Post-traitement des prévisions d'ensemble en météorologie par des méthodes d'apprentissage statistique

Gabriel Jouan

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Mathématiques et leurs Interactions (MI)*

Par
Gabriel JOUAN

Post-traitement des prévisions d'ensemble en météorologie par des méthodes d'apprentissage statistique.

Thèse présentée et soutenue à Rennes, le 17 Décembre 2021
Unité de recherche : IRMAR - UMR CNRS 6625

Rapporteurs avant soutenance :

Thomas ROMARY MCF, Mines Paris Tech, Paris
Sylvie PAREY Ingénieur, EDF/R&D, Palaiseau

Composition du Jury :

Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Président :	François COQUET	PR, ENSAI, IRMAR - UMR CNRS 6625, Bruz
Examinateurs :	Maëlle NODET	MCF, UVSQ - UMR 8100, Versailles
	Nicolas COURTY	PR, Université Bretagne Sud, IRISA, Vannes
	Thomas ROMARY	MCF, Mines Paris Tech, Paris
	Sylvie PAREY	Ingénieur, EDF/R&D, Palaiseau
Dir. de thèse :	Valérie MONBET	PR, Université Rennes 1, CNRS, IRMAR - UMR CNRS 6625, Rennes

Invité(s) :

Co-encadrants :	Anne CUZOL	MCF, Université Bretagne Sud, UMR 6205, LMBA, Vannes
	Goulven MONNIER	Ingénieur, Scalian, Rennes

Remerciements

Arrivé à cette étape de rédaction, les mots me manquent. Ces dernières lignes impliquent de conclure le travail le plus enrichissant, mais aussi le plus dur que j'ai eu à mener. Ce travail n'aurait pas pu être concrétisé sans les différentes personnes qui m'ont aiguillé dans ma recherche, ma façon de penser, mais également dans ma motivation.

Pour commencer, j'aimerais dire merci aux différentes personnes qui ont participé au projet que ce soit par l'intermédiaire des différentes réunions d'équipe, mais également dans la conception des solutions et applications abordées dans cette thèse.

Ensuite, je remercie ma directrice de thèse Valérie Monbet et mes encadrants Anne Cuzol et Goulven Monnier. Le travail effectué à vos côtés m'a permis d'engranger un grand nombre de connaissances, mais également de revoir ma façon de penser. Il me reste encore du chemin à faire, mais si j'ai pu au cours de cette thèse développer mon expertise en mathématiques appliquées, c'est grâce à vous. Ces quelques années ont été enrichissantes à la fois scientifiquement et humainement parlant. Durant la crise de COVID-19, j'ai pu compter sur votre présence et soutien malgré la situation sanitaire mondiale. Vous m'avez accompagné jusqu'au bout en me conseillant toujours avec bienveillance et pour cela, je tenais également à vous dire merci.

Aux membres de la communauté SWGEN, je tenais aussi à dire merci, notamment à Pierre Aillot, Philippe Naveau, Julie Bessac et Nicolas Raillard. Les échanges avec vous constituent mon premier retour d'expérience et ont guidé mes premiers pas sur ce grand sujet des ensembles de prévision. De plus, votre bienveillance et votre bonne humeur m'ont fait prendre un grand plaisir à participer aux différentes conférences scientifiques et réunions de travail.

Je remercie les membres de mon CSI, Maxime Taillardat et Nicolas Courty. Vos disponibilités pour suivre les travaux effectués chaque année représentent une importante contribution. Les discussions que nous avons eues au cours de ces réunions sur ce domaine de recherche ont toujours été un réel plaisir.

Merci au groupe Scalian pour le support et financement accordé à l'élaboration de

cette thèse. Plus particulièrement, je remercie les ingénieurs et jeunes chercheurs du LAB Scalian pour leurs temps, leurs idées et leurs bonnes humeurs. La réunion hebdomadaire du LAB a toujours été un événement incontournable de la semaine, permettant d'échanger sur divers projets et de bénéficier de l'expertise de chacun. Des réunions d'autant plus importantes et interminables, lorsqu'il s'agissait de comprendre le fonctionnement des congés N, N-1, RTT, RTTE. Je remercie également mes responsables LAB, Didier Rozzonelli et Rémi Poisvert. Depuis le début, vous avez toujours été d'un grand soutien et vous êtes montrés très compréhensifs au sujet de la durée des travaux de cette thèse.

Merci à Laurent Bessard, les ingénieurs Scalian et stagiaires informatiques pour leur expertise et le travail fourni dans la réalisation des tâches purement informatique de cette thèse. Ton management Laurent avec la motivation de Julien, Khalid et Alexis a été décisif pour aboutir à une super application des travaux de cette thèse. De plus, merci d'avoir été patient pour m'expliquer le fonctionnement de github/gitlab. Sans ton enseignement, je serais toujours perdu dans l'interminable labyrinthe des commandes git.

Merci à Thierry Daubos pour son temps et aussi ses connaissances en algorithme d'apprentissage. Ton savoir et ton retour sur ma rédaction m'ont permis de m'améliorer.

Merci à Valentin Resseguier pour son soutien moral et mathématique. Partager le bureau avec toi pendant ces quelques années, perturbées sur la fin par la COVID, a été d'un grand support. Ta rigueur m'a permis d'approfondir ma pensée scientifique. De plus, ton implication débordante et ta joie de communiquer sur tes travaux de recherche rendaient les journées plus enrichissantes.

Pour continuer, je remercie la structure d'accueil de l'agence Scalian de Rennes, mais aussi celle de l'IRMAR. Dans ces bâtiments, j'ai pu rencontrer des personnes bienveillantes, à la bonne humeur contagieuse. Notamment, je suis heureux d'avoir rencontré les personnes du CEN Simu de Scalian Rennes, dont leur capacité étonnante est de pouvoir transformer une journée maussade en bonne journée par une simple anecdote à la pause-café. Les vendredis restaurants, bien que logistiquement durs à organiser, ont toujours été d'excellents moments à partager en votre compagnie.

Je tiens également à remercier les doctorants présents aux alentours de l'IRMAR pour leur soutien et retour d'expérience. Merci à Marie Morvan pour sa bonne humeur et son expertise en mathématiques. Tes retours sur les travaux effectués ont permis d'affiner la rédaction de ce manuscrit.

Mes prochains remerciements concernent ma famille, mes amis de lycée, mes amis d'INSA. Je tiens à remercier du fond du coeur, mes parents, frères et soeurs qui ont toujours eu confiance en moi et sans qui ces travaux n'auraient pas eu lieu. Ensuite, mes amis de lycée ont toujours su être disponibles autour de moi, jusqu'à venir me chercher chez moi malgré ma condition d'ermite durant cette rédaction. Pour cela, je les remercie chaleureusement. À cela, j'ajoute de grands mercis à mon ancien colocataire et aussi à mon ancien binôme INSA, Gabriel et Gabriel qui ont été très présents pour m'épauler et s'intéresser à ces travaux. Je remercie également PM pour sa compagnie du midi permettant de penser à d'autres choses le temps d'un repas.

Avant de terminer, j'aimerais ajouter un remerciement à mon ancien directeur de département de génie mathématique de l'INSA Rennes, James Ledoux, qui m'a toujours soutenu du début jusqu'à la fin dans cette longue route qu'est la thèse. Je remercie également, toutes les personnes, amis et collègues qui m'ont soutenu, conseillé et aidé durant ces dernières années.

Enfin, j'adresse un immense merci à ma compagne Léa. Durant ces quelques années, tu as toujours su me soutenir et trouver les mots dont j'avais besoin. Ta présence a été un élément moteur important dans cette thèse et sa rédaction.

TABLE DES MATIÈRES

Introduction

Contexte

De nos jours, la prévision météorologique est d'une importance économique reconnue pour des secteurs comme l'agriculture, la production d'énergie renouvelable, la programmation d'opérations de maintenance, la gestion des risques ou encore pour l'organisation d'évènements de particuliers (vacances, loisirs, etc). Depuis le siècle dernier, des prévisions numériques peuvent être obtenues grâce à des modèles numériques basés sur les lois de la physique et initialisés à partir d'observations météorologiques. Cependant, ces modèles génèrent une trajectoire de prévision unique. Dès les années 60, les travaux de Lorenz 1963 et Lorenz 1965 ont mis en évidence le caractère chaotique de la météorologie. Le fait de décrire les états futurs de l'atmosphère à partir d'une seule trajectoire de prévision est donc restrictif. Bien que le coût numérique de production de plusieurs trajectoires soit important, il existe aujourd'hui des super calculateurs qui permettent de produire jusqu'à plusieurs dizaines de trajectoires. Ainsi, depuis quelques années, les centres de prévision météorologique produisent des ensembles de prévisions, autrement dit plusieurs trajectoires probables de l'état de l'atmosphère. Les ensembles de prévisions sont des simulations de Monte-Carlo obtenues en perturbant aléatoirement les conditions initiales et une partie des paramètres du modèle numérique (Epstein 1969 ; Leith 1974 ; Hoffman et Kalnay 1983 ; Buizza, Milleer et Tim N Palmer 1999 ; Stensrud, Bao et Warner 2000 ; Buizza, Leutbecher et Isaksen 2008). Les ensembles ainsi générés donnent une approximation de la distribution probabiliste des variables atmosphériques et permettent de quantifier l'incertitude des prévisions déterministes. Ces approximations peuvent présenter des biais et/ou des problèmes de sous ou sur dispersion (Bougeault et al. 2010 ; Park, Buizza et Leutbecher 2008 ; Hamill et Colucci 1997 ; Hamill et Whitaker 2006). La figure 1 donne une illustration de ces différents types d'erreurs. La série temporelle des observations de moyennes journalières de températures à Millau (France) est affichée en noir, et les boîtes à moustache décrivent la distribution des ensembles issus des prévisions à 3 jours. Les rectangles noirs mettent en évidence des situations spécifiques correspondant à un biais ou un problème de dispersion de l'ensemble de prévisions.

Les caractéristiques de ces erreurs peuvent être identifiées à partir de l'histogramme de rang (Anderson 1996 ; Talagrand, R. Vautard et Strauss 1997). L'histogramme correspond à la loi uniforme quand l'observation est une réalisation très probable de la distribution de l'ensemble. Un histogramme présentant une forme en "L" ou en "L" inversé est typique d'une distribution d'ensemble biaisée, c'est à dire une erreur en moyenne. Sur la figure 1, les temps 12 à 15 sont associés à cette forme et les boîtes à moustaches montrent un biais négatif. Les formes en $\cap$ (ou $\cup$) correspondent à des cas de sur (ou sous) dispersion des ensembles (Hamill et Colucci 1997) c'est à dire une erreur en variance. Des exemples de problèmes de dispersion sont montrés au début du mois pour la sur-dispersion et à partir du 24ème jour pour la sous-dispersion.

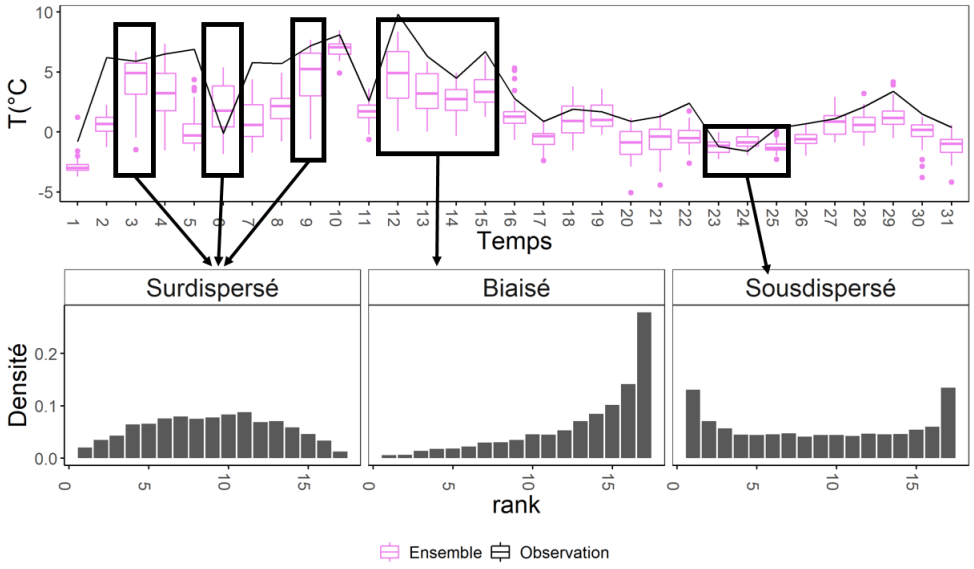


Figure 1 – Janvier 2015, 18H à Millau, ensemble de prévisions CEPMMT d'horizon de prévision 3 jours et observations de températures. Première ligne : série observée et boîtes à moustaches des ensembles de prévisions ; deuxième ligne : histogrammes de rangs illustrant des situations typiques d'erreurs.

Des méthodes de post-traitement statistique ont été construites dans l'objectif de corriger ces erreurs de distribution des ensembles. Cette approche est plus communément appelée calibration d'ensembles de prévisions. Elle vise à tirer parti du lien entre observations météorologiques et distributions des ensembles. Le problème de calibration a d'abord été posé dans un cadre univarié. L'objectif était de corriger la distribution d'ensemble d'une variable météorologique en un site à l'aide de modèles statistiques (Daniel

S Wilks 2018). Cependant, les phénomènes météorologiques sont spatio-temporels et la dépendance entre différentes variables météorologiques peut être forte. Ainsi, progressivement, de nouvelles approches de calibration ont été proposées pour mieux prendre en compte les divers types de dépendances (Schefzik et Möller 2018). Par exemple, (Clark et al. 2004 ; Feldmann, Scheuerer et Thorarinsdottir 2015 ; Rasp et Lerch 2018) ont développé des méthodes de calibration spatiale, (Pinson, Madsen et al. 2009 ; Scheuerer, Hamill et al. 2017) se sont intéressés plus particulièrement à la dépendance temporelle et (Ben Bouallègue et al. 2016) aux interactions entre variables météorologiques. Les approches de calibration abordées dans ces différents travaux visent à approcher des représentations réalistes de ces dépendances, les applications visées pouvant être par exemple : l'aide à la régulation du remplissage de bassins versant, la prédiction d'événements rares entraînant des précipitations importantes (Clark et al. 2004 ; Scheuerer, Hamill et al. 2017), ou encore l'amélioration de la production énergétique issue du secteur des énergies renouvelables (Pinson, Madsen et al. 2009 ; Ben Bouallègue et al. 2016).

Dans cette thèse, notre premier objectif est de proposer des méthodes de calibration multivariée pour la planification d'évènements ou d'opérations de maintenance. Les problèmes de calibration multivariée sont classiquement abordés de deux manières. La première consiste à considérer un ensemble de calibrations univariées, puis d'appliquer une procédure de réarrangement pour améliorer la structure multivariée des ensembles calibrés. L'alternative consiste à modéliser la loi jointe puis de corriger ses paramètres pour que la loi de l'ensemble coïncide avec celle des observations. La difficulté est alors de proposer un modèle de loi jointe suffisamment flexible pour bien décrire la distribution. Dans cette thèse, une alternative est proposée : le cadre multivarié est d'abord simplifié en discrétisant l'ensemble de définition de la loi jointe. On se ramène alors à un problème de classification avec une sortie univariée, chaque classe correspondant à une partie de la distribution jointe. Après un rappel des méthodes de calibration univariée et multivariée dans le chapitre 1, cette nouvelle méthode de calibration est l'objet du chapitre 2 :

Contribution 1 (chapitre 2)

- Une méthode originale de calibration multivariée est proposée dans laquelle l'espace des données est discrétisé de façon à transformer un problème de régression multi-sorties en un problème de classification à sortie unique.
- La méthode est testée sur des données réelles dans un contexte de planification d'évènements.
- Ces travaux ont donné lieu à une publication dans les actes d'une conférence internationale :
  Jouan, G. Cuzol, A., Monbet, V., Monnier, G. (2019) Weather type prediction at medium range from ensemble forecasts. 9th International workshop on Climate Informatics, Oct 2019, Paris, France.

Cette approche donne des résultats encourageants, mais les modèles obtenus restent difficiles à interpréter physiquement. Dans le chapitre 3, une méthode différente est donc proposée, partant de l'hypothèse que les erreurs des ensembles dépendent de régimes météorologiques. Par exemple, on peut imaginer que dans une situation de dépression hivernale les ensembles vont avoir tendance à surestimer la dispersion de l'intensité du vent alors que dans une situation anticyclonique d'été, la température de l'air peut être sous-estimée. Une méthode de calibration en deux étapes est alors proposée : une première étape de classification vise à identifier les régimes météorologiques, la seconde à corriger la distribution d'ensemble dans chaque régime. L'étape de classification est basée sur des modèles de mélange gaussiens. Cependant, la nature particulière des observations sous forme d'ensembles a nécessité d'adapter les méthodes d'inférence classiques pour ce type de modèles. Ces nouveaux développements sont exposés dans le chapitre 3 :

Contribution 2 (chapitre 3)

- Une méthode originale de calibration univariée est proposée dans laquelle la correction des ensembles dépend des régimes météorologiques.
- Des méthodes d'inférence ont été développées pour estimer les paramètres d'un modèle de mélange gaussien pour les données d'ensemble.
- Ces travaux ont donné lieu à un article soumis :
  Jouan, G., Cuzol, A., Monbet, V., Monnier, G. (2021) Gaussian mixture models for clustering and calibration of ensemble weather forecasts.
  et au développement d'un package R disponible sous github :
  `https://gitlab.com/gabrijou/gaussianmixturemodels`

Enfin, il est important de souligner que la plupart des méthodes de correction de la distribution d'ensemble se focalisent sur des horizons de prévision courts (moins de 48h). Néanmoins pour la programmation d'opérations de maintenance par exemple, il y a un réel intérêt à corriger les prévisions à moyenne échéance (3 à 10 jours) (Pinson, Madsen et al. 2009) bien que le problème soit plus difficile (Pinson et Girard 2012 ; Scheuerer et Hamill 2015a ; Bremnes 2019). Ainsi, **notre second objectif est de prouver l'applicabilité des méthodes de calibration proposées à des échéances de prévision de 3 à 10 jours.** Les méthodes de calibration présentées dans les chapitres 2 et 3 sont donc évaluées sur des données réelles dans ce contexte de prévision à moyen terme.

Finalement, **un troisième objectif opérationnel est de produire une interface web pour afficher les prévisions d'ensemble et fournir une information à l'utilisateur sur leur qualité**. Le développement de cette interface est décrit dans le chapitre 4. Dans un second temps, l'objectif serait d'intégrer à cet outil les méthodes de calibration proposées dans les chapitres 2 et 3.

Contribution 3 (chapitre 4)

— Une interface web a été développée pour fournir des prévisions météorologiques à moyen terme avec des informations sur la qualité de la prédiction issues des ensembles de prévision.

Chapitre 1

Méthodes statistiques appliquées à la calibration d'ensembles de prévision univariés et multivariés

Comme décrit dans l'introduction, les prévisions et ensembles de prévision sont soumis à d'importantes incertitudes. Les incertitudes rencontrées sur les ensembles de prévision se présentent généralement sous forme de biais ou de problèmes de dispersion. Ces types d'erreurs peuvent être corrigés à l'aide de modèles statistiques et d'observations météorologiques locales. L'étape de correction ainsi appliquée est connue sous le nom de calibration d'ensembles de prévision.

Durant les dix dernières années, plusieurs méthodes de calibration ont été proposées, allant du simple modèle linéaire aux approches d'apprentissage profond (voir Vannitsem et al. 2021 pour une revue de la littérature). La section 1.1 introduit le problème de calibration dans un contexte univarié en présentant des méthodes paramétriques et non paramétriques de référence dans ce domaine. Abordé en section 1.1.1, un des travaux précurseurs, qui fait maintenant office de référence dans l'étude des modèles de calibration univariée, est basé sur un post-traitement sous forme de régression statistique. Plus précisément, ce modèle défini par Klein, Lewis et Enger 1959 ; Gneiting, Raftery et al. 2005 est appelé Adaptation Statistique d'Ensemble (NGR, "Nonhomogeneous Gaussian Regression" ; EMOS, "Ensemble model output statistics"). Le modèle NGR construit une distribution permettant de prédire les futures quantités de variables météorologiques d'intérêt à l'aide de paramètres dépendant des ensembles. L'avantage principal de ce type de modèle réside en sa facilité d'implémentation et d'adaptation aux multiples variables météorologiques. Des variables telles que les vitesses du vent et les précipitations, ayant chacune une distribution atypique, sont régulièrement soumises à des incertitudes, provoquées par les approximations locales du modèle numérique. Or, ces variables représentent un fort intérêt économique pour la gestion de production énergétique ou la prévision de

risque. S. Baran et Lerch 2016 proposent de construire un modèle NGR à partir d'un mélange entre une loi normale tronquée positivement et une loi log-normale pour les vitesses du vent. Pour les précipitations, Scheuerer et Hamill 2015a proposent de les calibrer à l'aide d'une extension de la loi gamma. Les deux modèles NGR proposés pour les vitesses du vent et les précipitations sont étudiés en fin de section 1.1.1 en vue d'une première application de la calibration sur des données réelles.

Les erreurs de distribution des ensembles montrent parfois des comportements non linéaires difficiles à modéliser et corriger. Les approches linéaires telles que les modèles NGR atteignent ici leurs limites. Récemment, des méthodes non paramétriques, plus flexibles dans leur possibilité d'application, ont été étudiées par la communauté de calibration (Daniel S Wilks 2018). En particulier, l'approche de forêt aléatoire, qui effectue une régression non linéaire pour estimer les quantiles de la loi visée (QRF, "Quantile Random Forest", Meinshausen 2006), a été étudiée dans une cadre d'une calibration univariée par Taillardat, Mestre et al. 2016. Cet algorithme a apporté des résultats intéressants de calibration, permettant d'exploiter des informations sur l'ensemble, difficiles à modéliser par les modèles NGR, et d'obtenir un ensemble post-traité. L'approche de forêt aléatoire de régressions et d'estimation de quantiles QRF sera abordée en section 1.1.2.2.

Cependant, les phénomènes météorologiques sont spatio-temporels et la dépendance entre différentes variables météorologiques peut être forte. Ainsi, progressivement, de nouvelles approches de calibration ont été proposées pour mieux prendre en compte les divers types de dépendances (Schefzik et Möller 2018). Par exemple, Clark et al. 2004 ; Feldmann, Scheuerer et Thorarinsdottir 2015 ; Rasp et Lerch 2018 ont développé des méthodes de calibration spatiales, (Pinson, Madsen et al. 2009 ; Scheuerer, Hamill et al. 2017) se sont intéressé plus particulièrement à la dépendance temporelle et (Ben Bouallègue et al. 2016) aux interactions entre variables météorologiques. Les approches de calibration abordées dans ces travaux visent à approcher et reproduire des représentations réalistes de ces dépendances, pour servir dans des applications comme l'aide à la régulation du remplissage de bassins versants, à la prédiction d'événements rares entraînants des précipitations importantes (Clark et al. 2004 ; Scheuerer, Hamill et al. 2017), ou encore améliorer l'optimisation de la production d'énergies renouvelables (Pinson, Madsen et al. 2009 ; Ben Bouallègue et al. 2016). Les approches classiques de calibration multivariée sont basées sur les copules et copules empiriques. Ces méthodes visent à apprendre les dépendances des variables étudiées, pour ensuite les réintégrer dans des ensembles post-traités de manière univariée (Schefzik et Möller 2018). Le modèle

classique de Schaake shuffle (Clark et al. 2004), utilisant une approche de copule empirique, est introduit en 1.2. Cet algorithme est notamment connu pour sa facilité d'adaptation sur des problèmes aux dépendances complexes, mais ne fournit qu'une représentation simplifiée de celles-ci. Récemment, les approches de réseaux de neurones deviennent de plus en plus étudiées dans un contexte de calibration mutivariée. Notamment, l'aspect spatial présente un cadre idéal pour l'application de ce type de modèle pour la calibration d'ensembles de prévision de températures, vitesses du vent ou encore de couverture nuageuse totale (Rasp et Lerch 2018 ; Scher et Messori 2018 ; Bremnes 2020 ; A. Baran et al. 2021).

La section 1.3 pose un cadre d'application des modèles présentés pour la calibration d'ensembles de vitesses de vent et de précipitations. Les données d'ensemble de prévision issues du centre européen CEPMMT et observations SYNOP de stations in situ sont utilisées et ce pour trois différentes localisations spatiales et moyennes échéances de prévision. Les covariables décrivant les ensembles de prévision et nécessaires au modèle de forêt sont abordées. L'évaluation des performances de calibration des modèles s'effectue au travers de scores présentés dans cette section. Avant de conclure, les résultats des méthodes de calibration sont affichés et discutés concluant sur les difficultés rencontrées pour le problème de calibration multivariée.

1.1 Calibration d'ensembles univariés

Les ensembles de prévisions générés à partir d'un modèle numérique sont connus pour contenir des erreurs. En utilisant des observations issues de système de mesure de la météorologie locale, les erreurs de distributions des ensembles de prévisions peuvent être quantifiées et étudiées. Dès lors, le problème de calibration consiste à approcher la distribution des observations météorologiques locales conditionnées par les ensembles de prévision à l'aide de méthode statistique. La relation suivante résume l'objectif de la calibration :

$$f(X^*) = p(y|X) \tag{1.1}$$

où f représente un modèle statistique, $y \in \mathbb{R}$ une observation météorologique locale, X une variable ou un ensemble de variables construites à partir de l'ensemble de M prévisions (aussi appelées membres) noté $X^* \in \mathbb{R}^M$.

Dans la suite de cette section, l'approche classique pour un problème de calibration

univariée est présentée : le modèle paramétrique de régression gaussienne non homogène NGR. Le modèle NGR est très utilisé pour sa flexibilité et facilité d'implémentation. Cependant, ce genre d'approche paramétrique simple rencontre des difficultés à estimer des erreurs complexes de distributions d'ensembles. Pour dépasser cette limitation, une approche non paramétrique plus récente et plus précise a été proposée. Celle-ci se base sur le modèle de régression par forêt aléatoire de régression appliquée à l'estimation de quantiles.

1.1.1 Modèle NGR (EMOS)

Historiquement, le modèle de post-traitement linéaire "Non homogeneous Gaussian Regression" (NGR) ou "Ensemble model output statistics" (EMOS) a été proposé pour la régression d'ensembles de températures à localisation et échéance fixées. Dans un premier temps, le cadre général du modèle NGR est présenté. Puis, l'inférence associée est introduite. Enfin, les deux déclinaisons de modèles NGR utilisées pour la calibration des ensembles de vitesses du vent et des précipitations, sont exposées.

1.1.1.1 Définition

Le modèle NGR de Gneiting, Raftery et al. 2005 propose une régression non homogène pour approcher la relation conditionnelle entre les observations et l'ensemble. La méthode est constituée d'une première régression pour corriger les erreurs de biais entre l'observation réelle et l'observation conditionnée $y|X$. Ensuite, pour corriger les problèmes de dispersion, une seconde régression est effectuée entre la variance de la première régression et celle de l'ensemble. En reprenant les termes de Daniel S Wilks 2018 et à partir de la relation (1.1), le modèle NGR est défini par :

$$\begin{aligned} y &= \mu + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2) \\ \mu &= \beta_0 + \beta_1 x_1 + \cdots + \beta_M x_M \\ \sigma^2 &= \zeta_0 + \zeta_1 s^2 \end{aligned} \tag{1.2}$$

avec $s^2 = \frac{1}{M-1}\sum_{m=1}^{M}(x_m - \overline{x})^2$ la variance empirique non biaisée, $\overline{x} = \frac{1}{M}\sum_{m=1}^{M} x_m$ la moyenne empirique et $x_m \in \mathbb{R}$ une réalisation de l'ensemble X^*. Les coefficients des deux régressions sont notés $\beta_0, \cdots, \beta_M, \zeta_0, \zeta_1$. Contrairement à un modèle de régression classique, le modèle NGR prend en compte l'hétéroscédasticité de la variance des erreurs. Le

modèle (1.2) est adapté à la calibration d'ensembles constitués de membres déterministes, c'est-à-dire des membres issus de plusieurs modèles de prévision numérique non perturbés aléatoirement.

Dans le cas d'un ensemble échangeable, où chaque membre emporte la même information statistique liée à la distribution de l'ensemble, les coefficients de régression $\beta_1, \cdots, \beta_M$ sont identiques. L'expression du paramètre μ de (1.1) devient :

$$\mu = \beta_0 + \beta_1 \overline{x} \tag{1.3}$$

Le modèle NGR ainsi défini rend possible l'étude des probabilités d'apparition de quantiles spécifiques q :

$$p(y \leq q|X) = \Phi\left(\frac{q-\mu}{\sigma}\right) \tag{1.4}$$

avec $\Phi(.)$ la fonction de répartition de la loi normale centrée réduite. À partir de l'expression (1.4), une fonction génératrice peut être déduite afin de générer un nouvel ensemble d'observations conditionnées issu de la loi $y|X$. Cet ensemble est formé de M' quantiles correspondant à des probabilités généralement choisies équidistantes $\{\frac{1}{M'+1}, \cdots, \frac{M'}{M'+1}\}$. Un avantage à ne pas négliger est que le nouvel ensemble de M' réalisations peut être de taille différente par rapport à l'ensemble de M membres du modèle numérique.

Il existe plusieurs extensions à ce modèle, Hemri, Scheuerer et al. 2014 propose par exemple d'inclure une périodicité dans les covariables de la régression pour intégrer une dynamique saisonnière dans la correction des prévisions de températures. La flexibilité du modèle (1.1) rend possible sa déclinaison sous différentes hypothèses de lois conditionnelles $y|X$. Ainsi S. Baran 2014 présente différentes adaptations de NGR avec un mélange de lois normales tronquées dans l'étude des vitesses du vent. Dans le cas des prévisions de précipitations, Scheuerer et Hamill 2015a posent un modèle de régression gamma censuré pour gérer les prévisions et observations de précipitations.

1.1.1.2 Inférence

Pour un jeu de données d'observations et d'ensembles échangeables $\mathcal{X} = \{y_t, X_t^*\}_{1 \leq t \leq T}$ de taille T, la fonction de vraisemblance du modèle (1.2) s'exprime :

$$\mathcal{L}(\mathcal{X}; \mu, \sigma) = \prod_{t=1}^{T} p(y_t|X_t, \mu, \sigma) \tag{1.5}$$

avec $p(.)$ la densité de $y_t|X_t$ et μ, σ les paramètres du modèle NGR contenant les coefficients de régressions.

Gneiting, Raftery et al. 2005 montrent que l'approche du maximum de vraisemblance tend à produire une variance des ensembles corrigés sur-dispersée. Ils proposent une approche différente pour estimer les paramètres de régression en se basant sur la minimisation de la moyenne du score de probabilité des rangs continus (CRPS, "Continuous Ranked Probability Score", Matheson et Winkler 1976, Hersbach 2000). Le CRPS s'apparente à une mesure de distance entre une observation y_t et la distribution de l'ensemble conditionné $y_t|X_t$. Le CRPS moyen s'exprime de la façon suivante :

$$\begin{aligned} \overline{CRPS(\mathcal{X}; \mu, \sigma)} &= \frac{1}{T} \sum_{t=1}^{T} CRPS(y_t; \mu, \sigma) \\ CRPS(y_t; \mu, \sigma) &= \int_{-\infty}^{+\infty} (F(y_t \leq z|X_t, \mu, \sigma) - 1_{\{y_t \leq z\}})^2 dz \end{aligned} \tag{1.6}$$

avec F(.) la fonction de répartition de la distribution $y_t|X_t$.

Taillardat, Mestre et al. 2016 présentent plusieurs expressions analytiques de CRPS obtenues pour différentes lois conditionnelles $y_t|X_t$. Dans les travaux de S. Baran et Lerch 2016, la méthode d'estimation par maximum de vraisemblance est comparée à celle de minimisation de la moyenne du CRPS sur deux modèles NGR différents appliqués à la calibration des variables de vitesses du vent et de précipitations. Dans leurs conclusions, ils mettent en avant des propriétés très proches entre les deux méthodes d'inférence. Néanmoins, S. Baran et Lerch 2016 concluent sur le fait que l'estimation par minimisation de la moyenne du CRPS est plus appropriée dans le cas de données contenant des valeurs extrêmes ou aberrantes, car ce score est moins sensible.

1.1.1.3 Vitesses du vent

Pour modéliser les vitesses du vent (à valeurs dans $\mathbb{R}^+$), deux lois sont régulièrement utilisées. La première est une loi normale tronquée positivement, utilisée pour les vitesses du vent $y|X \sim \mathcal{N}_0(\mu_{TN}, \sigma^2_{TN})$ (Thorarinsdottir et Gneiting 2010). L'avantage de la loi normale tronquée est de pouvoir facilement représenter les situations classiques de vitesses du vent en utilisant directement les équations de régression (1.2) et (1.3) pour les

paramètres μ_{TN}, σ^2_{TN} :

$$\begin{aligned} \mu_{TN} &= \beta_{TN_0} + \beta_{TN_1}\overline{x} + \beta_{TN_2}x_{HRES} + \beta_{TN_3}x_{CTR} \\ \sigma^2_{TN} &= \zeta_{TN_0} + \zeta_{TN_1}s^2 \end{aligned} \tag{1.7}$$

avec $\{\beta_{TN_0}, \cdots, \beta_{TN_3}, \zeta_{TN_0}, \zeta_{TN_1}\}$ l'ensemble des coefficients de régression du modèle normal tronqué. x_{HRES} est une prévision haute résolution issue d'un modèle de prévision numérique disposant d'une approche physique mieux résolue et une résolution spatiale plus fine que celle du modèle de prévision d'ensemble. x_{CTR} représente la prévision de contrôle (prévision non perturbée) du modèle d'ensemble de prévisions numériques.

Néanmoins, cette loi peine à reproduire des situations de vent élevé (S. Baran 2014). Une autre modélisation à partir d'une loi log-normale $y|X \sim \text{log-}\mathcal{N}(\mu_{LN}, \sigma^2_{LN})$ est proposée par S. Baran et Lerch 2015 pour produire des situations réalistes de vent élevé. Dans ce cas les paramètres du modèle sont donnés par :

$$\begin{aligned} \mu_{LN} &= \log(\frac{\mu^2}{\sqrt{\sigma^2 + \mu^2}}) \\ \sigma^2_{LN} &= \log(1 + \frac{\sigma^2}{\sqrt{\mu^2}}) \end{aligned} \tag{1.8}$$

où μ et σ représentent les paramètres du modèle NGR classique avec les prévisions déterministes x_{HRES} et x_{CTR} ajoutées au modèle :

$$\begin{aligned} \mu &= \beta_{LN_0} + \beta_{LN_1}\overline{x} + \beta_{LN_2}x_{HRES} + \beta_{LN_3}x_{CTR} \\ \sigma^2 &= \zeta_{LN_0} + \zeta_{LN_1}s^2 \end{aligned} \tag{1.9}$$

où $\{\beta_{LN_0}, \cdots, \beta_{LN_3}, \zeta_{LN_0}, \zeta_{LN_1}\}$ représente l'ensemble des coefficients de régression du modèle log-normal.

Cependant, les deux lois présentées produisent parfois des ensembles avec des erreurs de dispersions et de biais. Pour le modèle normal tronqué, les erreurs peuvent sous-estimer les situations de vent fort, là où pour le modèle log-normal, les erreurs peuvent surestimer les situations de vent faible. Les travaux de S. Baran et Lerch 2016 repris par Bremnes 2019 montrent la pertinence de construire un mélange basé sur ces lois pour calibrer les ensembles de vitesses du vent. La densité de probabilité de ce mélange est exprimée de la

façon suivante :

$$f(y;\mu_{TN},\sigma^2_{TN},\mu_{LN},\sigma^2_{LN},w) = wf_0(y;\mu_{TN},\sigma^2_{TN}) + (1-w)g(y;\mu_{LN},\sigma^2_{LN}) \tag{1.10}$$

où f_0 représente la densité de loi normale tronquée et g celle de la loi log-normale. Le paramètre w représente la proportion attribuée à la densité de la loi normale tronquée.

L'inférence des paramètres de ce mélange passe par la minimisation du score moyen du $\overline{CRPS}(\mathcal{X};\mu_{TN},\sigma^2_{TN},\mu_{LN},\sigma^2_{LN},w)$. Le $CRPS$ du mélange (1.10) est exprimé dans S. Baran et Lerch 2016.

1.1.1.4 Précipitations

Les précipitations, ou cumuls de précipitations, sont représentées sous forme d'un couple de variables aléatoires. La première partie discrète illustre les situations sans précipitations, tandis que la partie continue permet de modéliser les hauteurs de précipitations, cumulées sur un intervalle d'heures. Dans ce cadre, Scheuerer et Hamill 2015a et S. Baran et Nemoda 2016 proposent un modèle NGR basé sur la loi Gamma censurée et décalée (CSG, "Censored Shifted Gamma") pour modéliser les précipitations sur une période définie. La loi Gamma modélise facilement les situations récurrentes de précipitations comme les longues périodes sans pluie. La partie censurée du modèle permet d'écarter les valeurs négatives dans les jeux de données, qui sont des artefacts dus aux erreurs de mesure et aux erreurs numériques. Le "shift" ou décalage pose un seuil de précipitation minimale permettant de séparer les artefacts de mesures et prévisions aux valeurs positives trop faibles pour être considérées comme des précipitations. Dans ces travaux, Taillardat, Fougères et al. 2019 montrent que cette méthode apporte des résultats convaincants sur des points spatiaux ne présentant pas ou peu d'événements extrêmes.

En reprenant les notions de la section 1.1.1.1 et les termes de Scheuerer et Hamill 2015a, la fonction de répartition du modèle CSG s'écrit :

$$F^0_{k,\theta,\delta}(y) = \begin{cases} F_{k,\theta}(y-\delta) & y \geq 0 \\ 0 & y < 0 \end{cases} \tag{1.11}$$

pour le décalage $\delta > 0$, $F_{k,\theta}(y-\delta)$ symbolisant la fonction de répartition de la loi Gamma

$\Gamma(k, \theta)$ avec comme paramètres :

$$\begin{aligned} k &= \frac{\mu_{Precip}^2}{\sigma_{Precip}^2} \\ \theta &= \frac{\sigma_{Precip}^2}{\mu_{Precip}} \end{aligned} \tag{1.12}$$

Scheuerer et Hamill 2015a exprime le décalage $\delta = \mu_{Precip} \log(\pi_{POP})$ où $\pi_{POP} = F_{k,\theta}(POP)$ représente la probabilité de précipitation.

Pour ce modèle NGR, les équations de régressions sont définies à l'aide des paramètres μ_{Precip} et σ_{Precip} :

$$\begin{aligned} \mu_{Precip} &= \frac{\mu_{cl}}{\beta_0} \log\left(1 + (e^{\beta_0} - 1)(\beta_1 + \beta_2 POP + \beta_3 \frac{\overline{x}}{\overline{x}_{cl}} + \beta_4 \frac{x_{HRES}}{x_{HRES_{cl}}} + \beta_5 \frac{x_{CTR}}{x_{CTR_{cl}}})\right) \\ \sigma_{Precip} &= \zeta_0 \sigma_{cl} \left(\frac{\mu_{Precip}}{\mu_{cl}}\right)^{\zeta_1} + \zeta_2 MD \end{aligned} \tag{1.13}$$

avec $\beta = (\beta_0, \cdots, \beta_5, \zeta_0, \zeta_1, \zeta_2)$ l'ensemble des coefficients de régressions. POP représente un seuil de précipitation minimal défini à partir des réalisations supérieures à zéro d'un ensemble, $\overline{x}$ la moyenne de l'ensemble, $MD = \frac{1}{m^2} \sum_{j'=1}^{M} \sum_{j=1}^{M} |x_{j'} - xj|$, x_{HRES} et x_{CTR} représentent des prévisions déterministes. μ_{cl} et σ_{cl} correspondent à la moyenne et l'écart type climatique estimés sur les observations, là où $\overline{x}_{cl}$, $x_{HRES_{cl}}$ et $x_{CTR_{cl}}$ sont les moyennes climatiques estimées sur les données associées.

Les équations de régressions de $(\mu_{Precip}, \sigma_{Precip})$ obtenues ont un lien non linéaire avec les ensembles de prévision et prévisions déterministes. Ce lien non linéaire offre davantage de flexibilité dans la forme de la loi Gamma associée aux précipitations mais rend les coefficients plus difficiles à interpréter. Pour un jeu de données d'observations et d'ensembles, l'inférence de ce modèle $\mathcal{X}$ est effectuée par la minimisation du score moyen $\overline{CRPS}(\mathcal{X}, \mu_{Precip}, \sigma_{Precip})$ (1.6). Le $CRPS(., \mu_{Precip}, \sigma_{Precip})$ est exprimé dans Scheuerer et Hamill 2015a à partir des paramètres $(\mu_{Precip}, \sigma_{Precip})$.

1.1.2 Modèle non paramétrique

Les modèles paramétriques NGR sont flexibles et faciles à implémenter. Néanmoins, la complexité des erreurs de distribution des ensembles peut s'exprimer à travers de multiples biais et erreurs de dispersions différents pour un même jeu de données étudié. Le

caractère paramétrique de ces modèles limite les performances des corrections dans cette situation. Avec l'augmentation récente des puissances de calcul et du nombre de jeux de données disponibles, les modèles non paramétriques sont devenus de plus en plus utilisés. Des modèles non paramétriques comme les régressions quantiles (Bremnes 2019), l'algorithme de forêt aléatoire (Taillardat, Mestre et al. 2016) ou encore les réseaux de neurones (Bremnes 2020) sont étudiés dans le domaine de la calibration. Le modèle de régression de forêt aléatoire appliqué à l'estimation de quantiles (QRF) permet une calibration d'ensemble de qualité, mais aussi une étude des covariables contribuant à la calibration. Dans cette thèse, on s'intéressera donc à ce modèle.

1.1.2.1 Forêt aléatoire

L'algorithme de forêt aléatoire proposé par Breiman 2001 génère différentes partitions des observations à partir d'un ensemble de règles tirées des covariables issues des ensembles. Pour cela, Breiman 2001 définit un nouvel ensemble de données aux individus indépendants $\mathcal{X} = \{y_t, x'_t\}_{1 \leq t \leq T}$ avec $x'_t \in \mathcal{B}$ une collection de covariables issue de X^*. Le sous-espace $\mathcal{B}$ contient les valeurs prises par les p covariables issues de l'ensemble de prévision X^*_t. À partir de ce jeu de données, le modèle de forêt aléatoire f construit un ensemble de collections de règles θ approchant la relation $y|X$ à partir des données de $\mathcal{X}$:

$$f(x') = \mathbb{E}[y|X = x'] \tag{1.14}$$

L'algorithme s'appuie sur N arbres ajustés indépendamment sur les données $\mathcal{X}$. Le modèle d'arbre proposé par Breiman et al. 1984 ajuste une collection de noeuds θ_n, $n \in \{1, \cdots, N\}$ sur les covariables $(x'_t)_{1 \leq t \leq T}$ sans émettre d'hypothèse sur la relation $y|X$. Cette collection de noeuds construit $l = \{1, \cdots, L\}$ feuilles représentant des sous-espaces rectangulaires $R_l \in \mathcal{B}$.

Dès lors, pour un nouveau x', la prédiction de chaque arbre $\mathcal{T}(\theta_n)$ est effectuée sous la forme de poids $w_t(x', \theta_n)$ décomptant les individus $(x'_{t'})$ $\forall t' \neq t \in \{1, \cdots, T\}$ associés à la feuille $R_{l(x',\theta_n)}$:

$$w_t(x', \theta_n) = \frac{1_{\{x'_t \in R_{l(x',\theta_n)}\}}}{\sum_{t' \neq t}^{T} 1_{\{x'_{t'} \in R_{l(x',\theta_n)}\}}} \tag{1.15}$$

La somme sur t des poids w_t est égale à 1. Pour terminer, la prédiction d'une observation

y sachant le jeu de covariables x' s'effectue par agrégation des N prédictions issues des arbres à l'aide de la moyenne empirique :

$$\begin{aligned} \hat{\mathbb{E}}[y|X = x'] &= \sum_{t=1}^{T} w_t(x') y_t \\ w_t(x') &= \sum_{n=1}^{N} \frac{w_t(x', \theta_n)}{N} \end{aligned} \tag{1.16}$$

L'algorithme de forêt aléatoire présenté précédemment n'est pas adapté à la calibration d'ensemble. En effet, cet algorithme ne prédit pas d'ensemble corrigé, mais plutôt une prédiction moyennée. Pour obtenir un ensemble post-traité par ce type de modèle, une modification est nécessaire.

1.1.2.2 Modèle QRF

Introduit par Meinshausen 2006 comme une extension des forêts aléatoires de Breiman 2001, l'algorithme de forêt aléatoire de régression appliqué à l'estimation de quantiles propose de remplacer le problème d'espérance (1.14) par un problème d'estimation de la fonction de répartition :

$$f(x', q) = F(y \leq q|X = x') = \mathbb{E}[1_{\{y \leq q\}}|X = x'] \tag{1.17}$$

avec q un quantile spécifique défini selon $y|X$.

En reprenant (1.16), la fonction de répartition estimée est obtenue :

$$\hat{F}(y \leq q|X = x') = \sum_{t=1}^{T} w_t(x) 1_{\{y \leq q\}} \tag{1.18}$$

La fonction de répartition estimée précédemment permet d'obtenir une expression pour la fonction quantile ou fonction génératrice. Dès lors, un nouvel ensemble de quantiles de la loi conditionnelle $y|X$ peut être généré.

1.1.2.3 Inférence

Traditionnellement, dans le cadre d'observations continues, on minimise l'erreur quadratique moyenne (MSE) estimée entre la valeur prédite $\hat{y}$ par le modèle et l'observation

y :

$$\begin{aligned} MSE &= \frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2 \\ \hat{y}_t &= \hat{\mathbb{E}}[y_t|X = x'_t] \end{aligned} \tag{1.19}$$

La MSE permet de mesurer de manière globale la qualité d'ajustement d'un modèle d'apprentissage sur un jeu de données. Néanmoins, pour l'ajustement de l'ensemble des θ du modèle de forêt, cette mesure ne suffit pas. Un autre score évaluant la qualité d'ajustement de l'ensemble $\mathcal{C}_i$ des groupes de données générés par une séparation du noeud θ_n^i est nécessaire, et ce pour tout $i \in \{1, \cdots, I_{Max}\}$. Dans le cadre d'une variable objectif continue, la variance empirique estimée pour un groupe de données $c \in \mathcal{C}_i$ est utilisée pour mesurer la qualité d'une séparation :

$$V(c) = \frac{1}{T_c - 1}\sum_{y_t \in c}(y_t - \overline{y}_t)^2 \tag{1.20}$$

avec $\overline{y}_t = \frac{1}{T_c}\sum_{y_t \in c} y_t$ la moyenne empirique des observations associée au groupe c et T_c le nombre d'observations se trouvant dans le groupe c.

A l'aide de la variance empirique V estimée pour chaque groupe $c \in \mathcal{C}_i$ d'observations généré par un noeud θ_n^i, Breiman et al. 1984 propose de déterminer la meilleure séparation maximisant le critère d'homogénéité H :

$$H(\theta_n^i) = V(\mathcal{C}_i) - \sum_{c \in \mathcal{C}_i} V(c) \tag{1.21}$$

Ensuite, l'opération de maximisation de l'homogénéité pour obtenir une nouvelle séparation θ_n^{i+1} est itérée jusqu'à ce que l'algorithme converge suivant un critère d'arrêt. Construire N ajustements d'arbres sur les données $\mathcal{X}$ engendre un risque que plusieurs arbres obtiennent la même collection de noeuds θ_n. Ces collections quasi identiques entraînent des prédictions très proches sous-estimant la réelle incertitude autour des prédictions. En contrepartie Breiman 1996 introduit le "bagging", une technique d'ajustement consistant à sélectionner des partitions d'individus et covariables des données $\mathcal{X}$ de manière aléatoire pour entraîner les arbres. Le "bagging" permet de construire un score d'importance mesurant l'apport individuel des p covariables de x' sur l'ajustement du modèle.

Taillardat, Mestre et al. 2016 ont montré les performances de calibration de

modèle QRF dans un contexte univarié pour la variable de température, pour les vitesses du vent et les précipitations. En particulier, ils ont pu observer que les ensembles post-traités par ce modèle étaient capables de représenter des phénomènes locaux, souvent mal résolus par les modèles de prévision numériques.

1.2 Calibration d'ensembles multivariés

Dans la suite, les observations et ensembles de prévision décrivent des conditions météorologiques multivariées prenant leurs valeurs dans un espace $\mathbb{R}^d$ où $d > 1$. Le nouvel espace introduit les situations physiques multivariées formées par un champ spatial, une trajectoire temporelle ou encore des variables météorologiques interagissant dans les prévisions et observations. La loi conditionnelle $y|X$ ne peut être approchée par les modèles introduits précédemment. Cette section a pour but de présenter une des méthodes classiques basée sur l'approche des copules empiriques. Cette méthode est régulièrement utilisée pour la calibration d'ensembles multivariés. Plus particulièrement, le modèle de Schaake shuffle sélectionné pour sa facilité d'adaptation et ses performances de calibration est présenté. Les notations de Schefzik et Möller 2018 sont reprises pour cette partie.

1.2.1 Copules empiriques

La modélisation paramétrique de distributions multivariées représente un challenge dont la complexité augmente rapidement avec la dimension des variables étudiées. Une approche naturelle à ce problème est de considérer une famille de copules décrivant les dépendances multivariées. Définie par Sklar 1959, une copule $C : [0,1]^d \to [0,1]$ est une fonction de répartition dont les d lois marginales sont uniformes. Sklar 1959 exprime la relation entre la fonction de répartition F d'un vecteur aléatoire et une unique copule C formée à partir de F et des $F_1, \cdots, F_d$ fonctions de répartition marginales.

Suivant le type de loi du processus étudié, on considère différentes familles de copules. Plusieurs cas d'applications de corrections d'ensembles bivariés utilisent des copules gaussiennes (Pinson, Madsen et al. 2009, Gill, Stephen et Galloway 2011, Möller, Lenkoski et Thorarinsdottir 2013).

Les copules paramétriques se trouvent être de très bon outil de modélisation de structure physique bivariée mais leur complexité augmente très rapidement avec la dimensionnalité des variables. Dès lors que $d > 2$, il devient difficile d'émettre une hypothèse sur

la structure multivariée. Dans cette situation et en s'aidant du théorème de Sklar 1959, Rüschendorf 2009 propose de déduire d'un échantillon de réalisations une estimation empirique de copule en passant par les fonctions de répartitions empiriques marginales. En posant un ensemble de réalisations $(x_1, \cdots, x_M)$ dans $\mathbb{R}^d$, la copule empirique E_M se définit de la manière suivante :

$$E_M(\frac{r_1}{M}, \cdots, \frac{r_d}{M}) = \frac{1}{M} \sum_{m=1}^{M} \prod_{i=1}^{d} 1_{\{rank(x_m^i) \leq r_i\}} \tag{1.22}$$

avec $\forall i \in \{1, \cdots, d\}$ r_i un entier allant de 0 à M. Les copules empiriques peuvent s'interpréter comme la distribution des rangs des réalisations $(x_1, \cdots, x_M)$. La définition (1.22) introduit un cadre pour employer des méthodes de calibration multivariées simples et flexibles (méthode de Schaake shuffle : Clark et al. 2004 ; Schefzik, Thorarinsdottir, Gneiting et al. 2013).

1.2.2 Méthode de Schaake shuffle

La méthode Schaake shuffle (SS) proposée par Clark et al. 2004 est une méthode basée sur les copules empiriques. L'algorithme vise à apprendre la structure multivariée d'un ensemble d'observations puis à la transmettre sur un ensemble dont les marginales ont été post-traitées de manière univariée sur chaque dimension $i \in \{1, \cdots, d\}$. La méthode de Schaake shuffle est constituée de quatre étapes clés reprises de Schefzik 2016 :

1. Définition de la structure de dépendance Pour un jeu de données d'observations et d'ensembles de M membres échangeables $\mathcal{X} = \{y_t, X_t^* = (x_{tm})_{1 \leq m \leq M}\}_{1 \leq t \leq T}$ avec $y_t \in \mathbb{R}^d$ et $x_{tm} \in \mathbb{R}^d$, il est défini $Y = (y_{m'})_{1 \leq m' \leq M'}$ un ensemble de M' observations où chaque $y_{m'}$ est sélectionné aléatoirement dans $\mathcal{X}$. À cela, une copule empirique E_M (1.22) est dérivée pour approcher la structure multivariée de l'ensemble d'observations Y. Cette opération revient à former $H = (\pi_i(m'))_{(1,1) \leq (m',i) \leq (M',d)}$ représentant l'ensemble des permutations $\pi_i(m') = rank(y_{m'}^i)$ induit par la statistique d'ordre $y_{(1)}^i \leq \cdots \leq y_{(M')}^i$.

2. Post-traitement univarié Dès lors que la structure de dépendance H est déterminée, une collection de modèles de calibration univariée $(\mathcal{M}_1, \cdots, \mathcal{M}_i, \cdots, \mathcal{M}_d)$ est construite pour chacune des dimensions $i \in \{1, \cdots, d\}$ à partir des données $\mathcal{X}^i = \{y_t^i, X_t^{*^i}\}_{1 \leq t \leq T}$.

3. Génération de l'ensemble Pour un nouvel ensemble $X^* = (x_m)_{1 \leq m \leq M}$ avec $x_m \in \mathbb{R}^d$, on construit un ensemble post-traité $\tilde{X}$ contenant une collection de quantiles

équidistants dont chaque quantile $\tilde{x}^i_{m'}$ est relié à la probabilité $\frac{m'}{M'+1}$ et ce $\forall(m', i) \in \{1, \cdots, M'\} \times \{1, \cdots, d\}$:

$$\tilde{x}^i_{m'} = F^{-1}_{\mathcal{M}_i}(\frac{m'}{M'+1}) \tag{1.23}$$

où $F^{-1}_{\mathcal{M}_i}(.)$ représente la fonction génératrice, soit l'inverse de la fonction de répartition obtenue pour le modèle $\mathcal{M}_i$.

4. Application de la structure de dépendance Dans cette dernière étape, l'ensemble multivarié $\hat{X}^*$ est prédit en appliquant les permutations stockées dans H sur les réalisations de $\tilde{X}$, $\forall(m', i) \in \{1, \cdots, M'\} \times \{1, \cdots, d\}$:

$$\hat{x}^i_{m'} = \tilde{x}^i_{\pi_i(m')} \tag{1.24}$$

L'avantage de la méthode de Schaake shuffle est de pouvoir tirer parti des nombreuses performances de calibration des modèles univariés. De plus, les ensembles post-traités par ces modèles se voient appliquer une structure multivariée apprise des observations incluant les représentations réelles des conditions météorologiques locales étudiées. La méthode de Schaake shuffle se base sur un ensemble d'observations Y et de M' membres et l'algorithme est contraint de produire des ensembles $\hat{X}$ de taille M'. Cependant, M' n'a pas l'obligation d'être égal à M. Cette méthode a pour avantage de tirer profit de la robustesse d'estimation des copules empiriques pour produire une estimation fidèle de la structure multivariée de $y|X$ sans hypothèse de loi particulière et ainsi fournir un ensemble multivarié X^*. Le modèle de Clark et al. 2004 bénéficie également d'une facilité d'adaptation numérique la rendant encore fréquemment utilisée par la communauté météorologique statistique (Scheuerer, Hamill et al. 2017, Wu et al. 2018, Schepen, Everingham et Wang 2020).

Néanmoins, un inconvénient de cette approche est la sélection aléatoire des observations formant l'ensemble Y lors de l'étape 1. L'ensemble d'observations ainsi formé n'est pas forcément représentatif de l'état de l'atmosphère lié à l'ensemble prédit par le modèle de prévision numérique X^*. Dans ses travaux, Schefzik 2016 propose le "similarity Schaake shuffle" (SimSS), une méthode basée sur une approche où l'on sélectionne les M' observations minimisant une similarité Δ calculée entre l'ensemble présent X^* et les ensembles passés des données $\mathcal{X}$. Schefzik 2016 utilise comme similarité la distance euclidienne entre moyennes empiriques et entre variances empiriques des marginales des

ensembles étudiés :

$$\Delta_t(X^*, X_t^*) = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(\overline{x}^i - \overline{x}_t^i)^2 + \frac{1}{d}\sum_{i=1}^{d}(s^{2^i} - s_t^{2^i})^2} \tag{1.25}$$

avec $\overline{x}^i$ et s^{2^i} respectivement la moyenne et variance empirique de la $i^{ème}$ marginale.

Avec cette amélioration, le modèle de Schaake shuffle approche mieux les structures multivariées des ensembles issus du modèle numérique.

Pour la section suivante, les modèles de calibration univariée NGR et QRF présentés précédemment sont couplés à l'extension du Schaake shuffle (SimSS) pour réaliser une correction d'ensembles multivariés. Ils sont alors nommés $NGR/SimSS$ et $QRF/SimSS$.

1.3 Application des modèles de calibration

Cette partie est consacrée à l'analyse de nos premiers résultats de calibration par modèles univariés (NGR et QRF) et multivariés ($NGR/SimSS$ et $QRF/SimSS$). Pour cela, les données d'ensemble du centre européen et des observations in situ des stations SYNOP sont introduites pour trois localisations spatiales différentes. Ensuite, les covariables tirées des ensembles de prévision et utilisées par les modèles de forêt aléatoire sont introduites.

1.3.1 Données

Les données d'ensemble utilisées proviennent de l'archive TIGGE ('Thropex Interactive Grand Global Ensemble'). L'archive est notamment disponible à l'adresse suivante : `https://apps.ecmwf.int/datasets/data/tigge`. Cette initiative regroupe initialement dix modèles d'ensemble de centres de prévisions météorologiques numériques différents regroupés suivant quatre centres d'archives : NCAR Etats-Unis, CEPMMT Angleterre, CMA Chine et plus récemment NCMRWF Inde (Park, Buizza et Leutbecher 2008 ; Bougeault et al. 2010 ; Swinbank et al. 2016). Plus particulièrement, les données du modèle d'ensemble du centre européen de prévision météorologique à moyen terme CEPMMT sont récupérées. Ces ensembles contiennent 50 membres échangeables générés à partir d'un modèle d'assimilation de données d'ensemble (EDA, 'Ensemble Data Assimilation') basé sur les perturbations de vecteurs singuliers des conditions initiales et d'un modèle physique et stochastique (Buizza, Leutbecher et Isaksen 2008 ; Buizza

2016). Ces données sont extraites pour l'ensemble des jours des années 2008 à 2018 avec deux lancements du modèle numérique par jour (6h et 18h) et des horizons de prévision de 3, 5 et 10 jours (3j, 5j et 10j). Les données sont récupérées pour trois stations spatiales sélectionnées pour leurs caractéristiques climatiques différentes. La première est située à proximité de la ville de Millau dans la chaîne de montagnes du massif Central. Cette station dispose d'un vent de forte intensité dû à l'élévation de la station, située à 340 m d'altitude, et est impactée par le climat méditerranéen. La seconde station est prise sur l'aéroport de la ville de Rennes avec une géolocalisation proche de l'océan Atlantique et de la Manche disposant donc d'un climat tempéré. La dernière prise à l'aéroport de Strasbourg possède une situation géographique éloignée dans les terres, proche de la frontière allemande, avec un climat continental.

Les variables météorologiques sélectionnées pour la définition des classes dans cette étude sont les cumuls de précipitations sur 12h (Precip, mm) et les vitesses du vent à 10m (VV, m.s^{-1}). Ces variables sont communément étudiées par la communauté de calibration d'ensemble de prévision pour leur intérêt économique dans des domaines tels que l'agriculture, les chantiers de voirie, l'actuariat, etc.

Les observations SYNOP sont fournies par les stations météorologiques maintenues par Météo France [1]. Les observations sont aussi connues pour contenir des biais systémiques provoqués par d'éventuelles erreurs de mesures des capteurs utilisés par les stations. Les réanalyses fournies par le modèle d'assimilation de données contiennent l'information des prévisions et observations passées et corrigées, de ce fait, elles représentent de bons candidates pour évaluer l'erreur contenues dans les observations SYNOP. Une méthode classique est d'appliquer une simple régression linéaire entre les réanalyses du modèle numérique prises en tant que référence et les observations (Glahn et Lowry 1972).

Pour la suite du chapitre, les observations et les ensembles de prévision sont définis de la façon suivante pour les deux variables météorologiques : $y = (y^{Precip}, y^{VV}) \in \mathbb{R}_+^d$ et $X^* = (X^{*^{Precip}}, X^{*^{VV}}) \in \mathbb{R}_+^{d \times M}$.

1.3.2 Ensemble de covariables

Les modèles de calibration basés sur les forêts aléatoires reposent sur un ensemble de covariables construit à partir des ensembles et d'informations annexes (horaire par exemple). Les membres des ensembles étant échangeables, ils ne peuvent pas être direc-

1. données disponibles à l'adresse suivante `https://donneespubliques.meteofrance.fr/?fond=produit&id_produit=90&id_rubrique=32`

tement utilisés comme covariables. Les travaux de Taillardat, Mestre et al. 2016 ont introduit un ensemble de covariables tirées de statistiques empiriques d'ensembles échangeables, obtenant de bons résultats avec le modèle de forêt QRF. Ces covariables ont été notamment reprises dans Straaten, Whan et Schmeits 2018 et Taillardat, Fougères et al. 2019. Des covariables similaires présentées dans le tableau 1.1 sont utilisées dans cette thèse pour entraîner le modèle de forêt.

Acronyme	Expression
HRES, CTR	x_{HRES}, x_{CTR} Membres déterministes
Mean	$\overline{x} = \frac{1}{M}\sum_{m=1}^{M} x_m$
Sigma	$\sigma = \sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(x_m - \overline{x})^2}$
skew	$\kappa^3 = \frac{1}{\sigma^3(M-1)}\sum_{m=1}^{M}(x_m - \overline{x})^3$
kurt	$\kappa^4 = -3 + \frac{1}{\sigma^4(M-1)}\sum_{m=1}^{M}(x_m - \overline{x})^4$
IQR	$IQR = \hat{F}_X^{-1}(\frac{2}{3}) - \hat{F}_X^{1}(\frac{1}{3})$
Q10,50,90	$q_p = \hat{F}_X^{-1}(p)$ avec $p = 0.1, 0.5, 0.9$
P0,03,1,3,5	$p_q = \hat{F}_X(q)$ avec $q = 0, 0.3, 1, 3, 5$
month, hour	Facteurs discriminants

Table 1.1 – Acronymes et expressions des covariables utilisées pour les modèles de calibration univariée QRF.

Les prévisions déterministes "HRES et CTR" (inclues dans les modèles NGR de vitesses du vent et de précipitations) sont également présentes comme covariables du modèle de forêt. la prévision HRES est fournie par un modèle de prévision numérique à la résolution supérieure à celle du modèle d'ensemble. Néanmoins, la haute résolution du modèle de prévision HRES ne dépasse pas l'échéance de 10 jours, ce qui est faible en comparaison à l'échéance du modèle d'ensemble, capable d'atteindre 15 jours. La prévision CTR est issue du modèle non perturbé d'ensemble de prévision. Les covariables "Mean, Sigma, skew, kurt, Q10, Q50, Q90" sont estimées à partir des membres des ensembles échangeables de vitesses du vent et de précipitations. Le groupe de covariables "IQR, P0, P03, P1, P3, P5" est estimé uniquement à partir des ensembles de précipitations. Les facteurs "month, hour" représentent des covariables qualitatives donnant le mois et l'heure des ensembles de prévision étudiés.

1.3.3 Scores de calibration

L'évaluation des résultats des méthodes mises en oeuvre dans ce chapitre passe par différents scores. Des scores tels les diagrammes de Talagrand (ou histogrammes de rangs) ou le CRPS estimé permettent d'évaluer la calibration des ensembles de prévision. Dans un cadre de données multivariées, on utilise l'extension multivariée du CRPS, connue sous le nom du score d'énergie (ES, "Energy score").

1.3.3.1 Histogramme de rangs

L'histogramme de rangs est une technique visuelle de contrôle de la qualité d'un ensemble de prévision ou ensemble de quantiles estimés. Pour un jeu de données d'observations et d'ensembles de M réalisations univariées $\mathcal{X} = \{y_t, X_t^*\}_{1 \leq t \leq T}$, avec $y_t \in \mathbb{R}$ et $X_t^* \in \mathbb{R}^M$, l'idée est de venir ranger chaque observation y_t dans un rang r approprié. En suivant Anderson 1996, Talagrand, R. Vautard et Strauss 1997, Hamill et Colucci 1997, un ensemble de rangs statistiques $\{r_1, \cdots, r_{M+1}\}$ est construit à partir de la statistique d'ordre $x_{t(1)} \leq \cdots \leq x_{t(M)}$ issue de l'ensemble X_t^* de la façon suivante $\forall (j,t) \in \{1, ..., M+1\} \times \{1, ..., T\}$:

$$r_j = \frac{1}{T} \sum_{t=1}^{T} \hat{p}(x_{t(j-1)} \leq y_t < x_{t(j)}), \tag{1.26}$$

avec $x_{t(0)} = -\infty$, $x_{t(M+1)} = +\infty$ et $\hat{p}(.)$ la probabilité estimée que l'observation soit rangée entre deux statistiques d'ordre de l'ensemble.

Un ensemble bien calibré devrait afficher un histogramme de rangs plat, semblable à celui d'une loi uniforme. Cependant, la réciproque n'est pas garantie Hamill 2001. Un histogramme de rang plat signifie que l'observation peut être proche de n'importe quel élément de l'ensemble de façon équiprobable. Il y a autant de chances que l'observation soit proche des éléments les plus petits de l'ensemble, que des éléments les plus grands de l'ensemble ou que des éléments au centre de l'ensemble. Ainsi, chaque élément de l'ensemble est utile et a le même niveau de représentativité que dans la réalité.

Un histogramme de rangs formant un U implique que les observations sont souvent à l'extérieur de l'ensemble ou proche de l'extérieur de l'ensemble. Les rangs des extrémités (les premiers et les derniers rangs) sont donc surreprésentés ce qui créera la forme en U indiquant un problème de sous-dispersion ou de biais conditionnel de l'ensemble selon Hamill 2001. Inversement, un histogramme de rang affichant un dôme ou $\cap$ indique

quant à lui un ensemble potentiellement sur-dispersé. Pour terminer, un histogramme de rang non symétrique formant un "L" tend vers la présence d'un biais dans la distribution de l'ensemble venant sous ou sur estimer les observations suivant l'orientation de ce "L". Les histogrammes de rangs font partie des techniques de référence dans l'évaluation des ensembles univariés qui ne cessent d'être étendues pour des objectifs d'étude d'ensembles multivariés (Gneiting, Stanberry et al. 2008) ou de multiples erreurs de distributions (Bröcker et Ben Bouallègue 2020).

1.3.3.2 Score de probabilité des rangs continus

Introduit de manière théorique dans la section 1.1.1.2, le CRPS est également un score propre utilisé pour évaluer les performances d'un modèle de calibration d'ensemble. Un score propre signifie qu'il est négativement orienté tel qu'une faible valeur indique une meilleure performance. En reprenant l'expression du CRPS (1.6), pour un ensemble de M réalisations $X^* = (x_m)_{1 \leq m \leq M}$ et une observation y, Grimit et al. 2006 définit le CRPS estimé par :

$$\widehat{CRPS}(X^*, y) = \frac{1}{M} \sum_{m=1}^{M} |x_m - y| - \frac{1}{2M^2} \sum_{m=1}^{M} \sum_{m'=1}^{M} |x_m - x_{m'}| \tag{1.27}$$

L'expression proposée permet d'estimer le CRPS sans aucun a priori de loi. Cependant, cette expression est coûteuse en termes d'estimation numérique dans un cas d'une taille d'échantillon ou nombre de membres M élevé. Une expression algébriquement équivalente proposée par Laio et Tamea 2007 permet de réduire les coûts d'estimation :

$$\widehat{CRPS}(X^*, y) = \frac{2}{M^2} \sum_{m=1}^{M} \Big((x_{(m)} - y)(M 1_{\{x_{(m)} \leq y\}} - m + \frac{1}{2}) \Big) \tag{1.28}$$

où $\{x_{(m-1)} \leq x_{(m)} \leq x_{(m+1)}\}_{2 \leq m \leq M}$ est l'ensemble réordonné par une statistique d'ordre croissante.

1.3.3.3 Score d'énergie

Lors de l'application de la méthode Schaake shuffle, une calibration multivariée est réalisée. Les scores présentés précédemment n'évaluent pas la structure multivariée des ensembles. Dans cette situation où l'observation y est à valeurs dans $y \in \mathbb{R}^d$, $d > 1$ et que la fonction de répartition F de l'ensemble X^* est d-variée, un score plus approprié

est le score d'énergie (ES, "Energy score", Gneiting et Raftery 2007 ; Scheuerer et Hamill 2015b) défini par :

$$ES(F, y) = \mathbb{E}_F\|x - y\| - \tfrac{1}{2}\mathbb{E}_F\|x - x'\| \tag{1.29}$$

avec $\|.\|$ la norme euclidienne, x et x' deux réalisations indépendantes de l'ensemble X^*.

En utilisant un ensemble de M réalisations $X^* = (x_m)_{1\leq m\leq M}$ avec $x_m \in \mathbb{R}^d$, l'estimateur du score d'énergie s'exprime de la façon suivante :

$$\widehat{ES}(X^*, y) = \frac{1}{M}\sum_{m=1}^{M}\|x_m - y\| - \frac{1}{2M^2}\sum_{i=1}^{M}\sum_{m'=1}^{M}\|x_m - x_{m'}\| \tag{1.30}$$

Comme pour le score CRPS, il est attendu que le score d'énergie converge vers 0 dès lors que l'observation et l'ensemble expriment une structure multivariée similaire.

1.3.4 Résultats

Dans cette section, les performances des modèles de calibration univariée (NGR et QRF) et multivariée ($NGR/SimSS$, $QRF/SimSS$) sont présentées et comparées aux ensembles sans post-traitement (RAW). L'étude de ces modèles se fera au travers des scores présentés dans la section 1.3.3. Les données introduites dans la section 1.3.1 sont séparées en une base d'apprentissage et une base de test. La même procédure de création de la base de test que dans Taillardat, Mestre et al. 2016 est appliquée : la base de test est composée de dates tirées aléatoirement dans l'intervalle d'années 2014 à 2018. Ainsi la base d'apprentissage est construite à partir des ensembles et des observations couvrant les années 2008 à 2013 sans discontinuités. Cela permet au modèle de forêt d'apprendre aussi la climatologie implicitement incluse dans les ensembles de prévision. Le modèle NGR est ici généralement entraîné sur une fenêtre glissante de longueur égale à un nombre de jours fixés (Gneiting, Raftery et al. 2005). Dans le cadre des vitesses de vent et des précipitations, cette taille est relativement importante dépassant le nombre de jours dans une année. De ce fait, pour une comparaison plus juste avec le modèle QRF, le modèle NGR est entraîné sur les mêmes nombres de données. Néanmoins, les données étant séparées en deux groupes suivant les horaires d'initialisations du modèle de prévision numérique, le modèle NGR est entraîné séparément sur chacun des groupes afin de prendre en compte plus facilement les erreurs liées aux différents horaires. Ensuite les scores sont estimés sur les résultats des modèles obtenus sur la base de test. L'opération

d'apprentissage et test est répétée $B = 30$ fois pour estimer l'incertitude de chaque score.

Ces modèles ont été implémentés en R. Les modèles de forêt sont initialisés à l'aide de paramètres sélectionnés au court d'une étape de validation croisée obtenant un nombre d'arbres $N = 300$ et un nombre de noeuds égal à 20. Concernant les modèles NGR, leurs coefficients de régression sont ajustés par minimisation de la log-vraisemblance à l'aide de l'algorithme d'optimisation de J. A. Nelder et Mead 1965 disponible sous R et régulièrement utilisé dans les approches NGR (Bremnes 2019, Courbariaux 2017).

1.3.4.1 Étude des méthodes de calibration univariée

Les histogrammes de rangs des variables météorologiques étudiées pour cette partie sont estimés à partir d'ensembles comportant $M = 50$ membres. Pour ce faire, de nouveaux ensembles sont générés, à l'aide des modèles NGR et QRF, en tirant des quantiles de la loi estimée liant les observations et ensembles. Plus précisément, l'ensemble de M probabilités $\{\frac{1}{M+1}, \cdots, \frac{M}{M+1}\}$ est considéré, générant des histogrammes de rangs à 50 rangs. Ceux-ci sont difficiles à visualiser en entier. Pour faciliter la représentation graphique, les rangs sont réduits à 17 au lieu de 50. L'échéance de 3 jours représentant l'information statistique la moins dégradée parmi les échéances disponibles, est sélectionnée pour afficher les histogrammes de rangs et obtenir une description des erreurs de calibration univariée des variables météorologiques aux différentes stations et horaires d'initialisation. Ensuite, le score du CRPS estimé évalue les ensembles des différents modèles. Ce score permet plus facilement de représenter et comparer les résultats de calibration suivant différents facteurs (localisations spatiales, échéances de prévisions, horaires, etc.).

Histogrammes de rangs. La figure 1.1 montre les résultats d'histogrammes de rangs pour la variable de vitesses du vent (VV) par station, et ce pour une échéance fixée à 3 jours à 6H. Les types d'erreurs de distributions des ensembles du modèle numérique (RAW) montrent globalement une forme en "L" inversée avec des rangs forts plus importants que les autres, indiquant un biais avec une tendance de l'ensemble à sous-estimer l'observation. Néanmoins, aux stations de Rennes et Millau, les ensembles RAW montrent aussi des rangs faibles légèrement plus importants que ceux du milieu. Cette remarque combinée au biais identifié précédemment indique un problème de sous-dispersion. Sur les ensembles post-traités par les modèles NGR et QRF, les histogrammes de rangs montrent des rangs rejoignant la ligne pointillée. Ainsi, les ensembles post-traités forment un histogramme de rangs presque uniforme et donc apportent une importante correction aux ensembles RAW.

Néanmoins, un biais du même type que celui des ensembles de *RAW* subsiste dans les ensembles post-traités des deux modèles.

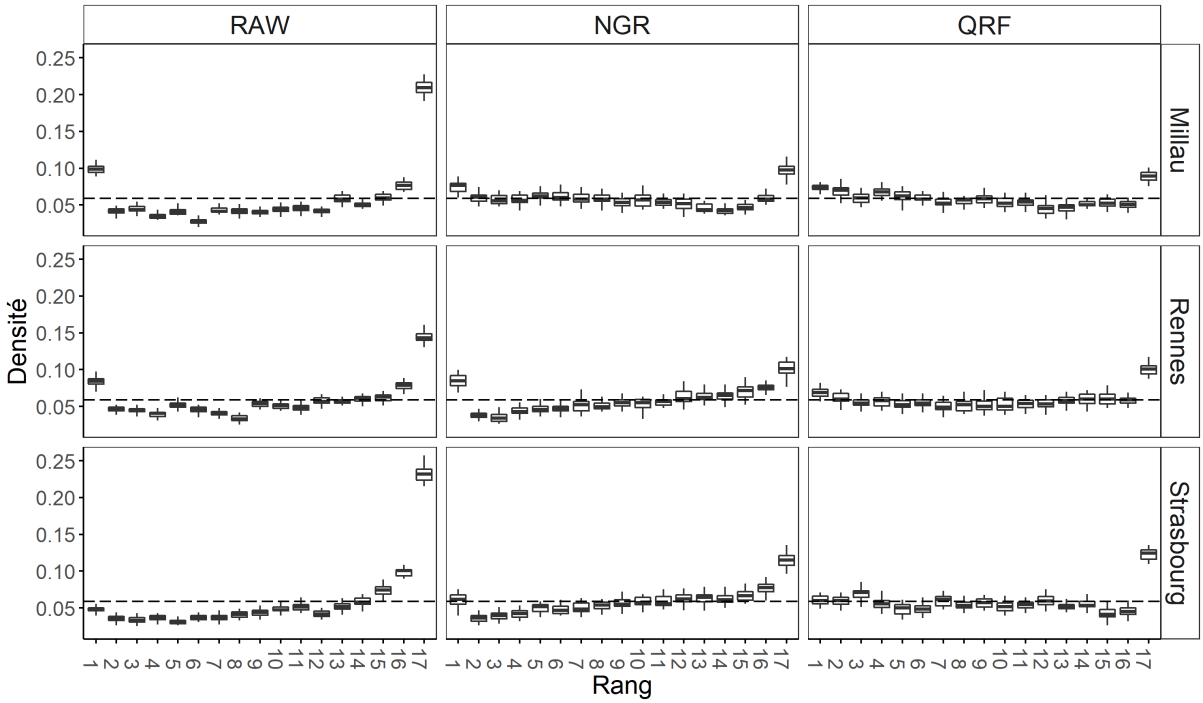


FIGURE 1.1 – Histogrammes de rangs des ensembles et observations des vitesses du vent (VV) à une échéance de prévision de 3 jours à 6H par variable météorologique, modèle et station. *Les modèles sont représentés par les colonnes et les stations par les lignes. La ligne en pointillés indique le seuil à atteindre pour former un histogramme uniforme. RAW : ensembles du modèle de prévisions numériques ; NGR : ensembles du modèle de régression non homogène ; QRF : ensembles du modèle de forêt aléatoire.*

Dans l'annexe A.1.1, la figure A.0a affiche les histogrammes de rangs pour les vitesses du vent (VV) toujours avec une échéance de 3 jours, mais avec un horaire d'initialisation de modèle numérique fixé à 18H. Ici aussi, les ensembles *RAW* montrent globalement le même type d'erreur. Cependant, les ensembles post-traités affichent des différences. Les ensembles issus de *NGR* obtiennent un biais inversé pour les stations de Millau et Rennes par rapport aux ensembles précédents. Les ensembles issus du modèle *QRF* ont un léger problème de sous-dispersion pour Rennes et de sur-dispersion pour Strasbourg. Ces différences d'erreurs d'ensembles post-traités indiquent un traitement différent suivant l'horaire d'initialisation et donc potentiellement un lien avec l'erreur des ensembles *RAW*.

Le modèle *QRF* semble fournir les meilleures corrections pour les ensembles de vitesses du vent et ce pour les différentes stations et échéances de prévision. Les conclusions

des résultats observés par échéance à 6H sont équivalentes à 18H. La dégradation des ensembles du modèle *NGR* au cours des échéances formant des erreurs de sous-dispersion est directement liée au mélange défini par le modèle de S. Baran et Lerch 2016. En effet, ce modèle emprunte des caractéristiques de la loi normale tronquée et de la loi log-normale pour former le mélange. Il y a deux inconvénients aux lois composant ce mélange. D'une part, le modèle normal tronquée tend à sous-estimer les observations. D'autre part, le modèle log-normal tend à les surestimer (S. Baran et Lerch 2015). À courte échéance, les erreurs sont rapidement compensées par le mélange et le côté informatif des statistiques empiriques des ensembles *RAW*. Mais dans le cas de moyennes échéances où l'information se dégrade, ces erreurs deviennent plus fortes.

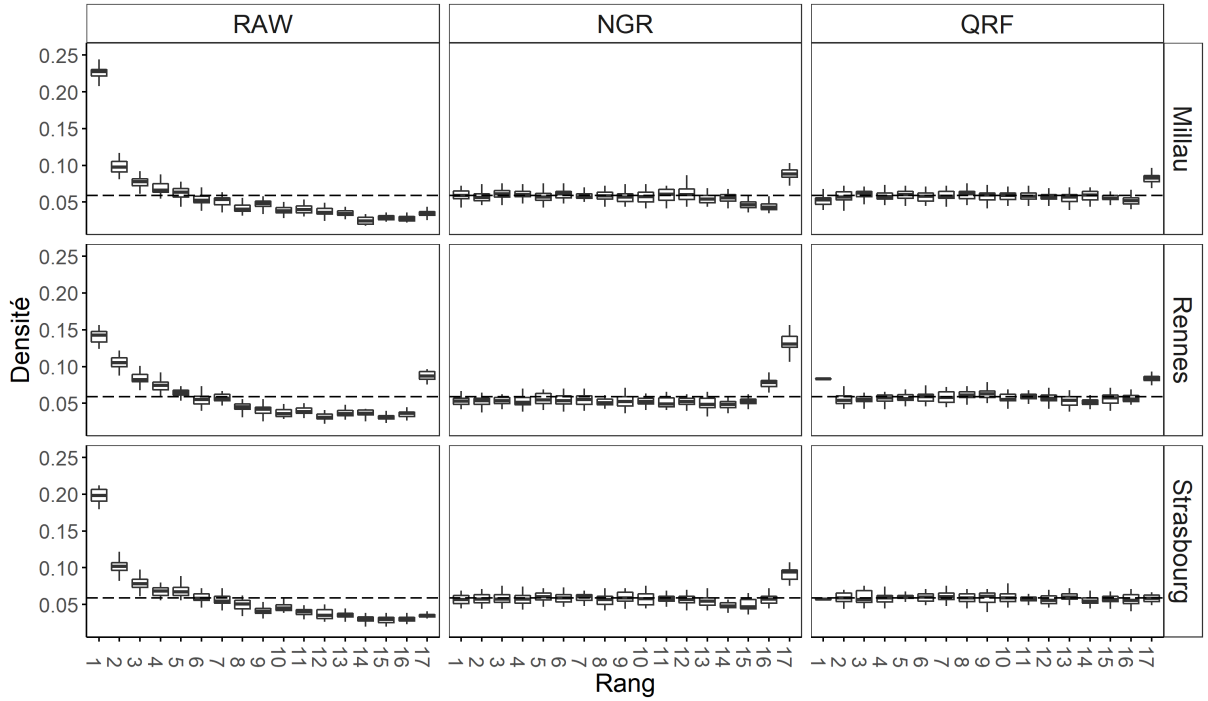


Figure 1.2 – Histogrammes de rangs des ensembles et observations de précipitations (Precip) à une échéance de prévision de 3 jours à 6H par variable météorologique, modèle et station. *Les modèles sont représentés par les colonnes et les stations par les lignes. La ligne en pointillés indique le seuil à atteindre pour former un histogramme uniforme. RAW : ensembles du modèle de prévisions numériques ; NGR : ensembles du modèle de régression non homogène ; QRF : ensembles du modèle de forêt aléatoire.*

Les figures 1.2 et A.0c montrent les histogrammes de rangs des ensembles à échéances de 3 jours aux horaires d'initialisation de modèle fixées respectivement à 6H et 18H pour la variable précipitation (Precip). Ici, les histogrammes des ensembles *RAW* montrent

une forme en "L" générale représentatif de la présence d'un biais dans les ensembles surestimant les observations de précipitation. De plus, sur la station de Rennes à 6H, les rangs forts sont aussi plus élevés que ceux du milieu laissant présager la présence d'une sous-dispersion. Les ensembles post-traités par le modèle NGR et QRF montrent une importante amélioration avec des rangs quasi uniformes à 18H et pour 6H une légère erreur subsiste indiquant un biais sous-estimant l'observation. Cette erreur est plus forte à l'intérieur des ensembles post-traités par le modèle NGR. Globalement à une échéance de 3 jours, le modèle numérique affiche des erreurs des distributions d'ensembles de précipitations surestimant les observations que les modèles de calibration corrigent voire sous-estiment légèrement.

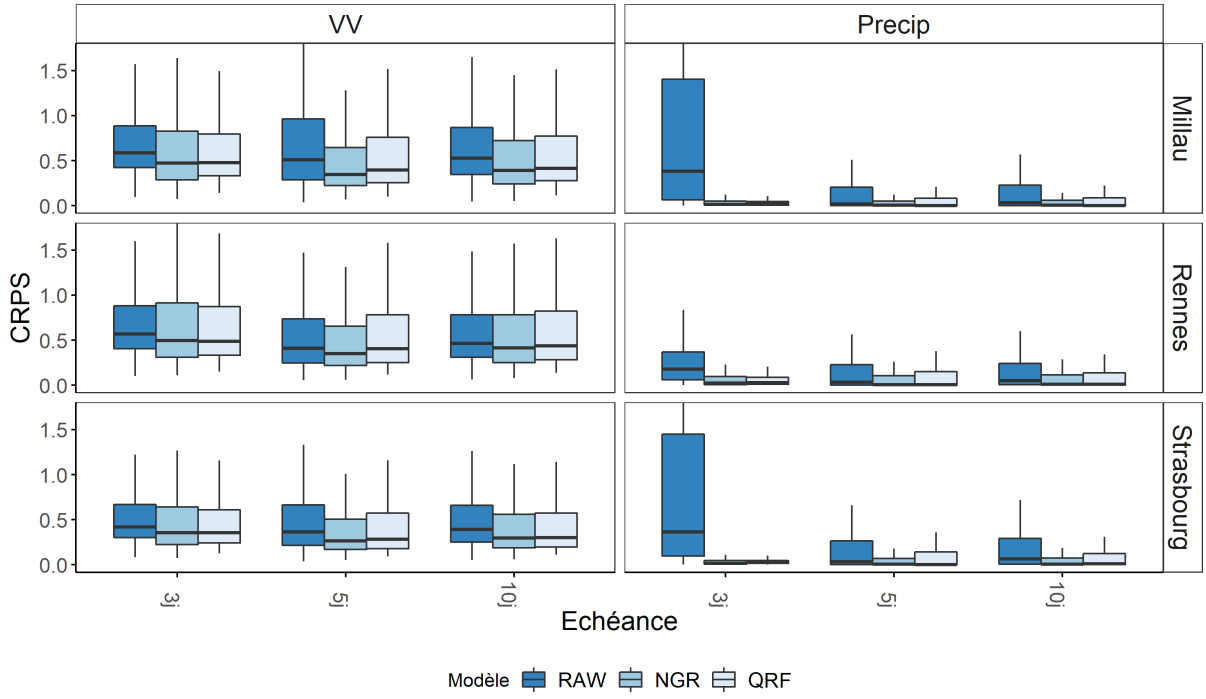


FIGURE 1.3 – Boites à moustaches des CRPS estimés par variable météorologique, modèle, localisation spatiale et échéance de prévision. *Les modèles sont en couleurs et les scores des stations sont affichés par lignes et des variables météorologiques par colonnes, VV : vitesses du vent ; Precip : précipitations. RAW : ensembles du modèle de prévisions numériques ; NGR : ensembles du modèle de régression non homogène ; QRF : ensembles du modèle de forêt aléatoire.*

Analyses CRPS. La figure 1.3 montre les scores CRPS résumés pour les deux horaires (6H et 18H) et ce par échéance de prévision et localisation spatiale. Ces figures affichent des scores plus élevés pour les ensembles RAW de vitesses du vent et de précipitations que ceux des modèles NGR et QRF et ce pour toutes échéances de prévision et localisations spatiales (excepté pour les vitesses du vent à Rennes). Les ensembles RAW de précipitations montrent des valeurs de CRPS plus importantes aux stations de Millau et Strasbourg que Rennes pour des échéances de 3 jours. Les stations de Millau et Strasbourg présentent des caractéristiques météorologiques locales amenant des précipitations importantes et difficiles à prévoir pour des horizons de 3 jours. Ajouté à cela, les ensembles des courtes échéances comme 3 jours disposent de distributions plus étroites générant des scores élevés en cas d'erreurs de prévision. La station de Rennes est quant à elle située dans une zone entourée par deux importantes étendues d'eau avec un climat plus doux, produisant de rares épisodes de précipitations importantes et difficiles à prévoir. Les résultats de CRPS confirment l'apport des post-traitements par les modèles univariés. De plus, le CRPS de la figure 1.3 montre une correction plus importante pour la variable de précipitations, et ce pour toutes localisations et échéances.

Le modèle NGR obtient des scores de calibration similaires ou légèrement supérieurs en moyennes à ceux du modèle QRF pour les vitesses du vent et les précipitations. Cependant, cette différence de score reste peu significative. Les histogrammes de rangs montrent que des erreurs de distributions subsistent de manière plus prononcée au sein des ensembles des modèles NGR que des modèles QRF. L'analyse des coefficients de régression des modèles NGR et des scores d'importance des covariables des modèles QRF présentée en annexe A.1.2 permet d'obtenir quelques indications sur les performances de calibration des modèles. Par exemple, les modèles révèlent des corrections avec une différence spatiale plus importante que la différence entre les échéances de prévision pour les ensembles de vitesse de vent. Les ensembles de vitesses de vent semblent être gouvernés par différents régimes suivant la localisation spatiale nécessitant des corrections particulières. Ces corrections prennent davantage l'information temporelle (aspect saisonnier et diurne des prévisions des vitesses de vent) et de la prévision déterministe HRES ainsi que de la médiane des ensembles. En parallèle, la calibration des précipitations est plus affectée par l'évolution des échéances avec un apport des statistiques décrivant la forme de la distribution des ensembles plus marqué que dans le cas des ensembles de vitesses du vent.

Pour conclure, les ensembles univariés générés par les modèles QRF et NGR affichent

de meilleures performances de calibration que celles des ensembles du modèle numérique malgré des erreurs de distributions subsistantes au sein des distributions d'ensembles post-traités. Les coefficients et scores d'importance des covariables des modèles NGR et QRF montrent que les résultats de calibration affichés par les histogrammes de rangs et CRPS semblent caractérisés en grande partie par l'information statistique des quantiles, moyenne, écart-type issus des ensembles ainsi que par la prévision déterministe HRES et les facteurs temporels suivant les différentes stations et échéances.

1.3.4.2 Résultats de calibration multivariée

Après avoir analysé l'apport des modèles univariés, les approches de calibration multivariée sont comparées grâce au score d'énergie.

Les résultats de l'approche $SimSS$ couplée aux modèles NGR et QRF sont présentés et comparés aux ensembles RAW par station et échéance de prévision dans la figure 1.4. Les scores d'énergie des modèles $NGR/SimSS$ et $QRF/SimSS$ sont inférieurs ou équivalents en moyenne au score d'énergie des ensembles RAW pour toutes localisations et échéances confondues. Les stations de Millau et Strasbourg affichent des différences plus marquées entre les modèles de calibration et les ensembles RAW. Les erreurs des ensembles multivariés à Millau et Strasbourg sont plus importantes qu'à Rennes. Il semblerait que la relation entre les erreurs des ensembles multivariés et la météorologie locale soit plus forte et mieux perçue par les modèles à ces localisations. Ensuite, le modèle $NGR/SimSS$ montre une distribution du score d'énergie plus élevé que celle du modèle $QRF/SimSS$. La structure multivariée dessinée par les variables de vitesses du vent et de précipitations semble être complexe à approcher par des modèles paramétriques couplés à une approche basée sur des copules empiriques. Cette différence est d'autant plus marquée pour les échéances de 3 et 5 jours. Pour une échéance de 10 jours, la calibration apportée par les modèles $NGR/SimSS$ et $QRF/SimSS$ devient quasi équivalente.

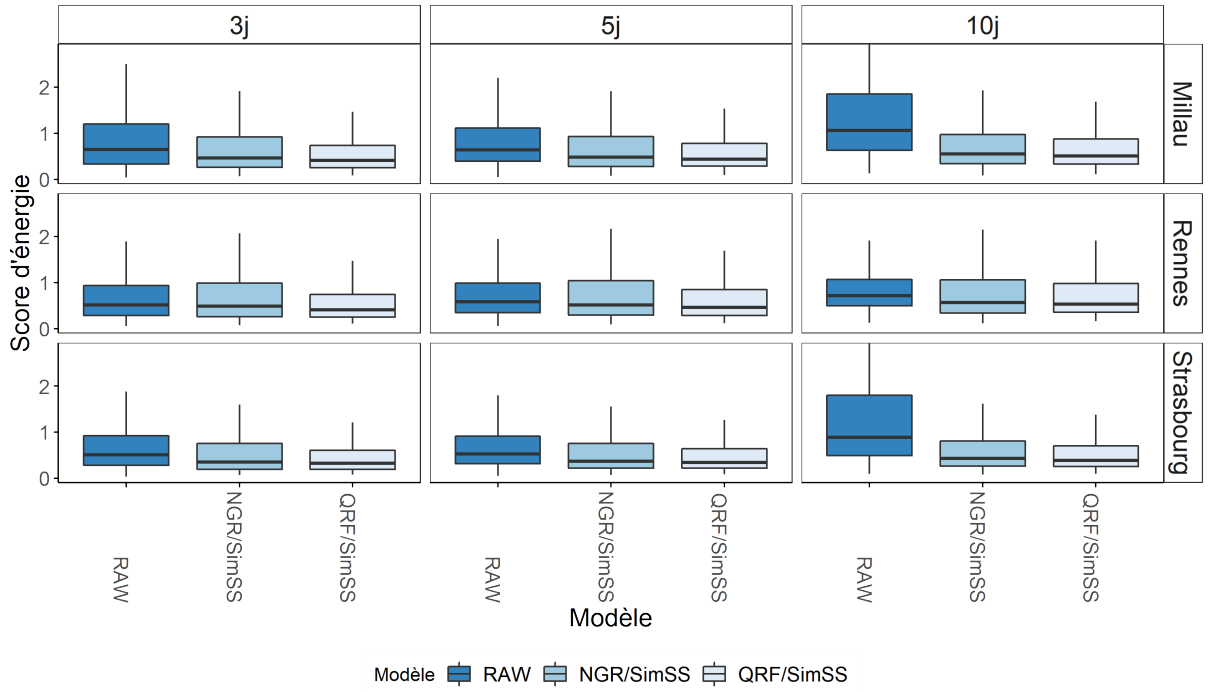


FIGURE 1.4 – Boites à moustaches des ES estimés par modèle, localisation spatiale et échéance de prévision. *Les stations sont représentées par lignes, échéances par colonnes et modèles par couleurs. RAW : ensembles du modèle de prévisions numériques ; NGR : ensembles du modèle de régression non homogène ; QRF : ensembles du modèle de forêt aléatoire.*

Pour conclure, l'algorithme $SimSS$ couplé aux modèles de régression NGR et de forêt QRF fournit des ensembles multivariés autant calibrés en moyenne que les ensembles RAW. Le modèle $QRF/SimSS$ obtient la meilleure calibration multivariée aux échéances de 3 et 5 jours. Cependant, les ensembles générés par les modèles $NGR/SimSS$ et $QRF/SimSS$ contiennent toujours des erreurs ne permettant pas de s'affranchir significativement des performances des ensembles RAW pour les échéances de 3 et 5 jours.

1.4 Conclusion

Dans chapitre, il a été question d'introduire différentes approches de calibration univariée et multivariée. La méthode de calibration univariée paramétrique de régression non homogène gaussienne (NGR) a été présentée, ainsi que deux extensions modélisant les vitesses du vent et les précipitations. Pour mieux mettre en perspective les performances

de ces modèles, l'approche non paramétrique de forêt aléatoire de régression appliquée à l'estimation de quantiles (QRF) a été étudiée. Dans le cas de la calibration multivariée, le modèle Schaake shuffle étendu par Schefzik 2016 ($SimSS$) basé sur les copules empiriques est présenté.

Ces modèles sont ensuite appliqués à la calibration d'ensembles dans un cadre de données de vitesses du vent et de précipitations pour des localisations et échéances de prévision différentes. De manière générale, les modèles ont montré des performances de calibration univariée supérieures à celles des ensembles du modèle numérique. Ensuite, les modèles NGR de S. Baran et Lerch 2016 et Scheuerer et Hamill 2015a ont obtenu des résultats de calibration proches, voire meilleurs, dans certains cas que ceux des modèles QRF. Les précipitations ont montré des erreurs importantes avec un caractère local, plus fort à une échéance de prévision de 3 jours. Pour cette même variable, la correction des erreurs apportée par les modèles a été d'autant plus remarquée. Néanmoins, les histogrammes de rangs révèlent que des erreurs continuent de résider dans les ensembles post-traités par les deux méthodes. L'analyse des coefficients de régression des modèles NGR et des covariables des modèles QRF permet de mettre en lumière certains liens entre les erreurs univariées de calibration et la physique des variables météorologiques étudiées. Néanmoins, les modèles utilisés sont limités à de simples résumés statistiques de la distribution des ensembles.

Dans leurs travaux, Taillardat, Mestre et al. 2016 indiquent une piste pour améliorer les résultats de calibration univariée du modèle QRF. L'idée est d'identifier d'autres variables aidant à décrire des phénomènes locaux reliés à la variable météorologique d'intérêt, et mal résolus par le modèle de prévision numérique. Par exemple, pour des stations montagneuses, en période hivernale, la réflexion des rayons solaires provoquée par la neige est régulièrement responsable des erreurs de prévision de température. Intégrer les données de prévision de rayonnement solaire et d'enneigement au modèle pourrait aider à la calibration des ensembles de températures pour ces stations. D'autres travaux plus récents proposent des modèles de calibration hybride entre approches non paramétriques et paramétriques (Scheuerer et Hamill 2019 ; Taillardat, Mestre et al. 2016). Typiquement, la méthode non paramétrique est utilisée en amont pour fournir une information *a priori* permettant d'enrichir le modèle paramétrique. De ce fait, le modèle hybride bénéficie des capacités non linéaires intéressantes pour approcher des distributions complexes et de la flexibilité des approches paramétriques.

Dans un second temps, la méthode de Schaake shuffle est appliquée à la correction

des erreurs de distribution formée par les variables de vitesses du vent et de précipitations. Plus particulièrement, la méthode se base sur les approches univariées introduites précédemment ($NGR/SimSS$, $QRF/SimSS$) pour corriger les erreurs de distributions des ensembles multivariés. Les résultats apportés par les modèles $NGR/SimSS$ et $QRF/SimSS$ montrent des performances de calibration multivariée meilleures ou équivalentes en moyenne face à celle des ensembles RAW. De plus le modèle $QRF/SimSS$ affiche de meilleurs résultats que ceux du modèle $NGR/SimSS$ à faibles échéances. Ce modèle dispose de plus de facilité à approcher des erreurs de distribution multivariées dans un contexte où la structure multivariée des ensembles est moins dégradée par les échéances. Cependant, ces résultats restent très proches des valeurs de scores obtenus par l'ensemble du modèle numérique initial. Cela illustre la difficulté à fournir des corrections satisfaisantes d'erreurs de distribution d'ensembles multivariés. De plus, les modèles de calibration multivariée utilisés ici rendent difficile l'étude des corrections apportées sur le plan physique. En effet, il est compliqué dans cette situation d'identifier des phénomènes physiques responsables des erreurs de distribution des ensembles multivariés de prévision.

Pour résumer, ce chapitre a abordé :

- Le problème de correction des erreurs de distribution d'ensemble de prévision appelé la calibration d'ensemble. Plus particulièrement, différentes méthodes classiquement employées dans un cadre de calibration univariée et multivariée ont été présentées ;
- La capacité des modèles de calibration univariée à fournir des résultats de calibration supérieurs à ceux du modèle numérique de prévision d'ensemble ;
- La complexité d'obtenir des résultats de calibration multivariée s'affranchissant significativement de ceux des ensembles du modèle numérique.

Chapitre 2

Prédiction de classes météorologiques à partir d'ensembles de prévision à moyen terme

---

En introduction, l'objectif opérationnel citait l'intérêt de la prévision de situations météorologiques dans la planification d'opérations de maintenance ou événementielles. Les prédictions des situations météorologiques définies à partir de plusieurs variables météorologiques font intervenir des méthodes de calibration multivariées. Les problèmes de calibration multivariée sont classiquement abordés de deux manières. La première consiste à considérer un ensemble de calibrations univariées, puis d'appliquer une procédure de réarrangement pour améliorer la structure multivariée des ensembles calibrés. L'alternative consiste à modéliser la loi jointe puis de corriger ses paramètres pour que la loi de l'ensemble coïncide avec celle des observations. La difficulté est alors de proposer un modèle de loi jointe suffisamment flexible pour bien décrire la distribution.

Dans ce chapitre, une alternative est proposée : le cadre multivarié est d'abord simplifié en discrétisant l'ensemble de définition de la loi jointe. On se ramène alors à un problème de classification avec une sortie univariée, chaque classe correspondant à une partie de la distribution jointe. Une approche classique de ce problème de classification passe par l'application de modèles prédisant directement une classe à partir des ensembles de prévision. Dans cette étude, nous comparons deux modèles. Le premier modèle non linéaire de forêt aléatoire de classification (RFC ; "Random Forest Classifier") de Breiman 2001 est un algorithme incontournable dans la résolution de ce type de problème. Ce modèle se base sur l'agrégation des prédictions d'algorithmes d'arbres de classification définis par Breiman et al. 1984. Plus particulièrement, l'approche par arbres de décision est un algorithme glouton inférant une partition de l'espace de la variable à prédire et générée à

partir d'une collection de règles issues des prédicteurs. L'algorithme de forêt dispose également d'un score permettant l'évaluation de la contribution des covariables dans la qualité de la prédiction fournie par le modèle. L'algorithme paramétrique de régression multinomiale Lasso (MLR; "Multinomial Lasso Regression") décrit par Friedman, Hastie et Tibshirani 2010 est considéré comme une alternative. C'est une extension du modèle linéaire classique de régression logistique pour un cadre de prédiction multi-classes. La partie Lasso ("Least Absolute Shrinkage and Selection Operator") permet la régularisation du modèle en sélectionnant un sous-ensemble de variables. De ce fait, les deux modèles retenus permettent d'évaluer les performances de classification et de comparer les covariables contribuant le plus à la prédiction pour un modèle non linéaire et un modèle linéaire.

L'utilisation des modèles de forêt et de régression multinomiale Lasso est appelée par la suite "approche de classification directe". Une autre méthode consiste à déduire des classes à partir des ensembles générés par les modèles de calibration multivariée étudiés dans le chapitre 1. Pour cela, l'algorithme de calibration multivariée de Schaake shuffle (Clark et al. 2004 ; Schefzik 2016, section 1.2.2). Les modèles de calibration univariés utilisés dans l'étape 2 de l'algorithme de Schaake shuffle sont l'algorithme de forêt aléatoire de régression appliquée à l'estimation de quantile (QRF, section 1.1.2.2), et le modèle paramétrique de régression gaussienne non homogène (NGR, section 1.1.1).

En section 2.1, la méthode de classification directe est présentée, se basant sur le modèle de forêt aléatoire (RFC) ou sur la régression multinomiale Lasso (MLR). La seconde méthode de classification déduite d'une calibration multivariée est aussi décrite. Ensuite, dans la section 2.2, l'application des méthodes de classification sur les données d'ensemble de prévision du modèle européen et d'observations SYNOP (présentées dans le chapitre précédent) est étudiée. La création de 4 classes, nommées "Bon", "Venteux", "Pluvieux" et "Venteux et Pluvieux", et construites à partir d'observations de vitesses du vent et précipitations, est présentée en section 2.2.1. Au travers de ces 4 classes, le but est d'obtenir une interprétation météorologique des classes d'observations prédites par une approche de classification utilisant l'information des ensembles de prévision. Les scores d'évaluation sont calculés à partir de la matrice de confusion permettant l'évaluation des résultats de classification, et les scores d'importance des covariables à partir du modèle de forêt aléatoire de classification. Ensuite, les résultats des méthodes de classification sont affichés et analysés.

2.1 Classification d'évènements météorologiques multivariés

La problématique de calibration multivariée transposée en une problématique de classification amène à définir deux nouvelles méthodes. L'idée de cette section est de poser le problème de classification et de définir les termes des différents éléments utilisés pour présenter les deux approches de prédiction de classes. Ensuite, une première approche naturelle pour la prédiction de ces classes à l'aide de modèle de classification par forêt aléatoire et par régression multinomiale Lasso est présentée. Une seconde méthode de classification utilisant les ensembles générés par les méthodes de calibration multivariée ($QRF/SimSS$ et $NGR/SimSS$) est introduite.

2.1.1 Définition du problème

Dans le contexte de calibration d'ensembles de prévision multivariés présentant des interactions entre d variables météorologiques, les notations du chapitre précédent sont reprises. Soit $y_t \in \mathbb{R}^d$ une observation et $X_t^* = (x_{tm})_{1 \leq m \leq M}$ un ensemble de M prévisions avec $x_{tm} \in \mathbb{R}^d$ issues de la variable aléatoire X_t. La variable discrète de classe $Z_t \in \{1, \cdots, K\}$ est définie où chaque $k \in \{1, \cdots, K\}$ pointe vers un sous-ensemble d'observations, i.e, $Z_t = k$ signifie que l'observation $y_t \in R_k$ où R_k représente un sous-espace rectangulaire de $\mathbb{R}^d$ délimité par un ensemble de valeurs fixe τ_k.

Ainsi, notre objectif complexe de correction des observations y_t suivant les ensembles X_t^* multivariés définis en (1.1) devient un objectif simplifié de classification supervisée :

$$f(X_t^*) = p(Z_t | X_t) \tag{2.1}$$

Contrairement à l'objectif de calibration (1.1), la définition (2.1) se place dans une problématique de prédiction de classe Z à partir des interactions entre les variables météorologiques des observations. Dans cette optique de classification, deux approches sont étudiées. L'idée est d'analyser la différence de résultats entre des modèles de références, accomplissant une classification directe, et des modèles proposant une classification issue d'ensembles post-traités par un modèle de calibration multivariée.

2.1.2 Classification directe

Dans cette section, le modèle non paramétrique de forêt aléatoire de classification est abordé, ensuite, le modèle paramétrique de régression multinomiale Lasso est présenté. Le jeu de données d'individus indépendants $\mathcal{X} = \{z_t, x'_t\}_{1 \leq t \leq T}$ est utilisé pour introduire les modèles, avec :

- z_t une réalisation de Z une variable aléatoire discrète symbolisant les K classes à valeurs dans $k \in \{1, \cdots, K\}$;
- $x'_t \in \mathcal{B} \subseteq \mathbb{R}^p$ un ensemble de p covariables caractérisant l'ensemble X^*_t.

2.1.2.1 Forêt aléatoire de classification

Le modèle de forêt aléatoire appliqué dans le cas d'une classification s'appuie directement sur le modèle de Breiman 2001 présenté dans la section 1.1.2.1. Cependant, ici la variable objectif Z est une variable discrète, donc la variance empirique n'est plus adaptée pour partitionner l'espace en régions homogènes. De ce fait, l'indice de Gini est utilisé. Il est défini pour un groupe de données c de la façon suivante :

$$G(c) = \sum_{k=1}^{K} p_c^k (1 - p_c^k) \tag{2.2}$$

avec p_c^k la proportion d'individus z_t présents dans un noeud θ_n^i du $n^{\text{ème}}$ arbre de classification et associée à la classe k. L'indice de Gini apporte une information sur la répartition des classes au sein d'un groupe $c \in \mathcal{C}_i$. Pour rappel $\mathcal{C}_i$ symbolise l'ensemble des groupes de données formés par une séparation du noeud θ_n^i. Plus l'indice est élevé, plus la répartition des classes est inégale. Avec cet indice, l'homogénéité H (1.21) est redéfinie :

$$H(\theta_n^i) = G(\mathcal{C}_i) - \sum_{c \in \mathcal{C}_i} \frac{T_c}{T_{\mathcal{C}_i}} G(c) \tag{2.3}$$

Comme dans le cas de la forêt aléatoire de régression, l'opération de maximisation de H est itérée jusque convergence de l'algorithme. Enfin, la fonction objectif évaluant la qualité de l'entraînement de forêt aléatoire pour un problème de classification, que l'on cherche à minimiser, est aussi à redéfinir. Pour cela, la proportion d'éléments mal prédits lors de l'entraînement, issue des données $\mathcal{X}$, est évaluée grâce au critère de précision $ACC = \frac{1}{T} \sum_{t=1}^{T} 1_{\{z_t \neq \hat{z}_t\}}$, $\hat{z}_t$ étant la classe prédite par l'algorithme.

2.1.2.2 Régression Multinomiale Lasso

Le modèle de régression multinomiale proposé par J. Zhu et Hastie 2004 est une généralisation du modèle logistique dans un cas où la variable objectif prend un nombre de classes K supérieur à 2. Dans cette situation, la loi $Z|X$ suit une loi $\mathcal{M}ultinomiale(\pi_1, \cdots, \pi_K)$, où chaque probabilité $\pi_k = p(Z = k|X = x')$ est modélisée par un modèle logit :

$$p(Z = k|X = x') = \frac{e^{\beta_{0k}+\beta_{1k}x'}}{\sum_{k'=1}^{K} e^{\beta_{0k'}+\beta_{1k'}x'}} \tag{2.4}$$

avec $(\beta_{0k}, \beta_{1k}) \in \mathbb{R}^{p+1}$ les coefficients de régression. À partir de ce modèle, Friedman, Hastie et Tibshirani 2010 proposent une pénalisation Lasso des coefficients de régression afin d'obtenir un modèle parcimonieux. Pour l'échantillon $\mathcal{X} = \{z_t, x'_t\}_{1\leq t\leq T}$, la log-vraisemblance pénalisée $\log \mathcal{L}_1$ s'écrit de la façon suivante :

$$\log \mathcal{L}_1(\mathcal{X}; \beta) = \frac{1}{T}\sum_{t=1}^{T}[\sum_{k=1}^{K} log(p(z_t = k|X = x'_t))] - \lambda \sum_{k=1}^{K} \|\beta_k\|_1 \tag{2.5}$$

avec $\beta = \{\beta_k = (\beta_{0k}, \beta_{1k})\}_{1\leq k\leq K}$ l'ensemble des coefficients de régression.

Pour la suite du chapitre, le modèle de forêt aléatoire de classification sera nommé **RFC**, et le modèle de régression multinomiale Lasso sera nommé **MLR**.

2.1.3 Classification issue d'une calibration multivariée

La méthode de classification issue d'une calibration multivariée s'appuie sur les ensembles tirés de la méthode de calibration multivariée de Schaake shuffle (SimSS) basée sur des modèles de calibration univariés. Les modèles de calibration sont ajustés sur un jeu de données $\mathcal{X} = \{y_t, X_t^*\}_{1\leq t\leq T}$ d'observations et d'ensembles de prévision. À partir de la connaissance des espaces formant les classes d'observations, les réalisations de l'ensemble post-traité peuvent être assignées à chaque région R_k. Ensuite, une méthode de vote à majorité est appliquée sur l'ensemble des réalisations classées pour obtenir une seule et unique prédiction. Pour résumer, la méthode de classification se décompose en trois étapes :

1. Application d'un modèle de calibration multivarié : Après avoir sélectionné une méthode de calibration univariée, les quatre étapes de la méthode Schaake shuffle (décrites dans la section 1.2.2) sont appliquées.

2. Prédiction d'un ensemble de classes : Dès lors que le nouvel ensemble multivarié post-traité $\hat{X} = (\hat{x}_m)_{1 \leq m \leq M'}$ est généré, un ensemble de classes $Z^* = (z_m)_{1 \leq m \leq M'}$ est déduit où chaque $z_m = k$ si $x_m \in R_k$.

3. Prédiction d'une nouvelle classe : Dans le but de prédire une unique classe z, une règle d'agrégation est utilisée sur l'ensemble de classes Z^* :

$$z = \arg\max_k \frac{1}{K} \sum_{m=1}^{M} 1_{\{x_m \in R_k\}} \tag{2.6}$$

La figure B.1 résume les différentes étapes de l'apprentissage du schéma de dépendance pour un cas d'ensembles et d'observations bivariés.

Dans l'étape 1, deux méthodes de calibration univariée seront utilisées dans la suite : le modèle paramétrique de régression gaussienne non homogène (NGR, décrit en 1.1.1.1) et le modèle non paramétrique de forêt aléatoire de régression appliqué à l'estimation de quantile (QRF, décrit en 1.1.2.2). La classification à l'issue des 3 étapes est alors nommée "$NGR/SimSS$" ou "$QRF/SimSS$".

2.2 Application des méthodes de classification

Maintenant que le problème de classification, ainsi que deux méthodes de résolution ont été définis, la prochaine étape est de comparer ces méthodes sur des données réelles. Dans un premier temps, les classes sont définies à partir des observations multivariées et de règles linéaires. Les scores tirés de la matrice de confusion et permettant d'évaluer les performances de classification des modèles dans une problématique à classes multiples sont présentés. Pour terminer, les résultats des modèles issus des deux méthodes appliquées sur les données réelles, utilisées dans le chapitre précédent en section 1.3.1, sont présentés et discutés. Pour rappel, les données réelles se composent des ensembles du centre européen CEPMMT, et des observations du réseau de stations SYNOP. Ces données sont sélectionnées pour les variables de vitesses de vent et de précipitations. Mais également, elles sont récupérées pour trois stations (Millau, Rennes et Strasbourg), deux horaires correspondant à l'heure d'initialisation du modèle de prévision (6H et 18H), et enfin trois horizons de prévision (3, 5 et 10 jours).

2.2.1 Définition des classes météorologiques

Comme mentionnée dans la section 2.1.1, les classes sont construites à partir d'observations y multivariées et de règles linéaires fixées. Notamment, ces règles forment une partition rectangulaire de l'espace de valeurs des observations. Un ensemble de $K = 4$ sous-espaces rectangulaires $(R_k)_{1 \leq k \leq K}$ est défini :

$$\begin{aligned} \text{Bon} \quad & R_1 = [0, \tau^{TP}[\times [0, \tau_1^{VV}] \\ \text{Venteux} \quad & R_2 = [0, \tau^{TP}[\times]\tau_1^{VV}, +\infty[\\ \text{Pluvieux} \quad & R_3 =]\tau^{TP}, +\infty[\times [0, \tau_1^{VV}] \\ \text{Venteux et pluvieux} \quad & R_4 =]\tau^{TP}, +\infty[\times]\tau_2^{VV} + \infty[\end{aligned} \tag{2.7}$$

où τ^{TP} est un seuil de précipitations et $(\tau_1^{VV}, \tau_2^{VV})$ deux seuils de vitesses de vent, le tout à fixer.

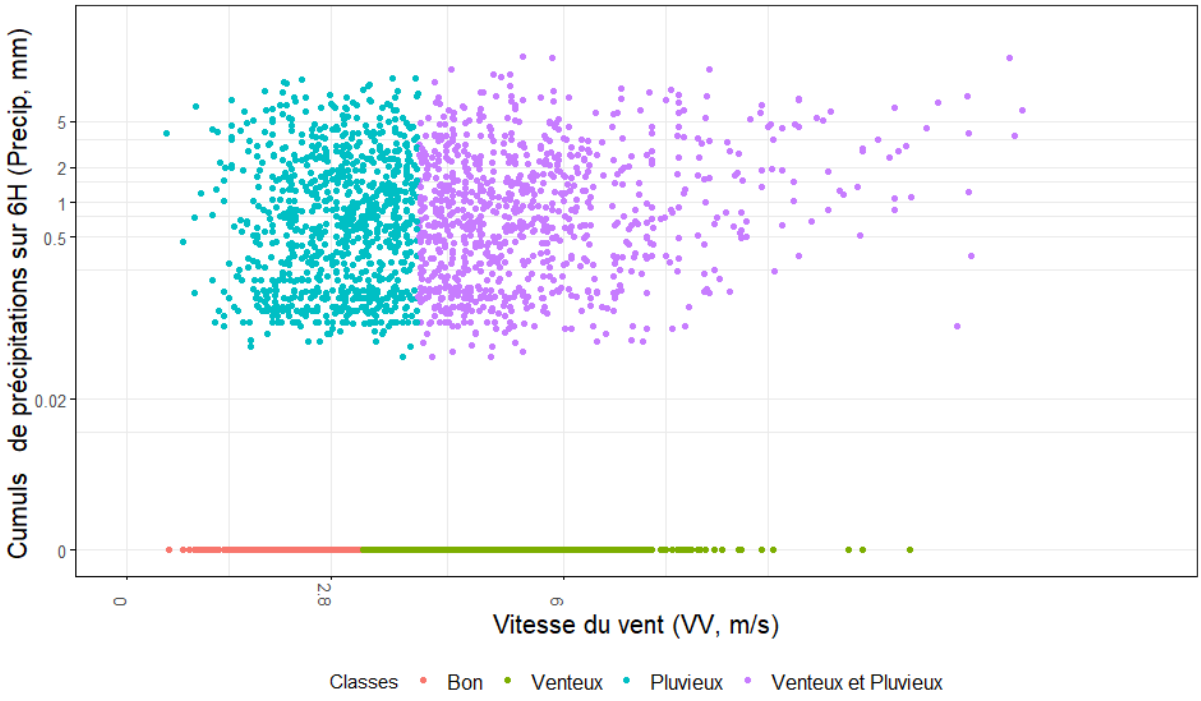


FIGURE 2.1 – Exemple de découpage par classes obtenu sur les observations de la ville de Rennes sur la période 2008-2018. *Les variables sont représentées en échelle logarithmique.*

Les seuils permettant de construire les classes (2.7) ont été choisis, dans un souci de planification d'événements nécessitant des précipitations et un vent faibles. Mais éga-

lement, ces seuils ont pour objectif de former des classes avec une population la plus homogène possible pour le jeu de données étudié. Le seuil de pluie minimale est maintenu à $\tau^{TP} = 0$ parmi les trois localisations sélectionnées à partir de la climatologie des sites. Ainsi, les prévisions avec une valeur supérieure à ce seuil sont assignées à des événements avec des précipitations. Les seuils de vent varient selon la localisation afin de garder un certain équilibre dans la répartition des classes. Pour Millau, les seuils de vent sont fixés à $\tau_1^{VV} = 4$ et $\tau_2^{VV} = 4$, pour Rennes à $\tau_1^{VV} = 2.8$ et $\tau_2^{VV} = 4$, et enfin pour Strasbourg à $\tau_1^{VV} = 2.1$ et $\tau_2^{VV} = 3$. La sélection de deux seuils de vent permet de garantir une certaine homogénéité dans le nombre d'individus par classe pour les sites de Rennes et Strasbourg, et de garder une interprétation météorologique entre les classes de vent faible et de vent fort selon s'il y a de la pluie ou non. La figure 2.1 montre la répartition des observations par espace R_k pour les données d'observations de Rennes. En moyenne, une répartition d'individus à 30% pour la classe Bon, 20% pour la classe Pluvieux, 30% pour celle Venteux et 20% pour la classe Venteux et Pluvieux est obtenue pour l'ensemble des sites.

2.2.2 Scores dérivés à partir de la matrice de confusion

La matrice de confusion est l'outil classiquement utilisé dans l'évaluation et la validation de problèmes de classification supervisée et non supervisée. Pour un ensemble d'individus observés et prédits $\mathcal{Z} = \{z_t, \hat{z}_t\}_{1 \leq t \leq T}$, la matrice de confusion est une matrice carrée, indexée par le nombre de classes K observées en ligne et prédites en colonne. Chaque élément n_{ij} de la matrice représente le nombre des individus observés z_t de la classe i dans la classe j, soit :

$$n_{ij} = \sum_{t=1}^{T} 1_{\{z_t = i, \hat{z}_t = j\}} \tag{2.8}$$

avec $(i, j) \in \{1, \cdots, K\} \times \{1, \cdots, K\}$.

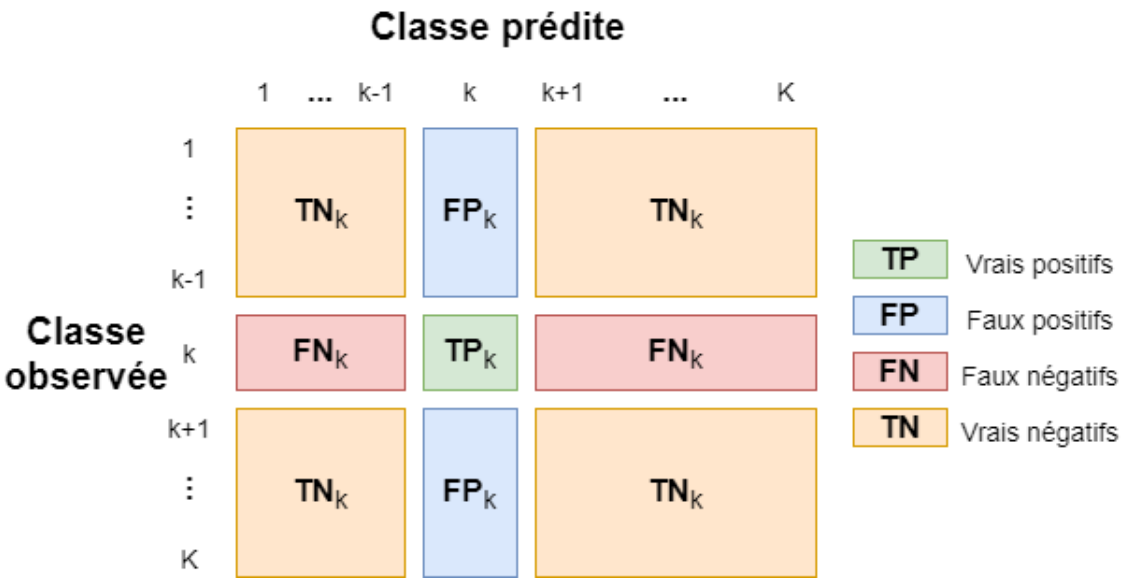


FIGURE 2.2 – Matrice de confusion estimée dans un cas à K classes. *Schéma tiré de* KRÜGER *2018.*

Sur les n_{ij} éléments, les quantités suivantes sont définies $\forall k \in \{1, \cdots, K\}$:

$$\begin{aligned} TN_k &= \sum_{i \neq k}^{K} \sum_{j \neq k}^{K} n_{ij} \\ FN_k &= \sum_{i=k}^{K} \sum_{j \neq k}^{K} n_{ij} \\ TP_k &= n_{kk} \\ FP_k &= \sum_{i \neq k}^{K} \sum_{j=k}^{K} n_{ij} \end{aligned} \tag{2.9}$$

Les quantités ci-dessus sont représentées dans la figure 2.2, et permettent de définir des scores d'analyse de la qualité de la classification (RIJSBERGEN 1979).

Le score d'analyse de la performance générale de classification d'une méthode est la précision globale (ACC, "Accuracy") définie par :

$$ACC = \frac{1}{T} \sum_{k=1}^{K} TP_k \tag{2.10}$$

La précision globale ACC est un score à maximiser. Plus les éléments prédits sont égaux

à ceux observés dans la base de données, plus l'ACC tendra vers 1. Néanmoins, ce score basé principalement sur la diagonale de la matrice de confusion est sensible à la proportion d'individus présents dans chaque classe. Juger la performance de classification d'un modèle uniquement sur ce type de score peut biaiser l'interprétation des résultats. Afin d'éviter cette situation, les scores de précision (PPV, "Positive Predictive Value") et rappel (TPR, "True Positive Rate" ou sensitivité), permettant l'analyse par classe k, sont étudiés. Ces deux scores se définissent de la façon suivante :

$$\begin{aligned} PPV_k &= \frac{TP_k}{TP_k + FP_k} \\ TPR_k &= \frac{TP_k}{TP_k + FN_k} \end{aligned} \tag{2.11}$$

Le score de précision (PPV) évalue la capacité de la méthode à classer correctement chaque individu de la base de données. Quant au score de rappel (TPR), il définit la probabilité de détection de chaque classe. Les scores PPV et TPR sont complémentaires dans leur interprétation. Une classification optimale donnera une valeur de 1 aux deux scores. Inversement, pour un modèle présentant des défauts de classification, ces scores tendront vers 0. Dans le cas où ces scores présentent des valeurs très différentes pour une classe étudiée, un problème de surapprentissage, ou de sous-apprentissage, doit être fortement envisagé.

2.2.3 Évaluation des méthodes de classification

Dans cette section, les résultats des méthodes de classification sont présentés et comparés à une classification déduite des ensembles du modèle de prévision numérique (RAW). La classe issue de l'ensemble du modèle numérique est prédite à l'aide d'un simple vote à la majorité. Les scores de la matrice de confusion sont calculés et affichés pour les modèles de la méthode de classification directe (MLR et RFC), et pour ceux de la méthode de classification issue d'une calibration multivariée ($NGR/SimSS$ et $QRF/SimSS$). Chaque score est estimé au travers de B=30 répétitions d'étapes d'apprentissage et de test du même jeu de données défini en section 1.3.4. Excepté le modèle $NGR/SimSS$, les autres modèles sont entraînés et testés pour les covariables issues des statistiques des ensembles et prévisions déterministes présentées en section 1.3.2. De plus, une covariable contenant les prédictions des classes issues de l'ensemble du modèle de prévision numérique, appelée "Raw Z", est ajoutée. Les paramètres du modèle de forêt aléatoire sont sélectionnés lors

d'une étape de validation croisée. Ainsi, un nombre d'arbres $N = 300$ et un nombre de noeuds égal à 20 ont été retenus. Le paramètre de pénalisation du modèle MLR est aussi sélectionné par validation croisée, le fixant à 0.0067.

Les résultats des scores de la matrice de confusion sont présentés pour chaque localisation et chaque échéance, et ce pour les deux horaires disponibles. Il importe d'indiquer pour le score de précision globale ACC que l'objectif des modèles avec $K = 4$ classes à prédire est de dépasser le seuil de $ACC = 0.25$, qui représente la performance qu'obtiendrait un modèle prédisant chacune des 4 classes de manière totalement aléatoire.

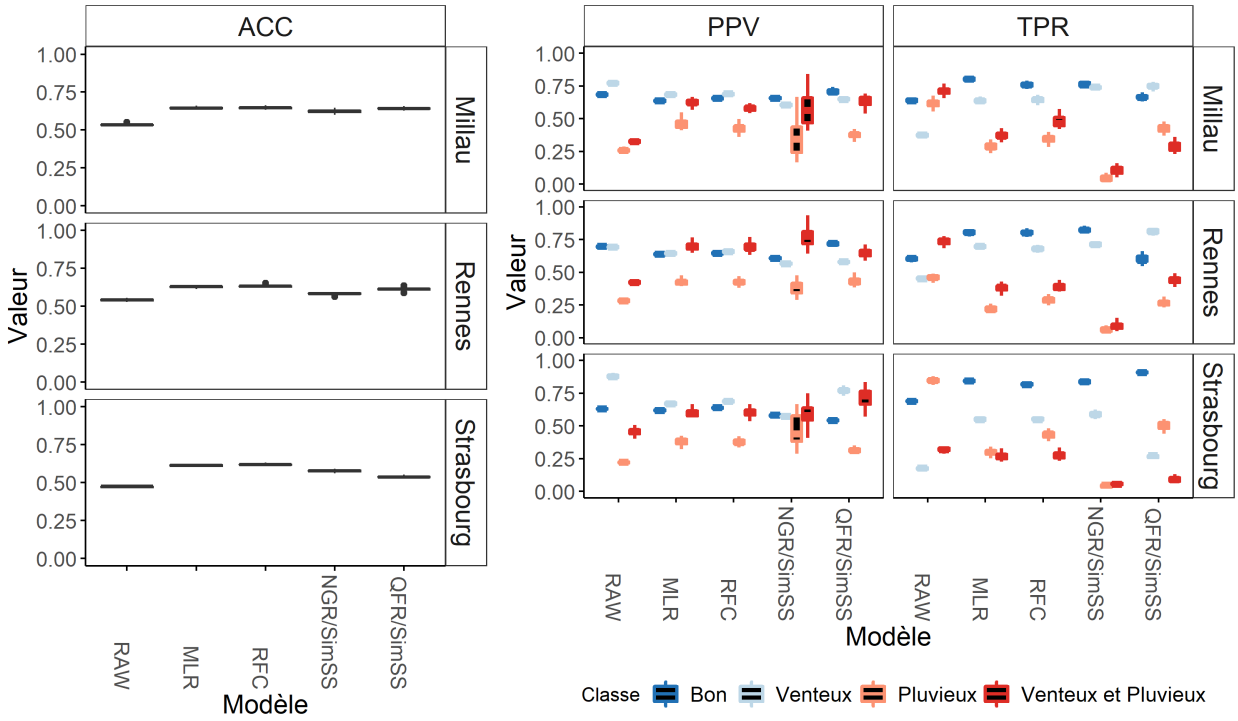


FIGURE 2.3 – Score de précision globale (ACC) par modèle et par station, scores de précision (PPV) et probabilité de détection (TPR) par modèle, par classe et par station pour une échéance de prévision de 3 jours. *Les stations sont représentées par des lignes, les scores par des colonnes et les classes par des couleurs.*

Performance de classification générale. Les figures 2.3, 2.4 et 2.5 présentent les scores estimés par modèles et station pour chaque échéance de prévision, respectivement, 3, 5 et 10 jours. Les performances de classification de la méthode RAW font office de référence. En première remarque, il est observé sur les figures 2.3, 2.4 et 2.5 que la précision globale (ACC) est maximisée par le modèle de forêt RFC, suivi de près par le

modèle de régression multinomial MLR, et ce pour toutes les localisations, avec un déclin de valeur avec les échéances de prévision. Les ACC des méthodes de classification issues de la calibration multivariée ($NGR/SimSS$ et $QRF/SimSS$) sont plus variées suivant les stations. Le modèle $QRF/SimSS$ affiche une meilleure classification à l'échéance 3 jours que celle du modèle $NGR/SimSS$, et une classification équivalente à l'échéance de 5 jours pour les stations de Millau et Rennes. Pour la station de Strasbourg, le modèle $NGR/SimSS$ montre toujours une qualité de classification supérieure à celle du modèle $QRF/SimSS$, et ce pour toutes les échéances. À une échéance de 10 jours, les performances de classification des ensembles du modèle $QRF/SimSS$ décroissent, passant en dessous de celles de $NGR/SimSS$ pour Rennes et sont équivalentes pour Millau.

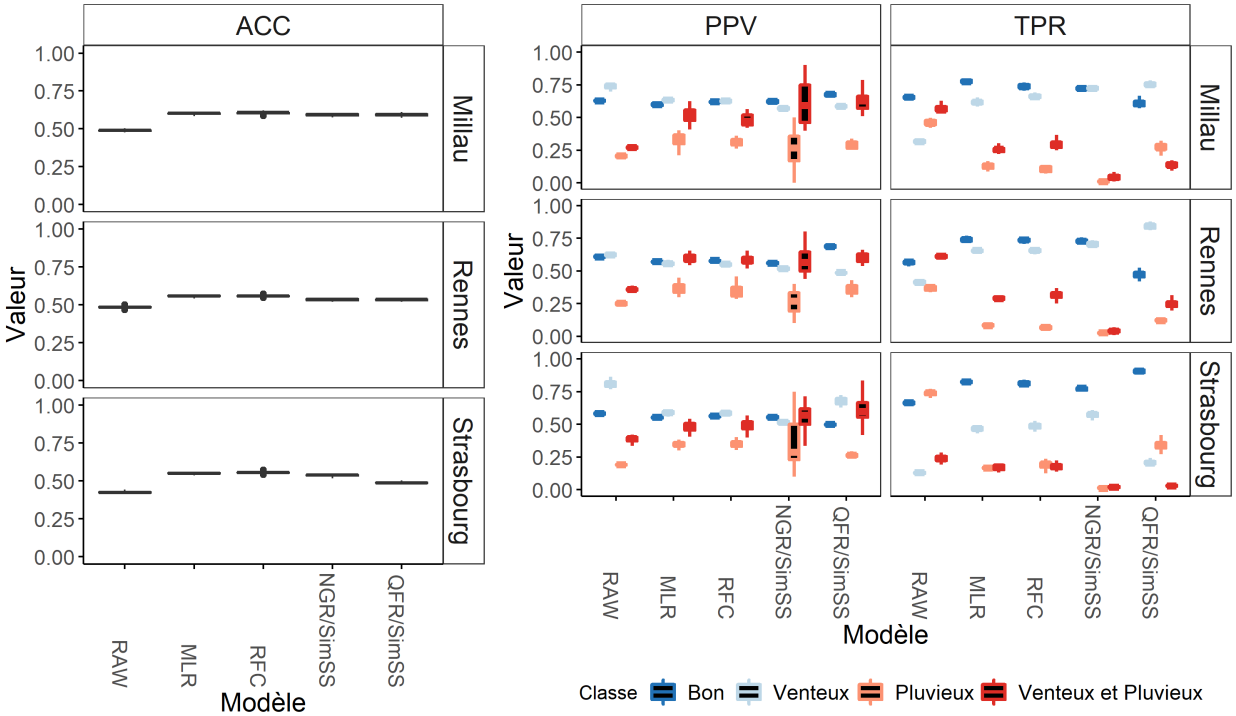


FIGURE 2.4 – Score de précision globale (ACC) par modèle et par station, scores de précision (PPV) et probabilité de détection (TPR) par modèle, par classe et par station pour une échéance de prévision de 5 jours. *Les stations sont représentées par des lignes, les scores par des colonnes et les classes par des couleurs.*

Analyse des performances de classification par classe. Les résultats par classe des scores de précisions (PPV) et de probabilité de détection (TPR) présentés au travers des figures 2.3, 2.4 et 2.5 permettent de comparer les performances de classification des

modèles en fonction des classes. Un premier résultat visible sur la figure 2.3, correspondant aux prédictions à l'échéance de 3 jours, est la séparation entre le score PPV des classes "Bon" et "Venteux" et ceux des classes "Pluvieux" et "Venteux et Pluvieux" obtenus par la méthode RAW. Pour le score TPR des classes prédites par la méthode RAW, la classe "Bon" obtient un score élevé, mais des fluctuations sont notées suivant les stations. En effet, les classes "Pluvieux" et "Venteux et Pluvieux" sont bien identifiées pour Millau,ce qui est moins le cas pour les classes "Venteux" et "Bon". Pour Rennes, les classes "Venteux" et "Pluvieux" sont moins bien détectées que les classes "Venteux et Pluvieux" et "Bon". Enfin, à Strasbourg, les classes "Bon" et "Pluvieux" sont mieux reconnues que les classes "Venteux" et "Venteux et Pluvieux". Chaque station affiche des différences de classification marquées en rapport avec la météorologie locale.

Toujours sur la figure 2.3, les modèles de la classification directe et de celle issue d'une calibration multivariée montrent une augmentation des scores PPV des classes de pluies, rejoignant ceux des classes sans pluie, avec des fluctuations suivant les stations. Les scores TPR montrent une séparation entre les classes avec et sans pluie. Ainsi, les classes avec pluie sont moins bien identifiées que celles avec pluie, et ce de manière générale. Les modèles (RFC et MLR) de la méthode de classification directe montrent des scores similaires avec des scores plus élevés pour les classes sans pluies. La diminution du score TPR et l'augmentation du score PPV pour les classes de pluies prédites par les modèles de la méthode directe indiquent une tendance à surapprendre ces classes, avec un manque d'information permettant de discriminer ces deux types de classes.

Le modèle $QRF/SimSS$ obtient quant à lui des résultats de classification similaires sur la figure 2.3, privilégiant les classes sans pluie et plus particulièrement la classe "Venteux" pour les stations de Millau et Rennes. Les résultats des modèles de calibration univariée (section 1.3.4.1) montraient des ensembles de précipitations du modèle numérique RAW sous-estimant les observations. Dans cette situation, le modèle QRF corrigeait les erreurs de sous-estimation, mais une légère erreur de sous-estimation apparaissait dans les nouveaux ensembles. Ce résultat pourrait être relié au fait que le modèle $QRF/SimSS$ prédit plus facilement les classes sans précipitations. Une hypothèse, concernant la mauvaise identification de la classe "Pluvieux" pour la plupart des modèles, est le manque d'information des covariables permettant de corriger l'erreur de prévision et caractériser correctement les classes. En conséquence, les modèles privilégient les classes de cardinal élevé, dont les classes sans pluie font partie.

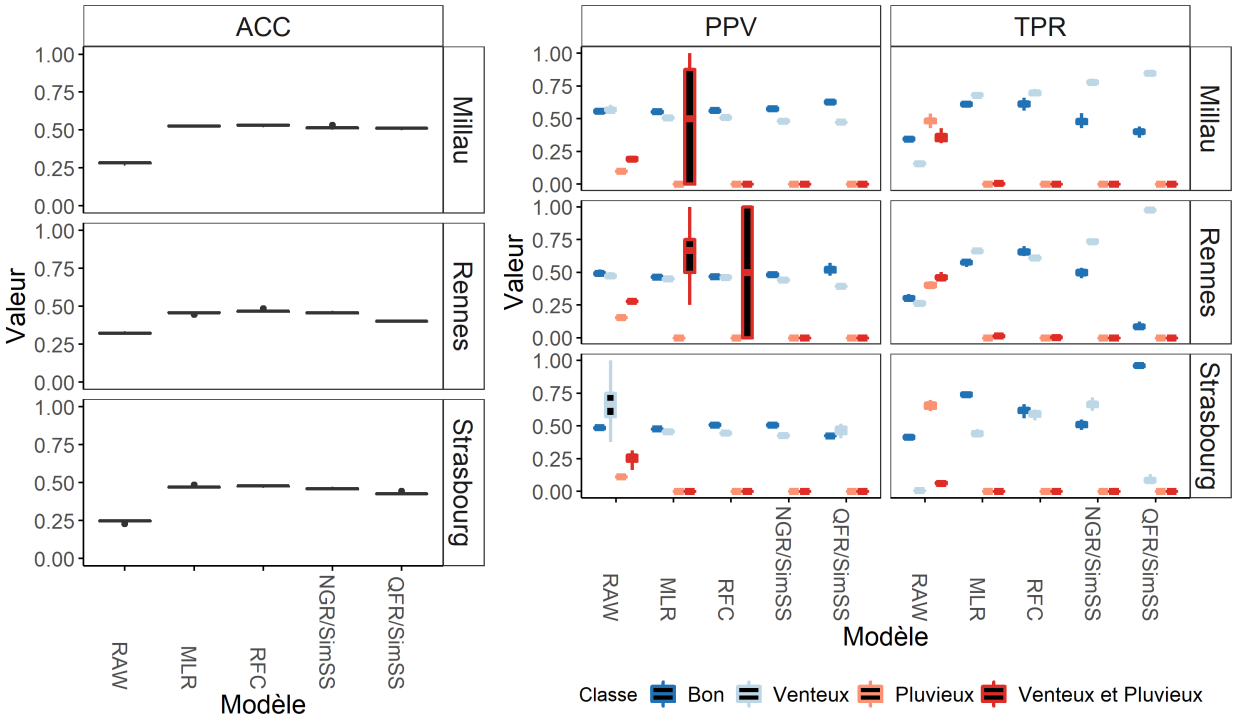


FIGURE 2.5 – Score de précision globale (ACC) par modèle et par station, scores de précision (PPV) et probabilité de détection (TPR) par modèle, par classe et par station pour une échéance de prévision de 10 jours. *Les stations sont représentées par des lignes, les scores par des colonnes et les classes par des couleurs.*

Les résultats des scores de classification sont donnés pour des échéances supérieures à 3 jours sur les figures 2.4 et 2.5 pour des ensembles de prévision aux échéances de 5 et 10 jours. Les scores de classification issus de la méthode *RAW* montrent un léger déclin suivant les échéances. Les performances de classification des modèles semblent être nettement plus impactées. En effet, la différence entre les classes pluie et sans pluie est encore plus marquée aux échéances de 5 et 10 jours. Le score PPV montre de grandes variations pour les classes de pluie du modèle $NGR/SimSS$ aux échéances 3 et 5 jours. Cette sensibilité semble indiquer la difficulté pour ce modèle à retrouver correctement les classes suivant les échantillons testés. De même qu'à 10 jours, une importante variation de ce score est notée pour la classe "Venteux et Pluvieux" du modèle MLR à Millau et des modèles MLR, RFC à Rennes. La variation de ce score montre que pour certains échantillons les modèles de classification directe arrivent à retrouver les éléments de cette classe. Néanmoins, le score TPR, restant proche de 0 pour ces cas, indique un problème de détection des classes de pluie. Les écarts de scores TPR montrent que les modèles tendent

à moins détecter les événements supérieurs au seuil de précipitations suivant les échéances de 5 et 10 jours. Il est également possible de voir qu'à une échéance de 10 jours, les scores de PPV et TPR sont presque nuls pour les classes de pluies prédites par les modèles.

Les scores d'importance de covariables du modèle RFC (résultats présentés dans l'annexe B.2.1) révèlent que les facteurs temporels ("month" et "hour") aident fortement à la prédiction des classes "Bon" et "Venteux". Les coefficients du modèle MLR (présentés dans l'annexe B.2.2) ajoutent l'importance de la prévision déterministe HRES et de la moyenne des ensembles des vitesses de vent dans la prédiction des classes citées, ceci pouvant varier suivant les localisations spatiales. L'analyse des scores et coefficients révèle également que la classe "Pluvieux" est plus caractérisée à faible échéance par les informations de précipitations. Notamment, les coefficients MLR soulignent l'impact des probabilités de dépassement de seuil de précipitations issues des ensembles. Enfin, la classe "Venteux et Pluvieux" bénéficie des informations des prévisions et ensembles de prévision des deux variables météorologiques, la rendant ainsi légèrement mieux identifiée par les modèles de classification directe que la classe "Pluvieux". Néanmoins, la forte dégradation de la qualité des ensembles de précipitations suivant les échéances impacte la prédiction de cette classe.

Les ensembles aux échéances de 3 et 5 jours contiennent assez d'informations statistiques et physiques pour que les modèles de classification directe puissent offrir une qualité de prédiction suffisante, et ce, pour toutes les classes. Dans le cas d'ensembles supérieurs ou égaux aux échéances de 10 jours, les modèles montrent des signes de surapprentissage important, omettant certaines classes comme les classes de pluies. Les précipitations représentent une variable météorologique locale et de court terme, difficile à prédire à des échéances moyennes comme 10 jours par les modèles numériques.

Pour résumer, les méthodes de classification directe offrent des résultats plus stables et plus performants que les méthodes de classification issues de calibrations multivariées. Les stations montrent un impact différent sur la qualité de prédiction des classes, affichant un lien local entre les erreurs de prévisions des vitesses de vent et des précipitations. Par exemple, à l'échéance de 3 jours pour Millau, les classes "Bon", "Venteux" et "Venteux et Pluvieux" sont mieux identifiées par le modèle RFC qu'à Rennes et Strasbourg. Enfin, les échéances dégradent l'information des ensembles de prévision, affichant nettement un déclin dans la qualité des prédictions effectuées par les modèles. Au-delà des échéances de 10 jours, représenter les prédictions de classes issues d'ensembles de précipitations et du modèle de prévision numérique semble être un choix plus adéquat, évitant les éventuelles

erreurs de surapprentissage des modèles de classification. La différence peu marquée entre les résultats des modèles de classification directe et ceux des modèles de classification issus d'une calibration multivariée tend à souligner un écart entre l'objectif de classification supervisée et celui de calibration des ensembles de prévision. Cet écart peut être dû au manque de lien entre les classes définies et les erreurs de distributions des ensembles.

2.3 Conclusion

L'objectif de ce chapitre était de proposer une transposition du problème de calibration multivariée en un problème de classification. Plus particulièrement, le problème de classification émet l'hypothèse que prédire une variable qualitative à partir des observations multivariées issues d'ensembles de prévision permet d'approcher la loi jointe, difficile à estimer avec des modèles de calibration multivariée. Pour résoudre ce problème de classification, deux méthodes ont été sélectionnées. La première, nommée méthode de classification directe, applique deux modèles classiques (régression et forêt) à la prédiction des classes. La seconde méthode prédit une classe à l'aide d'ensembles générés par un modèle de calibration multivariée. Les modèles de calibration utilisés ($NGR/SimSS$ et $QRF/SimSS$) sont construits à partir de l'algorithme Schaake shuffle ($SimSS$) et des méthodes de calibration univariée (NGR et QRF) étudiées dans le chapitre 1.

Les deux méthodes ont été appliquées sur des classes formées à partir de seuils d'intensité du vent et de précipitations, pour différentes stations, échéances et horaires d'initialisation du modèle numérique. Les résultats ont montré que l'approche par classification directe offre de meilleures performances de classification, indépendamment de l'échéance de prévision et de la station. Néanmoins, les performances de classification semblent limitées. En effet, certaines classes ont fait l'objet d'un surapprentissage en raison de la difficulté de correction à moyenne échéance d'une variable comme la précipitation. Ce problème est similaire à celui rencontré en augmentant l'échéance de la prévision. Les modèles de classification étudiés offrent la possibilité d'analyser les covariables contribuant à la prédiction des classes. Une contribution intéressante relevée est celle des facteurs temporels retenus par les modèles dans la prédiction des classes liées aux vitesses de vent (contribution également remarquée dans la calibration des vitesses de vent). Cet aspect temporel pourrait être approfondi dans la modélisation des classes. Par exemple, dans un premier temps une chaîne de Markov cachée serait appliquée pour modéliser les transitions entre classes de vitesses de vent et ainsi approcher une dynamique temporelle

simplifiée. Cependant, l'interprétation physique des résultats de prédiction de classe des différents modèles étudiés et de l'erreur de prévision multivariée corrigée indirectement reste complexe à mener.

L'application de ces deux méthodes de classification a permis d'obtenir des prédictions meilleures que celles obtenues à partir de la classification issue des ensembles du modèle de prévision numériques. Cependant, le fait que les classes d'observations sont générées de manière très linéaire amène plusieurs problématiques. La première est la difficulté à produire des classes proportionnées de manière équivalente tout en respectant une interprétation météorologique. La seconde est le manque de prise en compte des réelles dépendances des variables météorologiques non linéaires, ainsi que des erreurs de distributions d'ensembles se trouvant dans les données. Les erreurs de distributions d'ensembles multivariés représentent toujours un challenge à corriger à partir de modèles de calibration ou de classification supervisée. Une hypothèse de travail est que les ensembles ont des erreurs de distributions différentes pour des régimes météorologiques différents. À partir de là, une solution pourrait être de relier les dépendances de variables météorologiques et les erreurs de distribution des ensembles de prévision en sous-groupes d'ensembles à l'aspect météorologique similaire. Les sous-groupes seraient formés de façon non supervisé à l'aide des ensembles de prévision. Les observations serviraient à la vérification et correction des erreurs de prévisions des sous-groupes formés.

Pour résumer :

- Une méthode originale de calibration multivariée est proposée dans laquelle l'espace des données est discrétisé de façon à transformer un problème de régression multi-sorties en un problème de classification à sortie unique.
- Une approche de classification directe à base de modèles de régression multinomiale Lasso et forêt aléatoire est comparée à une seconde approche de classification issue d'une calibration multivariée.
- Les approches sont testées sur des données réelles dans un contexte de planification d'évènements.
- Les modèles de classification directe ont montré de bons résultats de prédiction de classes.
- Les travaux de ce chapitre ont donné lieu à une publication dans les actes d'une conférence internationale :
  Jouan, G. Cuzol, A., Monbet, V., Monnier, G. (2019) Weather type prediction at medium range from ensemble forecasts. 9th International workshop on Climate Informatics, Oct 2019, Paris, France.

Chapitre 3

Modèles de mélange gaussien pour la classification non supervisée et la calibration d'ensembles de prévisions météorologiques

---

Les ensembles de prévision issus des modèles numériques peuvent contenir des erreurs de distributions de différents types, allant de la présence de multiples biais aux erreurs de dispersion. La correction de ces erreurs est l'objectif principal des méthodes de calibration d'ensembles, mais la compréhension de ces erreurs de prévision est également un axe de recherche essentiel. Par exemple, les travaux récents de Bröcker et Ben Bouallègue 2020 montrent que les jeux de données d'ensembles peuvent contenir de multiples erreurs aux caractéristiques différentes, et que ces erreurs sont importantes à identifier, car elles impactent les problèmes de calibration.

La figure 3.1 montre des exemples de ces erreurs. Sur le graphique de la première ligne, la série temporelle des observations de moyennes journalières de températures à Millau (France) est affichée en noir, et les boîtes à moustache décrivent la distribution des ensembles issus des prévisions à 3 jours. Les rectangles noirs focalisent l'attention sur des situations spécifiques associées à différents types d'erreurs. Les caractéristiques de ces erreurs sont mises en avant par l'outil de vérification des histogrammes de rangs (Anderson 1996 ; Talagrand, R. Vautard et Strauss 1997). Si l'observation est indissociable des réalisations (appelées aussi membres) de l'ensemble, l'histogramme de rangs correspond à la loi uniforme. Un histogramme ayant une forme en "L" ou en "L" inversé est typique d'une distribution d'ensemble biaisée. Sur la figure 3.1, les temps 12 à 15 sont associés à cette forme. Il est clairement visible que les boîtes à moustaches sont biaisées négativement. Les formes en $\cap$ (ou $\cup$) correspondent à des cas de sur (ou sous) dispersion des ensembles (Hamill et Colucci 1997). Des exemples de situations

de problèmes de dispersion sont montrés sur la figure 3.1 au début du mois pour la sur-dispersion et à partir du 24ème jour pour la sous-dispersion.

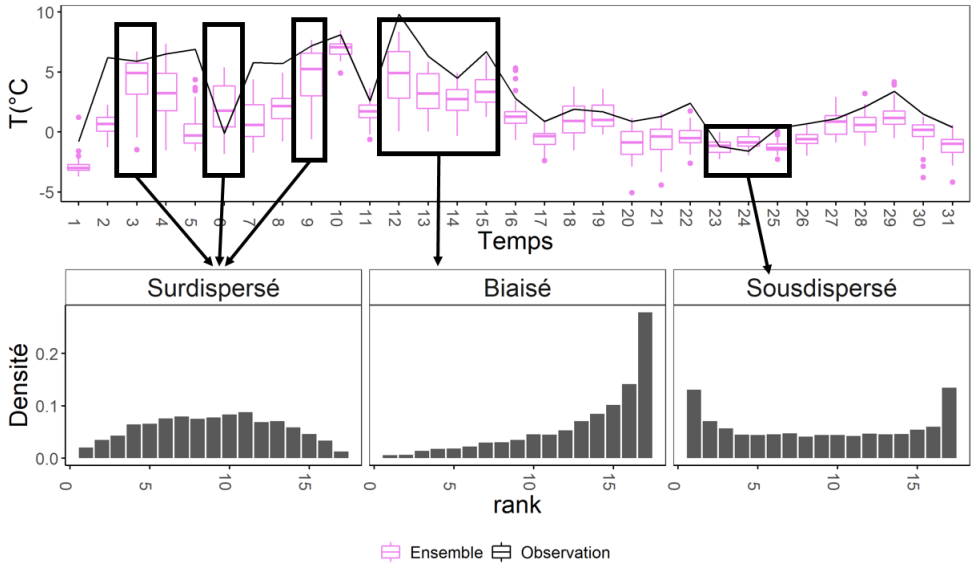


FIGURE 3.1 – Janvier 2015, 18H à Millau, ensemble de prévision CEPMMT d'horizon de prévision 3 jours et observations de températures. *Première ligne : série observée et boîtes à moustaches des ensembles de prévision ; deuxième ligne : histogrammes de rangs illustrant des situations typiques d'erreurs.*

En plus d'illustrer les types d'erreurs, la figure 3.1 suggère également un lien entre l'état de la variable météorologique d'intérêt et l'erreur de distribution des ensembles. Par exemple, les ensembles des temps $t = 12$ à 15 liés à un biais sont associés à des températures moyennes, alors que les situations sous-dispersées des temps t=23 à 25 correspondent à des températures plus faibles. Récemment, ALLEN, C. FERRO et KWASNIOK 2020 ; Sam ALLEN, C. A. FERRO et Frank KWASNIOK 2019 ont proposé une méthodologie de post-traitement des ensembles incorporants des informations de la circulation de l'atmosphère. Dans leur approche, les changements d'état à l'échelle synoptique de l'atmosphère sont inclus dans le modèle de calibration sous forme de régimes météorologiques. Contrairement à cette approche globale, BESSAC et al. 2016 a montré qu'il est intéressant de considérer des régimes locaux identifiés à l'aide de variables locales pour décrire leurs interactions.

Dans ce chapitre, l'objectif est de discriminer les ensembles de prévision en sous-groupes ayant des types de météorologie et des erreurs de prévision similaires, et d'utiliser

cette étape de classification pour améliorer la calibration. L'étape de classification non supervisée sera réalisée à l'aide de modèles de mélange adaptés aux données de type ensemble. Les modèles de mélange sont connus pour leur flexibilité, interprétabilité et facilité d'implémentation, trois critères essentiels dans les nouveaux challenges de la communauté de calibration (Vannitsem et al. 2021). Pour ce chapitre, le modèle de mélange sera appliqué dans le cadre de distributions gaussiennes donnant un modèle de mélange gaussien (GMM, 'Gaussian mixture model'), mais d'autres distributions peuvent être envisagées.

La généralisation du GMM présenté en 3.1 n'est pas évidente pour les données d'ensembles échangeables et nécessite différentes stratégies. La première contribution de ce chapitre est de proposer trois extensions du GMM pour apprendre des données d'ensemble de prévision. La section 3.1.2 introduit ces trois modèles et détaille leurs différentes caractéristiques. La première extension est d'ajuster le GMM sur des statistiques empiriques issues des ensembles. Les statistiques empiriques donnent une description synthétique de la distribution des ensembles pouvant conduire à une classification performante. Dans un besoin de comparaison, la seconde approche est de venir considérer les ensembles comme un "super échantillon". En d'autres mots, l'inférence du modèle est effectuée sur toute la base de données tout en abandonnant la structure de l'ensemble. Pour terminer, la dernière approche est basée sur un modèle considérant l'ensemble comme un vecteur de variables échangeables. Dans la section 3.1.3, les performances numériques des trois modèles et de leurs méthodes d'initialisation sont analysées au travers d'expérimentations variées.

En section 3.2, le modèle de mélange étendu pour les ensembles échangeables est testé sur des données d'ensemble du centre européen de prédiction météorologique à moyen terme (CEPMMT) avec comme objectifs l'interprétation et la correction des erreurs de prévision (calibration). Cette étape représente la seconde contribution de ce chapitre. Pour cela, un moyen de caractérisation automatique des types d'erreurs de calibration est nécessaire. La partie 3.2.1 introduit une méthode d'identification de ces types d'erreurs à l'aide de tests statistiques appliqués sur une variable normalisée et issue des histogrammes de rangs en reprenant des indications de Taillardat, Mestre et al. 2016. Ensuite, de multiples types d'erreurs de distributions d'ensembles sont simulés pour analyser les résultats de ces tests.

Dans la suite de cette section, le problème de calibration est alors traité en proposant une extension du modèle standard de calibration univariée NGR (introduit en 1.1.1.1) : un modèle de calibration est appliqué sur chaque classe issue du mélange, indépendamment

des autres.

Pour terminer, la section 3.3 présente les conclusions et les futures évolutions possibles.

3.1 Modèles de mélange gaussien pour la classification non supervisée d'ensembles

Le modèle de mélange gaussien présente un fort intérêt d'application dans un but de classification non supervisée par sa flexibilité d'adaptation et ses possibilités d'évolution. Dans un premier temps, le modèle de mélange gaussien classique est introduit dans le cadre classique. Ensuite, trois extensions du modèle de mélange gaussien sont proposées et discutées pour s'adapter au cas particulier des données d'ensemble. Pour terminer, les extensions proposées sont analysées sur des données simulées. Plus particulièrement, la sensibilité d'estimation des trois modèles est étudiée suivant différents types de méthodes d'initialisation et de paramétrage de données simulées.

3.1.1 Mélange gaussien

Un modèle de mélange gaussien (GMM) décrit la distribution d'un couple de variables composé d'une variable de classe latente et d'une variable continue observée (McLachlan et Peel 2004). La variable de classe Z est distribuée selon une distribution multinomiale $\mathcal{M}(\pi_1, \cdots, \pi_K)$ sur $\{1, \cdots, K\}$ et, pour tout $k \in \{1, \cdots, K\}$ la variable continue $X \in \mathbb{R}^d$ suit une distribution gaussienne sachant $Z = k$ avec les paramètres μ_k et Σ_k. La fonction de densité de probabilité (pdf) de X est donnée par :

$$f(x; \Psi) = \sum_{k=1}^{K} \pi_k \varphi(x; \mu_k, \Sigma_k) \tag{3.1}$$

où x représente une réalisation de X, $\Psi = (\pi_1, \cdots, \pi_K, \mu_1, \cdots, \mu_K, \Sigma_1, \cdots, \Sigma_K)$ l'ensemble des paramètres du modèle, et $\varphi(x; \mu, \Sigma)$ est la densité gaussienne de la sous-population k prenant l'ensemble des paramètres μ_k, Σ_k :

$$\varphi(x; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{\frac{d}{2}} \sqrt{det(\Sigma_k)}} e^{-\frac{1}{2}(x-\mu_k)^\intercal \Sigma_k^{-1}(x-\mu_k)} \tag{3.2}$$

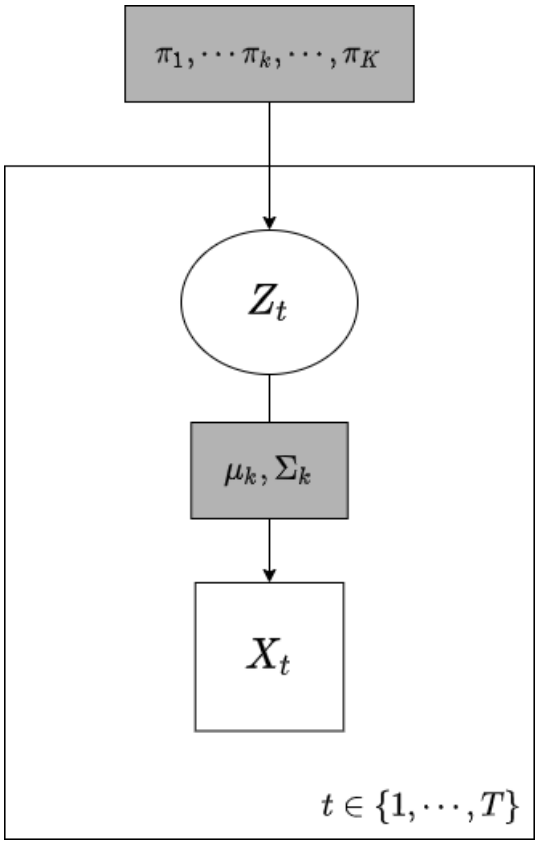


FIGURE 3.2 – Graphique acyclique orienté d'un modèle de mélange gaussien.
Le cercle contient la variable non observée et le carré représente la variable observée ; les paramètres du modèle sont affichés en gris.

Suivant un échantillon $\mathcal{X} = \{x_1, \cdots, x_n\}$ de n réalisations indépendantes de la variable aléatoire X, les paramètres inconnus $\Psi = (\pi_1, \cdots, \pi_K, \mu_1, \cdots, \mu_K, \Sigma_1, \cdots, \Sigma_K)$ sont estimés en maximisant la log-vraisemblance :

$$\log \mathcal{L}(\mathcal{X}; \Psi) = \sum_{t=1}^{T} \log \sum_{k=1}^{K} \pi_k \varphi(x_t; \mu_k, \Sigma_k). \tag{3.3}$$

Il existe deux approches principales dans la littérature statistique pour maximiser numériquement la vraisemblance dans les modèles avec des variables latentes : la montée du gradient ('Gradient Ascent') et l'algorithme d'Espérance-Maximisation (EM) (DEMPSTER, LAIRD et RUBIN 1977). Cependant, la montée de gradient peut être numériquement instable et, par conséquent, l'approche EM est généralement privilégiée lorsqu'on considère des modèles avec des composantes latentes. À chaque itération, l'algorithme Espérance-

Maximisation alterne deux étapes, l'étape E et l'étape M. À l'itération $[i]$ de l'algorithme EM, l'étape E calcule les probabilités *a posteriori* $\gamma_{tk}^{[i]}$ d'appartenance à un sous-groupe pour tous les individus $t \in \{1, \cdots, T\}$, sachant les valeurs actuelles des paramètres $\Psi^{[i-1]}$. Les probabilités *a posteriori* sont calculées grâce à :

$$\gamma_{tk}^{[i]} = \mathbb{P}(Z = k | x_t, \Psi^{[i-1]}) = \frac{\pi_k^{[i-1]} \varphi(x_t; \mu_k^{[i-1]}, \Sigma_k^{[i-1]})}{\sum_{\ell=1}^{K} \pi_\ell^{[i-1]} \varphi(x_t; \mu_\ell^{[i-1]}, \Sigma_\ell^{[i-1]})} \tag{3.4}$$

Ensuite, dans l'étape M, l'espérance conditionnelle de la log-vraisemblance étant donné la valeur actuelle des paramètres est maximisée. Cela conduit aux mises à jour suivantes des paramètres à l'itération $[i]$, pour tous les $k = 1, ..., K$:

$$\pi_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]}}{T} \tag{3.5}$$

$$\mu_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} x_t}{\sum_{t=1}^{T} \gamma_{tk}^{[i]}} \tag{3.6}$$

$$\Sigma_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} (x_t - \mu_k^{[i]})(x_t - \mu_k^{[i]})^\intercal}{\sum_{t=1}^{T} \gamma_{tk}^{[i]}} \tag{3.7}$$

L'annexe C.1 donne des descriptions supplémentaires sur l'étape E et l'étape M permettant d'aboutir aux estimateurs ci-dessus. Le cheminement introduit par l'algorithme EM pour l'approximation des paramètres du mélange gaussien est résumé dans l'algorithme 1.

Algorithm 1: Espérance-Maximisation (EM) pour le modèle de mélange gaussien

$\mathcal{X} = \{x_t\}_{1 \leq t \leq T}$ un jeu de données;
Initialisation $\Psi^{[0]} = \{\pi_k^{[0]}, \mu_k^{[0]}, \Sigma_k^{[0]}\}$, K, tol, $i = 1$, I_{max};
while $\epsilon > tol$ ***or*** $i > I_{max}$ **do**
  Etape-E Evaluer $\gamma_{tk}^{[i]} = \frac{\pi_k^{[i-1]} \varphi(x_t; \mu_k^{[i-1]}, \Sigma_k^{[i-1]})}{\sum_{k=1}^{K} \pi_k^{[i-1]} \varphi(x_t; \mu_k^{[i-1]}, \Sigma_k^{[i-1]})}$;
  Etape-M Mettre à jour $\Psi^{[i]}$ à travers
  $\pi_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]}}{T}$,
  $\mu_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} x_t}{\sum_{t=1}^{T} \gamma_{tk}^{[i]}}$,
  $\Sigma_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} (x_t - \mu_k^{[i]})(x_t - \mu_k^{[i]})^\intercal}{\sum_{t=1}^{T} \gamma_{tk}^{[i]}}$;
  Condition d'arrêt Mettre à jour ϵ et i
  $\epsilon = \log(\mathcal{L}(\Psi^{[i]})) - \log(\mathcal{L}(\Psi^{[i-1]}))$;
  $i = i + 1$;
end

Avant de passer à la description des extensions du GMM, il est important de définir une règle de classification pour la prédiction de sous-groupes pour chaque individu x_t. Pour cela, une règle reposant sur le maximum *a posteriori* (MAP) est utilisée avec les probabilités *a posteriori* γ_{tk} (3.4).

3.1.2 Modèles de mélanges pour ensembles de prévision

Dans les applications météorologiques, les modèles de prévision numérique du temps fournissent des ensembles au lieu d'échantillons standards. Plus précisément, chaque individu t est un ensemble de M réalisations $X_t^* = \{x_{t1}, \cdots, x_{tM}\}$ de la variable X. Ces réalisations sont appelées *membres*. Dans la suite, différentes solutions sont explorées pour adapter l'inférence du GMM aux observations d'ensemble. La première est inspirée d'une approche habituelle dans le traitement des ensembles météorologiques, qui consiste à travailler avec certaines statistiques empiriques de l'ensemble au lieu de l'ensemble entier. La deuxième est une application directe de la méthode d'inférence régulière où l'ensemble est considéré comme un super échantillon de taille $T \times M$. Dans la troisième, les membres sont considérés comme des réalisations d'un vecteur de variables échangeables $\{X_1, \cdots, X_M\}$ comme dans Diaconis et Freedman 1980 ; Courbariaux et al. 2019.

3.1.2.1 Statistiques empiriques d'ensembles

Afin de se délier des problématiques soulevées par les données d'ensembles échangeables, une méthode empruntée à la littérature de calibration (Gneiting, Raftery et al. 2005, Daniel S Wilks 2018) consiste à décrire l'ensemble échangeable par l'information des statistiques empiriques estimées sur ces membres.

Le vecteur S contenant les statistiques empiriques de l'ensemble X^* est défini. Le modèle de mélange gaussien est ensuite appliqué sur le couple de variables (Z, S) avec Z la variable de classe latente. Si S est composé de moments et M assez large, alors le théorème central limite assure que la distribution de S converge vers une loi normale, et ce pour chaque classe. Dans le cas particulier où S est composé de la moyenne empirique, la distribution de S suivant $Z = k$ est comparable à une loi normale $\mathcal{N}(\mathfrak{m}_k = \mu_k, \mathfrak{S}_k = \frac{\Sigma_k}{M})$. À l'aide de ces relations, les paramètres de la distribution de X peuvent être retrouvés avec les estimateurs de la moyenne empirique.

Dès lors, les étapes de l'algorithme EM sont appliquées sur le couple (Z, S). La probabilité *a posteriori* s'exprime de la façon suivante :

$$\gamma_{tk}^{[i]} = \mathbb{P}(Z = k|S_t, \Psi^{[i-1]}) = \frac{\pi_k^{[i-1]}\varphi(S_t; \mathfrak{m}_k^{[i-1]}, \mathfrak{S}_k^{[i-1]})}{\sum_{\ell=1}^{K} \pi_\ell^{[i-1]}\varphi(S_t; \mathfrak{m}_\ell^{[i-1]}, \mathfrak{S}_\ell^{[i-1]})}. \tag{3.8}$$

La probabilité *a posteriori* obtenue est une information importante pour discriminer les classes suivant les individus en utilisant la règle d'hétéroscédasticité de l'analyse discriminante de Fisher. Cependant, il faut garder à l'esprit que la précision de cette règle évolue suivant le rapport entre le signal et le bruit. Ensuite, si la moyenne de l'échantillon d'un ensemble est considérée au lieu d'un seul membre, on s'attend à ce que les sous-groupes soient mieux identifiés, car la variance de la moyenne de l'échantillon S dans chaque classe est plus petite que celle de X alors que les deux ont la même moyenne.

Les estimateurs, déduit de l'étape M, s'écrivent :

$$\pi_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]}}{T} \tag{3.9}$$

$$\mathfrak{m}_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} S_t}{\sum_{i=1}^{T} \gamma_{tk}^{[i]}} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} \sum_{m=1}^{M} x_{tm}}{M \sum_{t=1}^{T} \gamma_{tk}^{[i]}} \tag{3.10}$$

$$\mathfrak{S}_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} (S_t - \mathfrak{m}_k^{[i]})(S_t - \mathfrak{m}_k^{[i]})^\intercal}{\sum_{t=1}^{T} \gamma_{tk}^{[i]}} \tag{3.11}$$

L'algorithme 2 résume l'algorithme EM du modèle avec statistiques empiriques.

Algorithm 2: EM du modèle de mélange gaussien appliqué aux statistiques empiriques d'ensemble.

$\mathcal{X} = \{S_t = (\overline{x}_{t1}, \cdots, \overline{x}_{td})\}_{1 \leq t \leq T}$ un jeu de données de moyennes et variances empiriques de l'ensemble X_t^*;

Initialisation $\Psi^{[0]} = \{\pi_k^{[0]}, \mathfrak{m}_k^{[0]}, \mathfrak{S}_k^{[0]}\}$, K, tol, $i = 1$, I_{max};

while $\epsilon > tol$ ***or*** $i > I_{max}$ **do**

Etape-E Evaluer $\gamma_{tk}^{[i]} = \mathbb{P}(Z = k | S, \Psi^{[i-1]}) = \frac{\pi_k^{[i-1]} \varphi(S_t; \mathfrak{m}_k^{[i-1]}, \mathfrak{S}_k^{[i-1]})}{\sum_{\ell=1}^{K} \pi_\ell^{[i-1]} \varphi(S_t; \mathfrak{m}_\ell^{[i-1]}, \mathfrak{S}_\ell^{[i-1]})}$;

Etape-M Mettre à jour $\Psi^{[i]}$ à travers

$\pi_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]}}{T}$,

$\mathfrak{m}_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} S_t}{\sum_{i=1}^{T} \gamma_{tk}^{[i]}} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} \sum_{m=1}^{M} x_{tm}}{M \sum_{t=1}^{T} \gamma_{tk}^{[i]}}$,

$\mathfrak{S}_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} (S_t - \mathfrak{m}_k^{[i]})(s_t - \mathfrak{m}_k^{[i]})^\intercal}{\sum_{t=1}^{T} \gamma_{tk}^{[i]}}$;

Condition d'arrêt Mettre à jour ϵ et i

$\epsilon = \log(\mathcal{L}(\Psi^{[i]})) - \log(\mathcal{L}(\Psi^{[i-1]}))$;

$i = i + 1$;

end

Cependant, focaliser le modèle sur un seul moment peut être insuffisant pour obtenir une description précise de la distribution ciblée de X. Pour illustrer ce propos, dans le cas où seulement la moyenne empirique est utilisée, l'estimateur du maximum de vraisemblance échouera à approcher la variance conditionnelle de X suivant Z dans certaines situations. Typiquement, lorsque la taille de l'ensemble est insuffisante, une variance trop élevée autour de la moyenne empirique peut aboutir à des écarts d'estimation importants. Dans la section de simulation 3.1.3, le modèle basé sur les statistiques empiriques est ajusté sur $S_t = (\overline{x}_{t1}, \cdots, \overline{x}_{td}), (s_{t1}^2, \cdots, s_{td}^2)$ où $\overline{x}_{t\ell} = \frac{1}{T} \sum_{t=1}^{T} x_t^\ell$ est la moyenne empirique et $s_{t\ell}^2 = \frac{1}{T} \sum_{t=1}^{T} (x_{t\ell} - \overline{x}_{t\ell})^2$ la variance empirique associée à la dimension $\ell \in \{1, \cdots, d\}$ de l'ensemble X_t^* généré suivant la variable X de dimension d.

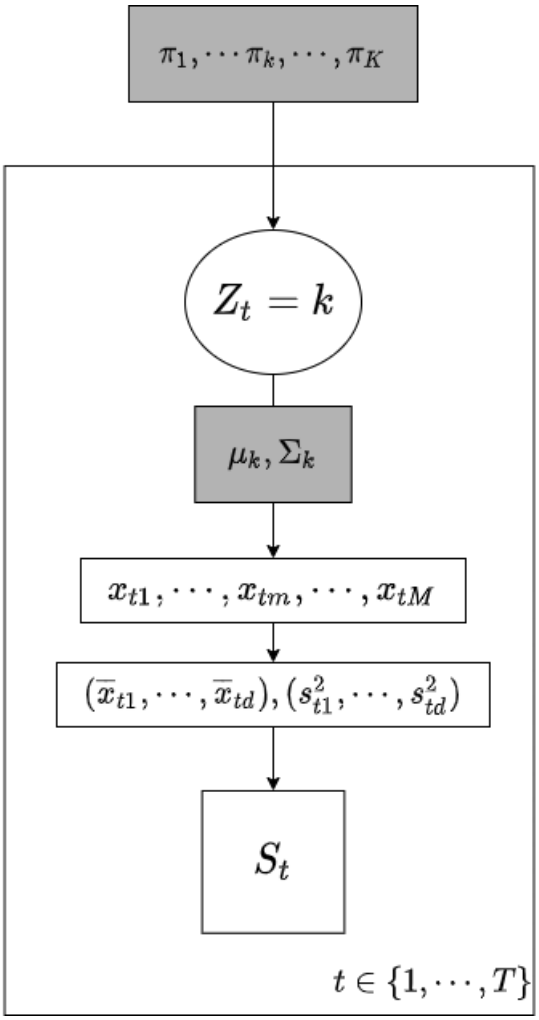


FIGURE 3.3 – Graphique acyclique orienté pour un mélange gaussien appliqué aux statistiques empiriques de données d'ensembles.
Le cercle contient la variable non observée et le carré représente la variable observée ; les paramètres du modèle sont affichés en gris.

Notez que la distribution de la variance empirique d'un échantillon de taille M est un Chi-deux avec $M - 1$ degrés de liberté et qu'elle est bien approchée par une distribution gaussienne pour des valeurs relativement faibles de M (typiquement supérieures à 20). De plus, lorsque la moyenne et la variance sont considérées comme des variables observées dans le GMM, l'algorithme EM renvoie directement $\mathfrak{m}_k$ qui est un vecteur à deux composantes composé d'une estimation de μ_k et d'une estimation de Σ_k dans un cas univarié. Néanmoins, dans un cas multivarié, la matrice de covariance peut être légèrement plus délicate à obtenir. Dans cette situation, la sous-matrice $d \times d$ des d premières dimensions

de l'estimateur $\mathfrak{S}_k$ de l'EM, multiplié par M, est utilisée comme estimateur de Σ_k. Cette sous-matrice contient l'estimation de la matrice de covariance de la variable gaussienne de paramètre μ_k et $\frac{\Sigma_k}{M}$ liée à la moyenne empirique. La figure 3.3 fait état du graphique acyclique décrivant les liens entre les variables et les paramètres du modèle de mélange proposé.

Dans la section 3.1.3, les statistiques empiriques conduiront à une bonne prédiction des sous-groupes et à de bonnes estimations lorsque la taille de l'ensemble est suffisamment grande. Cependant, pour un petit ensemble, les statistiques empiriques peuvent être mal estimées et les résultats obtenus avec le GMM ne sont pas aussi bons. Dans les prochaines sous-sections, une utilisation plus directe de l'ensemble est explorée.

3.1.2.2 Super échantillon

La façon la plus simple d'adapter l'algorithme EM aux données d'ensemble et de tirer parti du grand nombre d'observations est de considérer l'ensemble comme un "super échantillon" $\{x_{11}, \cdots, x_{TM}\}$ de $T \times M$ réalisations indépendantes de X. L'algorithme EM décrit ci-dessus peut alors être appliqué à cet échantillon. L'étape E et l'étape M sont calculées comme suit :

$$\gamma_{tmk}^{[i]} = \frac{\pi_k^{[i-1]} \varphi(x_{tm}; \mu_k^{[i-1]}, \Sigma_k^{[i-1]})}{\sum_{\ell=1}^{K} \pi_\ell^{[i-1]} \varphi(x_{tm}; \mu_\ell^{[i-1]}, \Sigma_\ell^{[i-1]})} \tag{3.12}$$

Ensuite, l'étape M renvoie l'expression suivante des estimateurs :

$$\pi_k^{[i]} = \frac{\sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{tmk}^{[i]}}{nM} \tag{3.13}$$

$$\mu_k^{[i]} = \frac{\sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{tmk}^{[i]} x_{tm}}{\sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{tmk}^{[i]}} \tag{3.14}$$

$$\Sigma_k^{[i]} = \frac{\sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{tmk}^{[i]} (x_{tm} - \mu_k^{[i]})(x_{tm} - \mu_k^{[i]})^\intercal}{\sum_{t=1}^{T} \sum_{m=1}^{M} \gamma_{tmk}^{[i]}} \tag{3.15}$$

Comparé au cas classique où le modèle ne dispose que d'une réalisation, le modèle basé sur des ensembles ou un "super échantillon" s'ajuste sur un jeu de données beaucoup plus grand donnant une variance d'estimateur plus faible. La figure 3.4 représente le graphique du modèle de mélange basé sur un "super échantillon". Ce graphique est très proche de celui du modèle de base 3.2 étant donné que l'apport principal est situé sur le réarrangement des données d'ensemble pour l'algorithme EM.

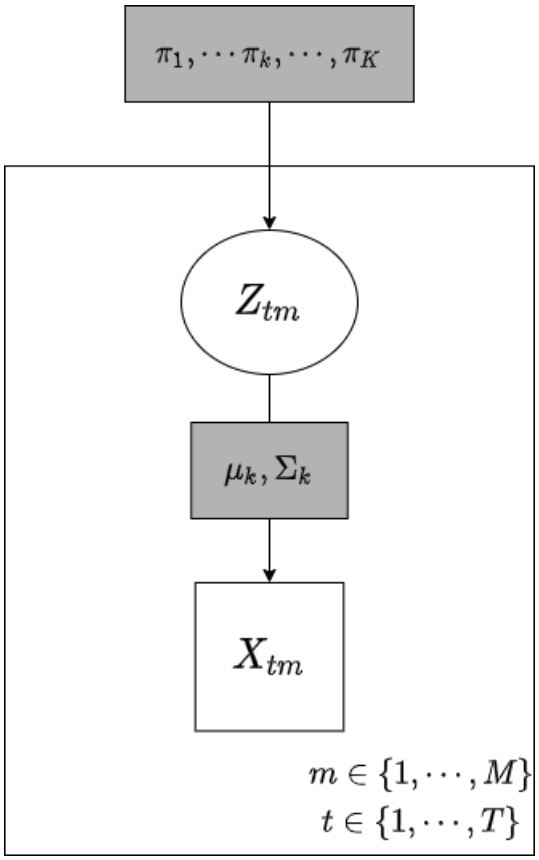


FIGURE 3.4 – Graphique acyclique orienté pour un mélange gaussien appliqué à chaque réalisation de données d'ensembles échangeables.
Le cercle contient la variable non observée et le carré représente la variable observée ; les paramètres du modèle sont affichés en gris.

La probabilité *a posteriori* (3.12) assigne une classe z_{tm} pour chaque membre x_{tm}, pouvant aboutir à un ensemble de classes différentes pour l'ensemble X_t^*. Autrement dit, pour un temps t, il est probable que l'ensemble X_t^* se voit attribuer différentes classes météorologiques par l'ensemble des classes z_{tm}. Or, pour de courtes échéances de prévision, avoir ce type de situation est peu réalisable. En effet, un ensemble issu du modèle de prévision d'ensemble avec des conditions initiales légèrement perturbées décrit une météorologie similaire à celle du modèle déterministe (BUIZZA, MILLEER et Tim N PALMER 1999, Daniel S WILKS 2005). Cependant, pour de longues échéances, il est observé que les membres peuvent évoluer suivant différents états de l'atmosphère. Dans ce cas, attribuer différentes classes au sein d'un même ensemble pourrait avoir du sens rejoignant les approches de calibration d'ensemble à l'aide de "multi-model outputs" de FRALEY,

Raftery et Gneiting 2010. Dans cette thèse, pour surpasser ce problème, une méthode de vote à majorité sur les z_{tm}, $\forall m \in \{1, \cdots, M\}$ est appliquée afin de prédire une unique classe associée à l'ensemble X_t^*. L'algorithme 3 résume l'algorithme EM pour le modèle "super échantillon".

Algorithm 3: EM du modèle de mélange gaussien pour un "super échantillon".

$\mathcal{X} = \{x_{tm}\}_{(1\leq t\leq T)\times(1\leq m\leq M)}$ un jeu de données de réalisations d'ensemble $X_t^* = (x_{t1}, \cdots, x_{tM})$;

Initialisation $\Psi^{[0]} = \{\pi_k^{[0]}, \mu_k^{[0]}, \Sigma_k^{[0]}\}$, K, tol, $i = 1$, I_{max};

while $\epsilon > tol$ ***or*** $i > I_{max}$ **do**

Etape-E Evaluer $\gamma_{tmk}^{[i]} = \frac{\pi_k^{[i-1]}\varphi(x_{tm};\mu_k^{[i-1]},\Sigma_k^{[i-1]})}{\sum_{\ell=1}^{K}\pi_\ell^{[i-1]}\varphi(x_{tm};\mu_\ell^{[i-1]},\Sigma_\ell^{[i-1]})}$;

Etape-M Mettre à jour $\Psi^{[i]}$ à travers

$\pi_k^{[i]} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{tmk}^{[i]}}{nM}$,

$\mu_k^{[i]} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{tmk}^{[i]}x_{tm}}{\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{tmk}^{[i]}}$,

$\Sigma_k^{[i]} = \frac{\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{tmk}^{[i]}(x_{tm}-\mu_k^{[i]})(x_{tm}-\mu_k^{[i]})^\intercal}{\sum_{t=1}^{T}\sum_{m=1}^{M}\gamma_{tmk}^{[i]}}$;

Condition d'arrêt Mettre à jour ϵ et i

$\epsilon = \log(\mathcal{L}(\Psi^{[i]})) - \log(\mathcal{L}(\Psi^{[i-1]}))$;

$i = i + 1$;

end

Une adaptation du GMM pour remédier à ce problème serait de forcer l'appartenance des membres à une même composante gaussienne k dans la vraisemblance complète de l'algorithme EM utilisée dans la fonction Q (C.2) de l'étape E. La probabilité *a posteriori* déduite de cette approche aurait la forme suivante :

$$\begin{aligned}\gamma_{tk}^{[i]} &= p(Z = z_t|(x_{t1}, \cdots, x_{tM}), \Psi^{[i-1]}) \\ &= \prod_{m=1}^{M}\gamma_{tmk}^{[i]} \\ &= \prod_{m=1}^{M}\left[\frac{\pi_k^{[i-1]}\varphi(x_{tm};\mu_k^{[i-1]},\Sigma_k^{[i-1]})}{\sum_{\ell=1}^{K}\pi_\ell^{[i-1]}\varphi(x_{tm};\mu_\ell^{[i-1]},\Sigma_\ell^{[i-1]})}\right]\end{aligned} \tag{3.16}$$

Cette probabilité *a posteriori* présente un avantage permettant d'ajuster directement les estimateurs en considérant M réalisations reliées à une des k classes latentes. Cependant, il s'avère que cette probabilité déroge à la contrainte $\sum_{k=1}^{K}\gamma_{tk}^{[i]} = 1$ dès lors que $M > 1$. En effet, si une réalisation m de la $k^{ème}$ composante se retrouve dans une queue de distribution

et proche de la distribution d'une autre composante $k' \neq k$, alors la probabilité $\gamma_{tmk}^{[i]}$ sera proche de zéro et entraînera le produit des M probabilités à tendre vers zéro (un exemple illustrant ce point est montré dans l'annexe C.2).

Dans la section 3.1.2.3, une autre adaptation du GMM classique est proposée pour forcer tous les membres d'un individu t à appartenir au même groupe.

3.1.2.3 Variables gaussiennes écheangeables

Une alternative au modèle de "super échantillon" précédent est de considérer l'ensemble comme un vecteur de M variables échangeables $(X_1, \cdots, X_m, \cdots, X_M)$ où X_m est le $m^{\text{ème}}$ membre supposé indépendant de X_ℓ si $m \neq \ell$ est échangeable. Pour reprendre Diaconis et Freedman 1980, un vecteur de variables est dit échangeable dès lors que sa distribution jointe est invariante face aux permutations des variables. Le cadre du modèle de mélange gaussien est appliqué à $(Z, X_1, \cdots, X_M)$ avec Z une variable discrète de classe latente. La distribution jointe du vecteur $(X_1, \cdots, X_M)$ est donnée par :

$$f_{X_1,...,X_M}(x_1, ..., x_M; \Psi) = \sum_{k=1}^{K} \pi_k \prod_{m=1}^{M} \varphi(x_m; \mu_k, \Sigma_k) \tag{3.17}$$

Par application de l'algorithme EM, la probabilité *a posteriori* s'écrit :

$$\gamma_{tk}^{[i]} = \mathbb{P}(Z = k | X_1, \cdots, X_M, \Psi^{[i-1]}) = \frac{\pi_k^{[i-1]} \prod_{m=1}^{M} \varphi(x_{tm}; \mu_k^{[i-1]}, \Sigma_k^{[i-1]})}{\sum_{\ell=1}^{K} \pi_\ell^{[i-1]} \prod_{m=1}^{M} \varphi(x_{tm}; \mu_\ell^{[i-1]}, \Sigma_\ell^{[i-1]})}. \tag{3.18}$$

La figure 3.5 donne une représentation graphique du lien entre la variable latente, les variables gaussiennes et les paramètres du modèle (3.18).

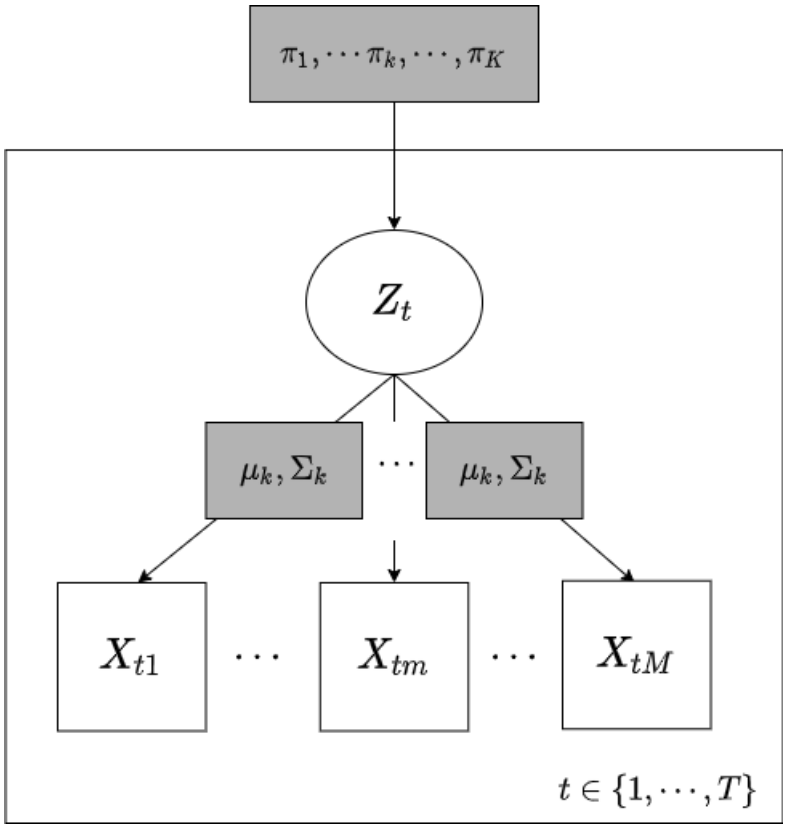


FIGURE 3.5 – Graphique acyclique orienté pour un mélange d'ensemble échangeable gaussien.
Le cercle contient la variable non observée et le carré représente la variable observée ; les paramètres du modèle sont affichés en gris.

Cette solution suppose l'appartenance des M réalisations d'ensemble à une même composante gaussienne k dans la définition du modèle (3.17). L'avantage de ce modèle est qu'il permet d'obtenir des probabilités respectant la contrainte $\sum_{k=1}^{K} \gamma_{tk}^{[i]} = 1$ grâce à la normalisation du produit des M marginales densités de (3.18). Ce même produit affecte la forme des probabilités obtenues. Pour être plus précis, les frontières entre classes s'affinent suivant le nombre de variables échangeables M. La conséquence directe de cet apport est que les ensembles proches du centre des classes auront un poids plus fort dans l'étape E de l'algorithme EM. De ce fait, la convergence de l'EM se retrouve améliorée.

Pour tout $k = 1, ..., K$, l'étape M conduit aux expressions d'estimateurs suivantes :

$$\pi_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]}}{T} \tag{3.19}$$

$$\mu_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} \sum_{m=1}^{M} x_{tm}}{M \sum_{t=1}^{T} \gamma_{tk}^{[i]}} \tag{3.20}$$

$$\Sigma_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} \sum_{m=1}^{M} (x_{tm} - \mu_k^{[i]})(x_{tm} - \mu_k^{[i]})^\intercal}{M \sum_{t=1}^{T} \gamma_{tk}^{[i]}} \tag{3.21}$$

Le modèle (3.17) supposé également l'indépendance entre les membres de l'ensemble. En pratique, les ensembles sont générés à l'aide de perturbations indépendantes des conditions initiales et paramètres du modèle de prévision numérique. Cependant, la physique développée au sein du modèle fait que malgré l'indépendance des perturbations générées, les membres obtenus sont corrélés entre eux. L'indépendance introduite dans le modèle (3.17) est donc une propriété forte et discutable qui pourrait être relâchée. Pour chaque k, le produit $\prod_{m=1}^{M} \varphi(x_m; \mu_k, \Sigma_k)$ serait remplacé par une densité gaussienne multivariée ayant une moyenne identique entre les membres $\mu_k e_M$, où e_M est un vecteur entièrement composé de 1. La covariance serait une matrice par blocs avec des matrices Σ_k sur la diagonale et une extra-matrice diagonale répétée sur les autres blocs. Cette amélioration ne sera pas incluse dans cette thèse. L'algorithme 4 résume l'algorithme EM du modèle de mélange gaussien au vecteur de variables échangeables.

Algorithm 4: EM du modèle de mélange gaussien pour un vecteur de variables échangeables.

$\mathcal{X} = \{X_t^* = (x_{t1}, \cdots, x_{tM})\}_{1 \leq t \leq T}$ un jeu de données d'ensemble de M réalisations;
Initialisation $\Psi^{[0]} = \{\pi_k^{[0]}, \mu_k^{[0]}, \Sigma_k^{[0]}\}$, K, tol, $i = 1$, I_{max};
while $\epsilon > tol$ ***or*** $i > I_{max}$ **do**
  Etape-E Evaluer
  $\gamma_{tk}^{[i]} = \mathbb{P}(Z = k | x_1, \cdots, x_M, \Psi^{[i-1]}) = \frac{\pi_k^{[i-1]} \prod_{m=1}^{M} \varphi(x_{tm}; \mu_k^{[i-1]}, \Sigma_k^{[i-1]})}{\sum_{\ell=1}^{K} \pi_\ell^{[i-1]} \prod_{m=1}^{M} \varphi(x_{tm}; \mu_\ell^{[i-1]}, \Sigma_\ell^{[i-1]})}$;
  Etape-M Mettre à jour $\Psi^{[i]}$ à travers
  $\pi_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]}}{T}$,
  $\mu_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} \sum_{m=1}^{M} x_{tm}}{M \sum_{t=1}^{T} \gamma_{tk}^{[i]}}$,
  $\Sigma_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} \sum_{m=1}^{M} (x_{tm} - \mu_k^{[i]})(x_{tm} - \mu_k^{[i]})^\intercal}{M \sum_{t=1}^{T} \gamma_{tk}^{[i]}}$;
  Condition d'arrêt Mettre à jour ϵ et i
  $\epsilon = \log(\mathcal{L}(\Psi^{[i]})) - \log(\mathcal{L}(\Psi^{[i-1]}))$;
  $i = i + 1$;
end

Pour la suite, les trois approches sont nommées respectivement dans l'ordre **Empirical statistics**, **Super sample** et **Exchangeable variables** et sont étudiées au travers de différentes expérimentations. Dans chaque expérience, des ensembles suivant un mélange gaussien aux paramètres définis sont simulés et les estimations des modèles introduits sont comparées aux vrais paramètres du mélange créé.

3.1.3 Validation par étude de simulation

Les performances des procédures proposées sont évaluées par une étude de simulation. En particulier, l'étude se concentre sur les propriétés de sélection des modèles et la capacité de chaque méthode à estimer correctement les paramètres.

3.1.3.1 Définition des expérimentations

Cadre expérimental. Deux cas avec $K = 4$ classes sont définis. Le premier cas appelé "Régulier" prend des paramètres d'exemples classiques tirés de la littérature des mélanges gaussiens (Biernacki, Celeux et Govaert 2003, Baudry et Celeux 2015). Les paramètres pour ce cas sont représentés ligne 1 et 3 du table (3.1). Le cas "Régulier" dispose de moyennes différentes avec des variances relativement faibles générant des classes

relativement bien séparées avec quelques chevauchements entre les individus de la composante 1 et 2. Dans le second cas nommé "Difficile", un fort chevauchement est simulé entre classes comportant des moyennes quasi identiques pour des variances différentes. Les situations introduites dans le cas "Difficile" apparaissent de manière récurrente dans les problématiques de calibration lorsque des ensembles présentent une erreur de dispersion (sur ou sous-dispersé) et recouvrent d'autres ensembles ne présentant pas les mêmes caractéristiques de distribution. Les paramètres de ce cas sont affichés ligne 2 et 4 du tableau (3.1).

		Paramètres des classes			
d	Cas	μ_1, σ_1	μ_2 , σ_2	μ_3 , σ_3	μ_4 , σ_4
1	Régulier	0, 1	2, 0.3	7, 1	10, 1
	Difficile	2, 2	3, 0.3	3, 1	4, 4
3	Régulier	(0,0,0), 1	(2,0,0), 0.3	(7,7,7), 1	(10,10,10), 1
	Difficile	(2,2,2), 2	(3,3,3), 0.3	(3,3,3), 1	(4,4,4), 4

TABLE 3.1 – Paramètre des composantes du mélange gaussien de chaque cas généré.

Chaque cas est généré suivant un contexte univarié (d=1) et multivarié (d=3). Pour le contexte trivarié, la matrice de covariance est définie de la façon suivante :

$$\Sigma_k = \begin{pmatrix} \sigma_k^2 & 0.5\sigma_k^{3/2} & 0.5\sigma_k^{3/2} \\ 0.5\sigma_k^{3/2} & \sigma_k^2 & 0.5\sigma_k^{3/2} \\ 0.5\sigma_k^{3/2} & 0.5\sigma_k^{3/2} & \sigma_k^2 \end{pmatrix}.$$

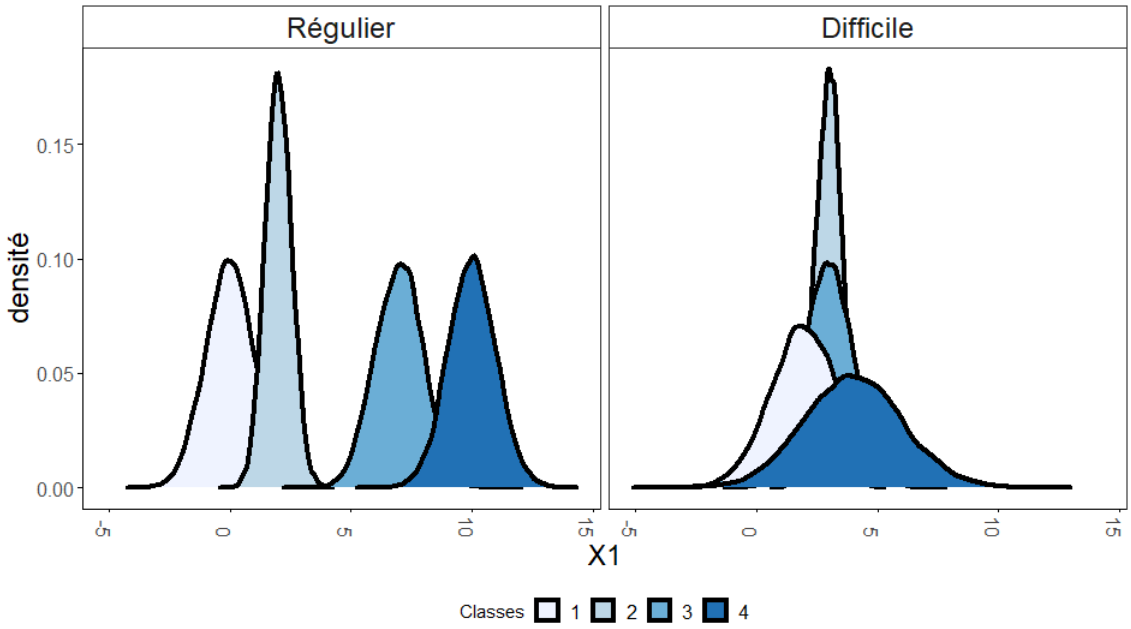


FIGURE 3.6 – Densité conditionnelle simulée pour le cas 'Régulier' (panneau gauche) et 'Difficile' (panneau droit) en dimension $d = 1$.

Une représentation des distributions ainsi générées est disponible dans un cas univarié sur la figure 3.6 et multivarié pour la figure C.2.

La vraisemblance du GMM n'est pas concave et présente généralement plusieurs maxima locaux. Il est donc important de procéder à une initialisation minutieuse de l'algorithme EM. Plusieurs approches ont été proposées dans la littérature. Certaines d'entre elles sont basées sur les k-means (ARTHUR et VASSILVITSKII 2006) ou sur des modèles de mélange comme par exemple l'initialisation stochastique EM (MCLACHLAN et KRISHNAN 2007) et le Small EM (BIERNACKI, CELEUX et GOVAERT 2003 ; BAUDRY et CELEUX 2015). Ces méthodes ont été testées sur les données simulées et comparées à une initialisation basée sur une classification hiérarchique (HC, FRALEY 1998). Plus de précisions sur les méthodes d'initialisation et leurs résultats peuvent être retrouvées dans l'annexe C.3.3. La méthode HC a été choisie car elle conduit à de meilleurs résultats de classification. Plus précisément, l'initialisation HC est effectuée sur 50 sous-échantillons aléatoires. Ensuite, l'ensemble des paramètres estimés maximisant la log-vraisemblance est sélectionné comme premier point de l'algorithme EM final. Le modèle basé sur les statistiques empiriques et celui basé sur le super échantillon sont initialisés indépendam-

ment. Le modèle à variables échangeables commence à partir de l'initialisation du modèle basé sur les statistiques empiriques.

Pour chaque jeu de données simulé, la capacité du critère d'information bayésien (BIC, SCHWARZ et al. 1978, annexe C.3.2) est évaluée pour la sélection du bon nombre de classes. L'algorithme EM est exécuté 30 fois pour différents nombres de classes ($K = 1$ à $K = 6$) et, à chaque répétition, le nombre associé au BIC minimum est sélectionné. Les résultats sont présentés sur la figure 3.7. Ensuite, pour un nombre fixe de sous-groupes, la qualité des estimations est mesurée par la racine de l'erreur quadratique moyenne (RMSE) entre les estimations et les vraies valeurs des paramètres. Les résultats sont discutés dans le paragraphe 3.1.3.2. Dans l'application de calibration considérée dans la section 3.2, il est important de pouvoir affecter les ensembles au bon sous-groupe. La performance de classification des modèles est évaluée par l'indice de précision globale (voir paragraphe 3.1.3.2). La moyenne de la RMSE et de la précision est calculée sur les 30 exécutions.

3.1.3.2 Évaluation des modèles

Les résultats des modèles de mélange sont présentés dans cette partie. Chaque cadre expérimental est évalué dans une problématique de sélection du nombre de classes et d'estimation des paramètres du mélange. L'estimation des modèles est étudiée pour différentes tailles d'échantillon $T \in \{50, 200, 1000\}$. Ce choix de différentes tailles est motivé par les longueurs du jeu de données régulièrement sélectionné dans l'application de modèles de calibration univariée et linéaire NGR. Lorsque T varie, le nombre de membres de l'ensemble reste fixé à $M = 50$. Ce nombre correspond au nombre de membres disponibles dans les données réelles utilisées dans la partie application. Les estimations sont aussi comparées pour différentes variations du nombre de membres $M \in \{1, 2, 5, 10, 25, 50\}$ pour une taille d'échantillon fixe $T = 200$.

Les scores introduits précédemment sont estimés au travers de $N = 30$ répétitions de chaque expérimentation. De plus, une procédure visant à ordonner les estimateurs et classes obtenues par modèle suivant les vrais paramètres est appliquée. Cette procédure aide à correctement comparer les vrais paramètres de chaque classe avec les estimations correspondantes. Pour cela, les estimateurs fournis par les modèles sont réordonnés par classe de façon à minimiser la distance euclidienne avec les vrais paramètres.

Sélection du nombre de classes. Considérons d'abord la sélection du nombre de classes. Pour le cas "Régulier" facile, le BIC fonctionne bien et le nombre de classes ($K = 4$)

est correctement sélectionné pour tous les modèles, dès que l'échantillon est suffisamment grand : $T \geq 200$ (résultats montrés dans l'annexe C.3.4.1, plus précisément sur la figure C.9). Pour le cas "Difficile" avec des sous-groupes se chevauchant fortement, la sélection du nombre de classes est plus difficile. La figure 3.7 montre le nombre de classes sélectionnées pour différentes tailles d'ensemble et une taille d'échantillon fixe $T = 200$. Tout d'abord, pour le modèle basé sur le super échantillon, le nombre de classes est toujours sous-estimé. Plus généralement, pour une taille d'ensemble faible, tous les modèles sous-estiment le nombre de classes. Le BIC favorise un modèle dans lequel les 2 sous-groupes de mêmes moyennes sont regroupés. Lorsque le nombre de membres augmente, les distributions des ensembles dans chaque sous-groupe sont mieux décrites et le BIC a de meilleures performances. En général, le modèle ajusté sur les statistiques empiriques conduit à une bonne sélection du nombre de sous-groupes pour les ensembles de taille supérieure à 10. Ces résultats sont proches de ceux du modèle basé sur les variables échangeables. Ensuite, pour le modèle basé sur les variables échangeables et les ensembles de tailles importantes, la sélection est plus efficace pour la dimension $d = 3$ que $d = 1$. L'information apportée sur la forme des sous-groupes et en dimension supérieure aide à identifier les sous-groupes. Enfin, la conclusion est que le modèle basé sur un vecteur de variables échangeables a les meilleures performances pour la sélection du nombre de sous-groupes.

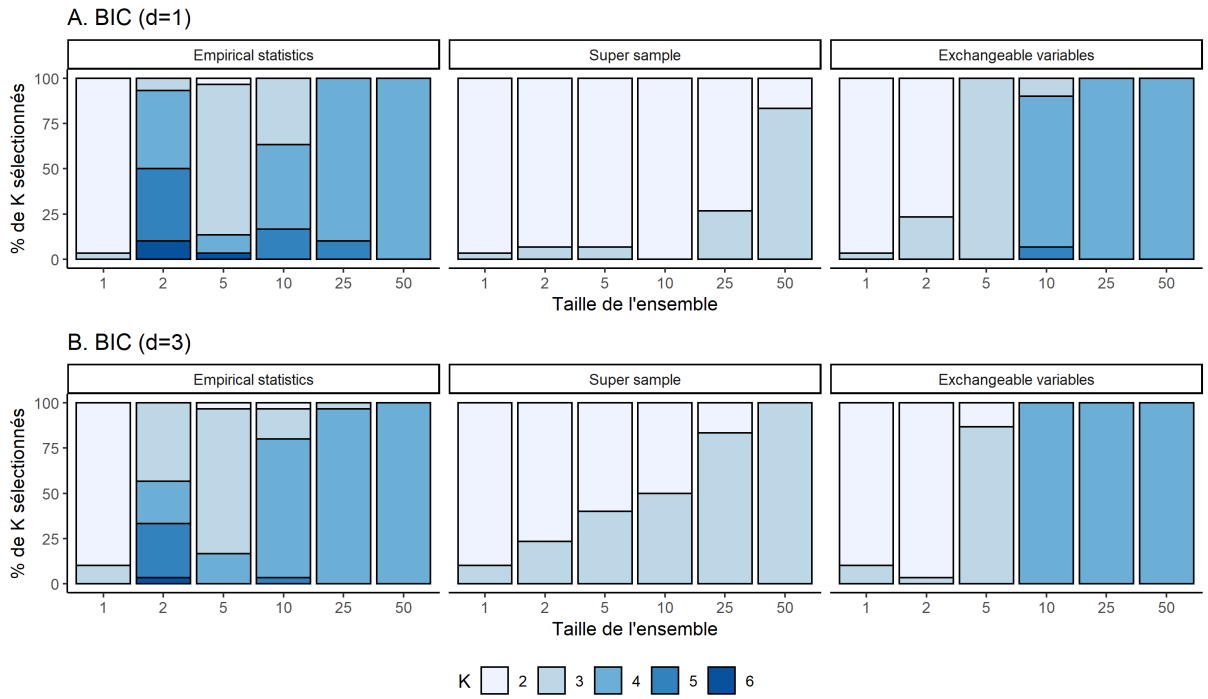


FIGURE 3.7 – Sélection du nombre de classes basée sur le score BIC pour le cas "Difficile" suivant différentes tailles d'ensemble et une taille d'échantillon fixée à $T = 200$.

Estimation des paramètres. Les performances de l'estimation des paramètres sont évaluées en calculant la RMSE entre les valeurs estimées et les vraies valeurs des paramètres. La distribution de la RMSE est illustrée par des boîtes à moustaches en figure 3.8 pour le cas "Régulier" avec différentes longueurs d'échantillon et en figure 3.9 pour le cas "Difficile" avec différentes tailles d'ensemble. Le nombre de classes est fixé à $K = 4$.

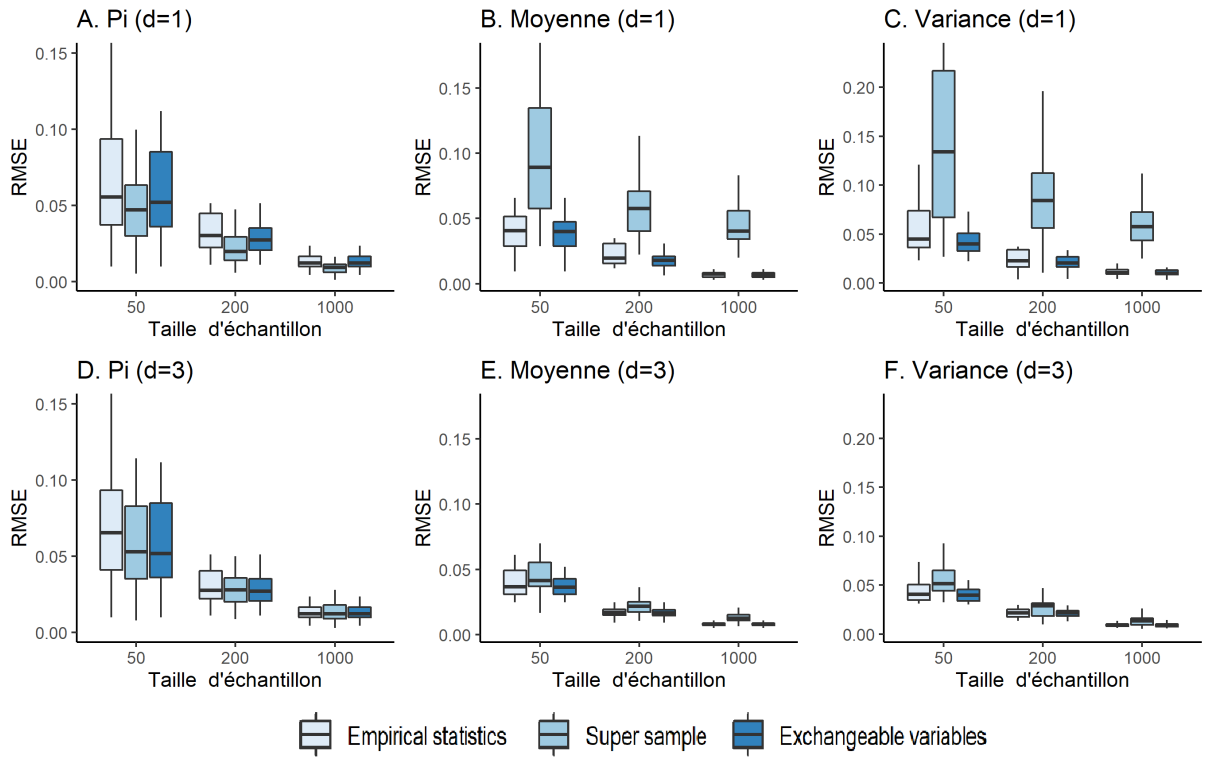


FIGURE 3.8 – Erreur sur l'estimation des paramètres pour différentes tailles d'échantillon et un nombre de membres fixé à $M = 50$ dans le cas "Régulier".

La figure 3.8 montre que la qualité de l'estimation augmente avec la taille de l'échantillon, comme on s'y attend pour les estimations par maximum de vraisemblance. Des résultats similaires sont obtenus pour le cas "Difficile" (résultats disponibles sur la figure C.10 de l'annexe C.3.4.2). Les estimations obtenues à partir du super échantillon ne sont rien d'autre que des estimations classiques du maximum de vraisemblance. Leurs propriétés peuvent donc être utilisées comme référence. Il ressort clairement des boîtes à moustaches de la figure 3.8 que le modèle basé sur les statistiques empiriques d'ensemble et le modèle ajusté selon l'hypothèse des variables échangeables conduisent à de meilleures estimations des paramètres de moyenne et de variance. En outre, leur estimation des proportions des sous-groupes est équivalente à celle basée sur le super échantillon. La principale différence entre le modèle basé sur le super échantillon et les deux autres, est que le super échantillon n'utilise pas la structure d'ensemble de l'ensemble de données. Les résultats montrent que l'ajout d'informations sur la distribution d'ensemble est utile pour obtenir de bonnes estimations des paramètres du GMM.

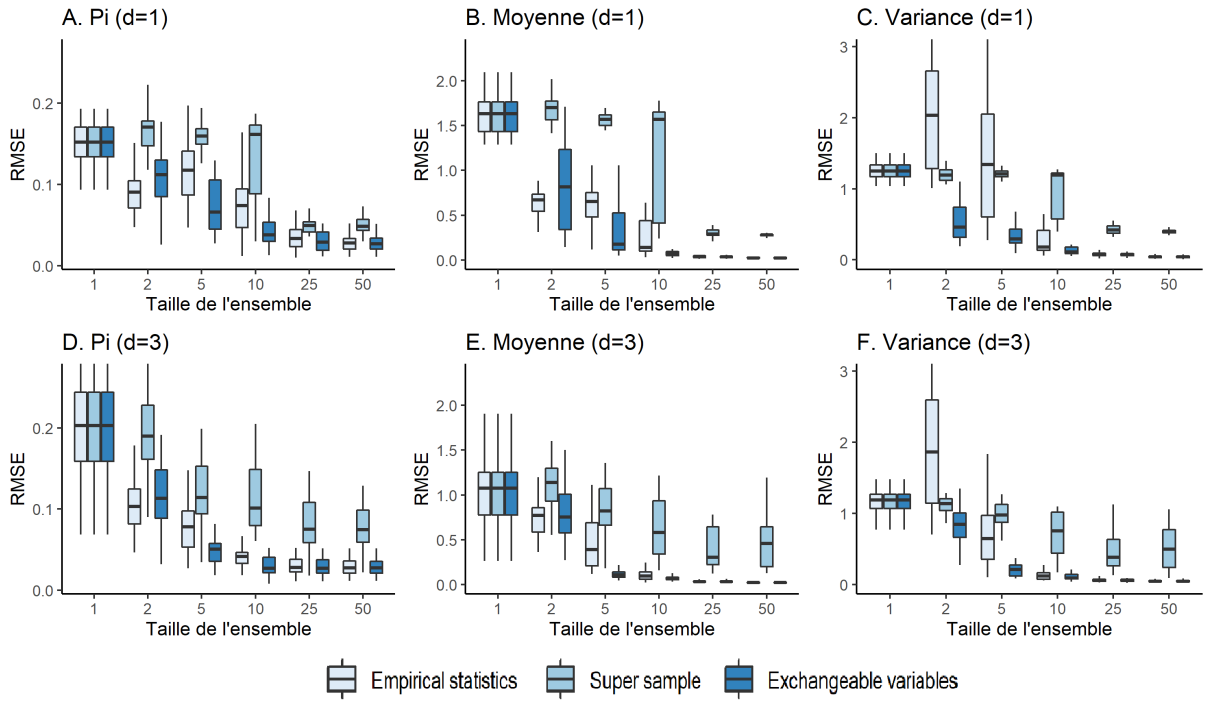


FIGURE 3.9 – Erreur sur l'estimation des paramètres pour différents nombres de membres et une taille d'échantillon fixée à $T = 200$ dans le cas "Difficile".

La figure 3.9 illustre l'impact du nombre de membres de l'ensemble pour une taille d'échantillon fixe $T = 200$. Comme pour la taille de l'échantillon, l'augmentation du nombre de membres de l'ensemble améliore généralement la qualité des estimations. Pour le cas "Difficile", l'information apportée par l'ensemble est vraiment notable. Ainsi, les modèles basés sur les statistiques empiriques et les variables échangeables conduisent à de meilleures estimations des paramètres que le modèle basé sur le super échantillon. Pour les grands ensembles, ces deux modèles ont des performances similaires. Cependant, les estimations du modèle basé sur les variables échangeables convergent plus rapidement. Pour tous les modèles, certaines erreurs sont importantes. Typiquement, il arrive que les deux classes avec les mêmes moyennes soient fusionnées et que la classe avec la moyenne la plus élevée soit divisée en deux groupes. Cela a un impact sur tous les paramètres : proportions, moyennes et variances.

Attribution des classes. Dans les cas où l'on ne parvient pas à identifier la bonne distribution de sous-groupes, la règle de prédiction de classe basée sur les probabilités *a posteriori* est également impactée. Cela est illustré sur la figure 3.10 où la précision de cette

règle est montrée pour le cas "Difficile". La précision mesure la qualité de la classification donnée par une observation. Comme pour les propriétés des estimations, il est clair que l'augmentation de la taille de l'ensemble permet d'améliorer la précision. Le modèle basé sur le super échantillon ne permet pas d'atteindre de bons taux de prédiction. Pour les grands ensembles (typiquement $M > 25$), les deux modèles qui prennent en compte la distribution de l'ensemble ont de bonnes performances similaires. Les résultats de la figure 3.10 montrent également que la précision peut être augmentée en utilisant une structure multivariée. En effet, la précision augmente plus rapidement avec M pour une dimension $d = 3$ que dans le cas $d = 1$.

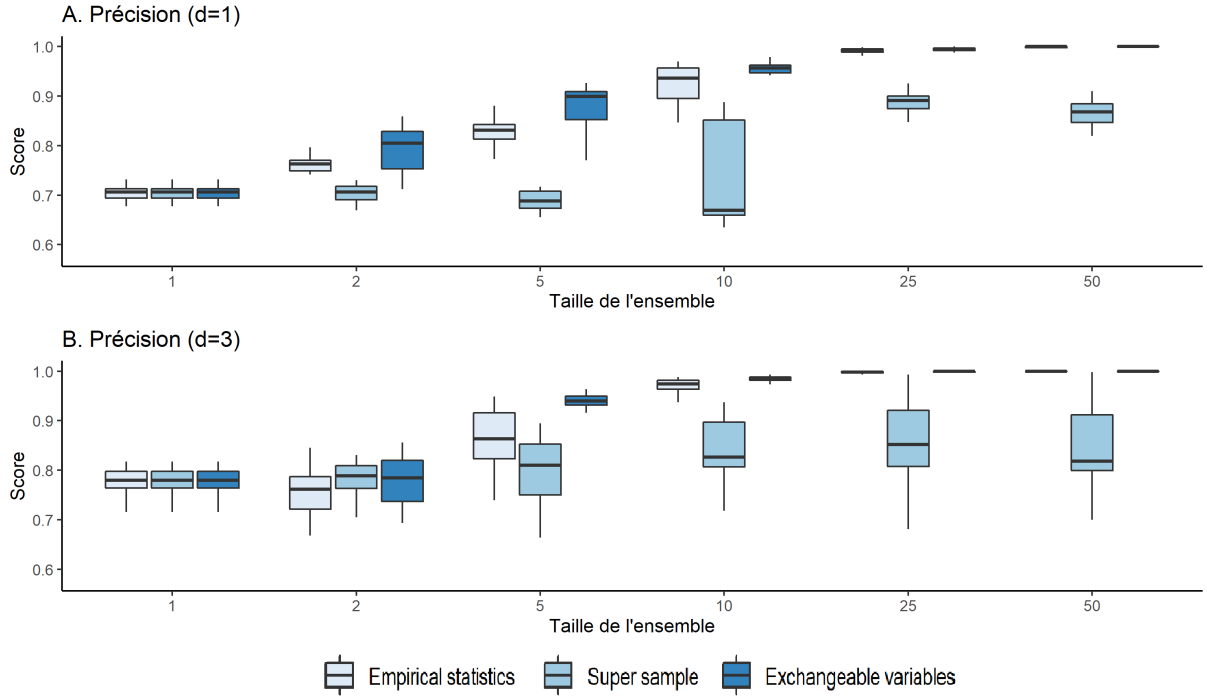


FIGURE 3.10 – Score de précision globale pour divers nombres de membres et une taille d'échantillon fixée à $T = 200$ dans le cas "Difficile".

En résumé, trois versions du GMM pour les échantillons d'ensemble ont été comparées sur des ensembles de données simulées. Les deux versions qui utilisent explicitement l'information sur la distribution de l'ensemble conduisent aux meilleures estimations des paramètres de la densité, quelle que soit la taille de l'ensemble ou encore des proportions de sous-groupes variables (résultats montrés dans l'annexe C.3.4.3). De plus, plus l'information fournie est structurée (distribution de l'ensemble, structure multivariée), meilleure est la précision. Enfin, dans les exemples considérés, une taille d'ensemble supérieure à 25 per-

met d'obtenir de très bons résultats pour la sélection du nombre de classes, l'estimation des paramètres et la discrimination des sous-groupes. Un aperçu des codes des différentes extensions des modèles de mélange gaussien et des affichages de résultats de simulation peut être retrouvé à cette adresse : `https://gitlab.com/gabrijou/gaussianmixturemodels`.

3.2 Calibration de prévisions météorologiques basée sur les classes issues du mélange

Maintenant qu'un modèle de mélange a été défini et validé pour ajuster les ensembles, la seconde partie de ce chapitre se focalise sur l'étude de l'apport des ensembles de prévision discriminés en sous-groupes dans une problématique de calibration. Les classes formées par le modèle de mélange regroupent des ensembles à la météorologie et au type d'erreur de prévisions similaires. Suivant la météorologie présente dans un sous-groupe, le type d'erreur de distribution des ensembles peut varier. Caractériser le type d'erreur relié à la météorologie étudiée dans un sous-groupe permet de fournir davantage d'information au prévisionniste sur la qualité des ensembles. Dans cet objectif de caractérisation, des tests d'hypothèse sont présentés et utilisés pour l'identification des types d'erreurs de distributions d'ensembles simulés dans la première section 3.2.1.

Ensuite, la problématique de calibration prenant compte les sous-groupes formés est introduite en section 3.2.2 à l'aide d'une extension du modèle de régression non homogène gaussienne (NGR ; Gneiting, Raftery et al. 2005). Le modèle étendu ainsi construit est étudié dans une application de calibration univariée et multivariée faisant intervenir la méthode de Schaake shuffle (Clark et al. 2004 ; Schefzik 2016) pour le cas multivarié. Dans un souci d'évaluation des performances des modèles, le score univarié d'analyse de probabilité des rangs continus (CRPS) et multivarié d'énergie score (ES ;Gneiting, Stanberry et al. 2008) sont utilisés. La section s'achève sur l'exploration des résultats des paramètres de mélange, la caractérisation des types d'erreurs et l'interprétation des coefficients du modèle NGR. Enfin, la correction apportée sur les données d'ensemble de la station Millau 6H à l'échéance de prévision de 3 jours est étudiée.

3.2.1 Caractérisation des types d'erreurs dans chaque classe

Un obstacle récurrent dans la modélisation du problème de calibration est l'identification des types d'erreurs de distributions d'ensembles de prévision dans un but de

meilleure adaptation des modèles. En s'appuyant sur l'histogramme de PIT ("probability integral transform") lié à l'histogramme de rang (outil visuel de vérification des ensembles de prévision présenté en section 1.3.3.1), il est possible de construire une procédure de tests d'hypothèse. Pour la suite de ce chapitre, il est important de rappeler les notations suivantes : un ensemble de prévision est défini selon $X^* \in \mathbb{R}^{M \times d}$ constitué de M membres échangeables de dimension d, et l'observation est notée $y \in \mathbb{R}^d$.

3.2.1.1 Histogramme PIT

Dans un cadre d'étude d'ensemble univarié, l'outil graphique des histogrammes de rangs est important pour émettre des hypothèses sur les relations entre les ensembles et les observations du jeu de données étudié. Une idée est d'utiliser l'histogramme PIT, une représentation continue de l'information des rangs d'ensembles pris par les observations (Gneiting et Katzfuss 2014). L'histogramme PIT est construit à partir de la variable $\tilde{R} \in [0, 1]$ résumant la relation entre la distribution des ensembles $\mathbf{F}$ et l'observation y :

$$\tilde{R} = \mathbf{F}(y) \tag{3.22}$$

Sous hypothèse de calibration la variable $\tilde{R}$ est considérée comme une variable aléatoire uniforme et donc d'espérance et variance connue. En suivant l'application de Taillardat, Mestre et al. 2016 du PIT histogramme sur les histogrammes de rangs, $\tilde{R}$ s'exprime de la façon suivante :

$$\tilde{R} = \frac{(rank_{X^*}(y) - 1)}{M} \tag{3.23}$$

où $rank_{X^*}(.)$ représente une fonction qui attribue un rang à l'observation y formée par les statistiques de rangs $x_{tm}^{(1)} \leq \cdots \leq x_{tm}^{(M)}$, $\forall m \in \{1, \cdots, M\}$.

Toujours sous des hypothèses d'ensembles calibrés, l'expression obtenue de $\tilde{R}$ (3.23) suit une variable uniforme d'espérance $\mathbb{E}[\tilde{R}] = \frac{1}{2}$ et de variance $\mathbb{V}[\tilde{R}] = \frac{M+2}{12M}$. En prenant $R \sim \mathcal{N}(0, 1)$ la variable normalisée de $\tilde{R}$, il est possible d'appliquer des tests de comparaisons de moyennes et de variances. Le but de ces tests est la caractérisation des types d'erreurs de distributions des ensembles de prévision d'un jeu de données.

3.2.1.2 Tests statistiques d'identification de biais et de dispersion

Les tests de comparaisons de moyenne et variance appliqués sur la variable normalisée de l'histogramme PIT mettent en place une mesure d'identification des erreurs de biais

et dispersion des distributions d'ensemble de prévision. Dans le cadre d'un échantillon de taille T de la variable normalisée R de l'histogramme PIT, le test de comparaison de moyenne (test de Student, $S\text{-}test(\mu_{R_k})$) est appliqué sous l'hypothèse nulle $H_0 : \mu_{R_k} = 0$. La contre-hypothèse est définie $H_1 : \mu_{R_k} \neq 0$ pour chaque sous-ensemble de données d'ensemble et d'observation appartenant au scénario k et ce $\forall k \in \{1, \cdots, K\}$. La statistique de test est définie par $t_S = \frac{\overline{\mu}_R - \mu_0}{s_R}$, où $\overline{\mu}_R$ est la moyenne empirique et s_R l'écart type empirique.

Ensuite, un test du ratio de variance (test de Chi-deux, $\chi\text{-}test(\sigma^2_{R_k})$) est effectué avec comme hypothèse nulle $H_0 : \sigma^2_{R_k} = 1$ contre l'hypothèse $H_1 : \sigma^2_{R_k} \neq 1$. La statistique de ce test s'écrit $q_\chi = \frac{(T-1)s^2{}_R}{\sigma_0^2}$. Si l'ensemble est calibré, la statistique de test t_S suit une loi normale $\mathcal{N}(\mu_R, \sigma^2_R)$, et q_χ une loi du chi-deux χ^2_{T-1} avec $T-1$ degrés de liberté. Dans le cas multivarié, la variable de rang normalisée R_k est générée et testée pour chaque variable $j \in \{1, \cdots, d\}$ et ce de manière indépendante.

Les rejets de l'hypothèse H_0 des tests sous regards du seuil de significativité α devraient indiquer la présence d'erreurs au sein de la distribution des ensembles. Plus précisément, rejeter l'hypothèse H_0 du $S\text{-}test$ de la variable normalisée R revient à identifier l'existence d'une forme en "L" caractérisant un biais au sein d'un histogramme de rang. Il est attendu que dans le cas d'un rejet de l'hypothèse H_0 du $\chi\text{-}test$, une erreur de dispersion de la distribution des ensembles serait identifiée.

3.2.1.3 Simulation de types d'erreurs d'ensemble

En vue d'évaluer les performances des tests dans l'identification des différents types d'erreurs, plusieurs comportements des ensembles avec leurs observations sont simulés dans cette section. Pour cela, un jeu de données $\mathcal{X}$ contenant T paires d'ensemble $X^* \in \mathbb{R}^{M \times d}$ de $M = 50$ membres et observation $Y \in \mathbb{R}^d$ de dimension $d = 1$ est défini. Les observations Y sont simulées selon une loi normale $N(0, 1)$. Afin de pouvoir générer différents comportements d'erreurs de distributions des ensembles, les ensembles sont générés selon une loi $N(\mu^*, \sigma^{2^*})$. Ainsi l'ensemble des paramètres $\{\mu^*, \sigma^{2^*}\}$ est pris dans $\{0, 1\} \times \{1, 0.3, 0.8, 1.5, 2, 5\}$ introduisant des déviations entre les ensembles et observations (Hamill 2001, Gneiting, Stanberry et al. 2008) :

— $\mu^* = 0$ et $\sigma^{2^*} < 1$ ensemble sous-dispersé ;
— $\mu^* = 0$ et $\sigma^{2^*} > 1$ ensemble sur-dispersé ;
— $\mu^* \neq 0$ et $\sigma^{2^*} = 1$ ensemble biaisé ;
— $\mu^* \neq 0$ et $\sigma^{2^*} < 1$ ensemble biaisé et sous-dispersé ;

— $\mu^* \neq 0$ et $\sigma^{2^*} > 1$ ensemble biaisé et sur-dispersé.

Les tests d'hypothèse paramétriques sont connus pour leur sensibilité face aux tailles d'échantillons étudiés. De ce fait, plusieurs tailles d'échantillons $T \in \{50, 200, 1000\}$ des ensembles et observations sont définis pour chaque type d'erreur. Les p-valeurs $P_{H_0}(t_S)$ et $P_{H_0}(q_\chi)$ de chaque test de comparaison sont obtenues pour des jeux de données de paires d'ensembles et observations. Pour chaque type d'erreur et chaque taille d'échantillon $N = 30$ jeux de données différents sont générés afin d'évaluer l'incertitude autour des p-valeurs estimées.

Application des tests d'hypothèses. La figure 3.11 montre les p-valeurs obtenues pour différents jeux de données. Le niveau de significativité $\alpha = 0.05$ est représenté en pointillés. Les p-valeurs du test de comparaison de moyenne affichent des résultats en dessous du seuil α concluant au rejet de H_0 dès lors qu'un biais est présent et ce pour différentes valeurs de T. Ce test montre une importante robustesse face aux tailles d'échantillons. De plus, il est capable de détecter efficacement la présence d'un biais dans les ensembles, et ce même en présence d'une erreur de dispersion supplémentaire. Quant aux p-valeurs du test de comparaison de variances, des fluctuations dépassant le seuil de significativité α sont à noter en présence d'erreur de dispersion pour des échantillons de taille faible, conduisant au rejet de H_1. Une sensibilité face aux tailles d'échantillons et aux mélanges des erreurs est observée autour des p-valeurs du χ-*test* pouvant mener à une erreur de rejet du type II dans une application sur données réelles.

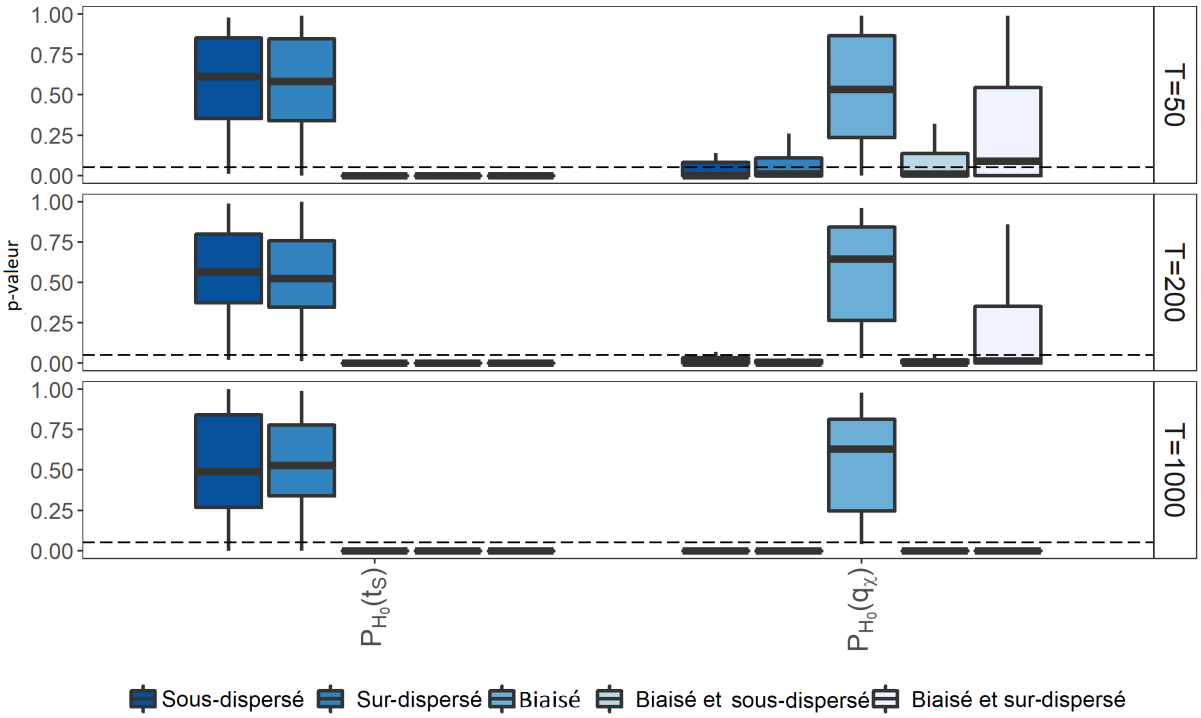


FIGURE 3.11 – P-valeurs des tests de comparaisons suivant différents types d'ensembles et tailles T d'échantillons générés. *La ligne en pointillés indique un seuil de significativité* $\alpha = 0.05$.

Les tests statistiques ont été évalués sur la caractérisation d'erreurs d'ensembles simulés dans un contexte gaussien. Dans la section 3.2.2, la procédure d'identification des erreurs sera testée sur des données réelles, dans le but de caractériser les types d'erreurs dans les sous-groupes d'ensembles issus de l'étape de classification.

3.2.2 Application aux données réelles

Dans cette section, une nouvelle méthode de calibration univariée est proposée, construite comme une extension du modèle linéaire NGR décrit en section 1.1.1.1. Une classification est d'abord réalisée sur les données d'ensemble à partir du modèle de mélange gaussien pour un vecteur de variables échangeables (décrit en 3.1.2.3), puis le modèle NGR est ajusté sur chaque sous-groupe de données. L'hypothèse motivant la construction de cette nouvelle méthode est que les différentes classes du modèle de mélange sont associées à des types d'erreurs de calibration spécifiques, et qu'il est donc utile de mettre en place

une calibration spécifique dans chaque classe. La méthode est testée et validée sur des données réelles.

Dans la pratique, les K classes sont d'abord obtenues sur des ensembles univariés ou multivariés à l'aide du modèle au vecteur de variables échangeables. Ensuite, K modèles NGR de calibration univariée sont ajustés indépendamment des autres. Cette nouvelle approche est appelée NGR_Z dans la suite de la section. Les performances prédictives des ensembles obtenus sont comparées avec un modèle NGR ajusté sur tout l'échantillon étudié, sans information provenant des classes. L'étude de la méthode de calibration proposée est étendue au contexte multivarié des données en proposant d'analyser les résultats d'application de l'algorithme Schaake Shuffle ($SimSS$; défini en section 1.2.2) couplé aux modèles NGR.

3.2.2.1 Données de prévisions d'ensemble et observations

Les données de prévision étudiées dans la partie application 1.3.1 des chapitres 1 et 2 sont réutilisées pour cette section. Les ensembles de prévision composant ces données sont issus du modèle CEPMMT fourni par le projet TIGGE. Pour éviter les effets saisonniers pouvant biaiser les résultats de sous-groupes fournis par le modèle de mélange, un sous-échantillon contenant uniquement les mois de janvier est extrait et utilisé dans ce chapitre. L'échantillon extrait contient donc les ensembles et observations des mois de janvier sur une période de 11 années de 2008 à 2018 pour deux horaires (6H et 18H) d'initialisation du modèle de prévisions numériques. L'ensemble de prévision est constitué de $M = 50$ membres échangeables.

Les variables météorologiques sélectionnées pour l'étude sont la température à 2 mètres du sol et les composantes zonale U et méridionale V du vent à 10 mètres du sol. Ces variables sont sélectionnées pour leurs corrélations physiques, leurs caractéristiques de distribution proches de la loi normale et potentiel économique. Ensuite, les échéances de prévision étudiées sont 3,5 et 10 jours, et les stations présentes dans l'échantillon sont Millau, Rennes et Strasbourg, comme dans l'étude de la section 1.3.1. Les observations utilisées sont prétraitées afin de retirer le biais systémique. Les modèles sont entraînés pour chaque échéance, station et horaire d'initialisation. Une base d'entraînement est construite à partir de 8 mois de janvier échantillonnés aléatoirement et une base de test est formée sur les 3 mois restants. Les scores sont ensuite calculés sur 30 répétitions de la procédure d'entraînement et test des modèles.

3.2.2.2 Score de calibration

Le score de probabilité des rangs continus (CRPS, "Continuous Ranked Probability Score") et le score d'énergie (ES, "Energy score") présentés en section 1.3.3.2 et 1.3.3.3 sont réutilisés ici pour pouvoir évaluer les performances de calibration univariée et multivariée. Ces scores ont montré quelques difficultés à comparer efficacement les performances prédictives des ensembles des différents modèles. Pour cela, le CRPS et ES sont étendus en score de compétence orienté (CRPSS ; ESS) pouvant évaluer la contribution entre deux modèles de calibration A et B :

$$CRPSS(A, B) = 1 - \frac{CRPS_A}{CRPS_B}. \tag{3.24}$$

Le score de compétence donne une mesure négative dès lors que le modèle B obtient des performances prédictives supérieures au modèle A. Les scores CRPSS et ESS peuvent seulement aider à la comparaison entre modèles dans le cas où ils partagent un dénominateur commun.

3.2.2.3 Calibration univariée de la température

Les performances prédictives des modèles sont comparées pour 3 types d'ensembles de prévision : le premier est nommé *Raw* faisant référence aux ensembles présents dans les données et n'ayant subi aucun post-traitement. Le type d'ensemble suivant nommé NGR est obtenu par post-traitements des ensembles à l'aide du modèle NGR. Le dernier type est NGR_Z indiquant les ensembles post-traités par l'approche ajustant un modèle NGR spécifique pour chaque échantillon d'ensemble et observations associé à une classe k. Ces notations sont conservées dans l'exploitation des résultats de calibration univariés et multivariés.

Le problème de calibration univariée de la température est étudié en premier, visant à corriger l'erreur de calibration de la température à 2 mètres du sol des ensembles de prévision. Deux types de classification sont comparés. La première nommée GMM_{uni} fait référence à des classes ajustées sur les données de température seulement. Le second type appelé GMM_{multi} construit des classes sur les ensembles multivariés composés des températures et composantes de vent (U,V). Les performances prédictives peuvent être comparées sur la figure 3.12, où le modèle NGR_Z est fixé comme référence pour calculer le score CRPSS. À rappeler qu'une valeur négative du CRPSS indique que le modèle NGR_Z obtient de meilleurs résultats de calibration comparés aux ensembles Raw et NGR.

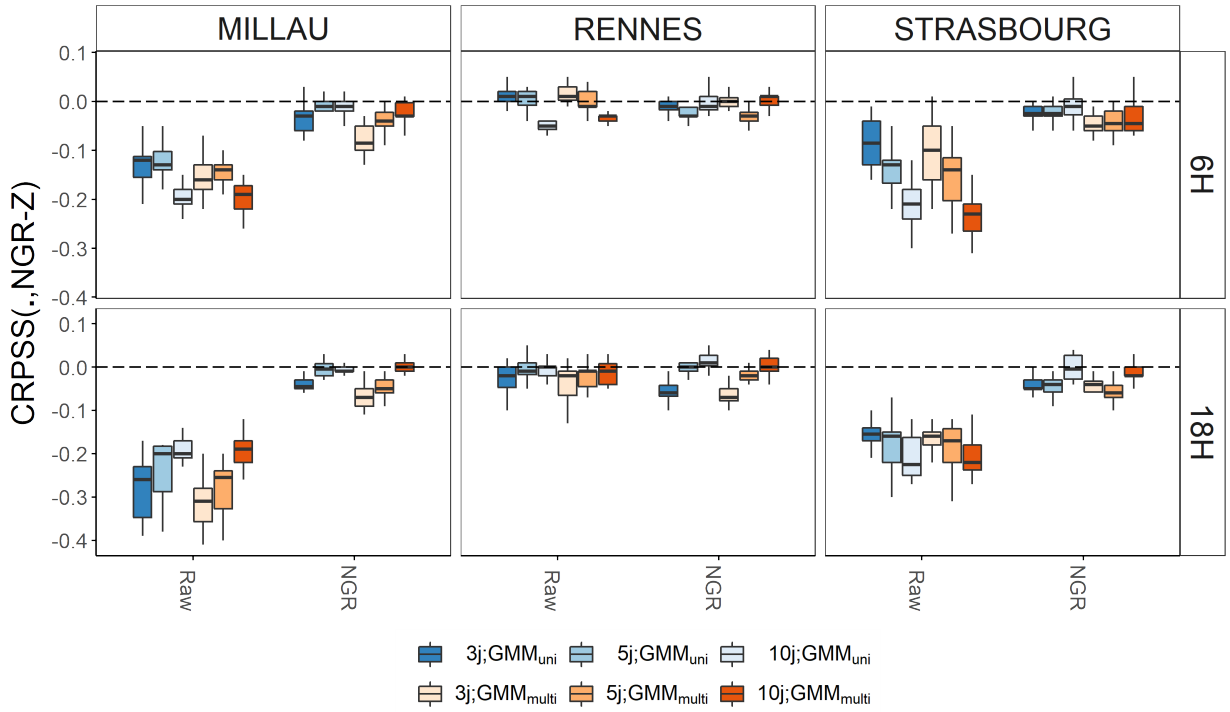


FIGURE 3.12 – Variable de température - CRPSS score des ensembles NGR_Z comparés aux ensembles *Raw* et NGR pour chaque station, heure et échéance de prévision, et ce pour les deux types de classification GMM_{uni} (classes ajustées sur les ensembles de température) et GMM_{multi} (classes ajustées sur les ensembles de températures et des composantes de vent (U,V)). *Les lignes pointillées délimitent un seuil de performance à atteindre pour dépasser le modèle* NGR_Z.

En premier lieu, des scores négatifs sont globalement observés montrant que les post-traitements améliorent les ensembles de prévision, et que les classes introduites dans le modèle NGR aident à l'amélioration de la qualité des ensembles. Cependant, cet effet n'est pas homogène pour les 3 stations qui affichent des météorologies locales différentes : l'amélioration n'est pas très claire pour la station de Rennes, là où elle est équivalente pour les stations de Millau et Strasbourg. Comme indiqué précédemment, Rennes est situé au Nord-Ouest de la France où le vent d'Ouest est dominant. Les températures d'hiver de cette région ne représentent pas de réel problème de prédiction. Dans des conditions dépressionnaires, la température se stabilise autour de 10 degrés Celsius. Dès lors que des conditions anticycloniques surviennent, la température diminue tombant aux alentours de 5 degrés Celsius. Ce type de conditions météorologiques reste relativement stable avec une durée moyenne tournant autour des 2 semaines. Quant aux conditions dépressionnaires,

elles sont assez faciles à prédire dû à leur direction venant de l'Ouest et donc relativement visible en avance. À Millau, la température se retrouve grandement impactée par le vent. Les valeurs peuvent s'effondrer brutalement dès que le vent commence à souffler en provenance du Nord le long de la vallée du Rhône. Pour Strasbourg, les épisodes d'événements météorologiques du type gel, neige et brouillard rendent la prévision difficile à fournir localement avec beaucoup de précision.

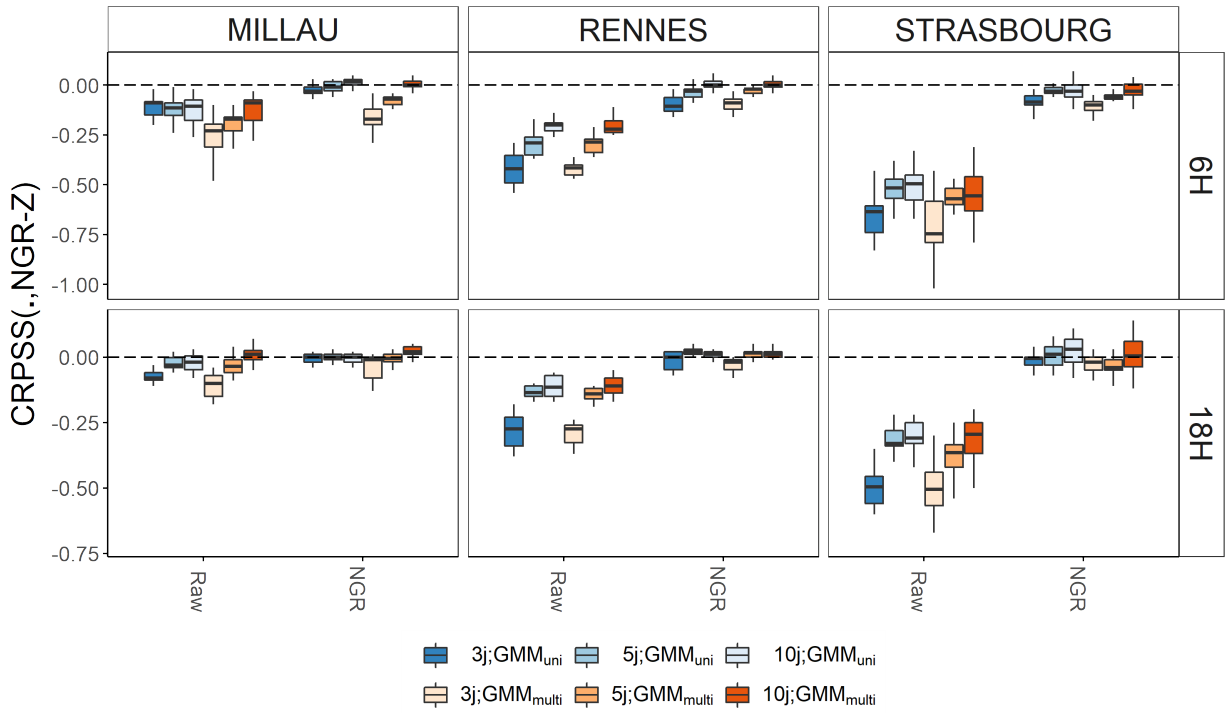


FIGURE 3.13 – Variable U, composante zonal du vent - CRPSS score des ensembles NGR_Z comparés aux ensembles *Raw* et NGR pour chaque station, heure et échéance de prévision, et ce pour les deux types de classification GMM_{uni} (classes ajustées sur les ensembles de température) et GMM_{multi} (classes ajustées sur les ensembles de températures et des composantes de vent (U,V)). *Les lignes pointillées délimitent un seuil de performance à atteindre pour dépasser le modèle* NGR_Z.

Il est aussi intéressant de voir l'amélioration de score entre GMM_{uni} (n'utilisant que les données d'ensemble de température pour ajuster les classes et nuancées en bleu dans la figure 3.12) et GMM_{multi} (utilisant les données d'ensemble de température et composante (U,V) du vent pour ajuster les classes et nuancées en rouge dans la figure 3.12). Pour la station de Millau à 18H, l'apport obtenu avec l'utilisation du modèle GMM_{multi} est observé tandis qu'il n'est pas visible pour Strasbourg. Dans le cas de la station Millau,

cet apport s'explique par les propriétés de la météorologie locale définie par les interactions entre la température et le vent permettant aux classes du modèle GMM_{multi} de fournir une information de taille au modèle NGR comparé aux classes utilisant seulement l'information de température.

Des expérimentations similaires de calibration univariée ont été menées indépendamment sur chaque composante de vent (U,V). Des conclusions proches à celles faites précédemment sont observées : la calibration issue des ensembles de composantes de vent se retrouve souvent améliorée par les informations des classes inclues dans le modèle NGR_Z. Les performances prédictives des modèles peuvent être comparées sur la figure 3.13 pour la composante U. Il est important de noter que pour la station de Rennes, les ensembles des composantes de vent montrent de fortes valeurs négatives (surtout pour U), ce qui n'était pas le cas pour les ensembles de température. Ce résultat est lié au fait que pour cette station, l'information des classes issues de toutes les variables météorologiques reste peu informative (pour cela, il suffit de remarquer la différence entre GMM_{uni} et GMM_{multi} pour la station de Rennes sur la figure 3.12). Les résultats de comparaisons de modèle de calibration pour la composante V peuvent être analysés sur la figure C.14. Pour cette composante, le CRPSS montre une amélioration importante entre les ensembles *Raw* et ceux post-traités par NGR. Cependant, la contribution des classes avec le modèle NGR est moins évidente et paraît plus discrète sur la composante V. La corrélation entre le vent et la température s'avère moins informative dans le cas des données de Rennes que pour Millau et Strasbourg.

3.2.2.4 Calibration multivariée

Les sous-groupes de classes formés à partir des dépendances entre la température et les composantes du vent (U,V) ont montré un apport non négligeable dans la calibration univariée. Une suite logique à cette étude est l'application de l'algorithme Schaake shuffle couplé aux modèles NGR et NGR_Z pour un objectif de calibration multivariée des ensembles de prévision.

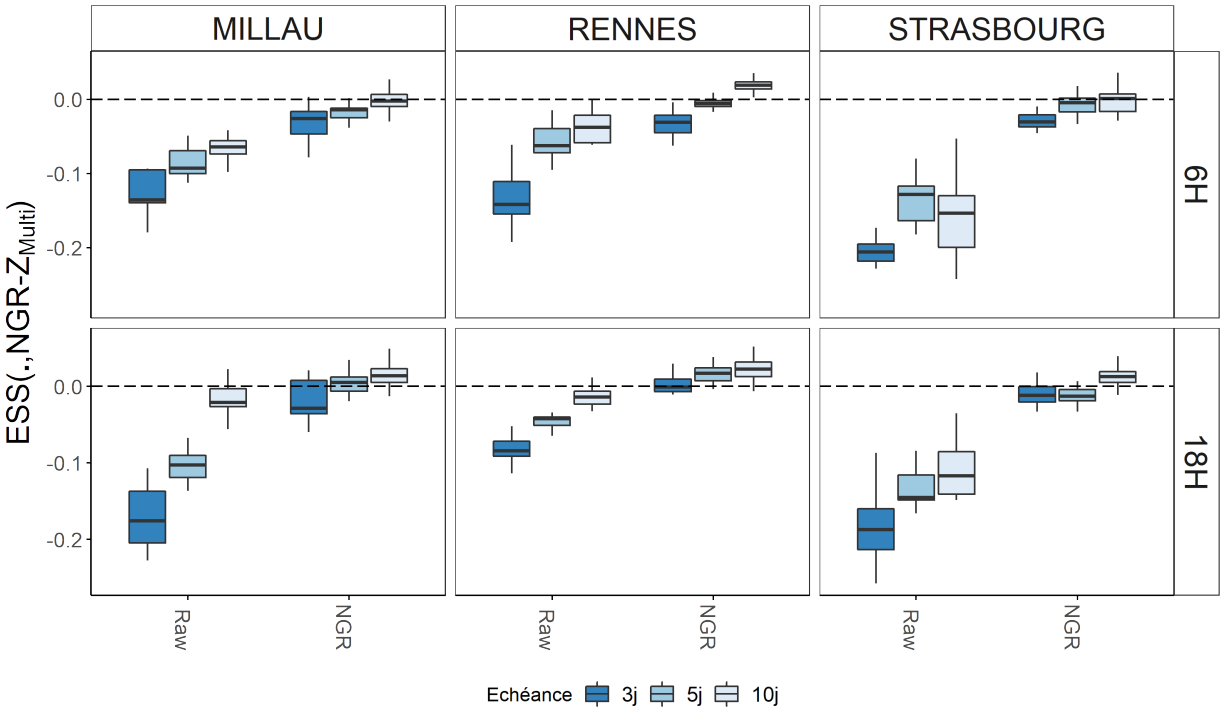


FIGURE 3.14 – ESS score des ensembles NGR_Z comparés aux ensembles *Raw* et NGR pour chaque station, heure et échéance de prévision, et ce pour le type de classes GMM_{multi} (classes ajustées sur les ensembles de températures et des composantes de vent (U,V)). *Les lignes pointillées délimitent un seuil de performance à atteindre pour dépasser le modèle* NGR_Z.

La figure 3.14 représente les résultats de comparaisons de modèles multivariés à l'aide du score ESS avec le modèle de calibration univariée NGR_Z en référence. Comparé aux ensembles sans post-taitement *Raw* et au modèle NGR standard, le modèle avec les classes NGR_Z couplé à l'algorithme Schaake shuffle montre globalement des performances supérieures voir équivalentes en moyenne par station et horaire à l'échéance de prévision de 3 jours. Cependant, les performances du modèle NGR_Z déclinent avec l'évolution des échéances devenant équivalent, voire moins bon que le simple modèle NGR. L'erreur de distribution des ensembles multivariés montre une complexité difficile à approcher par un simple modèle de calibration linéaire à une échéance de 3 jours. Néanmoins, le pouvoir prédictif de la distribution des ensembles étant suffisamment élevé à cette échéance, permets au modèle NGR_Z de corriger ces erreurs. Ensuite, l'augmentation des échéances provoque un déclin du lien entre les erreurs de prévisions des ensembles et la structure multivariée des variables. Les distributions étant très étirées à cette échéance forment un

espace de valeurs important rendant difficile la formation de sous-groupes caractéristiques de régimes météorologiques et d'erreurs dans les ensembles. Dans cette situation, les ensembles formés par NGR_Z peuvent induire des erreurs plus importantes qu'un simple modèle NGR.

Pour terminer, à l'aide des sous-groupes capables d'agglomérer les ensembles de météorologies et erreurs similaires, le modèle linéaire NGR_Z parvint à mieux corriger les erreurs de distribution d'ensembles multivariés à une échéance de 3 jours.

3.2.2.5 Analyse approfondie pour la station de Millau

Dans cette partie, l'analyse se restreint aux résultats de calibration pour la station Millau à 18H et 3 jours d'échéance de prévision. Une attention particulière est portée à l'interprétation des résultats des classes et de leur lien avec les distributions d'erreurs des ensembles, dans l'idée de comprendre le rôle que les classes ont dans le modèle de calibration NGR_Z.

		Modèles					
		GMM_{uni} Classes			GMM_{multi} Classes		
Variables	Paramètres	1	2	3	1	2	3
TMP	μ_k	-2.21	1.68	5.51	-1.30	3.19	3.39
	σ_k^2	3.33	2.42	3.47	4.38	5.05	4.42
	π_k	0.25	0.39	0.36	0.32	0.33	0.35
U10	μ_k				2.39	-0.40	2.56
	σ_k^2				2.70	6.80	6.70
V10	μ_k				-1.26	0.51	-0.98
	σ_k^2				3.29	2.42	2.63

TABLE 3.2 – Paramètres estimés des marginales du modèle du mélange gaussien pour Millau 18H avec une échéance de 3 jours. *GMM_{uni} représente les classes ajustées sur les ensembles de température et GMM_{multi} celles des ensembles multivariés de température et composantes de vent (U,V).*

Le tableau 3.2 révèle les paramètres estimés des classes obtenues par GMM_{uni} (classes associées aux ensembles de température) et GMM_{multi} (classes associées aux ensembles multivariés de température et composantes de vent (U,V)). Le modèle GMM est construit pour $K = 3$ classes. Pour le modèle GMM_{uni}, les moyennes des classes sont significativement différentes. Dans le cas du modèle multivarié GMM_{multi}, la classe 1 reste associée

aux faibles températures, tandis que les classes 2 et 3 sont définies pour des valeurs plus fortes de températures pour deux situations distinctes de vent. La classe 3 décrit un vent soufflant du Sud-Ouest (moyenne de V négative et positive pour U). La classe 2 obtient un vent de plus faible intensité provenant du Nord-Est. Cette dépendance entre la température et le vent au sein des classes multivariées semble aider la calibration univariée pour la station de Millau (comme montré dans la section précédente 3.2.2.3).

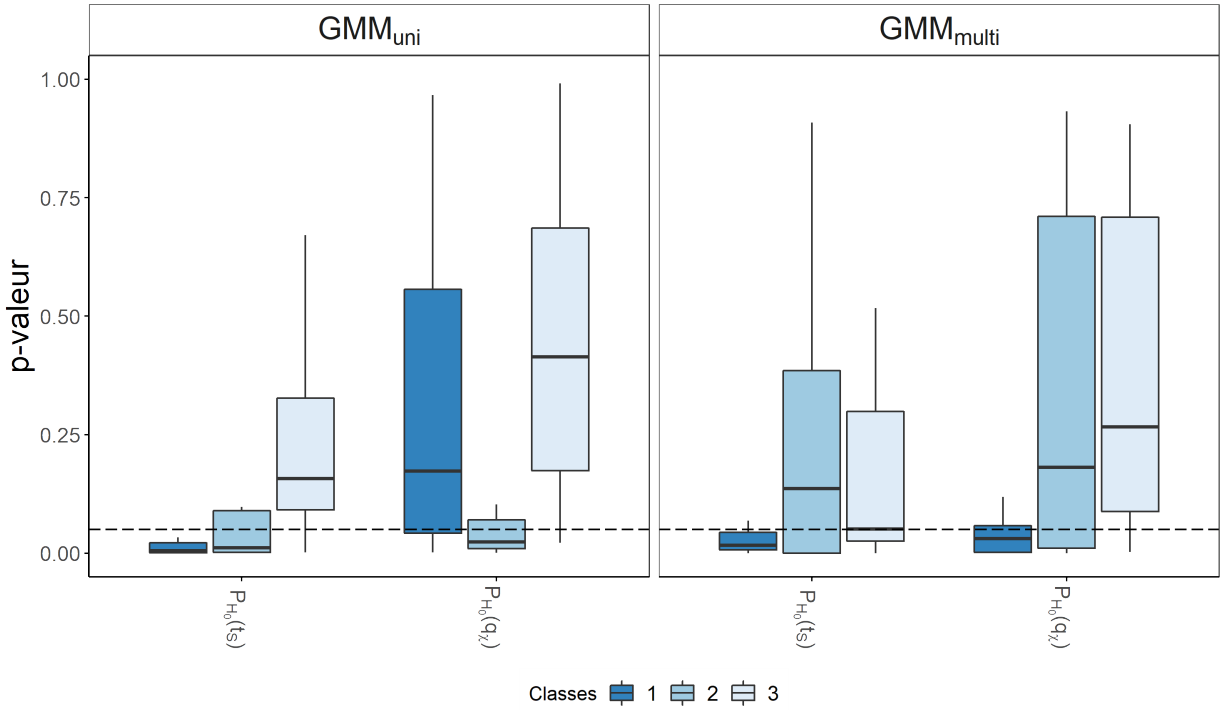


FIGURE 3.15 – P-valeurs des tests d'hypothèse appliqués sur les histogrammes PIT des ensembles (pour une échéance de 3 jours) et observations de température associées aux deux types de classes à Millau 18H. $P_{H_0}(t_S)$ *: p-valeur du test de comparaison de moyennes de Student ;* $P_{H_0}(q_\chi)$ *: p-valeur du test de comparaison de variances du Chi-deux ; les nuances de couleurs représentent les classes suivant les modèles* GMM_{uni} *(classes ajustées sur les ensembles de température) et* GMM_{multi} *(classes ajustées sur les ensembles de températures et des composantes de vent (U,V)). Les lignes pointillées représentent le seuil de significativité* $\alpha = 0.05$.

Les résultats de caractérisation du type d'erreur des distributions des ensembles à l'aide de tests par classes sont visibles sur la figure 3.15. Plus particulièrement, les p-valeurs des tests de comparaison de moyennes ($P_{H_0}(t_S)$) et de variances ($P_{H_0}(q_\chi)$) y sont représentées pour des ensembles de température avec une échéance de 3 jours et ce, pour les deux types de classes GMM_{uni} et GMM_{multi} en nuances de couleurs. Sur cette figure, la

classe 1 associée aux faibles températures est identifiée comme biaisée pour les deux types de classification GMM_{uni} et GMM_{multi} avec majoritairement des p-valeurs inférieures à α. Cette même classe associée au modèle GMM_{multi} est également identifiée comme possédant un problème de dispersion. Les tests de la classe 2 avec des valeurs moyennes de températures pour le modèle GMM_{uni} concluent sur la présence de biais et de problème de dispersion. Dans le cas de la classe 3 disposant de fortes valeurs de températures, les p-valeurs n'indiquent aucun type d'erreur en particulier.

Les classes 2 et 3 du modèle GMM_{multi} associées à de fortes valeurs de température montrent des probabilités aux fluctuations importantes dépassant le seuil α. Le franchissement de ce seuil empêche de conclure à la présence de types d'erreurs de distribution des ensembles par ces tests. Il faut noter que dans un cas de non-identification d'un type d'erreur de distribution, les tests ne permettent pas de conclure à la non-présence d'erreurs au sein d'un ensemble. De plus, la classe 1 de faible température du modèle GMM_{multi} semble récupérer certains individus responsables du problème de dispersion de la classe 2 du modèle GMM_{uni}. Ce point peut expliquer l'élévation de la moyenne de température de la classe 1 de GMM_{multi} par rapport à celle de GMM_{uni}.

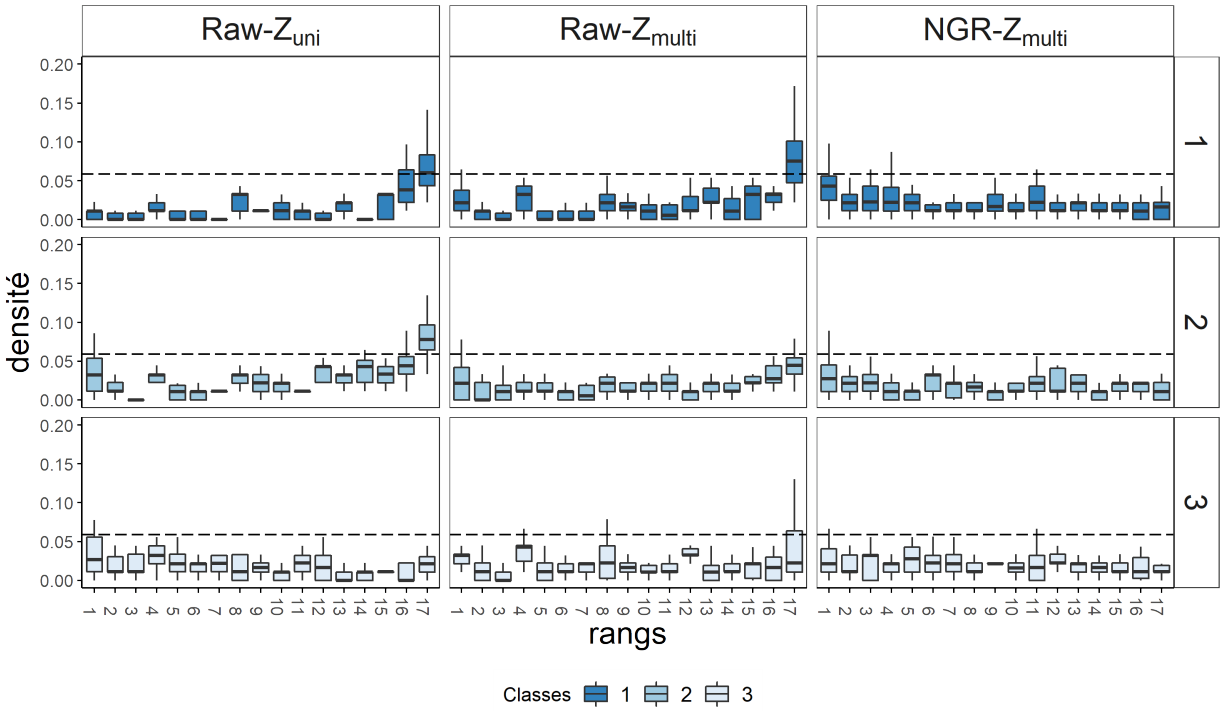


FIGURE 3.16 – Histogrammes de rangs des observations et des ensembles de température avec 3 jours d'échéance de prévision pour la station de Millau pour le mois de Janvier à 18H. *Première colonne : les ensembles Raw pour des classes univariées Z_{uni} ; Seconde colonne : les ensembles Raw pour des classes multivariées Z_{multi}. Les lignes pointillées délimitent le seuil de densité à atteindre et conserver pour former un histogramme de rangs uniforme.*

Afin d'étudier plus en détail les remarques émises lors de l'analyse des résultats de tests, les histogrammes de rangs sont affichés pour chaque type de classes (GMM_{uni} ou GMM_{multi}) sur la figure 3.16. Ces histogrammes de rangs sont générés à partir des ensembles *Raw*, post-traités et des observations pour la station Millau à 18H et échéance de prévision de 3 jours. La classe 1 pour les deux types de modèles est associée aux températures faibles pour lesquelles l'histogramme de rang semble former un L inversé correspondant à la présence d'un biais négatif dans la distribution des ensembles et donc appuyant les remarques des tests précédents. De plus, l'histogramme de rangs des ensembles *Raw* pour cette même classe du modèle GMM_{multi} semble montrer une légère élévation de rangs opposés à ceux formant un L avec quelques creux visibles dans les rangs centraux laissant présager une erreur de sous-dispersion et donc concorder avec le problème de dispersion remarqué dans les tests.

Pour les températures les plus fortes associées à la classe 2 du modèle GMM_{uni} et aux classes 2 et 3 du modèle GMM_{multi}, les histogrammes de rangs sont presque uniformes. Ils montrent que les ensembles *Raw* de ces classes sont de meilleure qualité pour ces situations météorologiques. Les histogrammes de rangs des ensembles post-traités par NGR_Z basé sur des classes issues des ensembles multivariés sont affichés dans la troisième colonne de la figure 3.16. Les histogrammes de rangs des températures ainsi obtenus sont proches des histogrammes uniformes pour chaque classe, représentant une calibration correcte au sein des ensembles post-traités.

	Moyenne		Variance	
Classes	Ordonnée	Pente	Ordonnée	pente
1	1.28	1.03	1.00	1.00
2	0.55	0.87	0.21	1.04
3	1.52	0.72	0.42	0.81

TABLE 3.3 – Coefficients NGR_Z de températures pour les données de Millau Janvier 2015 à 18H.

Les paramètres du modèle NGR_Z pour chaque classe sont affichés dans le tableau 3.3. Toutes les ordonnées sont positives indiquant que les moyennes et variances ont besoin d'être augmentées. La classe 1 ayant notamment un problème de dispersion détecté par les tests d'hypothèse, reçoit la plus forte augmentation de variance des différentes classes. Cette erreur de dispersion était également catégorisée comme une sous-dispersion par l'étude de l'histogramme de rangs. Les pentes sont légèrement inférieures ou égales à 1, ces coefficients peuvent restreindre les paramètres. Dans la classe 1, la moyenne et variance des ensembles sont augmentées ; ceci s'observe également dans la classe 2 mais avec un effet plus faible. Dans la classe 3, la moyenne n'est presque pas changée, mais la variance se retrouve diminuée.

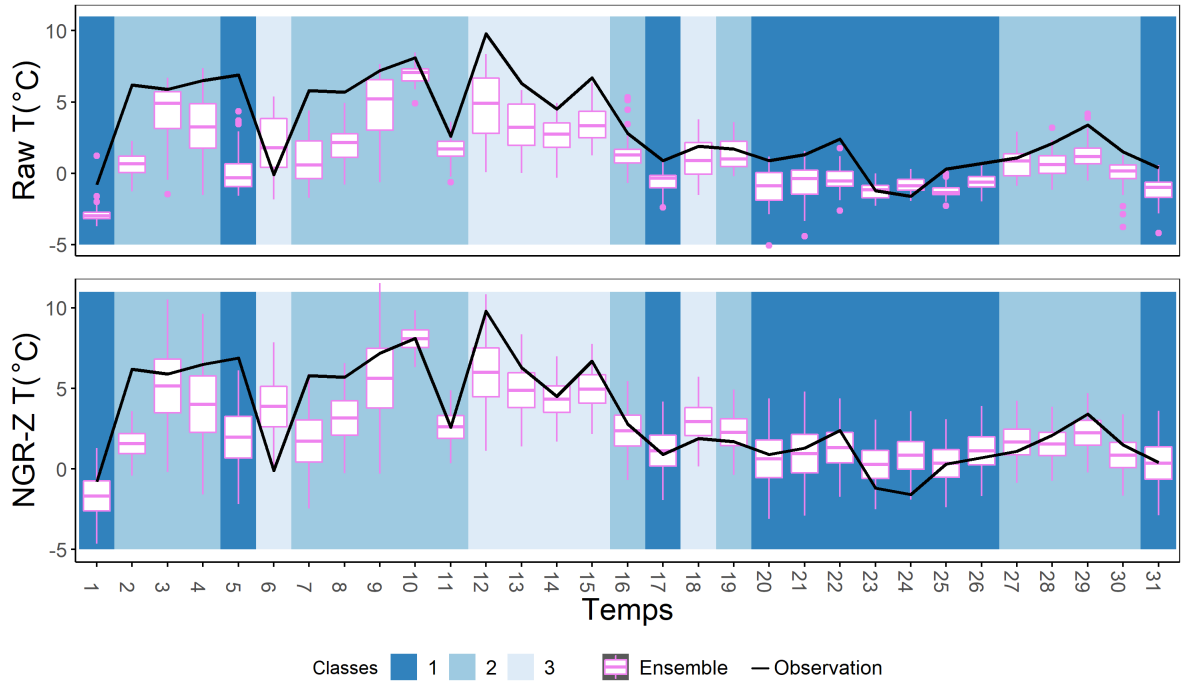


FIGURE 3.17 – Température à Millau, Janvier 2015 à 18H. *Première ligne : CEPMMT ensembles de prévision "Raw" (échéances de 3 jours) ; Seconde ligne : ensembles de prévision calibrés par le modèle NGR_Z. La ligne noire représente les observations. Les nuances de couleurs en fond représentent les classes ajustées par le modèle GMM_{multi}.*

Finalement, le résultat du type de classe GMM_{multi} est représenté avec différentes nuances sur la figure 3.17. La première ligne montre l'ensemble de prévision sans post-traitement (*Raw*). La seconde ligne affiche les résultats des ensembles post-traités par NGR_Z. En reprenant la première figure 3.1 où les différents types d'erreurs de calibration sont identifiés au travers de certains temps (biais, sur ou sous-dispersion), il est noté que ces situations sont associées à différentes classes ajustées. Par exemple, les ensembles biaisés (jours 12 à 15) font partie de la classe 3 tandis que les temps associés à la sous-dispersion (jours 23 à 25) sont dans la classe 1. Après calibration, le biais des ensembles de prévision est réduit pour les jours 12 à 13 (à noter qu'il n'est pas évident pour le jour 6 qui est aussi dans la classe 3 mais admet un biais positif). De plus, la dispersion des ensembles a été aussi augmentée dans la classe 1 (visible sur les jours 22 à 26).

3.3 Conclusion

Dans ce chapitre, trois extensions du modèle de mélange gaussien ont été présentées pour résoudre les problèmes d'ajustement sur des données d'ensemble de réalisations échangeables. Une étude de simulation a été menée pour comparer les nouveaux modèles et conclure sur leur efficacité à fournir des estimations pertinentes. La conclusion de cette étude est que le modèle aux statistiques empiriques et celui au vecteur de variables échangeables ont montré des performances similaires en réussissant à fournir de bons résultats d'estimations, et ce pour des ensembles de tailles acceptables (au moins dépassant les 25 membres). Dans le cas de petits ensembles, le modèle au vecteur de variables échangeables montre de meilleures performances que le modèle aux statistiques empiriques.

Pour ces travaux, les modèles ont été restreints par une hypothèse d'indépendance entre membres. Or cette hypothèse est discutable. Les travaux de O'Neill 2009 portant sur l'échangeabilité de variables aléatoires reviennent sur cette hypothèse. Il exprime le fait qu'il est difficile de maintenir dans la réalité la notion d'indépendance entre membres amplement utilisée et admet une certaine covariance existante entre variables échangeables. Il définit notamment une borne pour cette covariance entre membres en s'appuyant sur la distribution empirique du paramètre conditionnant l'ensemble (cas exploité dans ce chapitre). Pour de futurs travaux, la relaxation de cette hypothèse en considérant une structure de covariance entre membres serait intéressante à étudier.

D'autres distributions asymétriques ou disposant de longues queues pourraient être aussi considérées pour conduire l'ajustement des ensembles échangeables sur d'autres variables météorologiques comme les précipitations par exemple. De plus, une chaîne de Markov pourrait être utilisée pour modéliser les transitions entre classes et décrire une évolution temporelle comme un modèle de chaîne de Markov cachée.

La dernière partie de ce chapitre s'intéresse à l'intégration des nouvelles classes ainsi formées dans un modèle traditionnel NGR de calibration univariée appliqué aux ensembles de prévision à moyenne échéance. Le modèle proposé est appliqué à la calibration des variables de température et composante du vent (U,V) et ce pour trois différentes localisations en France. Le couplage des classes au modèle de calibration a permis de rendre plus flexible l'application du modèle et d'obtenir des résultats similaires, voire plus performants que le modèle standard sans classe. De plus, les classes apportent des éléments d'interprétation quant à la nature de l'erreur de calibration observée et corrigée par le modèle NGR. Enfin, le modèle NGR_Z est une méthode facile à déployer et qui peut être

facilement automatisée dans une application plus grande échelle.

Pour finir, il est tout à fait possible que d'autres modèles de calibration puissent être considérés pour un couplage avec les classes formées. Le modèle paramétrique 'Bayesian Model Averaging' (BMA) introduit par Raftery et al. 2005 pourrait être notamment utilisé. Cette approche offre plus de flexibilité que le modèle linéaire NGR en appliquant une méthode de calibration linéaire sur chaque membre des ensembles du sous-groupe formé. Des approches non paramétriques comme les modèles de forêts employés dans le chapitre 2 sont également envisageables intégrant l'information des classes comme covariable qualitative dans l'architecture du modèle.

Pour résumer :

- Des méthodes d'inférence ont été développées pour estimer les paramètres d'un modèle de mélange gaussien pour les données d'ensemble.
- Une méthode originale de calibration univariée est proposée dans laquelle la correction des ensembles dépend des régimes météorologiques.
- La méthode de calibration proposée permet d'approfondir l'interprétation des corrections effectuées.
- Des tests d'hypothèse peuvent être appliqués pour caractériser les types d'erreurs présents dans les sous-groupes d'ensembles formés.
- Ces travaux ont donné lieu à un article soumis :
  Jouan, G., Cuzol, A., Monbet, V., Monnier, G. (2021) Gaussian mixture models for clustering and calibration of ensemble weather forecasts.
  et au développement d'un package R disponible sous github :
  `https://gitlab.com/gabrijou/gaussianmixturemodels`

Chapitre 4

Développement d'une interface web pour les prévisions, ensembles de prévision et leur analyse

---

La prévision météorologique déterministe représente une information importante pour des secteurs d'activité comme la production de denrées alimentaires, d'énergies renouvelables, la gestion de risques, planification d'activités ou d'opérations de maintenance (Bauer, Thorpe et Brunet 2015). Dans des secteurs particuliers comme l'agriculture, lors de récoltes de céréales comme le blé s'étalant sur quelques semaines d'été, il est nécessaire d'avoir des conditions météorologiques précises marquées par une faible humidité et aucune précipitation. Ces conditions météorologiques garantissent l'accès des machines agricoles au champ et la qualité du grain évitant également un surcoût de séchage nécessaire à sa conservation. La contrainte météorologique impose aux agriculteurs une importante tâche de gestion de la location et logistique des machines agricoles. Cette tâche est impactée par l'information des prévisions à court et moyen terme. Ces contraintes ne se limitent pas qu'au secteur de l'agriculture. Dans ces travaux, Taillardat 2017 présente une activité de planification du séchage d'enrobé effectué par les travaux publics. Cette activée est également dépendante de la prévision de conditions météorologiques particulières. Les prévisions déterministes connues pour leurs incertitudes croissantes avec l'évolution des échéances de prévision font courir un risque non négligeable de changements imprévus de planning entraînant des pertes de revenus. Les risques générés par les erreurs de prévision justifient l'importance économique d'utiliser des ensembles de prévision pour évaluer les incertitudes des prévisions déterministes Y. Zhu et al. 2002.

C'est dans cet intérêt que Scalian (groupe d'ingénierie et de conseil en digital, `https://www.scalian.com/accueil/`) propose de développer un outil d'affichage et d'exploi-

tation des prévisions et des ensembles de prévision à moyen terme à destination des entreprises et particuliers. L'idée est qu'à travers une interface web, l'utilisateur dispose de fonctionnalités simples pour définir les événements météorologiques représentant les contraintes météorologiques qui le concernent. Ensuite, les probabilités d'occurrences de ces événements issues des ensembles de prévision seraient représentées pour différentes échéances. L'utilisateur peut ainsi créer un événement appelé "beau temps" en fixant, par exemple, une vitesse de vent moyenne inférieure à 3 m/s et des précipitations nulles. Les probabilités de cet événement seraient ainsi prédites à l'aide des traitements statistiques proposés dans le chapitre 2 et appliqués sur les ensembles. Mais aussi, à l'aide des contributions du chapitre 3, le prototype propose diverses informations décrivant l'ensemble. Plus particulièrement, les contributions du chapitre 3 regroupent les ensembles en classes dont les paramètres caractérisent les erreurs de distribution des ensembles des classes prédites, et ce pour chaque point spatial et échéance. Un indicateur de confiance de l'ensemble est déduit de cette caractérisation, et ce pour chaque classe. Cet indicateur est affiché avec les paramètres des classes pour laisser une interprétation des conditions météorologiques. Ce travail a été complété à l'aide des ingénieurs et stagiaires en informatique de Scalian ayant contribué aux multiples éléments d'architecture et de choix de technologies nécessaires pour l'édification de ce prototype.

La section 4.1 introduit les approches logicielles et applications web existantes et justifie le choix de développement d'une application web. La section 4.2 présente les choix technologiques ainsi que l'affichage de codages en couleur représentant les prévisions spatiales et l'affichage de courbes des prévisions selon l'échéance pour des coordonnées spécifiques sélectionnées par l'utilisateur. Enfin, les affichages et fonctionnalités ayant pour but d'intégrer les contributions des chapitres précédents sont proposés dans la section 4.3. Dans cette section, les fonctionnalités permettant la création d'événements météorologiques et l'affichage des probabilités associées sont introduites. La représentation des paramètres et indicateurs issus de la caractérisation des erreurs de distribution des ensembles des classes (ajustées par le modèle de mélange étendu du chapitre 3) est ensuite présentée. Le développement de l'application web étant récent, les données utilisées pour construire les affichages correspondent à des dates plus récentes que celles étudiées dans les chapitres précédents.

4.1 Solutions existantes

À l'heure actuelle, il existe de multiples outils destinés à présenter l'information des prévisions météorologiques. Ces outils se divisent en deux principales catégories : les logiciels et les applications web. Les logiciels comme Panoply (développé par GISS, `https://www.giss.nasa.gov/tools/panoply/`) proposent un panel varié d'outils d'affichage et de traitement applicables aux prévisions. Leur grande différence face aux applications web est qu'ils génèrent un environnement détaillé permettant de travailler à partir de fichiers de prévision sans avoir recours à une connexion internet ni aux langages de programmation. Quant aux applications web, elles offrent plus facilement une interface simple et parcimonieuse en matière de fonctionnalités, adaptable aux différents terminaux (ordinateur, téléphone portable, tablette, etc.), répondant ainsi aux besoins variés d'utilisateurs finaux.

4.1.1 Logiciel

Les logiciels comme Panoply sont largement utilisés dans l'exploration des fichiers de prévision météorologique et l'exploitation spatiales des variables météorologiques. Ces fonctionnalités utilisent de multiples extensions de fichiers comme NetCDF et Grib. Panoply fait partie des logiciels spécialisés dans la création de cartes et de déroulement temporel de celles-ci. Dans le cadre d'utilisation de prévisions d'ensemble, l'apport de M réalisations de champs météorologiques spatiaux pour une même date font apparaître de nouvelles problématiques de représentation. Dans l'optique de mieux représenter l'incertitude des prévisions à l'aide d'ensembles, Panoply montre certaines limites en exploitant les statistiques empiriques des ensembles de manière spatiale uniquement. Un état de l'art est donné dans Rautenhaus et al. 2015 où l'outil Met.3D est dévoilé. Cet outil est capable de représenter l'incertitude des ensembles en utilisant un affichage moderne 3D autour des diverses variables météorologiques. Les inconvénients d'une telle approche logicielle sont souvent liés à la difficulté de la prise en main de l'interface graphique, l'adaptation de l'interface sur divers types d'écrans et la mise à jour des méthodes scientifiques déployées autour des prévisions. Pour pallier ces problèmes, la solution a été le développement d'une application web.

4.1.2 Application Web

La représentation des prévisions météorologiques est un des domaines d'application du secteur informatique sans cesse renouvelée à l'aide des technologies web. En France et depuis 2004, le site Meteociel représente de manière spatiale et temporelle la plupart des produits de services météorologiques incluant prévisions, ensembles de prévision, observations, mais aussi un système d'archivage de ces données. Ce site regorge d'informations autour des prévisions. Cependant, un des points négatifs est l'ergonomie du site qui, bien que très étudiée à l'origine, n'a pas évolué au fil des nouvelles technologies web. Windy, un autre service du web récent et populaire, qui dispose aussi d'une application, offre une évolution de l'ergonomie de l'affichage des prévisions spatiales et temporelles. L'article de Popelka, Vondrakova et Hujnakova 2019 discute de l'importance de l'ergonomie des applications web d'affichage des données météorologiques. Notamment dans leur étude, ils comparent les nouveaux produits tels que Windy face à d'autres sites à l'ergonomie différente et montrent, à l'aide de listes de tâches à réaliser que les utilisateurs passent moins de temps à prendre en main des sites comme Windy.

C'est en prenant comme référence l'ergonomie de Windy que l'application de Scalian prend forme. Cette application innove par l'intégration de méthodes statistiques proposant de nouvelles analyses autour des prévisions et ensembles de prévision météorologique, le tout dans une page web capable de s'adapter à divers types de terminaux.

Dans la section suivante, les principales fonctionnalités et les affichages intégrés par l'application développée par Scalian seront abordés. Ces affichages représentent les prévisions de manière spatiale et temporelle en donnant accès à divers outils d'analyse.

4.2 Développement général

Afin de concevoir une interface utilisateur capable de représenter des prévisions météorologiques et les résultats des traitements, le choix d'une architecture et des techniques informatiques nécessaires à sa conception a dû être réalisé. Ensuite, les fonctionnalités générales comme l'affichage spatial à l'aide de codages en couleur des valeurs de prévision et l'affichage des courbes de prévision selon les échéances de prévision à un point spatial fixé ont été développés.

4.2.1 Architecture de l'application

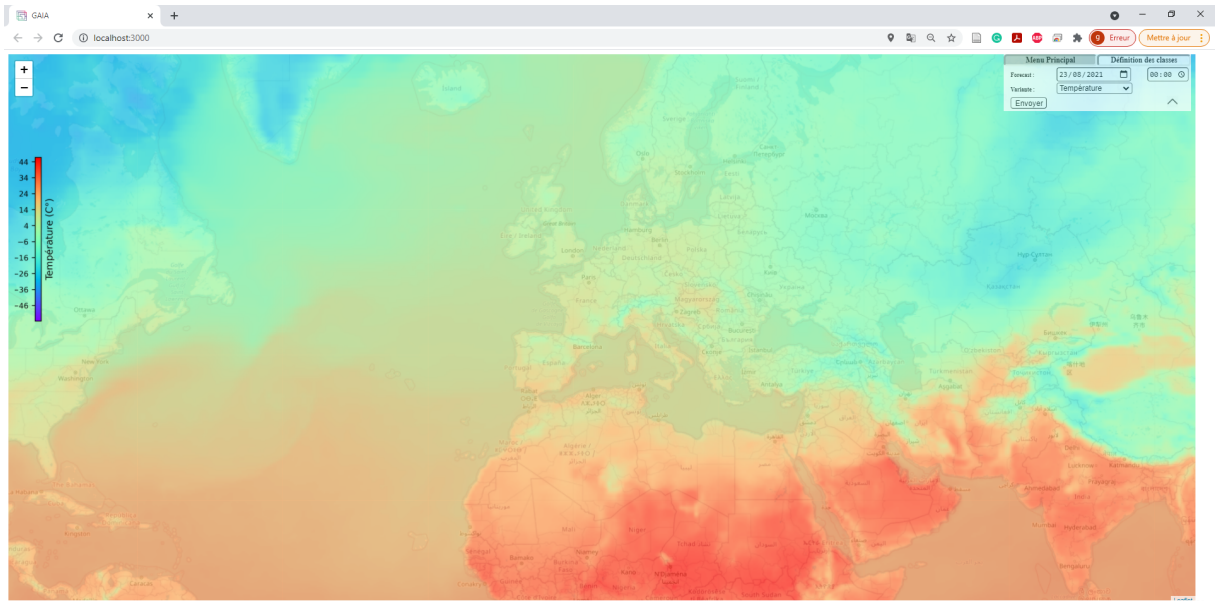
Le développement de l'application a nécessité la mise en place d'une architecture client-serveur. Ce type d'architecture a une partie cachée appelée le serveur opérant les données de prévision et effectuant les traitements les plus coûteux. Cette partie est développée avec le langage **Python** et le module **Flask**. Ces technologies sont libres et relativement aisées à prendre en main. De plus, Flask permet un développement asynchrone de la communication entre le serveur et le client, rendant l'application robuste face à un nombre important d'utilisateurs. Ensuite vient la partie client développée en **Javascript** et utilisant la librairie **React**. Le client contient l'interface principale faisant le lien entre l'utilisateur, les données de prévision et les traitements appliqués. Le développement en Javascript facilite la conception d'outils graphiques dynamique intégrant des interactions utilisateurs de type de zoom, sélection de zones, etc. La librairie React est une technologie récente devenue une référence dans la création de pages web dynamiques capables de s'adapter à des écrans de tailles variées. Le livre de Grinberg 2018 donne les clés du développement en Python et Flask et celui d'Eisenman 2015 aide à la prise en main de React.

4.2.2 Représentations classiques des prévisions

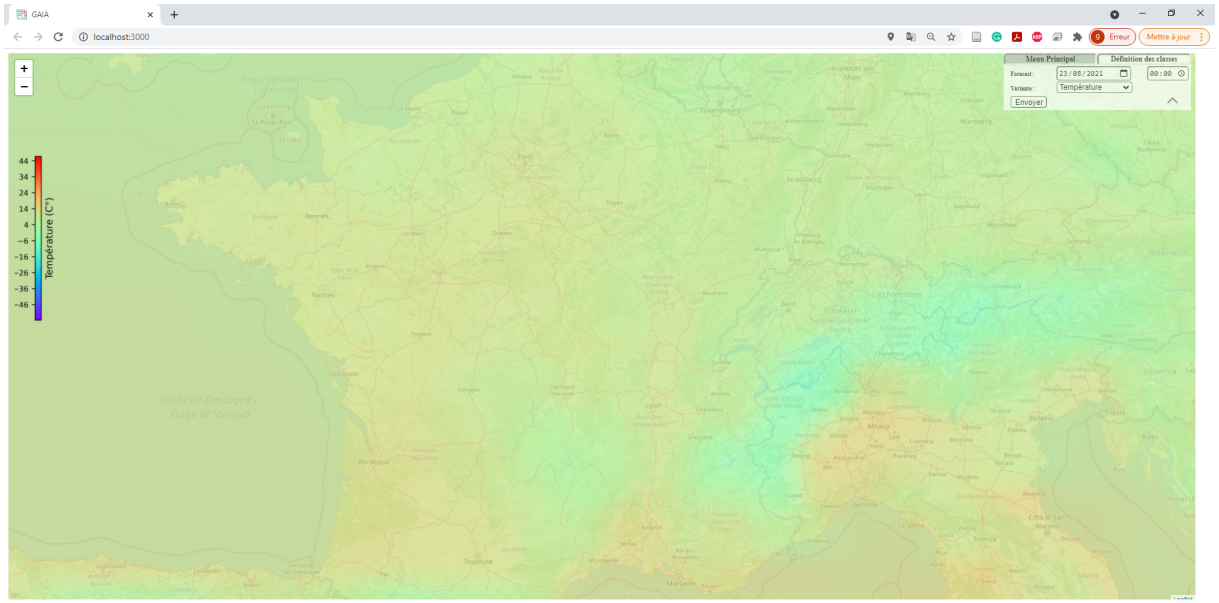
Usuellement, les données de prévision météorologique numérique sont représentées spatialement sur une grille régulière pour un temps fixe où chaque point est lié à une coordonnée longitude-latitude. L'espacement entre les points est défini par la résolution du modèle de prévision. Le codage en couleur des prévisions constitue un affichage représentant les variations de valeurs entre les points de cette grille. Cet outil attribue à chaque point de la grille un pixel de couleur défini à l'aide d'une plage représentant l'espace de valeurs de la grille de la variable météorologique étudiée.

La figure 4.1a affiche le résultat du codage en couleur obtenu pour une grille de prévision de température l'échéance de prévision de 0 jour de l'initialisation du 12/03/2021, superposée à un fond de carte géographique. La figure 4.1b illustre la fonctionnalité de grossissement intégré à la carte à l'aide librairie Leaflet de Javascript. La présentation de ce type d'affichage spatial est une nécessité dans le domaine des sites dédiés à la prévision météorologique. L'intégralité des logiciels d'affichage de prévision citée en section 4.1 possède un outil capable d'en concevoir.

En plus de leur dimension spatiale, les données de prévision météorologique intègrent



(a) Codages en couleurs des prévisions de températures d'une partie du globe et centrées sur l'Europe.



(b) Zoom sur la France.

FIGURE 4.1 – Prévisions initialisées le 12/03/2021 à 12H, codages en couleurs des prévisions de températures à une échéance de 0 jour.

une dimension temporelle d'échéances de prévision dont le pas est standardisé, mais peut dépendre aussi de la résolution du modèle. L'information contenue le long de ces courbes

de prévision selon les échéances est importante d'un point de vue décisionnel pour les utilisateurs. Cependant, elle pose une importante contrainte de limite d'affichage des informations due à sa grande dimension spatiale et temporelle. Contrairement aux codages en couleurs où l'échéance est fixée pour régler le problème de dimensions, ici chaque échéance est reliée afin de former une courbe. Cette courbe permet de donner une idée sur l'évolution temporelle de la prévision étudiée pour une variable météorologique et un point spatial sélectionné. Ce type de représentation a été inspiré par Windy.

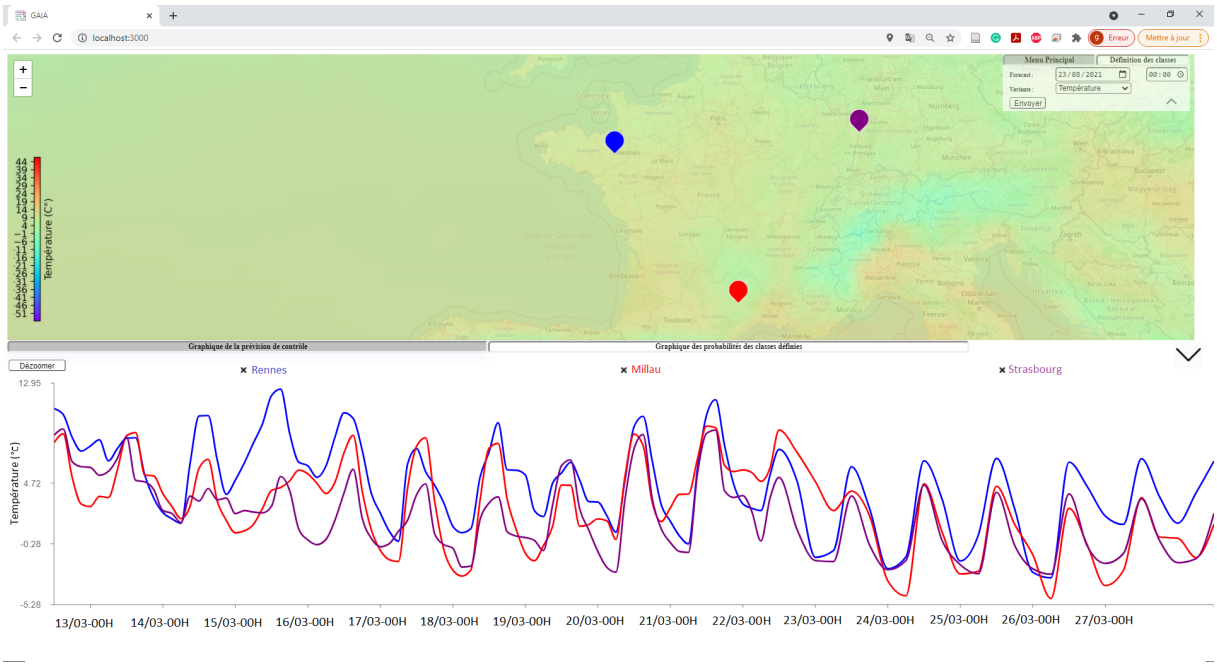


FIGURE 4.2 – Prévisions initialisées le 12/03/2021 à 12H, courbes de prévision de contrôle de la température selon les échéances de prévision espacées de 3H pour les stations de Rennes (en bleu), Millau (en rouge) et Strasbourg (en violet).

La figure 4.2 montre les courbes de prévision déterministe de contrôle de la variable température selon les échéances de prévision pour l'initialisation du 12/03/2021 à 12H par pas de 3H pour trois localisations fixées. Plus particulièrement, la figure 4.2 affiche une fonctionnalité de comparaisons spatiale des courbes d'échéances entre les points sélectionnés près de Rennes, Millau et Strasbourg.

4.3 Création des affichages associés aux contributions

Dans cette section, les différentes représentations graphiques créées dans le but d'intégrer les contributions des chapitres 2 et 3 dans une application web sont abordées. La première partie 4.3.1 présente le menu permettant à l'utilisateur de saisir une coordonnée spatiale et définir des événements météorologiques à partir de seuils appliqués sur les variables météorologiques sélectionnées. Les probabilités d'occurrence de ces événements sont estimées et affichées sur les ensembles, et ce pour les multiples échéances. La partie suivante traite de la représentation des résultats de l'extension du modèle de mélange proposé dans le chapitre 3. À partir de ces résultats, un affichage représentant les paramètres de la classe prédite pour l'ensemble étudié à une échéance et localisation sélectionnée est introduit. De plus, un indicateur de risque d'erreur des ensembles de prévision déduit pour chaque classe est présenté également dans cet affichage.

4.3.1 Affichages des probabilités d'événements météorologiques

Le chapitre 2 a présenté l'objectif de prédiction d'événements météorologiques définis pour une localisation spatiale sélectionnée. Dans ce chapitre, les ensembles issus du modèle numérique et les modèles de classification montrent des capacités de prédiction robustes de variables discrètes caractérisant des conditions météorologiques spécifiques, pour une échéance et localisation fixée. Sur la base de ces résultats, cette partie exploite la création d'événements météorologiques et l'affichage des probabilités prédites par les ensembles issus du modèle numérique dans un premier temps. Cette fonctionnalité, qui répond à un besoin exprimé par Scalian, contribuera au développement d'une application destinée aux particuliers. En premier lieu, une interface a été créée pour permettre à l'utilisateur de définir des événements météorologiques personnalisés par saisie de seuils sur différentes variables.

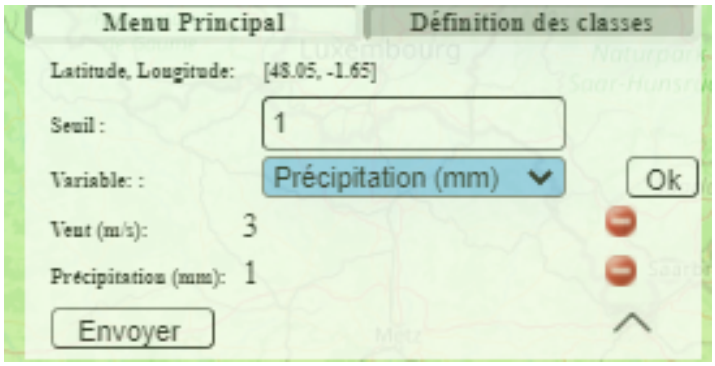
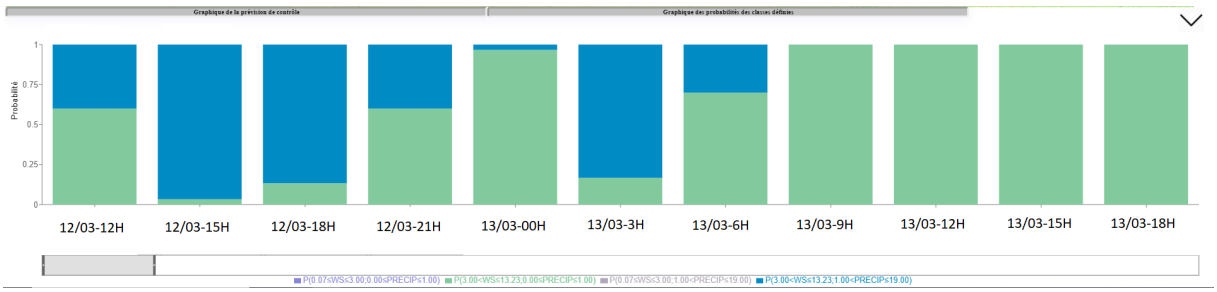


FIGURE 4.3 – Menu utilisateur pour définir des seuils d'études aux coordonnées spatiales sélectionnées (point proche de Rennes).

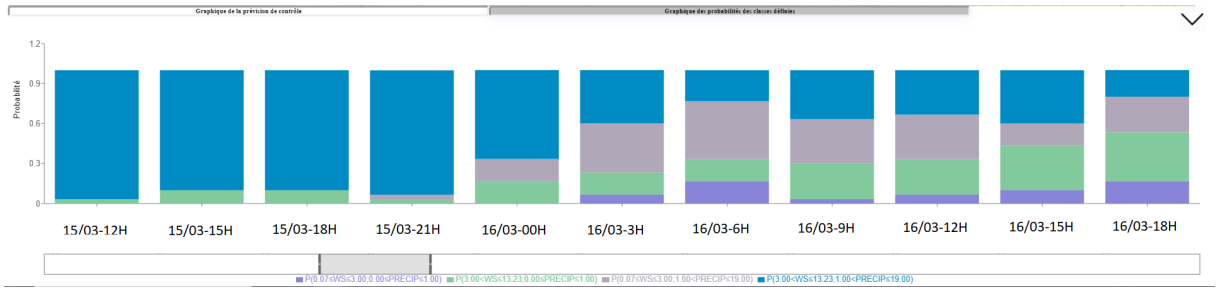
La figure 4.3 présente l'interface de saisie du seuil appliqué à une variable météorologique retenue pour une coordonnée spatiale sélectionnée (dans ce cas le point est situé proche de Rennes). Pour chaque variable météorologique sélectionnée, l'utilisateur ne peut saisir qu'un seuil séparant l'espace des valeurs en deux classes. Il est possible de sélectionner plusieurs variables à la fois. Par exemple, deux seuils sont saisis dans l'interface de la figure 4.3 générant 4 événements autour des variables de vitesses de vent et de précipitation :

1. Météo calme : vitesse vent $\leq$ 3 m/s, précipitations$\leq$ 1 mm ;
2. Météo venteuse : vitesse vent $>$ 3 m/s, précipitations $\leq$ 1 mm ;
3. Météo pluvieuse : vitesse vent $\leq$ 3 m/s, précipitations $>$ 1 mm ;
4. Météo venteuse et pluvieuse : vitesse vent $>$ 3 m/s, précipitations $>$ 1 mm.

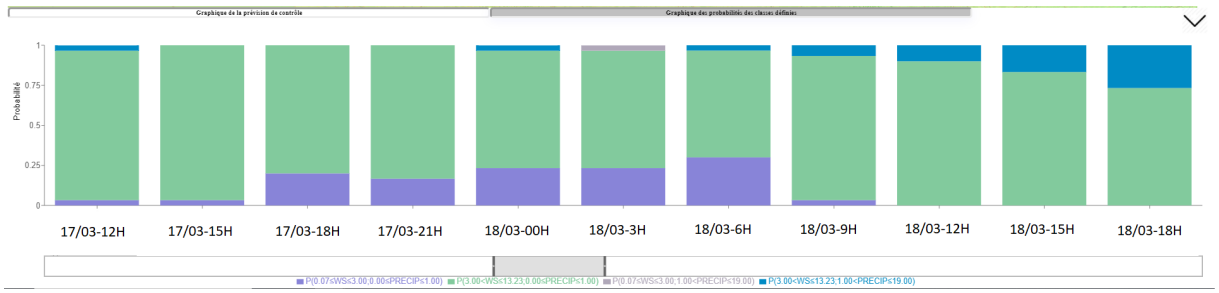
Dès lors que l'utilisateur saisit les variables et seuils souhaités, il peut transmettre ces informations au serveur qui calculera et renverra les probabilités estimées sur les ensembles de prévision.



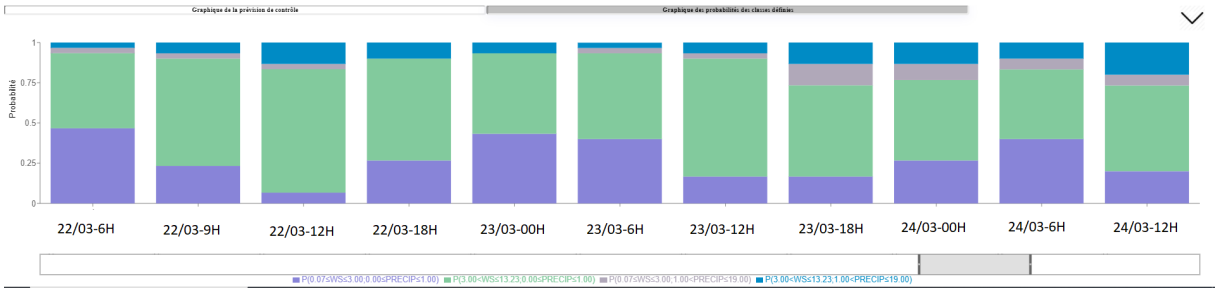
(a) Probabilités des classes pour des ensembles de prévision aux échéances de 0 à 1 jour.



(b) Probabilités des classes pour des ensembles de prévision aux échéances de 3 à 4 jours.



(c) Probabilités des classes pour des ensembles de prévision aux échéances de 5 à 6 jours.



(d) Probabilités des classes pour des ensembles de prévision aux échéances de 10 à 12 jours.

FIGURE 4.4 – Prévisions initialisées le 12/03/2021 à 12H, affichage des probabilités d'occurrences des 4 événements définis et estimées sur les ensembles de prévision à Rennes à différentes échéances. *WS : vitesse du vent, Precip : cumuls de précipitations. Les couleurs représentent les 4 événements météorologiques définis par les seuils :*

- *En violet météo calme,* $P(0.07 < WS \leq 3; 0 \leq PRECIP \leq 1)$ *;*
- *En vert météo venteuse,* $P(3 < WS \leq 13.23; 0 \leq PRECIP \leq 1)$ *;*
- *En gris météo pluvieuse,* $P(0.07 < WS \leq 3; 1 < PRECIP \leq 19)$ *;*
- *En bleu météo venteuse et pluvieuse,* $P(3 < WS \leq 13.23; 1 < PRECIP \leq 19)$.

Les figures 4.4a, 4.4b, 4.4c et 4.4d donnent un exemple d'affichage des probabilités d'occurrences de classes à l'aide de graphiques bâtons construits sous différentes échéances de prévision pour Rennes. Chaque portion colorée de bâton représente la va-

leur de la probabilité associée à un événement spécifique. La légende des couleurs est construite en notant la probabilité de l'événement en encadrant les noms de variables par le seuil et le minimum (ou le maximum suivant l'orientation de l'inégalité) des valeurs de l'ensemble. Dans l'exemple de la figure 4.3, l'événement de météo venteuse se note $P(3 < WS \leq 13.23; 0 \leq PRECIP \leq 1)$ où 13.23 est le maximum de valeur de vitesse de vent relevé dans l'ensemble et 0 le minimum de précipitations. Désormais, il est possible d'interpréter les prédictions de probabilités fournies par l'ensemble pour chaque événement et échéances à la coordonnée sélectionnée. Un exemple d'interprétation des probabilités prédites aux différentes échéances est que les figures 4.4a et 4.4b semble indiquer des événements oscillant entre une météo "Venteuse" (en vert) et "Pluvieuse" (en bleu) pour ensuite rentrer dans une zone de perturbations à 4 jours avec plusieurs probabilités différentes s'élevant. Les figures 4.4c et 4.4d montrent les probabilités d'une météo "Venteuse" majoritaire aux échéances de 5 et 10 jours. Il est également possible d'apercevoir pour des échéances supérieures à 10 jours que les probabilités d'événements, autres que celui majoritaire, apparaissent. La distribution des ensembles étant très étendue à partir de ces moyennes échéances, tends à contenir les multiples événements définis au sein des réalisations d'ensembles.

Les visualisations présentées dans cette partie génèrent des événements dépendant de la météorologie locale et exploitent l'information contenue dans les ensembles. Néanmoins, elles restent au stade de prototype et n'exploitent que les probabilités issues des ensembles du modèle numérique. L'intégration de l'apport des modèles de classification à la prédiction de ces probabilités d'événements reste à développer et constitue la prochaine évolution nécessaire pour valoriser les contributions du chapitre 2.

4.3.2 Représentation des informations de sous-groupes d'ensembles similaires

Dans le chapitre 3, il a été question de proposer une extension modèle de mélange capable de former des sous-groupes d'ensembles. Les sous-groupes formés permettent d'obtenir des informations sur les différentes conditions météorologiques disponibles dans les données étudiées, résumées dans les paramètres de moyennes et écarts types caractérisant les sous-groupes. En utilisant des observations météorologiques et des tests d'hypothèses, il est possible d'évaluer les types d'erreurs liés à ces sous-groupes. Représenter les informations issues des sous-groupes et des résultats de tests d'hypothèses d'ensembles et

observations révolues permettent de fournir à l'utilisateur des moyens pour mieux interpréter les conditions météorologies et possibles incertitudes associées à l'ensemble de prévision étudié. La suite de cette partie est consacrée à la présentation d'une solution d'affichage des sous-groupes prédits pour un nouvel ensemble de prévision ainsi que les paramètres associés. De plus, un critère de risque d'erreur créé pour faciliter l'interprétation des résultats de tests d'hypothèses est présenté.

La figure 4.5 donne un aperçu du type d'affichage créé pour représenter les résultats de prédiction de sous-groupes, sur une tranche d'échéance allant de 3 à 4 jours d'un ensemble initialisé le 15/05/2021 à 18H pour Rennes. L'ensemble par échéance est représenté par dix déciles sous la forme d'un nuage de niveau de gris ; plus les déciles sont proches et plus le gris du nuage s'intensifie. Pour développer ce prototype d'affichage, l'extension du modèle de mélange a été entraînée sur 3 mois de données précédents cette date, soit un cadre d'entraînement légèrement différent de celui qui a été étudié dans les chapitres précédents. Trois sous-groupes sont ainsi ajustés sur les ensembles des variables météorologiques de température à 2 mètres au-dessus du sol et de composantes de vent (U,V) à 10 mètres au-dessus du sol. Le sous-groupe prédit pour chaque échéance est représenté sous la forme de fond de couleur en arrière plan des déciles des ensembles. Néanmoins, le modèle de mélange étant ajusté de façon indépendante entre échéances, une étape d'harmonisation des sous-groupes a été nécessaire pour obtenir une meilleure capture des changements de météorologie au sein des sous-groupes et agréger ceux dont les conditions météorologiques sont similaires. Cette étape utilise la similarité utilisée dans le Shaake shuffle de la section 1.2.2 et définie par Schefzik 2016. La similarité utilise les moyennes et variances entre sous-groupes de différentes échéances pour définir une mesure et ainsi ordonner les sous-groupes suivant leur proximité de paramètres et ce par échéance. Dès lors la persistance des sous-groupes de conditions météorologiques similaires est mise en valeur comme sur la figure 4.5 suivant les échéances et les changements sont d'autant plus marqués, et visible sur la figure 4.6.

Plus encore, les tests d'hypothèses appliquées aux ensembles et observations univariés de la variable météorologique étudiée évaluent les typologies d'erreurs de distributions. Pour rappel, les deux tests, introduits au 3.2.1.2, posent une hypothèse nulle H_0 d'égalité des moyennes et une pour l'égalité des variances entre la distribution des ensembles de prévision et celles des observations de la variable météorologique étudiée. Le rejet d'une hypothèse H_0 de ces deux tests suivant le seuil de significativité α sélectionné entraîne la caractérisation d'un type d'erreur de distribution des ensembles. Cependant, afficher

les types d'erreurs détectées ou les p-valeurs de ces tests peut conduire à des difficultés d'interprétation pour des utilisateurs ne sachant pas comment interpréter les tests d'hypothèses ou les différentes typologies d'erreurs de distributions des ensembles. L'idée est de résumer les conclusions de ces tests en un indicateur facile à interpréter caractérisant l'état de l'ensemble ou pour être plus précis, caractérisant le risque de présence d'erreurs au sein de la distribution des ensembles. En se basant sur les rejets d'hypothèses nulles H_0 des tests avec le seuil de significativité α fixé à 0.05, l'indicateur de "risque d'erreur" est défini au travers de trois catégories, pour la variable météorologique étudiée :

1. \+ Faible : aucune hypothèse H_0 rejetée ;
2. \- Fort : une hypothèse H_0 rejetée ;
3. – Très fort : deux hypothèses H_0 rejetées ;

Il est important de noter que le "risque d'erreur" est construit à partir de rejets d'hypothèses nulles de tests possédant donc une marge d'erreur suivant le seuil de significativité α. De plus, cet indicateur ne peut pas conclure à l'absence d'erreur au sein de la distribution des ensembles, mais juste à la présence d'erreurs caractéristiques touchant la moyenne et la dispersion des distributions des ensembles de prévision. Néanmoins, ce genre d'indicateur peut renforcer la confiance de l'utilisateur dans les ensembles de prévision appartenant au sous-groupe étudié.

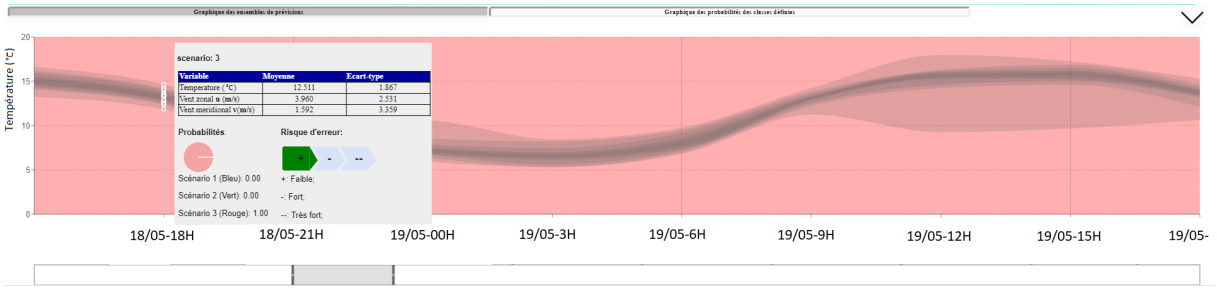
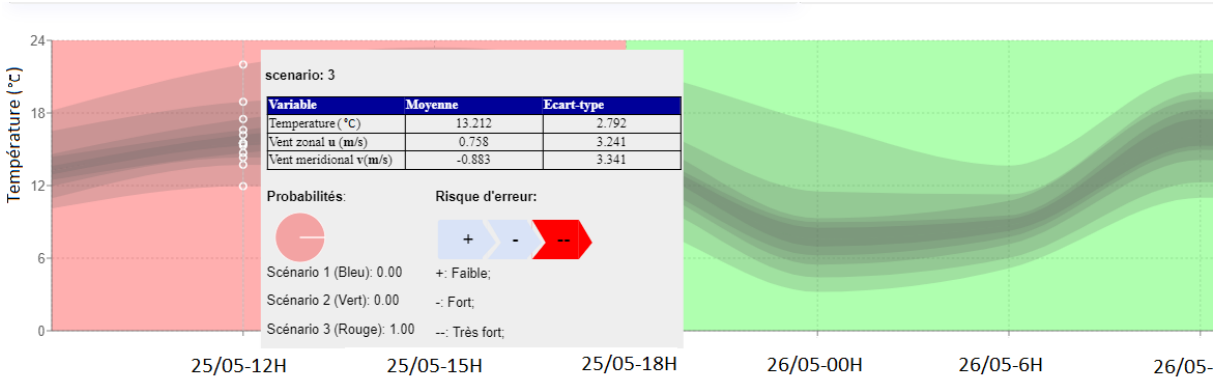


FIGURE 4.5 – Ensemble de prévision de températures à Rennes initialisé le 15/05/2021 à 18H, suivant les échéances de prévision de 3 à 4 jours avec une représentation de l'indicateur du risque d'erreur de l'ensemble et paramètres du sous-groupe prédit pour l'échéance du 18/05 18H. *Le nuage gris représente dix courbes des déciles de l'ensemble, l'espacement entre ces déciles fait varier l'intensité de gris passant du clair pour un nuage étiré au sombre pour un nuage étroit. Le fond de couleur représente les groupes ou scénarios issus du modèle de mélange (en bleu le sous-groupe 1, vert le sous-groupe 2 et rouge le sous-groupe 3). Les paramètres et résultats d'indicateur d'erreurs de ces groupes sont représentés par échéance dans un tableau dédié.*

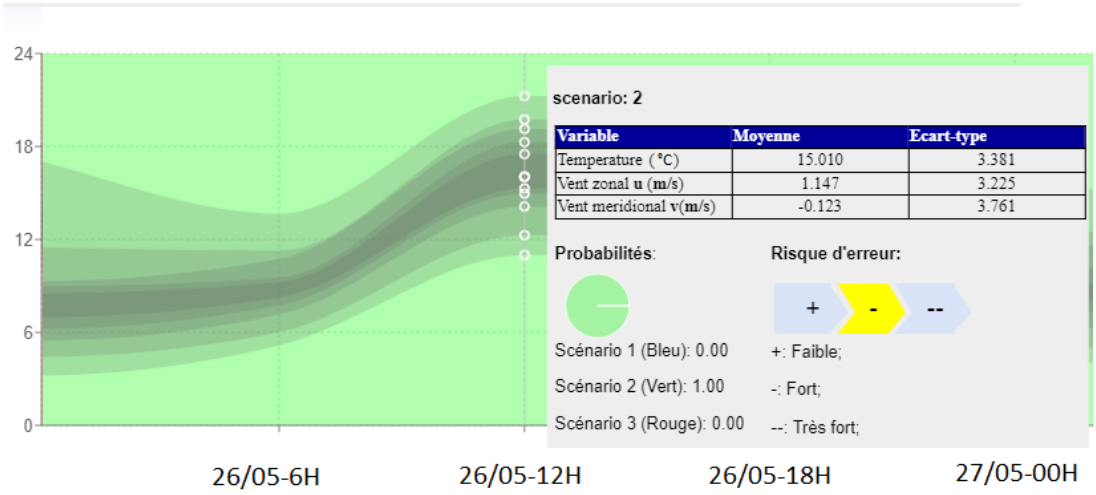
Dans la figure 4.5, le sous-groupe, "scénario 3" codé en rouge, est prédit pour les en-

sembles sur différentes échéances consécutives. En regardant l'ensemble à l'échéance du 18/05 18H, soit 3 jours après l'initialisation, les probabilités de prédiction de sous-groupes affichées dans le diagramme montrent que le modèle est unanime dans l'attribution du "scénario 3". La prédiction du scénario de manière unanime par le modèle laisse présager une certaine persistance des conditions météorologiques caractérisant ce scénario. Le tableau de paramètres affiche les conditions météorologiques liées au scénario 3 pour cette échéance, ce sous-groupe contient des ensembles avec un écart type de 1.8 °C et une moyenne proche de 12.5 °C associée à fort vent du Sud-Ouest (composantes U et V positives). Les échéances adjacentes montrent également des ensembles de température avec un écart type caractérisé par le faible écart entre les déciles formant le nuage de niveau de gris et une moyenne oscillant autour de 12.5 °C.

Sur le panneau d'information où se trouve le tableau, l'indicateur de risque d'erreur est représenté sous forme de pictogramme, avec une flèche colorée allant du vert (pour "+ Faible") au rouge (pour "– Très fort") en passant par l'orange (pour "- Fort") pour terminer. Dans le cas de l'échéance du 18/05 18H, le risque d'erreur indique la catégorie "+ Faible". Ce résultat peut s'interpréter comme la non-détection d'erreurs caractéristiques de distributions des ensembles de températures du sous-groupe 3 à cette échéance. Cet indicateur accroît la confiance dans les interprétations des conditions météorologiques et informations statistiques issues de l'ensemble.



(a) Exemple d'affichage des informations du sous-groupe 3 à l'échéance du 25/05 à 12H.



(b) Exemple d'affichage des informations du sous-groupe 2 à l'échéance du 26/05 à 12H.

FIGURE 4.6 – Ensemble de prévision de températures à Rennes initialisé le 15/05/2021 à 18H, suivant les échéances de prévision de 10 à 11 jours avec une représentation de l'indicateur du risque d'erreur de l'ensemble et paramètres du sous-groupe associé. *Le nuage gris représente dix courbes des déciles de l'ensemble, l'espacement entre ces déciles fait varier l'intensité de gris passant du clair pour un nuage étiré au sombre pour un nuage étroit. Le fond de couleur représente les groupes ou scénarios issus du modèle de mélange (en bleu le sous-groupe 1, vert le sous-groupe 2 et rouge le sous-groupe 3). Les paramètres et résultats d'indicateur d'erreurs de ces groupes sont représentés par échéance dans un tableau dédié.*

Pour prendre un autre exemple, les figures 4.6a et 4.6b montrent deux cas d'affichage de paramètres de différents sous-groupes aux échéances de 25/05 et 26/05 à 12H, soit entre 10 et 11 jours après l'initialisation du modèle d'ensemble du 15/05/2021 à 18H. La figure 4.6a présente des ensembles appartenant au sous-groupe 3 de moyenne 13.2 °C avec un écart type assez fort de 2.8 °C, conjugué à un vent faible provenant du Nord-Ouest et également de fort écart type. L'indicateur de risque d'erreur affiche sa plus forte catégorie de "– Très fort" laissant présager d'importantes déviations entre les prévisions météorologiques et la réalité. Comparé à la figure 4.5, il n'est pas surprenant de trouver à l'échéance de 10 jours un sous-groupe de même indice que celui à l'échéance de 3 jours affichant des paramètres et une catégorie de risque d'erreur différents. L'augmentation des échéances fait évoluer de manière générale les paramètres des trois sous-groupes ajustés à chaque échéance par le modèle.

Ensuite, à l'échéance de prévision du 26/05 à 12H, la figure 4.6b montre un changement de prédiction de sous-groupe pour les ensembles. Le nouveau sous-groupe prédit est le "Scénario 2" affichant des ensembles de prévision avec une température moyenne de 15.1 °C et avec un écart type de 3.38 °C, associés à un vent d'Ouest dominant. Pour ce sous-groupe, la catégorie de l'indicateur du risque d'erreur diminue d'un cran montrant "- Faible". En comparant les figures 4.6a et 4.6b, il ressort qu'à partir des échéances supérieures à 10 jours, les ensembles de prévision subissent une augmentation de la température moyenne et de l'écart type.

Ce risque d'erreur élevé témoigne de l'importante incertitude des distributions d'ensembles de prévision aux moyennes échéances. Il est possible d'observer une légère réduction du risque d'erreur entre l'échéance de la figure 4.6a et celle de la figure 4.6b. Cette réduction affichée par l'indicateur du risque d'erreur peut paraître contre-intuitive étant donné que l'échéance évolue. Une interprétation possible de l'augmentation de température et l'évolution du risque d'erreur sont potentiellement en lien avec le changement de vent entre les deux sous-groupes espacés d'un jour d'échéance de prévision. Le vent caractéristique du sous-groupe 3 de l'échéance du 25/05 à 12H provient du Nord-Ouest, un vent surnommé le "Noroît", souvent associé au passage de perturbations atmosphériques et à un risque d'erreur de prévision accru. Ensuite, la transition du sous-groupe 3 vers le sous-groupe 2 s'accompagne d'un vent d'Ouest pouvant indiquer une sortie de perturbation et donc une accalmie avec augmentation des températures et diminution du risque d'erreur.

À l'issue de cette partie, des exemples des paramètres de sous-groupes ajustés par l'extension du modèle de mélange présenté dans le chapitre 3 ont été présentés. La représentation de ces paramètres par échéance donne d'avantage d'informations à l'utilisateur quant aux conditions météorologiques regroupant les ensembles, et ce pour plusieurs échéances. De plus, un indicateur de risque d'erreur de distribution des ensembles de prévision de la variable météorologique étudiée a été introduit ainsi que sa représentation graphique. Les exemples analysés dans cette section soulignent l'intérêt de ce risque d'erreur, permettant à l'utilisateur de porter un regard critique sur la prévision d'ensemble. Un point non abordé ici et proposé dans le chapitre 3 est l'extension du modèle linéaire servant à la correction des erreurs de distribution des ensembles. Cette extension a montré des performances intéressantes de correction de l'erreur des ensembles, que les travaux futurs se chargeront de valoriser dans l'application web. L'utilisateur aurait le choix de représenter les ensembles du modèle de prévision numérique ou ceux post-traités selon l'extension du

modèle linéaire.

4.4 Conclusion

Dans ce chapitre, les différentes étapes de développement d'un prototype d'application web d'affichage des prévisions et ensembles de prévision à destination de professionnels ou particuliers ont été présentées. Le prototype proposé permet à l'utilisateur de bénéficier d'un suivi global des prévisions par la représentation spatiale en codage de couleur. L'utilisateur dispose également d'un suivi local en affichant les courbes de prévision par échéances, pour une ou des coordonnées spatiales sélectionnées. L'utilisateur est également libre d'étudier les probabilités d'occurrences d'événements liés à des conditions météorologiques spécifiques saisies pour une localisation sélectionnée. Enfin, des informations supplémentaires caractérisant la météorologie et les risques d'erreurs reliées à l'ensemble de prévision étudié sont fournies à l'utilisateur afin d'accroître son niveau décisionnel.

La figure 4.7 montre un exemple de représentation générale du prototype d'application web développée par Scalian pour des prévisions et ensembles de prévision du Centre Européen CEPMMT initialisé le 15 Mai 2021 à 18H. Dans cet exemple, un codage de couleur représente la prévision numérique de température spatialement répartie à l'échéance 0 soit la date d'initialisation du modèle. Dans le panneau en bas de la figure, les courbes de déciles des ensembles de températures sont affichés sous la forme d'un nuage de niveau de gris pour une tranche d'échéances du 25/05 12H jusqu'au 27/05 12H, soit les prévisions de 10 à 12 jours après l'initialisation. Les prévisions affichées dans le panneau en bas de la figure sont issues d'une coordonnée spatiale sélectionnée (point marqué en bleu foncé) proche de Rennes. Sur ce graphique, une barre glissante permet de naviguer au travers des échéances. Enfin, les paramètres découlant du sous-groupe ajusté par le modèle de mélange, ainsi que l'indicateur de risque d'erreur associé à ce sous-groupe sont représentés dans ce panneau.

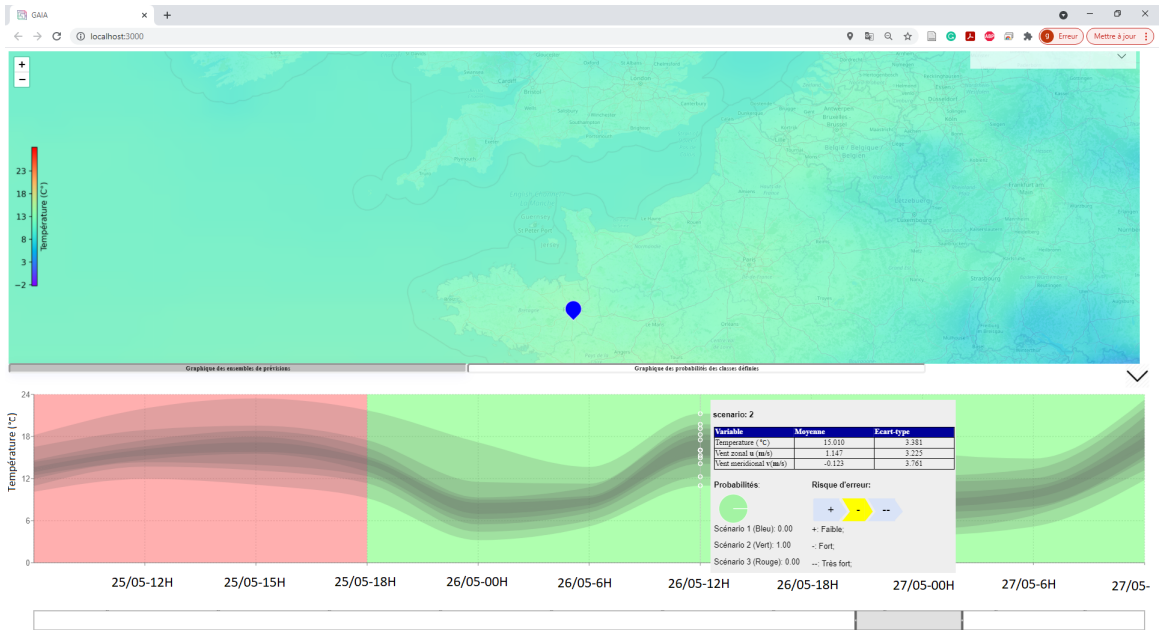


FIGURE 4.7 – Exemple d'affichage des prévisions et ensembles de prévision CEPMMT initialisé le 15 Mai 2021 18H.

Les informations et fonctionnalités fournies par le prototype développé pose une première pierre dans l'objectif de construction d'une application plus aboutie autour des prévisions et ensembles de prévision météorologique. Notamment, uniquement les prototypes d'affichage de résultats ou informations basés sur les ensembles de prévision du modèle numérique ont pu être présentés dans ce chapitre. La valorisation des modèles de classification offrant des résultats performants décrits dans le chapitre 2 fera l'objet de futurs travaux d'intégration à la chaîne de traitement. De la même manière, l'extension du modèle linéaire appliqué à la correction des erreurs de distributions des ensembles a montré de bons résultats dans le chapitre 3 et ne fait pas encore partie de la chaîne de traitement. Une évolution envisageable consisterait à afficher de manière automatique les ensembles post-traités par le modèle linéaire en laissant la possibilité à l'utilisateur d'afficher en parallèle les ensembles du modèle numérique.

La version actuelle ne permet que l'affichage du dernier ensemble de prévision fourni par le service météorologique avec trois variables météorologiques (température, précipitation, vitesse du vent). Ce nombre de variables limité sera augmenté dans les futures mises à jour, offrant aux utilisateurs un plus large panel d'étude et création d'événements. Pour permettre aux utilisateurs de comparer les prévisions révolues, la possibilité d'accéder

aux archives de prévision disponibles est dans les plans d'évolution du prototype. Enfin, l'ergonomie et le design seront également repensés pour améliorer le confort d'utilisation de l'application.

Pour résumer :

- Une interface web a été développée pour afficher des prévisions météorologiques à moyen terme.
- Inspirées par les travaux du chapitre 2, des fonctionnalités ont été développées, permettant aux utilisateurs de créer des événements météorologiques et d'y afficher leurs probabilités d'occurrences estimées à partir des ensembles, et ce suivant les échéances de prévisions.
- Des informations sur la qualité des ensembles de prévision sont également représentées en s'appuyant des travaux du chapitre 3.

Conclusion

Tout au long de ce manuscrit, les erreurs de distribution des ensembles de prévision ont fait l'objet de différentes études. Le chapitre 1 a présenté les algorithmes standard sélectionnés pour corriger des problèmes de calibration univariée et multivariée. Les résultats de ces modèles appliqués aux données d'ensemble de prévision moyen terme du centre Européen (CEPMMT), ont montré des apports non négligeables dans un contexte de calibration univariée des précipitations. Pour les vitesses de vent, les résultats des modèles se sont avérés équivalents ou légèrement meilleurs en moyenne que ceux des ensembles issus du modèle numérique. Dans le cadre de correction des erreurs des ensembles multivariés, l'approche non paramétrique a permis d'obtenir les meilleurs résultats de calibration en moyenne. Néanmoins, ces résultats restaient très proches de ceux des ensembles du modèle numérique témoignant de la difficulté de corriger les erreurs de distribution multivariée, et ce de manière significative.

Sur cette base, une nouvelle approche est proposée dans le chapitre 2. L'idée est de discrétiser l'ensemble de données de façon à transformer le problème de régression multi-sorties, appliquées à la calibration multivariée, en un problème de classification supervisée à sortie unique. Deux méthodes de classification sont ensuite étudiées. La première est une approche naturelle appliquant directement des modèles classiques de prédiction de classes. Cette approche peut être mise en difficulté par la combinaison du problème de classification et celui de correction des erreurs de distribution des ensembles de prévision. De ce fait, une seconde approche est proposée, déduisant une prédiction de classe issue d'ensembles post-traités par les modèles de calibration multivariée. Enfin, ces méthodes sont appliquées à des données réelles. Les modèles relevant de l'approche de classification directe ont montré de bonnes performances de classification. Leurs résultats dépassent ceux des classes prédites par les ensembles issus de calibration multivariée. L'écart de résultats peut être expliqué par la différence entre l'objectif de classification et celui de calibration. En effet, les classes construites sur les observations dans un but de planification d'événement font abstraction du lien réel existant entre l'erreur et le contexte météorologique contenu dans les ensembles de prévision. De plus, l'interprétation physique des résultats des modèles de classification directe reste complexe.

Les travaux du chapitre 3 proposent une méthode différente, partant de l'hypothèse que les erreurs des ensembles dépendent de régimes météorologiques. Une extension du modèle de calibration univariée est introduite en deux étapes : la première étape de classification vise à identifier les régimes météorologiques, la seconde à corriger la distribution d'ensemble dans chaque régime. L'identification des régimes météorologiques est effectuée en utilisant des ensembles de prévision regroupés, de façon non supervisée, à l'aide d'un modèle de mélange gaussien. Les données d'ensemble étant des individus atypiques, elles exigent des méthodes d'ajustement non conventionnelles. Pour cela, trois extensions du modèle de mélange gaussien ont été proposées. Ensuite, ces modèles proposés ont fait l'objet d'une évaluation dans une étape de simulation. Dans cette étape, différentes expérimentations au paramétrage de mélange et de données ont été effectuées. Les résultats obtenus montrent que deux méthodes fournissent une estimation particulièrement précise : l'une basée sur les statistiques empiriques et l'autre sur le vecteur de variables échangeables des membres de l'ensemble. Le modèle de mélange gaussien basé sur le vecteur de variables échangeables, ayant montré des performances légèrement supérieures, a été sélectionné pour l'application aux données réelles. Ensuite, l'extension du modèle utilisant les classes issues du modèle de mélange sélectionné est étudiée et comparée à un modèle NGR sans classe dans un cadre de données réelles. Les conclusions de cette étude montrent un réel apport des classes, dans la calibration univariée, mais aussi dans l'interprétation des corrections apportées. Les classes représentent différentes discriminations de l'erreur et des régimes météorologiques.

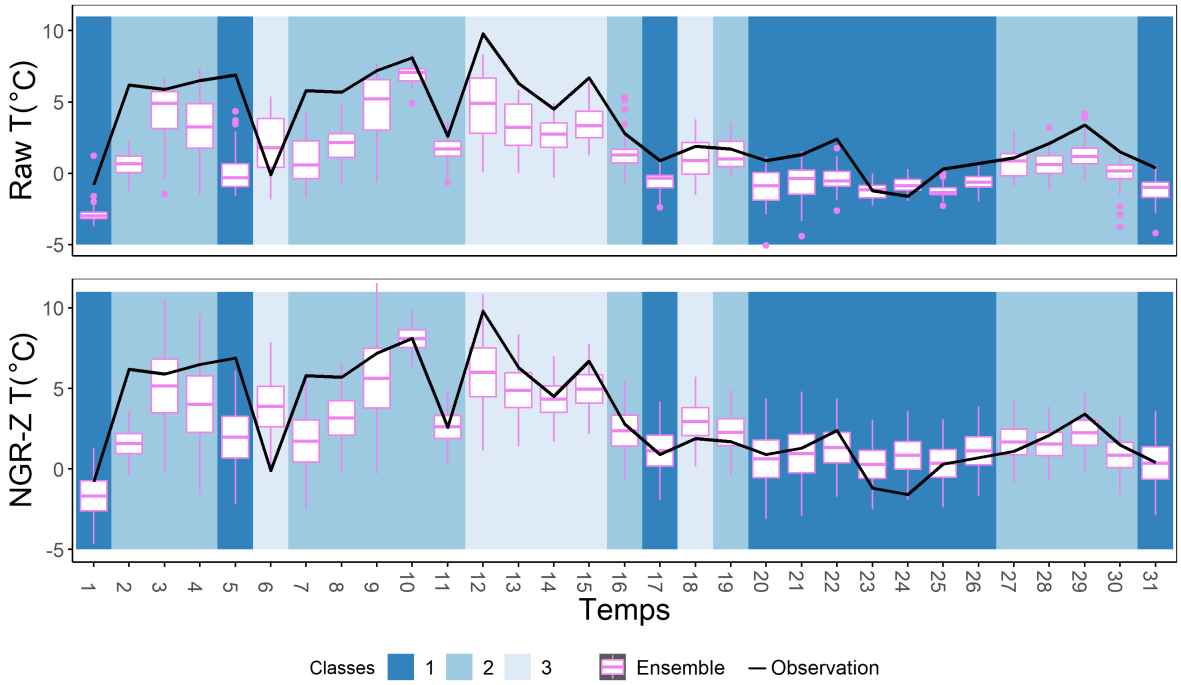


FIGURE 4.8 – Température à Millau, Janvier 2015 à 18H. *Première ligne : CEPMMT ensembles de prévision "Raw" (échéances de 3 jours) ; Seconde ligne : ensembles de prévision calibrés par le modèle NGR_Z. La ligne noire affiche les observations. Les nuances de couleurs en fond représentent les classes ajustées par le modèle GMM_{multi}.*

Enfin, Scalian propose de développer un outil d'affichage des prévisions et ensembles de prévision à destination des entreprises et particuliers. Ce prototype, présenté dans le chapitre 4, intègre différents aboutissements issus des chapitres précédents. En particulier, les fonctionnalités développées pour ce prototype permettent à l'utilisateur de définir des types de météorologies à partir de conditions spécifiques. Ensuite, les probabilités d'occurrence de ces types de météorologie sont affichées. Dès lors, l'utilisateur bénéficie des informations nécessaires pour l'aider à la planification d'événement. De plus, il est possible de représenter les courbes d'ensembles suivant les échéances avec les classes associées et prédites à partir du modèle de mélange étendu présenté précédemment. Les paramètres des classes, ainsi qu'un indicateur issu de la caractérisation des types d'erreurs sont présentés pour chaque échéance et localisation spatiale sélectionnée par l'utilisateur. Ces informations aident à évaluer la qualité des ensembles de prévision représentés.

Les chapitres précédents ont dégagé plusieurs axes de travaux futurs :

— **Temporalité des prévisions et ensembles de prévisions**. Les conclusions des chapitres 2 et 3 discutent de l'importance de considérer l'aspect temporel des don-

nées dans les approches proposées. La figure 4.8 illustre cet aspect à travers les séries d'observations, classes latentes et la distribution des ensembles de prévisions affichées à l'aide de boîtes à moustaches. Dans cette figure, les classes latentes représentées montrent une persistance temporelle avec une météorologie et type d'erreur différente entre les classes. Pour rester dans le cadre d'ajustement d'une variable qualitative, une approche envisagée serait d'utiliser une chaîne de Markov cachée pour modéliser les transitions entre classes et décrire une évolution temporelle des ensembles de prévision. Ensuite, pour reprendre les travaux de Möller et Gross 2016 et continuer sur l'axe des méthodes NGR étendues, un modèle autorégressif pourrait s'ajouter au modèle NGR pour la calibration univariée des ensembles de température dans les états décrits par la chaîne de Markov cachée.

— **Extensions des modèles de mélange gaussien proposés**. L'extension du modèle de mélange gaussien pour un vecteur de variables échangeables est restreinte par une hypothèse d'indépendance entre membres. Or cette hypothèse est discutable. Les travaux de O'Neill 2009 portant sur l'échangeabilité de variables aléatoires revient sur cette hypothèse. Il exprime le fait qu'il est difficile de maintenir, dans la réalité, la notion d'indépendance entre membres amplement utilisée et admet une certaine covariance existante entre variables échangeables. Pour de futurs travaux, la relaxation de cette hypothèse en considérant une structure de covariance entre membres serait intéressante, afin de complexifier le modèle aux variables échangeables et se rapprocher du cas réel. Également, d'autres distributions asymétriques ou disposant de longues queues pourraient être considérées pour conduire l'ajustement des ensembles échangeables sur d'autres variables météorologiques comme les précipitations. Pour finir sur les améliorations des travaux du chapitre 3, il est tout à fait possible que d'autres modèles de calibration puissent être envisagés pour un couplage avec les classes formées. Le modèle paramétrique 'Bayesian Model Averaging' (BMA) introduit par Raftery et al. 2005 pourrait être notamment utilisé. Cette approche offre plus de flexibilité que le modèle linéaire NGR en appliquant une méthode de calibration linéaire sur chaque membre des ensembles du sous-groupe formé. Mais encore, des approches non paramétriques comme les modèles de forêt aléatoire employés dans le chapitre 2 ou des réseaux de neurones de Rasp et Lerch 2018 ou Bremnes 2020 pourraient être appliqués à la calibration des sous-groupes formés par le modèle de mélange.

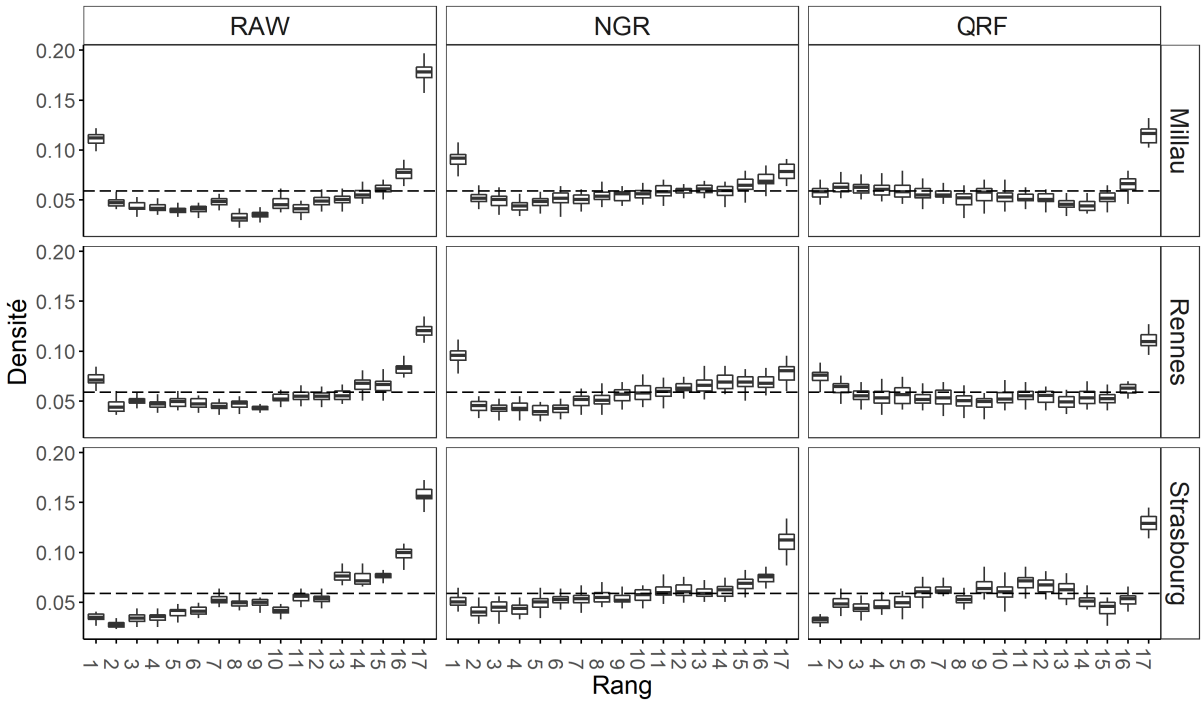
— **Application des modèles de mélange gaussien proposés sur les dimensions spatiales et temporelles**. Les dimensions spatiales et temporelles (échéances de prévisions) des ensembles ont été peu considérées dans les modèles proposés dans cette thèse. Concernant la dimension spatiale, la méthode de Schaake shuffle utilisé pour approcher les dépendances entre variables météorologiques a déjà fait ses preuves (Schefzik et Möller 2018) et pourrait donc être adaptée. Dans le cas temporel, modéliser la trajectoire de distribution des ensembles demande de composer les erreurs de prévision évoluant avec les échéances et l'aspect météorologique des données. En effet, à courte échéance, les ensembles ont tendance à être sous-dispersés puis en fin de trajectoire à finir sur-dispersés. L'approche de Scheuerer, Hamill et al. 2017 basée sur l'algorithme de Schaake shuffle redéfinit la similarité (1.25) proposée par Schefzik 2016. La nouvelle mesure prend en compte les aspects de distribution des ensembles. Cette extension de l'algorithme Schaake shuffle permettrait plus facilement de prendre en compte les différentes dimensions spatiales et temporelles en conservant au mieux la structure multivariée de la distribution des ensembles.

— **Cadre expérimental pour la simulation de types d'erreurs de distribution d'ensembles**. Dans la littérature de calibration, il existe peu de travaux concernant la simulation de différents types d'erreurs de distribution d'ensembles de prévision à partir d'un modèle numérique simplifié. Néanmoins, il existe quelques pistes. Ainsi, Daniel S Wilks 2005 propose de simuler un modèle de prévision numérique et un modèle d'ensemble de prévision simplifié approchant le système physique de Lorenz 1996 (L96). Le cadre ainsi proposé permet de simuler des ensembles sous-dispersés. Cette méthode a la particularité d'être flexible et relativement plus simple à manipuler que les modèles de prévision numérique déployés par les services météorologiques. Il serait intéressant de proposer d'autres schémas de paramètres du cadre expérimental posé par Daniel S Wilks 2005 pour simuler d'autres types d'erreurs de distribution des ensembles.

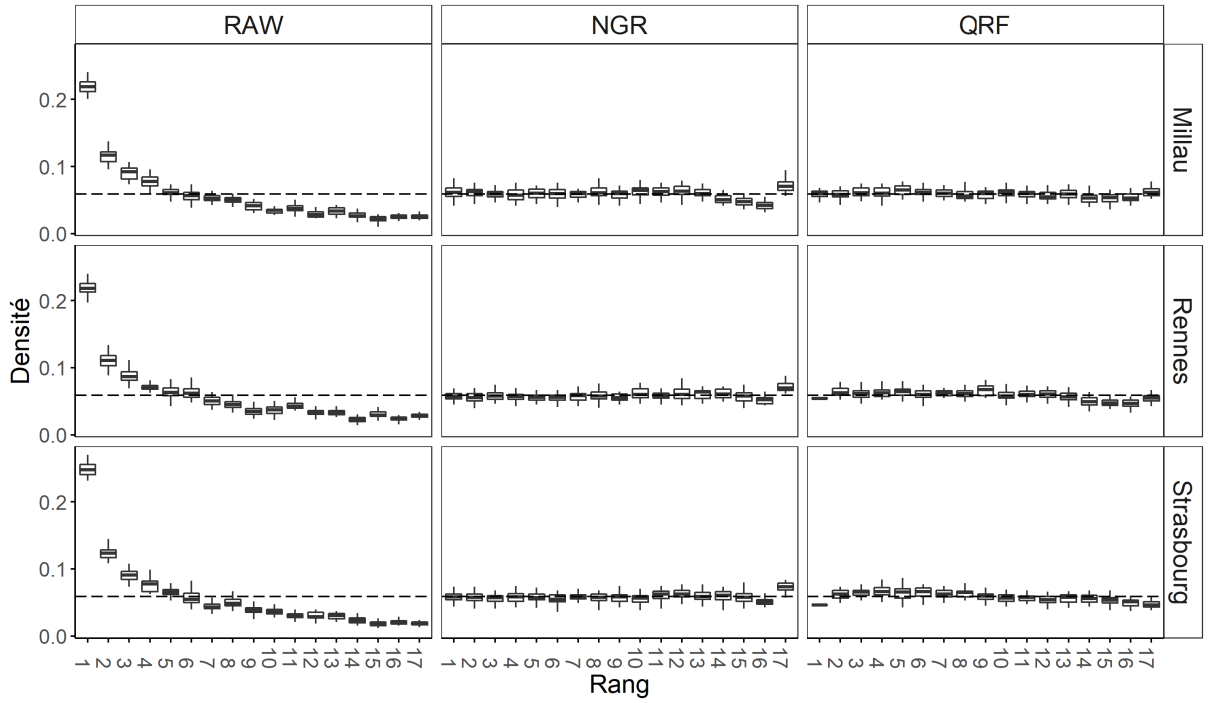
Annexe A

A.1 Compléments de résultats de calibration univariée

A.1.1 Histogrammes de rangs



(a) Histogrammes de rangs des ensembles et observations de vitesses de vent (VV) par modèle et station. *RAW : ensembles du modèle de prévisions numériques ; NGR : ensembles du modèle de régression non homogène ; QRF : ensembles du modèle de forêt aléatoire.*



(b) Precip

(c) Histogrammes de rangs des ensembles et observations de précipitations (Precip) par modèle et station. *RAW : ensembles du modèle de prévisions numériques ; NGR : ensembles du modèle de régression non homogène ; QRF : ensembles du modèle de forêt aléatoire.*

FIGURE A.0 – Histogrammes de rangs des ensembles et observations à une échéance de prévisions de 3 jours à 18H par variable météorologique, modèle et station. *Les modèles sont représentés par les colonnes et les stations par les lignes. La ligne en pointillés indique le seuil à atteindre pour former un histogramme uniforme.*

A.1.2 Analyse des covariables dans un cadre de calibration

Dans cette section, les résultats d'ajustement des coefficients du modèle de régression non homogène gaussienne (NGR) sont discutés. Ensuite, le score d'importance issu de l'algorithme de forêt aléatoire (QRF) est introduit. Ce score permet d'évaluer le rôle des covariables dans les performances de régression de l'algorithme QRF. Avant de terminer cette section, les résultats de ce score sont analysés.

A.1.2.1 Coefficients des modèles NGR

Le tableau A.1 affiche un exemple des coefficients de régressions et paramètre du mélange de lois appliqué à la calibration des vitesses du vent pour des ensembles d'échéances

3 jours et à 6H ce pour les stations Millau, Rennes et Strasbourg. Le paramètre w montre une importante part allouée à la régression log-normale laissant présager à l'importante proportion de situations capables d'évoluer en vent fort au sein des données étudiées. Le mélange permet de comparer deux régimes de vent moyen visible sur les coefficients β_{TN_0}, β_{LN_0} et issues des prévisions caractérisant chacune des stations. Ainsi, Rennes et Strasbourg ont un vent légèrement plus fort dans la régression log-normale que la régression normale tronquée. Ces coefficients sont quasiment identiques pour Millau, une station dont la localisation et l'altitude induisent des vents forts de manière récurrente comparés aux deux autres stations. Dans le modèle de mélange, il est également possible de comparer la contribution de l'ensemble avec les coefficients $(\beta_{TN_1}, \beta_{LN_1})$ de celles des prévisions HRES $(\beta_{TN_2}, \beta_{LN_2})$ et CTR $(\beta_{TN_3}, \beta_{LN_3})$. De ce fait, l'ensemble obtient un poids plus important pour la régression log-normale. Ensuite, la prévision HRES semble compléter l'information de l'ensemble pour Millau et Strasbourg. Quant à la station de Rennes, la prévision CTR est généralement privilégiée à l'intérieur de la régression log-normale. Quant à la régression normale tronquée, elle privilégie davantage la prévision HRES. Les coefficients des régressions de variances des lois du mélange $(\zeta_{TN_0}, \zeta_{LN_0})$ et $(\zeta_{TN_1}, \zeta_{LN_1})$ montrent également une plus grande contribution de la variance de la loi log-normale que la variance de la loi normale tronquée pour constituer celle des ensembles issus du mélange.

Échéance	Station	w	β_{TN_0}	β_{TN_1}	β_{TN_2}	β_{TN_3}	ζ_{TN_0}	ζ_{TN_1}
3 jours	Millau	0.25	1.69	0.10	0.03	0.05	0.08	0.11
	Rennes	0.29	0.82	0.40	0.11	0.01	0.21	0.27
	Strasbourg	0.29	1.09	0.08	0.53	0.13	0.01	0.67

(a) Coefficients de la loi normale tronquée et paramètre w du mélange.

Échéance	Station	β_{LN_0}	β_{LN_1}	β_{LN_2}	β_{LN_3}	ζ_{LN_0}	ζ_{LN_1}
3 jours	Millau	1.64	0.40	0.10	0.04	0.61	0.32
	Rennes	1.04	0.61	0.01	0.14	0.37	0.39
	Strasbourg	1.34	0.24	0.14	0.03	0.48	0.34

(b) Coefficients de la loi log-normale.

TABLE A.1 – Coefficients du mélange de loi normale tronquée et log-normale ajusté sur les observations et les ensembles de vitesses du vent pour une échéance de prévision 3 jours et à 6H, le tout pour différentes stations.

En autre exemple, les statistiques climatiques issues des observations et coefficients du modèle NGR pour les précipitations sont disponibles dans le tableau A.3 pour des ensembles à l'échéance de 3 jours à 6H. Une forte valeur de β_0 supérieur à la moyenne climatique μ_{cl} est observée. De plus, le coefficient β_2 élevé montre l'importance du seuil de précipitation POP dessiné par les réalisations des ensembles (généralement entre 0 et 1 mm) face aux coefficients $(\beta_1, \beta_3, \beta_4, \beta_5)$ de l'ordre de 10^{-2} et associés aux prévisions. Les histogrammes de rangs précédents ont montré des ensembles de précipitations tirés à partir de ce modèle avec des valeurs proches des observations voir les sous-estimant. Dans cette même situation, les ensembles RAW ont affiché de fortes surestimations des observations. Un exemple d'interprétation est que les faibles coefficients $(\beta_1, \cdots, \beta_5)$ font que la partie linéaire de l'équation (1.13) devient inférieure à 1. Ce résultat mène à une valeur moyenne de μ_{Precip} plus faible que la moyenne climatique des observations μ_{cl} et ce pour toutes les stations.

Station	μ_{cl}	σ_{cl}
Millau	0.29	1.35
Rennes	0.32	1.10
Strasbourg	0.33	1.18

(a) Moyennes et écarts types climatiques issus des observations.

Échéance	Station	β_0	β_1	β_2	β_3	β_4	β_5	ζ_0	ζ_1	ζ_2
3 jours	Millau	9.88	0.05	0.15	0.004	0.010	0.011	0.18	0.59	0.28
	Rennes	11.3	0.07	0.37	0.016	0.011	0.005	0.13	0.62	0.25
	Strasbourg	11.5	0.23	1.15	0.15	0.007	0.016	0.21	0.24	0.67

(a) Coefficients NGR.

TABLE A.3 – Statistiques climatiques et coefficients du modèle NGR des observations et ensembles de précipitations pour une échéance de prévision 3 jours à 6H, le tout pour différentes stations.

Les coefficients des modèles NGR de vitesses du vent et de précipitations pour des ensembles à 6H aux horizons de prévision de 5 jours et 10 jours sont disponibles dans les tableaux A.4 et A.5. Le paramètre w estimé aux échéances 5 et 10 jours affiche une part plus faible allouée aux régressions normales tronquées du modèle de mélange des vitesses du vent. Les coefficients β_{TN_0} se retrouvent également augmentés prenant des valeurs de vents moyens supérieures aux coefficients β_{LN_0} à 10 jours. Une augmentation significative est également notée pour les coefficients de la régression de la variance en

lien avec le caractère dispersé des ensembles à une échéance de 10 jours. Néanmoins cette augmentation touche les coefficients de variance de la loi normale tronquée et une diminution est observée pour ceux de la loi log-normale. Le modèle de mélange tend à vouloir restreindre l'élargissement de la distribution des ensembles *RAW* à cette échéance en composant avec la variance d'une loi normale tronquée et diminuant celle de la loi log-normale. Concernant le modèle *NGR* des précipitations, l'importance des seuils de précipitations *POP* issues des ensembles augmente ainsi que la valeur du coefficient ζ_0 à une échéance de 10 jours. L'écart entre la moyenne climatique μ_{cl} des observations et le coefficient β_0 diminue. Le modèle tend à recentrer l'ensemble autour des observations avec une augmentation de dispersion comparé aux ensembles des courtes échéances comme 3 jours.

Échéance	Station	w	β_{TN_0}	β_{TN_1}	β_{TN_2}	β_{TN_3}	ζ_{TN_0}	ζ_{TN_1}
5 jours	Millau	0.17	2.24	0.48	0.16	0.001	0.97	0.001
	Rennes	0.39	0.52	0.89	0.03	0.001	0.25	0.31
	Strasbourg	0.38	1.15	0.32	0.35	0.02	0.63	0.19
10 jours	Millau	0.18	2.06	0.79	-	0.04	1.47	0.14
	Rennes	0.08	3.85	2.46	-	0.17	0.46	0.12
	Strasbourg	0.13	1.39	0.76	-	0.01	1.11	0.001

(a) Coefficients de la loi normale tronquée et paramètre w du mélange.

Echéance	Station	β_{LN_0}	β_{LN_1}	β_{LN_2}	β_{LN_3}	ζ_{LN_0}	ζ_{LN_1}
5 jours	Millau	1.35	0.47	0.001	0.001	0.10	0.40
	Rennes	1.22	0.50	0.003	0.06	0.56	0.47
	Strasbourg	1.16	0.35	0.02	0.04	0.07	0.51
10 jours	Millau	1.61	0.36	-	0.012	0.41	0.14
	Rennes	0.88	0.69	-	0.076	0.05	0.70
	Strasbourg	1.65	0.25	-	0.01	0.42	0.26

(b) Coefficients de la loi log-normale.

TABLE A.4 – Coefficients du mélange de loi normale tronquée et log-normale ajusté sur les observations et les ensembles de vitesses du vent à 6H, le tout pour différentes stations aux échéances de prévisions 5 et 10 jours. *Coefficient β_4 inexistant aux échéances de 10 jours, la prévision HRES est indisponible.*

Echéances	Stations	β_0	β_1	β_2	β_3	β_4	β_5	ζ_0	ζ_1	ζ_2
5 jours	Millau	12.92	0.39	1.93	0.39	0.01	0.06	0.14	0.27	0.28
	Rennes	8.43	0.02	0.44	0.06	0.02	0.026	0.30	0.31	0.37
	Strasbourg	10.9	0.04	0.05	0.06	0.02	0.002	0.14	0.38	0.30
10 jours	Millau	2.70	0.001	1.01	1.45	-	0.82	1.08	0.63	0.001
	Rennes	1.97	0.002	0.60	0.93	-	1.22	0.86	0.006	0.003
	Strasbourg	2.66	0.003	0.71	1.21	-	0.47	1.11	0.50	0.001

TABLE A.5 – Coefficients du modèle gamma censuré et décalé prenant les observations et les ensembles de précipitations initialisés à 6H, le tout pour différentes stations aux échéances de prévisions 5 et 10 jours. *Coefficient β_4 inexistant aux échéances de 10 jours, la prévision HRES est indisponible.*

L'analyse des coefficients des modèles NGR révèle des caractéristiques de modèle NGR des vitesses de vent différentes suivant la localisation spatiale et l'échéance de prévision. Par exemple, les coefficients du modèle NGR suggèrent différents régimes de vitesses de vent pour Rennes et Strasbourg à une échéance de 3 jours. Ensuite, l'évolution des échéances entraîne une augmentation des écarts de vitesses de vent entre les régimes, mais également une réduction de l'impact des régimes avec les plus fortes vitesses dans la régression (excepté Strasbourg). Dans le cas du modèle NGR des précipitations, la partie "non linéaire" du modèle rend plus complexe l'interprétation des coefficients de régressions. Néanmoins, les valeurs ajustées suivant les modèles suggèrent un changement des caractéristiques du modèle plus dominé par l'évolution des échéances que les localisations spatiales.

A.1.2.2 Score d'importance du modèle de forêt aléatoire

L'évaluation des contributions des covariables est une information non négligeable dans l'étude des résultats de calibration. Cette information peut être prise en compte par l'intermédiaire du score d'importance estimé lors de l'application de l'algorithme de forêt aléatoire de BREIMAN 2001. Cet algorithme fournit deux scores : l'importance de permutation et l'importance de Gini. Les deux scores sont généralement très similaires en termes de conclusions. Néanmoins, l'importance de Gini peut mener à des divergences dans une situation de surapprentissage alors que l'importance de permutation est légèrement plus robuste. Pour simplifier l'affichage des scores, seule l'importance de permutation sera considérée dans cette partie. Les notations de la section 1.1.2.1 sont reprises pour définir ce score, soit le jeu de données $\mathcal{X} = \{y_t, x'_t\}_{1 \leq t \leq T}$ avec $x'_t \in \mathcal{B} \subseteq \mathbb{R}^p$ un ensemble de p

covariables caractérisant l'ensemble X_t^* et y_t une observation de Y est étudiée.

Pour reprendre Gregorutti, Michel et Saint-Pierre 2017, la notion d'importance d'un prédicteur x_p se définit par l'évaluation du lien existant entre la variable visée Y et x_p. Lors de l'ajustement du modèle de forêt aléatoire f de Breiman 2001 pour N arbres f_n avec $n \in \{1, \cdots, N\}$, l'étape de "bagging" brise ce lien en construisant une collection de N sous-échantillons $\{\mathcal{X}_n\}_{1 \leq n \leq N}$ de taille $T' < T$ tirés aléatoirement de $\mathcal{X}$. Pour chaque sous-échantillon $\mathcal{X}_n$, une permutation aléatoire de la $p^{ème}$-covariable s'opère ensuite, générant le nouveau sous-échantillon $\mathcal{X}_n^p$. Dès lors, la différence d'erreur d'estimation du modèle peut être évaluée pour chaque covariable p à partir des sous-échantillons $\mathcal{X}_n^p$ et ceux non modifiés $\mathcal{X}_n$. Le score d'importance entre $\mathcal{X}_n^p$ et $\mathcal{X}_n$ est donc défini :

$$\mathcal{I}mp(x_p) = \frac{1}{N} \sum_{n=1}^{N} [\mathcal{R}(f_n, \mathcal{X}_n^p) - \mathcal{R}(f_n, \mathcal{X}_n)] \tag{A.1}$$

avec la fonction $\mathcal{R}(f, \mathcal{X})$ représentant la fonction objectif jugeant la qualité de l'apprentissage du modèle f à travers le jeu de données $\mathcal{X}$. Dans le cas d'une régression avec une variable à prédire continue comme Y, la fonction objectif $\mathcal{R}$ s'exprime comme une moyenne des moindres carrés (MSE) décrite précédemment (1.19).

Dès lors, la fonction $\mathcal{I}mp(x_p)$ fournit une mesure orientée du lien existant entre la covariable x_p et la variable à prédire. Un résultat négatif indiquerait que la covariable nuit à la prédiction de la variable ciblée. Inversement, un résultat positif montrerait la contribution positive de cette covariable à la prédiction.

A.1.2.3 Résultats du score d'importance

Les scores d'importance des modèles QRF appliqués aux covariables des ensembles de vitesses du vent et de précipitations sont visibles sur la figure A.1 par station moyennant les échéances et la figure A.2 par échéance moyennant les stations, et ce pour les deux horaires d'initialisations confondus. Pour les vitesses du vent, le facteur temporel de mois ("month") contribue fortement pour toutes les stations et échéances. Le modèle de forêt met en avant le lien entre les erreurs présentes dans la distribution des ensembles de vitesses du vent et la saisonnalité du vent (Taillardat, Mestre et al. 2016). Les horaires intégrés dans la seconde covariable temporelle "hour" étant 6h et 18H permettent au modèle d'appréhender la dimension temporelle du vent diurne importante pour différencier des scénarios de vitesses du vent (Pinson et Hagedorn 2012, Ailliot et al. 2015).

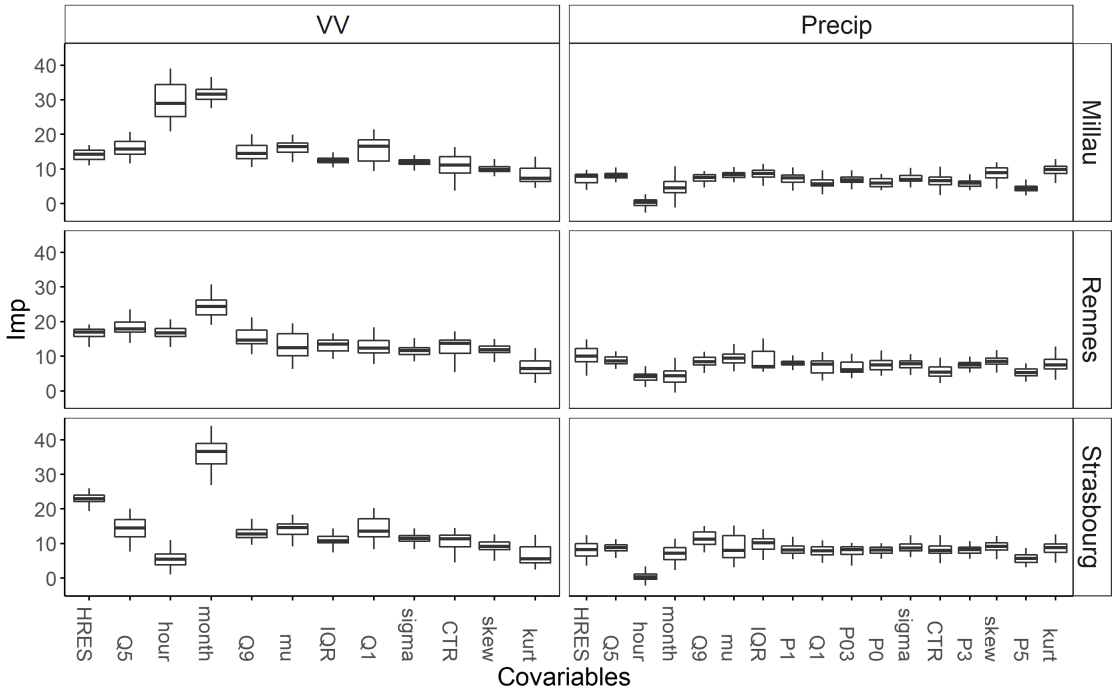


FIGURE A.1 – Importance des prédicteurs des modèles QRF par variables météorologiques et station toutes échéances confondues. *Les stations sont représentées par lignes, variables météorologiques par colonnes, VV : vitesses du vent ; Precip : précipitations.*

Les scores montrent un lien local entre cette covariable et la calibration des ensembles. En effet, une forte importance est observée pour la station de Millau, une moyenne pour Rennes et une faible pour Strasbourg, et ce pour les différentes échéances. Millau est une station à une altitude élevée comparée aux deux autres localisations. La dynamique du vent local dispose d'un comportement plus difficile à caractériser par le modèle de prévision numérique grande échelle justifiant de l'importance de ce type de facteurs temporels pour les modèles de calibration. Ensuite, les covariables de prévision déterministe "HRES" et la médiane de l'ensemble "Q5" semblent être privilégiées par le modèle de calibration avec quelques fluctuations dépendantes de la station ou de l'échéance. Ces deux covariables concentrent les informations essentielles pour obtenir des performances calibration correcte des vitesses du vent suivant les différentes stations et échéances. Les covariables statistiques de quantiles ("Q1", "Q9"), de moyenne ("mu"), d'écart type ("sigma") et du moment d'asymétrie d'ordre 3 ("skew") complètent les informations fournies par la prévision déterministe "HRES" et la médiane "Q5" de l'ensemble pour la calibration des ensembles de vitesses du vent, moyennant quelques fluctuations suivant les stations et les

échéances. Néanmoins, la covariable statistique du moment d'acuité d'ordre 4 "kurt" issue des ensembles se détache pour son faible score au sein du modèle des vitesses du vent. Les résultats de calibration du modèle QRF rejoignant ceux du modèle NGR, il n'est pas anodin d'obtenir que seuls les trois premiers moments empiriques suffisent pour estimer la forme de la distribution des ensembles de vitesses du vent.

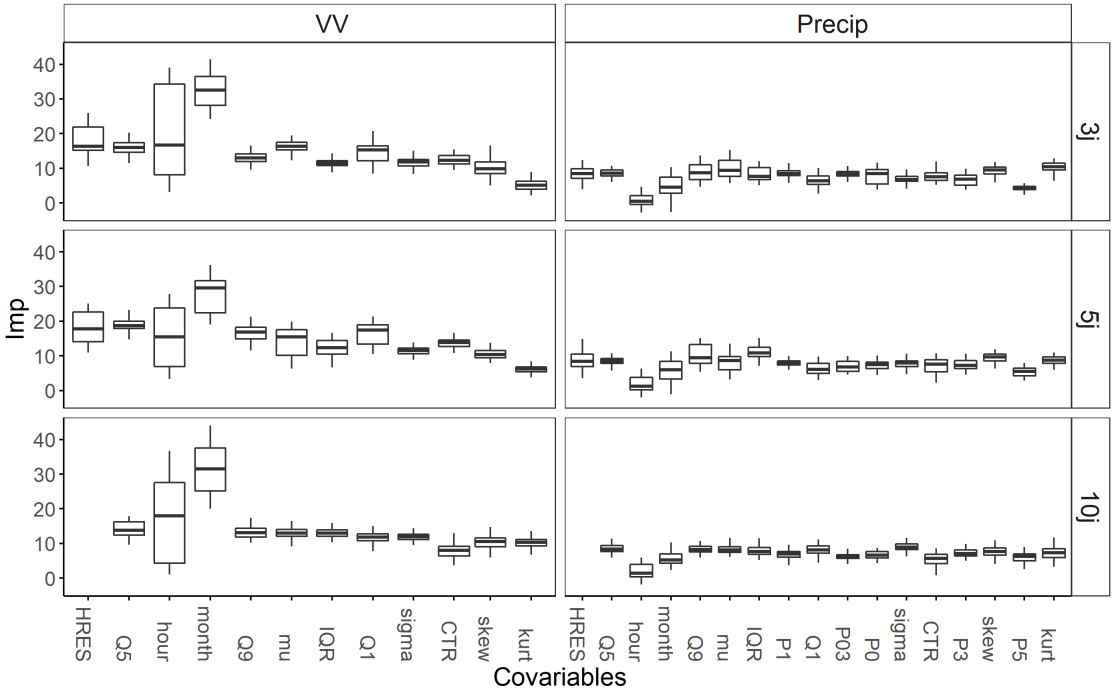


FIGURE A.2 – Importance des prédicteurs des modèles QRF par variables météorologiques et échéances pour toutes stations confondues. *Les échéances sont représentées par les lignes, les variables météorologiques par les colonnes. Les boîtes à moustaches sont vides pour la covariable HRES car la prévision est indisponible pour des échéances égales et supérieures à 10 jours. VV : vitesses du vent ; Precip : précipitations.*

Concernant la variable de précipitations, les covariables de prévision déterministes et d'informations statistiques tirées des ensembles de prévision indiquent des contributions très proches des unes des autres, et ce sans grande fluctuation suivant les différentes stations et échéances. Les covariables de facteurs temporels montrent un très faible impact dans la caractérisation des erreurs de distributions des ensembles de précipitations. Néanmoins, il est possible d'apercevoir que les scores d'importance des covariables ("IQR", "Q9", "skew" "kurt" et "sigma") semblent être plus élevés suivant certaines localisations et échéances. Par exemple, Millau affiche des contributions plus importantes pour les covariables de "skew" et "kurt" affinant la forme des distributions des ensembles post-traités.

Ces covariables permettent notamment d'améliorer la quantification d'incertitude autour de la prévision d'événements rares comme l'arrivée de fortes précipitations (événement particulièrement sensible pour la localisation de Millau).

Pour résumer, les modèles de forêt aléatoire de régression QRF montrent des similitudes avec les modèles NGR. En effet, le modèle QRF appliqué à la calibration des vitesses de vent montre des scores d'importance plus sensibles au changement de localisation spatiale. Également pour cette variable, les statistiques décrivant la forme de la distribution des ensembles sont moins impactantes pour le modèle que l'aspect temporel ou encore la prévision déterministe haute résolution (HRES) et la médiane des ensembles. Dans le cas du modèle QRF appliqué aux précipitations, la variabilité des scores est plus observée lors du changement d'échéance de prévision. De plus, les scores d'importance mettent en avant les covariables décrivant l'aspect de la distribution des ensembles de précipitations contrairement au modèle des ensembles des vitesses de vent.

Annexe B

B.1 Classification issue d'une calibration multivariée

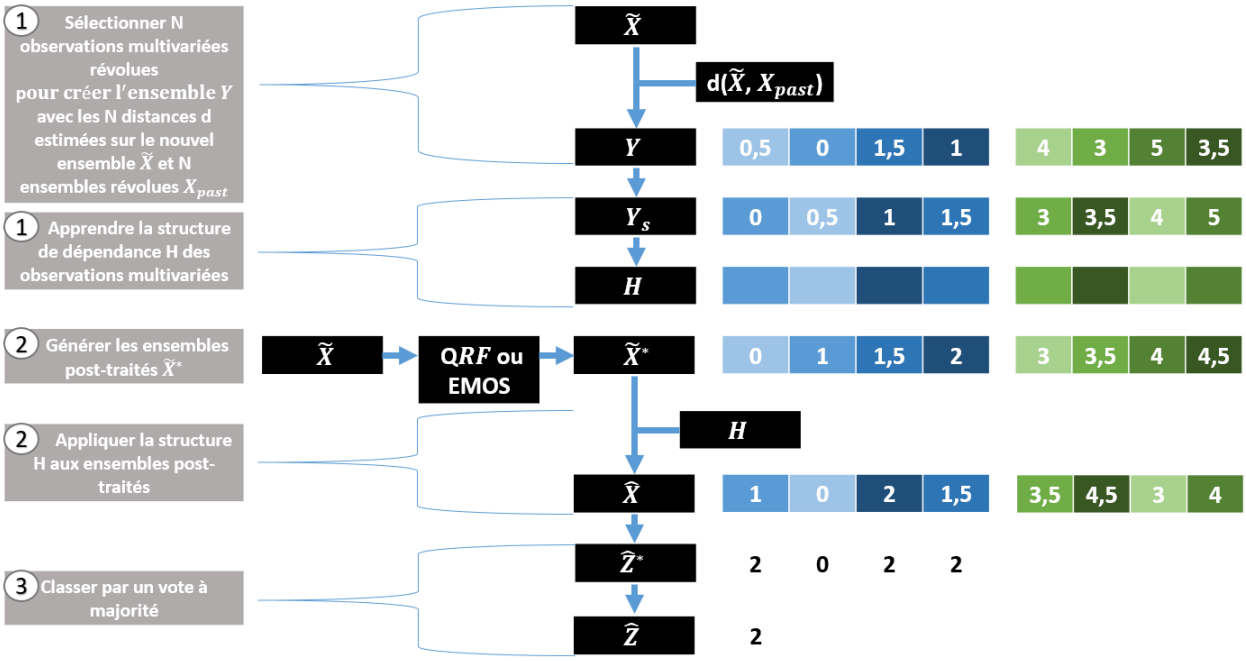


Figure B.1 – Schéma des trois étapes du couplage entre les modèles de calibration univariée et multivariée pour des ensembles et des observations bivariées. *La partie bleue représente des données issues de la première dimension et vert de la seconde.*

La figure B.1 représente les grandes étapes de la méthode décrite en section 1.3.1. Le dégradé de couleur visible sur les deux dimensions montre l'ordonnancement des données qui est appris par la structure **H**.

B.2 Analyse des contributions des covariables

L'analyse des contributions des covariables par modèle est intéressante aussi bien pour comprendre que pour améliorer les performances de classification des modèles. Les covariables sélectionnées par les modèles permettent aux prévisionnistes de mieux comprendre les informations météorologiques où la topologie locale influençant la qualité de prédiction des modèles. Dans cette section, un exemple des contributions des covariables des modèles de classification directe est donné au travers de l'affichage des importances du modèle RFC et des coefficients du modèle MLR. Le score d'importance du modèle de forêt est également représenté par classe, ainsi que les coefficients de régression multinomiale. Avant cela, le score d'importance du modèle de forêt aléatoire introduit en annexe A.1.2.2, est présenté dans un cadre de classification.

Lorsque le problème étudié se présente sous la forme d'une classification avec une variable discrète Z à prédire, comme dans la section 2.1, la fonction $\mathcal{R}$ de l'équation (A.1) du score d'importance prend la forme d'une précision globale 2.10 estimée à partir des TP_k 2.9 de la matrice de confusion qui traduit l'écart entre l'échantillon prédit $\{\hat{z}_1, \cdots, \hat{z}_T\}$ où $\hat{z}_t = f(x'_t)\ \forall t \in \{1, \cdots, T\}$ et l'échantillon de la variable cible $\{z_1, \cdots, z_T\}$.

Pour rappel, une valeur positive de l'importance indique une contribution de la covariable étudiée dans la prédiction de la variable objectif. Une valeur négative indique au contraire que l'information de la covariable est redondante au sein du modèle, et nuit à l'objectif de prédiction de classe.

B.2.1 Mesure d'importance des covariables du modèle de forêt

L'importance des covariables de la table 1.1, utilisées par les modèles de calibration univariés QRF, est présentée dans la partie résultats du chapitre précédent. Dans ce chapitre, une covariable contenant les classes prédites par l'ensemble du modèle de prévision numérique nommé "RAW Z" a été ajoutée à l'ensemble des covariables. Pour les modèles QRF, cette covariable n'a que très peu d'impact.

L'évaluation de l'importance des covariables du modèle RFC est affichée des scores par classe, comme présentée sur les figures B.2 par station , B.3 par échéance pour les vitesses de vent et sur les figures B.4 par station, B.5 par échéance pour les précipitations. Pour faciliter l'affichage, les covariables qualitatives de "RAW Z", "month" et "hour" ont été répliquées par variable météorologique.

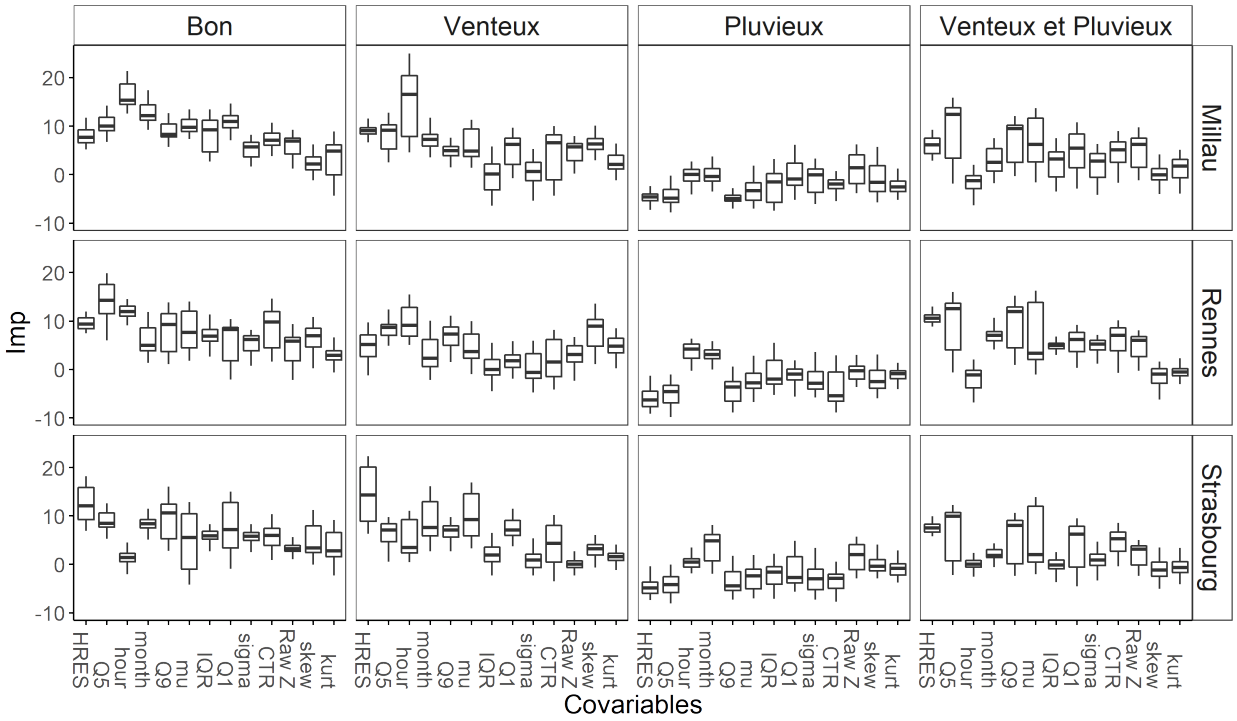


FIGURE B.2 – Importance des prédicteurs de la variables de vitesses du vent pour le modèle *RFC* par classe et par station, toutes échéances confondues. *Les stations sont représentées par des lignes,les classes par des colonnes.*

Sur les figures B.2 et B.3, les scores associés aux facteurs temporels ("month" et "hour") sont plus élevés les classes sans pluie ("Bon" et "Venteux") pour les localisations de Millau et Rennes. Du côté de Strasbourg, seul le caractère saisonnier des ensembles (décrit par la covariable "month") montre un score élevé pour les classes "Bon" et "Venteux". L'étude des covariables du modèle QRF avait révélé que l'aspect saisonnier touchait particulièrement la calibration des vitesses du vent. Ensuite, les covariables de la prévision déterministe HRES et de la médiane des ensembles ("Q5") ont des scores équivalents voir plus élevés que les autres covariables (avec néanmoins quelques fluctuations suivant les stations et échéances) suivant les classes "Bon", "Venteux" et "Venteux et Pluvieux". De plus, le moment d'asymétrie d'ordre 3 ("skew") des vitesses du vent est sélectionné pour décrire les classes aux stations de Millau et Rennes. Cette information est également partagée avec le modèle de calibration QRF pour ces stations et mêmes classes.

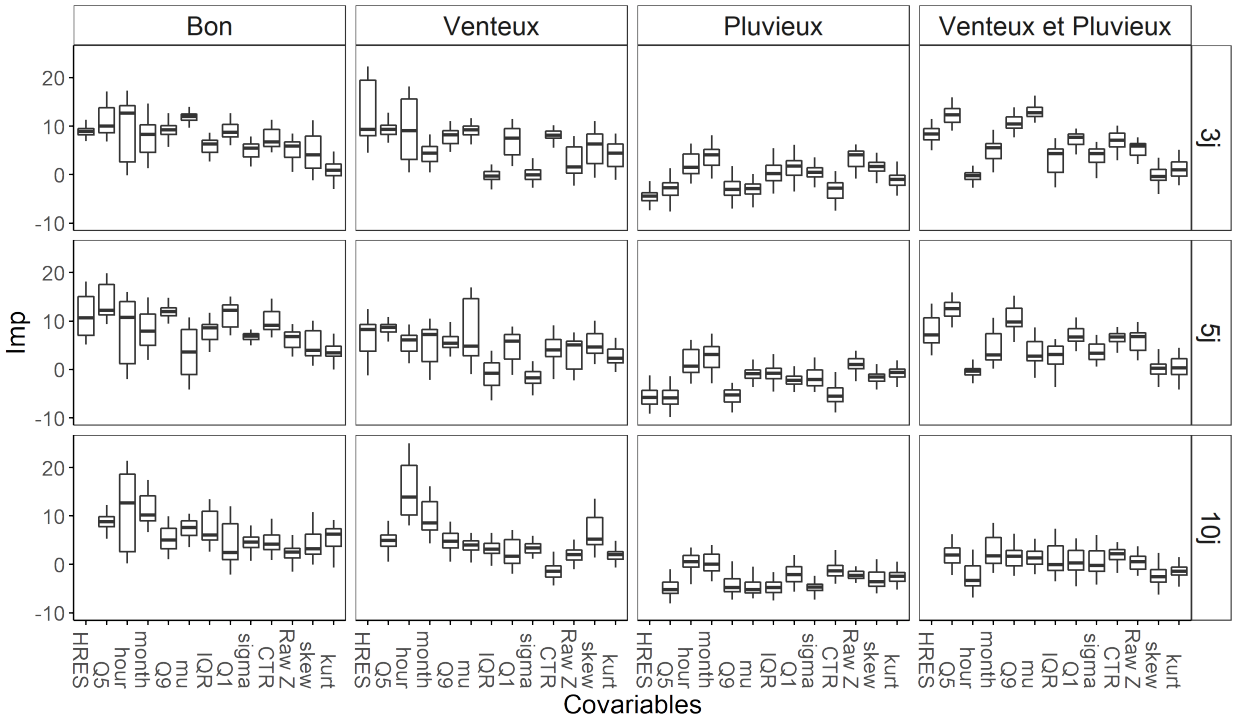


FIGURE B.3 – Importance des prédicteurs de la variable de vitesses du vent pour le modèle *RFC* par échéances et classes pour toutes stations confondues. *Les échéances sont représentées par lignes, classes par colonnes. Boîte à moustache vide pour la covariable HRES car la prévision est indisponible pour des échéances égales et supérieures à 10 jours.*

De plus, la séparation entre les classes de pluie ("Pluvieux" et "Venteux et Pluvieux") et les classes sans pluie pour les facteurs temporels, ainsi que pour les autres covariables, est d'autant plus forte pour l'échéance de 10 jours. Les scores des covariables liées à la variable des vitesses du vent et affichés sur les figures B.2 par station et B.3 par échéance, montrent que les classes "Bon" et "Venteux" sont plus sensibles aux changements de météorologie locale, et que la classe "Venteux et pluvieux" est plus sensible à l'évolution des d'échéances. Par exemple, la prévision déterministe HRES contribue le plus dans la prédiction des classes sans pluie à Strasbourg, alors qu'à Rennes les classes sans pluies sont principalement prédites par la médiane de l'ensemble et par l'heure. Un point attendu et visible sur la figure B.2 des covariables de vitesses de vent est l'observation de scores négatifs pour la classe "Pluvieux" montrant la difficulté du modèle à identifier les situations de pluies sans vent en se basant sur ces covariables.

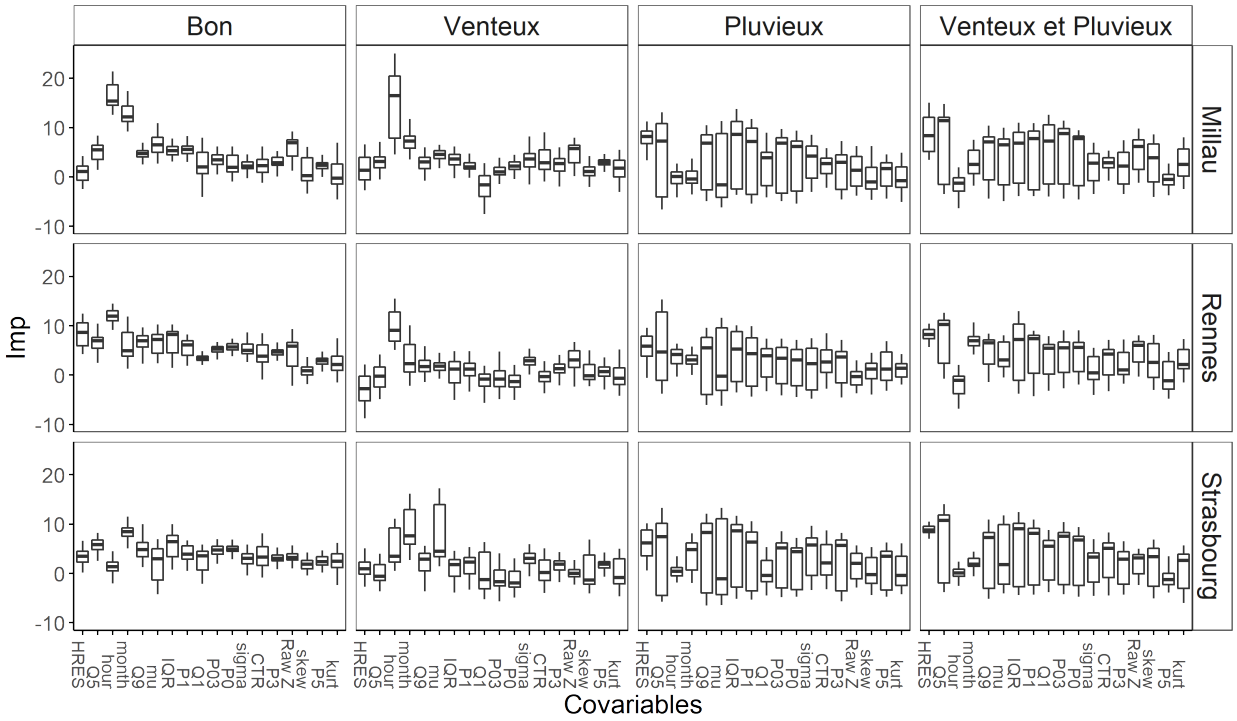


FIGURE B.4 – Importance des prédicteurs du modèle *RFC* par classe et station pour la variable météorologique des précipitations, toutes échéances confondues. *Les stations sont représentées par des lignes, les classes par des colonnes.*

Sur les figures B.4 et B.5, il est intéressant de noter que les scores associés aux covariables issues des ensembles de précipitations sont plus sensibles aux changements d'échéance que de localisation spatiale. Ensuite, la figure B.5 montre que la covariable "IQR", décrivant l'écart des valeurs des membres des ensembles de précipitations, est importante pour la prédiction des classes pour chaque station à faible échéance. Le moment d'acuité d'ordre 4 ("kurt") des vitesses du vent et de précipitations, ainsi que la covariable "Raw Z" des classes prédites par l'ensemble du modèle de prévision numérique obtiennent les scores d'importances les plus faibles, et ce indépendamment des stations et des échéances. L'information de ces covariables est jugée redondante et peu corrélée à l'objectif de prédiction du modèle de forêt, alors que dans le modèle de calibration QRF des précipitations, la covariable "kurt" a un score d'importance plus élevé. De manière générale, une grande partie des covariables importantes pour le modèle de calibration QRF sont retrouvées par le modèle de classification RFC. Néanmoins des variations plus importantes des scores d'importance sont observées suivant les échéances pour le mo-

dèle RFC que pour le modèle QRF. Cette variation semble indiquer que les objectifs de calibration et de classification ont des erreurs légèrement différentes à corriger.

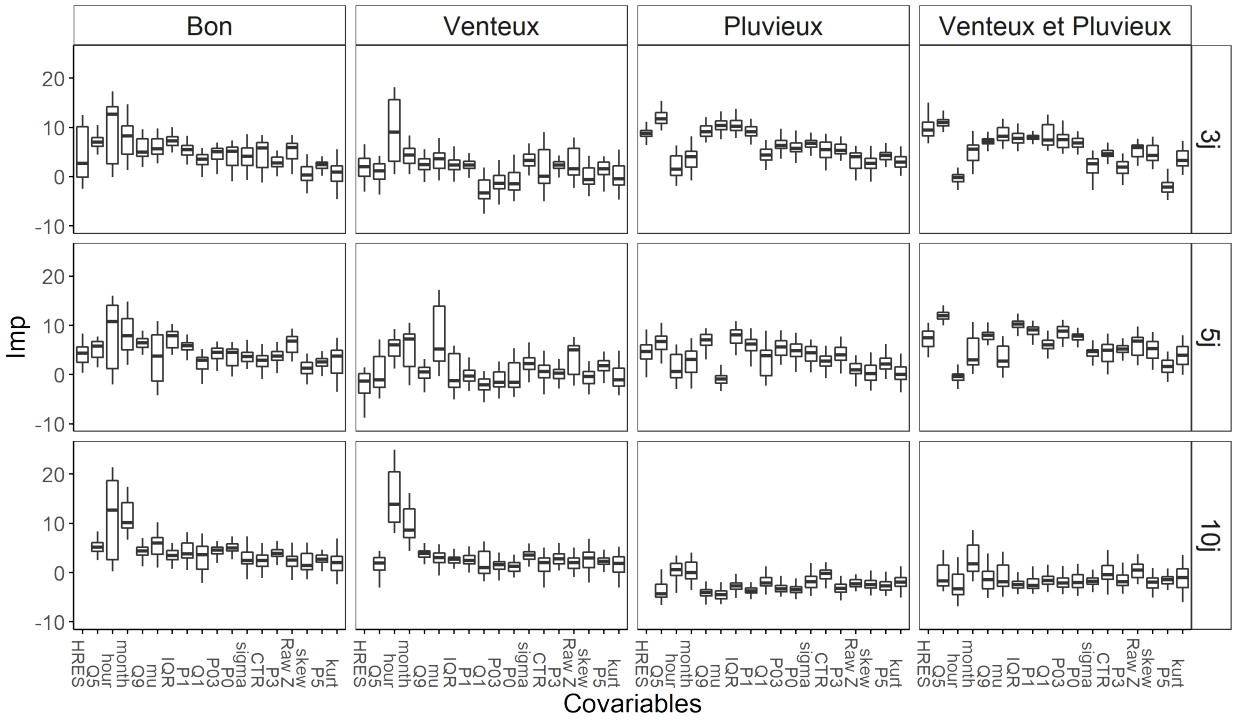


FIGURE B.5 – Importance des prédicteurs de la variable de précipitations pour le modèle RFC par échéances et classes pour toutes stations confondues. *Les échéances sont représentées par lignes, classes par colonnes. Boîte à moustache vide pour la covariable HRES car la prévision est indisponible pour des échéances égales et supérieures à 10 jours.*

La plupart des covariables issues des ensembles de précipitations montrent un pattern similaire, avec d'importantes dispersions sur les boîtes à moustaches des classes de pluie, visible sur la figure B.4. Cette dispersion est provoquée par la décroissance de l'importance entre les échéances de 3 et 5 jours et celle de 10 jours pour les classes "Pluvieux" et "Venteux et Pluvieux", illustrée par la figure B.5. La décroissance suivant les échéances des scores des covariables issues des ensembles de précipitations pour les classes supérieures au seuil de précipitations illustrent la dégradation du lien multivarié entre les variables météorologiques aidant le modèle de forêt à prédire les classes d'observations. Les informations statistiques tirées des ensembles de prévision sont très dispersées à cet horizon et peinent à caractériser la distribution des précipitations en comparaison à celles des vitesses du vent. De ce fait, le modèle de classification maximise son objectif de prédiction

en se concentrant sur les classes différenciées par la vitesse du vent, plus simples à relier aux informations des ensembles de vitesses du vent à une échéance de 10 jours. Ce résultat souligne le déclin de performance de classification des modèles suivant les échéances, et rejoint les résultats observés dans la partie précédente.

L'analyse des scores d'importances des covariables rejoignent les conclusions des modèles QRF du chapitre précédent. Les covariables du modèle RFC mettent en avant l'aspect temporel par station des ensembles, ainsi que la prévision HRES et les statistiques de quantiles, moyennes, écarts types, IQR et probabilités de précipitations identifiées dans les travaux de Taillardat, Mestre et al. 2016. Les scores par échéances révèlent la perte d'information caractérisant le lien entre les précipitations et vitesses du vent. Pour terminer, l'importance par classe montre que la classe "Pluvieux" est difficilement décrite par les covariables de vitesses du vent et peu par celles de précipitations dont la contribution décroît suivant les échéances.

B.2.2 Coefficients pénalisés de la régression multinomiale Lasso

Le modèle par régression multinomiale Lasso utilisé pour la classification sélectionne les covariables en faisant converger vers zéro les coefficients de celles ayant la contribution la plus faible.

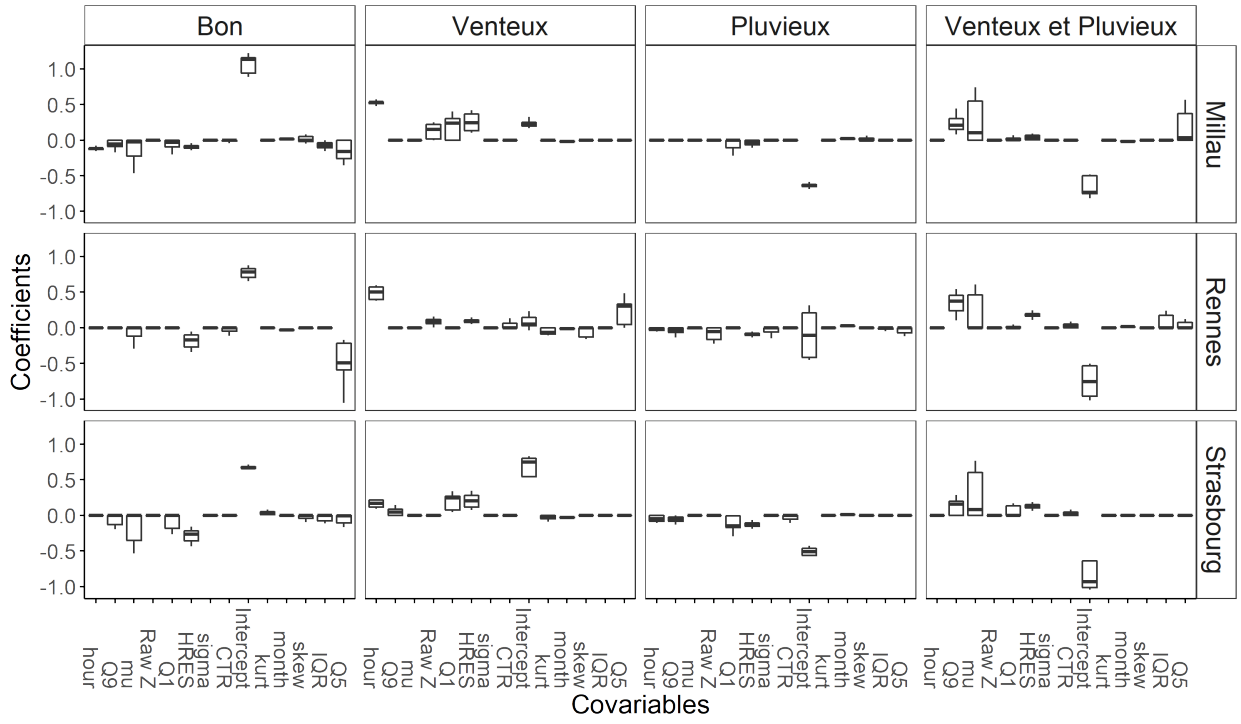


FIGURE B.6 – Coefficients MLR des covariables de vitesses du vent par classe et par station, toutes échéances confondues. *Les stations sont représentées par des lignes, et les classes par des colonnes.*

Les coefficients du modèle MLR sont présentés sur les figures B.6, B.7 par classe et station, respectivement pour les covariables de vitesses de vent et de précipitations. De même que les coefficients du modèle MLR selon les classes et échéances sont affichés sur la figure B.8 pour les covariables de vitesses de vent et la figure B.9 pour les covariables de précipitation. Sur ces figures, le modèle MLR montre des coefficients supérieurs à zéro pour des covariables également identifiées comme importante pour le modèle QRF. Par exemple, le coefficient de la prévision déterministe "HRES" de la variable de vitesses du vent est remarqué supérieur à zéro.

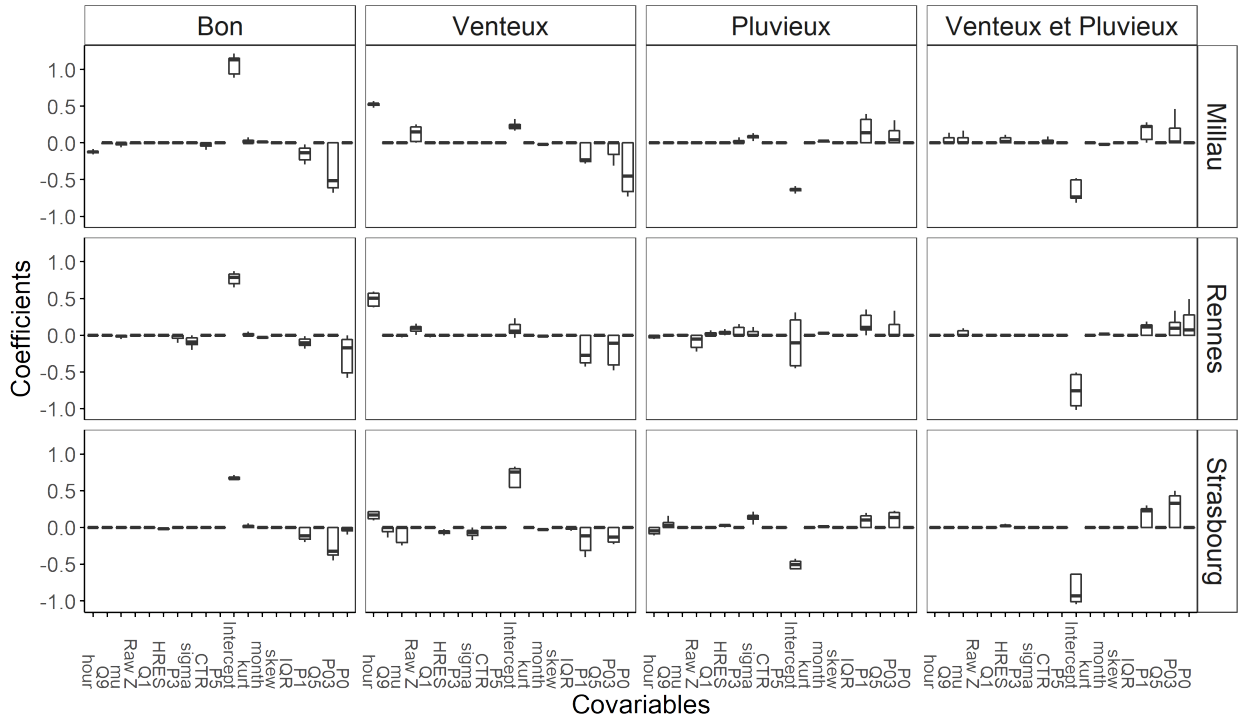


FIGURE B.7 – Coefficients MLR des covariables de précipitations par classe et par station, toutes échéances confondues. *Les stations sont représentées par des lignes, et les classes par des colonnes.*

Cependant, des nouvelles covariables apparaissent comme utiles pour la prédiction des classes par le modèle MLR. En effet, il est intéressant de voir que le modèle MLR utilise fortement les covariables de probabilité de dépassement de seuil de précipitations issues des ensembles ("P0", "P03","P1",), et proches du seuil défini pour construire les classes de pluies. Le facteur temporel d'heure obtient également un coefficient élevé pour les stations de Millau et Rennes, et un coefficient moins élevé pour Strasbourg, indépendamment des échéances. Ce résultat est également partagé par le modèle RFC. Les coefficients associés aux statistiques des ensembles "Q5, Q1, Q9" sont également élevés, mais varient en fonction des stations et classes.

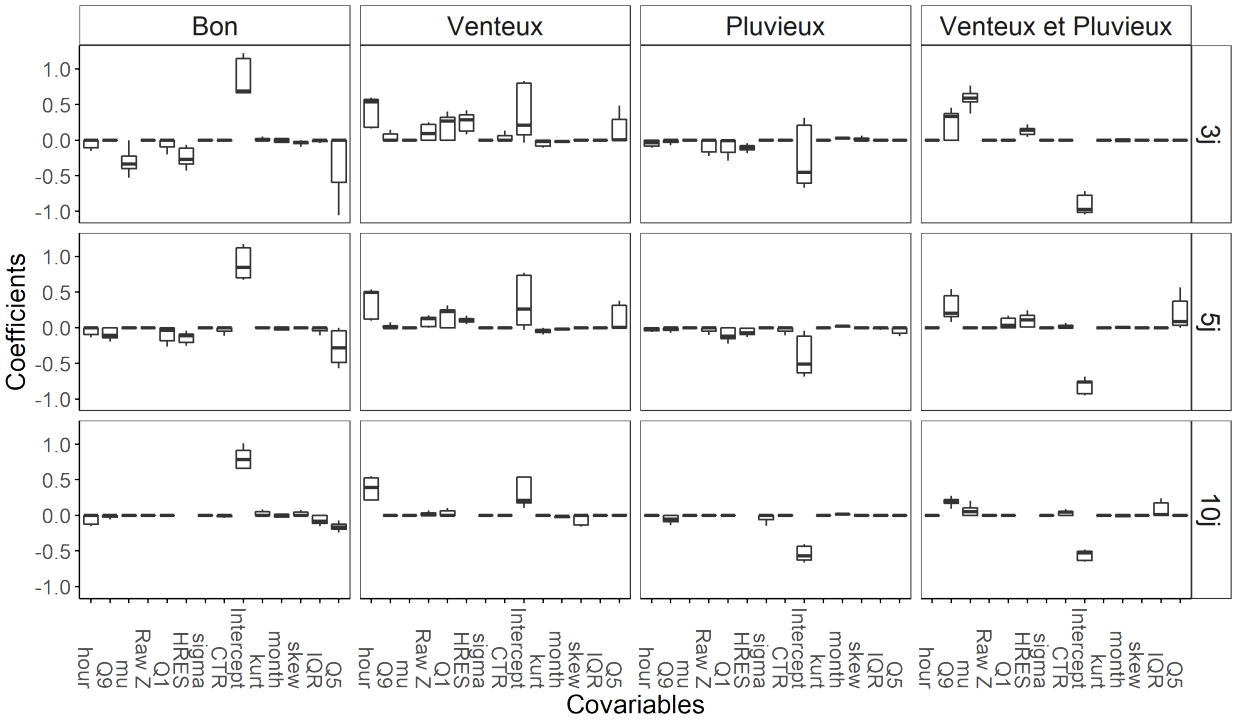


FIGURE B.8 – Coefficients MLR des covariables de vitesses du vent par classes, échéances, et ce pour toutes stations confondues. *Les échéances sont représentées par lignes, classes par colonnes.*

De plus, les coefficients MLR des figures par classe et échéance, révèlent une séparation entre les valeurs des coefficients d'une covariable associé à deux classes et les valeurs de coefficients de cette même covariable pour deux autres classes différentes. Par exemple, les covariables de probabilités "P1, P03" liées aux précipitations ont des valeurs de coefficients négatives pour les classes "Bon" et "Venteux" et des valeurs positives pour les classes "Pluvieux" et "Venteux et Pluvieux" pour Millau et Rennes aux échéances de 3 et 5 jours. Quant aux covariables "HRES, Q5, mu" des vitesses du vent et le facteur d'heure, elles séparent les classes "Venteux" et "Venteux et Pluvieux" des classes "Bon" et "Pluvieux". Les coefficients ajustés par le modèle MLR permettent d'identifier les covariables utiles à la discrimination des classes avec ou sans pluie, et des classes avec différentes vitesses de vent.

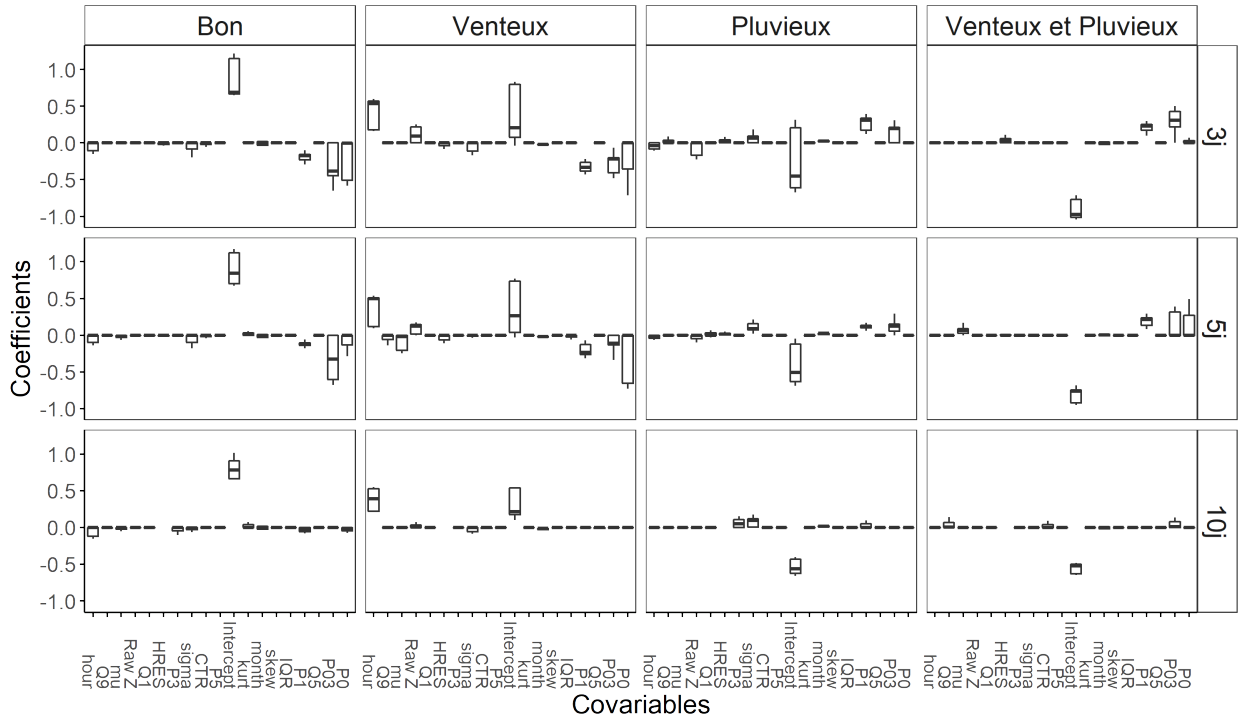


FIGURE B.9 – Coefficients MLR des covariables de précipitations par classes, échéances, et ce pour toutes stations confondues. *Les échéances sont représentées par lignes, classes par colonnes.*

Annexe C

C.1 Algorithme Espérance-Maximisation

Initié par Dempster, Laird et Rubin 1977, l'algorithme d'Espérance-Maximisation (EM) est une méthode numérique proposée pour des problèmes de maximisation de vraisemblance aux données incomplètes ou manquantes. Pour expliciter cette sous-partie, les travaux de Bilmes et al. 1998 et McLachlan et Krishnan 2007 sont repris. Les données $\mathcal{X}$ issues d'un mélange peuvent être décrites comme incomplète de par le manque d'information caractérisant les classes latentes. Pour y remédier, l'existence de données non observées $\mathcal{Z}$ composé de T réalisations $\{z_t\}_{1\leq t\leq T}$, $z_t \in \{1, ..., K\}$ suivant Z, une variable aléatoire discrète de classe latente discriminant chaque composante du mélange gaussien. Le nouvel ensemble $(\mathcal{X}, \mathcal{Z})$ forme un jeu de données complet. En reprenant la log-vraisemblance (3.3), le problème de maximisation de la vraisemblance incomplète devient :

$$\begin{aligned}\Psi^* &= \arg\max_{\Psi} \quad \log(\mathcal{L}(\mathcal{X}, \mathcal{Z}; \Psi)) \\ &= \arg\max_{\Psi} \quad \sum_{t=1}^{T} \log\left(f(x_t, z_t; \Psi)\right) \\ &= \arg\max_{\{\pi_k, \mu_k, \Sigma_k\}_{1\leq k\leq K}} \quad \sum_{t=1}^{T}\sum_{k=1}^{K} 1_{\{z_t=k\}} \log\left(\pi_k \varphi(x_t; \mu_k, \Sigma_k)\right)\end{aligned} \tag{C.1}$$

Désormais, l'idée de l'algorithme EM est de déduire une expression des estimateurs de Ψ au travers d'une approche itérative. Pour cela, chaque itération $i \in \{1, ..., I_{max}\}$ de l'algorithme EM est divisée en deux étapes clés. Après l'initialisation de l'ensemble des paramètres $\Psi^{[0]}$, la première étape de l'algorithme (appelé étape "E") évalue la vraisemblance complète $\mathcal{L}(\mathcal{X}, \mathcal{Z}; \Psi^{[i]})$ conditionnée aux paramètres de l'itération passée contenus dans $\Psi^{[i-1]}$. La seconde étape (appelé étape "M") met à jour l'ensemble des paramètres $\Psi^{[i]}$ en maximisant l'expression obtenue de l'étape E".

C.1.1 Etape E

Cette étape se traduit par l'expression de l'espérance de la log-vraisemblance des données complètes conditionnée à $\Psi^{[i-1]}$ et les données connues $\mathcal{X}$. La fonction suivante est définie :

$$\begin{aligned} Q(\Psi^{[i]}, \Psi^{[i-1]}) &= \mathbb{E}_Z \left[\mathcal{L}(\mathcal{X}, \mathcal{Z}; \Psi^{[i]}) | \mathcal{X}, \Psi^{[i-1]} \right] \\ &= \sum_{t=1}^{T} \sum_{k=1}^{K} \left[log(\pi_k^{[i]} \varphi(x_t; \mu_k^{[i]}, \Sigma_k^{[i]})) \mathbb{E}_Z[1_{\{z_t=k\}} | \mathcal{X}, \Psi^{[i-1]}] \right] \\ &= \sum_{t=1}^{T} \sum_{k=1}^{K} \left[log(\pi_k^{[i]} \varphi(x_t; \mu_k^{[i]}, \Sigma_k^{[i]})) p(Z = z_t | x_t, \Psi^{[i-1]}) \right] \end{aligned} \tag{C.2}$$

Le terme $p(Z = z_t | x_t, \Psi^{[i-1]})$ obtenu représente la probabilité *a posteriori* d'être dans le régime $z_t = k$ sachant la réalisation x_t et les paramètres estimés à l'itération précédente $\Psi^{[i-1]}$.

En appliquant le théorème de Bayes avec les expressions (3.1), la probabilité $\gamma_{tk}^{[i]}$ est définie :

$$\begin{aligned} \gamma_{tk}^{[i]} &= p(Z = z_t | X, \Psi^{[i-1]}) \\ &= \frac{p(x_t | z_t = k, \Psi^{[i-1]})}{p(x_t | \Psi^{[i-1]})} \\ &= \frac{\pi_k^{[i-1]} \varphi(x_t; \mu_k^{[i-1]}, \Sigma_k^{[i-1]})}{\sum_{k=1}^{K} \pi_k^{[i-1]} \varphi(x_t; \mu_k^{[i-1]}, \Sigma_k^{[i-1]})} \end{aligned} \tag{C.3}$$

De ce fait, l' expression de la fonction Q s'écrit :

$$Q(\Psi^{[i]}, \Psi^{[i-1]}) = \sum_{t=1}^{T} \sum_{k=1}^{K} \left[\gamma_{tk}^{[i]} log(\pi_k^{[i]}) + \gamma_{tk}^{[i]} log(\varphi(x_t; \mu_k^{[i]}, \Sigma_k^{[i]})) \right] \tag{C.4}$$

C.1.2 Etape M

La nouvelle étape M de maximisation de la fonction Q par rapport au paramètre $\Psi^{[i]}$ est introduite. L'idée est de maximiser l'espérance conditionnelle calculée à l'étape E en rappelant que les paramètres $\{\pi_k^{[i]}\}_{1 \leq k \leq K}$ sont soumis à la contrainte $\sum_{k=1}^{K} \pi_k^{[i]} = 1$. Le

problème de maximisation de la fonction Q s'exprime :

$$\begin{aligned}\Psi^{*[i]} &= \underset{\Psi^{[i]}}{\arg\max} \quad Q(\Psi^{[i]}, \Psi^{[i-1]}) \\ &= \underset{\{\pi_k^{[i]}, \mu_k^{[i]}, \Sigma_k^{[i]}\}_{1\leq k\leq K}}{\arg\max} \quad \sum_{t=1}^{T}\sum_{k=1}^{K}\left[\gamma_{tk}^{[i]} log(\pi_k^{[i]}) + \gamma_{tk}^{[i]} log(\varphi(x_t; \mu_k^{[i]}, \Sigma_k^{[i]}))\right] \\ \text{s.t.} \sum_{k=1}^{K} \pi_k^{[i]} &= 1\end{aligned} \tag{C.5}$$

L'objectif (C.5) montre un problème d'optimisation de la fonction Q suivant $\Psi^{[i]}$ décomposable en une somme de deux fonctions, l'une dépendante des paramètres $\{\pi_k^{[i]}\}_{1\leq k\leq K}$ et l'autre des paramètres $\{\mu_k^{[i]}, \Sigma_k^{[i]}\}_{1\leq k\leq K}$.

Le premier paramètre maximisé est $\{\pi_k^{[i]}\}_{1\leq k\leq K}$. Si l'on considère la fonction Lagrangienne appliquée à la première partie du problème de maximisation (C.5), l'expression suivante est obtenue :

$$L(\lambda, \pi_1^{[i]}, ..., \pi_K^{[i]}) = \sum_{t=1}^{T}\left[\sum_{k=1}^{K}\gamma_{tk}^{[i]} log(\pi_k^{[i]})\right] + \lambda\left(\sum_{k=1}^{K}\pi_k^{[i]} - 1\right) \tag{C.6}$$

Résoudre l'équation du gradient nul de L selon l'ensemble $\pi_1^{[i]}, ..., \pi_K^{[i]}$ revient à résoudre (C.6) pour chaque π_k. On en déduit comme expression pour l'estimateur $\pi_k^{[i]}$ du paramètre π_k :

$$\pi_k^{[i]} = -\frac{\sum_{t=1}^{T}\gamma_{tk}^{[i]}}{\lambda} \tag{C.7}$$

Le multiplicateur Lagrangien λ reste à exprimer. Sur la fonction (C.5), γ représente à l'instant $t \in \{1, ..., T\}$ les probabilités normalisées *a posteriori*. De ce fait, pour chaque individu t la fonction γ est soumis à la contrainte $\sum_{k=1}^{K}\gamma_{tk}^{[i]} = 1$. En sommant sur k les deux côtés de l'équation (C.7), l'expression du paramètre λ devient $\lambda = -T$, donnant comme expression finale :

$$\pi_k^{[i]} = \frac{\sum_{t=1}^{T}\gamma_{tk}^{[i]}}{T} \tag{C.8}$$

Ensuite, l'équation de la seconde partie de la somme (C.5) est reprise avec les paramètres

des composantes gaussiennes $\mu_k^{[i]}, \Sigma_k^{[i]}$ et premièrement dérivée selon $\mu_k^{[i]}$:

$$\begin{aligned} \frac{\partial}{\partial \mu_k^{[i]}} \left[\sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_{tk}^{[i]} log(\varphi(x_t | \mu_k^{[i]}, \Sigma_k^{[i]})) \right] = 0 \\ \frac{\partial}{\partial \mu_k^{[i]}} \left[\sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_{tk}^{[i]} log(c) - \frac{1}{2}(x_t - \mu_k^{[i]})^\intercal \Sigma_k^{-1} (x_t - \mu_k^{[i]}) \right] = 0 \\ \frac{1}{2} \sum_{t=1}^{T} 2\gamma_{tk}^{[i]} {\Sigma_k^{[i]}}^{-1} (x_t - \mu_k^{[i]}) = 0 \end{aligned} \tag{C.9}$$

avec $c = (2\pi)^{\frac{d}{2}} det(\Sigma_k^{[i]})^{\frac{1}{2}}$ définissant une constante ne dépendant pas de $\mu_k^{[i]}$.

Une expression pour l'estimateur du vecteur moyen μ_k est déduite de l'équation précédente :

$$\mu_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} x_t}{\sum_{t=1}^{T} \gamma_{tk}^{[i]}} \tag{C.10}$$

Résoudre l'équation (C.5) selon $\Sigma_k^{[i]}$ nécessite de dériver des expressions complexes comme le déterminant de la matrice symétrique et positive $\Sigma_k^{[i]}$. La propriété suivante est introduite $det(\Sigma_k^{[i]}) = \frac{1}{det({\Sigma_k^{[i]}}^{-1})}$, permettant de dériver (C.5) à l'aide de ${\Sigma_k^{[i]}}^{-1}$. La dérivée du logarithme du déterminant de ${\Sigma_k^{[i]}}^{-1}$ s'exprime :

$$\begin{aligned} \frac{\partial log(det({\Sigma_k^{[i]}}^{-1}))}{\partial {\Sigma_k^{[i]}}^{-1}} &= \frac{det({\Sigma_k^{[i]}}^{-1})(2\Sigma_k^{[i]} - diag(\Sigma_k^{[i]}))}{det({\Sigma_k^{[i]}}^{-1})} \\ &= 2\Sigma_k^{[i]} - diag(\Sigma_k^{[i]}) \end{aligned} \tag{C.11}$$

De plus, $\forall (t,k) \in \{1..., T\} \times \{1..., K\}$, l'expression $(x_t - \mu_k^{[i]})^\intercal {\Sigma_k^{[i]}}^{-1} (x_t - \mu_k^{[i]})$ peut être récrite comme la trace entre ${\Sigma_k^{[i]}}^{-1}$ et $N_t^k = (x_t - \mu_k^{[i]})(x_t - \mu_k^{[i]})^\intercal$. Cette expression comportant le paramètre ${\Sigma_k^{[i]}}^{-1}$ se voit aussi dériver de la façon suivante :

$$\begin{aligned} \frac{\partial tr({\Sigma_k^{[i]}}^{-1} N_t^k)}{\partial {\Sigma_k^{[i]}}^{-1}} &= N_t^k + N_t^{k\intercal} - diag(N_t^k) \\ &= 2N_t^k - diag(N_t^k) \end{aligned} \tag{C.12}$$

En repartant de (C.11) avec la dérivée partielle ${\Sigma_k^{[i]}}^{-1}$ et les expressions définies pré-

cédemment, il est obtenu :

$$\begin{aligned}
\frac{\partial}{\partial {\Sigma_k^{[i]}}^{-1}} \left[\sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_{tk}^{[i]} log(\varphi(x_t|\mu_k^{[i]}, \Sigma_k^{[i]})) \right] &= 0 \\
\frac{\partial}{\partial {\Sigma_k^{[i]}}^{-1}} \left[\frac{1}{2} \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_{tk}^{[i]} (log(a) + log(det({\Sigma_k^{[i]}}^{-1})) - tr({\Sigma_k^{[i]}}^{-1} N_t^k)) \right] &= 0 \\
\sum_{t=1}^{T} \gamma_{tk}^{[i]} \left[2\Sigma_k^{[i]} - diag(\Sigma_k^{[i]}) - 2N_t^k + diag(N_t^k) \right] &= 0 \\
\sum_{t=1}^{T} \gamma_{tk}^{[i]} \left[\Sigma_k^{[i]} - N_t^k \right] &= 0
\end{aligned} \tag{C.13}$$

avec $a = (2\pi)^{\frac{d}{2}}$ renvoyant une constante ne dépendant pas de $\Sigma_k^{[i]}$.

Ainsi, l'estimateur de la matrice de covariance Σ_k s'exprime :

$$\Sigma_k^{[i]} = \frac{\sum_{t=1}^{T} \gamma_{tk}^{[i]} (x_t - \mu_k^{[i]})(x_t - \mu_k^{[i]})^\intercal}{\sum_{t=1}^{T} \gamma_{tk}^{[i]}} \tag{C.14}$$

Les travaux de Redner et Walker 1984 montrent que la convergence linéaire de $\Psi^{[i]}$ à chaque itération i vers un ensemble de paramètres Ψ^* est assurée. En composant avec la log-vraisemblance incomplète, la fonction Q (C.6) et l'inégalité de Gibbs, il est possible de montrer l'inégalité entre la différence de vraisemblance du mélange gaussien et la différence des fonctions Q suivant les paramètres de l'itération actuelle i et précédente $i-1$: $\log(\mathcal{L}(\Psi^{[i]})) - \log(\mathcal{L}(\Psi^{[i-1]})) \geq Q(\Psi^{[i]}, \Psi^{[i-1]}) - Q(\Psi^{[i-1]}, \Psi^{[i-2]})$. Sous regard de l'expression précédente, à chaque itération i, Q est maximisée suivant $\Psi^{[i]}$ de telle manière que $Q(\Psi^{[i]}, \Psi^{[i-1]}) \geq Q(\Psi^{[i-1]}, \Psi^{[i-2]})$. Cette propriété implique et garantit la convergence monotone de la vraisemblance le long des itérations EM. Néanmoins, la log-vraisemblance du problème n'est pas convexe et donc l'algorithme ne peut qu'approcher des optimums locaux sous regards des hypothèses des données et de l'initialisation des paramètres Ψ^0.

C.2 Compléments de modèle

Dans cette sous-partie d'annexe, un exemple d'évolution des probabilités a posteriori γ suivant différentes tailles d'ensembles M est montré sur la figure C.1 et ce $\forall t \in \{1, \cdots, T = 10000\}$. Cet exemple est construit à l'aide d'ensemble généré suivant deux composantes gaussiennes univariées $\mathcal{N}(\mu_1 = -1, \sigma_1^2 = 1)$ et $\mathcal{N}(\mu_2 = -1, \sigma_2^2 = 0.7)$ avec des probabilités

a priori égales $\pi_1 = 0.5$ et $\pi_2 = 0.5$.

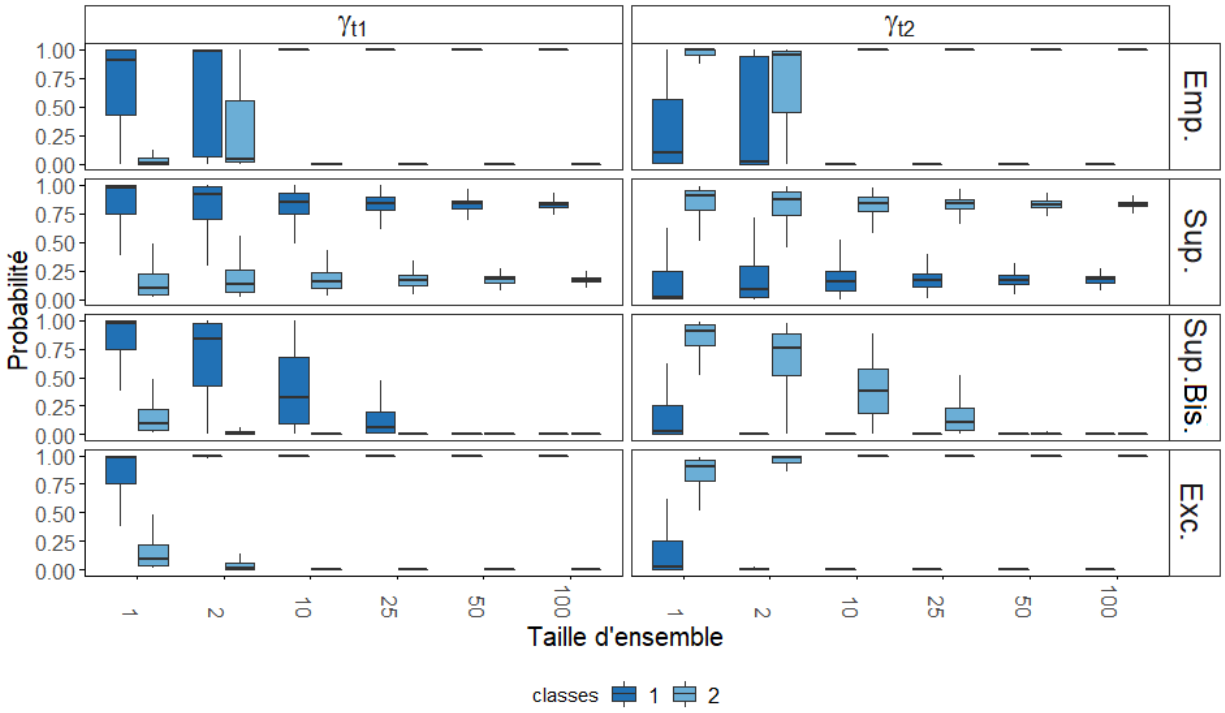


FIGURE C.1 – Evolution de la probabilité a posteriori γ par modèle pour un mélange gaussien à $K = 2$ composantes et pour différentes tailles d'ensemble avec une taille d'échantillon fixé à $T = 10000$. *Emp. : modèle avec statistiques empiriques, Sup. : modèle transformant l'ensemble en super échantillon, Sup. Bis. : modèle considérant les ensembles du super échantillon comme associé à une même classe, Exc. : modèle avec vecteur de variables gaussiennes échangeables.*

L'évolution des probabilités *a posteriori* γ est affiché sur la figure C.1 pour chaque modèle, et ce, pour différentes valeurs de tailles d'ensembles $M \in \{1, 2, 10, 25, 50, 100\}$. Les probabilités des modèles surnommés 'Emp.', 'Sup.', 'Sup. Bis.' et 'Exc.' correspondent dans l'ordre aux probabilités (3.8) du modèle aux statistiques empiriques, (3.12) du modèle au "super échantillon", (3.16) du modèle au "super échantillon" d'ensemble associé à une même classe et (3.18) du modèle au vecteur de variables échangeables (présenté dans la section suivante 3.1.2.3). Sur cette même figure, les probabilités des modèles 'Emp.', 'Sup.' et 'Exc.' conservent leur complémentarité lorsque M augmente sans déroger à la contrainte de la somme des probabilités selon K égale à 1. Quant aux résultats associés au modèle 'Sup. Bis.', ils laissent paraître une "dégénérescence" des probabilités (3.16) qui tendent vers 0 suivant l'augmentation du nombre de membres. Ces probabilités transgressent la

contrainte imposée $\sum_{k=1}^{K} \gamma_{tk}^{[i]} = 1$ rendant impossible d'obtenir une estimation valide des paramètres.

C.3 Annexe des résultats de simulation

C.3.1 Vues des composantes gaussiennes trivariées

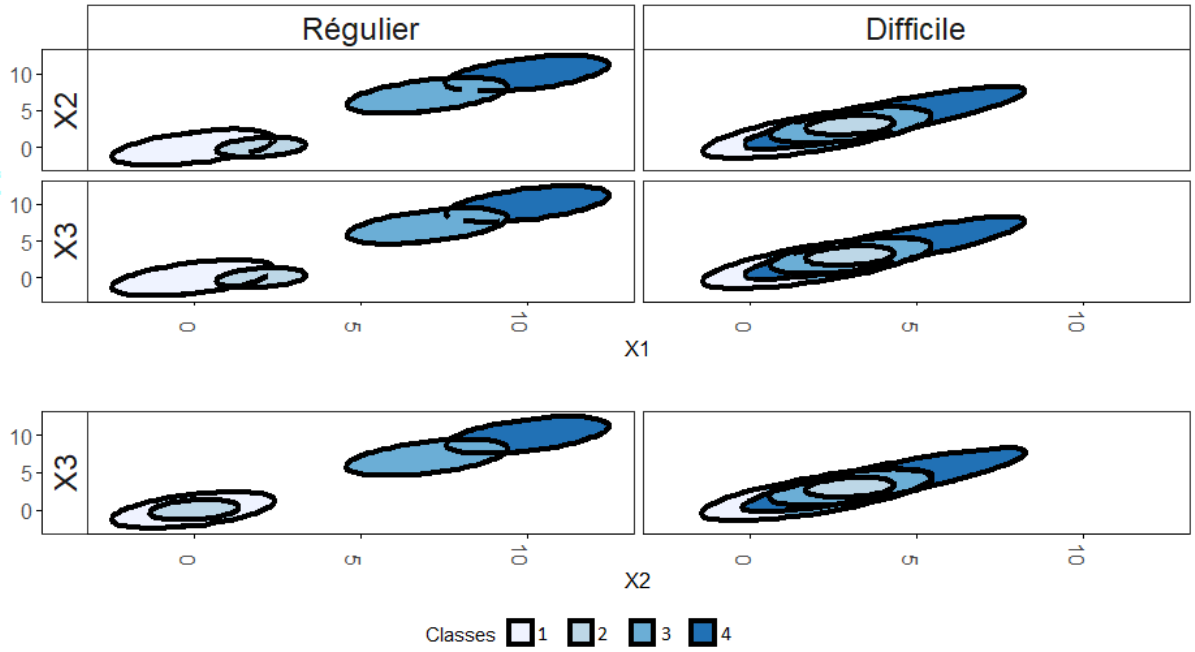


FIGURE C.2 – Coupes deux dimensions des $K = 4$ distributions gaussiennes trivariées simulées dans le cas "Régulier" et "Difficile".

C.3.2 Critère de sélection de modèles

Dans cette sous-partie, un critère de sélection de modèles est présenté permettant de sélectionner un nombre de classes K adéquat. La question de la sélection du nombre de classes est une problématique récurrente au sein de la communauté des modèles de mélange (BOZDOGAN 1993, M.-S. YANG et NATALIANI 2017, LEROUX 1992) touchant les approches de classification non supervisée. L'idée est de trouver un critère dont l'optimum permet de sélectionner un nombre K^* de classes correspondant au mieux au modèle et jeu

de données étudié. Pour cela, le modèle de mélange $\mathcal{M}_K$ ajusté avec K classes est définis tel que $\mathcal{M}_1$,...,$\mathcal{M}_K$,...$\mathcal{M}_{K_{max}}$ forme une séquence de modèles ajustés selon une séquence de nombre de classes $K \in \{1, ..., K_{max}\}$. Wong 1982 a montré par simulation qu'un des choix possible pour K_{max} est $(T)^{0.3}$. Là où Bozdogan 1993 déduits que $K_{max} = (\frac{T}{2})^{\frac{1}{2}}$ en se basant sur l'inégalité reliant le nombre d'observations et le nombre de paramètres à respecter pour éviter une surparamétrisation du modèle de mélange. La règle de Wong 1982 tends à être plus amplement plus faible que celle de Bozdogan 1993 lorsque T augmente ce qui peut justifier de choisir $K_{max} = (T)^{0.3}$ pour des raisons computationnelles.

Dans la continuité des travaux d'Akaike 1973, Schwarz et al. 1978 définit un critère basé sur l'intégration de la distribution jointe entre les données et l'ensemble des paramètres du modèle dans un cadre bayésien. Ce critère nommé BIC ("Bayesian Information Criterion") est donné par :

$$BIC(K) = -2log(\mathcal{L}(\Psi^{MLE})) + \nu_K log(TM) \tag{C.15}$$

T étant le nombre d'individus et M le nombre de membres échangeables de l'ensemble de la base de données $\mathcal{X}$ et $\mathcal{L}$ la vraisemblance du modèle utilisé. Leroux 1992 et McLachlan et Peel 2004 montrent que le BIC parvient généralement à sélectionner le nombre de composantes K du mélange. Il est aussi important de considérer les travaux de Biernacki, Celeux et Govaert 2000 où ils dévoilent au travers de simulation d'un mélange d'une loi gaussienne et d'une loi uniforme que dans un cadre de modèle ne correspondant pas aux données, le BIC surestime le nombre de composantes fixé à 2. Baudry 2009 suppose que le BIC tend à minimiser la divergence de Kullback-Leibler forçant un grand nombre de composantes à se rapprocher de la distribution des données. Sous les conditions particulières de régularité et d'identifiabilité du mélange gaussien, Keribin 2000 conclut à l'aide de simulations que le critère BIC est consistent et tend à estimer le bon nombre de composantes K du mélange.

C.3.3 Choix de la méthode d'initialisation

Plusieurs méthodes d'initialisation de l'algorithme EM du modèle de mélange gaussien sont présentées dans cette section. Puis, l'idée est d'évaluer ces méthodes à travers les expérimentations (3.1.3.1). L'objectif de cette évaluation est de sélectionner la meilleure méthode pour initialiser et évaluer les extensions de modèles de mélange gaussien proposées.

C.3.3.1 Initialisation de l'EM

Suivant le choix de la méthode d'initialisation, les paramètres initiaux estimés Ψ^0 peuvent se retrouver proche d'un attracteur local situé loin des vrais paramètres. Celeux et Govaert 1992 dévoilent que dans une configuration de mélange gaussien aux covariances sphériques et de volumes égaux, une estimation EM revient à effectuer un **k-means** (Lloyd 1982) sur les données. Il est naturel et fréquent de choisir la méthode du *k-means* comme méthode d'initialisation des paramètres de l'EM.

L'algorithme *k-means* bénéficie d'une grande simplicité d'application aux problèmes de classification non supervisée. Néanmoins, cet algorithme est initialisé en prenant aléatoirement K individus de $\mathcal{X}$ comme centres initiaux pour débuter son partitionnement. Cette initialisation aléatoire peut amener à un attracteur local éloigné des vrais paramètres sous certaines configurations. Arthur et Vassilvitskii 2006 proposent une version améliorée nommée k-means++ (**kM**) avec comme objectif de résoudre les problèmes d'attractions d'optimums locaux suite à l'initialisation non contrôlée de l'algorithme. L'idée étant d'initialiser le premier centre μ_1 avec un individu de $\mathcal{X}$ sélectionné aléatoirement. Ensuite, le centre suivant $1 < k \leq K$: μ_k est initialisé avec l'individu le plus éloigné de l'ensemble des centres précédents $\{\mu_{k'}\}_{1 \leq k' < k}$. Le critère d'éloignement est la maximisation d'une similarité entre le centre précédent et les individus, le tout au carré et normalisé.

En autre méthode d'initialisation basée sur les distances entre individus, l'algorithme de classification hiérarchique (**HC**) de Fraley 1998 permet d'obtenir de performants résultats d'estimation. Cette technique repose sur le calcul de la matrice de dissimilarité des individus de la base de données. Ensuite, ces individus sont partitionnés en K classes suivant leur plus proche voisin minimisant les distances de la matrice. Une initialisation des estimateurs est ensuite tirée de chacun de ces groupes. Le calcul de la matrice de dissimilarité peut être coûteux avec un grand jeu de données. Dans ce cas, l'algorithme HC est appliqué sur différents sous-échantillons. Puis, l'initialisation donnant les meilleurs paramètres (celle maximisant la la vraisemblance du modèle de mélange) est sélectionnée. Cette technique ayant fait ses preuves est couramment déployée dans l'algorithme MCLUST d'ajustement de modèle de mélange gaussien (Scrucca et al. 2016).

Biernacki, Celeux et Govaert 2003 proposent **Small EM** en méthode d'initialisation alternative basée sur l'EM du modèle étudié. Cette approche consiste en un lancement d'un grand nombre N d'EM avec un faible nombre d'itérations (généralement $I_{max} \leq 5$) et prenant une initialisation aléatoire. Ensuite, le modèle $\mathcal{M}^*$ maximisant l'ensemble des log-vraisemblances $\{\mathcal{L}_n(\Psi^{MLE})\}_{1 \leq n \leq N}$ des N modèles ajustés est sélectionné

le modèle. Le but n'est pas de faire converger les N modèles mais plutôt d'étudier lequel obtient la meilleure direction. À partir de la direction fournie par les meilleurs paramètres sélectionnés par Small EM, l'EM est initialisée avec un nombre d'itérations plus grand maximisant les chances de converger vers les vrais paramètres.

Celeux 1985 apporte une version d'initialisation différente et toujours basée sur l'EM appelée **stochastique EM (SEM)**, une approche Monte-Carlo appliquée à l'estimation des paramètres initiaux. Dans SEM, une étape intermédiaire dite stochastique "S" se place avant l'étape E. L'étape S tire aléatoirement un échantillon z_t, $\forall t \in \{1, ..., T\}$, suivant une variable multinomiale Z_t au vecteur de probabilités $(\gamma_1(t), ..., \gamma_1(K))^\intercal$, $\forall k \in \{1, ..., K\}$. $\gamma_k(t)$ représente la probabilité *a posteriori* définie en (C.3). De ce fait, SEM peut être vu comme une méthode générant une chaîne de Markov stationnaire avec une distribution indépendante de son état initial. Les paramètres maximisant la log-vraisemblance après un certain d'itération SEM sont utilisés pour initialiser l'EM finale.

C.3.3.2 Évaluation des méthodes

Cette sous-partie se concentre sur l'évaluation des méthodes d'initialisation de l'algorithme EM des modèles introduits en section 3.1.2. Les méthodes kM, HC, Small EM, SEM sont comparées au travers des expérimentations décrites en 3.1.3.1. Les résultats de sélection du nombre de classes, la performance d'estimation et la précision de classification sont comparés et discutés entre méthodes d'initialisation. L'algorithme kM et HC ne sont pas prévus pour prendre directement les données d'ensemble empêchant d'initialiser les modèles de mélange gaussien au vecteur de variables échangeables. Pour cela, les techniques basées sur les statistiques empiriques et la mise en place de "super-échantillon" élaborées en section 3.1.2.1 et 3.1.2.2 sont appliquées pour l'ajustement des algorithmes kM et HC. Ces modèles d'initialisations porteront les noms suivants **kM Emp.**, **HC Emp.** pour l'initialisation à partir de données statistiques et **kM Sup.**, **HC Sup.** pour celles prenant un "super échantillon" de données d'ensemble. Pour les modèles de mélange gaussien *Empirical statistics* et *Super Sample*, les deux méthodes d'initialisation kM et HC sont directement appliquées aux données. Chacune de ces méthodes est étudiée sur ses performances de sélection du nombre de classes et ses capacités à fournir des estimations correctes des vrais paramètres.

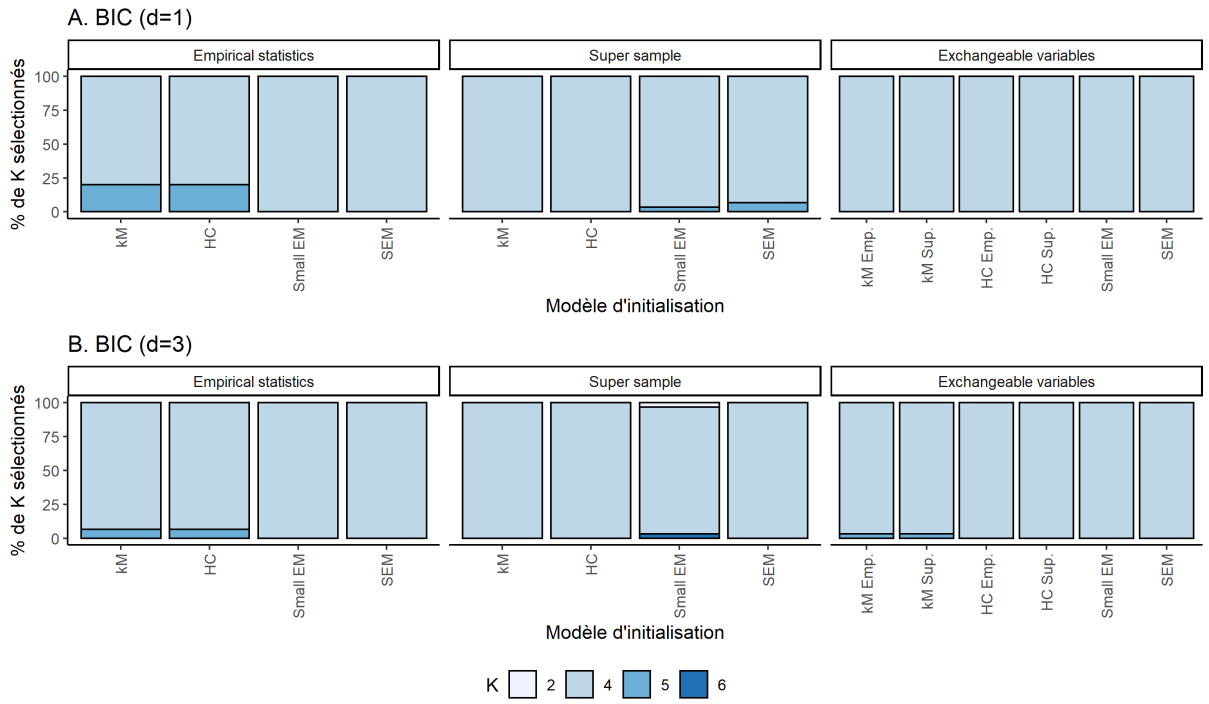


FIGURE C.3 – Sélection du nombre de classes basée sur le score BIC pour le cas "Régulier" par modèle d'initialisation pour des longueurs d'échantillon et d'ensemble fixées $T = 200$ et $M = 50$. *kM : algorithme k-means++ ; HC : méthode de classification hiérarchique ; kM Emp, HC Emp : méthodes prenant les statistiques empiriques des ensembles en données ; kM Sup, HC Sup : méthodes transformant les données d'ensemble de dimension* $T \times M \times d$ *en super échantillon de dimension* $(T \times M) \times d$.

Sélection du nombre de classes. La majeure partie des méthodes d'initialisation pour les différents modèles de mélanges performe une sélection de classe correcte dans un cas "Régulier" de données simulées présenté sur la figure C.3. Dans le cas de composantes gaussiennes difficiles à identifier, les résultats affichés en figure C.4 montrent plus de variétés dans le nombre de classes sélectionné par les modèles d'initialisation. Notamment, les méthodes HC et Small EM obtiennent des sélections correctes pour le modèle de statistiques empiriques (*Empirical statistics*). Le modèle au vecteur de variables échangeables (*Exchangeable variables*) montre aussi de bons résultats avec les méthodes HC prenant des statistiques empiriques, mais aussi avec les méthodes Small EM, SEM et kM prenant un "super-échantillon". Quant au modèle basé sur le "super-échantillon" (*Super sample*), une sous-estimation du nombre de classes est observée et sera abordée plus en détail dans la sous-partie 3.1.3.2.

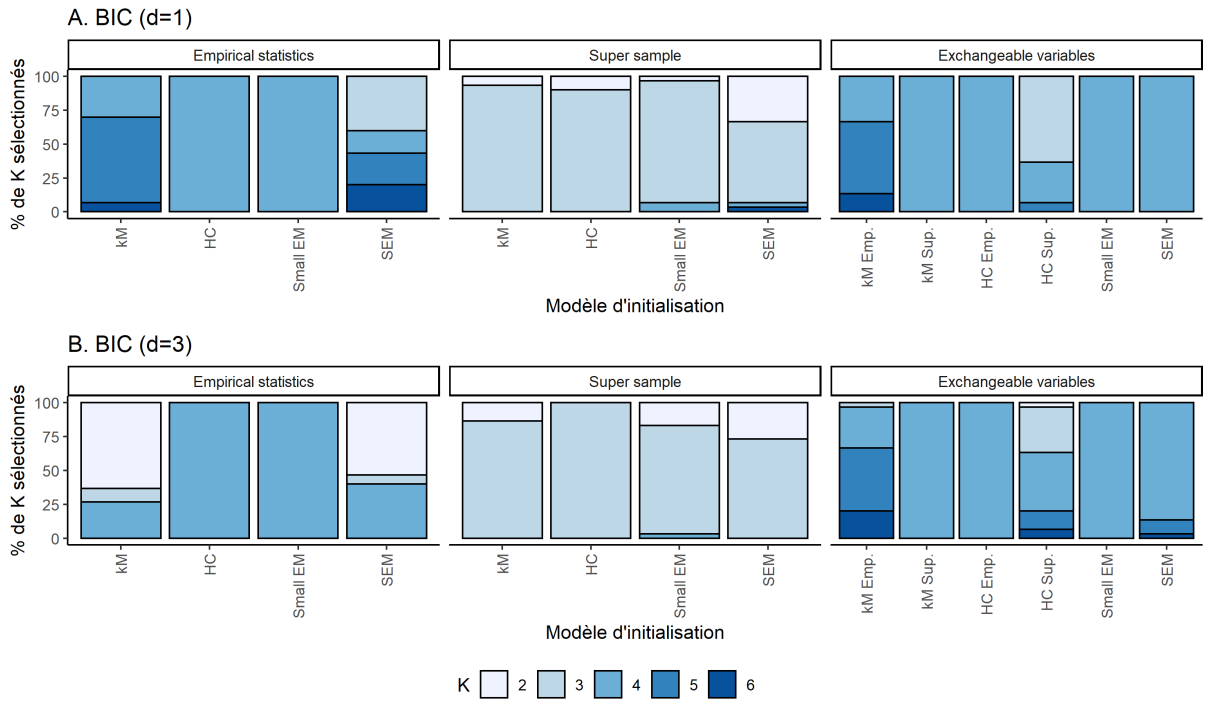


FIGURE C.4 – Sélection du nombre de classes basée sur le score BIC pour le cas "Difficile" par modèle d'initialisation pour des longueurs d'échantillon et d'ensemble fixées $T = 200$ et $M = 50$. *kM : algorithme k-means++ ; HC : méthode de classification hiérarchique ; kM Emp, HC Emp : méthodes prenant les statistiques empiriques des ensembles en données ; kM Sup, HC Sup : méthodes transformant les données d'ensemble de dimension* $T \times M \times d$ *en super échantillon de dimension* $(T \times M) \times d$.

La méthode *k-means++* obtient des résultats moins performants pour le modèle avec les statistiques empiriques en surestimant en dimension univariée et sous-estimant en trivarié. Cet effet de la dimension sur l'estimation sera plus exploité dans la section d'évaluation des modèles. La méthode *k-means++* étant initialisée sur une recherche des centres les plus éloignés tend à privilégier une séparation de l'espace des ensembles plus par leurs moyennes que leurs variances. Ce modèle est moins adapté à une problématique de recherche de classes ayant des moyennes proches et variances différentes. L'algorithme SEM montre aussi de faibles résultats pour ce modèle. SEM est initialisé en associant des classes aléatoirement à chaque individu pour ensuite en extraire des paramètres. Cette initialisation renvoie à l'itération 0 des paramètres très proches et sensibles à la dimension des données pouvant rapidement mener à des optimums locaux rendant les résultats d'EM erronées.

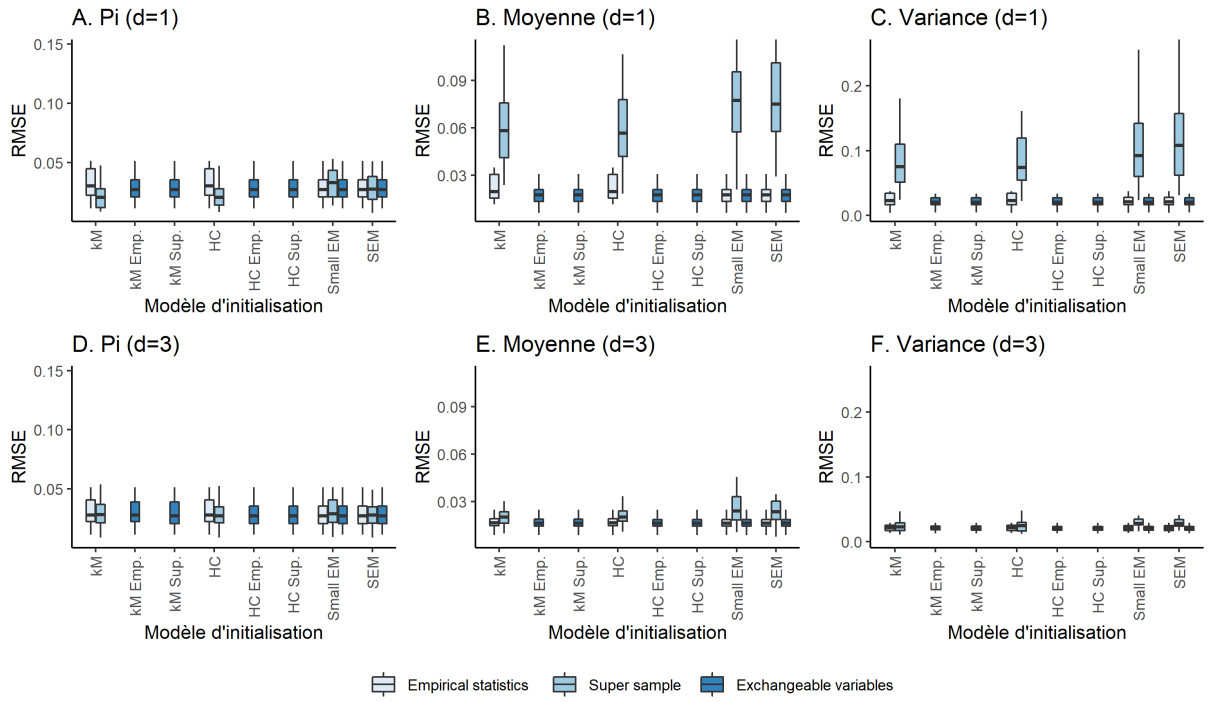


FIGURE C.5 – Erreur sur l'estimation des paramètres pour le cas "Régulier" avec des longueurs d'échantillon et d'ensemble fixées $T = 200$ et $M = 50$. *kM : algorithme k-means++ ; HC : méthode de classification hiérarchique ; kM Emp, HC Emp : méthodes prenant les statistiques empiriques des ensembles en données ; kM Sup, HC Sup : méthodes transformant les données d'ensemble de dimension $T \times M \times d$ en super échantillon de dimension $(T \times M) \times d$.*

Evaluation des estimateurs. Les résultats des RMSE du cas "Régulier" représentés dans la figure C.5 font état de référence et montre des performances similaires entre méthodes d'initialisation pour chaque modèle. Contrairement au cas "Régulier", les résultats d'estimation pour les données du cas "Difficile" montrés dans les RMSE de la figure C.6 affichent une importante variabilité des performances suivant les méthodes d'initialisation.

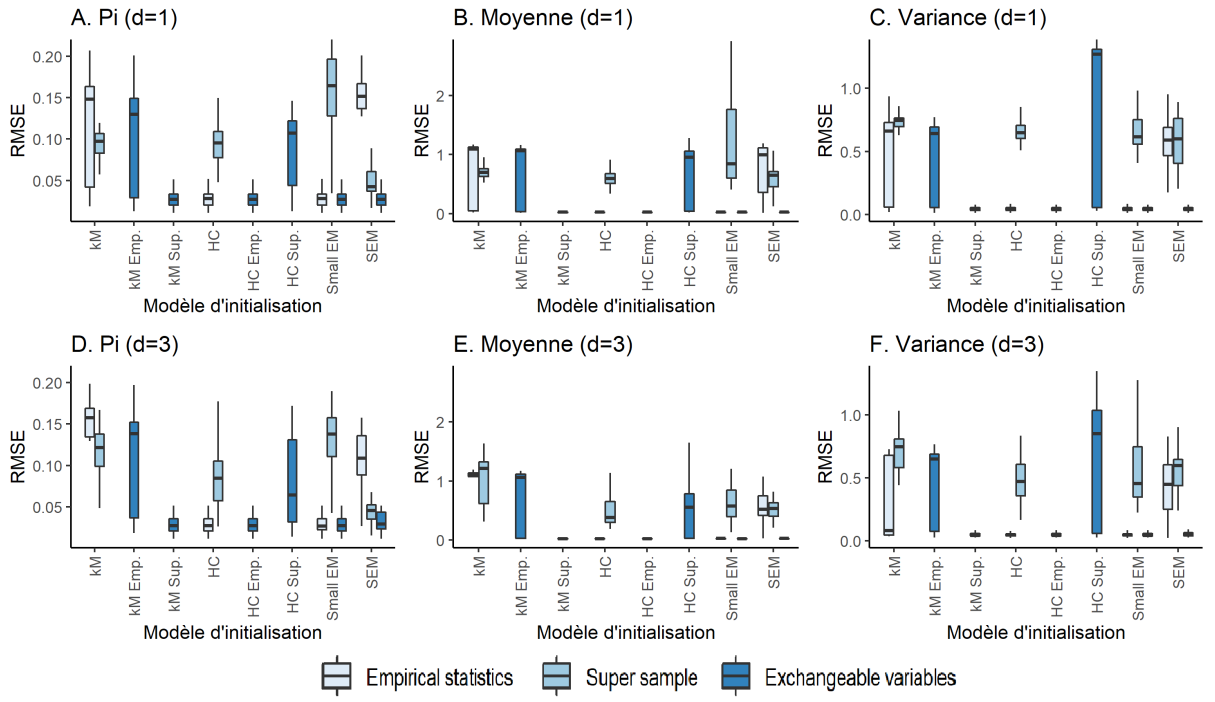


FIGURE C.6 – Erreur sur l'estimation des paramètres pour le cas "Régulier" avec des longueurs d'échantillon et d'ensemble fixées $T = 200$ et $M = 50$. *kM : algorithme k-means++ ; HC : méthode de classification hiérarchique ; kM Emp, HC Emp : méthodes prenant les statistiques empiriques des ensembles en données ; kM Sup, HC Sup : méthodes transformant les données d'ensemble de dimension* $T \times M \times d$ *en super échantillon de dimension* $(T \times M) \times d$.

Il devient plus évident sur cette figure que les algorithmes *k-means++* et SEM obtiennent les plus fortes RMSE parmi les différents paramètres estimés et dimensions ainsi que les plus faibles estimations pour les trois modèles étudiés. Les méthodes HC et Small EM montrent toujours des performances d'estimations intéressantes pour le modèle avec les statistiques empiriques. Dans le cas du modèle au "super-échantillon", seule l'initialisation HC permet d'acquérir des performances correctes d'estimations pour les deux dimensions. Enfin, des candidats d'initialisations commencent à sortir du lot pour le modèle *Exchangeable variables*. En effet, la méthode HC prenant des statistiques empiriques et la méthode Small EM affichent de bons résultats parmi les estimateurs pour les différentes dimensions.

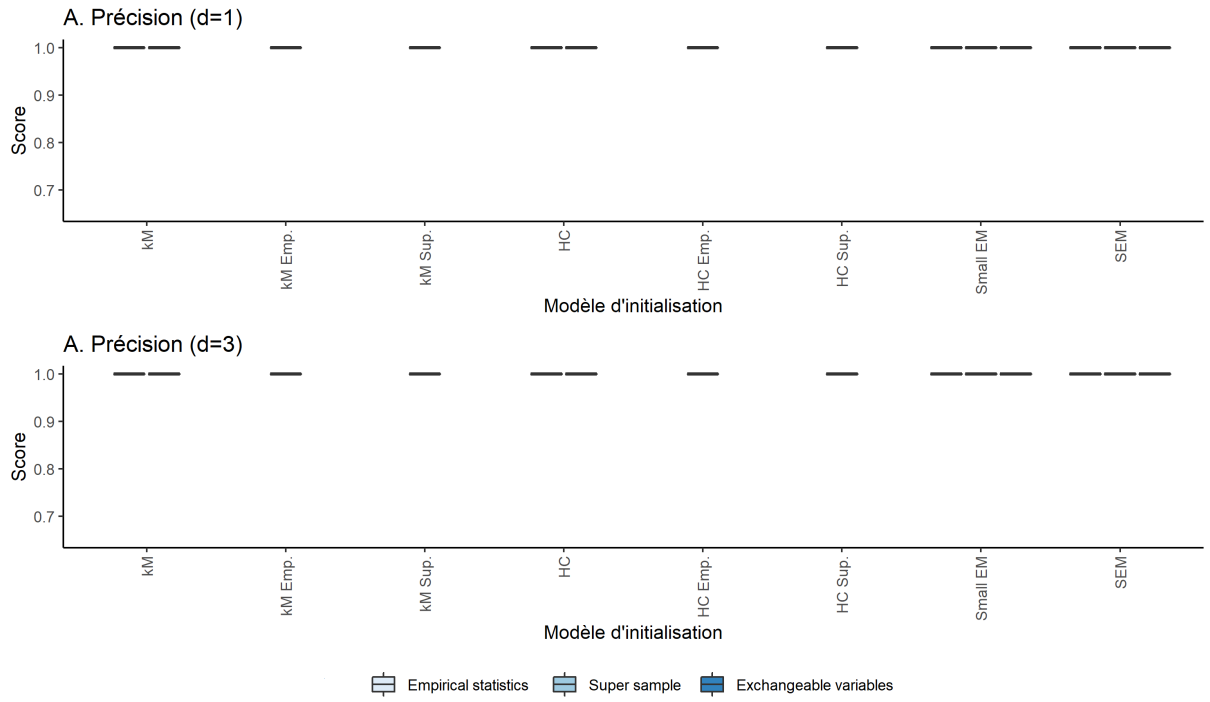


FIGURE C.7 – Score de précision globale pour le cas "Régulier" par modèle d'initialisation pour des longueurs d'échantillon et d'ensemble fixées $T = 200$ et $M = 50$. *kM : algorithme k-means++ ; HC : méthode de classification hiérarchique ; kM Emp, HC Emp : méthodes prenant les statistiques empiriques des ensembles en données ; kM Sup, HC Sup : méthodes transformant les données d'ensemble de dimension $T \times M \times d$ en super échantillon de dimension $(T \times M) \times d$.*

Performance de classification. Dans un cas "Régulier", les précisions globales de la figure C.7 sont affichées par méthodes d'initialisation des modèles de mélanges et dimension de variable pour une longueur de jeu de données et nombre de membres respectivement fixés à $T = 200$ et $M = 50$. Après analyse des scores de précisions, les méthodes d'initialisation montrent des modèles de mélange assignant correctement les classes pour chaque individu de la base.

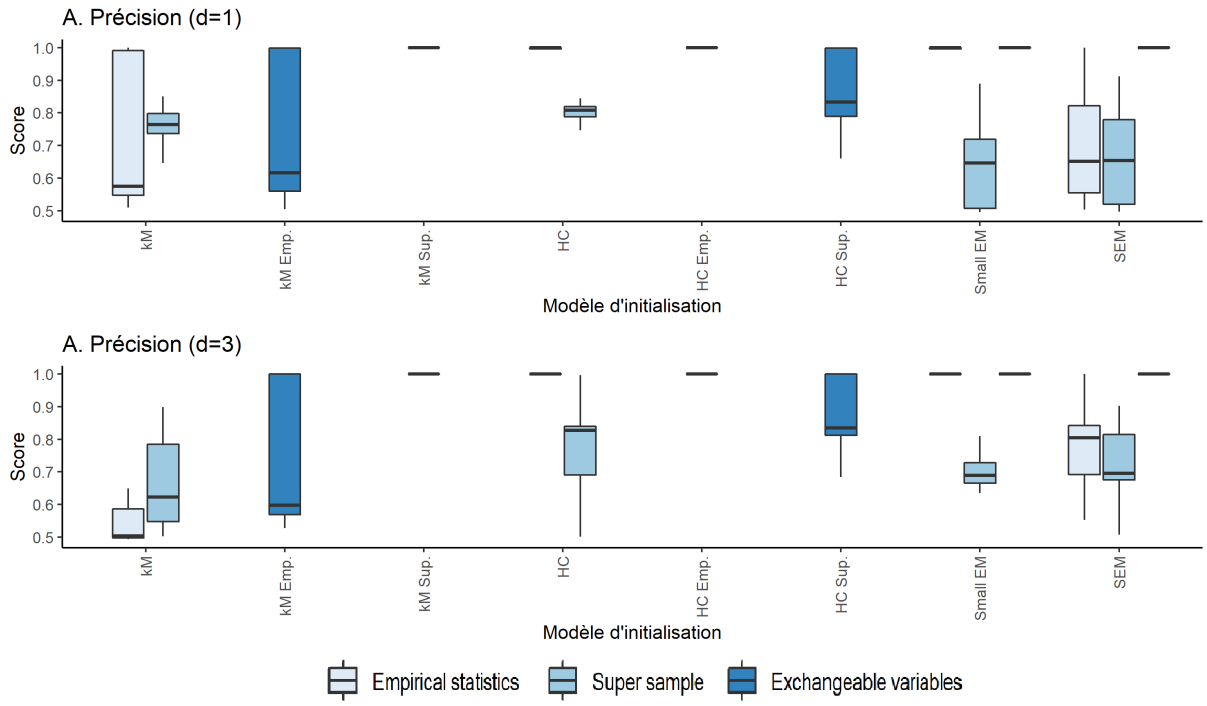


FIGURE C.8 – Score de précision globale pour le cas "Difficile" par modèle d'initialisation pour des longueurs d'échantillon et d'ensemble fixées $T = 200$ et $M = 50$. *kM : algorithme k-means++ ; HC : méthode de classification hiérarchique ; kM Emp, HC Emp : méthodes prenant les statistiques empiriques des ensembles en données ; kM Sup, HC Sup : méthodes transformant les données d'ensemble de dimension* $T \times M \times d$ *en super échantillon de dimension* $(T \times M) \times d$.

Dans le cas "Difficile", les performances de classification montrées dans la figure C.8 affichent des conclusions similaires à celles faites pour l'estimation des RMSE. Plus particulièrement, les méthodes HC (HC Emp. pour le modèle au vecteur de variables échangeables) et Small EM montrent les meilleurs scores parmi les modèles *Empiricals statistics*, *Super sample* et *Exchangeables variables* et ce pour les deux dimensions de variables évaluées. Les algorithmes kM et SEM montrent des scores plus faibles pour le modèle de mélange gaussien aux statistiques empiriques en accord avec les RMSE supérieurs aux autres méthodes.

En conclusion, les méthodes HC, HC Emp. et Small EM initialisant les modèles aux statistiques empiriques, "super-échantillon" et au vecteur de variables échangeables montrent les meilleurs résultats d'estimation. Cependant, l'algorithme de classification hiérarchique (HC) offre une initialisation plus stable pour le modèle au "super-échantillon" que le modèle Small EM. De plus, l'algorithme HC présente une facilité de déploiement,

tandis que Small EM dépend des performances de computations de l'EM mis en place. Pour la suite du chapitre, la méthode d'initialisation sélectionnée est l'algorithme de classification hiérarchique (HC). Dans le cas de l'initialisation du modèle au vecteur de variables échangeables, la méthode HC est utilisée avec les statistiques empiriques d'ensemble.

C.3.4 Compléments de résultats de simulation des modèles de mélange gaussien

Dans un cadre de méthode d'initialisation fixée, des compléments de résultats d'ajustements des paramètres par les modèles de mélange étudiés sont présentés ici, suivant différentes expérimentations.

C.3.4.1 Cas "Régulier" avec des tailles d'ensembles variables

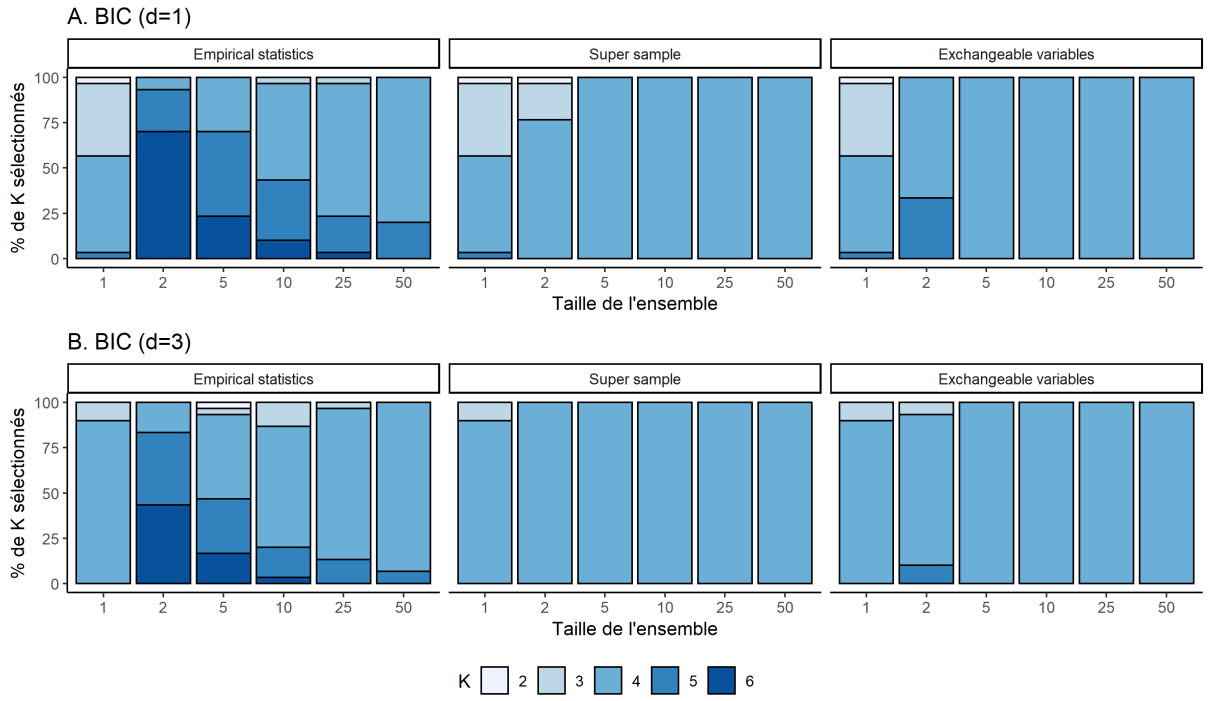


FIGURE C.9 – Sélection du nombre de classes à l'aide du score BIC dans le cas "Régulier" de données simulées pour différentes valeurs de tailles d'ensembles et une taille d'échantillon fixée $T = 200$.

C.3.4.2 Cas "Difficile" avec des tailles d'échantillon variables

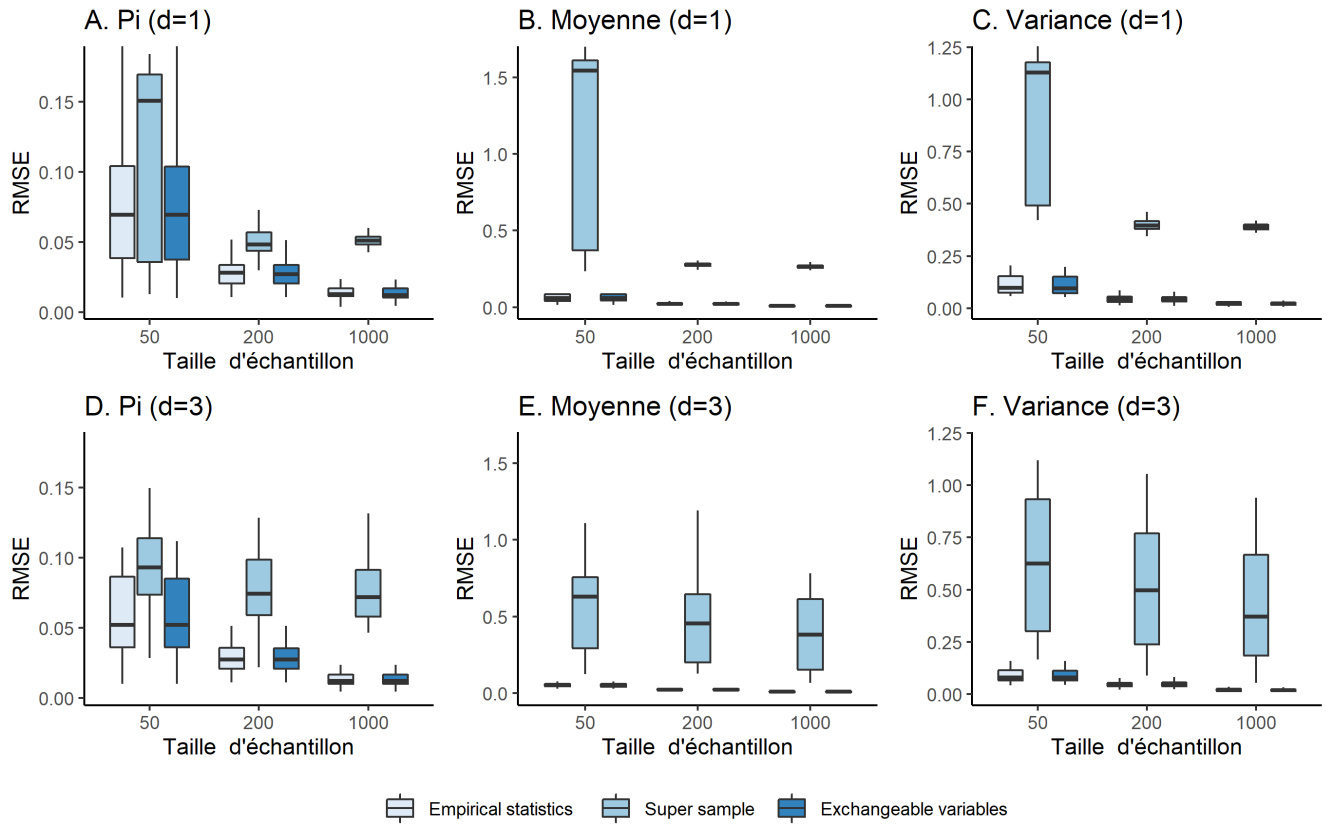
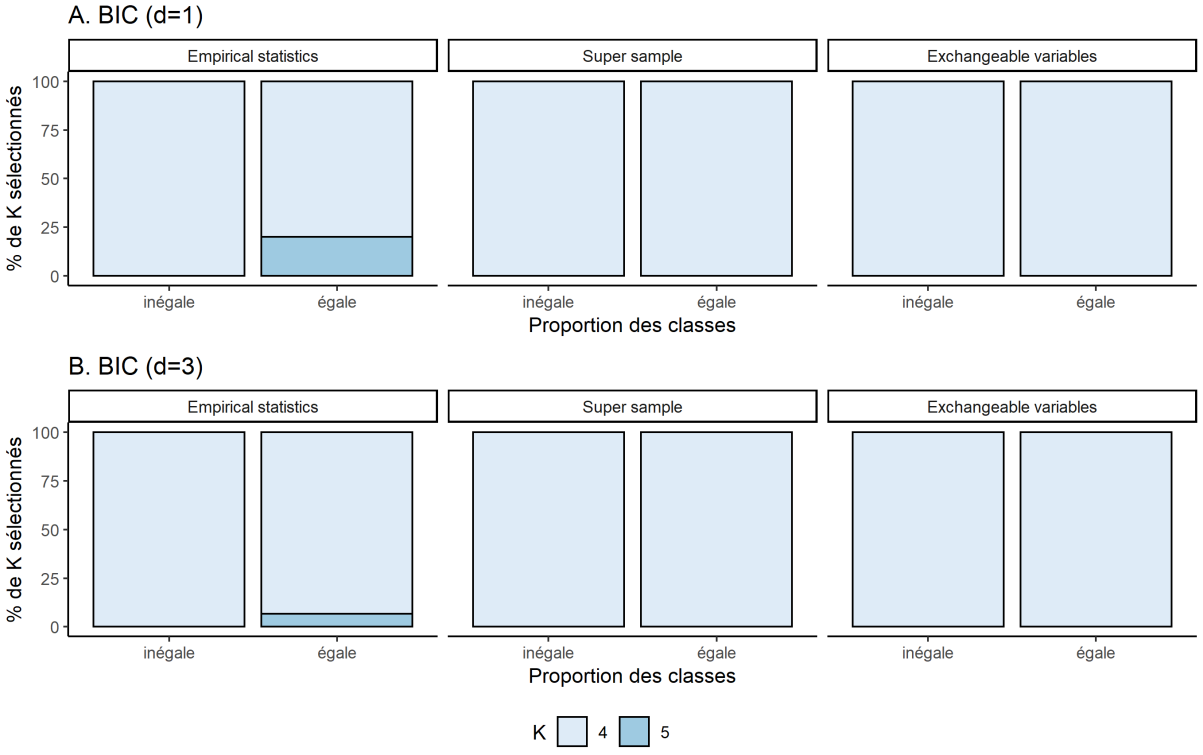


FIGURE C.10 – Erreur sur l'estimation des paramètres des modèles dans le cas "Difficile" de données simulées pour différentes valeurs de tailles d'échantillon et une taille d'ensemble fixée $M = 50$.

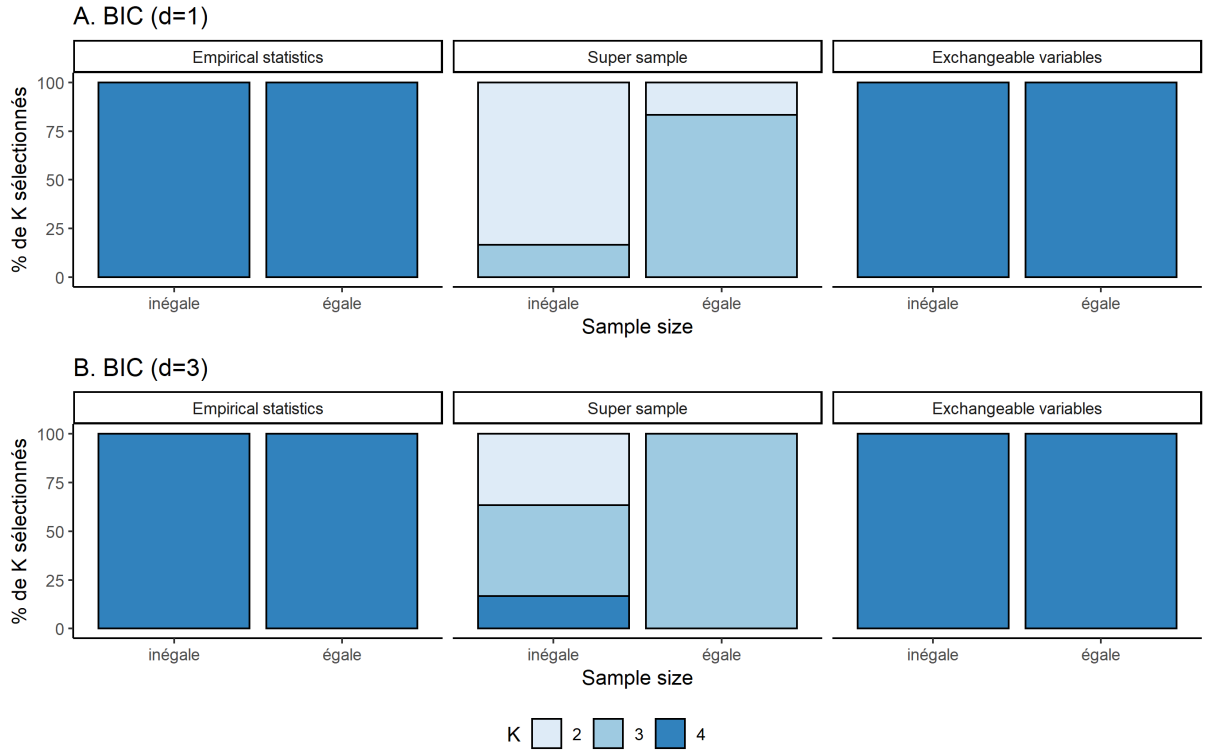
C.3.4.3 Sous-groupes aux proportions variables

Dans cette partie, les extensions de modèle de mélange gaussien sont évaluées dans un cadre de données avec différentes proportions. De ce fait, un ensemble aux proportions égales les unes des autres et un aux proportions inégales $\pi \in \{(0.25, 0.25, 0.25, 0.25), (0.2, 0.3, 0.1, 0.4)\}$ sont mis en place.

Sélection du nombre de classes. Les résultats de sélection sont présentés par modèles et dimensions suivant des proportions de classes différentes sur la figure C.11a dans un cas "Régulier", et sur la figure C.11b dans un cas "Difficile". Lorsque les proportions des individus entre classes deviennent inégales dans le cas "Régulier", la sélection dans un cas "Régulier" ne semble pas être perturbée et reste correcte entre modèles. Seuls les modèles de mélange gaussien au vecteur de variables échangeables et aux statistiques empiriques montrent de bonnes performances de sélection du nombre de classes dans un cas "Difficile". Alors que pour le modèle de mélange gaussien *Super Sample*, les résultats affichent des difficultés à retrouver le bon nombre K en le sous-estimant.



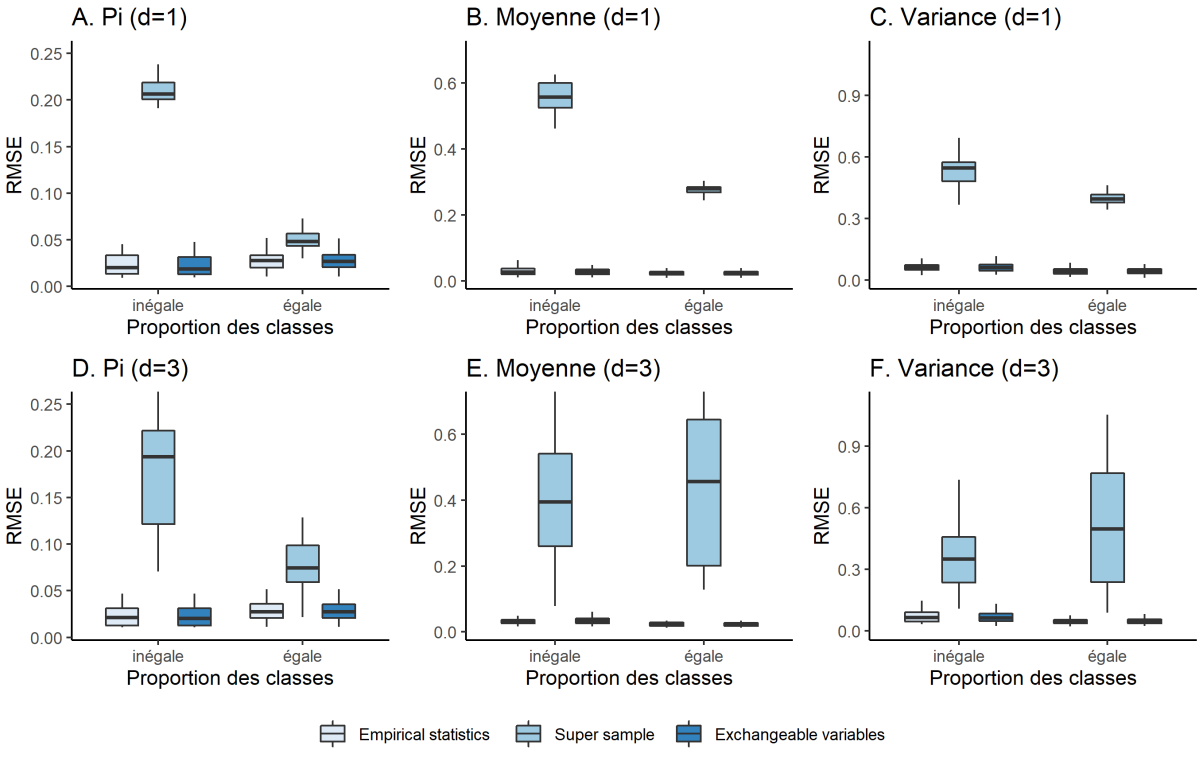
(a) Régulier



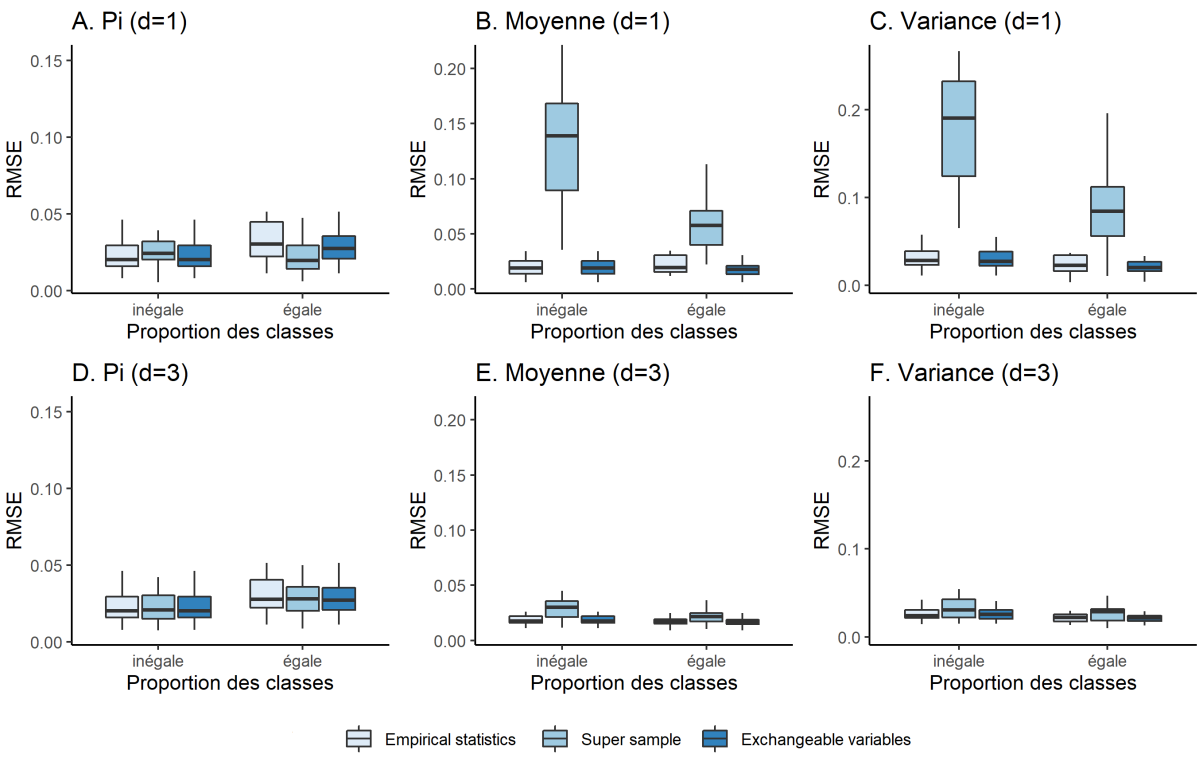
(b) Difficile

FIGURE C.11 – Sélection du nombre de classes basée sur le score BIC suivant différentes proportions de classes π, des tailles d'échantillon et d'ensemble fixées à $T = 200$ et $M = 50$. *égale* : $\pi \in (0.25, 0.25, 0.25, 0.25)$, *inégale* : $\pi \in (0.2, 0.3, 0.1, 0.4)$.

Estimation des paramètres. La figure C.12 en annexe, affiche les résultats des scores RMSE des modèles pour les deux cas de données simulées pour différentes proportions d'individus par classes et une taille d'échantillon et d'ensemble fixé à $T = 200$ et $M = 50$. Les cas "Régulier" et "Difficile" affichent peu de variations de RMSE avec toujours des scores assez faibles entre l'estimation avec proportion égale et inégale pour les modèles aux statistiques empiriques et avec les variables échangeables. Quant au modèle *Super Sample*, les RMSE montrent une augmentation entre les données simulées avec des proportions égales et celles inégales. Le modèle de mélange gaussien *Super Sample* considérant chaque réalisation comme un individu ayant sa propre classe est plus sensible au paramètre π dès lors que les proportions entre classes deviennent inégales. Quelques irrégularités sont à noter entre les dimensions, notamment dans le cas "Régulier" en dimension $d = 3$ où le modèle *Super Sample* affiche des RMSE des estimateurs moyens et de la matrice covariance légèrement plus faible avec moins d'incertitude dans le cas inégal qu'avec des proportions égales.



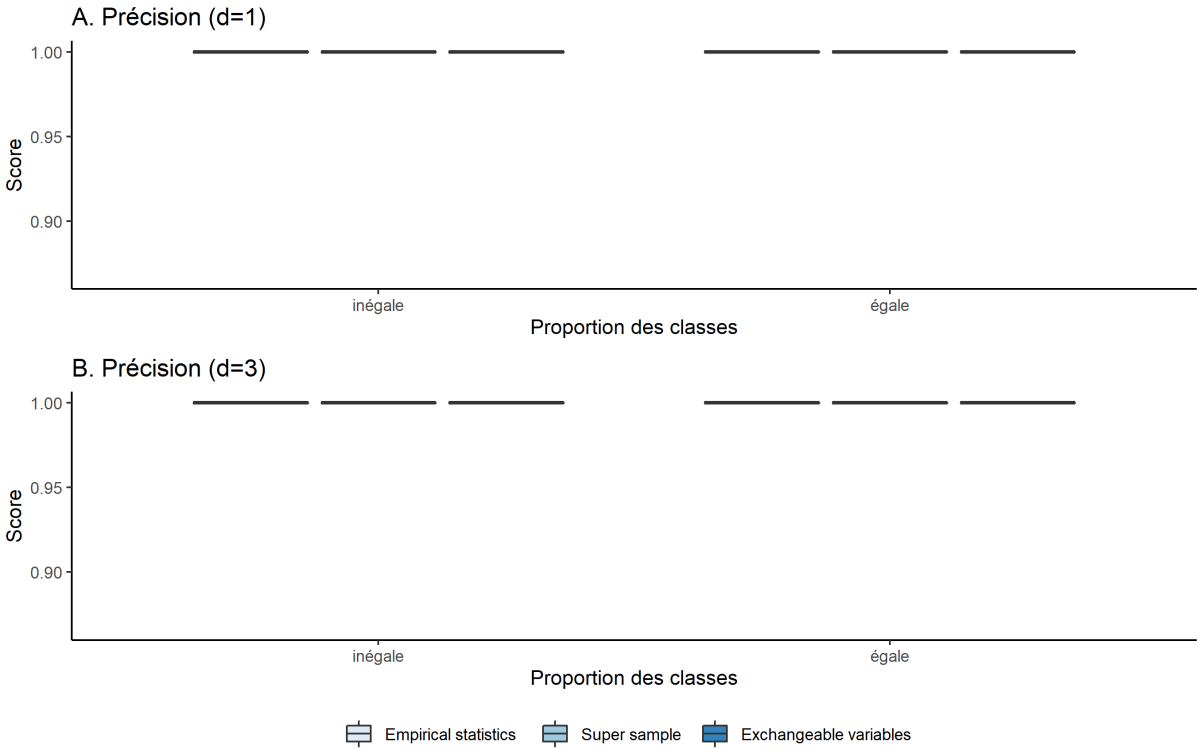
(a) Régulier



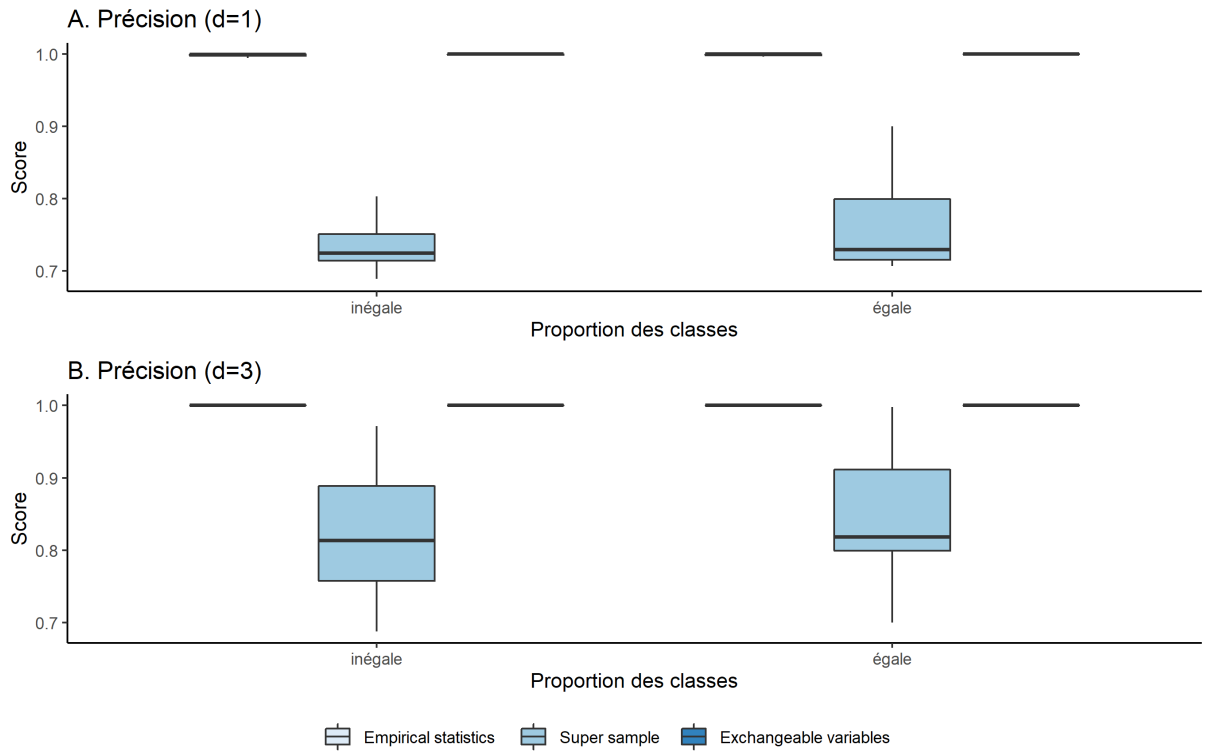
(b) Difficile

FIGURE C.12 – Racine de la moyenne des moindres carrés suivant différentes proportions de classes π, des tailles d'échantillon et d'ensemble fixées à $T = 200$ et $M = 50$. *égale :* $\pi \in (0.25, 0.25, 0.25, 0.25)$, *inégale :* $\pi \in (0.2, 0.3, 0.1, 0.4)$.

Attribution des classes. Les précisions par modèles dans un cas "Régulier" et "Difficile" de données simulées sont représentées dans les figures C.12 avec des proportions de classes différentes. Ces scores indiquent que dans un cas "Régulier", tous les modèles de mélanges arrivent à recouvrer correctement les individus des sous-groupes formés. Cependant, dans le cas "Difficile" de données simulées, le modèle *Super sample* n'offre pas une prédiction correcte des sous-groupes, comparées à celles des modèles *Empirical statistics* et *Excheangeable variables*.



(a) Régulier



(b) Difficile

FIGURE C.13 – Score de précision globale suivant différentes proportions de classes π, des tailles d'échantillon et d'ensemble fixées à $T = 200$ et $M = 50$. *égale* : $\pi \in (0.25, 0.25, 0.25, 0.25)$, *inégale* : $\pi \in (0.2, 0.3, 0.1, 0.4)$.

Pour résumer, l'évaluation des performances des modèles de mélanges étendus pour des proportions de sous-groupes variables montre des résultats très corrects pour le modèle de mélange gaussien aux statistiques empiriques et celui au vecteur de variables échangeables. Cette conclusion rejoint les observations faites dans la section 3.1.3.2.

C.4 Calibration univariée de la composante de vent méridionale (V)

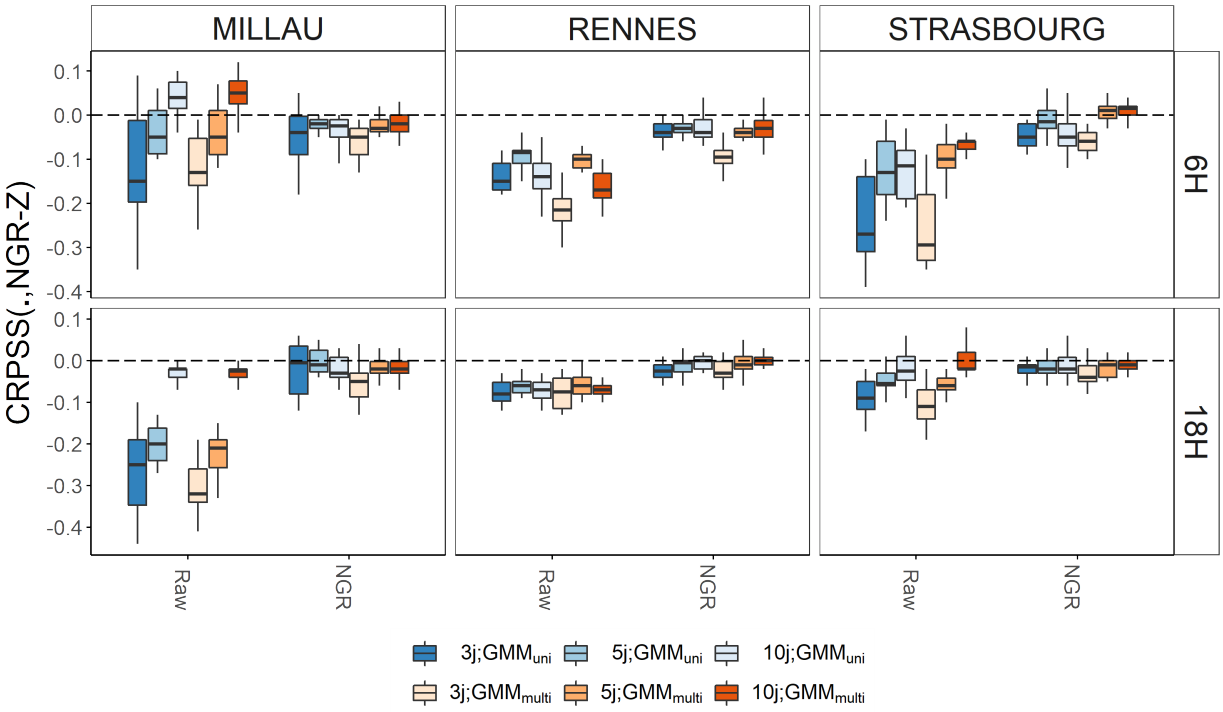


FIGURE C.14 – Variable V, composante méridionale du vent - CRPSS score des ensembles NGR_Z comparés aux ensembles *Raw* et *NGR* pour chaque station, heure et échéance de prévision, et ce pour les deux types de classification GMM_{uni} (classes ajustées sur les ensembles de température) et GMM_{multi} (classes ajustées sur les ensembles de températures et des composantes de vent (U,V)). *Les lignes pointillées délimitent un seuil de performance à atteindre pour dépasser le modèle* NGR_Z.

BIBLIOGRAPHIE

Ailliot, Pierre et al. (2015), « Non-homogeneous hidden Markov-switching models for wind time series », in : *Journal of Statistical Planning and Inference* 160, p. 75-88.

Akaike, Htrotugu (1973), « Maximum likelihood identification of Gaussian autoregressive moving average models », in : *Biometrika* 60.*2*, p. 255-265.

Allen, S, CAT Ferro et F Kwasniok (2020), « Recalibrating wind-speed forecasts using regime-dependent ensemble model output statistics », in : *Quarterly Journal of the Royal Meteorological Society* 146.*731*, p. 2576-2596.

Allen, Sam, Christopher AT Ferro et Frank Kwasniok (2019), « Regime-dependent statistical post-processing of ensemble forecasts », in : *Quarterly Journal of the Royal Meteorological Society* 145.*725*, p. 3535-3552.

Anderson, Jeffrey L (1996), « A method for producing and evaluating probabilistic forecasts from ensemble model integrations », in : *Journal of Climate* 9.*7*, p. 1518-1530.

Arthur, David et Sergei Vassilvitskii (2006), *k-means++ : The advantages of careful seeding*, rapp. tech., Stanford.

Balenzuela, Mark P et al. (2020), « A Variational Expectation-Maximisation Algorithm for Learning Jump Markov Linear Systems », in : *arXiv preprint arXiv :2004.08564*.

Baran, Agnes et al. (2021), « Machine learning for total cloud cover prediction », in : *Neural Computing and Applications* 33.*7*, p. 2605-2620.

Baran, Sándor (2014), « Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components », in : *Computational Statistics & Data Analysis* 75, p. 227-238.

Baran, Sándor et Sebastian Lerch (2015), « Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting », in : *Quarterly Journal of the Royal Meteorological Society* 141.*691*, p. 2289-2299.

— (2016), « Mixture EMOS model for calibrating ensemble forecasts of wind speed », in : *Environmetrics* 27.*2*, p. 116-130.

Baran, Sándor et Dóra Nemoda (2016), « Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting », in : *Environmetrics* 27.*5*, p. 280-292.

Baudry, Jean-Patrick (2009), « Sélection de modèle pour la classification non supervisée. Choix du nombre de classes. », Thesis.

Baudry, Jean-Patrick et Gilles Celeux (2015), « EM for mixtures », in : *Statistics and computing* 25.*4*, p. 713-726.

Baudry, Jean-Patrick, Cathy Maugis et Bertrand Michel (2012), « Slope heuristics : overview and implementation », in : *Statistics and Computing* 22.*2*, p. 455-470.

Bauer, Peter, Alan Thorpe et Gilbert Brunet (2015), « The quiet revolution of numerical weather prediction », in : *Nature* 525.*7567*, p. 47.

Ben Bouallègue, Zied et al. (2016), « Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach », in : *Monthly Weather Review* 144.*12*, p. 4737-4750.

Bessac, Julie et al. (2016), « Comparison of hidden and observed regime-switching autoregressive models for (u, v)-components of wind fields in the Northeast Atlantic », in : *Advances in Statistical Climatology, Meteorology and Oceanography* 2.*1*, p. 1-16.

Biernacki, Christophe, Gilles Celeux et Gérard Govaert (2000), « Assessing a mixture model for clustering with the integrated completed likelihood », in : *IEEE transactions on pattern analysis and machine intelligence* 22.*7*, p. 719-725.

— (2003), « Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models », in : *Computational Statistics & Data Analysis* 41.*3-4*, p. 561-575.

Bilmes, Jeff A et al. (1998), « A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models », in : *International Computer Science Institute* 4.*510*, p. 126.

Birgé, Lucien et Pascal Massart (2007), « Minimal penalties for Gaussian model selection », in : *Probability theory and related fields* 138.*1-2*, p. 33-73.

Bougeault, Philippe et al. (2010), « The THORPEX interactive grand global ensemble », in : *Bulletin of the American Meteorological Society* 91.*8*, p. 1059-1072.

Bouttier, François et Graeme Kelly (2001), « Observing-system experiments in the ECMWF 4D-Var data assimilation system », in : *Quarterly Journal of the Royal Meteorological Society* 127.*574*, p. 1469-1488.

Bozdogan, Hamparsum (1993), « Choosing the number of component clusters in the mixture-model using a new informational complexity criterion of the inverse-Fisher information matrix », in : *Information and classification*, Springer, p. 40-54.

Breiman, Leo (1996), « Bagging predictors », in : *Machine learning* 24.*2*, p. 123-140.

— (2001), « Random forests », in : *Machine learning* 45.*1*, p. 5-32.

Breiman, Leo et al. (1984), « Classification and regression trees. Wadsworth Int », in : *Group* 37.*15*, p. 237-251.

Bremnes, John Bjørnar (2019), « Constrained quantile regression splines for ensemble postprocessing », in : *Monthly Weather Review* 147.*5*, p. 1769-1780.

— (2020), « Ensemble postprocessing using quantile function regression based on neural networks and Bernstein polynomials », in : *Monthly Weather Review* 148.*1*, p. 403-414.

Bröcker, Jochen et Zied Ben Bouallègue (2020), « Stratified rank histograms for ensemble forecast verification under serial dependence », in : *Quarterly Journal of the Royal Meteorological Society* 146.*729*, p. 1976-1990.

Buizza, Roberto (2016), « Weather prediction in a world of uncertainties : should ensembles simulate the effect of model approximations ? », in : *ECMWF/WWRP Workshop : Model Uncertainty*, ECMWF, Reading.

Buizza, Roberto, Martin Leutbecher et Lars Isaksen (2008), « Potential use of an ensemble of analyses in the ECMWF Ensemble Prediction System », in : *Quarterly Journal of the Royal Meteorological Society : A journal of the atmospheric sciences, applied meteorology and physical oceanography* 134.*637*, p. 2051-2066.

Buizza, Roberto, M Milleer et Tim N Palmer (1999), « Stochastic representation of model uncertainties in the ECMWF ensemble prediction system », in : *Quarterly Journal of the Royal Meteorological Society* 125.*560*, p. 2887-2908.

Celeux, Gilles (1985), « The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem », in : *Computational statistics quarterly* 2, p. 73-82.

Celeux, Gilles et Gérard Govaert (1992), « A classification EM algorithm for clustering and two stochastic versions », in : *Computational statistics & Data analysis* 14.*3*, p. 315-332.

Celeux, Gilles et Gilda Soromenho (1996), « An entropy criterion for assessing the number of clusters in a mixture model », in : *Journal of classification* 13.*2*, p. 195-212.

Christensen, HM, IM Moroz et TN Palmer (2015), « Simulating weather regimes : Impact of stochastic and perturbed parameter schemes in a simple atmospheric model », in : *Climate Dynamics* 44.*7*, p. 2195-2214.

Clark, Martyn et al. (2004), « The Schaake shuffle : A method for reconstructing space–time variability in forecasted precipitation and temperature fields », in : *Journal of Hydrometeorology* 5.*1*, p. 243-262.

Courbariaux, Marie (2017), « Contributions statistiques aux prévisions hydrométéorologiques par méthodes d'ensemble », Thesis.

Courbariaux, Marie et al. (2019), « Post-processing Multiensemble Temperature and Precipitation Forecasts Through an Exchangeable Normal-Gamma Model and Its Tobit Extension », in : *Journal of Agricultural, Biological and Environmental Statistics* 24.*2*, p. 309-345.

Couso, Inés et Didier Dubois (2018), « A general framework for maximizing likelihood under incomplete data », in : *International Journal of Approximate Reasoning* 93, p. 238-260.

De Finetti, B (1931), *Funzione Caratteristica Di un Fenomeno Aleatorio, vol. 4 of 6.*

Dempster, Arthur P, Nan M Laird et Donald B Rubin (1977), « Maximum likelihood from incomplete data via the EM algorithm », in : *Journal of the Royal Statistical Society : Series B (Methodological)* 39.*1*, p. 1-22.

Diaconis, Persi et David Freedman (1980), « Finite exchangeable sequences », in : *The Annals of Probability*, p. 745-764.

Directorate, ECMWF (2012), « Describing ECMWF's forecasts and forecasting system », in : *EcMWF Newsletter* 133, p. 11-13.

Eisenman, Bonnie (2015), *Learning react native : Building native mobile apps with JavaScript*, " O'Reilly Media, Inc."

Epstein, Edward S (1969), « Stochastic dynamic prediction », in : *Tellus* 21.*6*, p. 739-759.

Feldmann, Kira, Michael Scheuerer et Thordis L Thorarinsdottir (2015), « Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression », in : *Monthly Weather Review* 143.*3*, p. 955-971.

Fisher, Ronald Aylmer (1992), « Statistical methods for research workers », in : *Breakthroughs in statistics*, Springer, p. 66-70.

Fraley, Chris (1998), « Algorithms for model-based Gaussian hierarchical clustering », in : *SIAM Journal on Scientific Computing* 20.*1*, p. 270-281.

Fraley, Chris et Adrian E Raftery (1998), « How many clusters ? Which clustering method ? Answers via model-based cluster analysis », in : *The computer journal* 41.*8*, p. 578-588.

Fraley, Chris, Adrian E Raftery et Tilmann Gneiting (2010), « Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging », in : *Monthly Weather Review* 138.*1*, p. 190-202.

Friedman, Jerome, Trevor Hastie et Rob Tibshirani (2010), « Regularization paths for generalized linear models via coordinate descent », in : *Journal of statistical software* 33.*1*, p. 1.

Friedman, Jerome, Trevor Hastie, Rob Tibshirani et al. (2021), « Package 'glmnet' », in : *CRAN R Repositary*.

Gagne, David John, Amy McGovern et Ming Xue (2014), « Machine learning enhancement of storm-scale ensemble probabilistic quantitative precipitation forecasts », in : *Weather and Forecasting* 29.*4*, p. 1024-1043.

Gill, Simon, Bruce Stephen et Stuart Galloway (2011), « Wind turbine condition assessment through power curve copula modeling », in : *IEEE Transactions on Sustainable Energy* 3.*1*, p. 94-101.

Glahn, Harry R et Dale A Lowry (1972), « The use of model output statistics (MOS) in objective weather forecasting », in : *Journal of Applied Meteorology and Climatology* 11.*8*, p. 1203-1211.

Gneiting, Tilmann et Matthias Katzfuss (2014), « Probabilistic forecasting », in : *Annual Review of Statistics and Its Application* 1, p. 125-151.

Gneiting, Tilmann et Adrian E Raftery (2007), « Strictly proper scoring rules, prediction, and estimation », in : *Journal of the American Statistical Association* 102.*477*, p. 359-378.

Gneiting, Tilmann, Adrian E Raftery et al. (2005), « Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation », in : *Monthly Weather Review* 133.*5*, p. 1098-1118.

Gneiting, Tilmann, Larissa I Stanberry et al. (2008), « Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds », in : *Test* 17.*2*, p. 211.

Gregorutti, Baptiste, Bertrand Michel et Philippe Saint-Pierre (2017), « Correlation and variable importance in random forests », in : *Statistics and Computing* 27.*3*, p. 659-678.

Grimit, Eric P et al. (2006), « The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification », in : *Quarterly Journal of the Royal Meteorological Society : A journal of the atmospheric sciences, applied meteorology and physical oceanography* 132.*621C*, p. 2925-2942.

Grinberg, Miguel (2018), *Flask web development : developing web applications with python*, " O'Reilly Media, Inc."

Hamill, Thomas M (2001), « Interpretation of rank histograms for verifying ensemble forecasts », in : *Monthly Weather Review* 129.*3*, p. 550-560.

Hamill, Thomas M et Stephen J Colucci (1997), « Verification of Eta–RSM short-range ensemble forecasts », in : *Monthly Weather Review* 125.*6*, p. 1312-1327.

Hamill, Thomas M et Jeffrey S Whitaker (2006), « Probabilistic quantitative precipitation forecasts based on reforecast analogs : Theory and application », in : *Monthly Weather Review* 134.*11*, p. 3209-3229.

Hamill, Thomas M, Jeffrey S Whitaker et Xue Wei (2004), « Ensemble reforecasting : Improving medium-range forecast skill using retrospective forecasts », in : *Monthly Weather Review* 132.*6*, p. 1434-1447.

Hemri, Stephan, Thomas Haiden et Florian Pappenberger (2016), « Discrete postprocessing of total cloud cover ensemble forecasts », in : *Monthly Weather Review* 144.*7*, p. 2565-2577.

Hemri, Stephan, Michael Scheuerer et al. (2014), « Trends in the predictive performance of raw ensemble weather forecasts », in : *Geophysical Research Letters* 41.*24*, p. 9197-9205.

Herman, Gregory R et Russ S Schumacher (2018a), « "Dendrology" in numerical weather prediction : What random forests and logistic regression tell us about forecasting extreme precipitation », in : *Monthly Weather Review* 146.*6*, p. 1785-1812.

— (2018b), « Money doesn't grow on trees, but forecasts do : Forecasting extreme precipitation with random forests », in : *Monthly Weather Review* 146.*5*, p. 1571-1600.

Hersbach, Hans (2000), « Decomposition of the continuous ranked probability score for ensemble prediction systems », in : *Weather and Forecasting* 15.*5*, p. 559-570.

Hewitt, Edwin et Leonard J Savage (1955), « Symmetric measures on Cartesian products », in : *Transactions of the American Mathematical Society* 80.*2*, p. 470-501.

Hoffman, Ross N et Eugenia Kalnay (1983), « Lagged average forecasting, an alternative to Monte Carlo forecasting », in : *Tellus A : Dynamic Meteorology and Oceanography* 35.*2*, p. 100-118.

Jan, Mr (2019), « Workshop on Predictability, dynamics and applications research using the TIGGE and S2S ensembles », in : *ECMWF*.

Jordan, Alexander (2016), « Facets of forecast evaluation », Thesis, Karlsruher Institut für Technologie (KIT).

Jouan, Gabriel et al. (oct. 2019), « Weather types prediction at medium-range from ensemble forecasts », in : *9th International workshop on Climate Informatics*, Paris, France, URL : https://hal.archives-ouvertes.fr/hal-02425230.

Keribin, Christine (2000), « Consistent estimation of the order of mixture models », in : *Sankhyā : The Indian Journal of Statistics, Series A*, p. 49-66.

Klein, William H, Billy M Lewis et Isadore Enger (1959), « Objective prediction of five-day mean temperatures during winter », in : *Journal of Atmospheric Sciences* 16.*6*, p. 672-682.

Koenker, Roger et al. (2018), « Package 'quantreg' », in : *Cran R-project. org*.

Krüger, Frank (2018), « Activity, context, and plan recognition with computational causal behavior models », Thesis, Universiät Rostock. Fakultät für Informatik und Elektrotechnik.

Laio, Francesco et Stefania Tamea (2007), « Verification tools for probabilistic forecasts of continuous hydrological variables », in : *Hydrology and Earth System Sciences* 11.*4*, p. 1267-1277.

Leith, CE (1974), « Theoretical skill of Monte Carlo forecasts », in : *Monthly Weather Review* 102.*6*, p. 409-418.

Leroux, Brian G (1992), « Consistent estimation of a mixing distribution », in : *The Annals of Statistics*, p. 1350-1360.

Leutbecher, Martin et al. (2017), « Stochastic representations of model uncertainties at ECMWF : State of the art and future vision », in : *Quarterly Journal of the Royal Meteorological Society* 143.*707*, p. 2315-2339.

Lloyd, Stuart (1982), « Least squares quantization in PCM », in : *IEEE transactions on information theory* 28.*2*, p. 129-137.

Lorenz, Edward N (1963), « Deterministic nonperiodic flow », in : *Journal of atmospheric sciences* 20.*2*, p. 130-141.

— (1965), « A study of the predictability of a 28-variable atmospheric model », in : *Tellus* 17.*3*, p. 321-333.

— (1996), « Predictability : A problem partly solved », in : *Proc. Seminar on predictability*, t. 1, 1.

Lourens, Spencer et al. (2013), « Bias in estimation of a mixture of normal distributions », in : *Journal of biometrics & biostatistics* 4.

Maraun, Douglas (2016), « Bias correcting climate change simulations-a critical review », in : *Current Climate Change Reports* 2.*4*, p. 211-220.

Matheson, James E et Robert L Winkler (1976), « Scoring rules for continuous probability distributions », in : *Management science* 22.*10*, p. 1087-1096.

McGovern, Amy et al. (2017), « Using artificial intelligence to improve real-time decision-making for high-impact weather », in : *Bulletin of the American Meteorological Society* 98.*10*, p. 2073-2090.

McLachlan, Geoffrey J et Thriyambakam Krishnan (2007), *The EM algorithm and extensions*, t. 382, John Wiley & Sons.

McLachlan, Geoffrey J et David Peel (2004), *Finite mixture models*, John Wiley & Sons.

Mcnicholas, Paul David et Thomas Brendan Murphy (2008), « Parsimonious Gaussian mixture models », in : *Statistics and Computing* 18.*3*, p. 285-296.

Meinshausen, Nicolai (2006), « Quantile regression forests », in : *Journal of Machine Learning Research* 7.*Jun*, p. 983-999.

Messner, Jakob W et al. (2014), « Extending extended logistic regression : Extended versus separate versus ordered versus censored », in : *Monthly Weather Review* 142.*8*, p. 3003-3014.

Möller, Annette et Jürgen Gross (2016), « Probabilistic temperature forecasting based on an ensemble autoregressive modification », in : *Quarterly Journal of the Royal Meteorological Society* 142.*696*, p. 1385-1394.

Möller, Annette, Alex Lenkoski et Thordis L Thorarinsdottir (2013), « Multivariate probabilistic forecasting using ensemble Bayesian model averaging and copulas », in : *Quarterly Journal of the Royal Meteorological Society* 139.*673*, p. 982-991.

Naveau, Philippe et Julie Bessac (2018), « Forecast evaluation with imperfect observations and imperfect models », in : *arXiv preprint arXiv :1806.03745*.

Nelder, John A et Roger Mead (1965), « A simplex method for function minimization », in : *The computer journal* 7.*4*, p. 308-313.

Nityasuddhi, Dechavudh et Dankmar Böhning (2003), « Asymptotic properties of the EM algorithm estimate for normal mixture models with component specific variances », in : *Computational statistics & data analysis* 41.*3-4*, p. 591-601.

O'Neill, Ben (2009), « Exchangeability, correlation, and Bayes' effect », in : *International statistical review* 77.*2*, p. 241-250.

Palmer, Tim (2019), « The ECMWF ensemble prediction system : Looking back (more than) 25 years and projecting forward 25 years », in : *Quarterly Journal of the Royal Meteorological Society* 145, p. 12-24.

Palmer, TN et al. (2009), « Stochastic parametrization and model uncertainty », in :

Park, Young-Youn, Roberto Buizza et Martin Leutbecher (2008), « TIGGE : Preliminary results on comparing and combining ensembles », in : *Quarterly Journal of the Royal Meteorological Society* 134.*637*, p. 2029-2050.

Pearson, Karl (1894), « Contributions to the mathematical theory of evolution », in : *Philosophical Transactions of the Royal Society of London. A* 185, p. 71-110.

Pinson, Pierre et Robin Girard (2012), « Evaluating the quality of scenarios of short-term wind power generation », in : *Applied Energy* 96, p. 12-20.

Pinson, Pierre et Renate Hagedorn (2012), « Verification of the ECMWF ensemble forecasts of wind speed against analyses and observations », in : *Meteorological Applications* 19.*4*, p. 484-500.

Pinson, Pierre, Henrik Madsen et al. (2009), « From probabilistic forecasts to statistical scenarios of short-term wind power production », in : *Wind Energy : An International Journal for Progress and Applications in Wind Power Conversion Technology* 12.*1*, p. 51-62.

Popelka, Stanislav, Alena Vondrakova et Petra Hujnakova (2019), « Eye-tracking evaluation of weather web maps », in : *ISPRS International Journal of Geo-Information* 8.*6*, p. 256.

Raftery, Adrian E et al. (2005), « Using Bayesian model averaging to calibrate forecast ensembles », in : *Monthly weather review* 133.*5*, p. 1155-1174.

Rasp, Stephan et Sebastian Lerch (2018), « Neural networks for postprocessing ensemble weather forecasts », in : *Monthly Weather Review* 146.*11*, p. 3885-3900.

Rautenhaus, Marc et al. (2015), « Three-dimensional visualization of ensemble weather forecasts–Part 1 : The visualization tool Met. 3D (version 1.0) », in : *Geoscientific Model Development* 8.*7*, p. 2329-2353.

RColorBrewer, Suggests et Maintainer Andy Liaw (2018), « Package 'randomForest' », in : *University of California, Berkeley : Berkeley, CA, USA*.

Redner, Richard A et Homer F Walker (1984), « Mixture densities, maximum likelihood and the EM algorithm », in : *SIAM review* 26.*2*, p. 195-239.

Richardson, Lewis Fry (2007), *Weather prediction by numerical process*, Cambridge university press.

Rijsbergen, CJ (1979), « v.(1979) », in : *Information retrieval* 2.

Roy, Anandarup et Swapan K Parui (2014), « Pair-copula based mixture models and their application in clustering », in : *Pattern recognition* 47.*4*, p. 1689-1697.

Rüschendorf, Ludger (2009), « On the distributional transform, Sklar's theorem, and the empirical copula process », in : *Journal of statistical planning and inference* 139.*11*, p. 3921-3927.

Schefzik, Roman (2016), « A similarity-based implementation of the Schaake shuffle », in : *Monthly Weather Review* 144.*5*, p. 1909-1921.

— (2017), « Ensemble calibration with preserved correlations : unifying and comparing ensemble copula coupling and member-by-member postprocessing », in : *Quarterly Journal of the Royal Meteorological Society* 143.*703*, p. 999-1008.

Schefzik, Roman et Annette Möller (2018), « Ensemble postprocessing methods incorporating dependence structures », in : *Statistical Postprocessing of Ensemble Forecasts*, Elsevier, p. 91-125.

Schefzik, Roman, Thordis L Thorarinsdottir, Tilmann Gneiting et al. (2013), « Uncertainty quantification in complex simulation models using ensemble copula coupling », in : *Statistical science* 28.*4*, p. 616-640.

Schepen, Andrew, Yvette Everingham et Quan J Wang (2020), « On the joint calibration of multivariate seasonal climate forecasts from GCMs », in : *Monthly Weather Review* 148.*1*, p. 437-456.

Scher, Sebastian et Gabriele Messori (2018), « Predicting weather forecast uncertainty with machine learning », in : *Quarterly Journal of the Royal Meteorological Society* 144.*717*, p. 2830-2841.

Scheuerer, Michael et Thomas M Hamill (2015a), « Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions », in : *Monthly Weather Review* 143.*11*, p. 4578-4596.

— (2015b), « Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities », in : *Monthly Weather Review* 143.*4*, p. 1321-1334.

— (2019), « Probabilistic forecasting of snowfall amounts using a hybrid between a parametric and an analog approach », in : *Monthly Weather Review* 147.*3*, p. 1047-1064.

Scheuerer, Michael, Thomas M Hamill et al. (2017), « A method for preferential selection of dates in the S chaake shuffle approach to constructing spatiotemporal

forecast fields of temperature and precipitation », in : *Water Resources Research* 53.*4*, p. 3029-3046.

Schwarz, Gideon et al. (1978), « Estimating the dimension of a model », in : *The annals of statistics* 6.*2*, p. 461-464.

Scrucca, Luca et al. (2016), « mclust 5 : clustering, classification and density estimation using Gaussian finite mixture models », in : *The R journal* 8.*1*, p. 289.

Siegert, Stefan, Jochen Bröcker et Holger Kantz (2012), « Rank histograms of stratified Monte Carlo ensembles », in : *Monthly weather review* 140.*5*, p. 1558-1571.

Singer, Saša et John Nelder (2009), « Nelder-mead algorithm », in : *Scholarpedia* 4.*7*, p. 2928.

Sklar, Abe (1959), « Fonction de répartition dont les marges sont données », in : *Inst. stat. univ. Paris* 8, p. 229-231.

Stensrud, David J, Jian-Wen Bao et Thomas T Warner (2000), « Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems », in : *Monthly Weather Review* 128.*7*, p. 2077-2107.

Straaten, Chiem van, Kirien Whan et Maurice Schmeits (2018), « Statistical postprocessing and multivariate structuring of high-resolution ensemble precipitation forecasts », in : *Journal of Hydrometeorology* 19.*11*, p. 1815-1833.

Swinbank, Richard et al. (2016), « The TIGGE project and its achievements », in : *Bulletin of the American Meteorological Society* 97.*1*, p. 49-67.

Taillardat, Maxime (2017), « Non-parametric Methods of post-processing for Ensemble Forecasting », Thesis, Université Paris Saclay (COmUE).

Taillardat, Maxime, Anne-Laure Fougères et al. (2019), « Forest-based and semiparametric methods for the postprocessing of rainfall ensemble forecasting », in : *Weather and Forecasting 2019*.

Taillardat, Maxime, Olivier Mestre et al. (2016), « Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics », in : *Monthly Weather Review* 144.*6*, p. 2375-2393.

Talagrand, O., R. Vautard et B Strauss (1997), « Evaluation of probabilistic prediction systems », in : *Workshop on Predictability, 20-22 October 1997*, ECMWF, Shinfield Park, Reading : ECMWF, p. 1-26, url : `https://www.ecmwf.int/node/12555`.

Thorarinsdottir, Thordis L et Tilmann Gneiting (2010), « Probabilistic forecasts of wind speed : Ensemble model output statistics by using heteroscedastic censored

regression », in : *Journal of the Royal Statistical Society : Series A (Statistics in Society)* 173.*2*, p. 371-388.

Thorarinsdottir, Thordis L, Michael Scheuerer et Christopher Heinz (2016), « Assessing the calibration of high-dimensional ensemble forecasts using rank histograms », in : *Journal of computational and graphical statistics* 25.*1*, p. 105-122.

Thorarinsdottir, Thordis L et Nina Schuhen (2018), « Verification : assessment of calibration and accuracy », in : *Statistical postprocessing of ensemble forecasts*, Elsevier, p. 155-186.

Vannitsem, Stéphane et al. (2020), « Statistical postprocessing for weather forecasts–review, challenges and avenues in a big data world », in : *Bulletin of the American Meteorological Society*, p. 1-44.

— (2021), « Statistical Postprocessing for Weather Forecasts : Review, Challenges, and Avenues in a Big Data World », in : *Bulletin of the American Meteorological Society* 102.*3*, E681-E699.

Vautard, Robert (1990), « Multiple weather regimes over the North Atlantic : Analysis of precursors and successors », in : *Monthly weather review* 118.*10*, p. 2056-2081.

Wilks, Daniel S (2005), « Effects of stochastic parametrizations in the Lorenz'96 system », in : *Quarterly Journal of the Royal Meteorological Society : A journal of the atmospheric sciences, applied meteorology and physical oceanography* 131.*606*, p. 389-407.

— (2018), « Univariate ensemble postprocessing », in : *Statistical postprocessing of ensemble forecasts*, Elsevier, p. 49-89.

Wilks, DS (2011), « On the reliability of the rank histogram », in : *Monthly Weather Review* 139.*1*, p. 311-316.

Williams, RM, CAT Ferro et Frank Kwasniok (2014), « A comparison of ensemble post-processing methods for extreme events », in : *Quarterly Journal of the Royal Meteorological Society* 140.*680*, p. 1112-1120.

Wong, M Anthony (1982), « A hybrid clustering method for identifying high-density clusters », in : *Journal of the American Statistical Association* 77.*380*, p. 841-847.

Wu, Limin et al. (2018), « Comparative evaluation of three Schaake shuffle schemes in postprocessing GEFS precipitation ensemble forecasts », in : *Journal of Hydrometeorology* 19.*3*, p. 575-598.

Yang, Miin-Shen et Yessica Nataliani (2017), « Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters », in : *Pattern Recognition* 71, p. 45-59.

Yang, Yuhong (2005), « Can the strengths of AIC and BIC be shared ? A conflict between model indentification and regression estimation », in : *Biometrika* 92.*4*, p. 937-950.

Ypma, Jelmer, Hans W Borchers et Dirk Eddelbuettel (2014), « nloptr : R Interface to NLopt », in : *R package version* 1.*4*.

Yuen, RA et al. (2018), « Package 'ensembleMOS' », in :

Zhu, Ji et Trevor Hastie (2004), « Classification of gene microarrays by penalized logistic regression », in : *Biostatistics* 5.*3*, p. 427-443.

Zhu, Yuejian et al. (2002), « The economic value of ensemble-based weather forecasts », in : *Bulletin of the American Meteorological Society* 83.*1*, p. 73-84.

TABLE DES FIGURES

LISTE DES TABLEAUX

DOCTORAT BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

Titre : Post-traitement des prévisions d'ensemble en météorologie par des méthodes d'apprentissage statistique.

Mot clés : Météorologie, problème de calibration, méthode de classification, mélange gaussien, planification d'événements

Résumé : Aujourd'hui, la plupart des centres de prévision météorologique produisent des prévisions d'ensemble. Ces données fournissent une description plus complète de l'atmosphère qu'une exécution unique du modèle météorologique. Cependant, elles peuvent souffrir d'erreurs de biais et de sous/sur-dispersion. Pour les corriger, des méthodes statistiques sont employées. Cette approche est appelée la calibration d'ensembles de prévisions.

Dans cette thèse, nous proposons une méthode originale de calibration multivariée dans laquelle l'espace des données est discrétisé de façon à transformer un problème de régression multi-sorties en une classification à sortie unique. Les résultats de cette approche sont encourageants. Néanmoins, elle pose des difficultés quant à l'interprétation physique des modèles. Par la suite, une méthode de calibration en deux étapes est alors présentée : une première étape de classification vise à identifier les régimes météorologiques à l'aide d'une extension d'un modèle de mélange gaussien, la seconde à corriger la distribution d'ensemble dans chaque régime. Pour terminer, une interface web a été développée autour des prévisions et des ensembles de prévisions à moyen terme avec comme cas d'usage la planification d'évènements.

Title: Post-processing of ensemble forecasts in meteorology using statistical learning methods.

Keywords: Meteorology, calibration problem, classification method, gaussian mixture, event planning

Abstract: Nowadays, most weather forecasting centers produce ensemble forecasts. They give a more complete description of the atmosphere than a unique run of the meteorological model. However, they may suffer from bias and under/over dispersion errors that need to be corrected. In order to correct these errors, statistical methods are used. This approach is called ensemble forecast calibration.

In this thesis, we propose an original multivariate calibration method in which the data space is discretised in order to transform a multi-output regression problem into a single output classification. The results of this approach are encouraging. The results of this approach are encouraging. However, the physical interpretation of the models remains difficult. Subsequently, a two-step calibration method is then presented: a first classification step aims at identifying the weather regimes using an extension of a Gaussian mixture model, the second at correcting the ensemble distribution in each regime. Finally, a web interface has been developed around forecasts and medium-term forecast sets with the use case of event planning.