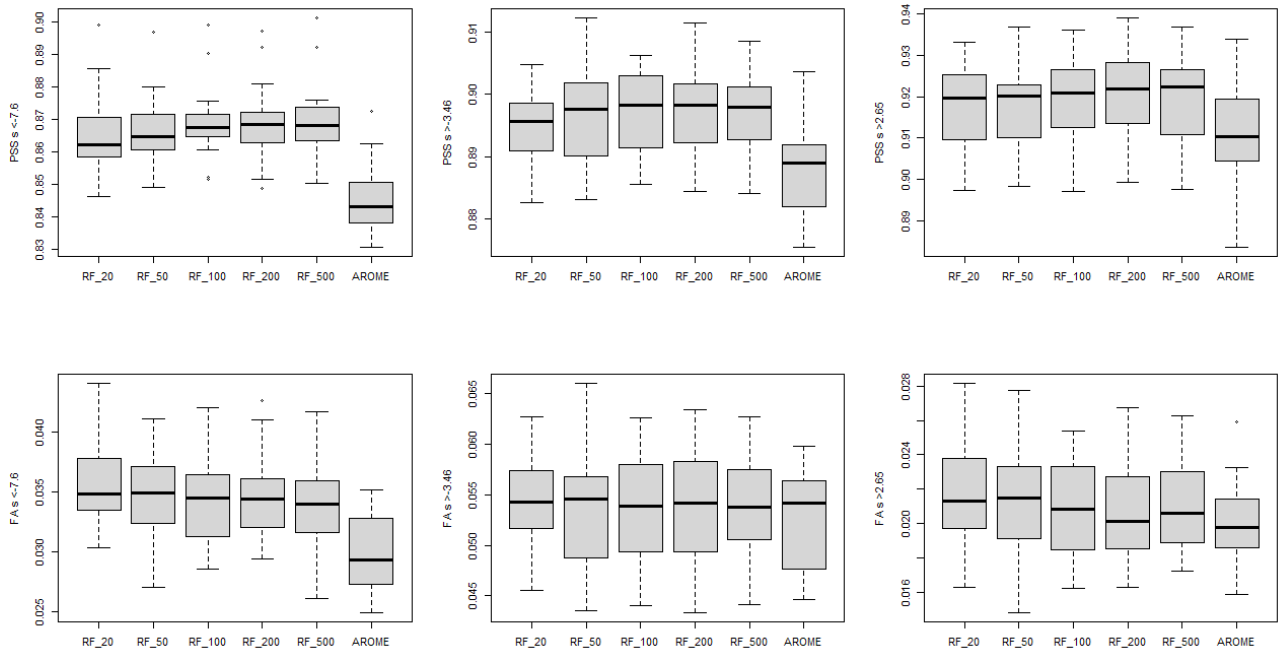

*Illustration 3.13: AO5 Bretagne-Sud Bouée Nord – **U** – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot PSS (première ligne PSS U < -7.6 m/s, PSS U > -3.46 m/s et PSS U > 2.65 m/s) et FA (deuxième ligne FA U < -7.6 m/s, FA U > -3.46 m/s et FA U > 2.65 m/s) pour l'**échantillon de test.***

*Tableau 3.6: AO5 Bretagne-Sud Bouée Nord – **U** – Modèles de forêt aléatoire – Scores de validation croisée sur l'**échantillon d'apprentissage** (en vert, le modèle choisi)*

Modèle	RMSE	ECT	BIAIS	MAE	PSS U < -7.6 m/s	FA U < -7.6 m/s	PSS U > -3.46 m/s	FA U > -3.46 m/s	PSS U > 2.65 m/s	FA U > 2.65 m/s
RF_20	0.537	0.537	0.001	0.39	0.943	0.016	0.954	0.025	0.965	0.008
RF_50	0.511	0.511	0.001	0.376	0.945	0.015	0.955	0.025	0.966	0.008
RF_100	0.502	0.502	0.001	0.37	0.946	0.015	0.956	0.024	0.967	0.007

RF_200	**0.498**	**0.498**	**0**	**0.368**	**0.945**	**0.015**	**0.956**	**0.024**	**0.968**	**0.007**
RF_500	0.496	0.496	0.001	0.367	0.945	0.015	0.956	0.024	0.968	0.007
AROME	1.327	1.319	-0.148	1.014	0.844	0.029	0.89	0.053	0.909	0.019

*Tableau 3.7: AO5 Bretagne-Sud Bouée Nord – **U** – Modèles de forêt aléatoire – Scores de validation croisée sur l'**échantillon de test** (en vert, le modèle choisi)*

Modèle	RMSE	ECT	BIAIS	MAE	PSS U < -7.6 m/s	FA U < -7.6 m/s	PSS U > -3.46 m/s	FA U > -3.46 m/s	PSS U > 2.65 m/s	FA U > 2.65 m/s
RF_20	1.209	1.208	-0.003	0.913	0.865	0.036	0.895	0.054	0.918	0.022
RF_50	1.191	1.19	-0.001	0.898	0.867	0.035	0.897	0.053	0.918	0.021
RF_100	1.185	1.185	-0.001	0.894	0.869	0.034	0.897	0.053	0.92	0.021
RF_200	**1.182**	**1.182**	**-0.002**	**0.891**	**0.869**	**0.035**	**0.897**	**0.053**	**0.921**	**0.021**
RF_500	1.18	1.18	-0.001	0.889	0.87	0.034	0.897	0.054	0.92	0.021
AROME	1.327	1.318	-0.151	1.014	0.845	0.03	0.888	0.053	0.911	0.02

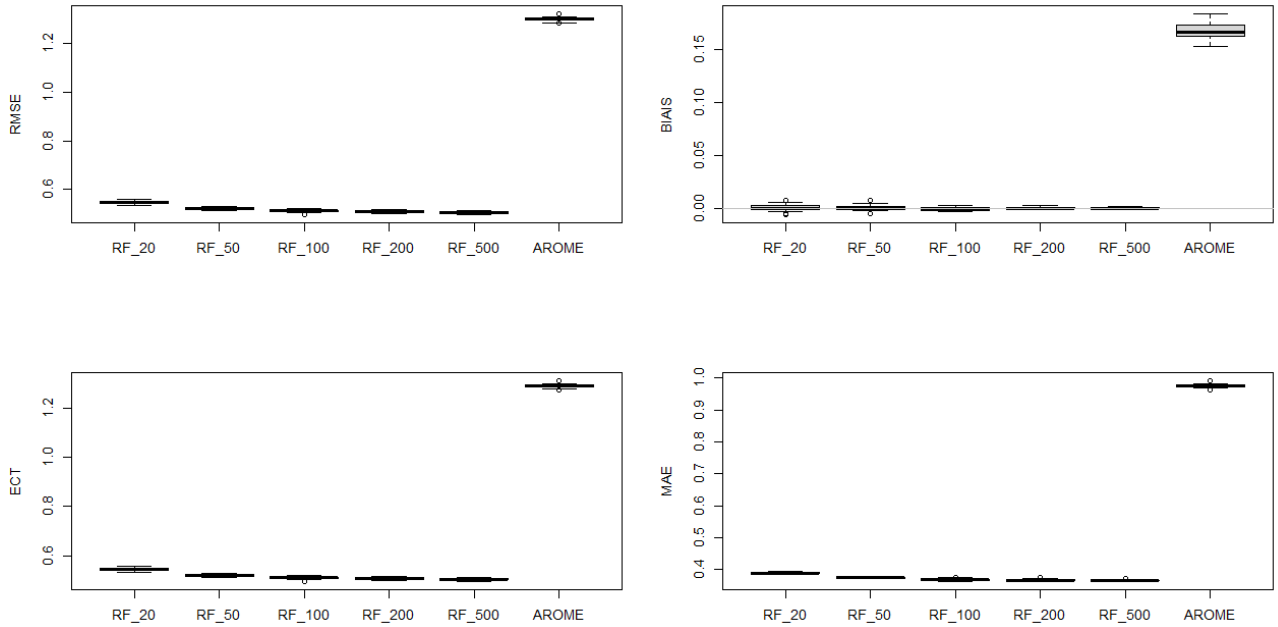

*Illustration 3.14: AO5 Bretagne-Sud Bouée Nord – **V** – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'**échantillon d'apprentissage.***

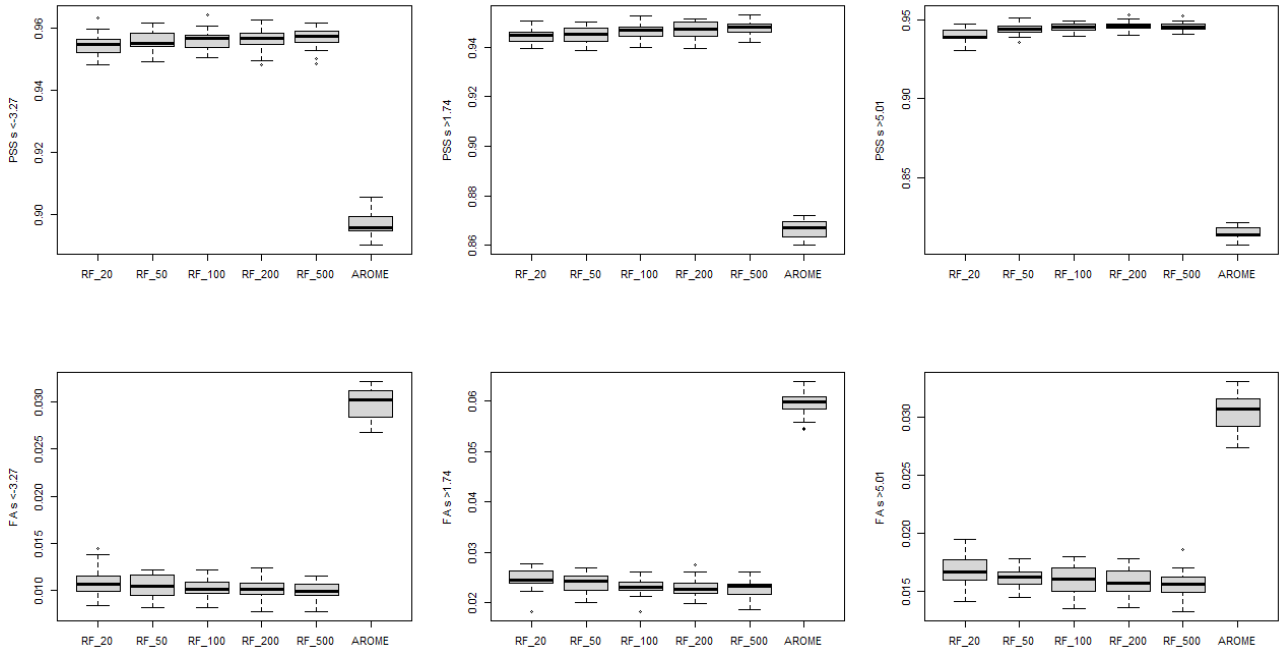

*Illustration 3.15: AO5 Bretagne-Sud Bouée Nord – **V** – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot PSS (première ligne PSS V < -3.27 m/s, PSS V > 1.74 m/s et PSS V > 5.01 m/s) et FA (deuxième ligne FA V < -3.27 m/s, FA V > 1.74 m/s et FA V > 5.01 m/s) pour l'**échantillon d'apprentissage.***

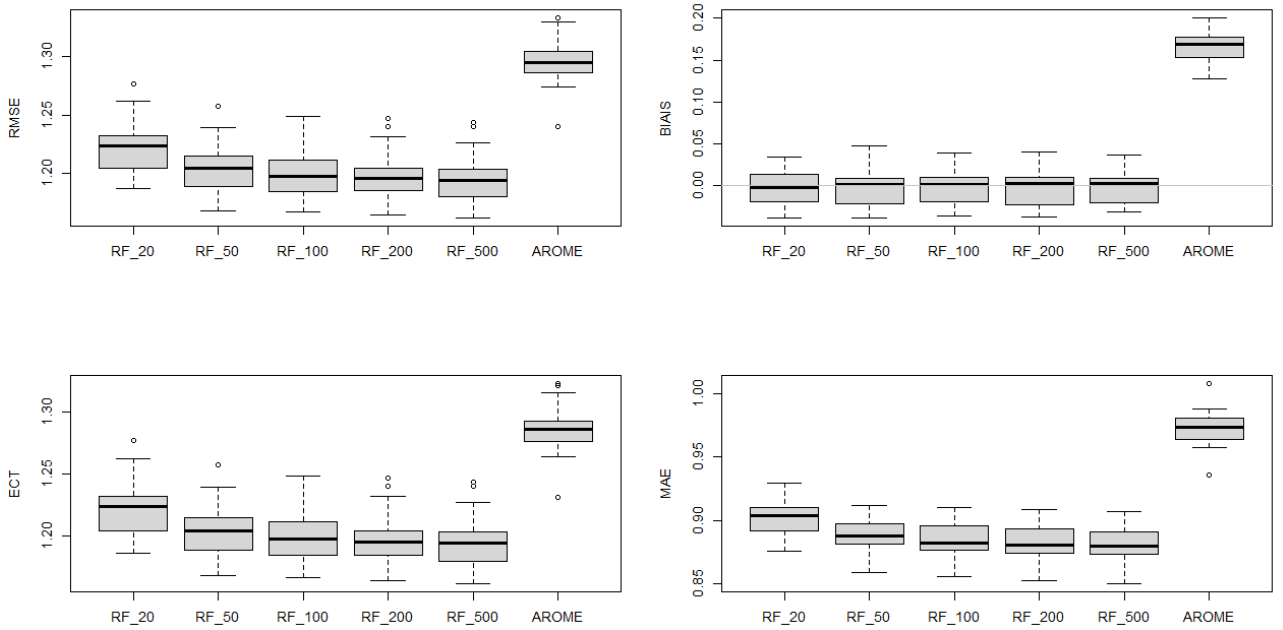

*Illustration 3.16: AO5 Bretagne-Sud Bouée Nord – **V** – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'**échantillon de test.***

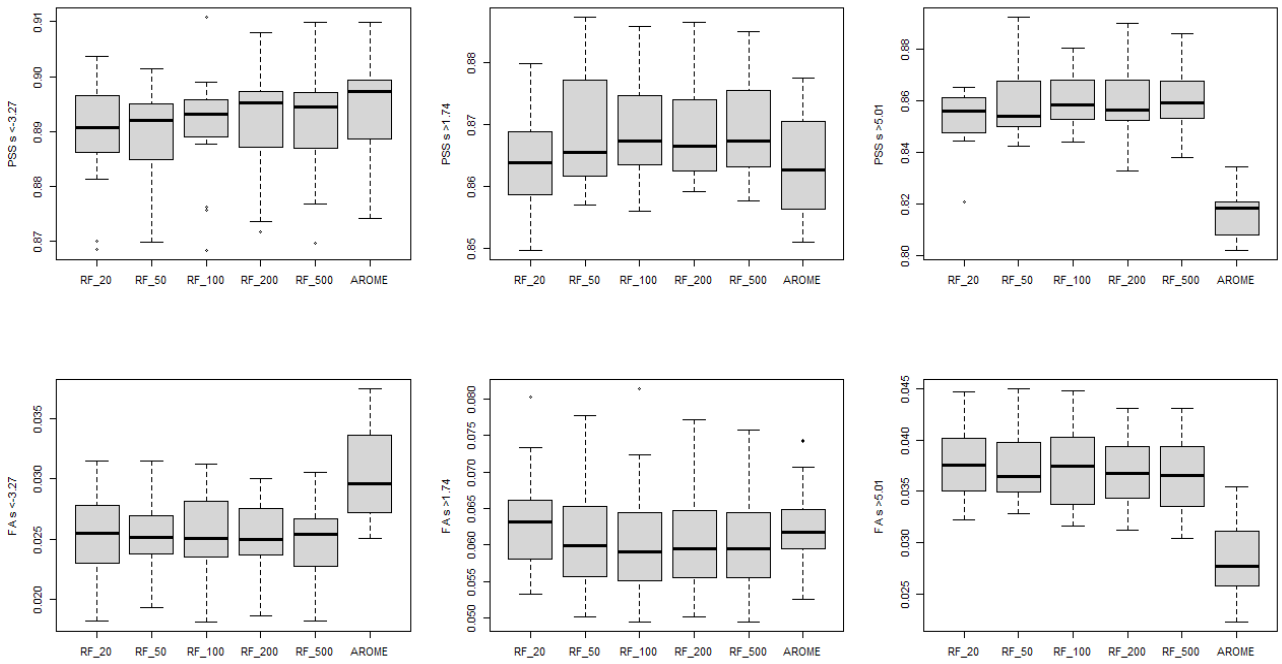

*Illustration 3.17: AO5 Bretagne-Sud Bouée Nord – **V** – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot PSS (première ligne PSS V < -3.27 m/s, PSS V > 1.74 m/s et PSS V > 5.01 m/s) et FA (deuxième ligne FA V < -3.27 m/s, FA V > 1.74 m/s et FA V > 5.01 m/s) pour l'**échantillon de test.***

*Tableau 3.8: AO5 Bretagne-Sud Bouée Nord – **V** – Modèles de forêt aléatoire – Scores de validation croisée sur l'**échantillon d'apprentissage** (en vert, le modèle choisi)*

Modèle	RMSE	ECT	BIAIS	MAE	PSS V < -3,27 m/s	FA V < -3,27 m/s	PSS V > 1.74 m/s	FA V > 1.74 m/s	PSS V > 5.01 m/s	FA V > 5.01 m/s
RF_20	0.546	0.546	0.001	0.389	0.955	0.011	0.944	0.025	0.94	0.017
RF_50	0.521	0.521	0.001	0.375	0.956	0.01	0.945	0.024	0.944	0.016
RF_100	0.512	0.512	0	0.369	0.956	0.01	0.946	0.023	0.945	0.016
RF_200	**0.508**	**0.508**	**0.001**	**0.367**	**0.956**	**0.01**	**0.947**	**0.023**	**0.946**	**0.016**
RF_500	0.505	0.505	0.001	0.366	0.957	0.01	0.948	0.023	0.946	0.016
AROME	1.299	1.288	0.167	0.976	0.897	0.03	0.867	0.059	0.815	0.03

*Tableau 3.9: AO5 Bretagne-Sud Bouée Nord – **V** – Modèles de forêt aléatoire – Scores de validation croisée sur l'**échantillon de test** (en vert, le modèle choisi)*

Modèle	RMSE	ECT	BIAIS	MAE	PSS V < -3,27 m/s	FA V < -3,27 m/s	PSS V > 1.74 m/s	FA V > 1.74 m/s	PSS V > 5.01 m/s	FA V > 5.01 m/s
RF_20	1.222	1.222	-0.004	0.903	0.89	0.025	0.864	0.063	0.853	0.038
RF_50	1.205	1.204	-0.003	0.889	0.89	0.025	0.869	0.061	0.858	0.037
RF_100	1.2	1.2	-0.002	0.885	0.891	0.026	0.869	0.061	0.86	0.037
RF_200	**1.198**	**1.197**	**-0.003**	**0.883**	**0.892**	**0.025**	**0.869**	**0.061**	**0.859**	**0.037**
RF_500	1.196	1.195	-0.002	0.882	0.892	0.025	0.869	0.061	0.86	0.037
AROME	1.296	1.285	0.167	0.972	0.895	0.03	0.863	0.062	0.816	0.028

Pour le **modèle choisi comme optimal (RF_200) pour U et V**, l'importance des variables est donnée respectivement par les tableaux 3.10 et 3.11.

Tableau 3.10: AO5 Bretagne-Sud bouée Nord ***– U –***
Importance des variables explicatives *pour le modèle de forêt aléatoire avec 200 arbres*

Variable	RegHH_DD	RegMM	TKEARO_100	SQRTTKEARO_100	TPWARO_850	PC1_DD	TARO_100	PC3_DD	PC4_DD	HUARO_2	PC5_DD	VARO_100	PC2_DD	PMERARO	VARO_500	UARO_500	UARO_100
Importance	87.4	469.65	495.59	496.96	593.31	605.27	622.26	627.1	648.21	654.61	657.54	679.36	737.16	746.17	757.65	797.09	822.98

Tableau 3.11: AO5 Bretagne-Sud bouée Nord ***– V –***
Importance des variables explicatives *pour le modèle de forêt aléatoire avec 200 arbres*

Variable	RegHH_DD	RegMM	SQRTTKEARO_100	TKEARO_100	PC1_DD	PC2_DD	TARO_100	TPWARO_850	PC4_DD	UARO_100	PC3_DD	UARO_500	HUARO_2	VARO_500	PC5_DD	PMERARO	VARO_100
Importance	80.75	430.27	473.15	477.71	535.49	550.06	579.25	583.69	621.33	647.75	659.97	660.09	664.54	686.75	751.54	757.77	794.43

Pour U et V, à l'exception de la variable de prise en compte du cycle diurne (RegHH_DD), toutes les variables ont un poids important pour le modèle de forêt aléatoire.

Cependant :

- pour U, les variables UARO_100, UARO_500, VARO_500, PMERARO et PC2_DD ont une importance relativement plus élevée que les autres variables explicatives,
- pour V, ce sont les variables VARO_100, PMERARO et PC5_DD qui s'illustrent avec une importance relativement plus élevée que les autres variables.

Ainsi, on retrouve la variable PMERARO avec une grande importance pour chacune des composantes U et V du vent.

3.3.2 Reconstitution de DD à 100 m

La direction du vent est reconstituée à partir des estimations des composantes U et V :

$$DD = \left(90 - \frac{180}{\pi} \times \arctan 2\left(V, U\right)\right) \% \left(360\right) \quad (1)$$

Les scores du paragraphe précédent ont permis de sélectionner le modèle RF_200 comme modèle optimal pour ces deux composants. Pour choisir la meilleure reconstitution de la direction du vent, nous établissons DD avec une combinaison de U et V issues du modèle RF_200 et AROME à l'aide de la formule (1). L'ensemble des scores de ce paragraphe est calculé sur l'échantillon de test concaténé.

Pour cela, nous analysons les RMSE de DD par classe de direction (18 secteurs de 20 degrés) représentés dans le tableau 3.12.

Tableau 3.12: AO5 Bretagne-Sud Bouée Nord – DD –
Scores ***RMSE de DD (en °) par secteur de direction*** *sur l'****échantillon de test concaténé*** *(en vert, le modèle choisi)*

SECTEUR	SECT 00	SECT 20	SECT 40	SECT 60	SECT 80	SECT 100	SECT 120	SECT 140	SECT 160	SECT 180	SECT 200	SECT 220	SECT 240	SECT 260	SECT 280	SECT 300	SECT 320	SECT 340
Intervalle	[350°, 10°[	[10°, 30°[	[30°, 50°[	[50°, 70°[	[70°, 90°[	[90°, 110°[	[110°, 130°[	[130°, 150[	[150°, 170°[	[170°, 190°[	[190°, 210[	[210°, 230°[	[203°, 250°[	[250°, 270°[	[270°, 290°[	[290°, 310°[	[310°, 330°[	[330°, 350°[
NB DATA	2701	2148	3188	3921	3058	1399	1090	817	964	1432	3089	3790	4122	3999	5921	5623	4384	2594
RF_200	10.49	10.31	7.13	**4.37**	**2.45**	**2.61**	**4.86**	**8.8**	**10.74**	9.98	**8.14**	**6.36**	**4.45**	2.15	2.03	**4.15**	**6.76**	9.47
AROME	10.68	**9.85**	7.55	5.05	2.63	2.98	5.92	9.74	11.18	10.57	9.28	7.27	4.5	**1.94**	2.09	4.93	7.7	10.37
U.RF_200 V.AROME	**10.1**	10.34	7.54	4.8	2.51	2.9	5.34	8.94	**10.74**	**9.66**	8.33	6.67	4.48	2.08	2.16	4.61	6.95	**9.45**
U.AROME V.RF_200	11.02	9.86	**7.1**	4.6	2.54	2.66	5.4	9.69	11.21	11	9.01	6.97	4.5	2	**2**	4.55	7.51	10.41

Ces scores sont majoritairement en faveur de DD reconstituée avec U et V du modèle de RF_200. Ce modèle a les meilleurs RMSE pour 11 des 18 secteurs de direction, dont la quasi-totalité des secteurs dominant (secteurs avec des valeurs élevées sur la ligne NB DATA du tableau). Le modèle dont la direction est issue de U RF_200 et V AROME est celui qui se comporte le mieux après DD issue de RF_200 pour U et V d'après ces scores de RMSE par secteur.

Cette analyse concorde bien avec les scores de corrélation, sur l'échantillon de test concaténé, qui met en avant le modèle RF_200 et le modèle U.RF_200 V.AROME avec un score quasiment identique. Les scores de corrélation sont en effet les suivant :

- 0.901 pour DD issue de RF_200,
- 0.896 pour DD issue de AROME,
- 0.902 pour DD issue de U.RF_200 V.ARO,
- 0.896 pour DD issue de U.ARO V.RF_200.

Enfin nous avons visualisé les QQ-Plots des différentes combinaisons de U et V pour la reconstitution de DD (illustration 3.18). Ils sont tous très proches les uns des autres, et ne permettent pas de démarquer de manière évidente un modèle parmi les quatre. **Par conséquent, nous avons décidé que la direction du vent sera reconstituée à partir des paramètres U et V étendue à l'aide du modèle RF_200.**

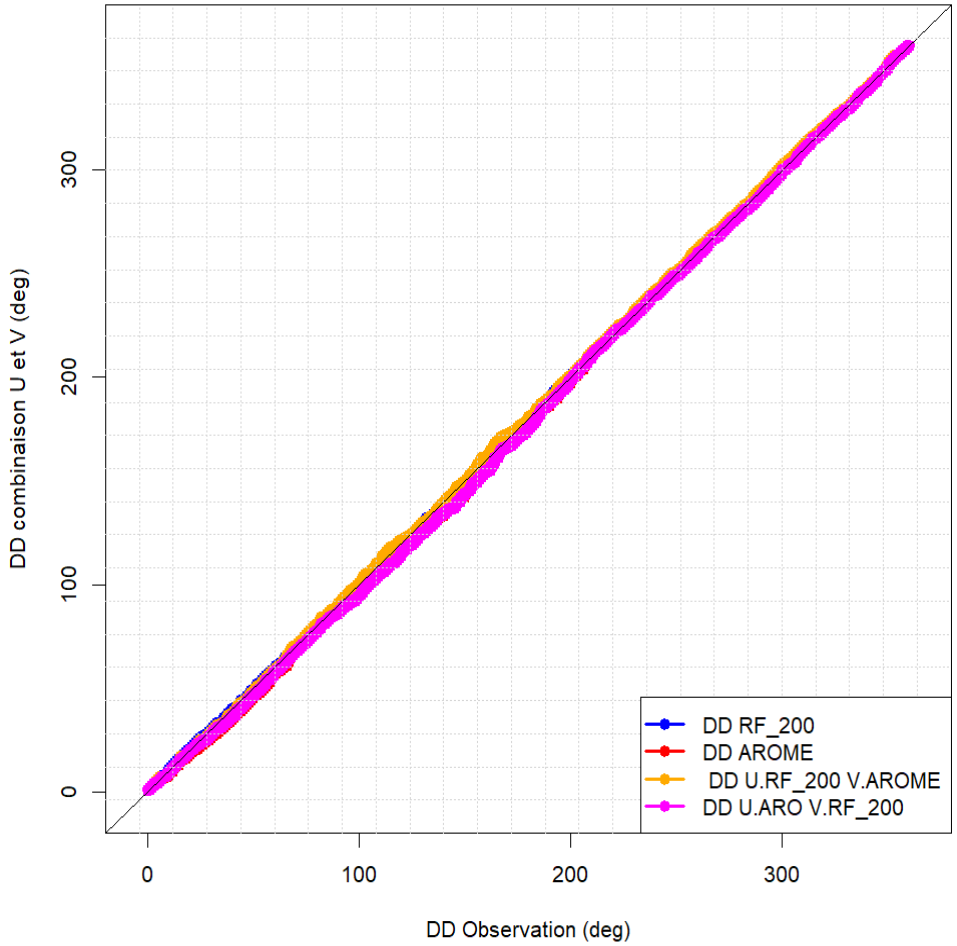

Illustration 3.18: AO5 Bretagne-Sud Bouée Nord – DD –
*QQ-Plots de la **reconstitution de DD** avec combinaison de U et V sur l'**échantillon de test concaténé.***

4 Principaux résultats sur l'extension de la série temporelle d'observations horaires

Les résultats de modélisation de FF et DD ont permis de choisir le modèle de forêt aléatoire à 200 arbres (RF_200) pour étendre les séries temporelles d'observations horaires.

Pour étendre les séries, nous réalisons l'apprentissage avec RF_200 sur l'intégralité des données de la période d'observation. Une fois les paramètres du modèle estimés à travers l'apprentissage, on les applique aux prédicteurs sélectionnés sur la période de reconstitution des séries horaires (du 01/01/2000 00H au 07/10/2020 23H) : c'est l'extension statistique.

Nous prolongeons l'extension statistique sur la période d'observation afin d'évaluer sa capacité de restitution des cycles diurne et saisonnier, et des roses des vents par rapport aux observations. Ensuite, nous examinons deux scores complémentaires :

- un score sur les RMSE de FF (en m/s) par secteur de direction sur la période d'observation
- et un score sur les données de puissances électriques sur la période d'observation

afin de qualifier davantage les modèles (extension statistique et AROME) en vue de la livraison.

4.1 Restitution des cycles

Les illustrations 4.1 et 4.2 présentent respectivement la restitution des cycles diurne et annuel/saisonnier de l'extension statistique pour FF et DD. Une comparaison avec l'observation et AROME est réalisée sur la période d'observation.

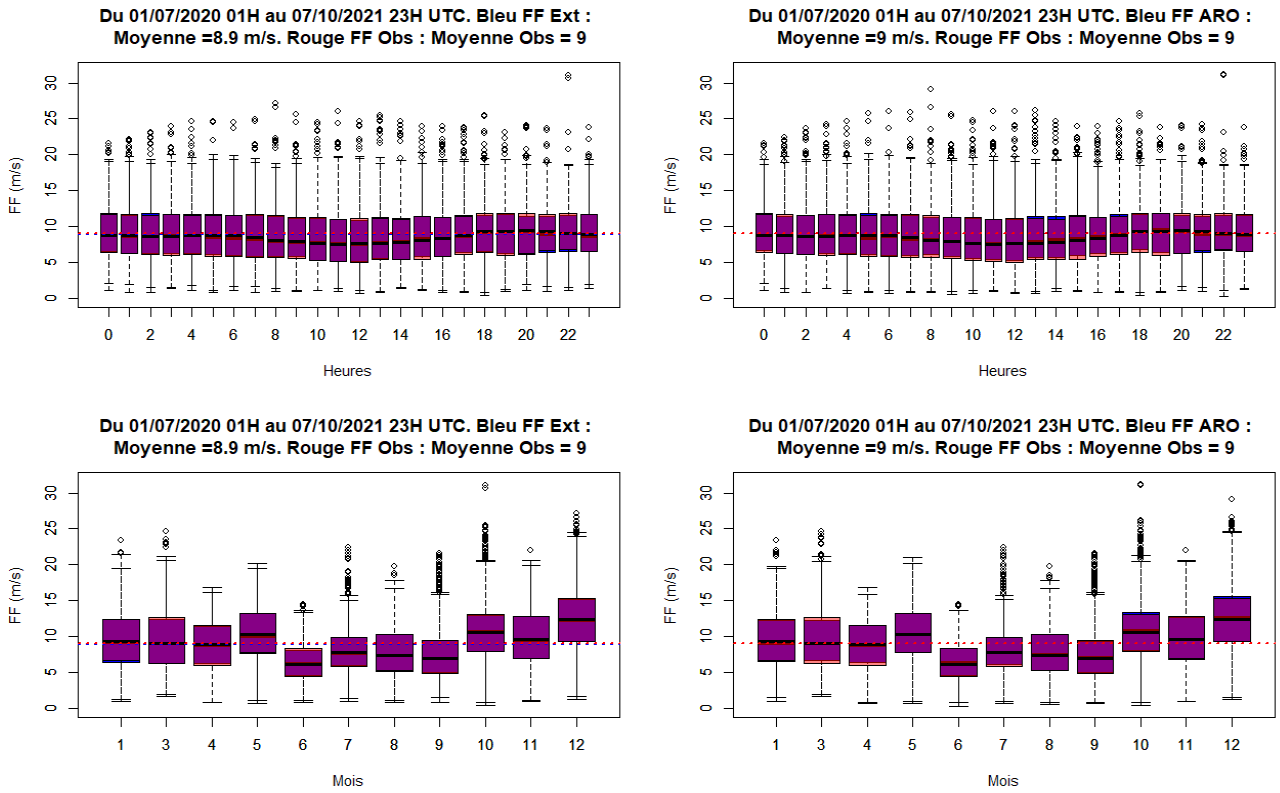

*Illustration 4.1: AO5 Bretagne-Sud Bouée Nord – **FF** – **Restitution des cycles de FF sur la période d'observation**. Sur la première ligne de gauche à droite, on retrouve le cycle diurne de l'extension statistique et AROME superposé à l'observation (rouge), sur la deuxième ligne de gauche à droite, le cycle annuel de l'extension statistique et AROME superposé à l'observation (rouge).*

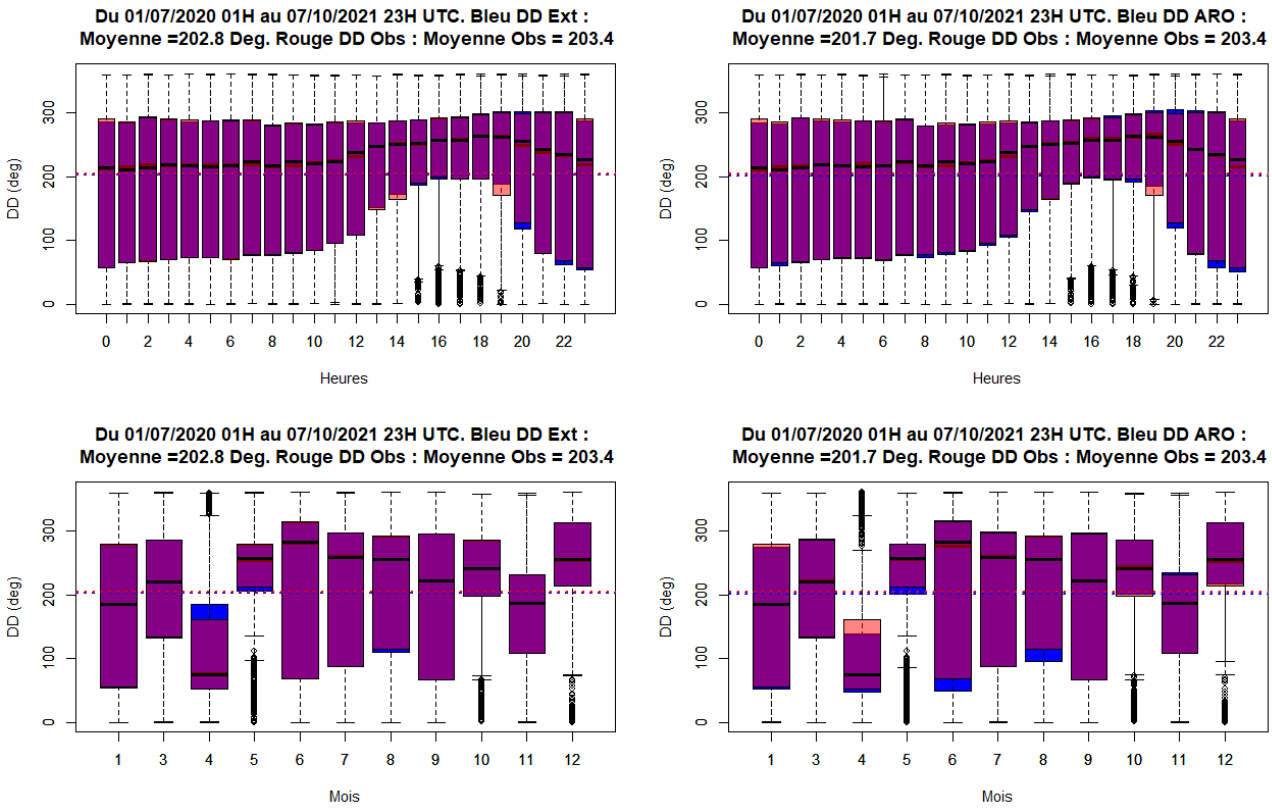

*Illustration 4.2: AO5 Bretagne-Sud Bouée Nord – **DD** – **Restitution des cycles de DD sur la période d'observation**. Sur la première ligne de gauche à droite on retrouve le cycle diurne de l'extension statistique et AROME superposé à l'observation (rouge), sur la deuxième ligne de gauche à droite le cycle annuel de l'extension statistique et AROME superposé à l'observation (rouge).*

Globalement les deux modèles (extension statistique et AROME) restituent bien les cycles diurne et annuel pour la force et la direction du vent.

Pour FF, l'extension statistique est très légèrement plus proche de l'observation qu'AROME sur plusieurs heures (00H, 01H, 07H, 08H, 10H, 11H, 13H, 14H, 16H) pour le cycle diurne, et pour plusieurs mois (mars, avril et juillet) pour le cycle annuel.

Pour DD, le cycle diurne des modèles (extension statistique et AROME) sont très proches du cycle diurne de l'observation. Quant au cycle saisonnier de DD, il est légèrement meilleur pour l'extension statistique que pour AROME notamment sur les mois de janvier, juin, août, octobre et décembre.

On note cependant que le cycle annuel ne tient pas compte du mois de février (mois 2), car il n'y a pas eu de mesure d'observation pour ce mois.

4.2 Restitution des roses des vents

L'illustration 4.3 présente la restitution de rose des vents des modèles : extension statistique, AROME, et deux autres modèles issues de la combinaison de force et direction des deux modèles précédents. Une comparaison est faite avec l'observation sur la période d'observation.

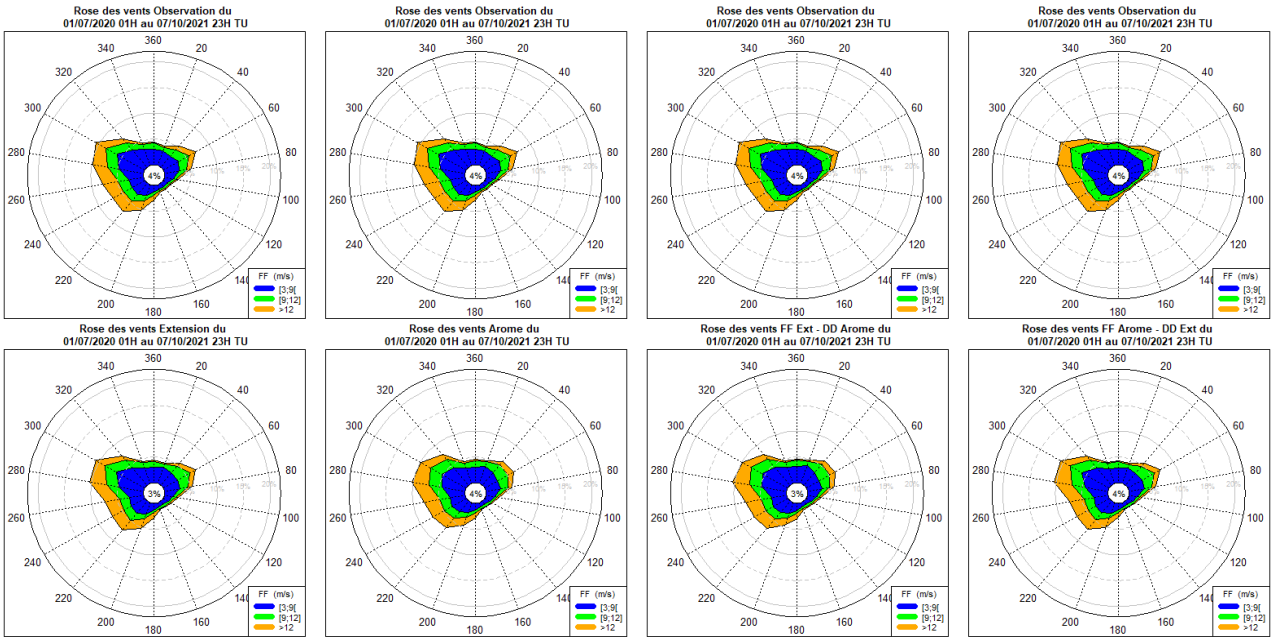

*Illustration 4.3: AO5 Bretagne-Sud Bouée Nord – Roses des vents sur la **période d'observation** (du 01/07/2020 01H au 07/10/2021 23H) : sur la 1ère ligne on retrouve l'observation ; sur la 2ème ligne (de gauche à droite) on retrouve respectivement les roses issues de FF Extension et DD Extension, FF AROME et DD AROME, FF Extension et DD AROME, FF AROME et DD Extension*

Le tableau 4.1 présente les scores B95+ associés aux roses de l'illustration 4.3.

*Tableau 4.1: AO5 Bretagne-Sud Bouée Nord – Scores B95+ des roses des vents de l'extension statistique sur la **période d'observation** (du 01/07/2020 01H au 07/10/2021 23H)*

Modèle	C1	C2	C3	C4 (corrélation circulaire)
Extension statistique	**94.46**	**97.13**	**96.18**	**0.992**
AROME	93.60	94.98	98.40	0.985
FF.Ext – DD.ARO	92.67	94.84	96.18	0.985
FF.ARO – DD.Ext	**95.66**	**97.41**	**98.40**	**0.992**

Au vu des scores B95+ et selon l'aspect visuel des roses, ce sont les modèles de l'extension statistique et issu de FF AROME et DD extension statistique qui apparaissent comme meilleurs modèles (modèle restituant au mieux la rose des vents de l'observation). Cependant, les B95+ sont légèrement avantageux au modèle issu de FF AROME – DD extension.

4.3 Scores complémentaires pour FF

Dans l'ensemble, les différents scores de qualités sont en faveur de l'extension statistique (DD et FF reconstituées avec le modèle statistique de forêt aléatoire). Afin d'apprécier davantage le choix du modèle de FF entre l'extension statistique et AROME, deux scores supplémentaires ont été calculés sur la période d'observation : il s'agit notamment du RMSE par secteur (RMSE sur les classes de direction par secteur de 20 degrés) qui est présenté dans le tableau 4.2, et les scores de qualité sur la puissance théorique obtenue à l'aide d'une courbe de charge théorique[1] pour les deux modèles (extension statistique et AROME) présentés dans le tableau 4.3.

1 Les caractéristiques de l'éolienne choisie sont disponibles ici : https://github.com/IEAWindTask37/IEA-10.0-198-RWT

*Tableau 4.2: AO5 Bretagne-Sud Bouée Nord – Scores **RMSE de FF (en m/s) par secteur de direction** sur la **période d'observation** (en vert les meilleurs scores)*

SECTEUR	SECT _00	SECT _20	SECT _40	SECT _60	SECT _80	SECT _100	SECT _120	SECT _140	SECT _160	SECT _180	SECT _200	SECT _220	SECT _240	SECT _260	SECT _280	SECT_ 300	SECT _320	SECT _340
Intervalle	[350°, 10°[	[10°, 30°[	[30°, 50°[	[50°, 70°[	[70°, 90°[	[90°, 110°[	[110°, 130°[	[130°, 150[	[150°, 170°[	[170°, 190°[	[190°, 210[	[210°, 230°[	[203°, 250°[	[250°, 270°[	[270°, 290°[	[290°, 310°[	[310°, 330°[	[330°, 350°[
NB DATA	423	392	495	699	492	247	172	146	153	255	486	664	658	708	942	997	680	455
RMSE FF EXTENSION STATISTIQUE	**0.88**	**0.82**	**0.9**	**0.83**	**0.9**	**0.95**	**0.86**	**0.87**	**0.91**	**0.82**	**0.91**	**0.9**	**0.96**	**0.95**	**0.88**	**0.9**	**0.9**	**0.94**
RMSE FF AROME	1.23	1.16	1.2	1.12	1.24	1.41	1.32	1.37	1.33	1.21	1.3	1.27	1.4	1.38	1.28	1.27	1.25	1.27

Les scores RMSE sur FF par secteur de direction sont en faveurs de l'extension statistique.

*Tableau 4.3: AO5 Bretagne-Sud Bouée Nord – **Scores de qualité des données de puissances électriques** sur la **période d'observation** (en vert les meilleurs scores)*

SCORE	RMSE	ECT	BIAIS	MAE	PSS P < 2.5 MW	PSS P < 5.0 MW	PSS P > 5.0 MW	PSS P > 7.5 MW	FA P < 2.5 MW	FA P < 5.0 MW	FA P > 5.0 MW	FA P > 7.5 MW
EXTENSION STATISTIQUE	**1.098**	**1.098**	**-0.021**	**0.646**	**0.852**	**0.87**	**0.87**	**0.869**	**0.049**	**0.072**	**0.058**	**0.05**
AROME	1.472	1.47	-0.074	0.891	0.771	0.821	0.821	0.817	0.068	0.084	0.095	0.072

Les scores de qualité des données de puissances sont tous en faveurs de l'extension statistique. L'illustration 4.4 montre le QQ-Plot associés aux données de puissances. Il est majoritairement favorable à l'extension statistique, notamment pour les puissances entre 0 et 6 MW environ.

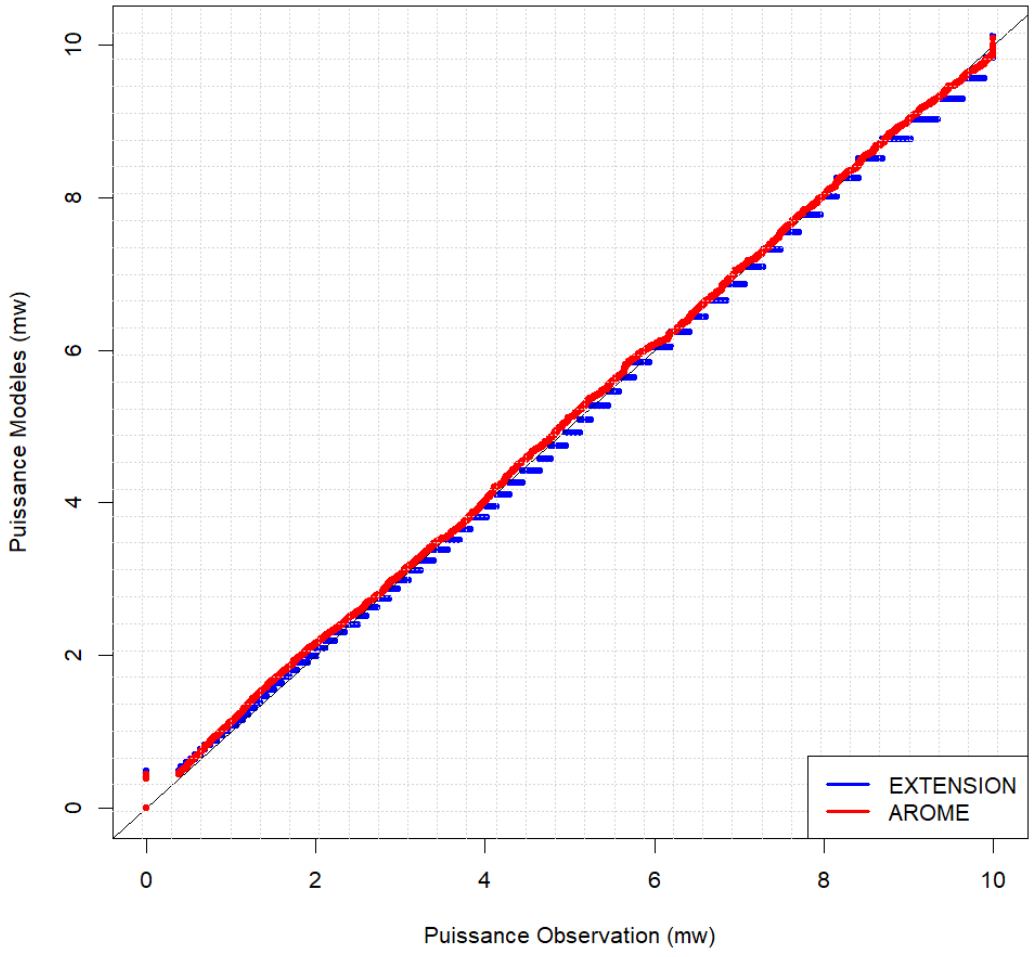

Illustration 4.4: AO5 Bretagne-Sud Bouée Nord – QQPlot des données de puissances sur la période d'observation pour l'extension statistique (bleu) et AROME (rouge)

4.4 Choix du modèle pour la livraison

En tenant compte des scores présentés dans l'ensemble des chapitres 3 et 4, nous choisissons de qualifier l'**extension statistique** (issue de la forêt aléatoire pour la force et la direction du vent) **de modèle le plus approprié pour l'extension de la série de mesure du LiDAR Nord de la zone de l'AO5 Bretagne sud.**

Le modèle AROME sera également mis à disposition des porteurs de projets.

L'illustration 4.5 présente la rose des vents des deux modèles sur la période d'extension, c'est-à-dire du 01/01/2000 00H TU au 07/07/2020 23H TU.

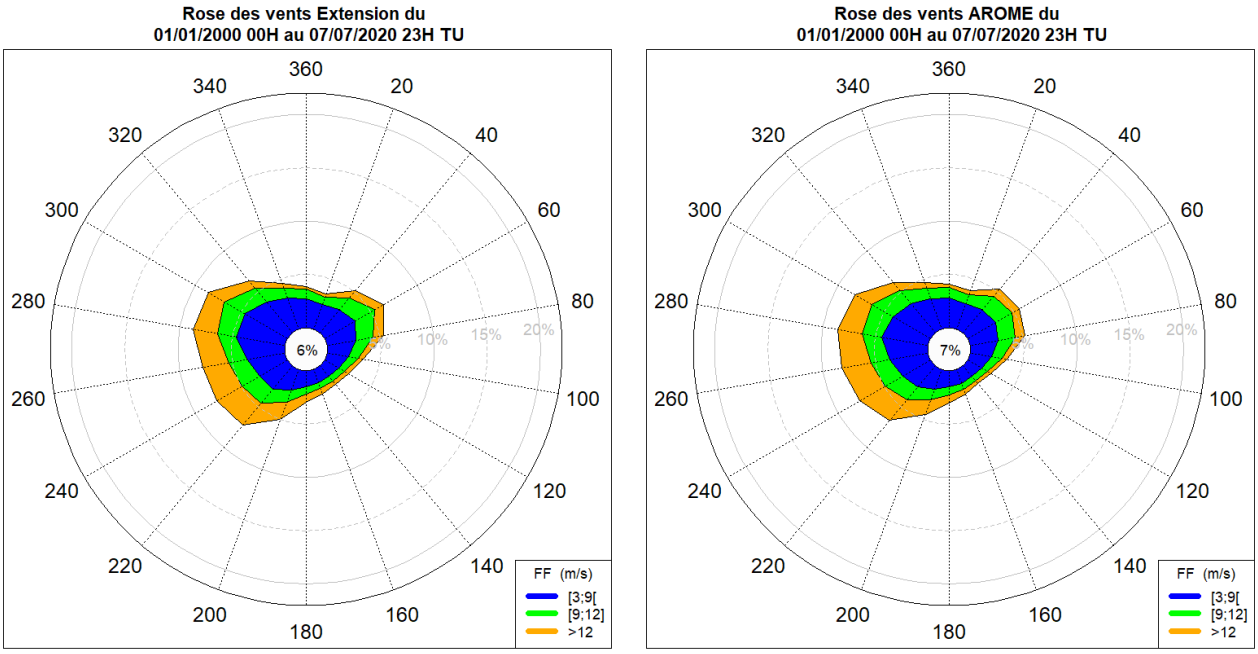

*Illustration 4.5: AO5 Bretagne-Sud Bouée Nord – **Roses des vents sur la période d'extension** du 01/01/2000 00H TU au 07/07/2020 23H TU : à gauche on retrouve la rose de l'extension statistique et à droite la rose d'AROME.*

5 Livraison de l’extension de la série temporelle d’observations horaires

La livraison des données consiste en deux fichiers .csv (le séparateur de colonne est « , », le séparateur de décimale est « . ») contenant :

- les données du 01/01/2000 00H TU au 07/07/2020 23TU,
- une ligne d’en-tête précisant le nom des colonnes,
- les colonnes suivantes :
  - DATE sous la forme AAAAMMJJHH (année, mois, jour, heure TU),
  - FF100 pour la force du vent en m/s à 100 m,
  - DD100 pour la direction du vent en degrés à 100 m,
  - CODE, pour le code qualité associé à la donnée valant :
    - Pour la livraison de l’extension statistique

      - e si les données FF100 et DD100 sont estimées statistiquement

      - b si au moins l’une des données (FF100 ou DD100) est issue du modèle AROME

      - r si au moins l’une des données (FF100 ou DD100) est reconstituée à partir des moyennes horaires des 3 jours précédents et des 3 jours suivants le manque.

    - Pour la livraison AROME

      - b si les deux données (FF100 et DD100) sont issues du modèle AROME,

      - br si l’une des données (FF100 ou DD100) issue du modèle AROME et l’autre (FF100 ou DD100) est reconstituée à partir des moyennes horaires des 3 jours précédents et des 3 jours suivants le manque,

      - r si les deux données (FF100 ou DD100) sont reconstituées à partir des moyennes horaires des 3 jours précédents et des 3 jours suivants le manque.

Dans le cas de cette livraison :

- pour l’extension statistique, on dénombre 178 612 données estimées statistiquement (code e), 982 données issues du modèle AROME (code b), et 262 données reconstituées (code r) sur la période d'extension.
- pour AROME, seuls les codes qualité b et r ont été rencontrés. En effet, il y a 262 données reconstituées à partir des moyennes horaires pour FF et DD, et 179 594 données issues du modèle AROME.

Nous avons appliqué les mêmes critères que pour l’observation, à savoir

- La force du vent à 100 m est arrondie à la première décimale,
- La direction du vent à 100 m est arrondie à l’entier (entre 1° et 360°).

Le nom des fichiers sont :

- ***AO5_Bretagne_Sud_Bouee_Nord_extSerieLidarH100M.csv*** pour l’extension statistique ;
- ***AO5_Bretagne_Sud_Bouee_Nord_AROME****_extSerieLidarH100M.csv pour* AROME.

6 Annexe 1 : Description des modèles statistiques

Dans cette annexe nous décrivons les modèles statistiques utilisés lors de notre étude préliminaire (*cf.* Annexe 2 : Étude d'optimisation de la méthode d'extension) : arbre binaire, modèle linéaire général, modèle linéaire avec anamorphose, forêt aléatoire et réseau de neurone.

6.1 Arbre binaire de décision

Cette méthode statistique est basée sur un découpage, par des hyperplans parallèles aux axes, de l'espace engendré par les variables explicatives. Nommés initialement partitionnement récursif ou segmentation, les développements importants de Breiman et col. (1984) les ont fait connaître sous l'acronyme de CART : Classification and Regression Tree ou encore de C4.5 (Quinlan, 1993) dans la communauté informatique. L'acronyme correspond à deux situations bien distinctes selon que la variable à expliquer, modéliser ou prévoir est qualitative (discrimination ou classification, en anglais) ou quantitative (régression, notre cas).

Les solutions obtenues sont présentées sous une forme graphique simple à interpréter, même pour des néophytes, et constituent une aide efficace pour l'aide à la décision.

Le paramètre de réglage de ce modèle est la règle permettant de décider qu'un nœud est terminal : il devient ainsi une feuille. Ce point est le plus délicat. Il correspond à la recherche d'un modèle parcimonieux. Un arbre trop détaillé, associé à une sur-paramétrisation, est instable et donc probablement plus défaillant pour la prévision d'autres observations.

Un graphique représente la décroissance ou éboulis de la déviance (ou du taux de mal classés) en fonction du nombre croissant de feuilles dans l'arbre. Quand l'amélioration du critère est jugé trop petite ou négligeable, on élague l'arbre au nombre de feuilles obtenues.

Cette méthode ne requière pas d'hypothèse sur les distributions des variables.

Nous utilisons ce modèle essentiellement dans un but exploratoire des données, même si nous proposons une validation de ce modèle.

Référence : *https://www.math.univ-toulouse.fr/~besse/**Wikistat**/pdf/st-m-app-cart.pdf*

6.2 Forêt aléatoire

La méthode d'approche est différente et plus coûteuse que celle des arbres binaires de décision. Elle est décrite dans le paragraphe 2.3.1 du corps de ce rapport.

6.3 Modèle linéaire général

Cette méthode cherche à exprimer l'espérance d'une variable réponse Y (qui est équivalente à sa moyenne, ou plus précisément sa moyenne attendue) en fonction d'une combinaison linéaire des variables explicatives X^i et d'un terme d'erreur (*i.e.*, de bruit) non contrôlé qui doit impérativement suivre une distribution normale et de même variance.

Référence : *www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modlin-mlg.pdf*

6.4 Modèle linéaire avec anamorphose

Cette méthode utilise une régression linéaire multiple barycentrique avec les contraintes fortes sur les coefficients du modèle qui doivent être soit tous positifs soit tous négatifs. La routine DLSEI de la librairie mathématique SLATEC est utilisée pour résoudre cette contrainte. Le modèle est établi par succession de 2 apprentissages.

Tout d'abord la variable à prédire (l'observation) est transformée en une distribution gaussienne à travers l'application d'une loi de weibull et d'une loi normale aux données d'observations.

Ensuite, un premier apprentissage est réalisé sur chacune des variables FF, FXI, T et TKE… (une régression par variable) en utilisant comme prédicteur pour une variable donnée, les points de grilles voisins du site considéré comme prédictand. Parmi des milliers de points à moins de 100 km autour du point central, 16 points sont sélectionnés comme prédicteurs pour participer à la régression de ce premier apprentissage. En effet, les prédicteurs sont classés en utilisant le critère de maximisation du coefficient de corrélation multiple. La sélection des 16 prédicteurs est effectuée en prenant les premiers de ce classement.

Le second apprentissage est réalisé par la suite sur les prédictions du premier apprentissage. C'est donc les prédictions issues de la régression de chaque variable FF, FXI, T et TKE… qui sont utilisées comme prédicteur de cette deuxième régression.

Référence : https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwjB0qDT1uD0AhWF4YUKHdEYB1kQFnoECAMQAQ&url=http%3A%2F%2Fwww.math.u-bordeaux.fr%2F~bbercu%2FEvent%2FEDFWorkshop%2FTalkEDFFarges.pdf&usg=AOvVaw3Ldkf4XFpK9xCyYHjfFlSh

6.5 Réseau de neurone

6.5.1 Réseau multicouche

Le modèle de réseau de neurone est basé sur une représentation schématique des neurones biologiques. Les neurones font partie d'un réseau structuré organisé en couches tout en ayant un échange d'information entre les couches. Le modèle a pour vocation d'entraîner un réseau avec des données disponibles afin d'avoir la meilleure corrélation possible entre les données d'entrée et celles estimées. Il existe plusieurs modèles de réseaux de neurones, dont le perceptron à une couche et le perceptron multicouche. Pour modéliser des phénomènes complexes, les réseaux les plus utilisés sont les réseaux multicouches. Ce sont ces réseaux qui sont utilisés dans notre cas d'études.

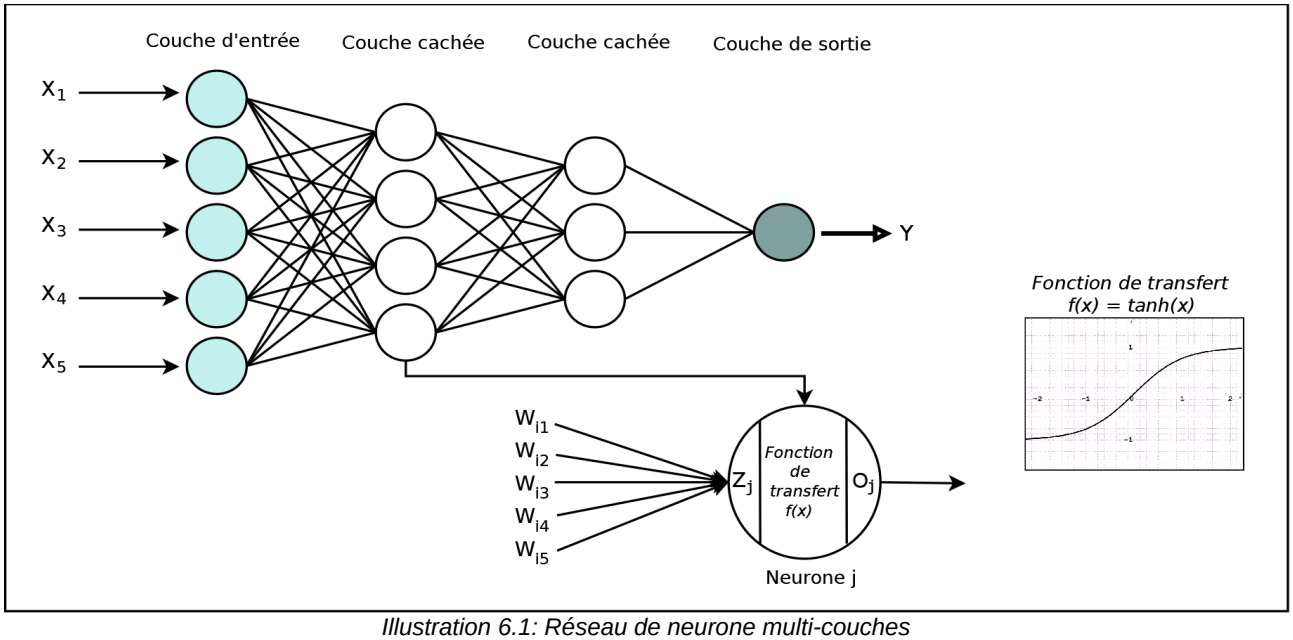

Illustration 6.1: Réseau de neurone multi-couches

La figure de l'illustration 6.1 montre un réseau multicouche. La couche d'entrée est constituée de toutes les données d'entrées du modèle : nombre de neurones égale à nombre de variables d'entrées du modèle (chacun étant connecté à un seul neurone de la couche d'entrée).

Ensuite, chaque neurone *j* reçoit une série de signaux (sorties) des neurones *i* situés aux couches précédentes. Le fonctionnement du réseau est dirigé par la formule suivante :

$$Z_j = \sum_{i=1}^{n_i} W_{ij} . X_i + b_j \quad .$$

Les W_{ij} représentent les poids (pondérations) respectifs des connexions entre les neurones *i* de la couche précédente et les neurones *j* de la couche actuelle. Les b_j sont des biais additionnés à la somme des pondérations pour produire un résultat intermédiaire. Ce résultat est ensuite modulé par une fonction de transfert *f* (ou fonction d'activation), puis transmis aux neurones de la couche suivante. La fonction d'activation *f* permet de relier la sortie O_j d'un neurone donné à la formule de pondération Z_j :

$$O_j = f(Z_j).$$

Il en existe plusieurs types, leurs rôles étant de délinéariser la sortie d'un neurone, permettant ainsi de modifier spatialement la représentation des données et d'avoir par la suite une nouvelle approche sur des données.

Ainsi, ce processus se répète jusqu'à la couche de sortie. Cependant, la couche de sortie comporte toujours autant de neurones que de variables à prédire.

Finalement pour contrôler la sortie du réseau de neurones dans la phase d'apprentissage, la distance entre les valeurs de sortie et les valeurs attendues (valeurs à prédire) est mesurée par la fonction de perte (loss fonction) aussi appelé fonction de coût :

$$E = \frac{1}{2n} \times \sum_{k=1}^{n} (Y_k - \hat{Y}_k)^2 \quad .$$

où Y_k représente la valeur observée et $\hat{Y}_k$ la valeur estimée, *n* est le nombre de données et *E* l'erreur globale mesuré.

L'erreur mesurée est aussitôt propagée dans le réseau pour ajuster les poids W_{ij} . Cette technique couramment utilisée dans la mise en place des réseaux multi-couches est connu sous le nom de la technique de rétro-propagation du gradient. L'algorithme de descente du gradient de l'erreur a pour but de minimiser la fonction de coût, en convergeant de manière itérative vers une configuration optimale des poids de chaque connexion du réseau.

Il existe plusieurs algorithmes pour réaliser la descente du gradient. On les appelle **Optimizer** dans l'implémentation Keras. Parmi ces algorithmes, nous utilisons notamment les optimiseurs

- **SGD :** Stochastic Gradient Descent, qui est une méthode de descente de gradient utilisant des échantillons sélectionnés au hasard au lieu de l'ensemble des données pour chaque itération dans le processus de descente.
- **Adam :** qui est aussi une méthode de descente de gradient stochastique basée sur l'estimation adaptative des moments du premier et du second ordre.
- **RAdam :** Rectified Adam, est une variante de l'algorithme Adam introduisant un terme pour rectifier la variance du taux d'apprentissage adaptatif.

Tous ces algorithmes de descente du gradient sont implémentés dans Keras. Cependant, il n'y a pas vraiment l'un qui est meilleur que l'autre, cela dépend surtout du problème à traiter ainsi que la configuration du réseau auquel ils sont utilisés.

Le schéma de la figure 6.2 est une illustration du processus de fonctionnement de la fonction de perte ainsi que la mise à jour des paramètres du réseau.

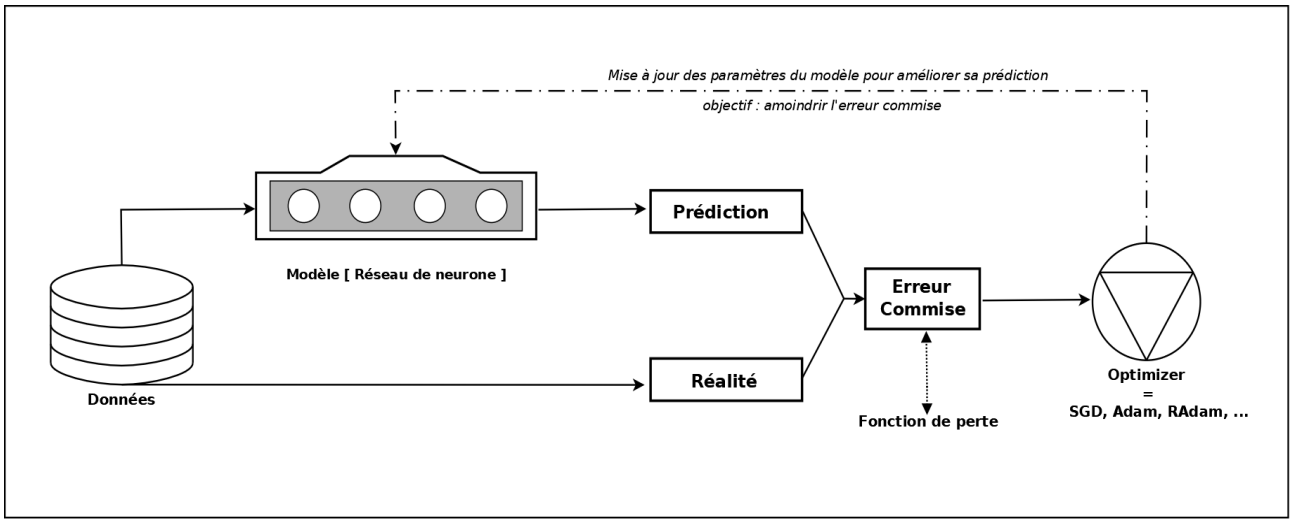

Illustration 6.2: Processus de fonctionnement de la fonction de perte et de la mise à jour des paramètres du réseau

6.5.2 Entraînement du réseau

Une fois le prototype du modèle définit, il est important de régler les **hyperparamètres** (paramètres réglés pendant les exécutions successives de l'entraînement du modèle) dans la phase d'apprentissage. Les paramètres les plus importants à ajuster sont entre autres :

- le nombre de couche du modèle ainsi que le nombre de neurone dans chaque couche,
- le choix de la fonction d'activation pour chaque couche,
- le choix de l'Optimizer ainsi que les différents paramètres d'entraînement notamment le taux d'apprentissage (**Learning Rate** ou LR), le nombre d'itération (ou **Epochs**), et la taille du lot (également appelé le **Batch Size**).

En effet, en pratique la descente du gradient se fait par lot (appelé batch) constitué de mini-lot (aussi appelé mini-batch). Il est bien plus efficace de calculer la perte pour un mini-lot que pour l'ensemble entier des données d'apprentissage.

- Un Batch (ou lot) : est l'ensemble d'exemples utilisés dans une itération de l'entraînement du modèle (aussi appelé Epochs), c'est-à-dire pour une mise à jour du gradient. En principe il est constitué de mini-batch qui prennent leurs valeurs dans l'intégralité du jeu de donnée d'entraînement.

- Un mini-batch (ou mini-lot) : est un petit sous-ensemble, sélectionné aléatoirement, du lot complet d'exemple (intégralité des données d'apprentissage) exécutés simultanément dans une même itération d'apprentissage. En pratique, on ne le choisit pas directement, il se déduit directement de la taille du lot qu'on aurait défini au préalable.

- Le Batch Size (ou taille du lot) : est le nombre d'exemple (échantillon) présenté au modèle. En fonction de la taille défini, chaque lot est constitué de façon aléatoire et propager dans le réseau.

Par exemple, si on dispose de 3750 données dans le jeu d'entraînement, en configurant un Batch Size égale à 250. Pour chaque itération (Epochs), l'algorithme va décomposer les 3750 données en 15 mini-lots puis entraîne le réseau successivement avec chaque mini-lot jusqu'à ce que tous les échantillons se propagent dans le réseau.

- Epochs (ou itération) : est le cycle d'apprentissage complet sur l'intégralité de l'ensemble de données de manière à ce que chaque exemple ait été vu une fois par le modèle.

- Learning Rate (ou taux d'apprentissage) : est une grandeur scalaire (généralement comprise entre 0 et 1) utilisée pour entraîner le modèle via la descente de gradient. À chaque itération, l'algorithme de descente du gradient multiplie le taux d'apprentissage par le gradient. Le produit ainsi généré est appelé pas de gradient. En pratique, il pilote la vitesse de convergence de la descente du

gradient. D'une part, plus sa valeur est faible, plus la convergence est lente (mais sécurisée). D'autre part, plus sa valeur est élevée, plus la convergence est rapide (mais au risque de diverger). Il faut donc trouver un compromis.

Très concrètement une fois qu'on a fixé le nombre de couche et de neurone dans la couche, on joue sur le Learning Rate, le Batch Size et le nombre d'Epochs pour avoir un réseau optimal (compromis entre convergence, temps de calcul et complexité du modèle) en examinant la courbe d'apprentissage. En effet, la courbe d'apprentissage donne une indication sur de l'état d'apprentissage du modèle : modèle bien ajusté, en sur-apprentissage ou en sous apprentissage.

Si l'entraînement ne conduit pas à un modèle bien ajusté, on réajuste les paramètres du modèle jusqu'à trouver le modèle optimal.

7 Annexe 2 : Étude d'optimisation de la méthode d'extension

7.1 Préambule

Pour étudier le potentiel éolien des sites d'implantation, les études réalisées pour la DGEC se font généralement en utilisant 1 an (voire moins d'un an comme le cas de la livraison intermédiaire de l'AO4 Normandie) de données d'observations horaires de vent issue de campagne de mesure. Ces données d'observations sont ensuite étendues (en utilisant des modèles statistiques) sur une période reconstituée autour de 20 ans selon la profondeur de la base de données climatologique AROME de Météo-France.

À travers les expériences des études précédentes pour la DGEC (Dunkerque et Oléron) et de deux nouvelles études élaborées à partir d'une profondeur de données plus longue et décrite ci-dessus, nous cherchons à fiabiliser et optimiser la méthode d'extension à mettre en œuvre pour les futures campagnes de dérisquage éolien off-shore.

En effet, lors des deux premières campagnes de dérisquage pour la DGEC, nous avons pu mettre en évidence l'apport de la modélisation statistique à travers l'utilisation des modèles linéaire général et de forêt aléatoire comme méthodes d'extension des séries horaires. Dans cette étude, nous testons un modèle supplémentaire (réseau de neurone), une version du modèle linéaire généralisé par anamorphose gaussienne, une nouvelle méthode de sélection des variables explicatives (ACP) et de nouvelles variables explicatives en ayant comme objectif de pouvoir mettre au point un modèle unique que l'on pourra utiliser dans les études d'extension de série à venir.

7.2 Protocole pour les stations d'études

Pour deux stations de test (dont les noms et les coordonnées resteront confidentiels dans ce rapport mais que l'on nommera « Station de test 1 » et « Station de test 2 » dans la suite), on dispose de plus de 4 ans de mesure (2017-2020) d'observation sodar sur terre. Avec ces données d'observation de vent, on se place dans le même contexte que les études DGEC : on entraîne les modèles statiques uniquement sur 1 an d'observation, ensuite on réalise l'extension (reconstitution de l'observation avec les modèles statistiques grâce aux prédictions des modèles), puis on compare cette extension statistique à l'observation sur la période non apprise. Cela permet ainsi de valider un modèle statistique robuste qui est en mesure d'être au plus proche de l'observation, et d'évaluer par la suite l'erreur commise par ce modèle validé.

L'objectif de cette annexe est de présenter de manière synthétique les résultats des différentes expériences qui ont été faites pour choisir le modèle statistique final pour l'extension.

À savoir que la plupart des expérimentations ont été réalisées sur les deux stations de test mais pour ne pas alourdir ce rapport, les détails et illustrations sont majoritairement fournis pour la station de test 1 et seul un petit commentaire supplémentaire est ajouté pour la station de test 2.

7.3 Identification du point AROME de référence

Comme décrit dans la méthodologie employée (*cf.* chapitre 2), avant d'établir les modèles statistiques, on identifie d'abord le point AROME de référence dont les paramètres seront utilisés comme variables explicatives (prédicteurs) des modèles statistiques. Les caractéristiques de vent de ce point doivent donc être les plus proches de celles du point d'observation.

Le choix du point AROME de référence s'effectue en s'appuyant sur les indicateurs de qualité (B95+, voir section 2.2.1 pour la définition) des roses de vent. L'illustration 7.1 présente les roses de vent du point d'observation et des 4 points AROME voisins pour la station de test 1.

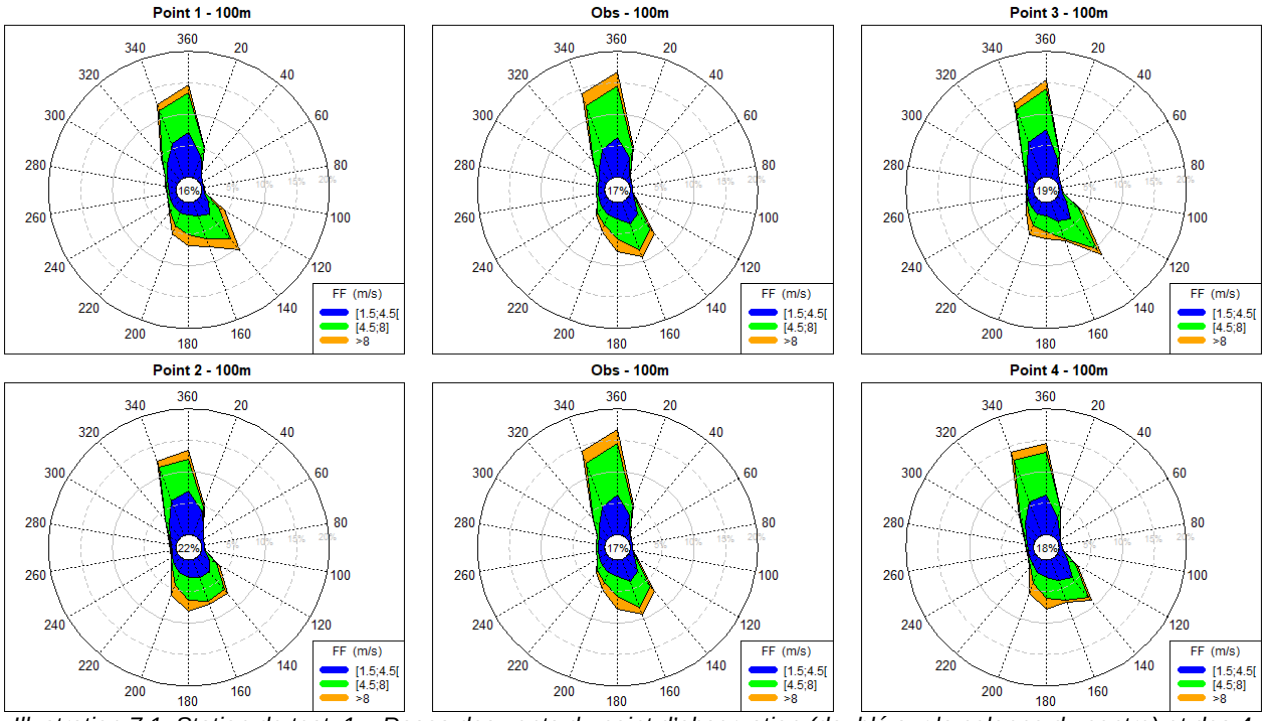

Illustration 7.1: Station de test 1 – Roses des vents du point d'observation (doublé sur la colonne du centre) et des 4 points AROME voisins (colonne de gauche et droite) sur la période du 01/01/2017 00H TU au 31/12/2020 23H TU

En visualisant ces roses de vent, globalement, ce sont les points AROME 1 et 4 qui semblent un peu plus proches de l'observation que les points 2 et 3. En effet, les vents du sud sont beaucoup moins bien restitués par les points 2 et 3 que les deux autres points.

Cependant, il est difficile de trancher visuellement la ressemblance des points 1 et 4 par rapport au point d'observation pour les vents du sud. C'est donc les scores B95+ qui permettront de faire la différence.

Le tableau suivant présente les scores B95+ des 4 points AROME voisins du point d'observation.

Tableau 7.1: Station de test 1 – Scores B95+ des 4 points AROME voisins du point d'observation (en vert les scores décisifs)

ID	C1	C2	C3	C4 (corrélation circulaire)
Point AROME voisin 1	83.76	85.57	97.62	0.48
Point AROME voisin 2	77.58	80.55	85.03	0.39
Point AROME voisin 3	77.56	78.74	91.79	0.18
Point AROME voisin 4	85.69	86.82	94.72	0.44

On rappelle que le critère C1 est le critère global qui met en contribution tous les 18 secteurs de la direction du vent (DD) et les 4 classes de la force du vent (FF). Le critère C2 est celui qui permet d'étudier la qualité de la modélisation des fréquences de direction. Il est donc à privilégier pour la mise en évidence de DD. Quant au critère C3, il permet d'étudier la qualité de la modélisation des fréquences de force du vent. C3 est donc favorable pour mettre en évidence FF. Et enfin le critère C4, appliqué uniquement sur la direction du vent, est la corrélation circulaire.

Les points AROME 1 et 4 ont des scores relativement proches l'un de l'autre pour l'ensemble des critères de qualité, ce sont donc les critères favorables à DD (notamment C2) qui ont été décisifs pour le choix du point de référence. Et **c'est le point 4 qui a été choisi comme référence**, car il a un meilleur score C2 que le point 1 avec des corrélations circulaires quasiment identiques.

Le même travail a été réalisé sur la station de test 2, et c'est le point 1 qui a été sélectionné comme référence avec une rose de vent plus proche de l'observation que celle des trois autres points voisins.

Une fois le point AROME de référence sélectionnée, on procède à la modélisation de FF puis de DD.

7.4 Modélisation de la force du vent à 100 m

Les modèles statistiques sont ajustés non pas sur la force du vent FF observé, mais sur son écart avec la force du vent AROME. D'après les expériences antérieures, cela permet de mieux modéliser les valeurs extrêmes qui affectent généralement certains modèles statistiques. Par conséquent, **la variable à prédire est :** $Y = FFmFFARO = FF_{OBS} - FF_{ARO}$.

7.4.1 Sélection des variables explicatives

Une étape importante dans la modélisation de la force du vent (FF) est le choix des variables explicatives. Ce choix est basé sur la connaissance météorologique de l'origine du vent dont plusieurs paramètres sont accessibles en sortie du modèle AROME. Les cycles saisonnier et diurne sont également pris en considération du fait de l'information qu'ils apportent sur la variabilité temporelle du vent.

7.4.1.1 Paramètres calendaires

Du fait de phénomènes météorologiques de plus ou moins grande ampleur (effet de brise, prépondérance des tempêtes en hiver…), la vitesse et la direction du vent varient en fonction de l'heure du jour et du mois de l'année. On cherche donc à établir à partir de l'heure et du mois des observations (et des sorties AROME) des comportements moyens, plus stables sur des plages horaires et saisonnières fixes. Dans le cas des stations de test 1 et 2, nous avons pu dégager les plages suivantes :

- La variable catégorielle notée **RegHH_FF** représente la variable du cycle diurne (illustration 7.2 pour la station de test 1). Elle est composée de 2 facteurs correspondant à 2 plages horaires : [10H – 20H] et [21H – 09H] pour la station de test 1, et [06H – 17H] et [18H – 05H] pour la station de test 2.
- La variable catégorielle notée **RegMM** représente la variable saisonnière. Elle est composée de 4 facteurs correspondants aux 4 saisons météorologiques : Hiver (décembre, janvier, février), Printemps (mars, avril, mai), Été (juin, juillet, août), et Automne (septembre, octobre, novembre). Ce sont les mêmes catégories pour les deux stations.

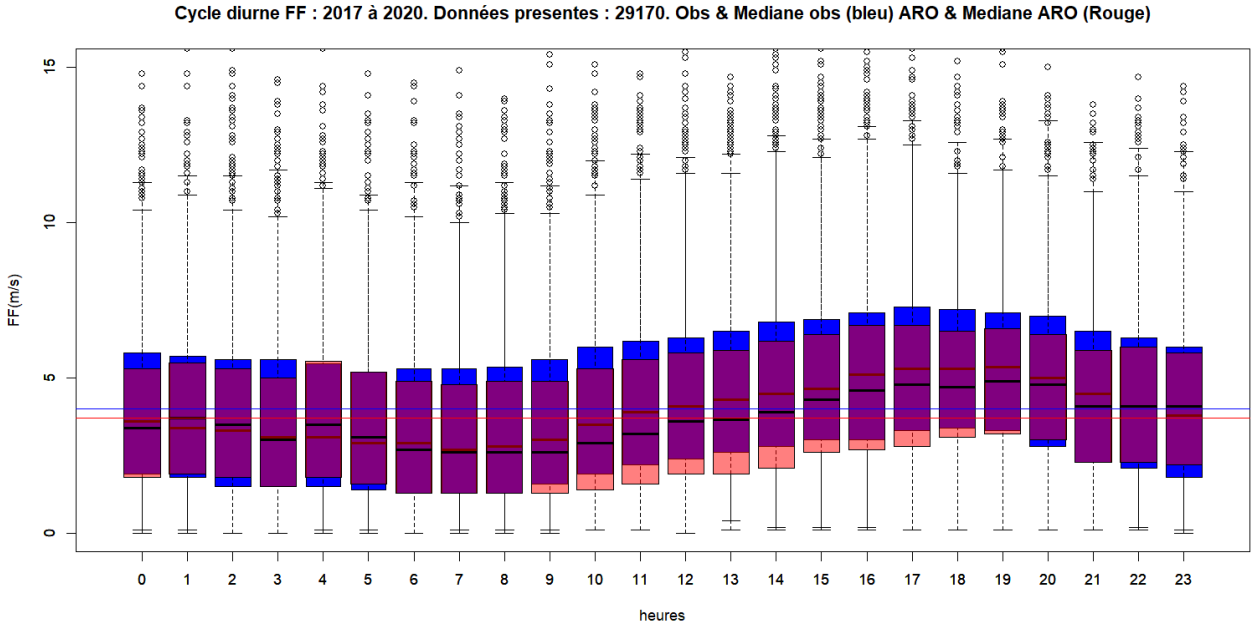

Illustration 7.2: Station de test 1 – Cycle diurne de la force du vent observée (bleu) et AROME (rouge) de 2017 à 2020.

7.4.1.2 Paramètres météorologiques en sortie AROME

Pour les paramètres accessibles directement en sortie AROME, on examine les liens entre la variable à prédire **Y** et les variables explicatives $\mathbf{X^i}$, mais aussi entre les variables explicatives. On s'assure donc qu'il y

ait assez de corrélation entre **Y** et chaque $\mathbf{X^i}$, tout en évitant d'avoir une forte corrélation entre les $\mathbf{X^i}$ (moins de 0,96). Cela permet d'éviter le problème de colinéarité pour la mise en place des modèles linéaires.

Pour cela, on a donc adopté une technique de sélection des variables explicatives en 2 phases. Tout d'abord on sélectionne toutes les variables explicatives bien corrélées avec la variable à prédire, même si celles-ci sont fortement corrélées entre elles. Ensuite on divise cette première sélection en 2 groupes en dissociant les variables explicatives fortement corrélées entre elles. Le premier groupe de cette dissociation est conservé dans le choix des variables explicatives finales, puis on réalise une analyse en composantes principales (**ACP**) sur le deuxième groupe afin de réduire sa dimension tout en conservant le maximum d'information, mais aussi de les décorréler avec les variables explicatives du premier groupe déjà admis. Ainsi les variables explicatives finales seront constituées de celles du premier groupe et des composantes principales de l'ACP.

Les variables pré-sélectionnées sont les suivantes :

- **FF :** force du vent AROME à plusieurs niveaux de hauteur. C'est une variable quantitative notée FFARO_NIV où NIV={10, 100, 250, 500} est le niveau de hauteur en mètre auquel elle a été prise en compte,
- **SECTEUR :** direction du vent AROME à 100 mètres par secteur de 20 degrés. C'est une variable catégorielle à 18 facteurs (0, 20, 40… jusqu'à 340, l'intervalle des 20° étant centré sur la valeur des facteurs) noté SECTEURARO_100,
- **T :** température AROME à plusieurs niveaux. Variable quantitative noté TARO_NIV avec NIV={2, 100, 250, 500},
- **PMER :** pression AROME réduite au niveau de la mer. Variable quantitative noté PMERARO,
- **TKE et SQRTKE :** la turbulence AROME au niveau NIV={10, 100, 160}. Variable quantitative noté TKEARO_NIV et sa racine carrée notée SQRTKEARO_NIV. On exploite la racine carrée TKE de manière à privilégier un lien linéaire avec la force du vent de l'observation,
- **HU :** humidité AROME relative à 2 mètres. Variable quantitative noté HUARO_2,
- **TPW850 :** température pseudo-adiabatique potentielle du thermomètre mouillé AROME à 850 hectopascals. C'est une variable quantitative qui caractérise une masse d'air. Elle est notée TPW850.

7.4.1.2.1 Résultats de l'analyse en composantes principales

Pour la station de test 1, l'ACP a été réalisée sur 11 des variables explicatives préalablement sélectionnées dans le deuxième groupe évoqué ci-dessus. Il s'agit notamment de : FFARO_50, FFARO_500, TARO_50, TARO_100, TARO_250, TKEARO_100, TKEARO_160, SQRTTKEARO_10, SQRTTKEARO_50, SQRTTKEARO_160, et PMERARO. La figure 7.3 montre la part de variance expliquée par l'ACP. Cette figure est associée au tableau 7.2 qui donne le détail pour chaque composante principale.

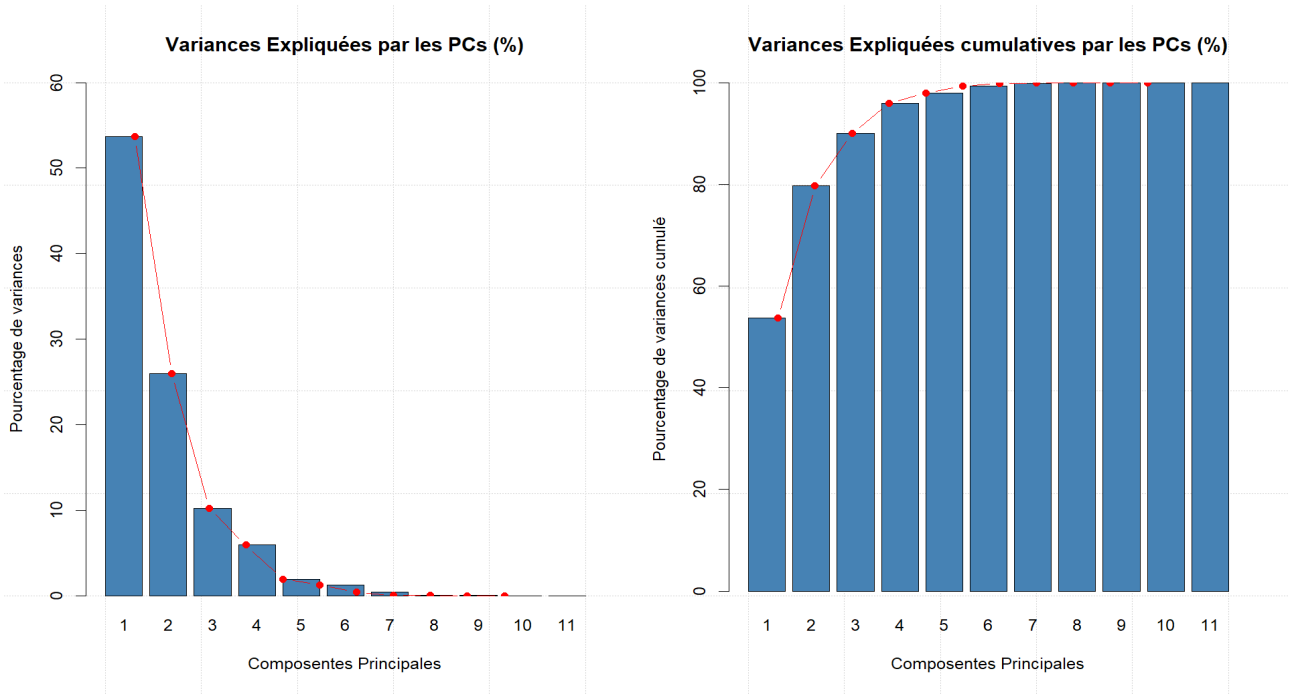

Illustration 7.3: Station de test 1 – Part de variance expliquée de l'ACP

Tableau 7.2: Station de test 1 – Tableau des variances expliquées par l'ACP (en vert les composantes conservées)

Composante principale	valeur propre	pourcentage de variance (%)	pourcentage de variance cumulative (%)
PC1	5.90673407247875	53.69758247708	53.69758247708
PC2	2.85788353236657	25.9807593851508	79.6783418622308
PC3	1.13247164812364	10.2951968011241	89.9735386633549
PC4	0.656785890026829	5.97078081842577	95.9443194817807
PC5	0.221518742956181	2.01380675414712	97.9581262359278
PC6	0.146704214854232	1.33367468049303	99.2918009164209
PC7	0.0504935297812545	0.459032088920499	99.7508330053414
PC8	0.0147102751125461	0.13372977375042	99.8845627790918
PC9	0.00867267446628389	0.078842495148036	99.9634052742398
PC10	0.00349898393380089	0.0318089448527356	99.9952142190926
PC11	0.000526435899818015	0.00478578090743654	100

Les 5 premières composantes principales expliquent 98 % de la variance des données de l'ACP. À partir de la composante principale 6, la part de variance expliquée est trop faible pour chacun de ces composantes principales. L'analyse a donc retenu les 5 premiers axes factoriels comme dimension intrinsèque des données (les autres dimensions étant considérées comme du bruit). Les composantes principales 1, 2 et 3 sont celles qui résument le maximum d'information avec respectivement 54, 26 et 10 % de la variance des données. Il ne faut cependant pas négliger les composantes 4 et 5 qui, malgré leur part de variance inférieure aux 3 premiers, peuvent renfermer de l'information pertinente. Pour cela, il est important de visualiser la contribution de chaque variable pour les axes factoriels de l'ACP. La figure suivante 7.4 montrent la contribution de chaque variable pour chacun des axes factoriels de l'ACP.

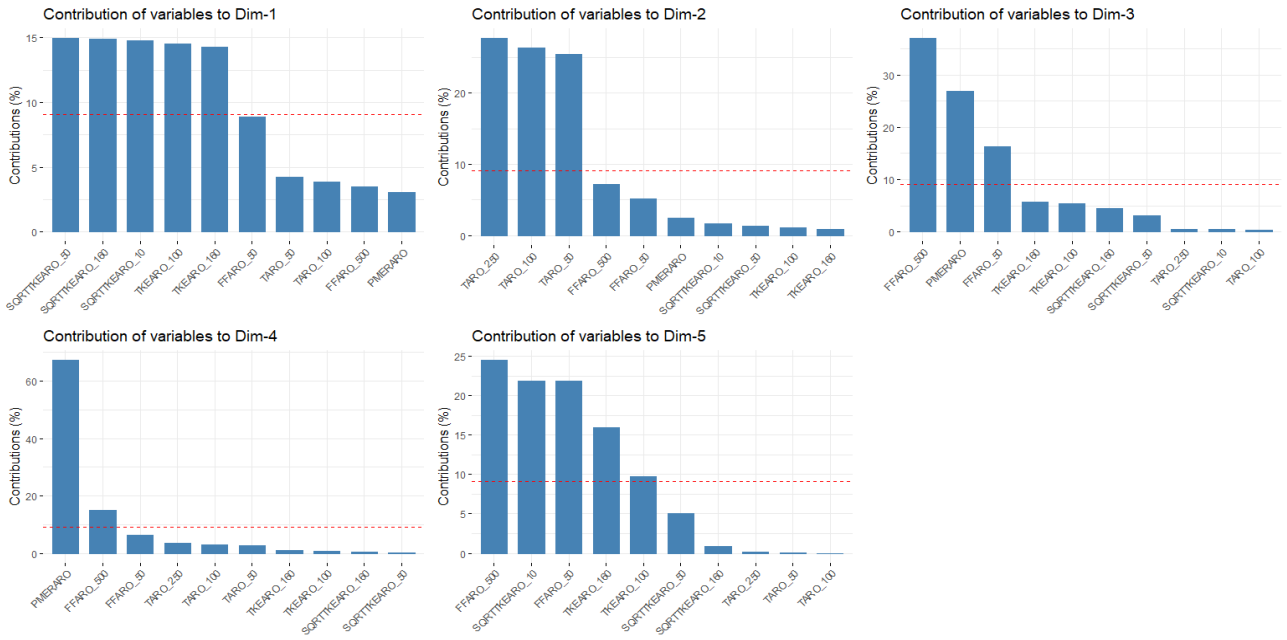

Illustration 7.4: Station de test 1 – FF – Contribution des variables à l'ACP (une figure pour chaque dimension)

La figure 7.4 permet d'avoir un aperçu de la contribution de chaque variable sur chacun des axes factoriels. Les contributions des variables dans la définition d'un axe principal donné, sont exprimées en pourcentage. Les lignes en pointillées rouges sur les graphiques de l'illustration 7.4 indiquent la contribution moyenne attendue. **Si la contribution des variables était uniforme, la valeur attendue serait** $1/length(variables)=1/11=9\%$. Pour une composante donnée, une variable avec une contribution supérieure à ce seuil peut être considérée comme importante pour contribuer à la composante.

On s'aperçoit que la moitié des variables de l'ACP ont une contribution au-delà de la moyenne pour la définition de la première dimension (avec une corrélation plutôt forte pour la turbulence AROME et sa racine carrée). La deuxième dimension montre une bonne corrélation des températures AROME avec cet axe. La troisième dimension est plutôt liée à la force du vent AROME et à la pression AROME réduite au niveau de la mer. Quant à la quatrième dimension, elle est liée uniquement à 2 variables de manière significative : la pression AROME réduite au niveau de la mer (qui contribue à plus de 67 %), et la force du vent AROME au niveau 500 mètres. Plusieurs variables contribuent également à la définition de la cinquième dimension, notamment la force du vent AROME et la turbulence AROME et sa racine carrée. Elle n'est donc pas à négliger, cependant l'interprétabilité de cet axe (ainsi que tous les autres axes factoriels) dépendra de la qualité de représentation des variables sur celui-ci.

Pour mieux apprécier la qualité de la représentation et la proximité des variables sur chaque axe factoriel, on a représenté chaque variable sur tous les plans factoriels (cercle de corrélation des variables) comme le montre les deux figures suivantes (illustrations 7.5 et 7.6).

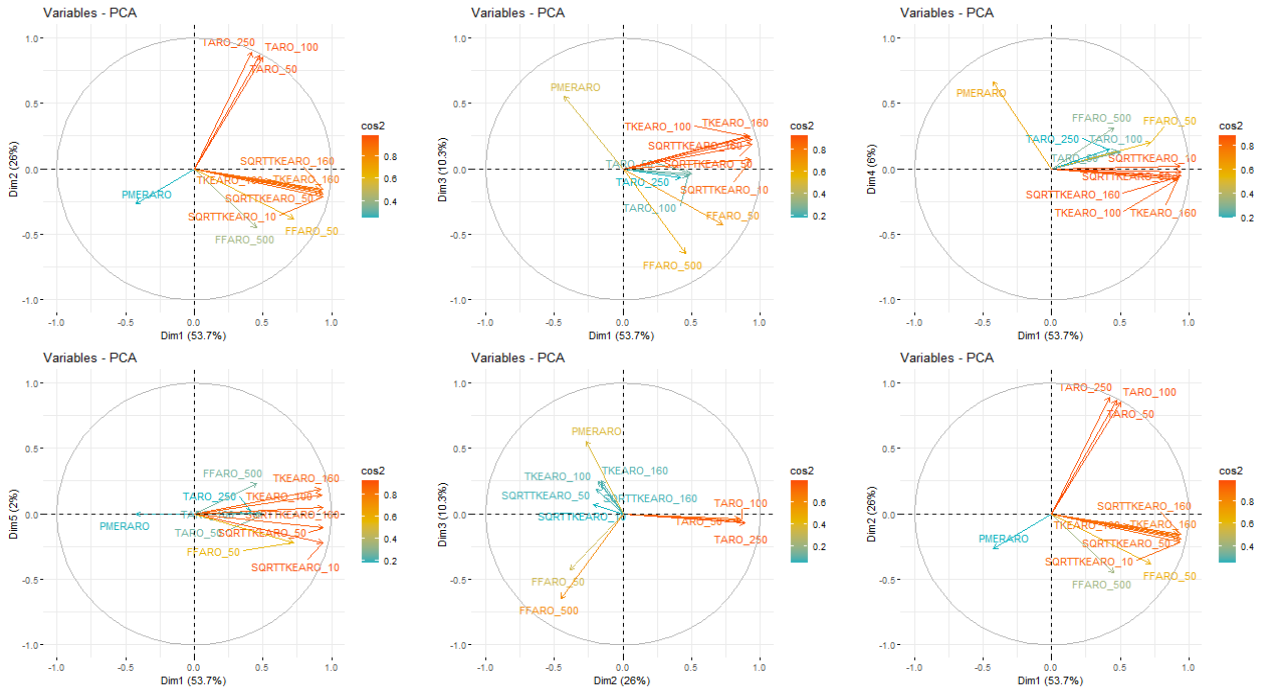

Illustration 7.5: Station de test 1 – FF – Représentation des variables sur les plans factoriels de l'ACP (première partie des plans) : première ligne dimension 1-2, dimension 1-3 et dimension 1-4 ; deuxième ligne dimension 1-5, dimension 2-3, et dimension 1-2 (à nouveau)

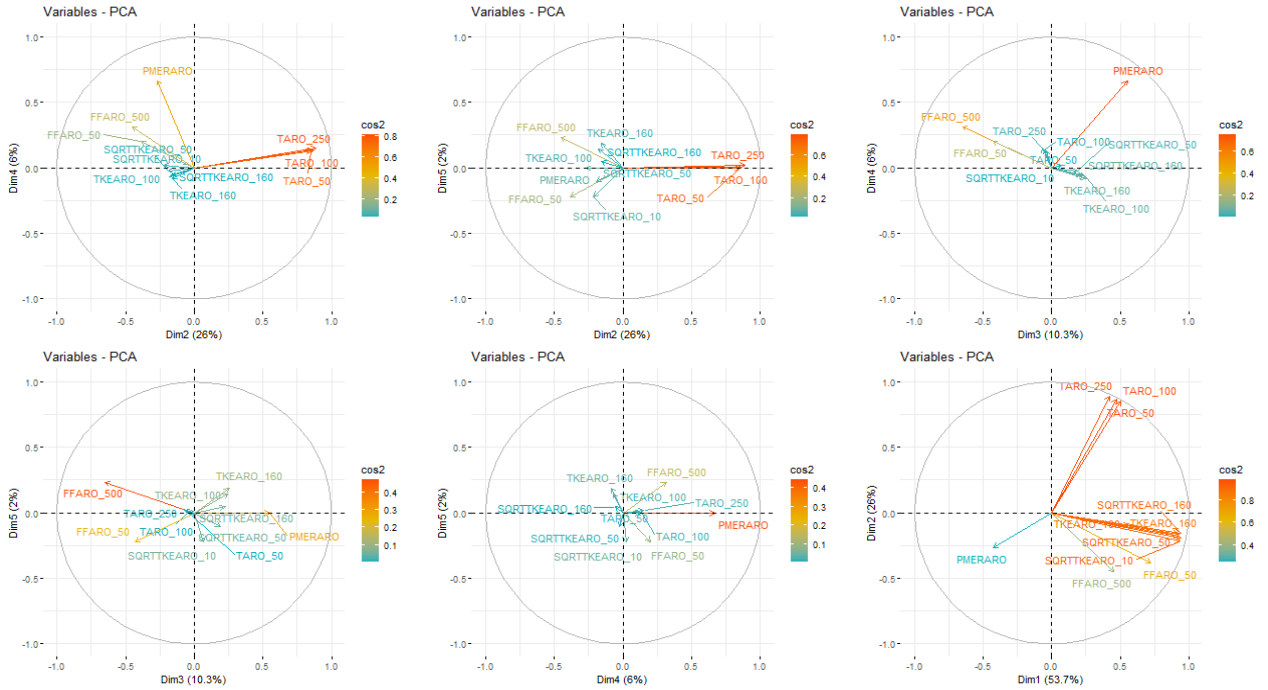

Illustration 7.6: Station de test 1 – FF – Représentation des variables sur les plans factoriels de l'ACP (deuxième partie des plans) : première ligne dimension 2-4, dimension 2-5 et dimension 3-4 ; deuxième ligne dimension 3-5, dimension 4-5, et dimension 1-2

Sur le cercle de corrélation, la qualité de représentation des variables est donnée par le cosinus carré (noté cos2) :

- Un cos2 élevé (proche de 1) indique une bonne représentation de la variable sur les axes principaux en considération. Dans ce cas, la variable est positionnée à proximité de la circonférence du cercle de corrélation.
- Un faible cos2 (proche de 0) indique que la variable n'est pas parfaitement représentée par les axes principaux. Dans ce cas, la variable est proche du centre du cercle.

Ainsi, en tenant compte de la qualité de représentation des variables et de leurs contributions pour chaque axe factoriel, les plans les plus pertinents à interpréter sont les plans constitués par : l'axe 1 et 2 (Dim1-Dim2), l'axe 1 et 3 (Dim1-Dim3), l'axe 2 et 3 (Dim2-Dim3), l'axe 3 et 4 (Dim3-Dim4).

- Le premier plan (axes factoriels 1 et 2) explique 80 % de la variance des données. Il montre une bonne corrélation entre les TKE et SQRTKE qui sont plutôt liées à l'axe 1, et une bonne corrélation entre les températures AROME qui sont liées aux 2 axes mais un peu plus à l'axe 2 qu'à l'axe 1.
- Le second plan (axes factoriels 1 et 3, avec 64 % de l'inertie des données) confirme toujours la bonne corrélation des TKE et SQRTKE liées à l'axe 1. Par ailleurs, il oppose (en termes de comportement) les forces du vent AROME à la PMER (qui sont tous bien liées aux 2 axes).
- Le plan constitué par l'axe 2 et 3 explique quant à lui 36 % de la variance des données. Il confirme d'une part la bonne corrélation des températures AROME (liées à l'axe 2), d'autres part l'opposition entre PMER et FF.
- Le plan constitué par l'axe 3 et 4 avec 16 % de l'inertie des données, fait ressortir uniquement l'opposition entre les variables FF et PMER. En effet, ces variables ont une liaison plutôt bonne avec les 2 axes.
- Et enfin comme évoqué plus haut, l'axe 5 ne doit pas non plus être négligé malgré ses 2 % de variance expliquée. En effet, dans tous les plans pour lesquels il intervient, les phénomènes que l'on vient d'évoquer se vérifient par rapport à l'axe auquel il est associé, notamment : que les TKE et SQRTKE sont fortement corrélées entre eux (axe 1), que les températures AROME sont fortement corrélées entre eux (axes 2), que FF et PMER ont des comportements opposés (axe 3), que PMER est fortement lié à l'axe 4 (axe 4).

On pourrait se demander si les trois premières composantes principales n'auraient pas suffi pour expliquer les données de l'ACP. Mais l'arbre binaire plus tard, permet de confirmer la pertinence des composantes principales jusqu'au cinquième.

Le travail a été réalisé pour les deux stations de test, elles se comportent de façon similaire. **À l'issue nous conservons donc 5 potentielles variables explicatives supplémentaires (PC1 à PC5) qui sont ajouté aux autres variables explicatives du premier groupe.**

7.4.1.2.2 Sélection finale des paramètres accessibles en sortie AROME

Maintenant qu'on est sûr de récupérer de l'information dans les données à travers l'ACP, on peut donc intégrer les composantes principales (notés PC1_FF, PC2_FF, PC3_FF, PC4_FF et PC5_FF, contenant les coordonnées des individus sur les différents axes factoriels) aux potentielles variables explicatives.

Les figures suivantes (illustrations 7.7 à 7.9) présentent la corrélation entre la variable à prédire et les variables explicatives pour chacune des 2 phases de la sélection préalablement mentionnées.

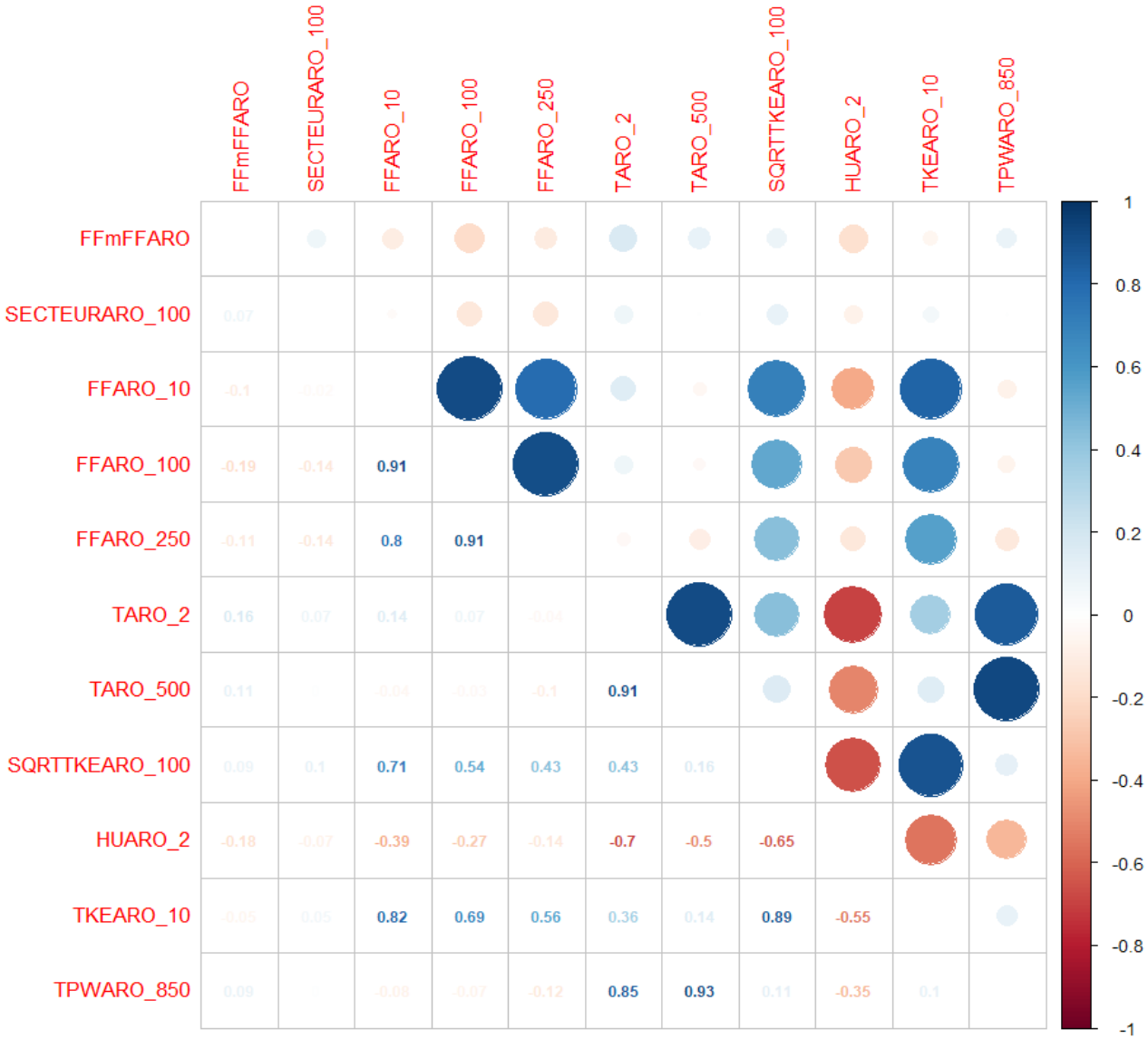

Illustration 7.7: Station de test 1 – Premier groupe (variables explicatives conservées pour la sélection finale) pour la prédiction de FF

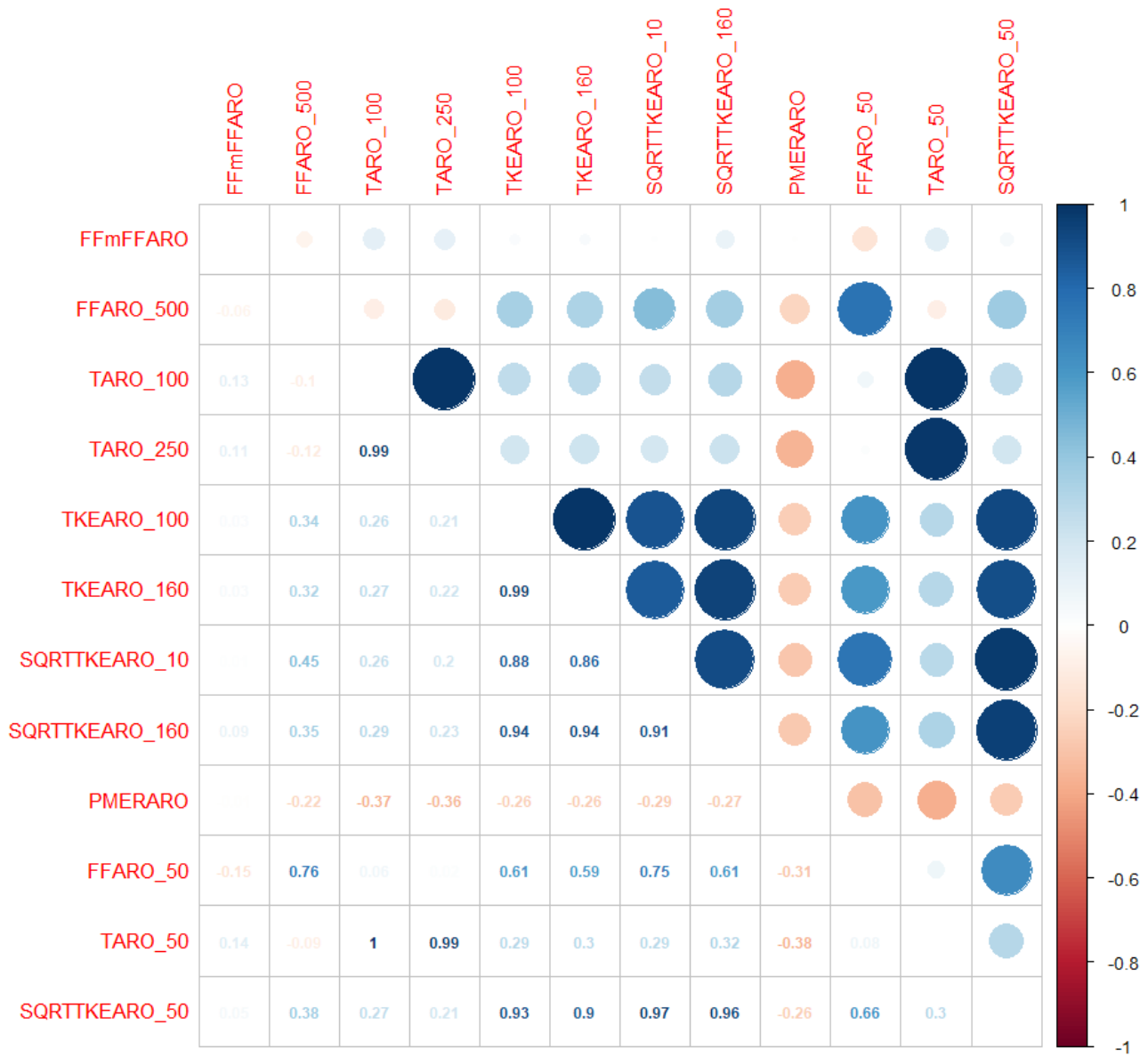

Illustration 7.8: Station de test 1 – Deuxième groupe (variables de l'ACP) pour la prédiction de FF

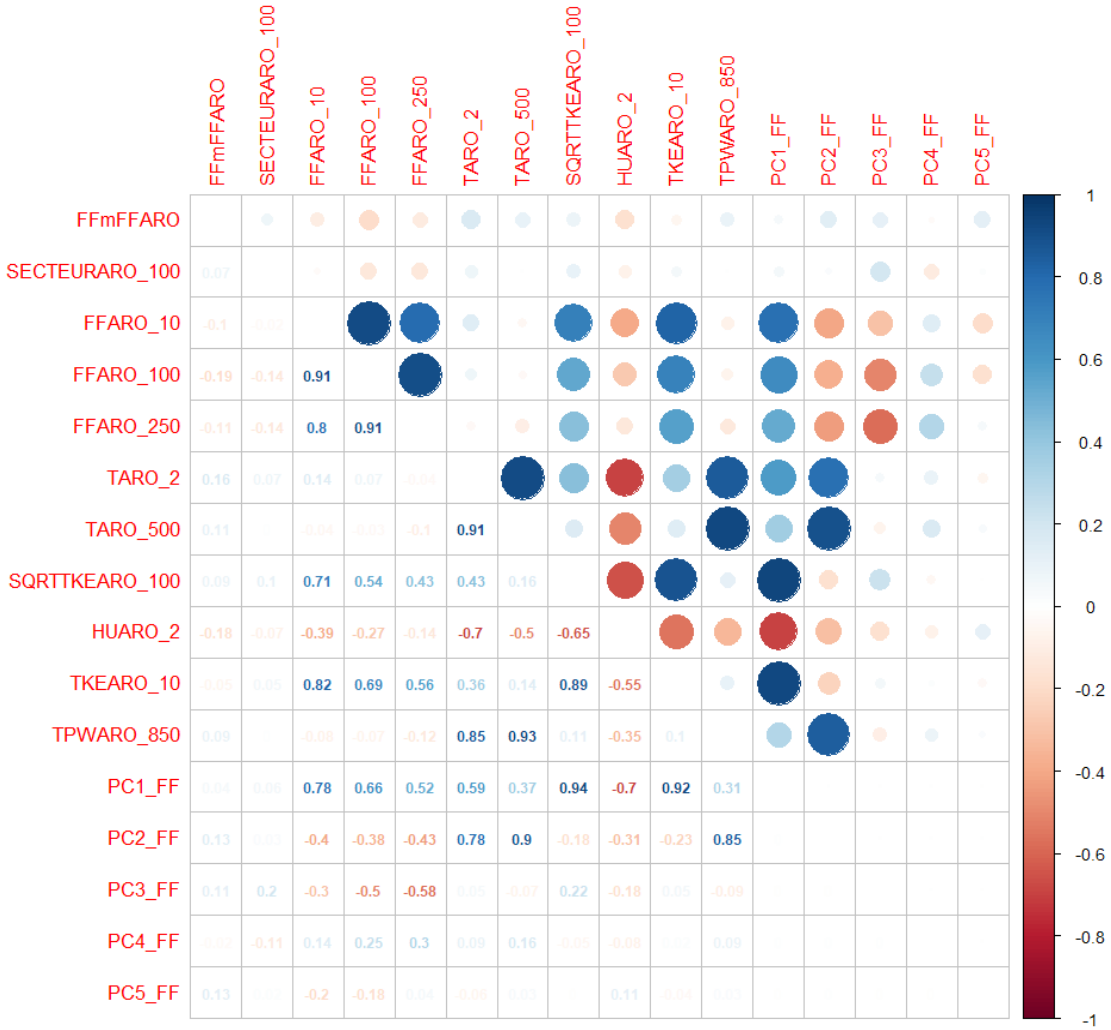

Illustration 7.9: Station de test 1 – Variables explicatives finales (avec les composantes principales de l'ACP) pour la prédiction de FF

Au final, on se retrouve avec 17 potentielles variables explicatives qui seront utilisées pour établir les modèles statistiques (modèles linéaires, arbre binaire, forêts aléatoires, réseaux de neurones). Il s'agit notamment de 12 variables issues ou calculées du modèle AROME et des 5 composantes principales de l'ACP.

Les potentielles variables explicatives de la station de test 1 sont : RegMM, RegHH_FF, SECTEURARO_100, FFARO_10, FFARO_100, FFARO_250, TARO_2, TARO_500, SQRTTKEARO_100, HUARO_2, TKEARO_10, TPWARO_850, PC1_FF, PC2_FF, PC3_FF, PC4_FF et PC5_FF.

Pour la seconde station, nous avons également 17 potentielles variables explicatives à l'issue de cette sélection : RegMM, RegHH_FF, SECTEURARO_100, FFARO_100, FFARO_250, TARO_2, TARO_50, TKEARO_100, SQRTTKEARO_10, HUARO_2, PMERARO, TPWARO_850, PC1_FF, PC2_FF, PC3_FF, PC4_FF, PC5_FF.

7.4.2 Échantillonnage

L'échantillonnage est très important dans le processus de validation d'un modèle statistique. Pour valider les modèles statistiques définis, nous utilisons la technique de **validation croisée** qui consiste à itérer l'estimation de l'erreur sur plusieurs échantillons de validation puis d'en calculer la moyenne.

Dans la définition des modèles statistiques, **l'année choisie pour l'apprentissage des données est 2017** pour la station de test 1. Le choix de cette année découle de l'analyse de la distribution du vent AROME en comparaison avec l'observation. L'année 2017 présente aussi beaucoup plus de données que les autres années.

Les données de cet échantillon (2017) sont ainsi séparées en 2 sous-échantillons :

- Un sous-échantillon d'apprentissage constitué par 70 % des données initiales pour ajuster les modèles statistiques. Pour le tirage de ces 70 %, les données initiales sont d'abord séparées en 4 groupes selon leurs appartenances aux 4 classes (hiver, printemps, été, automne) de la variable définissant le cycle saisonnier (RegMM). Ainsi, **70 % des données de chaque classe est tirée aléatoirement** (l'idée étant d'avoir les données d'hiver, du printemps, d'été et d'automne à parts égales dans chaque échantillonnage).
- Un sous-échantillon de test constitué par le reste des données non choisies dans le sous-échantillon d'apprentissage. Ces données sont donc constituées de 30 % des données de chaque classe (hiver, printemps, été, automne).

Le processus est itéré 20 fois afin d'avoir un sous-échantillon d'apprentissage et de test différent à chaque itération.

Il en est de même pour la station de test 2 sauf que l'apprentissage a été réalisé sur 2018 pour cette station.

7.4.3 Études des modèles statistiques

Les modèles statistiques sont tous ajustés sur le même échantillon et évalués par validation croisée. Les scores et critères de qualité intervenant dans la validation des modèles sont les suivants : RMSE, ECT, BIAIS, MAE, PSS, FA, BIC, les courbes de fiabilité QQ-Plot et les corrélations (*cf.* section 2.3.4 pour leur définition).

Pour PSS et FA, ces scores sont calculés à 3 niveaux de seuils : PSS1 et FA1 pour la détection des vents faibles ($FF>3m/s$), PSS2 et FA2 pour la détection des vents moyens ($FF>9m/s$), et PSS3 et FA3 pour la détection des vents forts ($FF>12m/s$).

Une fois la 4e étape de la méthode décrite à la section 2.3.2 appliquée à chaque modèle statistique, 4 nouvelles étapes sont mis en œuvre :

1. Comparaison des modèles optimaux (un par méthode) obtenus par estimation de l'erreur de prévision sur l'échantillon test.
2. Choix de la méthode retenue en fonction de ses capacités de prévision, de sa robustesse mais aussi, éventuellement, de l'interprétabilité du modèle obtenu.
3. Ré-estimation du modèle avec la méthode, le modèle et sa complexité optimisée à l'étape précédente sur les 1 an de données d'observation (au lieu de 70 % de ces données).
4. Exploitation du modèle sur la base complète et de nouvelles données.

On rappelle que **la variable à prédire est : Y =** $FFmFFARO$ **= FF_{obs} - FF_{ARO}** pour tous les modèles statistiques testés. On présente ci-après les résultats des tests d'apprentissages pour toutes nos expérimentations : arbres binaires, forêt aléatoire, modèle linéaires puis réseau de neurones.

7.4.3.1 Les arbres binaires de décision

Tout d'abord, il faut noter que l'arbre binaire ne permet pas d'avoir une prédiction continue de la variable à expliquer. Cependant, il permet surtout de comprendre plus ou moins le rôle de chacunes des variables explicatives ainsi que leur contribution en cascade des éléments de l'arbre. Cela permet donc d'avoir une appréciation supplémentaire concernant le choix des variables explicatives établit dans la phase exploratoire (cf § 7.4.1.2.2). Ce modèle a donc été établi sur l'échantillon tout entier (année d'apprentissage 2017 pour la station de test 1 et 2018 pour la station de test 2).

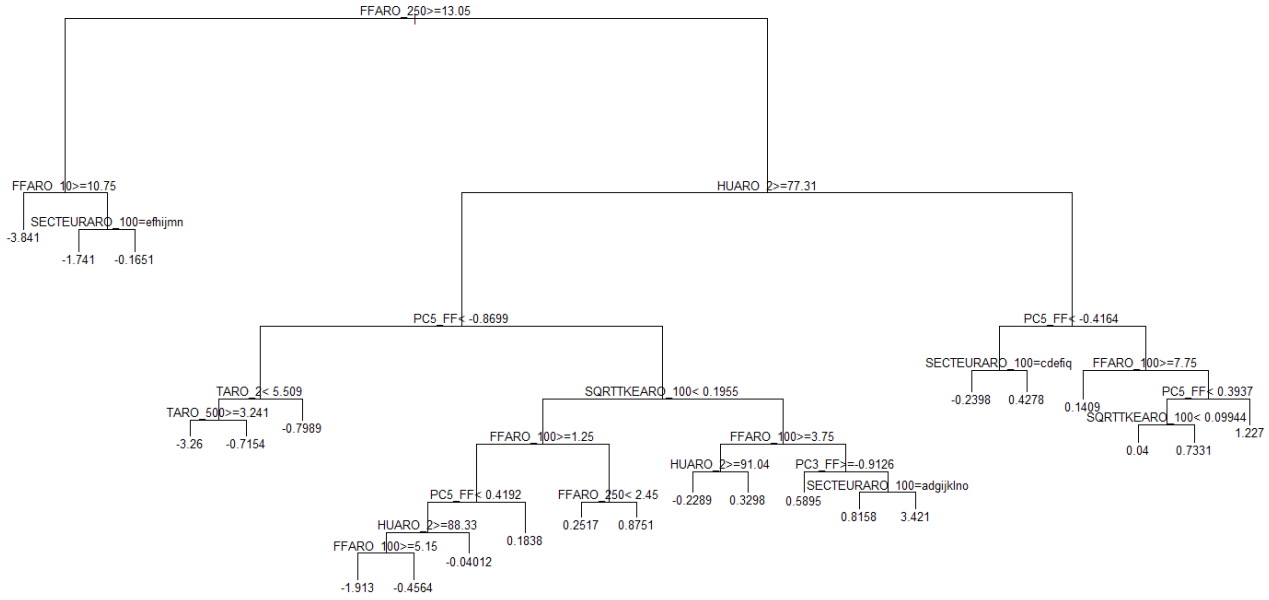

Illustration 7.10: Station de test 1 – FF –Modèle d'arbre de décision (arbre optimal)

Pour la définition du modèle, on part d'un arbre avec le plus de branches possibles, puis on cherche à optimiser sa taille et à l'élaguer par la suite. Le paramètre de complexité de l'arbre (cp) minimisant l'erreur estimée sur l'échantillon d'apprentissage vaut $cp = cp.optim = 0.003546972$. Une fois l'arbre élagué à travers l'utilisation de ce paramètre, le modèle final (optimal) est un arbre constitué de 22 nœuds comme le montre le graphique 7.10.

Les données se séparent à plusieurs niveaux de l'arbre autour de la force du vent et de la composante principales 5 de l'ACP. Ces variables ont donc une grande importance. Les variables génératrices de séparation des données dès le haut de l'arbre (tel que HUARO) ont également une grande importance.

L'indice d'importance fournit un classement des variables (de la plus importante à la moins importante). Il permet de ne retenir uniquement les variables explicatives permettant d'obtenir une meilleure prédiction. Le tableau 7.3 ci-dessous présente ainsi l'importance des variables (par ordre croissant d'importance de gauche à droite) fourni par le modèle pour la station de test 1.

Tableau 7.3: Station de test 1 – FF – Importance des variables explicatives pour l'arbre binaire de décision

Variable	RegMM	PC4_FF	PC2_FF	TARO_500	TPWARO_850	RegHH_FF	PC3_FF	SECTEURARO_100	TARO_2	TKEARO_10	SQRTTKEARO_100	HUARO_2	PC1_FF	PC5_FF	FFARO_10	FFARO_250	FFARO_100
Importance	15.92	60.07	89.58	140.33	153.23	197.13	211.1	220.98	364.02	495.85	537.15	602.28	604.48	622.29	656.89	1029.42	1084.74

Pour la station de test 2, la force du vent et la composante principales 5 de l'ACP (PC5_FF) s'illustrent à nouveau avec une plus grande importance pour le modèle. Cependant, pour cette station les autres variables d'importance sont la direction du vent SECTEURARO_100, et la turbulence AROME ainsi que sa racine carrée. Le tableau 7.4 présente l'importance des variables pour la station de test 2.

Tableau 7.4: Station de test 2 – FF – Importance des variables explicatives pour l'arbre binaire de décision

Variable	**HUARO_2**	**RegHH_FF**	**TPWARO_850**	**PC3_FF**	**TARO_50**	**PC2_FF**	**TARO_2**	**PMERARO**	**PC4_FF**	**RegMM**	**PC1_FF**	**TKEARO_100**	**SQRTTKEARO_10**	**FFARO_250**	**FFARO_100**	**SECTEURARO_100**	**PC5_FF**
Importance	17.98	35.99	221.48	266.12	322.17	326.4	344.15	374.84	455.26	827.25	836.21	929.64	1045.15	1079.75	1487.97	1744.34	2201.45

7.4.3.2 AROME brut et les modèles linéaires

Plusieurs modèles linéaires ont été testés afin de sélectionner par la suite celui ou ceux qui donnent les meilleures estimations.

Le modèle **brut** est défini par la force du vent AROME à 100 m noté **AROME.BRUT**.

Quant aux modèles linéaires, 7 modèles ont été définis (avec des niveaux de complexités différents) et ayant tous $FFmFFARO$ comme variable à prédire. C'est donc les variables explicatives qui diffèrent pour chaque modèle linéaire donné. Il s'agit notamment du LM0, GLM, GLM.STEP, GLM_PCA, GLM_PCA.STEP, GLM_MIXTE et GLM_MIXTE.STEP. Par exemple pour la station de test 1, on a :

- **Modèle nommé LM0 :** c'est le modèle linéaire simple qui a seulement la force du vent AROME comme prédicteur. On a donc **FFmFFARO = f(FFARO_100)**.
- **Modèle GLM :** c'est le modèle dont les prédicteurs sont ceux du premier groupe évoqué dans la phase exploratoire (§ 7.4.1.2.2 de la sélection des variables explicatives) associés aux variables de prise en compte de l'heure et du mois (RegHH_FF et RegMM). Ces 2 derniers interviennent dans tous les autres modèles linéaires par la suite. On a donc **FFmFFARO = f(RegHH_FF, RegMM, SECTEURARO_100, FFARO_10, FFARO_100, FFARO_250, TARO_2, TARO_500, SQRTTKEARO_100, HUARO_2, TKEARO_10, TPWARO_850)^2**. Le carré (^2) dans l'expression désigne l'interaction 2 à 2 des prédicteurs.
- **Modèle GLM.STEP :** c'est la modèle établit à partir de la sélection automatique des interactions significatives des prédicteurs du modèle GLM ci-dessus.
- **Modèle GLM_PCA :** c'est le modèle dont les prédicteurs sont uniquement les composantes principales de l'ACP associées aux variables de prise en compte de l'heure et du mois. On a **FFmFFARO = f(RegHH_FF, RegMM, PC1_FF, PC2_FF, PC3_FF, PC4_FF, PC5_FF)^2**.
- **Modèle GLM_PCA.STEP :** comme pour le GLM, il s'agit du modèle établit à partir de la sélection automatique des interactions significatives des prédicteurs de GLM_PCA.
- **Modèle nommé GLM_MIXTE :** c'est le modèle ayant comme prédicteurs ceux du modèle GLM plus les composantes principales de l'ACP. **FFmFFARO = f(RegHH_FF, RegMM, SECTEURARO_100, FFARO_10, FFARO_100, FFARO_250, TARO_2, TARO_500, SQRTTKEARO_100, HUARO_2, TKEARO_10, TPWARO_850, PC1_FF, PC2_FF, PC3_FF, PC4_FF, PC5_FF)^2**.
- **Modèle GLM_MIXTE.STEP :** modèle issu de la sélection automatique des interactions significatives des prédicteurs de GLM_MIXTE.

Les deux modèles GLM_PCA et GLM_PCA.STEP ont été introduits non pas pour être choisi comme modèle d'extension des séries horaires, mais pour constater l'apport d'information des variables issues de l'ACP (notamment par rapport à AROME.BRUT).

Les graphiques 7.11 à 7.14 montrent respectivement les scores de qualités (RMSE, ECT, BIAIS, MAE, PSS et FA) des différents modèles (évalués par validation croisée) sur l'échantillon d'apprentissage et de test. Sur chacun des graphiques, les modèles sont positionnés dans cet ordre : LM0, GLM, GLM.STEP, GLM_PCA, GLM_PCA.STEP, GLM_MIXTE et GLM_MIXTE.STEP.

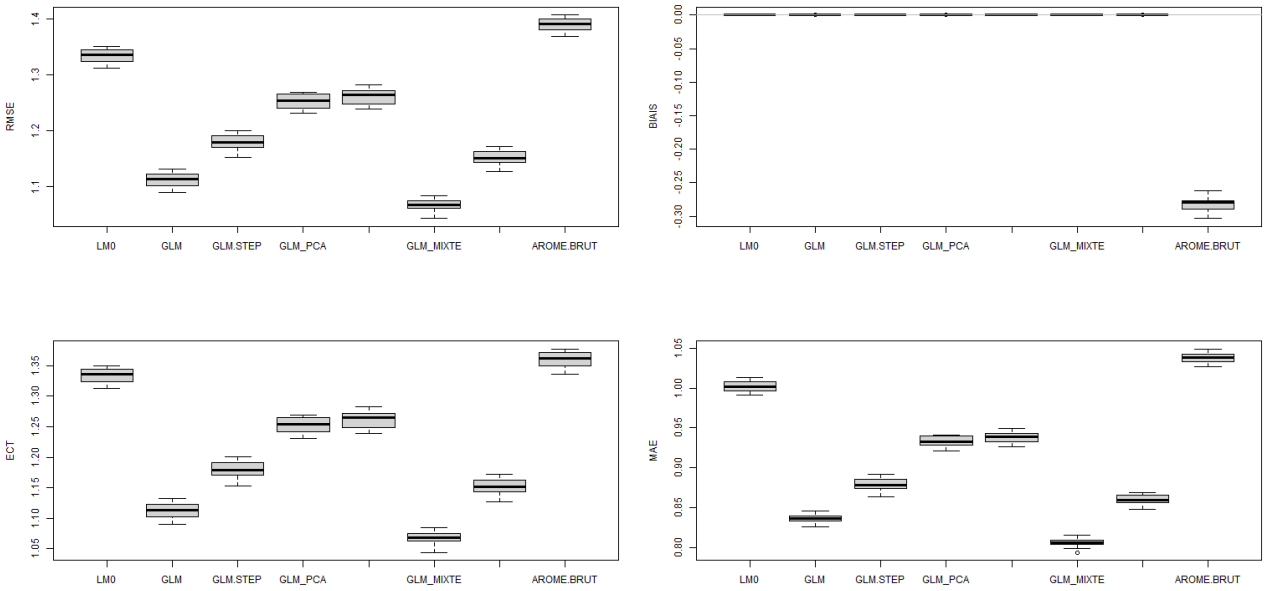

Illustration 7.11: Station de test 1 – FF – Modèles linéaires et AROME.BRUT - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'échantillon d'apprentissage

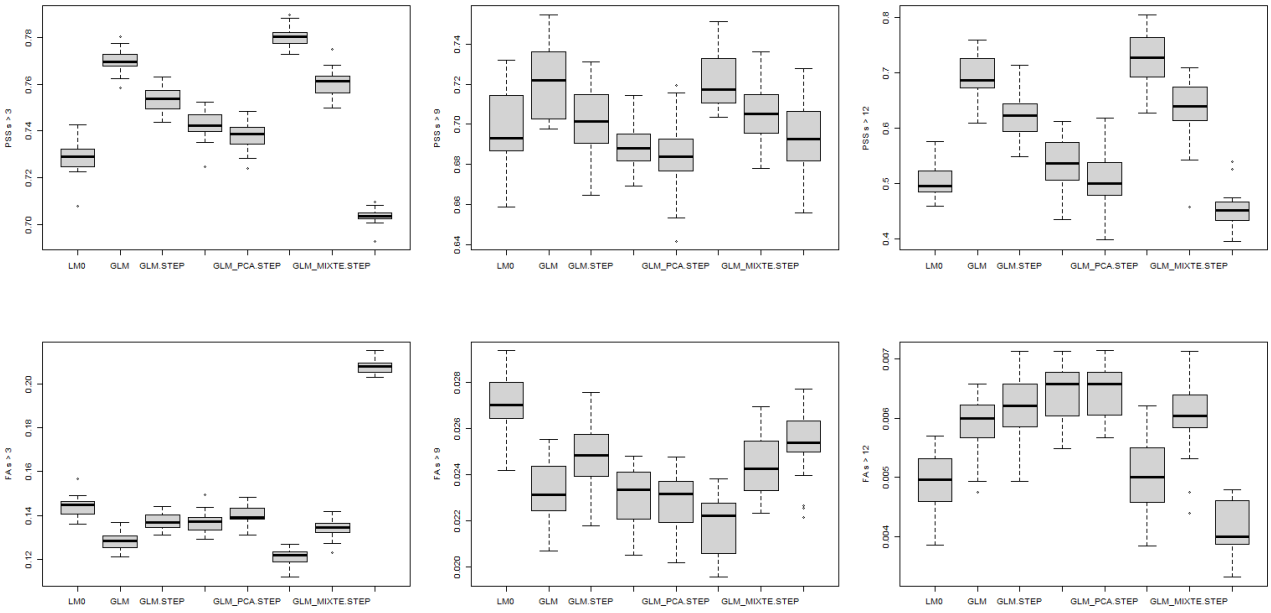

Illustration 7.12: Station de test 1 – FF – Modèles linéaires et AROME.BRUT - Box-plot PSS (première ligne pss > 3 m/s, pss > 9 m/s et pss > 12 m/s noté respectivement pss1, pss2 et pss3) et FA (deuxième ligne fa > 3 m/s, fa > 9 m/s et fa > 12 m/s notée respectivement fa1, fa2 et fa3) pour l'échantillon d'apprentissage

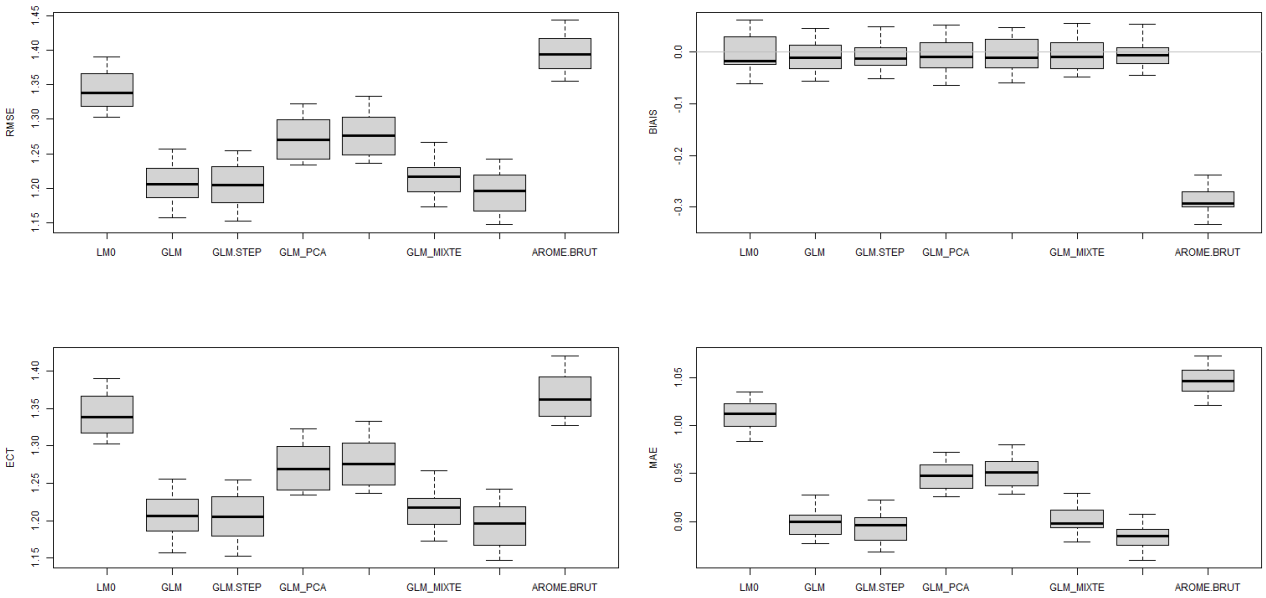

Illustration 7.13: Station de test 1 – FF – Modèles linéaires et AROME.BRUT - Box-plot RMSE, ECT, BIAIS et MAE pour l'échantillon de test

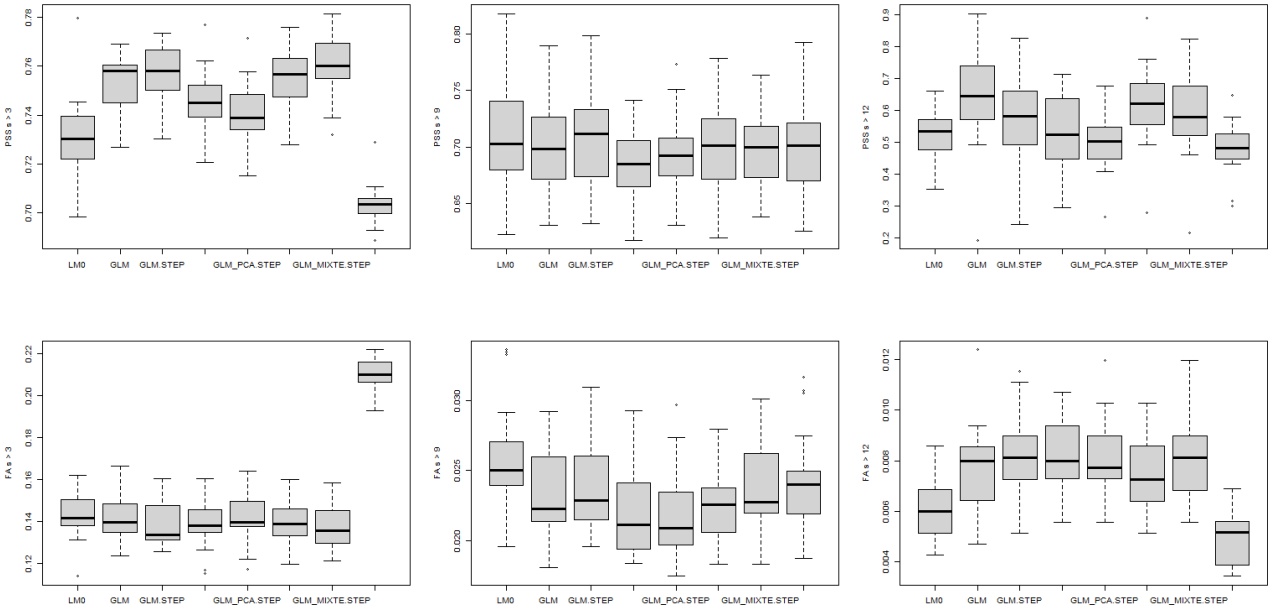

Illustration 7.14: Station de test 1 – FF – Modèles linéaires et AROME.BRUT - Box-plot PSS et FA pour l'échantillon de test

Dans tous les cas (apprentissage, test) les modèles linéaires améliorent AROME.BRUT. Une amélioration plus prononcée pour les modèles GLM, GLM_MIXTE et leur STEP que pour le modèle linéaire simple et les modèles avec ACP.

Le tableau 7.5 récapitule les **scores de validation croisées sur l'échantillon d'apprentissage.**

Tableau 7.5: Station de test 1 – FF – Modèles linéaires et AROME.BRUT - Scores de validation croisée sur l'échantillon d'apprentissage (en bleu les modèles avec les meilleurs scores)

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS2	FA2	PSS3	FA3
LM0	1.334	1.334	0	1.002	0.728	0.144	0.698	0.027	0.501	0.005
GLM	**1.113**	**1.113**	**0**	**0.836**	**0.77**	**0.128**	**0.722**	**0.023**	**0.692**	**0.006**
GLM.STEP	1.179	1.179	0	0.879	0.754	0.137	0.702	0.025	0.623	0.006
GLM_PCA	1.253	1.253	0	0.933	0.743	0.137	0.688	0.023	0.535	0.006
GLM_PCA.STEP	1.262	1.262	0	0.939	0.738	0.14	0.685	0.023	0.5	0.006
GLM_MIXTE	**1.068**	**1.068**	**0**	**0.806**	**0.781**	**0.121**	**0.721**	**0.022**	**0.729**	**0.005**
GLM_MIXTE.STEP	1.151	1.151	0	0.86	0.761	0.134	0.705	0.024	0.631	0.006
AROME.BRUT	1.389	1.36	-0.282	1.038	0.704	0.208	0.694	0.025	0.452	0.004

Les scores d'apprentissage sont plutôt favorables aux modèles GLM et GLM_MIXTE avec un léger avantage pour le GLM_MIXTE. Cependant pour l'apprentissage, les scores sont appliqués à l'échantillon qui a été utilisé pour construire les modèles. Il est donc important de confronter ces scores à ceux des échantillons tests (les modèles n'ayant pas encore vu les données de test).

En examinant les scores de test des illustrations 7.13 et 7.14, l'avantage revient aux 2 modèles GLM et GLM_MIXTE.STEP. Mais il est très difficile de les départager visuellement avec les graphiques (les scores étant du même ordre de grandeur). On note que le GLM_MIXTE.STEP possède un meilleur score que le GLM pour la détection des vents faibles (PSS et FA supérieur à 3 m/s). D'un autre côté, il est moins bon que le GLM pour la détection de vent fort (PSS et FA supérieur à 12 m/s). Ainsi, ces critères ne permettent pas de les départager. La différence va donc se faire au niveau des scores BIC et des scores de corrélation (pearson, kendall et spearman).

Le tableau 7.6 présente les **scores par validation croisée sur l'échantillon de test**, le critère BIC pour chaque modèle linéaire ainsi que les corrélations.

Tableau 7.6: Station de test 1 – Modèles linéaires et AROME.BRUT - Scores de validation croisée sur l'échantillon de test (en bleu et vert, les modèles avec les meilleurs scores)

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS2	FA2	PSS3	FA3	BIC	Cor Pearson	Cor Kendall	Cor Spearman
AROME.BRUT	1.396	1.366	-0.288	1.047	0.703	0.21	0.699	0.024	0.482	0.005		0.88	0.7	0.87
LM0	1.342	1.341	-0.007	1.011	0.731	0.143	0.705	0.026	0.53	0.006	19614	0.88	0.7	0.87
GLM	**1.208**	**1.207**	**-0.01**	**0.9**	**0.754**	**0.142**	**0.7**	**0.023**	**0.644**	**0.008**	**19614**	**0.91**	**0.73**	**0.9**
GLM.STEP	1.205	1.205	-0.008	0.894	0.756	0.14	0.708	0.024	0.582	0.008	17673	0.91	0.73	0.9
GLM_PCA	1.272	1.272	-0.005	0.948	0.745	0.138	0.686	0.022	0.531	0.008	18465	0.9	0.72	0.89
GLM_PCA.STEP	1.278	1.278	-0.005	0.951	0.74	0.141	0.692	0.022	0.503	0.008	18337	0.89	0.71	0.89
GLM_MIXTE	1.217	1.217	-0.007	0.902	0.756	0.14	0.701	0.023	0.621	0.007	20573	0.9	0.73	0.9
GLM_MIXTE.STEP	**1.192**	**1.192**	**-0.004**	**0.884**	**0.76**	**0.138**	**0.7**	**0.024**	**0.585**	**0.008**	**17673**	**0.91**	**0.74**	**0.91**

L'analyse du tableau des scores nous permet de dire que le modèle GLM_MIXTE.STEP est le meilleur des modèles linéaires. En effet, il a un meilleur score BIC que le GLM, les scores de corrélation étant quasiment identiques pour les 2 modèles. On rappelle que le score BIC est le critère d'information bayésien. Son calcul dépend de la taille de l'échantillon et du nombre de paramètre du modèle statistique. Plus il est petit, meilleur est le modèle.

Cependant, on garde en tête que le modèle GLM_MIXTE.STEP a un score de détection des vents forts (FF>12 m/s) un peu moins bon que le GLM. Par conséquent **à ce stade, on garde les 2 modèles GLM_MIXTE.STEP et GLM** en attendant la comparaison inter-modèles où l'analyse des courbes de fiabilité QQ-Plot et la restitution des cycles annuels et diurnes pourraient faire la différence entre les modèles.

La même méthode a été appliquée pour la station de test 2 et nous avons obtenu des résultats similaires, c'est-à-dire les deux modèles GLM et GLM_MIXTE.STEP donnent des scores meilleurs dans l'ensemble que les autres modèles linéaires.

7.4.3.3 Les forêts aléatoires

Plusieurs modèles de forêts aléatoires ont été testés avec différents niveaux d'arbres. Pour choisir le modèle optimal, on optimise le nombre d'arbre dans la forêt de sorte que les estimations du modèle soient optimales, tout en restant un modèle parcimonieux.

Pour ces modèles, nous utilisons l'ensemble des 17 variables explicatives (voir section 7.4.1.2.2).

Les figures suivantes (illustrations 7.15 à 7.18) présentent les scores de validation croisée (apprentissage, test) des modèles avec les différents niveaux d'arbres.

Les modèles RFA, RFB, RFC, RFD et RFE correspondent respectivement aux modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres.

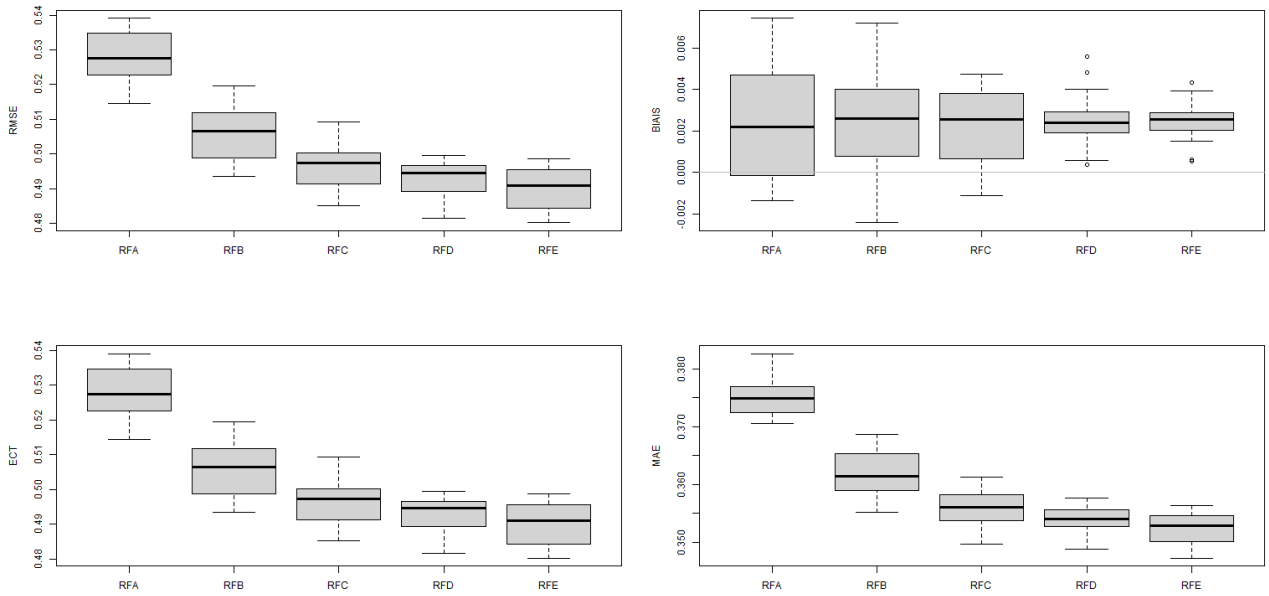

Illustration 7.15: Station de test 1 – FF – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'échantillon d'apprentissage

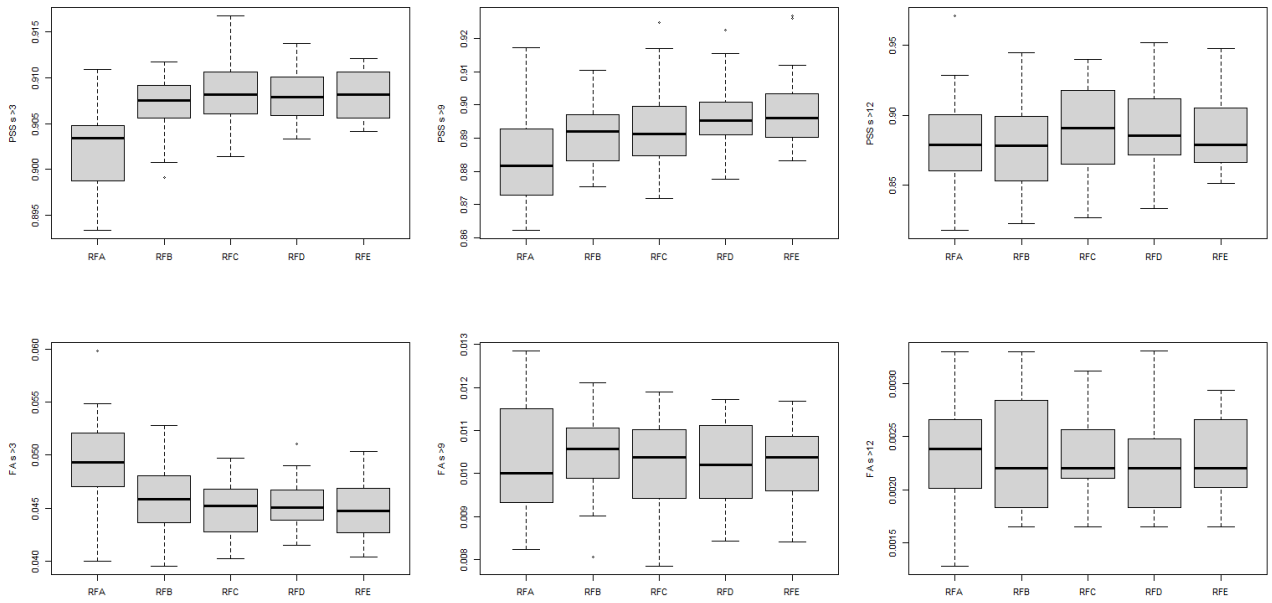

Illustration 7.16: Station de test 1 – FF – Modèles de forêt aléatoire à 20, 50, 100, 200 et 500 arbres - Box-plot PSS (première ligne pss > 3 m/s, pss > 9 m/s et pss > 12 m/s) et FA (deuxième ligne fa > 3 m/s, fa > 9 m/s et fa > 12 m/s)pour l'échantillon d'apprentissage

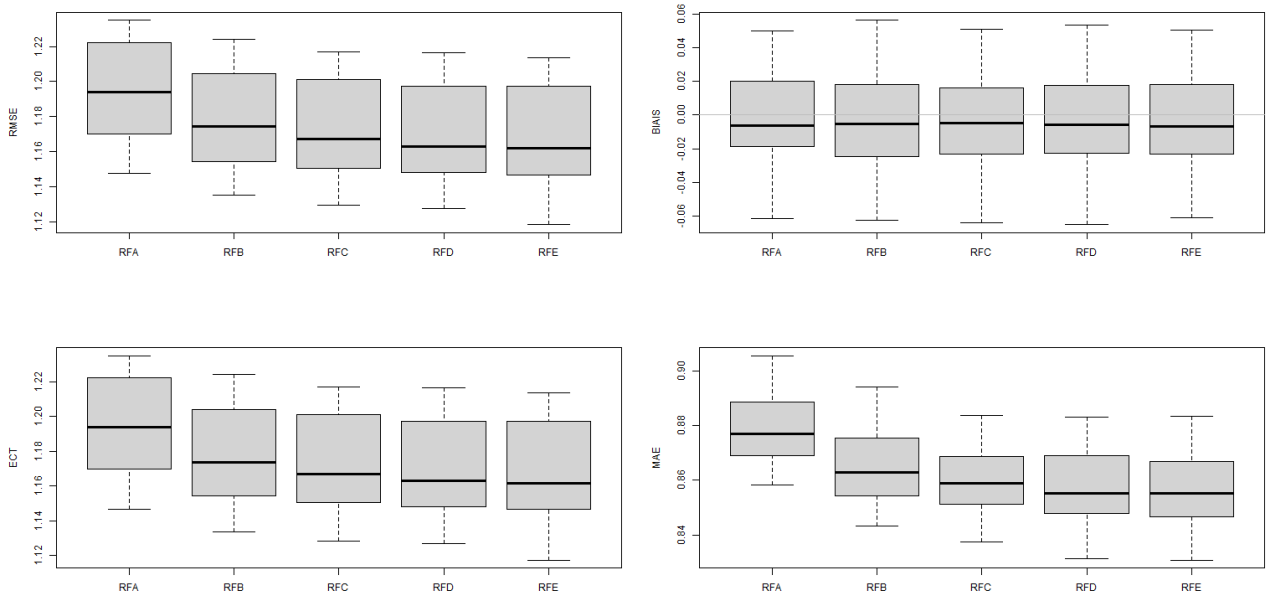

Illustration 7.17: Station de test 1 – FF – Modèles de forêt aléatoire - Box-plot RMSE, ECT, BIAIS et MAE pour l'échantillon de test

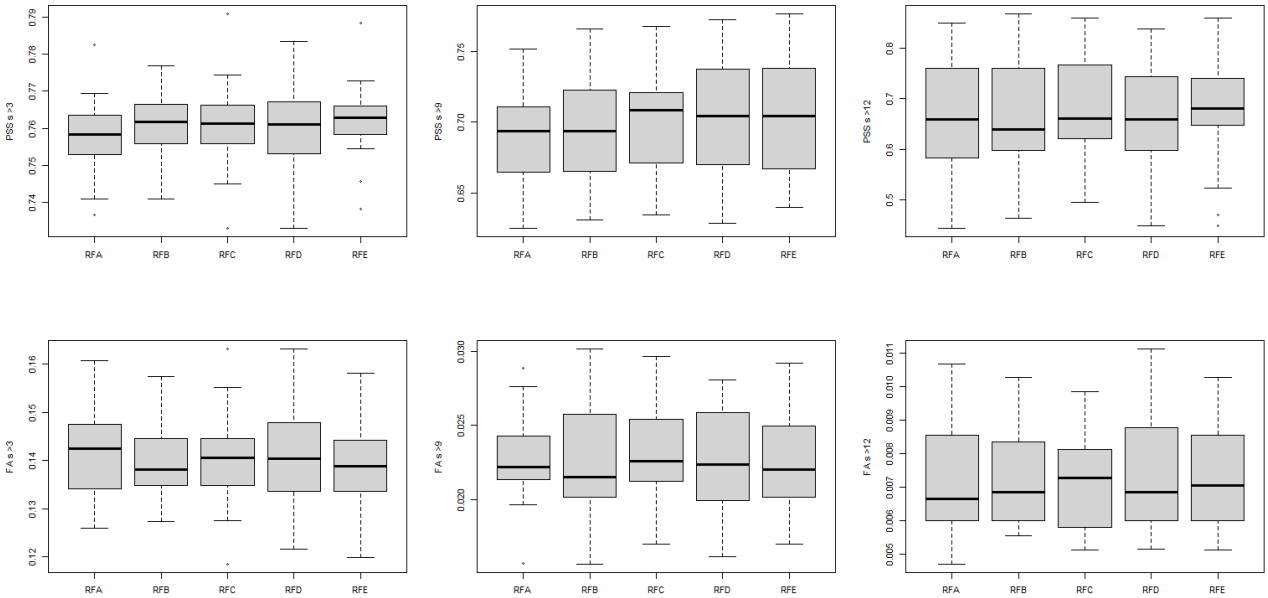

Illustration 7.18: Station de test 1 – FF – Modèles de forêt aléatoire - Box-plot PSS et FA pour l'échantillon de test

Les scores des modèles RFC, RFD et RFE sont très proches les unes des autres (aussi bien pour l'apprentissage que pour le test). Ils sont meilleurs que ceux des 2 autres modèles (RFA et RFB qui ont un nombre inférieur d'arbre). Sachant que l'évolution des scores est très minime à partir du modèle RFC (100 arbres) malgré l'augmentation importante du nombre d'arbre, **nous avons décidé de conserver le modèle RFC soit 100 arbres pour le choix du modèle de forêt aléatoire**. Les tableaux 7.7 et 7.8 présentent le récapitulatif des scores de validation croisée sur l'échantillon d'apprentissage puis de test pour la station de test 1.

Tableau 7.7: Station de test 1 – FF – Modèles de forêt aléatoire - Scores de validation croisée sur l'échantillon d'apprentissage (en vert, le modèle choisi)

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS2	FA2	PSS3	FA3
RFA	0.528	0.528	0.002	0.375	0.902	0.05	0.883	0.01	0.883	0.002
RFB	0.506	0.506	0.003	0.362	0.907	0.046	0.891	0.01	0.881	0.002
RFC	**0.497**	**0.497**	**0.002**	**0.356**	**0.908**	**0.045**	**0.893**	**0.01**	**0.892**	**0.002**
RFD	0.493	0.493	0.002	0.354	0.908	0.045	0.897	0.01	0.889	0.002
RFE	0.49	0.49	0.002	0.352	0.908	0.045	0.899	0.01	0.888	0.002

Tableau 7.8: Station de test 1 – FF – Modèles de forêt aléatoire - Scores de validation croisée sur l'échantillon de test (en vert, le modèle choisi)

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS2	FA2	PSS3	FA3
RFA (20 arbres)	1.194	1.194	-0.002	0.879	0.758	0.142	0.689	0.023	0.656	0.007
RFB (50 arbres)	1.177	1.177	-0.002	0.865	0.761	0.14	0.697	0.023	0.657	0.007
RFC (100 arbres)	**1.173**	**1.172**	**-0.003**	**0.86**	**0.761**	**0.14**	**0.701**	**0.023**	**0.676**	**0.007**
RFD (200 arbres)	1.169	1.169	-0.003	0.858	0.761	0.14	0.704	0.023	0.654	0.007
RFE (500 arbres)	1.168	1.167	-0.003	0.856	0.762	0.139	0.706	0.023	0.673	0.007

Pour les deux stations (test 1 et test 2), le modèle sélectionné comme optimal est une forêt à 100. Pour ce modèle, l'importance des variables est donné par le tableau 7.9 pour la station de test 1. C'est la moyenne de l'importance des variables pour chacun des 20 échantillonnages.

Tableau 7.9 : Station de test 1 – FF –
Importance des variables explicatives pour le modèle de forêt aléatoire avec 100 arbres

Variable	RegHH_FF	RegMM	TARO_500	TPWARO_850	PC2_FF	TKEARO_10	FFARO_10	PC1_FF	TARO_2	PC3_FF	PC4_FF	SQRTTKEARO_100	HUARO_2	FFARO_250	SECTEURARO_100	PC5_FF	FFARO_100
Importance	71.51	161.47	436.73	449.7	459.29	490.03	510.67	511.67	554.37	609.32	624.89	625.21	730.94	754.4	810.83	933.45	965.36

Comme pour l'arbre binaire de décision, la force du vent AROME, la composante principale 5 de l'ACP ainsi que HUARO_2 ont une importance plus élevée que les autres variables explicatives pour ce modèle. En plus de ces variables, c'est le SECTEURARO_100 qui s'illustre comme variable ayant une grande importance pour le modèle de forêt aléatoire.

7.4.3.3.1 Comparaison du modèle de forêt aléatoire avec et sans ACP

Nous avons également testé les modèles de forêt aléatoire avec les variables explicatives sans les composantes principales de l'ACP, c'est-à-dire avec les variables explicatives suivantes de la station de test 1 : RegMM, RegHH_FF, SECTEURARO_100, FFARO_10, FFARO_100, FFARO_250, TARO_2, TARO_500, SQRTTKEARO_100, HUARO_2, TKEARO_10, TPWARO_850.

Les graphiques des illustrations 7.19 à 7.22, ainsi que les tableaux 7.10 et 7.11 présentent les scores de validation croisée pour le modèle avec ACP (RF_100) et le modèle sans ACP (noté RF_100.SANS.ACP).

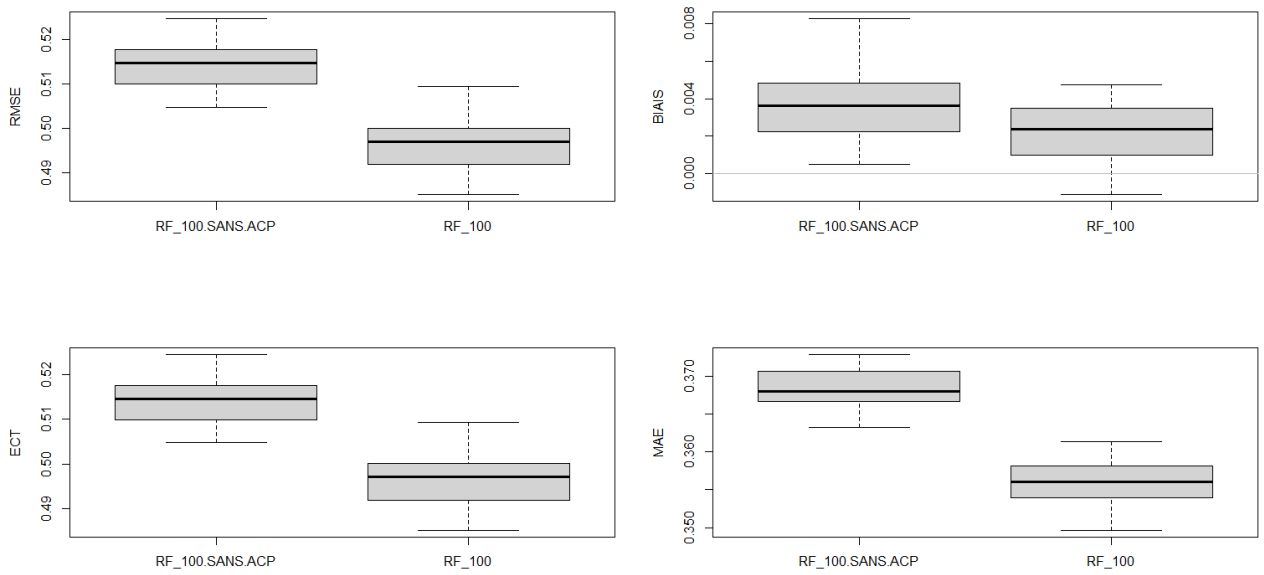

Illustration 7.19: Station de test 1 – FF – Modèles de forêt aléatoire à 100 arbres avec et sans ACP - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'échantillon d'apprentissage

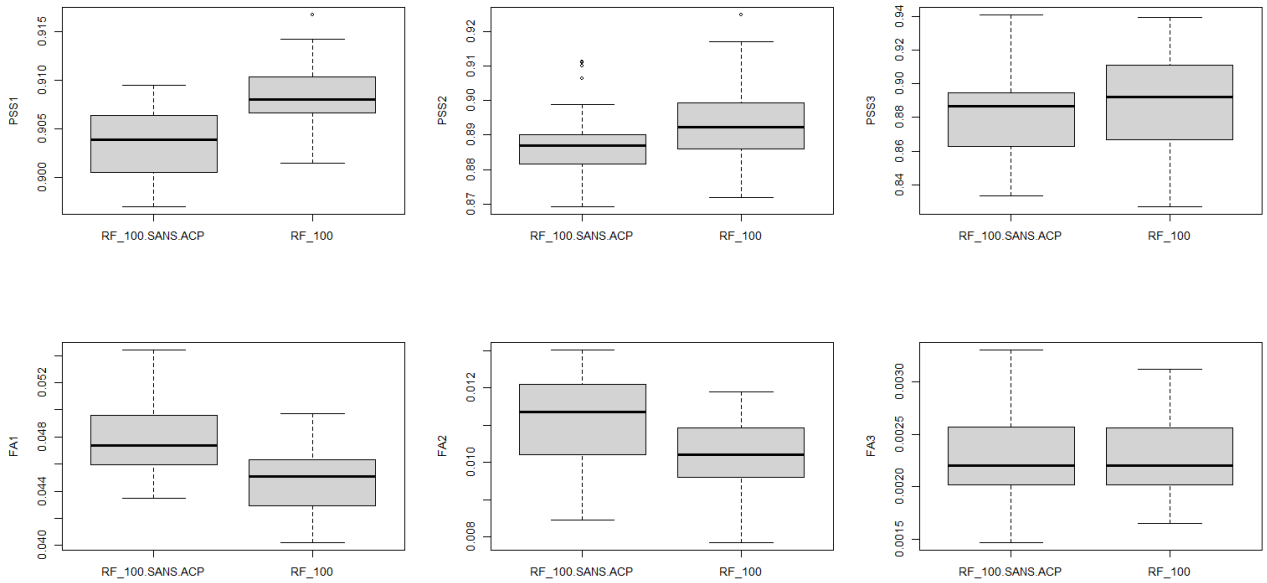

Illustration 7.20: Station de test 1 – FF – Modèles de forêt aléatoire à 100 arbres avec et sans ACP - Box-plot PSS (première ligne pss > 3 m/s, pss > 9 m/s et pss > 12 m/s) et FA (deuxième ligne fa > 3 m/s, fa > 9 m/s et fa > 12 m/s) pour l'échantillon d'apprentissage

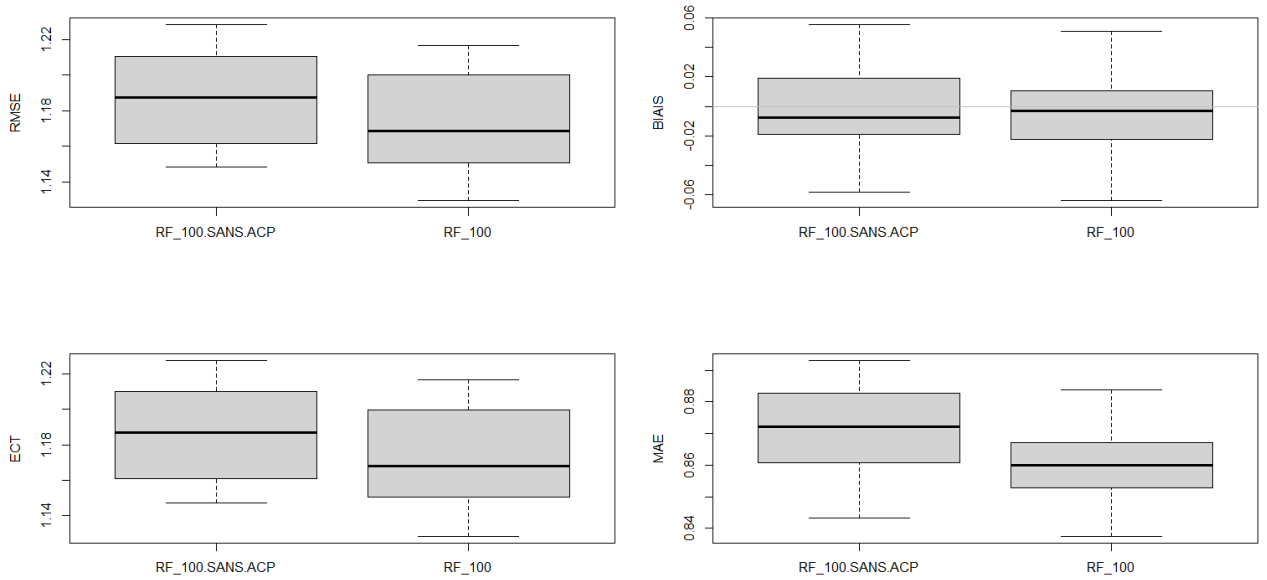

Illustration 7.21: Station de test 1 – FF – Modèles de forêt aléatoire à 100 arbres avec et sans ACP - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'échantillon de test

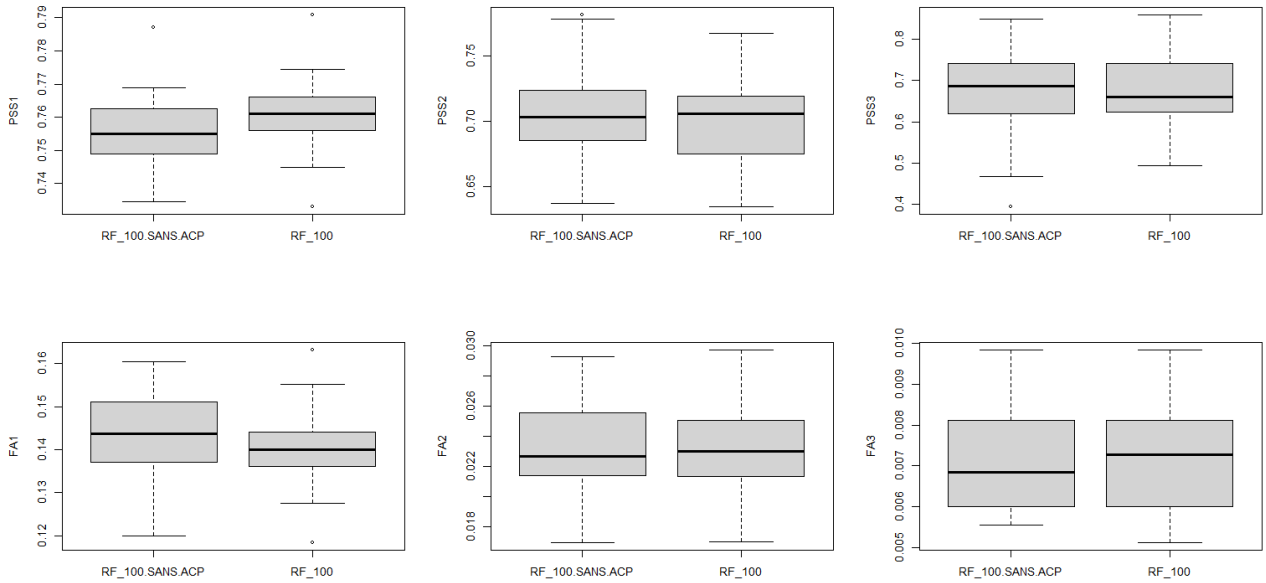

Illustration 7.22: Station de test 1 – FF – Modèles de forêt aléatoire à 100 arbres avec et sans ACP - Box-plot PSS (première ligne pss > 3 m/s, pss > 9 m/s et pss > 12 m/s) et FA (deuxième ligne fa > 3 m/s, fa > 9 m/s et fa > 12 m/s) pour l'échantillon de test

Tableau 7.10: Station de test 1 – FF – Modèles de forêt aléatoire avec et sans ACP – Scores de validation croisée sur l'échantillon d'apprentissage (en vert, le modèle choisi)

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS2	FA2	PSS3	FA3
RF_100	**0.497**	**0.497**	**0.002**	**0.356**	**0.908**	**0.045**	**0.893**	**0.01**	**0.892**	**0.002**
RF_100.SANS.ACP	0.514	0.514	0.004	0.368	0.904	0.048	0.89	0.011	0.883	0.002

Tableau 7.11: Station de test 1 – FF – Modèles de forêt aléatoire avec et sans ACP – Scores de validation croisée sur l'échantillon de test (en vert, le modèle choisi)

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS2	FA2	PSS3	FA3
RF_100	**1.173**	**1.172**	**-0.003**	**0.86**	**0.761**	**0.14**	**0.701**	**0.023**	**0.676**	**0.007**
RF_100.SANS.ACP	1.188	1.187	-0.001	0.871	0.756	0.143	0.703	0.023	0.671	0.007

L'interprétation des différents scores de la station de test 1 présentée ci-dessus montrent que les modèles sans ACP sont légèrement moins bons. **Nous avons donc décidé de conserver comme meilleur des modèles de forêt aléatoires, le modèle RFC à 100 arbres (noté RF_100) qui possèdent les composantes principales de l'ACP comme variables explicatives.**

Pour la station de test 2, le même protocole a été appliqué, et c'est aussi le modèle de forêt à 100 arbres qui a été considéré comme meilleur.

7.4.3.4 Les réseaux de neurones

Pour ce nouveau modèle statistique, les variables explicatives sont les mêmes que celles utilisées pour les modèles précédents. Il y a cependant, un traitement préalable à réaliser avant de soumettre les données aux réseaux de neurones. Il s'agit de la **standardisation** des données. En effet, les réseaux de neurones sont très susceptibles aux données ayant des variances élevées. La standardisation permet de ramener la variance de toutes les données à une valeur égale à 1.

D'autre part, les données catégorielles (notamment RegHH_FF, RegMM et SECTEURARO_100) sont aussi soumises à un traitement préalable : **l'encodage**. Il faut en effet les encoder sous un format numérique

reconnaissable par le réseau de neurone. Pour cela, nous utilisons le module OneHotEncoder de scikit-learn, qui code les variables catégorielles sous forme de vecteurs binaires (constitués de 0 et 1).

7.4.3.4.1 Configuration du modèle retenu

Le processus d'entraînement du modèle nous a conduit à utiliser plusieurs configurations du réseau avant de trouver le modèle optimal (moins complexe, avec des prédictions satisfaisantes). Le tableau 7.12 montre les grandes étapes des différentes configurations qui ont été testées :

Tableau 7.12: Station de test 1 – FF – Tableau récapitulatif du processus d'entraînement du réseau

Optimizer	NB Nœud C1	NB Nœud C2	Activation C1	Activation C2	Activation Sortie	Learnig Rate	Epoch	RMSE App	RMSE Test
SGD	38	2*38	sans activation	sans activation	sans activation	0.0013	570	1.175	1.258
SGD	38	2*38	sans activation	sans activation	sans activation	0.013	570	1.174	1.261
SGD	10	2	sans activation	sans activation	sans activation	0.013	570	1.172	1.260
SGD	2*38	2	relu	relu	relu	0.000013	3000	1.327	1.369
SGD	10	2	relu	relu	relu	0.013	570	1.274	1.327
SGD	10	2	softplus	softplus	softplus	0.0013	870	1.320	1.362
SGD	10	2	softplus	softplus	softplus	0.013	570	1.295	1.333
SGD	10	2	elu	elu	elu	0.013	570	1.140	1.233
SGD	10	2	selu	selu	selu	0.013	570	1.120	1.237
SGD	10	2	tanh	tanh	tanh	0.013	570	1.136	1.231
SGD	10	2	tanh	tanh	softplus	0.013	570	1.283	1.331
SGD	2*38	1*38	tanh	tanh	selu	0.013	570	1.065	1.222
SGD	10	2	tanh	tanh	selu	0.013	570	1.112	1.221
SGD	**10**	**2**	**tanh**	**tanh**	**selu**	**0.013**	**870**	**1.100**	**1.217**

Les colonnes NB Nœud C1 et NB Nœud C2 correspondent respectivement au nombre de neurone dans la première et la deuxième couche cachée. RMSE App et RMSE Test désignent respectivement le RMSE sur le jeu de données d'apprentissage et de test par validation croisée.

À la suite du processus d'entraînement du modèle, la configuration suivante a été adoptée pour le choix du réseau final :

- Le nombre de couche intermédiaire (cachée) est égal à 2 couches Dense, avec 10 neurones dans la première couche cachée et 2 neurones dans la deuxième couche cachée. Une couche Dense est une couche de neurones classique, complètement connectée avec la couche précédente et la couche suivante.
- La fonction d'activation tangente hyperbolique

$$\tanh(x)=(\exp(x)-\exp(-x))/(\exp(x)+\exp(-x))$$

  a été utilisée pour la couche d'entrée et les 2 couches cachées, et la fonction d'activation *selu* pour la couche de sortie.

  La fonction *selu* (Scaled Exponential Linear Unit) est définie de la façon suivante :

  $selu(x) = \alpha * x$ si $x>0$

  $selu(x) = \alpha * \beta * (\exp(x)-1)$ si $x<0$

  avec α=1.05070098, et β=1.67326324.

- Les Optimizer SGD, Adam et RAdam ont tous été utilisés à leur tour pour réaliser la descente du gradient avec le paramétrage ci-dessus du modèle. Pour chacun des Optimizers la valeur du taux d'apprentissage (learning rate) conservée vaut 0.013.
- Les autres paramètres d'entraînement du modèle ont été définies comme suit :

- le nombre d'Epochs conservé à 870,

- le Batch Size conservé à 380.

Le graphique de l'illustration 7.23 présente l'architecture du réseau final conservé.

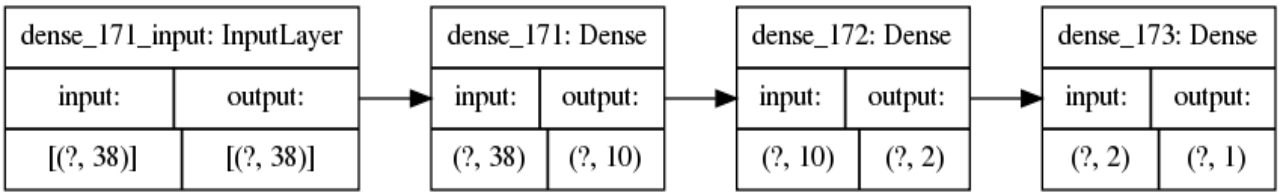

Illustration 7.23: Station de test 1 – Architecture du réseau conservé pour la modélisation de FF

Sur cette figure, la couche d'entrée reçoit 38 variables. Le nombre de variable explicative (17) n'a pas changé. Il s'agit en effet de 14 variables numériques (paramètres AROME) et des 3 variables catégorielles (RegHH_FF, RegMM et SECTEURARO_100) qui ont été encodé sous forme de vecteurs binaires. C'est cet encodage qui transforme les 3 trois variables catégorielles en 24 vecteurs binaires, conduisant ainsi le nombre de données d'entrée du réseau à 38 variables.

Comme évoqué ci-dessus, le nombre de neurone de la couche d'entrée est toujours égal au nombre de données (variables) d'entrées. Ensuite, il faut regarder l'output de chaque couche qui correspond au nombre de neurone dans cette couche.

7.4.3.4.2 Courbes d'entraînement du modèle de réseau de neurones

Les illustrations 7.24, 7.25 et 7.26 montrent les courbes d'entraînement du modèle sur 1 des 20 échantillons d'apprentissage.

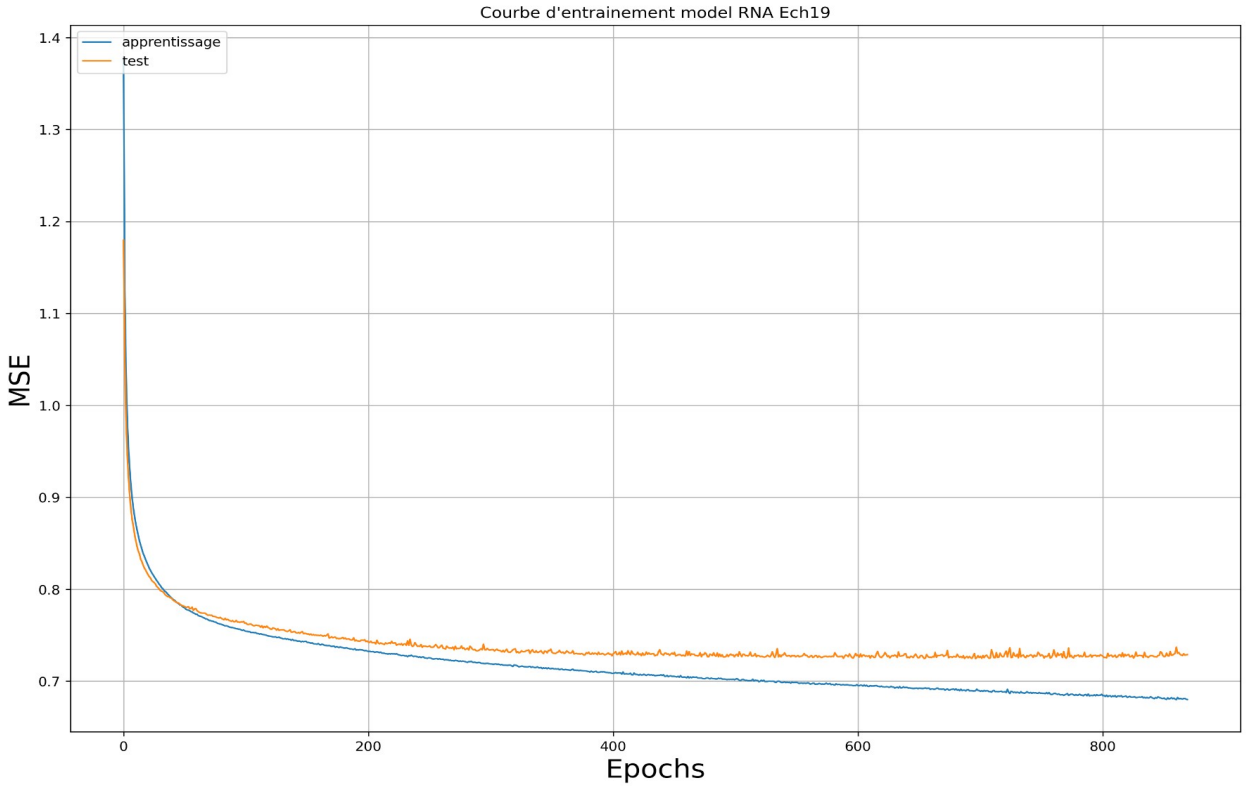

Illustration 7.24: Station de test 1 – FF – Courbe d'entraînement - modèles RNA avec l'optimizer SGD

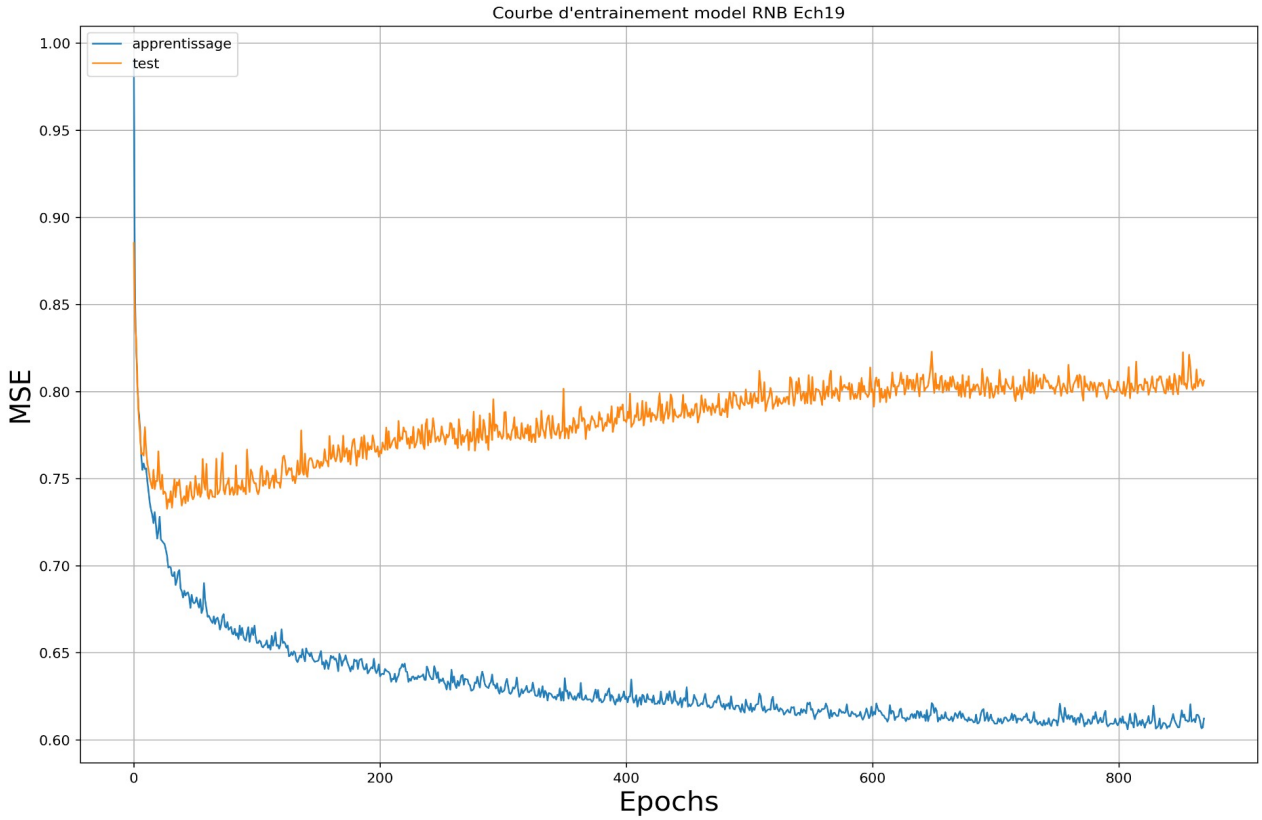

Illustration 7.25: Station de test 1 – FF – Courbe d'entraînement - modèles RNB avec l'optimizer Adam

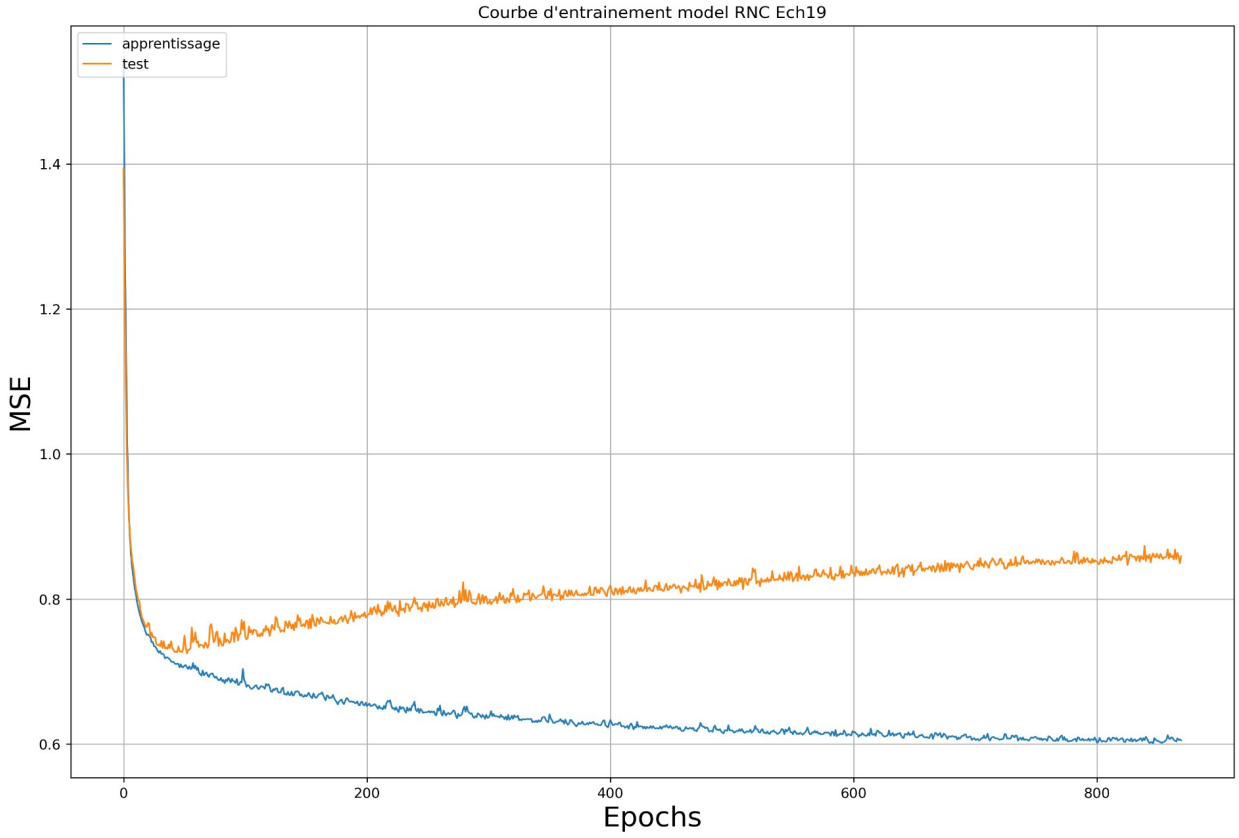

Illustration 7.26: Station de test 1 – FF – Courbe d'entraînement - modèles RNC avec l'optimizer RAdam

On analyse la courbe d'entraînement selon la décroissance du MSE (Mean Squared Error ou erreur quadratique moyenne). Pour limiter le risque de sur-apprentissage, on arrête l'apprentissage juste au moment où l'erreur de test ne décroît plus simultanément avec l'erreur d'apprentissage (ou du moins lorsque l'erreur de test ne se remet pas à augmenter).

En adoptant cette technique, le modèle RNA est celui qui a été choisi comme modèle final à la suite du processus d'entraînement. Comme on peut le voir, les modèles RNB et RNC ont tous les deux sur-appris lorsqu'on avance le nombre d'Epoch jusqu'à 870. En effet, si on devait arrêter l'apprentissage pour ces deux modèles, le nombre d'Epoch idéal se trouverait autour de 20, et dans ce cas le MSE serait un peu plus élevé pour ces deux modèles que pour le modèle RNA. Par conséquent, avec les réglages appliqués, le meilleur modèle est bien le RNA avec l'Optimizer SGD.

7.4.3.4.3 Scores de validation croisée pour la prédiction de FF

Les illustrations 7.27, 7.28, 7.29 et 7.30 présentent respectivement les distributions des scores de qualités (RMSE, ECT, BIAIS, MAE, PSS et FA) des différents modèles RNA, RNB et RNC sur l'échantillon d'apprentissage puis de test (par validation croisée).

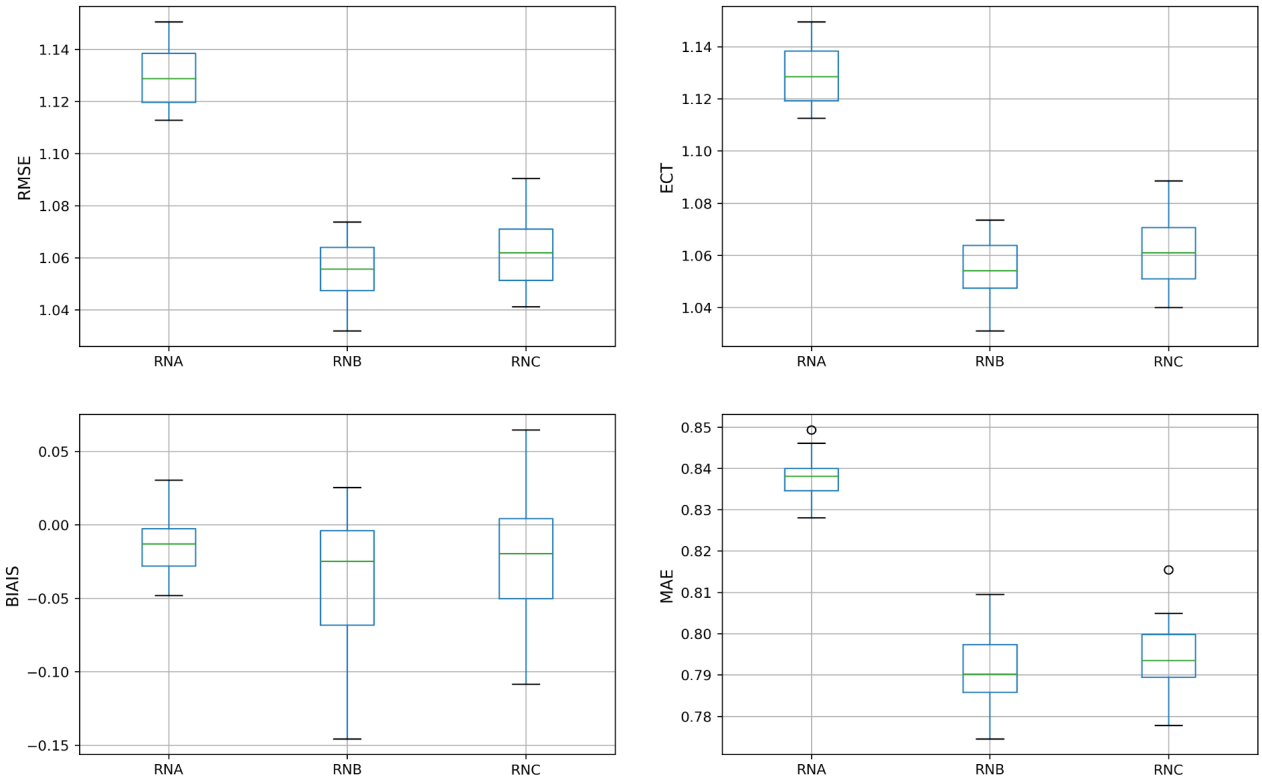

Illustration 7.27: Station de test 1 – FF – Réseau de neurone - Box-plot RMSE, ECT, BIAIS et MAE pour l'échantillon d'apprentissage

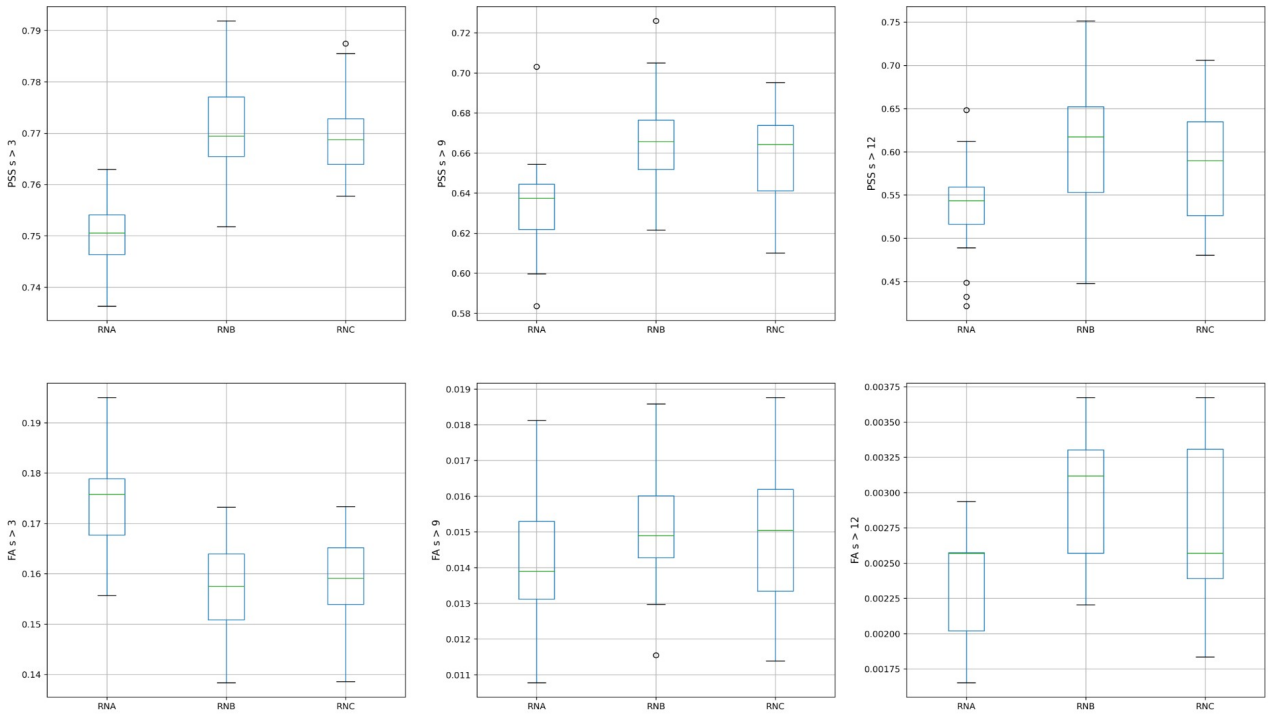

Illustration 7.28: Station de test 1 – FF – Réseau de neurone - Box-plot PSS et FA pour l'échantillon d'apprentissage

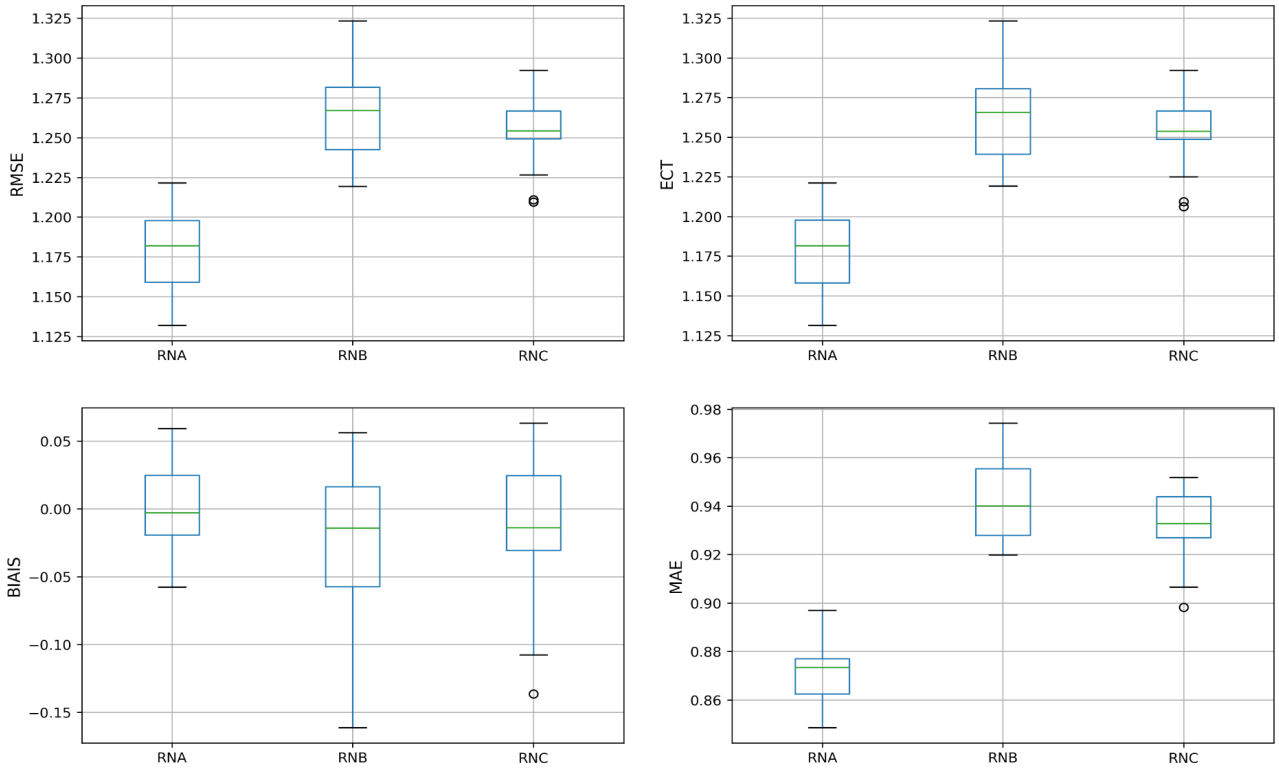

Illustration 7.29: Station de test 1 – FF – Réseau de neurone - Box-plot RMSE, ECT, BIAIS et MAE pour l'échantillon de test

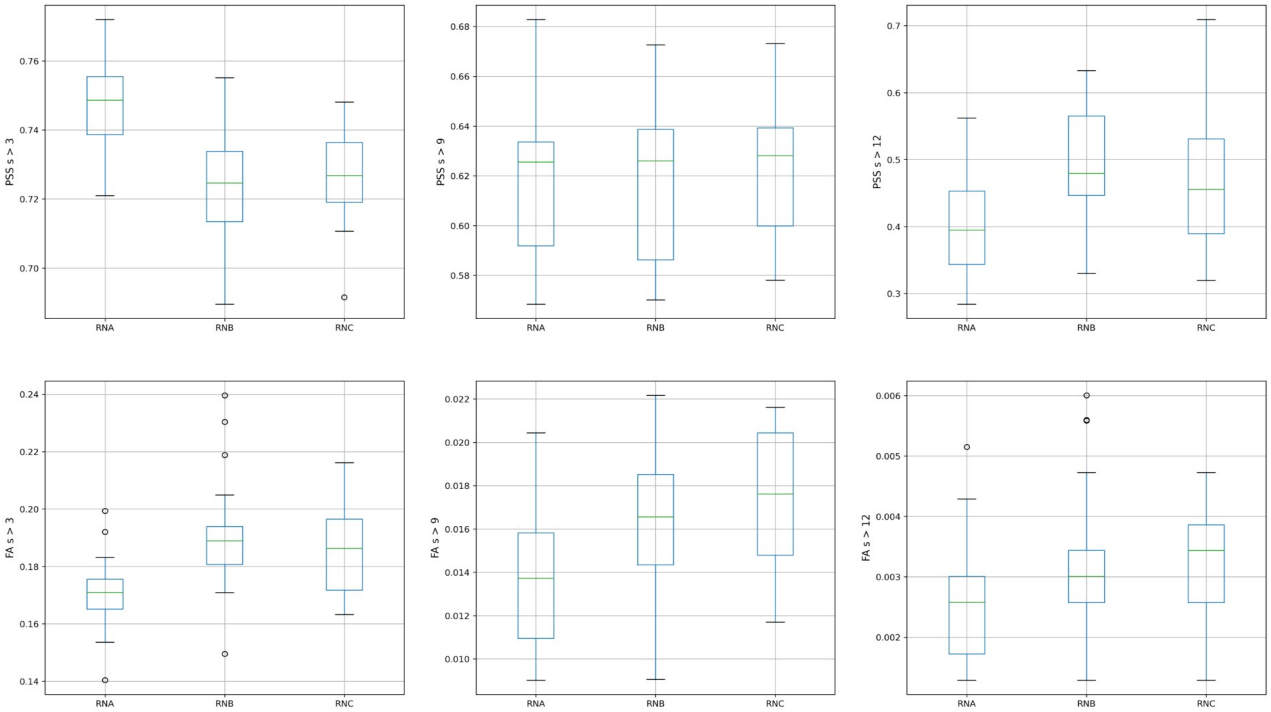

Illustration 7.30: Station de test 1 – FF – Réseau de neurone - Box-plot PSS et MAE pour l'échantillon de test

Les tableaux 7.13 et 7.14 présentent le récapitulatif des scores de validation croisée sur l'échantillon d'apprentissage et de test.

Tableau 7.13: Station de test 1 – FF – Réseau de neurone - Scores de validation croisée sur l'échantillon d'apprentissage

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS2	FA2	PSS3	FA3
RNA	**1..115**	**1.114**	**-0.033**	**0.827**	**0.750**	**0.173**	**0.634**	**0.014**	**0.536**	**0.002**
RNB	1.047	1.046	-0.016	0.782	0.770	0.156	0.665	0.015	0.612	0.002
RNC	1.053	1.052	-0.011	0.785	0.769	0.157	0.658	0.014	0.588	0.002

Tableau 7.14: Station de test 1 – FF – Réseau de neurone - Scores de validation croisée sur l'échantillon de test

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS2	FA2	PSS3	FA3
RNA	**1..182**	**1.181**	**-0.023**	**0.871**	**0.746**	**0.172**	**0.624**	**0.015**	**0.452**	**0.002**
RNB	1.276	1.275	-0.005	0.949	0.727	0.180	0.637	0.018	0.503	0.003
RNC	1.271	1.271	-0.006	0.943	0.726	0.183	0.631	0.018	0.502	0.003

Comme on a pu constater à travers les courbes d'entraînements des modèles, les scores d'apprentissages sont meilleurs pour les modèles RNB et RNC que pour le modèle RNA. À l'inverse ce sont les scores de test du modèle RNA qui sont meilleurs et en restant du même ordre de grandeur que ses scores d'apprentissage. En effet, avec les réglages communs définis pour les trois modèles, **le modèle RNA est le seul à être bien ajusté. Les deux autres modèles (RNB et RNC) ont sur-appris.**

Lorsqu'on analyse uniquement les scores du modèle RNA, ils sont très proches des scores du modèle de forêt aléatoire sans pour autant les égaler. La principale faiblesse de ce modèle est sa capacité de prédiction des vents forts. En effet, les scores de test PSS et FA du modèle RNA pour les seuils S2 (FF > 9 m/s) et S3 (FF > 12 m/s) sont beaucoup moins bons que ceux du modèle de forêt.

De part la complexité de la mise en œuvre et des résultats recueillis, les réseaux de neurones ont été testés uniquement sur la station de test 1.

7.4.3.5 Modèle linéaire avec anamorphose

Ce modèle linéaire avec anamorphose (noté LM_ANA) a été testé en prenant en compte une loi de Weibull pour l'anamorphose sur FF. Des variables explicatives différentes ont également été utilisés (des FF à différents niveaux de pression).

Lors de nos premiers tests sur la station de test 1, les prédictions de la méthode ont produit des forces de vent uniquement comprises entre 2 m/s et 8 m/s (sachant que les observations vont de 0 à 17 m/s).

De part ces premiers résultats et l'étendu des travaux de R&D restant pour améliorer cette méthode, nous avons décidé de ne pas retenir ce modèle dans la suite de notre étude.

7.4.4 Comparaison inter-modèles sur l'échantillon de test

Dans la phase précédente, pour la station de test 1, deux modèles linéaires, un modèle de forêt aléatoire et un modèle de réseau de neurone ont été sélectionnés. Il s'agit du modèle GLM, GLM_MIXTE.STEP, RFC (qu'on notera RF_100 dans la suite), et RNA.

Pour la station de test 2, 2 modèles linéaires (GLM, GLM_MIXTE.STEP) et 1 modèle de forêt aléatoire (RF_100) ont été sélectionnés, les réseaux de neurones n'ayant pas été utilisés pour la station de test 2.

Par la suite, une comparaison a été établie entre ces modèles afin de choisir le modèle final qui sera utilisé pour l'extension de la série d'observation. Pour cela, les scores de validation croisée sur l'échantillon de test, la courbe de fiabilité (QQ-Plot) et la restitution de cycle (diurne, annuel) pour chacun des modèles ont été examinés.

Les illustrations 7.31 et 7.32 présentent les scores de validation croisée sur l'échantillon de test.

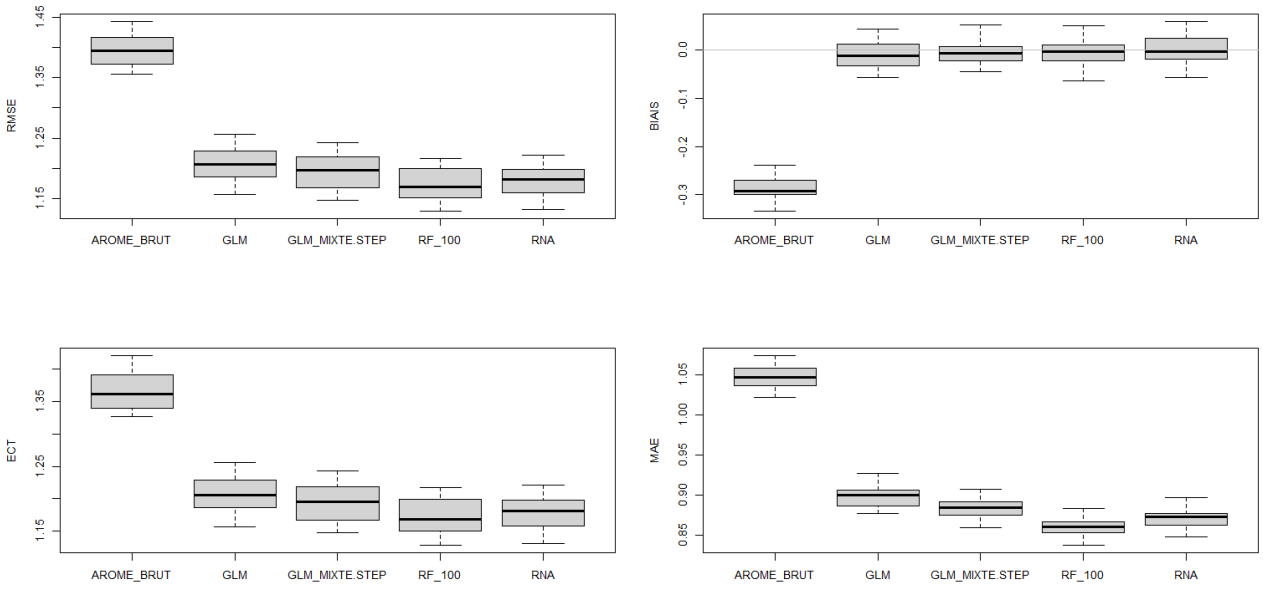

Illustration 7.31: Station de test 1 – FF – Comparaison inter-modèles - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'échantillon de test

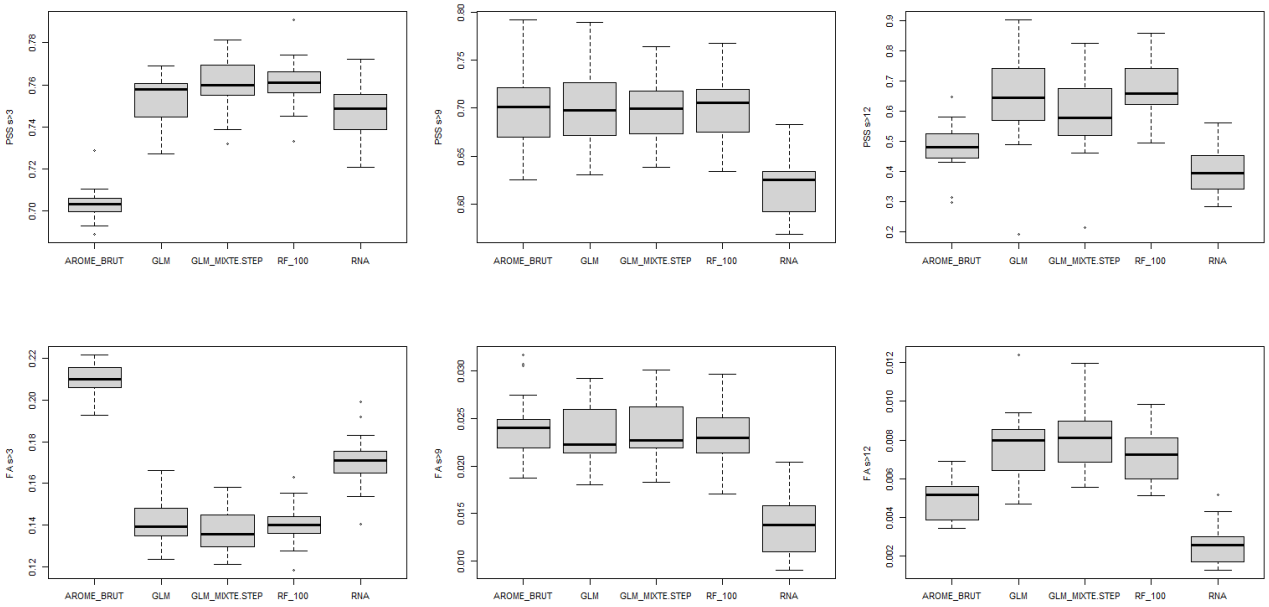

Illustration 7.32: Station de test 1 – FF – Comparaison inter-modèles - Box-plot PSS (première ligne pss > 3 m/s, pss > 9 m/s et pss > 12 m/s) et FA (deuxième ligne fa > 3 m/s, fa > 9 m/s et fa > 12 m/s) pour l'échantillon de test

Les scores RMSE, ECT, BIAIS et MAE des 4 modèles GLM, GLM_MIXTE.STEP, RF_100 et RNA sont très proches et dans tous les cas l'utilisation des modèles statistiques améliorent les sorties de AROME.BRUT. Cependant, le modèle RNA a des scores de détection de vents moyen (PSS > 9 m/s) et fort (PSS > 12 m/s) moins bons que tous les autres modèles. Par conséquent, il ne pourra pas être retenu comme modèle final d'extension.

Par ailleurs, le modèle RF_100 semble avoir des indicateurs légèrement meilleurs que ceux des deux modèles linéaires GLM et GLM_MIXTE.STEP (ces 2 derniers ayant été confrontés dans la phase précédente). Le tableau 7.15 récapitule les scores de validations sur l'échantillon de test.

Tableau 7.15: Station de test 1 – FF – Scores de comparaison inter-modèles sur l'échantillon de test (en vert le modèle sélectionné)

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS2	FA2	PSS3	FA3	Corrélation linéaire avec l'observation
AROME.BRUT	1.396	1.366	-0.288	1.047	0.703	0.21	0.699	0.024	0.482	0.005	0.882
GLM	1.208	1.207	-0.01	0.9	0.754	0.142	0.7	0.023	0.644	0.008	0.906
GLM_MIXTE.STEP	1.192	1.192	-0.004	0.884	0.76	0.138	0.7	0.024	0.585	0.008	0.908
RF_100	**1.173**	**1.172**	**-0.003**	**0.86**	**0.761**	**0.14**	**0.701**	**0.023**	**0.676**	**0.007**	**0.911**
RNA	1.182	1.181	-0.023	0.871	0.746	0.172	0.624	0.015	0.452	0.002	0.910

Au vu des scores ci-dessus, **le modèle RF_100 se positionne comme le meilleur modèle pour la station de test 1.** Selon la même démarche c'est également **RF_100 qui a été choisi comme meilleurs modèle pour la station de test 2**. Pour affiner cette analyse, les courbes de fiabilités (QQ-Plot) ainsi que le cycle diurne et annuel ont été visualisés (illustrations 7.33, 7.34 et 7.35).

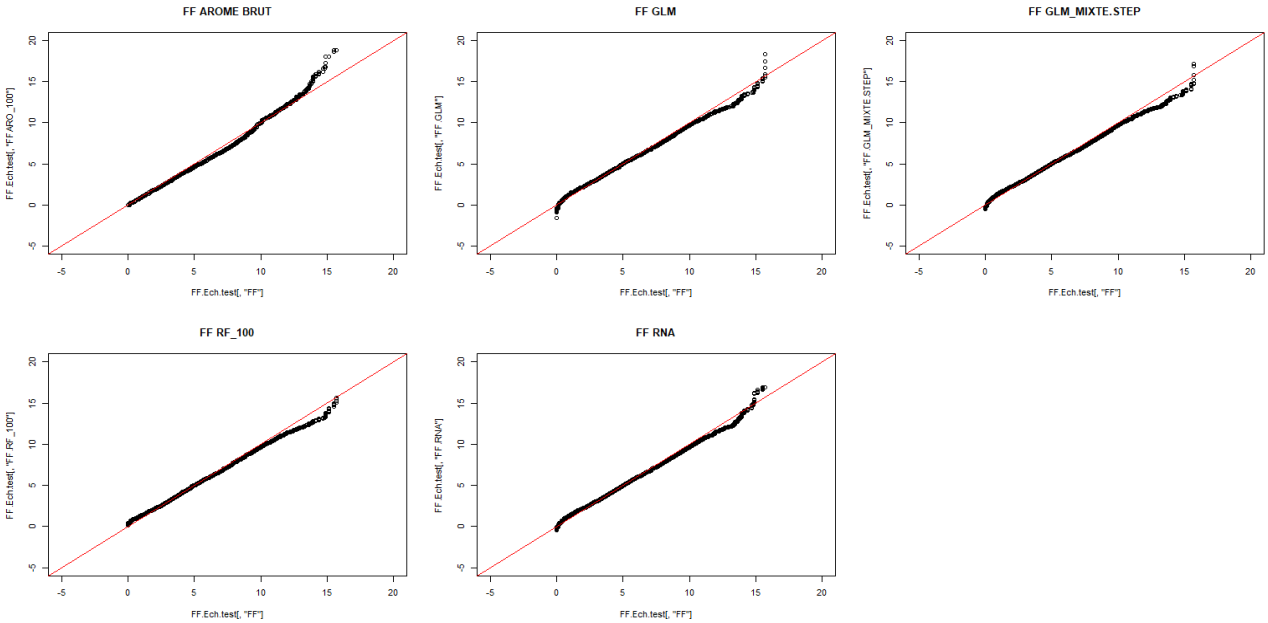

Illustration 7.33: Station de test 1 – FF – QQ-Plot de FF pour les modèles GLM, GLM_MIXTE.STEP, RF_100, RNA et AROME

Sur la courbe QQ-Plot de AROME, on remarque que celui-ci est très proche des observations jusqu'à environ 14 m/s mais surestime les fortes vitesses de vent.

Les courbes de fiabilités penchent également en faveur du modèle RF_100. En effet, le QQ-Plot du RF_100 est meilleur que ceux du GLM, GLM_MIXTE.STEP et RNA. Il se démarque notamment en début et fin de distribution où il est plus proche de la diagonale. Il faudra retenir que la correction par rapport à AROME est un peu trop forte avec notamment une légère sous-estimation des vents entre 12 et 15 m/s environ.

Les QQ-Plots sont également favorables au modèle RF_100 pour la station de test 2.

Pour aller encore loin dans l'analyse, les cycles diurne et annuel des modèles superposés aux observations ont été examinés (illustrations 7.34 et 7.35).

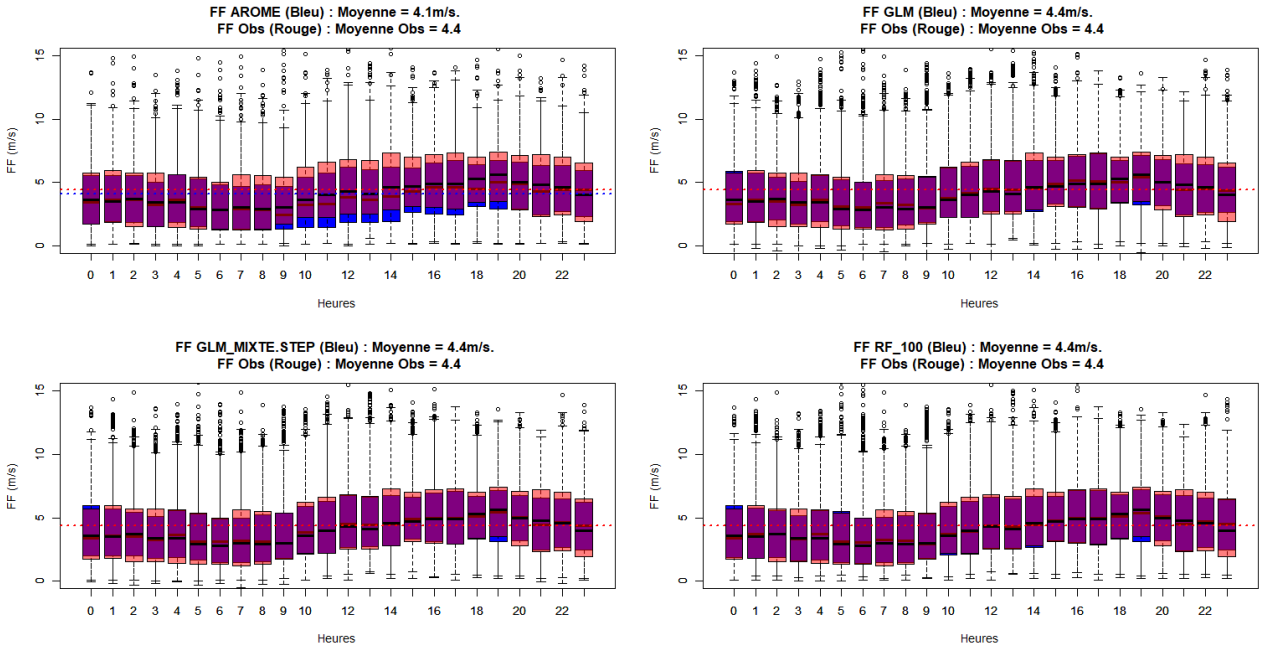

Illustration 7.34: Station de test 1 – FF – Cycle diurne des modèles statistiques (bleu) superposé aux observations (rouge)

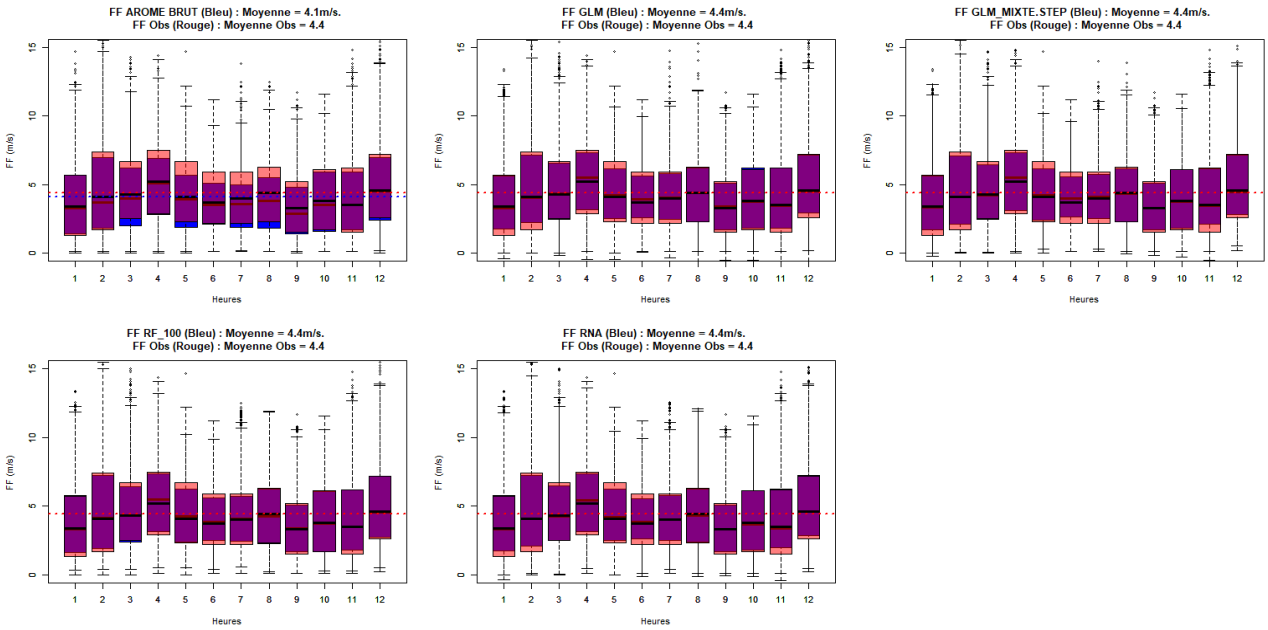

Illustration 7.35: Station de test 1 – FF – Cycle annuel des modèles statistiques (bleu) superposé aux observations (rouge)

Les modèles statistiques améliorent tous la distribution de la force du vent AROME (notamment le biais annuel). Cependant, ils restituent tous le cycle diurne et annuel quasiment de la même façon. En effet, ces graphiques ne permettent pas de faire une distinction flagrante visuellement entre les distributions des 4 modèles statistiques (GLM, GLM_MIXTE.STEP, RF_100 et RNA). Par conséquent, l'avantage reste toujours pour le modèle RF_100 dans le choix du modèle final pour l'extension.

Au vu des analyses précédentes, **nous avons décidé de choisir le modèle RF_100 (qui présente les meilleurs indicateurs par rapport aux autres modèles) pour l'extension de la série d'observation horaire de la force du vent de la station de test 1**.

Concernant la station de test 2, le constat était le même. C'est le modèle RF_100 qui présente les meilleurs indicateurs de qualité.

7.4.4.1 Test des prédicteurs de régimes de temps

Après avoir choisi d'étendre la série d'observation avec le modèle de forêt aléatoire avec des prédicteurs issus du modèle AROME, nous avons testé les données de régimes de temps comme prédicteurs pour les 2 stations de test. Ces données caractérisent (plus ou moins) la variabilité basse fréquence de l'atmosphère. Elles sont issues de la classification en anomalie de pression ramenée au niveau de la mer (PMER) des réanalyses ERA5.

Nous utilisons 8 variables constituées par les corrélations de chaque jour aux centroïdes des régimes PMER pour constituer 4 prédicteurs. Les 8 variables utilisés sont les suivantes :

- H_Z0 : corrélation au régime d'hiver NAO+ (Zonal)
- H_AR : corrélation au régime d'hiver Dorsale (Atlantic Ridge)
- H_EA : corrélation au régime d'hiver Blocage (European Blocking)
- H_AL : corrélation au régime d'hiver NAO- (Atlantic Low)
- E_GA : corrélation au régime d'été Anticyclone Groenland
- E_AL : corrélation au régime d'été minimum atlantique (Atlantic Low)
- E_EA : corrélation au régime d'été Blocage (European Blocking)
- E_Z0 : corrélation au régime d'été Zonal

Ces 8 variables constituées par les régimes été/hiver sont liées 2 à 2 en utilisant le tableau 7.16 de pondération.

Tableau 7.16: Table de pondération utilisée pour lier les régimes de temps 2 à 2

Pondération	Janvier	Février	Mars	Avril	Mai	Juin	Juillet	Août	Septembre	Octobre	Novembre	Décembre
H_Z0	1	1	1	0.5	0.25	0	0	0	0.25	0.5	1	1
E_Z0	0	0	0	0.5	0.75	1	1	1	0.75	0.5	0	0
H_AR	1	1	1	0.5	0.25	0	0	0	0.25	0.5	1	1
E_GA	0	0	0	0.5	0.75	1	1	1	0.75	0.5	0	0
H_EA	1	1	1	0.5	0.25	0	0	0	0.25	0.5	1	1
E_EA	0	0	0	0.5	0.75	1	1	1	0.75	0.5	0	0
H_AL	1	1	1	0.5	0.25	0	0	0	0.25	0.5	1	1
E_AL	0	0	0	0.5	0.75	1	1	1	0.75	0.5	0	0

Une fois les pondérations appliquées aux données, nous obtenons les 4 prédicteurs (RG_Zonal, RG_Dorsal, RG_Blocage et RG_AtantiqueLow) en utilisant les associations suivantes :

- RG_Zonal = H_Z0 + E_Z0
- RG_Dorsal = H_AR + E_GA
- RG_Blocage = H_EA + E_EA
- RG_AtlantiqueLow = H_AL + E_AL.

Afin d'apprécier l'apport de ces prédicteurs, ils ont été ajoutés aux 17 variables explicatives pour les 2 stations, c'est-à-dire :

- RegMM, RegHH_FF, SECTEURARO_100, FFARO_10, FFARO_100, FFARO_250, TARO_2, TARO_500, SQRTTKEARO_100, HUARO_2, TKEARO_10, TPWARO_850, PC1_FF, PC2_FF, PC3_FF, PC4_FF et PC5_FF pour la modélisation de FF de la station de test 1,
- RegMM, RegHH_FF, SECTEURARO_100, FFARO_100, FFARO_250, TARO_2, TARO_50, TKEARO_100, SQRTTKEARO_10, HUARO_2, PMERARO, TPWARO_850, PC1_FF, PC2_FF, PC3_FF, PC4_FF et PC5_FF pour la modélisation de FF de la station de test 2.

Les illustrations 7.36 à 7.39 ainsi que les tableaux 7.17 et 7.18 présentent les scores de validation croisée pour le modèle de forêt aléatoire pour FF sans les prédicteurs de régimes de temps (RF_100) et avec les prédicteurs de régimes de temps (noté RF_100.AVEC.REGI) pour la station de test 2.

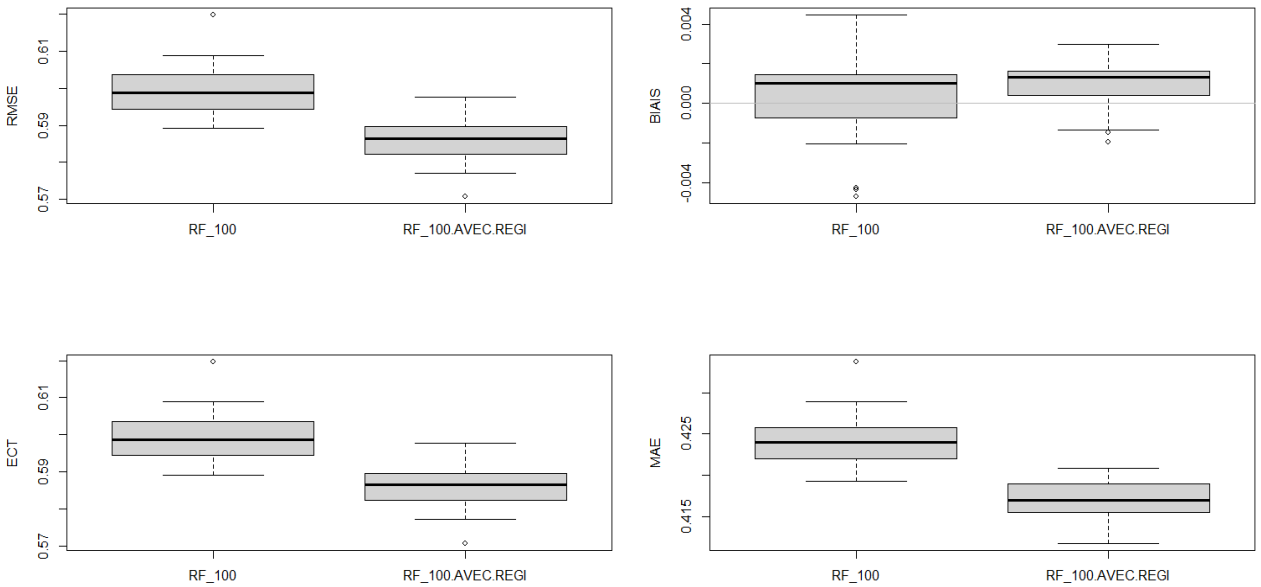

Illustration 7.36: Station de test 2 – FF – Modèles de forêt aléatoire à 100 arbres avec et sans prédicteurs de régimes de temps - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'échantillon d'apprentissage

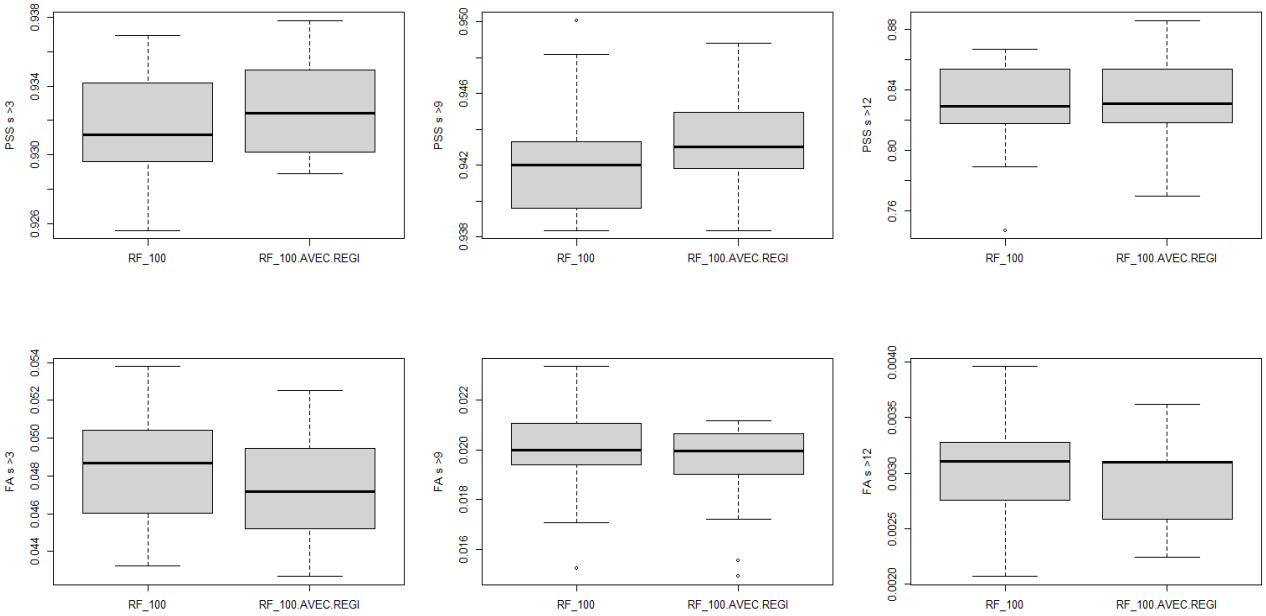

Illustration 7.37: Station de test 2 – FF – Modèles de forêt aléatoire à 100 arbres avec et sans prédicteurs de régimes de temps - Box-plot PSS (première ligne pss > 3 m/s, pss > 9 m/s et pss > 12 m/s) et FA (deuxième ligne fa > 3 m/s, fa > 9 m/s et fa > 12 m/s) pour l'échantillon d'apprentissage

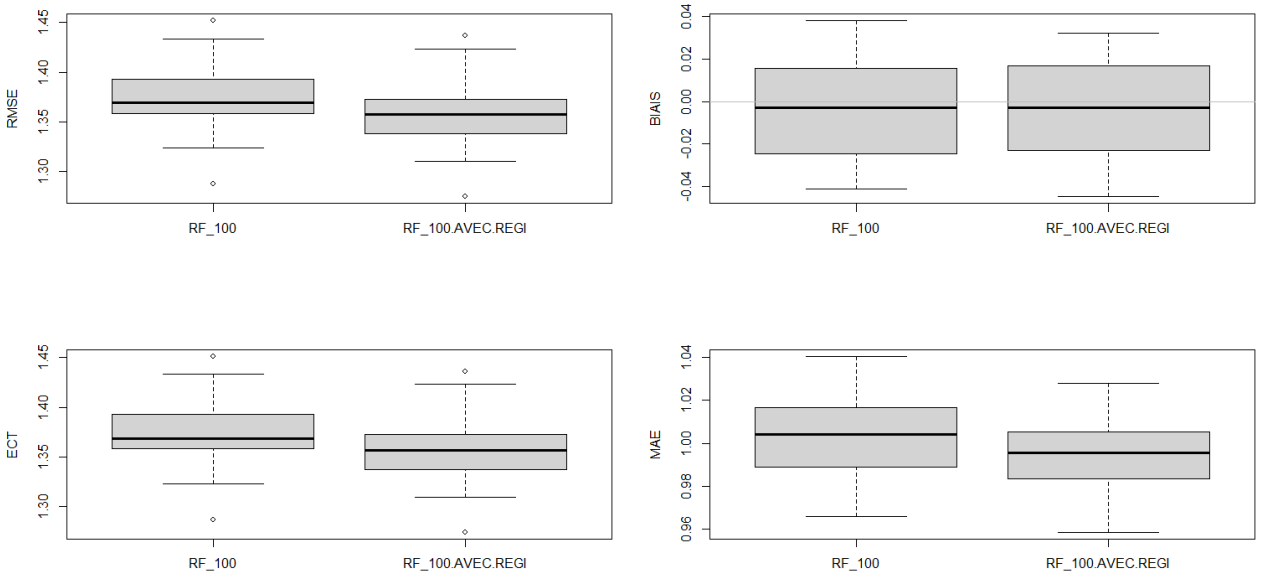

Illustration 7.38: Station de test 2 – FF – Modèles de forêt aléatoire à 100 arbres avec et sans prédicteurs de régimes de temps - Box-plot RMSE (haut gauche), ECT (bas gauche), BIAIS (haut droite) et MAE (bas droite) pour l'échantillon de test

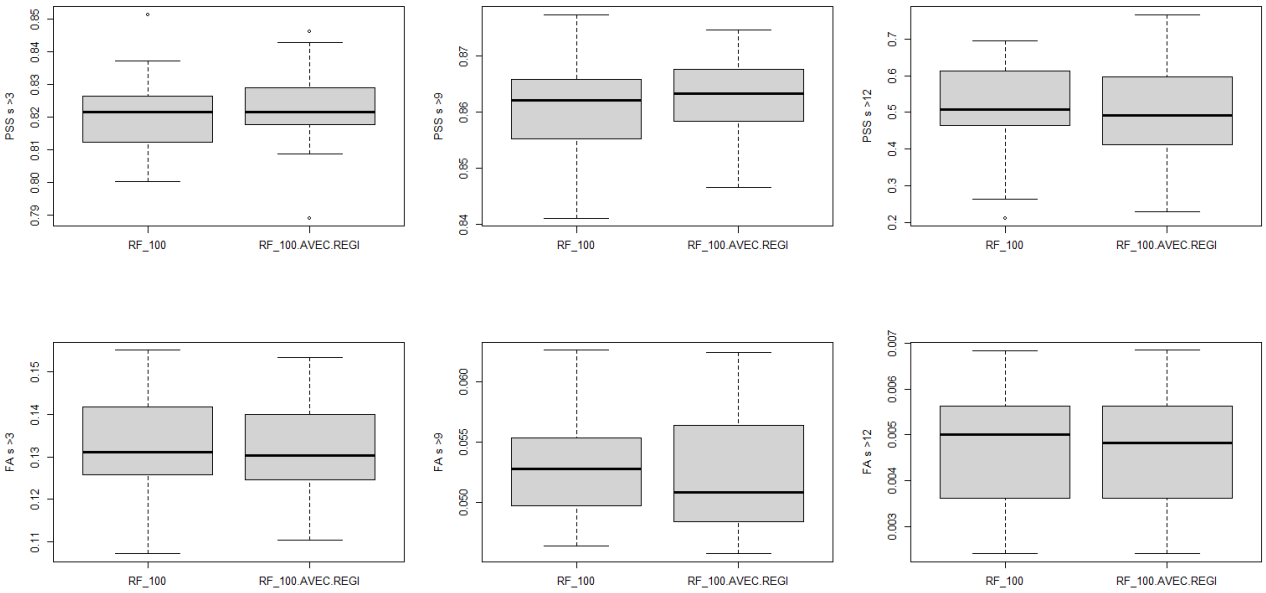

Illustration 7.39: Station de test 2 – FF – Modèles de forêt aléatoire à 100 arbres avec et sans prédicteurs de régimes de temps - Box-plot PSS (première ligne pss > 3 m/s, pss > 9 m/s et pss > 12 m/s) et FA (deuxième ligne fa > 3 m/s, fa > 9 m/s et fa > 12 m/s) pour l'échantillon de test

Tableau 7.17: Station de test 2 – FF – Modèles de forêt aléatoire avec et sans prédicteurs de régimes de temps - Scores de validation croisée sur l'échantillon d'apprentissage (en vert, le modèle choisi)

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS2	FA2	PSS3	FA3
RF_100	**0.6**	**0.6**	**0**	**0.424**	**0.931**	**0.049**	**0.942**	**0.02**	**0.829**	**0.003**
RF_100.AVEC.REGI	0.586	0.586	0.001	0.417	0.933	0.047	0.943	0.019	0.831	0.003

Tableau 7.18: Station de test 2 – FF – Modèles de forêt aléatoire avec et sans prédicteurs de régimes de temps - Scores de validation croisée sur l'échantillon de test (en vert, le modèle choisi)

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS2	FA2	PSS3	FA3
RF_100	**1.373**	**1.372**	**-0.003**	**1.004**	**0.821**	**0.132**	**0.861**	**0.053**	**0.509**	**0.005**
RF_100.AVEC.REGI	1.357	1.357	-0.003	0.994	0.822	0.131	0.863	0.052	0.491	0.005

L'ajout des prédicteurs de régimes de temps aux modèles de FF, U et V ne montre pas un apport notable. De ce fait et afin de ne pas augmenter encore le nombre de prédicteurs (déjà 17 pour les deux stations de test, risque de non parcimonie), **nous avons décidé de conserver les modèles de forêts aléatoires à 100 arbres pour FF.**

7.5 Modélisation de la direction du vent à 100 m

Dans l'élaboration des modèles statistiques pour la direction du vent DD, la démarche reste la même que celle présentée pour la force du vent FF. La différence majeure se situe au niveau de la variable à prédire (DD), qui est une variable circulaire. Par conséquent, DD est décomposée en 2 composantes U et V qui sont définies par :

$$U = \sin\left(DD * pi/180\right) * FF$$

$$V = \cos\left(DD * pi/180\right) * FF$$

Ainsi, on a une variable à prédire pour chacun des composantes U et V de DD, puis on estime la direction du vent à partir des estimations de U et V à travers la formule suivante :

$$DD = \left(90 - \left(180 * atan2\left(V, U\right)/pi\right)\right) \quad .$$

Comme pour la force du vent, les modèles statistiques de chacune des composantes U et V sont ajustés non pas sur U et V de l'observation, mais sur l'écart de U et V observé avec U et V du vent AROME.

Par conséquent, **les variables à prédire sont :**

Y_1 = UmUARO = $U_{OBS} - U_{ARO}$, et **Y_2 = VmVARO = $V_{OBS} - V_{ARO}$.**

La méthodologie étant la même que pour les études de la force du vent, nous allons présenter uniquement les résultats principaux pour la direction du vent.

Notamment, nous ne détaillons pas la sélection des modèles « champion » pour chaque catégorie de modèle statistique et nous présentons directement les résultats de l'inter-comparaison de ces modèles « champion ».

7.5.1 Sélection des variables explicatives

Les figures suivantes (illustrations 7.40 à 7.42) montrent les différentes phases de la sélection des variables explicatives pour les composantes U et V de DD pour la station de test 1.

À la suite du processus de sélection des variables explicatives pour les composantes U et V de DD, il y a 18 variables qui ont été utilisées pour établir les modèles statistiques (linéaires, arbre binaire, forêt aléatoire et réseau de neurone) pour chacune des deux stations de test.

Les variables explicatives de la station de test 1 sont : RegMM, RegHH_DD, UARO_10, UARO_100, UARO_250, VARO_100, VARO_500, SQRTTKEARO_10, SQRTTKEARO_100, TARO_2, TARO_500, HUARO_2, TPWARO_850, PC1_DD, PC2_DD, PC3_DD, PC4_DD, PC5_DD ;

et celles de la station de test 2 sont RegMM, RegHH_DD, UARO_100, UARO_500, VARO_10, VARO_500, SQRTTKEARO_100, SQRTTKEARO_160, TARO_100, TARO_500, HUARO_2, PMERARO, TPWARO_850, PC1_DD, PC2_DD, PC3_DD, PC4_DD, PC5_DD.

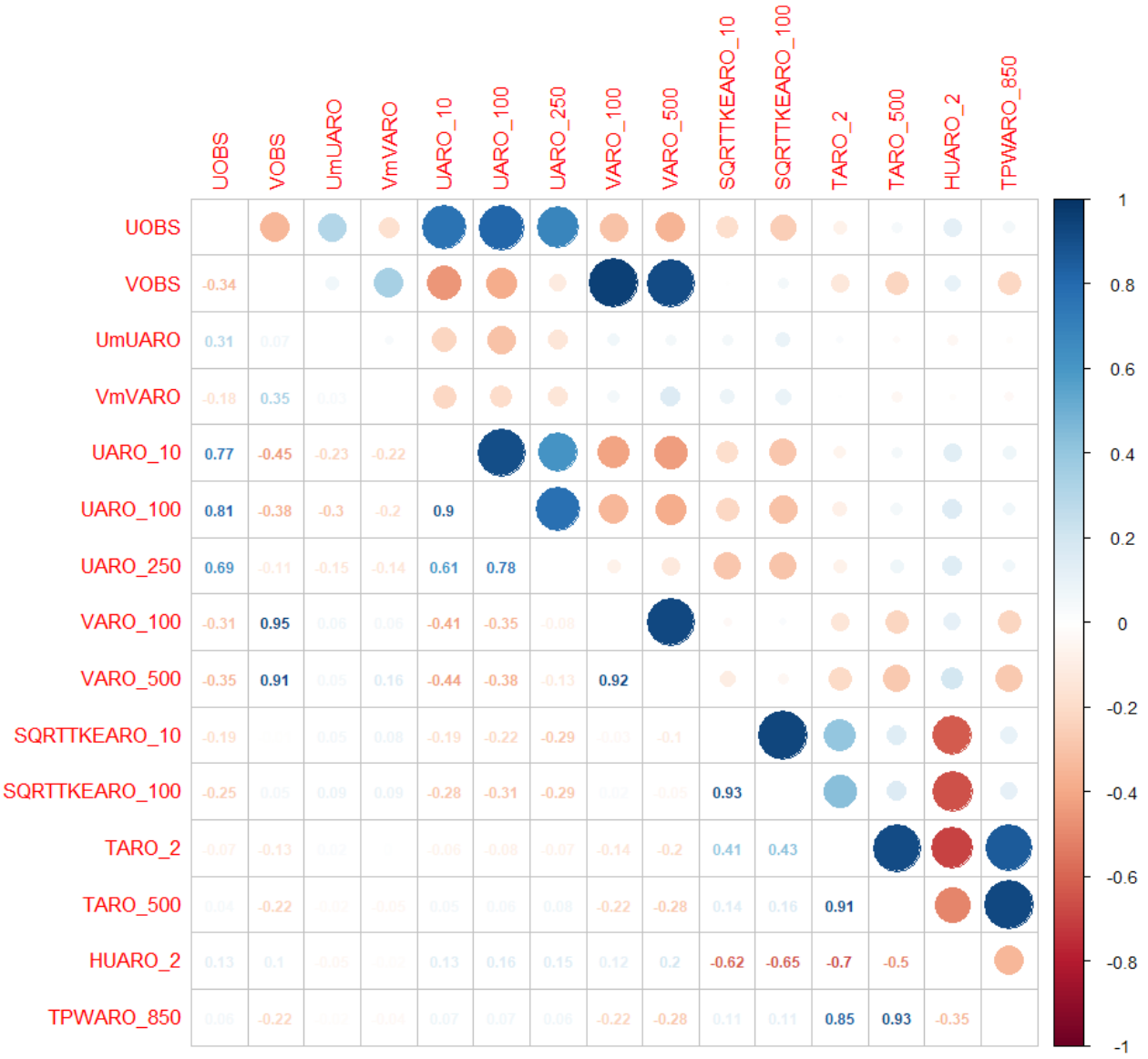

Illustration 7.40: Station de test 1 – Premier groupe (variables explicatives conservées pour la sélection finale) pour la prédiction de DD

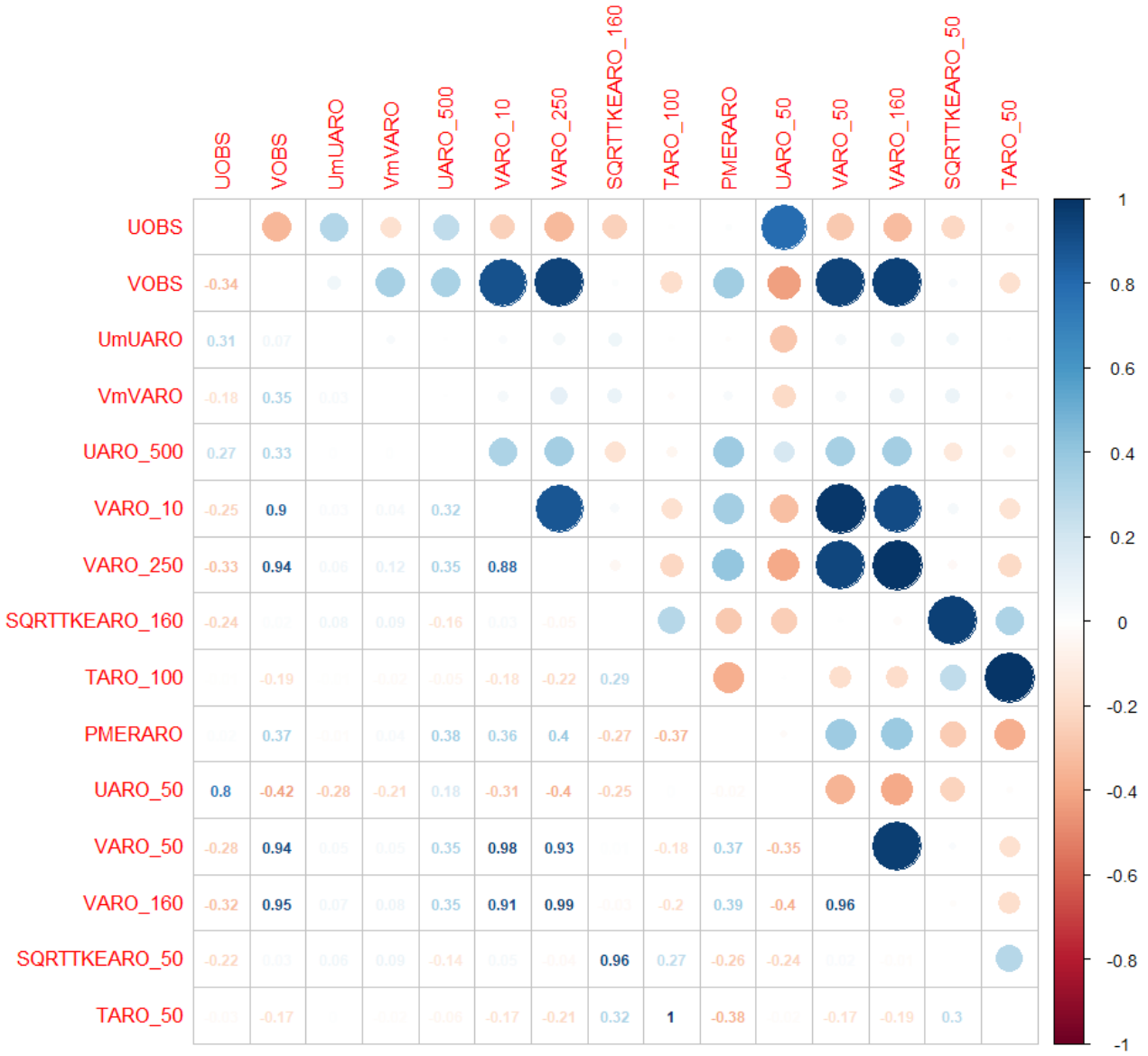

Illustration 7.41: Station de test 1 – Deuxième groupe (variables de l'ACP) pour la prédiction de DD

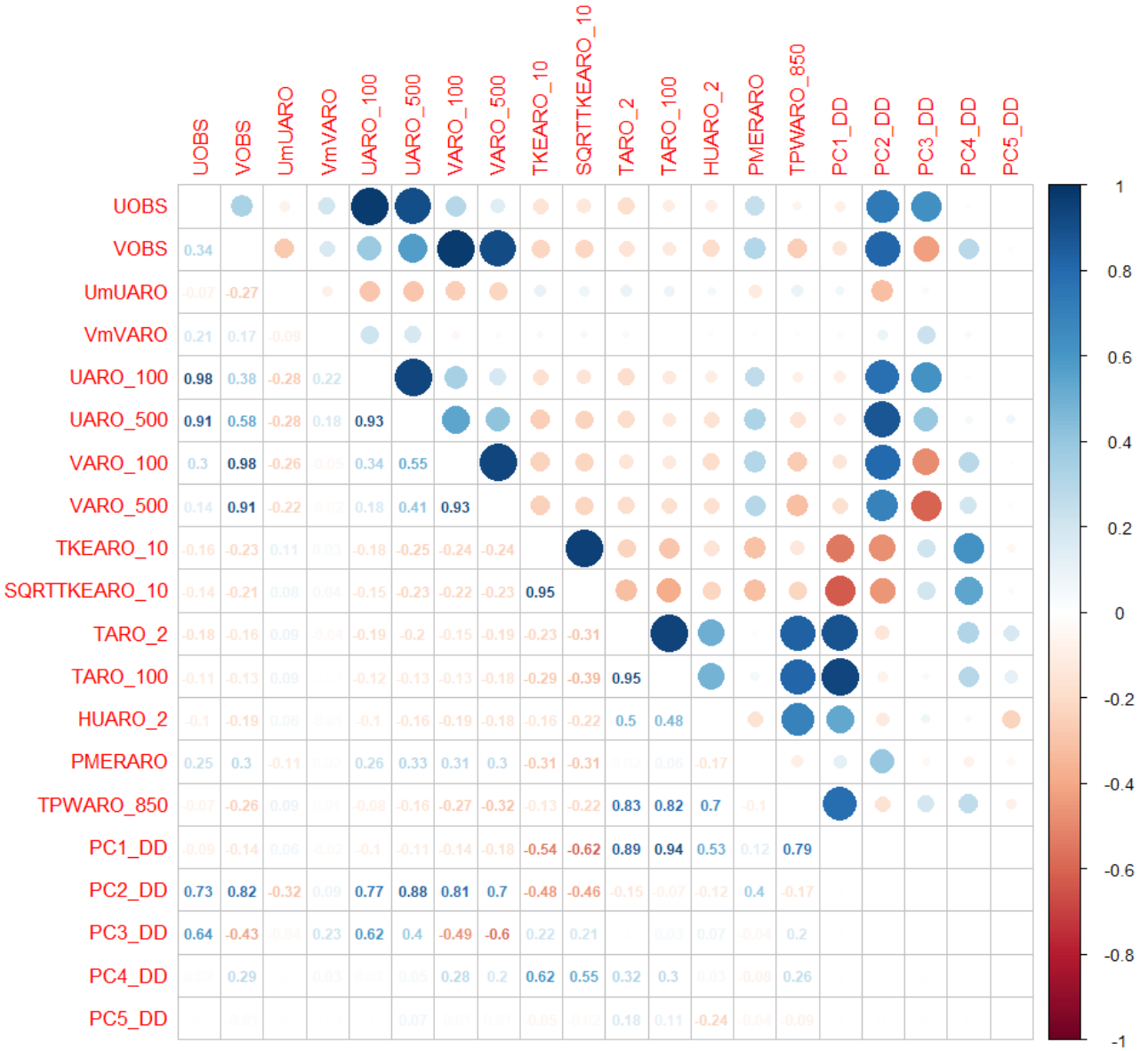

Illustration 7.42: Station de test 1 – Variables explicatives finales (avec les composantes principales de l'ACP) pour la prédiction de DD

7.5.2 Études des modèles statistiques

L'établissement des modèles statistiques pour l'estimation de la direction du vent (composantes U et V) suit la même démarche que celle de la force du vent.

Comme pour la force du vent, les réseaux de neurone n'ont pas été utilisés pour la modélisation des composantes U et V de la direction du vent de la station de test 2.

7.5.2.1 Scores de test pour la prédiction de U et V

7.5.2.1.1 Scores pour la composante U

Les illustrations 7.43 à 7.46 présentent les distributions des scores de test (par validation croisée) de la prédiction de U de la station de test 1 pour les réseaux de neurones puis AROME, les modèles linéaires et les forêts aléatoires « champion ».

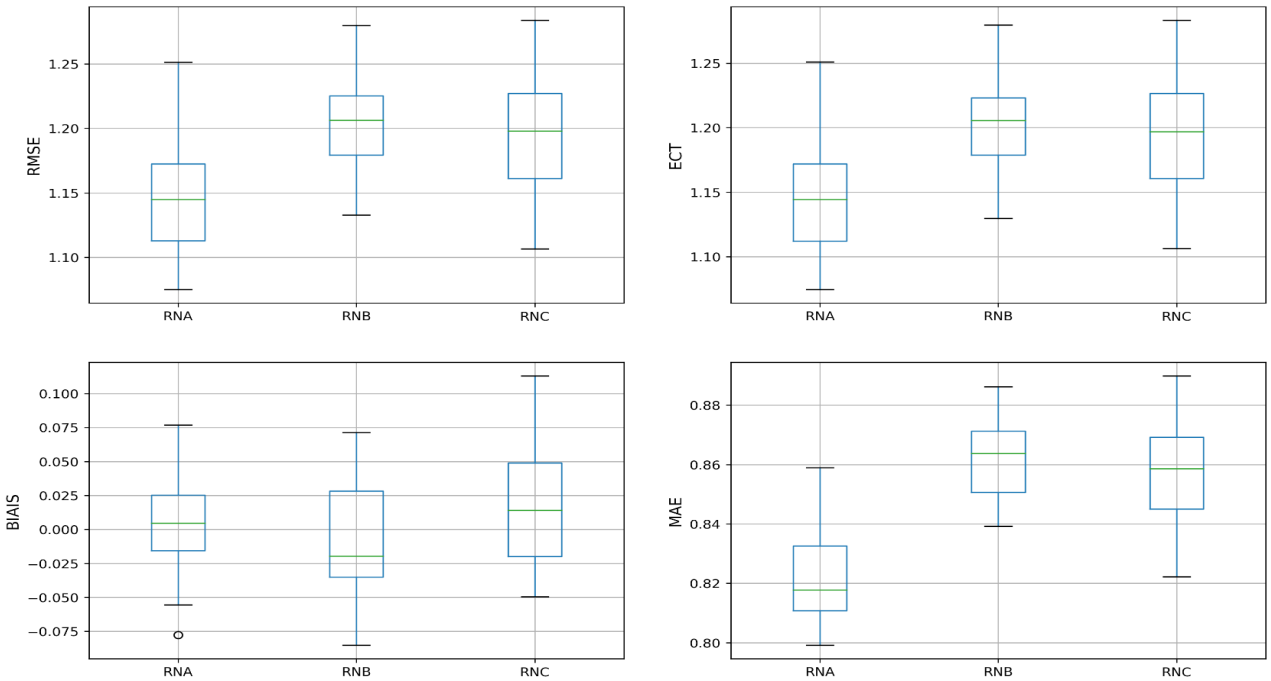

Illustration 7.43: Station de test 1 – U – Réseau de neurone - Box-plot RMSE, ECT, BIAIS et MAE pour l'échantillon de test

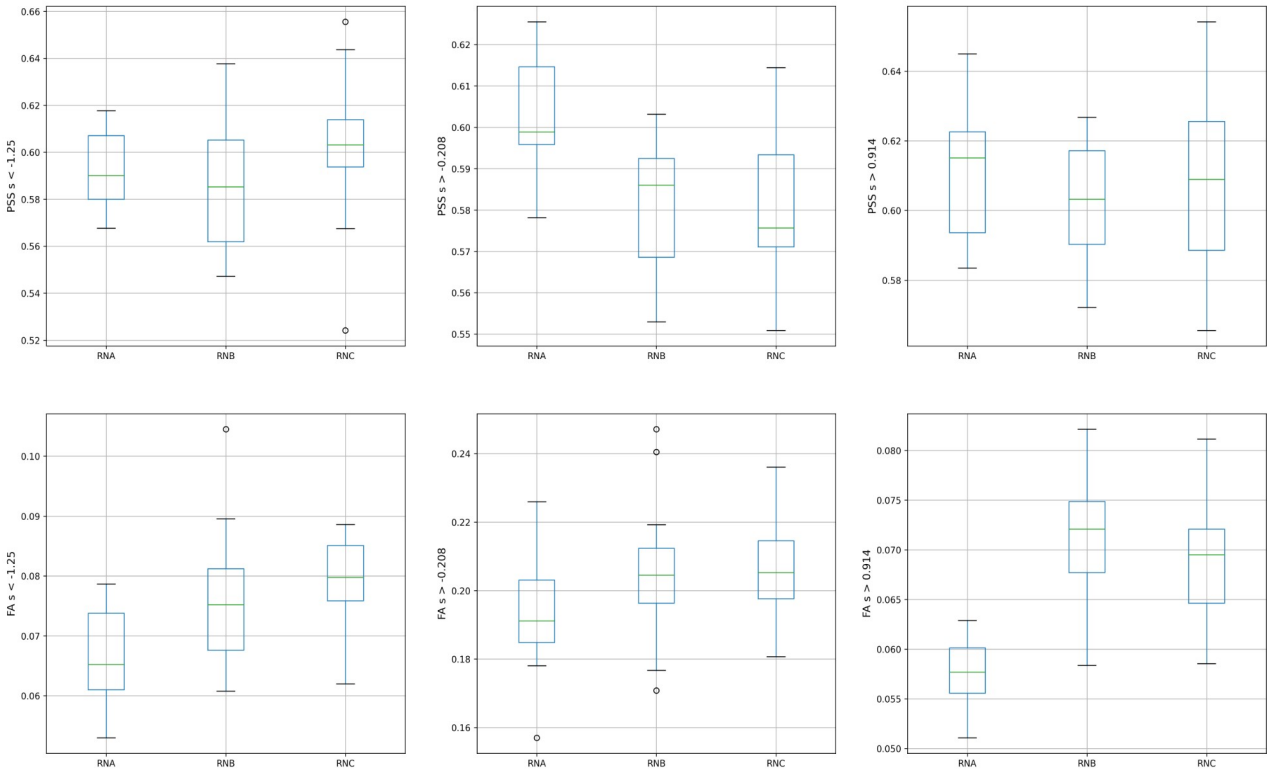

Illustration 7.44: Station de test 1 – U – Réseau de neurone - Box-plot PSS et FA pour l'échantillon de test

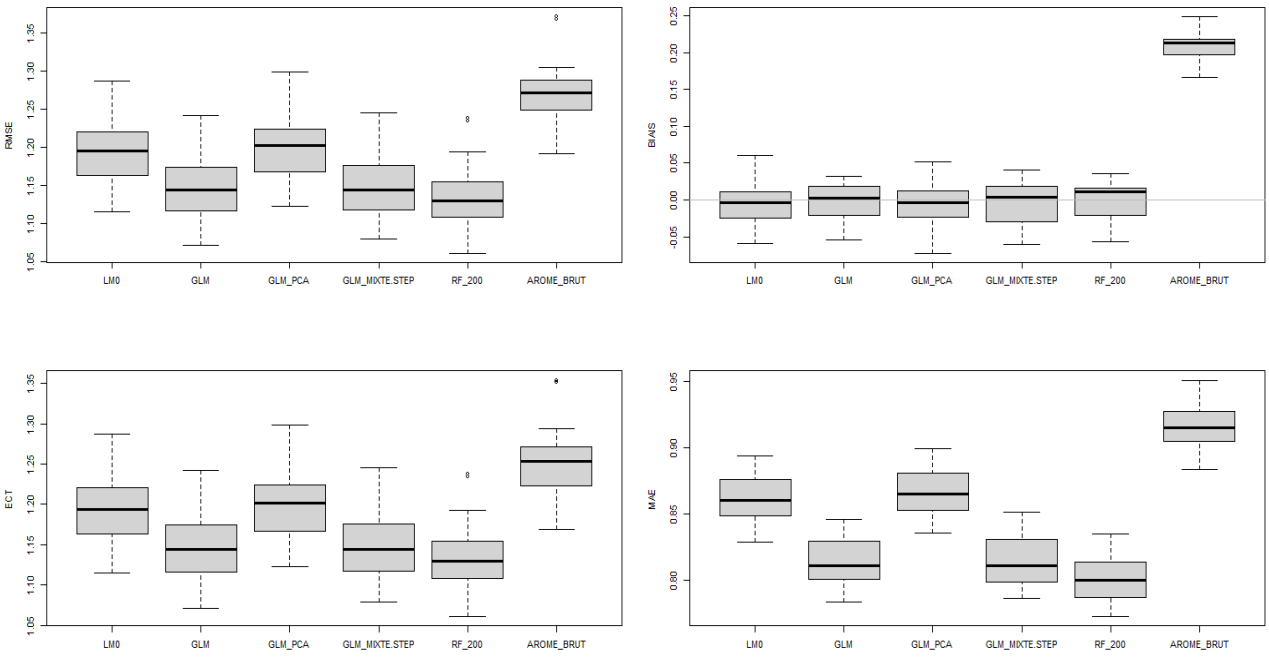

Illustration 7.45: Station de test 1 – U – Modèles linéaires, forêt et AROME - Box-plot RMSE, ECT, BIAIS et MAE pour l'échantillon de test

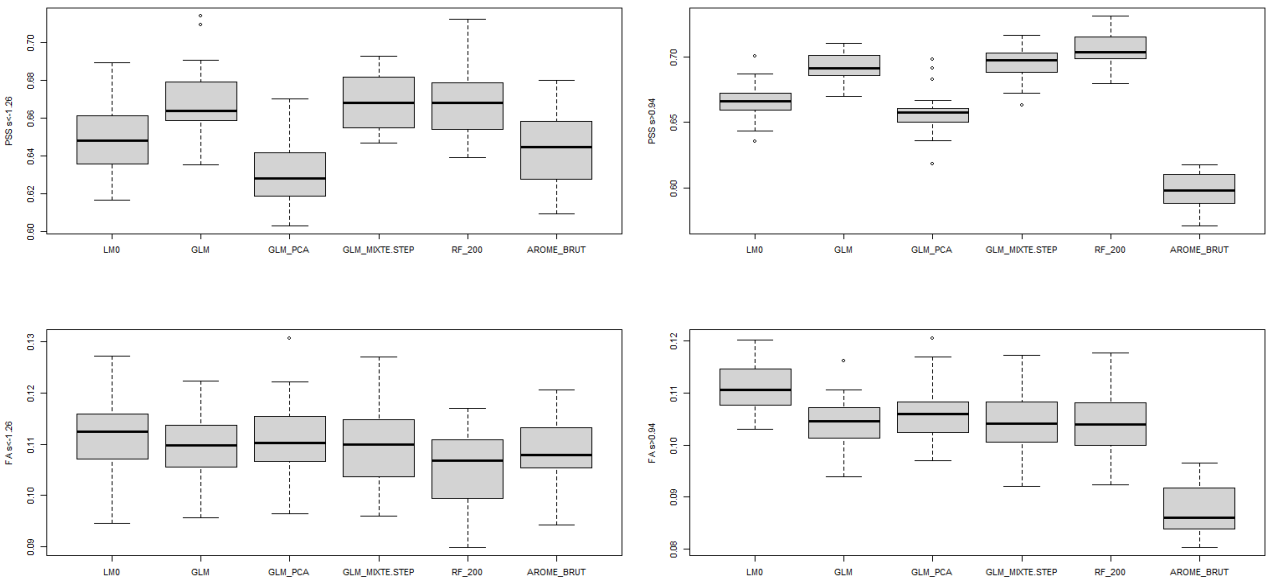

Illustration 7.46: Station de test 1 – U – Modèles linéaires, forêt et AROME - Box-plot PSS et FA pour l'échantillon de test

Comme pour la force du vent, les modèles RNB et RNC ont sur-appris. Par conséquent, on analyse uniquement les résultats du modèle RNA.

Les scores RMSE, ECT, BIAIS et MAE du modèle RNA sont du même ordre de gradeur que ceux des meilleurs modèles (GLM, GLM_MIXTE.STEP et RF_200) pour la prédiction de U. Mais visuellement ces scores sont relativement meilleurs pour les modèles GLM, GLM_MIXTE.STEP et RF_200 que pour RNA.

Pour les scores PSS et FA, ils sont beaucoup moins bons pour le modèle RNA que pour les modèles GLM, GLM_MIXTE.STEP et RF_200. En effet pour le calcul de ces scores, nous avons cherché à connaître le taux de bonne prédiction et de fausse alarme de la prédiction pour les valeurs et U inférieur à son 1er quantile, et les valeurs U supérieur à son 3^{e} quantile.

Le tableau 7.19 permet de résumer les scores de test détaillés pour les différents modèles.

Tableau 7.19 : Station de test 1 – U – Modèles linéaires, forêt, Réseau de neurone – Scores de validation croisée sur l'échantillon de test

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS3	FA3
LM0	1.195	1.194	-0.004	0.861	0.648	0.112	0.666	0.111
GLM	1.147	1.147	-0.004	0.813	0.669	0.109	0.691	0.104
GLM_PCA	1.204	1.204	-0.007	0.865	0.629	0.111	0.658	0.106
GLM_MIXTE.STEP	1.147	1.147	-0.005	0.814	0.668	0.11	0.696	0.105
RF_200	**1.139**	**1.138**	**-0.001**	**0.802**	**0.668**	**0.105**	**0.707**	**0.104**
AROME.BRUT	1.272	1.254	0.21	0.916	0.644	0.108	0.598	0.088
RNA	1.149	1.149	0.003	0.821	0.592	0.066	0.610	0.057

Le tableau confirme la hiérarchie constatée des modèles statistiques lors de l'analyse des box-plots. Les meilleurs modèles restent le GLM, GLM_MIXTE.STEP et le RF_200, avec un avantage pour le RF_200.

Le même constat est fait pour la station de test 2 : selon les indicateurs de qualité, les meilleurs modèles restent le GLM, GLM_MIXTE.STEP et le RF_200, avec un avantage pour le RF_200.

7.5.2.1.2 Scores pour la composante V

Les box-plots suivants (illustrations 7.47, 7.48, 7.49 et 7.50) présentent les scores de test (par validation croisée) de la prédiction de la composante V de la station de test 1.

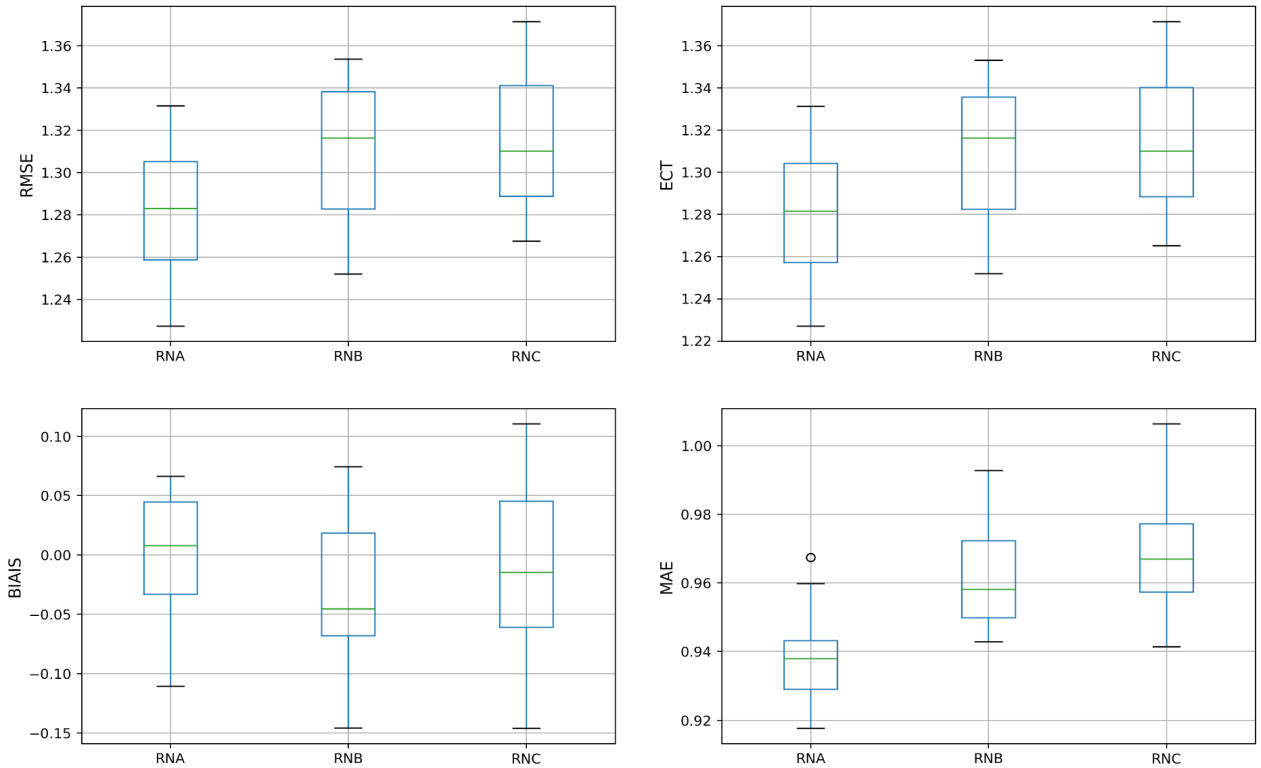

Illustration 7.47: Station de test 1 – V – Réseau de neurone - Box-plot RMSE, ECT, BIAIS et MAE pour l'échantillon de test

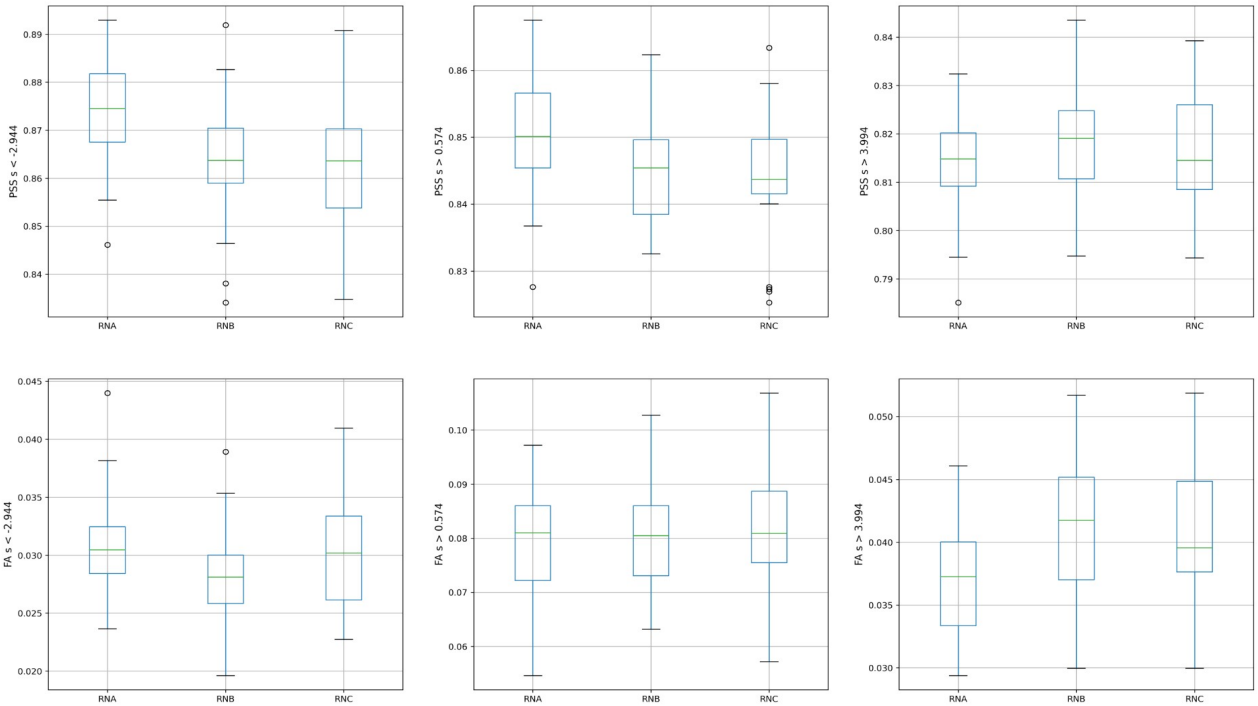

Illustration 7.48: Station de test 1 – V – Réseau de neurone - Box-plot PSS et FA pour l'échantillon de test

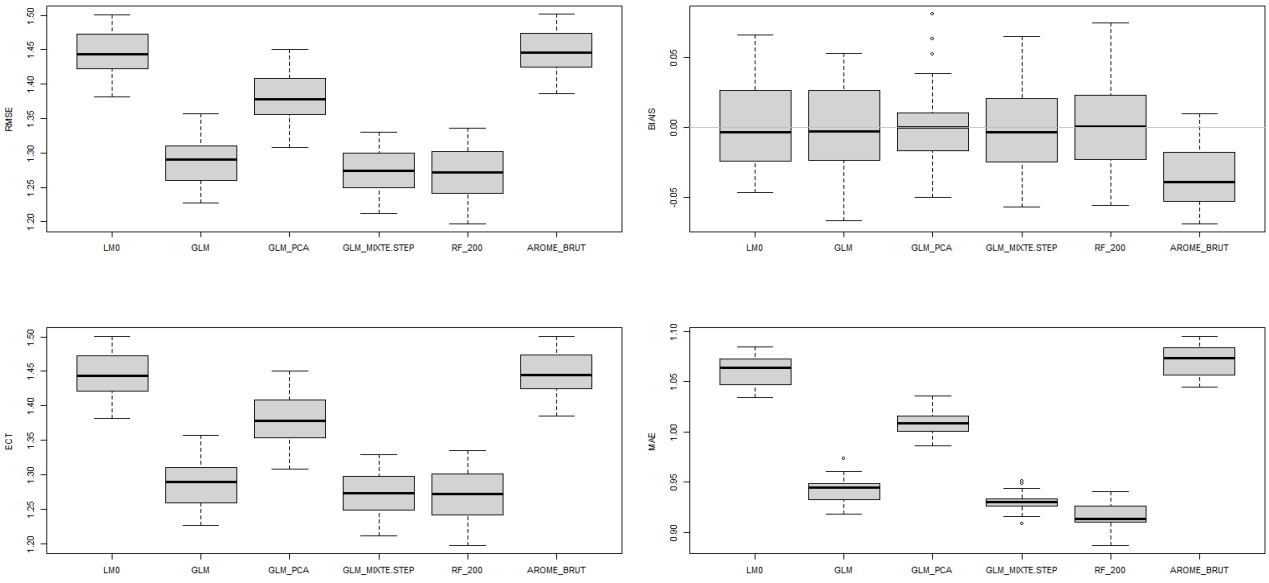

Illustration 7.49: Station de test 1 – V – Modèles linéaires, forêt et AROME - Box-plot RMSE, ECT, BIAIS et MAE pour l'échantillon de test

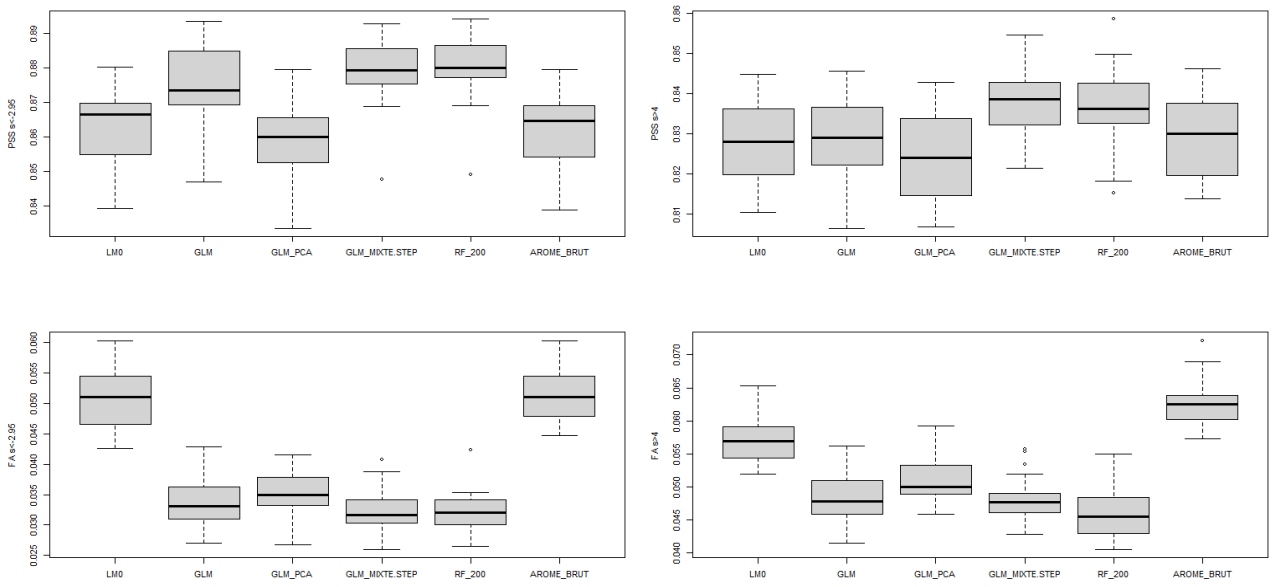

Illustration 7.50: Station de test 1 – V – Modèles linéaires, forêt et AROME - Box-plot PSS et FA pour l'échantillon de test

Le tableau 7.20 récapitule les scores de validation croisée sur le jeu de test pour tous les modèles de V.

Tableau 7.20 : Station de test 1 – V – Modèles linéaires, forêt, Réseau de neurone – Scores de validation croisée sur l'échantillon de test

Modèle	RMSE	ECT	BIAIS	MAE	PSS1	FA1	PSS3	FA3
LM0	1.443	1.443	0.002	1.061	0.863	0.051	0.828	0.057
GLM	1.29	1.29	0	0.944	0.876	0.033	0.829	0.048
GLM_PCA	1.378	1.378	0.004	1.009	0.86	0.035	0.824	0.051
GLM_MIXTE.STEP	1.273	1.273	0	0.93	0.879	0.032	0.838	0.049
RF_200	**1.269**	**1.269**	**0.003**	**0.915**	**0.88**	**0.032**	**0.836**	**0.046**
AROME.BRUT	1.446	1.445	-0.036	1.072	0.862	0.051	0.83	0.063
RNA	1.281	1.280	0	0.937	0.874	0.031	0.813	0.037

La même analyse réalisée sur les scores de test de U a été faite sur les scores de V, et les tendances n'ont pas changé pour les deux stations. Les meilleurs modèles pour la prédiction de V restent le GLM, GLM_MIXTE.STEP et RF_200, avec toujours un avantage pour le RF_200.

7.5.2.1.3 Roses des vents après reconstitution de DD à partir de U et V

Pour les deux stations de test, les scores RMSE, ECT, MAE, BIAIS, PSS et FA ont permis de sélectionner les trois modèles GLM, GLM_MIXTE.STEP et RF_200 comme meilleurs modèles statiques pour la prédiction de U et V.

Cependant, il est très difficile de départager les modèles sur les critères d'évaluations identifiés pour l'estimation statistique de direction du vent. **Nous avons donc décidé de sélectionner le modèle statistique d'estimation de DD sur sa capacité de restitution de la rose des vents** : c'est-à-dire que parmi les roses des vents établies à partir des modèles statistiques, celle qui se rapprochera le plus de la rose des vents de l'observation verra sa méthode choisie pour l'estimation de DD. Le FF des roses des vents est celui du modèle RF_100 pour une comparaison optimale sur DD.

Les graphiques de l'illustration 7.51 montrent les roses des vents issues de la combinaison des différents modèles d'estimation des composantes U et V de DD.

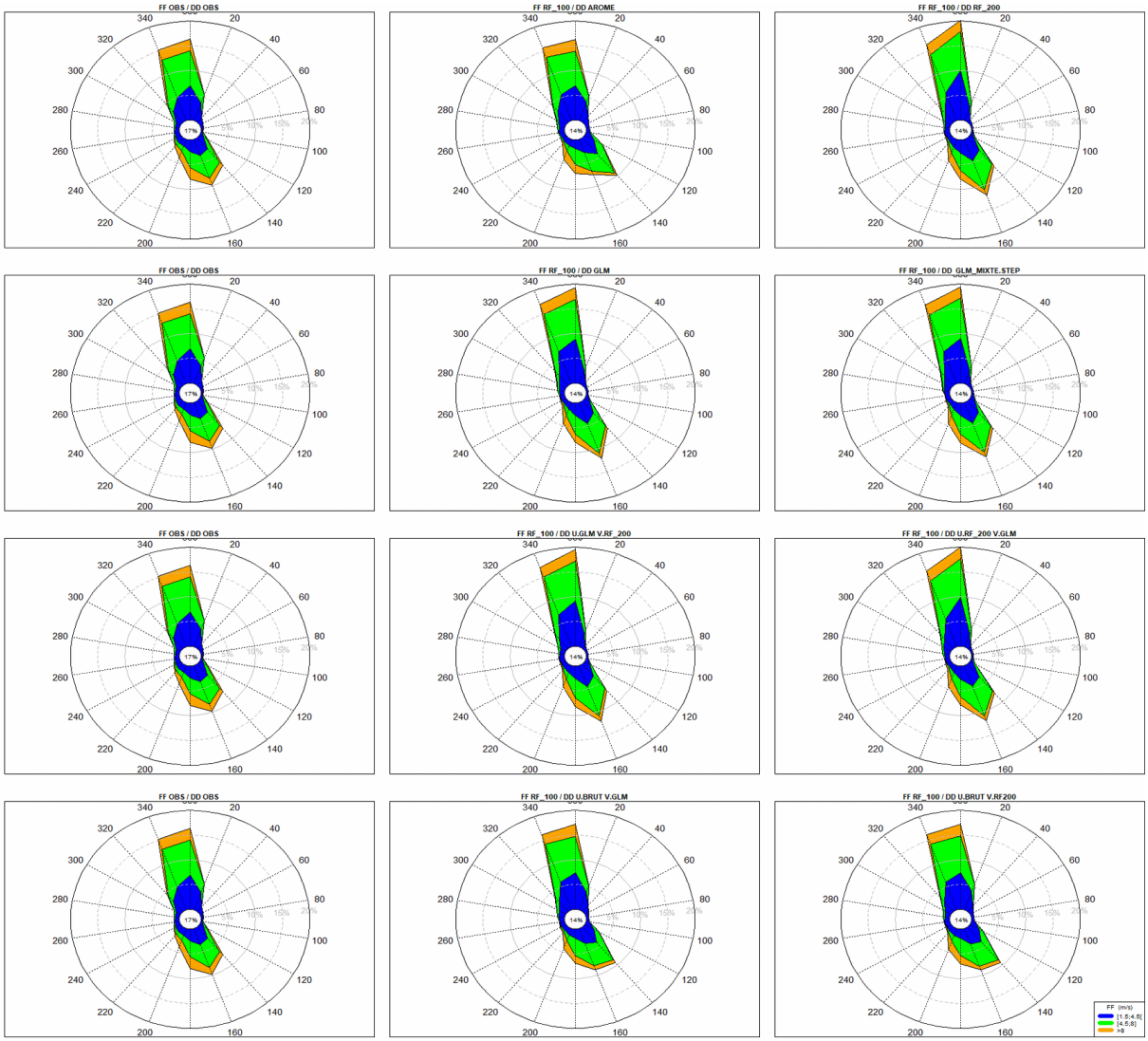

Illustration 7.51: Station de test 1 – Roses des vents issues des combinaisons de U et V des modèles statistiques: sur la 1ère colonne l'observation ; sur la 2ème colonne (de haut en bas) on retrouve DD AROME brut (U.BRUT et V.BRUT), DD GLM (U.GLM et V.GLM), U.GLM et V.RF_200, U.BRUT et V.GLM; et sur la 3ème colonne on retrouve DD RF_200 (U.RF_200, V.RF_200), DD GLM_MIXTE.STEP (U.GLM_MIXTE.STEP, V.GLM_MIXTE.STEP), U.RF_200 et V.GLM, U.BRUT et V.RF_200

Pour la station de test 1, les secteurs prépondérants pour le choix du modèle statistique sont les secteurs situés dans le nord (notamment de 340° à 360°), et dans le sud (notamment de 140° à 200°) où AROME brut restitue moins bien la rose de vent. Cependant, à part les roses où on utilise la composante U de AROME (roses issues de U.BRUT V.GLM, et U.BRUT V.RF_200), les roses issues des modèles statistiques sont très proches les unes des autres. En effet, elles sont meilleures que les roses impliquant AROME pour les vents du sud, et moins biens que ce dernier pour les vents du nord. Ainsi, c'est les scores B95+ (notamment le critère C2 qui concerne la direction, et le critère C4 qui concerne la corrélation circulaire) qui seront décisifs pour choisir le modèle statistique d'estimation de DD.

Le tableau 7.21 présente le récapitulatif des scores B95+ pour les différents modèles.

Tableau 7.21: Station de test – Scores B95+ des roses de vents issues de la combinaison de U et V des modèles statistiques (en vert le modèle choisi)

Modèle	C1	C2	C3	C4 (corrélation circulaire)
OBS - BRUT	85.65	89.31	92.72	19.93
OBS - RF_200	**85.34**	**89.59**	**92.72**	**64.19**
OBS - GLM	84.94	89.08	92.72	61
OBS - GLM_MIXTE.STEP	84.94	89.28	92.72	61.27
OBS - U.GLM-V.RF_200	84.82	89.03	92.72	61.57
OBS - U.RF_200-V.GLM	85.08	89.32	92.72	63.77
OBS - U.BRUT-V.GLM	86.74	90.59	92.72	22.2
OBS - U.BRUT-V.RF_200	87.17	91.1	92.72	21.25

Les scores sont favorables au modèle RF_200 qui se démarque avec un très bon C2 et la meilleure corrélation circulaire C4. Par conséquent, **nous avons décidé de choisir le modèle RF_200 pour l'estimation de DD pour la station de test 1**.

La reconstitution de DD à partir de U et V pour la station de test 2a donné des résultats similaires. C'est en effet le modèle RF_200 qui a été choisi pour l'estimation de DD.

Des tests sur le réseau de forêt à 200 arbres pour U et V avec et sans les prédicteurs de régimes de temps ont également été réalisés pour le deux stations de test. Comme pour la force du vent, l'apport des régimes de vent n'est pas suffisamment notable pour rajouter ces variables explicatives supplémentaires.

7.6 Synthèse du choix des modèles

Suite aux études des différents modèles statistiques, **le modèle de forêt aléatoire à 100 arbres (RF_100) a été retenu pour étendre la série horaire de FF et le modèle de forêt à 200 arbres (RF_200) pour étendre la série de DD pour chacune des deux stations de test**.

Les modèles statistiques d'estimation de FF et DD ont été appris sur les données AROME de l'année 2017 pour la station de test 1, et celles de l'année 2018 pour la station de test 2.

Le graphique ci-dessous (illustration 7.52) montre les roses finales d'extension entre 2018 et 2020 de la station de test 1.

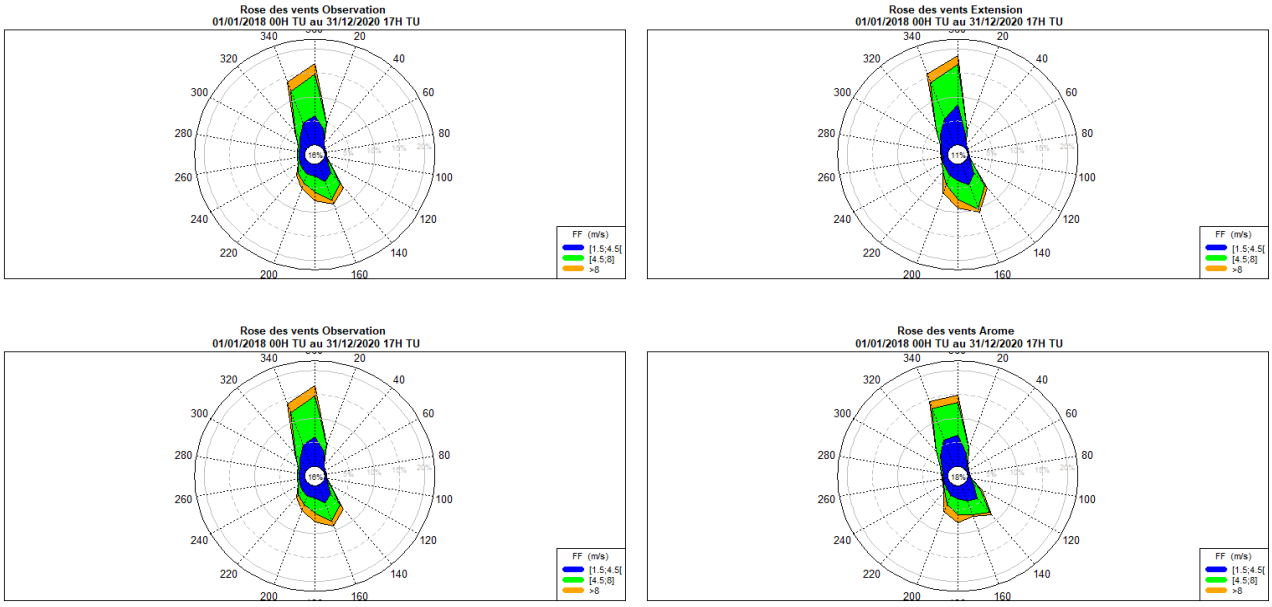

Illustration 7.52: Station de test 1 – Roses des vents finales de l'extension entre l'année 2018 et 2020

Sur cette période d'extension (2018 à 2020), la rose des vents de l'extension restitue mieux les vents du sud que AROME. En revanche, il surestime légèrement la force du vent dans le secteur nord, et sous estime la proportion de vent très faibles (FF < 1.5 m/s). **Globalement c'est la rose issue du modèle statistique qui est plus proche de l'observation.**

Comme pour la station de test 1, c'est la rose issue du modèle statistique qui est plus proche de l'observation dans l'ensemble pour la station de test 2. L'illustration 7.53 présente les roses finales d'extension entre 2017 et 2020 (sans les données de l'année d'apprentissage 2018) de la station de test 2.

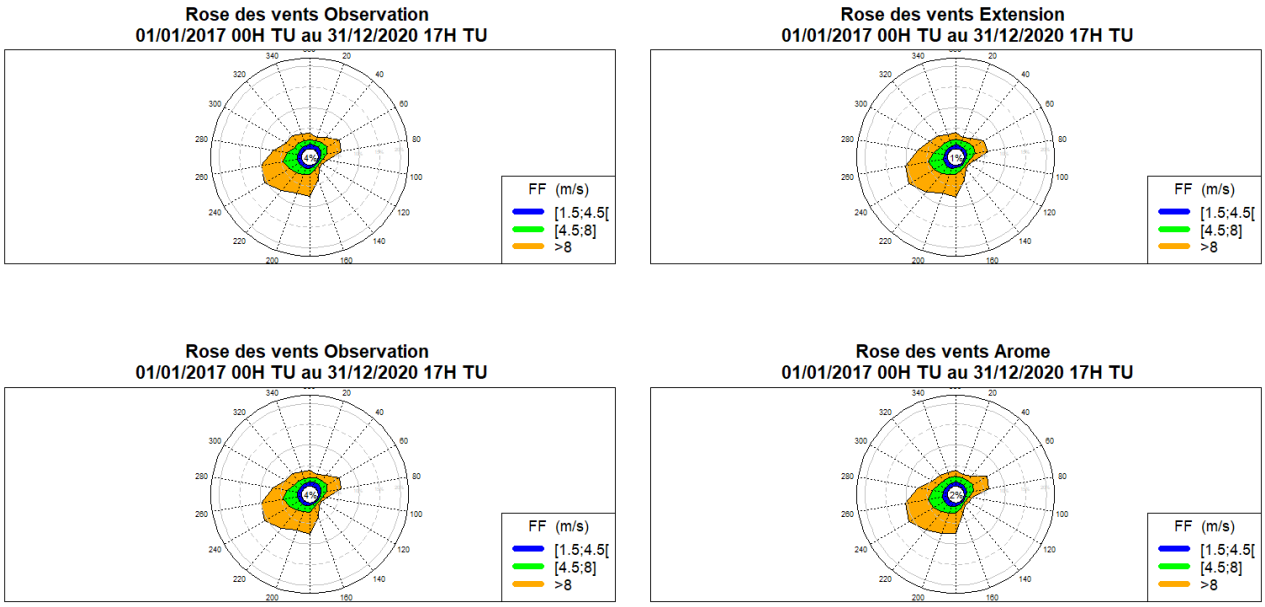

Illustration 7.53: Station de test 2 – Roses des vents finales de l'extension entre l'année 2017 et 2020 (sans les données d'apprentissage de l'année 2018)

7.7 Limites de l'extension

Pour la station de test 1, l'extension avec le modèle statistique améliore la rose des vents AROME, autant sur la force (secteur 340° à 0°) que sur la direction (secteur 140° à 180°). Cependant, la fréquence des vents très faibles restent toujours moindres pour l'extension comme on peut l'apercevoir sur les graphiques de l'illustration 7.54 qui présentent les histogrammes de FF par classe de vent par pas de 5 m/s.

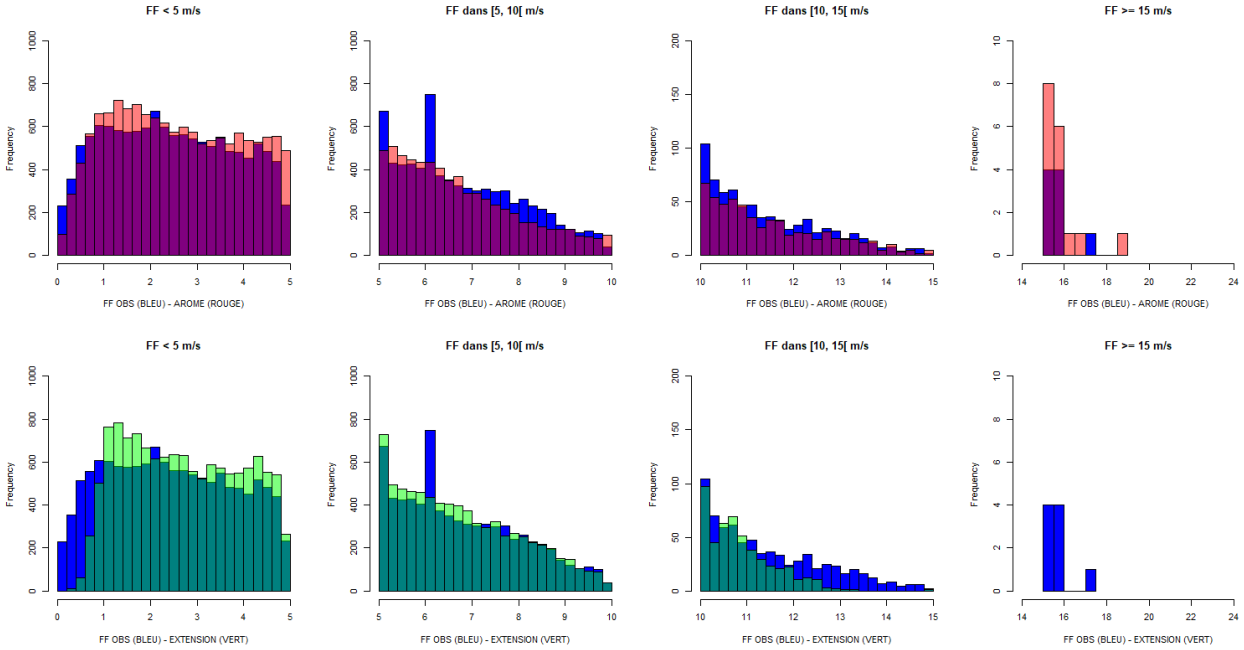

Illustration 7.54: Station de test 1 – Histogrammes de FF par classe de vent par pas de 5 m/s pour les années 2018 à 2020. En haut, observations (bleu) comparées à AROME (rouge). En bas, observations comparées à l'extension (vert).

Les vents très faibles sont moins fréquents pour l'extension que pour AROME et l'observation. Ils correspondent globalement à des forces de vents inférieurs à 1.5 m/s. Par ailleurs, les vents forts ($FF > 15\,m/s$) sont également moins fréquents (voire inexistant pour la période de 2018-2020) pour l'extension comparativement à AROME et l'observation où l'on trouve peu de vent fort.

L'illustration 7.55 présente les histogrammes de FF par classe de vent par pas de 5 m/s pour la station de test 2.

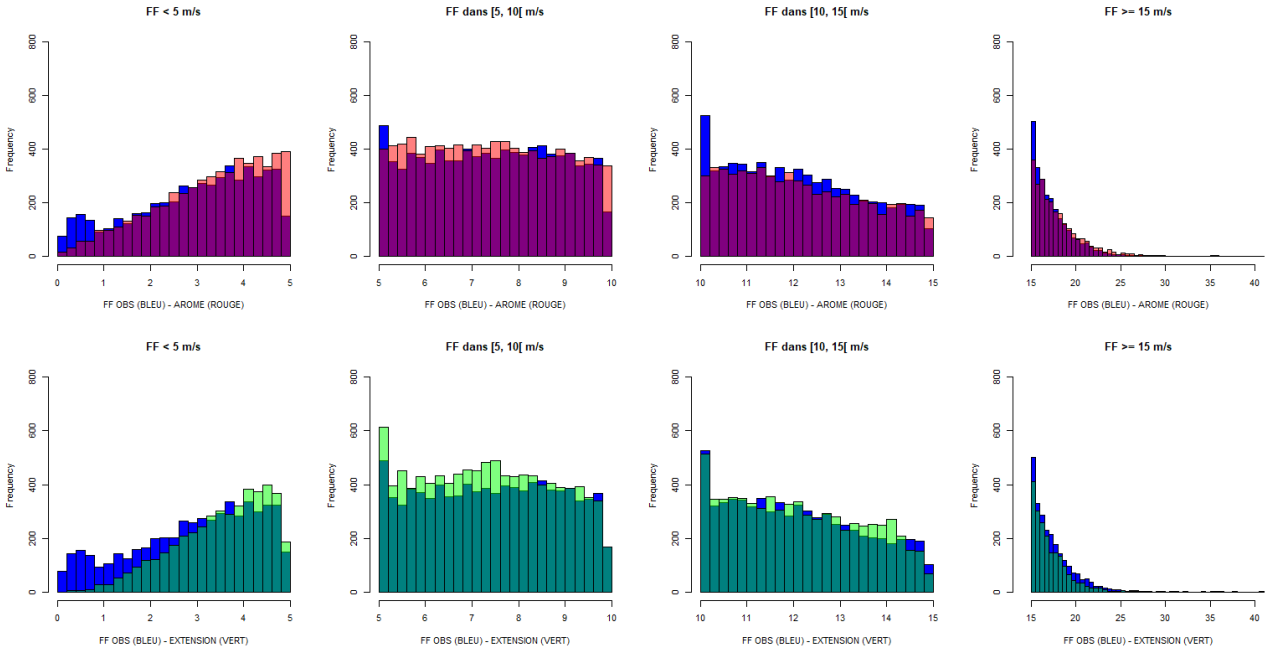

Illustration 7.55: Station de test 2 – Histogrammes de FF par classe de vent par pas de 5 m/s pour les années 2017 à 2020 (sans les données d'apprentissage de l'année 2018). En haut, observations (bleu) comparées à AROME (rouge). En bas, observations (bleu) comparées à l'extension (vert).

Pour la station de test 2, les vents très faibles sont également sous-estimés par l'extension avec le modèle statistique. Aussi les vents forts sont légèrement moins fréquents que pour AROME, mais se comportent un

peu mieux qu'avec les vents forts que la station de test 1. On note cependant qu'il y a plus de vents forts dans les observations de la station de test 2 que dans celles de la station 1.

7.8 Conclusion de l'étude d'optimisation

L'objectif de cette étude complète sur les stations de test était de valider l'utilisation d'une méthode statistique unique et optimale pour étendre la série horaire d'observation.

En complément des méthodes statistiques utilisées sur les deux campagnes de mesures précédentes (Dunkerque et Oléron), nous avons exploré d'autres variables explicatives (régime de temps), une autre méthode de sélection des variables explicatives (l'ACP) et deux autres méthodes statistiques (modèle linéaire avec anamorphose et réseau de neurone). Excepté le modèle linéaire avec anamorphose, toutes les méthodes testées ont donné des résultats satisfaisants mais ce sont les forêts aléatoires qui se sont démarquées autant sur la force que sur la direction du vent.

Dans la suite de nos études, nous utiliserons la forêt aléatoire pour étendre nos séries de mesure de vent à 100 m. Pour chaque extension, une comparaison au modèle AROME sera effectuée pour qualifier l'apport de la forêt aléatoire.

FIN DE DOCUMENT