Statistical Post-processing of Deterministic and Ensemble Wind Speed Forecasts on a Grid

Michaël Zamo

▶ **To cite this version:**

Michaël Zamo. Statistical Post-processing of Deterministic and Ensemble Wind Speed Forecasts on a Grid. Earth Sciences. Université Paris Saclay (COmUE), 2016. English. NNT : 2016SACLA029 . tel-01598119

HAL Id: tel-01598119
https://theses.hal.science/tel-01598119

Submitted on 29 Sep 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NNT : 2016SACLA029

Thèse de doctorat de l'Université Paris-Saclay préparée à AgroParisTech

Ecole doctorale n°581
Agriculture Alimentation BIologie Environnement Santé
Spécialité de doctorat : Statistiques appliquées

par

M. Michaël Zamo

Statistical Post-processing of Deterministic and Ensemble Wind Speed Forecasts on a Grid

Thèse présentée et soutenue à Météo-France, Toulouse, le 15 décembre 2016.

Composition du Jury :

M.	Philippe BESSE	Professeur INSA UMR 5219, Toulouse	Président du jury
M.	Denis ALLARD	Directeur de recherche INRA UR 546, Avignon	Rapporteur
M.	Tilmann GNEITING	Professeur HITS, Heidelberg	Rapporteur
M.	Philippe ARBOGAST	ICPEF CNRM UMR 3589, Toulouse	Examinateur
M.	Vivien MALLET	Chargé de recherche INRIA, Paris	Examinateur
M.	Mathieu VRAC	Directeur de recherche LSCE UMR 8212, Gif-sur-Yvette	Examinateur
Mme.	Liliane BEL	Professeur AgroParisTech, Paris	Directrice de thèse
M.	Olivier Mestre	IDTM Météo-France	Co-encadrant de thèse

The pessimist complains about the wind;
the optimist expects it to change;
the realist adjusts the sails.

William Arthur Ward

Acknowledgements

First, I would like to thank my two thesis supervisors, Liliane Bel and Olivier Mestre. Liliane managed to supervise this thesis from the far Paris, and nevertheless used her meticulousness to much improve the readability of this manuscript. Thank you also for providing me six months of work by just uttering one seemingly innocent question starting with "why don't you...?". Olivier, with his famous negociation skills, offered me the opportunity to work on this thesis, freed from the dreadful burden of operations (and their most dreadful COMOD forms...), and also for his experience in applied Statistics and in supervising researches. Thanks again to both of you for sharing your kindness, good humor and experience.

I am also grateful to Tillmann Gneiting and Denis Allard for reporting on this manuscript written in English by a French guy. This is an achievement in itself.

A great thank you also to those who accepted to take part in the several committees during these three years, Jean-Yves Dauxois, Philippe Arbogast, Vivien Mallet, Philippe Besse and Mathieu Vrac.

I acknowledge also the help, at various levels, from my colleagues from COMPAS at Météo-France, notably Joël Stein, Véronique Mathiot, Isabelle Sanchez, Harold Petithomme and Marina Covre, whose knowledge in their own specific fields make them a very skilled team to work with. Without mentioning the pleasure to work in such a pleasant environment.

I was also very pleased to exchange ideas during these three years with very wise people, during the many work travels accross France and Europe. A special mention to Tilmann Gneiting, Christopher Ferro, Aurélien Ribes, Sébastien Gerchinovitz, Nicolas Morié, Maxime Taillardat, Ivan Kojadinovic and, last but not least, Philippe Naveau.

Contents

Résumé en français **5**
0.1 Introduction 5
0.2 Chapitre 2: Improved Gridded Wind Speed Forecasts with Block MOS . . . 8
0.2.1 Motivation 8
0.2.2 Résultats 8
0.3 Chapitre 3: Estimation of the CRPS with Limited Information 10
0.3.1 Motivation 10
0.3.2 Résultats 11
0.4 Chapitre 4: Aggregation of Probabilistic Wind Speed Forecasts 12
0.4.1 Motivation 12
0.4.2 Résultats 13
0.5 Conclusion 15

1 Introduction and Summary **17**
1.1 Introduction 18
1.2 Chapter 2: Improved Gridded Wind Speed Forecasts with Block MOS . . . 20
1.2.1 Motivation 20
1.2.2 State of the art of wind speed MOS and gridded MOS 20
1.2.3 Results 22
1.3 Chapter 3: Estimation of the CRPS with Limited Information 24
1.3.1 Motivation 24
1.3.2 State of the art of the estimation of the CRPS 25
1.3.3 Results 26
1.4 Chapter 4: Aggregation of Probabilistic Wind Speed Forecasts 27
1.4.1 Motivation 27
1.4.2 State of the art of EMOS and combination thereof 28
1.4.3 Results 29
1.5 Conclusion 30

2 Improved Gridded Wind Speed Forecasts with Block MOS **33**

2.1 Introduction . . . 34
2.2 Gridding 10-m windspeed measurements . . . 36
2.2.1 Methodology . . . 37
2.2.2 Data description . . . 37
2.2.3 Verification strategy . . . 38
2.2.4 Results about the best interpolation strategy . . . 39
2.3 Improving wind speed forecasts on a grid by block regression . . . 46
2.3.1 Data . . . 47
2.3.2 Block MOS . . . 49
2.3.3 Results . . . 50
2.4 Conclusion . . . 58

3 Estimation of the CRPS with Limited Information **61**
3.1 Introduction . . . 62
3.2 Review of available estimators of the CRPS . . . 65
3.3 Study with simulated data . . . 66
3.3.1 CRPS estimation with a random ensemble . . . 67
3.3.2 CRPS estimation with an ensemble of quantiles . . . 70
3.4 Real data examples . . . 77
3.4.1 Raw and calibrated ensemble forecast data sets . . . 80
3.4.2 Issues estimating the CRPS of real data . . . 80
3.4.3 Issues on the choice between QRF and NR . . . 83
3.5 Conclusion and discussion . . . 84
Appendix . . . 86
3.A What is elicited when the 1-norm CRPS of an ensemble is minimized? . . . 86
3.B Relationships between the estimators of the CRPS . . . 87
3.B.1 Equality of $\widehat{crps}_{Fair}$ and $\widehat{crps}_{PWM}$ . . . 87
3.B.2 Equality of $\widehat{crps}_{NRG}$ and $\widehat{crps}_{INT}$ . . . 88
3.B.3 Relationship between $\widehat{crps}_{PWM}$ and $\widehat{crps}_{NRG}$ . . . 88

4 Aggregation of Probabilistic Wind Speed Forecasts **89**
4.1 Introduction . . . 90
4.2 Theoretical framework and verification strategy . . . 91
4.2.1 The individual sequence prediction framework . . . 91
4.2.2 Sequential aggregation of step-wise CDFs . . . 92
4.2.3 Verification strategy . . . 94
4.3 Aggregation methods . . . 97
4.3.1 Inverse CRPS weighting . . . 98
4.3.2 Sharpness-calibration paradigm . . . 98
4.3.3 Minimum CRPS . . . 99
4.3.4 Exponential weighting . . . 99

4.3.5 Exponentiated gradient 99
4.4 The experts and the observation 100
4.4.1 The TIGGE data set and experts 100
4.4.2 The calibrated experts 101
4.4.3 The observation 102
4.5 Results 103
4.5.1 CRPS, reliability and sharpness 103
4.5.2 Spatio-temporal characteristics of the most reliable aggregated forecast 107
4.6 Discussion about calibration and aggregation procedures 108
4.7 Conclusion and perspectives 108
Appendix 119
4.A Formula for the gradient of the CRPS 119
4.B Bounds for the EWA aggregation 121
4.C Time series of the regret at each lead time 123
4.D Maps of rank histograms of raw ensembles 127
4.E Time series of the aggregation weights 135

5 Conclusion and Perspectives **141**

Bibliography **145**

Résumé en français

Abstract La présente thèse concerne l'amélioration des prévisions météorologiques grâce à des méthodes de post-traitement statistique, dans un contexte opérationnel. Des prévisions de vitesse de vent sont construites sur la grille d'un modèle, ce qui pose des problèmes de temps de traitement dus au nombre potentiellement élevé de points de grille dans les modèles de prévision météorologique actuels. Ce résumé esquisse d'abord la procédure de post-traitement des prévisions météorologiques et les motivations du sujet choisi. Ensuite, chaque partie du présent travail est présentée en termes de motivations des problèmes spécifiques traités et de principaux résultats.

0.1 Introduction

La plupart des services météorologiques nationaux et certaines entreprises privées emploient des modèles de prévision numérique du temps (PNT) pour produire des prévisions météorologiques et des prévisions dérivées pour des domaines météo-sensibles, tels que certains risques économiques (Alexandridis and Zapranis 2012; Bertrand et al. 2015), la production d'énergie renouvelable (Costa et al. 2008; Dubus et al. 2014; Zamo et al. 2014a,b), la consommation électrique (Taylor and Buizza 2002), la qualité de l'air (Mallet and Sportisse 2006; Besse et al. 2007) ou l'agriculture (Cantelaube and Terres 2005; Baker and Kirk 2007; Trnka et al. 2011), entre autres. Chaque modèle de PNT commet des erreurs qui ne sont pas totalement aléatoires et, à ce titre, peuvent être partiellement réduites grâce à des méthodes de post-traitement statistique qui ont été utilisées dès le début de l'ère de la prévision météorologique moderne (Glahn and Lowry 1972).

L'approche la plus usuelle pour post-traiter les prévisions météorologiques consiste à construire automatiquement une relation statistique entre les observatiosn passées et les paramètres associés prévus par PNT. Pour cela, les nombreux outils statistiques développés

dans les domaines de la fouille de données ou de l'apprentissage machine sont mis à contribution (Hastie et al. 2009; Kuhn and Johnson 2013; Alpaydin 2014). La relation trouvée est ensuite appliquée aux prévisions futures afin d'améliorer leur performance, ou de prévoir des paramètres observés non directement prévus par un modèle de PNT. Cette approche est appelée Adaptation Statisque (AS) pour une prévision de PNT déterministe, ou Adaptation Statistique d'ensemble (Gneiting et al. 2005) lorsque la prévision post-traitée s'appuie sur une gamme de valeurs générées par un modèle de PNT d'ensemble (Leutbecher and Palmer 2008). Les méthodes d'AS produisent une valeur scalaire qui peut être la moyenne attendue de l'observation (Jacks et al. 1990; Zamo et al. 2014a) ou un quantile quelconque (Friederichs and Hense 2008), voir également Gneiting (2011a). Les méthodes d'AS d'ensemble prévoient tout ou partie de la distribution de l'observation (Oesting et al. 2013; Zamo et al. 2014b). Bien que des modèles de PNT de plus en plus sophistiqués et performants ont été développés durant les dernières décennies (Lynch 2008; Inness and Dorling 2013; Bauer et al. 2015), les méthodes d'AS et d'AS d'ensemble parviennent encore à améliorer la performance des prévisions (Ruth et al. 2009; Hemri et al. 2014; Taillardat et al. 2016).

Puisque aucun résultat théorique ne permet de choisir *a priori* les meilleures méthodes d'AS ou d'AS d'ensemble pour un problème donné, plusieurs méthodes d'AS ou d'AS d'ensemble sont comparées, afin de choisir la meilleure, selon un critère de performance. Cependant, plusieurs études (Bates and Granger 1969; Clemen 1989) ont montré que combiner plusieurs prévisions (post-traitées ou pas) d'un ou plusieurs modèles de PNT (Fritsch et al. 2000) peut améliorer les performances de prévision pour des prévisions déterministes (Vislocky and Fritsch 1995) ou d'ensemble (Baudin 2015). Là encore, plusieurs manières de combiner les prévisions existent (Cesa-Bianchi et al. 2006; Allard et al. 2012; Gneiting et al. 2013), parmi lesquelles il faut choisir grâce à une comparaison empirique (Palm and Zellner 1992; Mallet et al. 2007; Gerchinovitz et al. 2008).

Étant généralement ajustées pour obtenir de bonnes performances de prévision sur une période d'un an environ, les méthodes d'AS et d'AS d'ensemble gomment les biais de long terme des modèles de prévision. Des méthodes de filtrage ou de corrections en ligne du biais permettent d'améliorer encore les prévisions post-traitées. Ces corrections dynamiques permettent de corriger les biais sur des intervalles de temps courts, ou résultant de possibles changements dans l'implémentation du modèle de PNT post-traité (Fritsch et al. 2000; Woodcock and Engel 2005; Glahn 2014).

Météo-France, le service météorologique national français, a la charge de la surveillance et de la prévision de l'évolution de l'atmosphère, et de la production des alertes de risques météorologiques. Des alertes efficaces requièrent des prévisions météorologiques précises, afin d'éviter les manques et les fausses alarmes autant que faire se peut. Puisque les méthodes d'AS et d'AS d'ensemble requièrent des mesures, elles sont principalement empoyées pour les emplacements des stations météorologiques, sur les surfaces continentales.

Actuellement, des prévisions avec AS et AS d'ensemble sont disponibles pour plusieurs paramètres et modèles de PNT uniquement à l'emplacement des stations météorologiques, et à un pas de temps de 3 heures. Les utilisateurs finaux dans l'industrie et les secteurs du service (électricité, aéronautique,...), tels que le Conseil supérieure de la Météorologie (CSM), un organisme officiel rassemblant d'importants utilisateurs des prévisions météorologiques, demandent des prévisions améliorées en dehors des emplacements des stations météorologiques et à une résolution temporelle plus fine. Une évolution naturelle est de constuire des AS et AS d'ensemble sur la grille de modèles de PNT, avec une résolution temporelle d'une heure. Cela exige de revoir les méthodes actuelles du fait d'un changement d'échelle de la quantité de données traitées : environ 600 emplacements pour des AS ou AS d'ensemble construites pour des stations météorologiques, contre plusieurs milliers de points de grille pour un modèle de PNT typique. L'objectif de ce travail est de construire des prévisions améliorées pour la vitesse du vent mesurée chaque heure à dix mètres de hauteur, aux points de grille de différents modèles de PNT sur la France. La vitesse du vent a été choisie du fait de son impact potentiellement important pour la sécurité des personnes et des biens, et de la complexité des champs de vitesse de vent près du sol, ce qui constitue un bon test des méthodes de post-traitement. Le présent travail s'est déroulé à DOP (Données et outils de prévision), l'équipe opérationnelle de Météo-France en charge du post-traitement statistique des prévisions météorologiques. Étant donnée la visée opérationnelle de DOP, une attention constante a été d'utiliser des méthodes compatibles avec les contraintes opérationnelles, telles des contraintes de temps, de ressources informatiques et de capacités de stockage. Le travail a été effectué avec le langage statistique R (R Core Team 2015), pour des raisons de compatibilité avec le logiciel de post-traitement statistique développé en interne.

Ce résumé se poursuit en présentant, pour chaque chapitre de la thèse, les motivations pour les problèmes spécifiquement étudiés et les résultats de notre travail. La Section 0.2 aborde la question de la construction d'AS pour des prévisions de vitesse de vent sur une grille. La Section 0.3 présente des détails sur l'estimation du score de probabilité de rangs continus (CRPS), une mesure de performance très utilisée pour les prévisions probablistes d'une observation scalaire. Les résultats obtenus dans cette seconde partie sont utilisés dans la partie suivante de la thèse. La Section 0.4 introduit le dernier sujet, à savoir la comparaison de plusieurs manières de combiner différents ensembles post-traités par des méthodes d'AS d'ensemble. Enfin, la Section 0.5 conclue et résume les principaux sujets et résultats.

0.2 Chapitre 2: Improved Gridded Wind Speed Forecasts with Block MOS

0.2.1 Motivation

Sur les continents, les observations ne sont souvent disponibles qu'aux emplacements des stations météorologiques. De ce fait, les AS et AS d'ensemble ne sont généralement produites qu'à ces emplacements, qui peuvent ne pas coïncider avec les localisations intéressant les utilisateurs finaux. Un objectif naturel est de construire des prévisions post-traitées sur les grilles d'un modèle de PNT. Cela peut être effectué en interpolant des AS ou AS d'ensemble depuis les emplacements des stations vers les emplacements des points de grille, ou en appliquant les méthodes d'AS et d'AS d'ensemble sur une grille avec des mesures précédemment interpolées comme observations. Cette thèse utilise la seconde approche, qui permet de construire une archive d'observations en points de grille utile pour des études climatologiques et pour entraîner des méthodes d'AS et d'AS d'ensemble pour plusieurs modèles de PNT, ce qui est fait par la suite. De plus, comme indiqué dans Bosart (2003) et Novak et al. (2014), une analyse précise et en temps réel s'avère utile pour les prévisionnistes pour maintenir une connaissance suffisante des biais des AS, les corriger et ajouter leur expertise à des prévisions automatiques.

Des domaines étendus et/ou des grilles fines peuvent contenir des milliers de points de grille, ce qui demande des méthodes d'AS et d'AS d'ensemble rapides ou une optimisation soigneuse des codes pour obéir aux délais alloués pour produire et diffuser les prévisions aux utilisateurs finaux. Cela motive dans cette partie de la thèse une étude des moyens d'accélérer les traitements opérationnels.

Deux objectifs sont atteints : produire des pseudo-mesures de vitesse de vent en points de grille sur la France; et construire des méthodes rapides d'AS pour post-traiter les prévisions de vitesse de vent sur des milliers de points de grille sur la France, comme demandé par les utilisateurs finaux et les membres du CSM.

0.2.2 Résultats

Plusieurs techniques statistiques sont testées pour construire des mesures en points de grille, en interpolant la vitesse du vent mesurée aux emplacements des stations météorologiques vers une grille régulière de maille environ 0,025° (soit 2,5 km). Ces nouvelles analyses de vitesse de vent sont construites car les analyses existantes à Météo-France n'ont pas été jugées suffisamment précises à cause d'un défaut de vents forts et parce qu'elles n'étaient disponibles, jusqu'en Avril 2015, qu'au pas de temps tri-horaire, ce qui est insuffisant pour construire des AS horaires comme demandé. La principale difficulté ici est que la vitesse du

vent en surface constitue un phénomène très inhomogène et non stationnaire, dans l'espace et dans le temps. Des effets locaux, tels que l'effet Venturi, la friction de surface ou des effets d'abri par des obstacles interagissent d'une manière très complexe, s'ajoutant aux variations additionnelles selon les conditions météorologiques du jour. Par exemple, dans une traine, des cellules convectives s'accompagnent de courants de densité, c'est-à-dire de zones localisées de masses d'air descendantes qui accroissent la vitesse du vent sur quelques kilomètres. D'un autre côté, le long du front associé à une dépression, de fortes vitesses de vent existent sur plusieurs centaines de kilomètres de long et quelques dizaines de kilomètres de large. Mais des courants de densité peuvent également exister dans un front, renforçant le vent localement. Ces phénomènes, ainsi que d'autres tels que les brises côtières, les vents orographiques et d'autres, rendent difficile la modélisation des structures des champs de vent près de la surface. Dans notre étude, certaines techniques d'interpolation modélisent explicitement la dépendance spatiale, d'autres non. Afin de capturer la bonne longueur de corrélation du champ de vent, l'interpolation est faite sur des domaines de taille variable. La meilleure technique d'interpolation statistique, parmi 48 testées, fait appel aux splines plaque mince de régression (SPMR), ajustées sur la vitesse de vent mesurée au même instant en chaque station météorologique de France. La fonction d'interpolation s'avère très parsimonieuse, comportant seulement deux composantes additives: une première spline fonction de la plus récente prévision de vent moyen du modèle local de haute résolution de Météo-France, AROME, et une seconde spline fonction des coordonnées tri-dimensionnelle des points. Ce n'est pas surprenant : la modélisation de la dépendance entre points de grille est rendue inutile par l'usage des prévisions d'AROME, qui contiennent une information sur la structure spatiale du champ de vent. D'un autre côté, du fait de la complexité du champ de vitesse de vent, des méthodes classiques d'interpolation telles que le krigeage ne peuvent modéliser efficacement la dépendance spatiale du vent.

Une comparaison par validation croisée montre que la nouvelle analyse est plus précise que l'analyse AROME existante. La vérification des performances de l'analyse de vitesse de vent d'AROME n'avait jamais été effectuée. Cette étude montre que vérifier l'analyse AROME en utilisant des mesures de vitesse de vent non assimilées aboutit à des performances très différentes que celles obtenues lors d'une vérification avec les données assimilées. Un code implémentant ce banc d'essai de méthodes d'interpolation a été écrit, afin de l'appliquer à d'autres paramètres (la force des rafales de vent a également été interpolée). Les mesures en points de grille obtenues couvrent une période de quatre ans, débutant le 1er janvier 2011, au pas horaire. Cette longue archive sert d'observation pour ajuster les méthodes d'AS et d'AS d'ensemble testées dans la suite de ce travail.

Avec ces mesures en points de grille comme observation, plusieurs méthodes d'AS sont ensuite comparées pour post-traiter ARPEGE, le modèle global de Météo-France avec une maille de 10 km sur la France. La régression linéaire multiple sert d'AS simple de référence. Les forêts aléatoires constituent une méthode d'AS plus complexe. Cependant, alors que l'information nécessaire pour prédire avec la régression linéaire multiple nécessite quelques

kilo-octets de stockage sur disque et un temps de chargement en mémoire de quelques millisecondes, l'information nécessaire pour utiliser les forêts aléatoires requiert plusieurs méga-octets de stockage et un temps de chargement d'environ une demie-seconde. La France comportant plusieurs centaines de points de grille, des solutions pour traiter une telle quantité de données dans des délais opérationnels sont développées. Premièrement, des "AS par bloc" sont employées, à savoir que chaque forêt aléatoire est ajustée sur des points de grille proches simultanément. Pour les applications opérationnelles, la France sera divisée en blocs contigus, chacun étant post-traité avec sa propre AS. Ce traitement par bloc permet de réduire le nombre d'objets R utilisés pour stocker les forêts aléatoires, sans réduire les performances de prévision en comparaison de forêts aléatoires ajustées pour chaque point de grille séparément. Deuxièmement, les forêts aléatoires sont optimisées en réduisant le nombre et la profondeur de leurs arbres constitutifs, à nouveau sans diminution de la performance de prévision. Ces deux améliorations réduisent le nombre et la taille de stockage des objets R à charger en mémoire lors des applications. Le temps de chargement se voit réduit d'un facteur 10, soit un temps total estimé d'une demie minute pour toute la France. Les forêts aléatoires obtiennent de meilleures performances que la régression linéaire multiple. Ces AS par bloc avec des forêts aléatoires deviendront opérationnelles courant 2017, pour la vitesse du vent et la force des rafales.

Ce chapitre reproduit l'article Zamo et al. (2016) paru dans *Weather and Forecasting*.

0.3 Chapitre 3: Estimation of the CRPS with Limited Information

0.3.1 Motivation

Dans le Chapitre 4, plusieurs prévisions probabilistes sont combinées linéairement afin d'obtenir une prévision plus performante. Pour certaines méthodes de combinaison testées, les poids de la combinaison dépendent d'une mesure de performance des prévisions probablistes, le score de probabilité des rangs continus (CRPS). Certaines distributions prévues sont paramétriques et donc totalement connues, ce qui autorise un calcul analytique exact du CRPS. D'autres distributions prévues sont non paramétriques, et donc partiellement connues, le CRPS devant alors être estimé. Afin de calculer de manière précise les poids de la combinaison, le CRPS des prévisions probabilistes non paramétriques doit être estimé précisément, ce qui motive l'étude de ce chapitre.

Lorsque une distribution prévue est partiellement connue à travers un ensemble de M valeurs (appelées "membres" en Météorologie), comme c'est le cas dans le Chapitre 4, la précision de l'estimation du CRPS et le temps de calcul de ce score augmentent avec M. Une motivation de ce chapitre est d'estimer la précision de l'estimation du CRPS avec le

nombre de membres, afin de dégager un bon compromis entre une estimation précise et un nombre de membres réduit, ce qui accélère les calculs.

Nous avons établi des bornes d'erreur de l'estimation du CRPS, qui se sont avérées d'un intérêt pratique limité car demandant de connaître la distribution prévue. De ce fait, l'approche utilisée dans cette partie du travail est de simuler des cas pour lesquels la distribution prévue est partiellement connue et de calculer les estimations associées du CRPS. Les simulations sont effectuées en supposant que la distribution prévue consiste en une distribution paramétrique dont le CRPS possède une forme analytique. Cela permet une étude empirique des biais et variance d'estimation.

0.3.2 Résultats

Quatre estimateurs du CRPS proposés dans la littérature sont comparés. On démontre qu'ils se réduisent à deux, et une relation est établie entre ces deux estimateurs. Ces résultats ne requièrent aucune hypothèse. En particulier, l'ensemble des M valeurs peut ne pas être issu d'un tirage aléatoire selon une distribution, et aucune hypothèse de stationnarité ou d'échangeabilité n'est requise.

Suite à ces résultats, des simulations permettent d'étudier la précision des deux estimateurs avec un ensemble de valeurs aléatoires issues de la distribution prévue ou un ensemble de quantiles d'ordres connus de la distribution prévue. D'autres études ne concernaient que des ensembles aléatoires. Cette étude avec des ensembles de quantiles est motivée par le Chapitre 4, où de tels ensembles sont employés. Les simulations montrent que l'estimateur le plus précis dépend du type d'ensemble prévu : un échantillon aléatoire ou un ensemble de quantiles d'ordres connus. Pour un ensemble aléatoire, les simulations confirment que le "fair CRPS" de Ferro et al. (2008) est non biaisé, et elles dégagent des indications empiriques sur la variabilité de l'estimateur pour des faibles nombres de paires de prévisions et d'observations. Dans le cas d'un ensemble de quantiles, l'estimation la plus précise s'obtient en choisissant les ordres optimaux introduits dans Bröcker (2012), mais le fair CRPS ne constitue pas un bon estimateur. La présence de valeurs égales dans un ensemble de quantiles, possible avec certaines méthodes d'AS d'ensemble employées dans le Chapitre 4, peut fortement réduire la précision de l'estimation du CRPS. Une méthode d'interpolation est proposée pour rétablir une meilleure précision. À notre connaissance, cette possible existence d'ex aequos dans des prévisions d'AS d'ensemble n'a pas été pointée dans les études précédentes, ni son impact sur l'estimation du CRPS. Suite à ces simulations, des recommendations sont émises pour estimer précisément le CRPS d'une prévision probabiliste lorsque la distribution prévue n'est pas exhaustivement connue. Bien que le caractère non biaisé du "fair CRPS" d'un ensemble aléatoire peut être démontré, il n'a pas été possible de justifier théoriquement ces recommendations dans le cas d'un ensemble de quantiles d'ordres connus. Mais une large gamme de distributions prévues ayant été utilisées durant

les simulations, cela conduit à penser que les recommendations sont assez générales. De plus, une explication qualitative mais logique de la bonne précision d'estimation dans le cas d'un ensemble de quantiles est avancée. Cette explication ne dépend pas de la distribution prévue, ce qui renforce la présomption sur la généralité de nos résultats.

0.4 Chapitre 4: Aggregation of Probabilistic Wind Speed Forecasts

0.4.1 Motivation

Les utilisateurs finaux des prévisions météorologiques auraient intérêt à utiliser des prévisions probabilistes, mais ne sont pas satisfaits de leur performance. Des prévisions fiables constitueraient une aide pour évaluer l'incertitude de la prévision et pour la prise de décision en contexte incertain. Cette partie de la thèse compare dans un banc d'essai plusieurs méthodes pour combiner de manière linéaire plusieurs prévisions d'ensemble opérationnelles à Météo-France et ces mêmes prévisions d'ensemble post-traitées avec des méthodes d'AS d'ensemble. La combinaison d'AS d'ensemble vise à améliorer la prévision par rapport à une seule AS d'ensemble. Le point principal est ici la méthode de combinaison plus que les méthodes d'AS d'ensemble, dont l'apport a été étudié dans une autre thèse à Météo-France (Taillardat et al. 2016). Dans la présente thèse, trois moyens permettent d'améliorer la performance de la prévision combinée. D'abord, l'apprentissage de plusieurs méthodes d'AS d'ensemble sur des périodes glissanes ou fixes fournit une vaste gamme de performances. Ensuite, l'utilisation de plusieurs méthodes d'ensembles et de différents ensembles vise à produire des prévisions avec des structures d'erreurs différentes dans le temps et l'espace. Enfin, les poids de la combinaison sont calculés sur une période glissante, afin que la méthode de combinaison puisse s'adapter rapidement à l'évolution des performances et puisse pondérer fortement les prévisions avec les meilleures performances.

Du fait de la variabilité importante de l'état de l'atmosphère dans le temps et l'espace (Slingo and Palmer 2011; Holton and Hakim 2012; Rohli et al. 2013), la plupart des méthodes d'AS ou d'AS d'ensemble requièrent une archive d'observations et de prévisions associées couvrant plusieurs années et un nombre suffisant de situations météorologiques différentes. Le code du modèle de PNT doit rester contant sur ces années d'apprentissage, de sorte que les structures d'erreurs de prévision puissent être modélisées par les méthodes d'AS et d'AS d'ensemble. Les modèles de PNT sont généralement modifiés une ou deux fois par an, de sorte que de légères modifications des performances s'accumulent au fil du temps ou que d'importantes améliorations de performances surviennent. Ces changements finissent par rendre caduque la relation statistique construite précédemment entre les observations et les prévisions. Par exemple, Erickson et al. (1991) étudie l'impact des

changements d'un modèle de PNT sur la performance d'une AS pour différents paramètres, et montre que la performance de l'AS pour la couverture nuageuse est diminuée. Le schéma de combinaison proposé plus haut tente de traiter ce problème. Une autre solution serait de construire des reprévisions, c'est-à-dire prévoir le passé avec la nouvelle version du modèle (Wilks and Hamill 2007; Hagedorn et al. 2008; Hamill et al. 2008) puis de redévelopper une AS. Cela s'avère toutefois cher en ressources, et la combinaison de prévisions pourrait rendre cela inutile, comme suggéré dans Miller et al. (1992).

Le post-traitement est effectué sur une grille, avec pour observation les mesures de vitesse de vent en points de grille présentées au Chapitre 2. Les quatre ensembles employés proviennent du projet TIGGE (Bougeault et al. 2010; Swinbank et al. 2016), qui rassemble les prévisions d'ensemble de différents services météorologiques.

0.4.2 Résultats

Deux méthodes d'AS d'ensemble sont employées : la régression non homogène (Hemri et al. 2014) et les forêts aléatoires quantiles (Meinshausen 2006). Les forêts aléatoires quantiles (QRF) montrent de bonnes performances pour la vitesse du vent, comme montré dans une précédente étude (Taillardat et al. 2016). Ces bonnes performances découlent du caractère non paramétrique des QRF, ce qui leur permet d'ajuster la distribution de la vitesse du vent de manière plus flexible que la régression non homogène (NR). Comme QRF requiert une longue période d'apprentissage, elles sont toutefois combinées avec NR, moins flexible mais pouvant être entraînée sur une période plus courte grâce à son nombre réduit de paramètres d'ajustement. Conformément à Hemri et al. (2014), qui montre que c'est le meilleur choix pour la vitesse du vent, une distribution normale tronquée en zéro de la racine carrée de la vitesse du vent est choisie pour NR. Plusieurs versions sont ajustées sur des fenêtres d'apprentissage de tailles différentes, afin de construire des distributions prévues qui évoluent plus ou moins rapidement. Cela vise à permettre à la combinaison d'adapter ses poids à des régimes de temps plus ou moins stables.

Puisque QRF ne permet de produire que des fonctions de répartition (CDF) en escalier, les différentes versions de NR sont également transformées en CDF en escalier pour rendre possible la combinaison des prévisions. Cela s'inspire de la thèse de Baudin (2015), qui combine d'une manière similaire les valeurs prévues *triées par valeur croissante* de plusieurs ensembles mis en commun. Cela signifie que les prévisions probabilistes individuelles sont des CDFs en escalier avec une seule marche. Ces prévisions individuelles de Baudin (2015) ne sont pas identifiables au cours du temps, ce qui est requis par la théorie de la prévision avec avis d'experts employée dans ce travail et le nôtre. Par exemple, la valeur minimale à différents instants peut être issue d'un membre différent et/ou d'un ensemble différent. Étant un ensemble (post-traité ou brut) en soi, chacune de nos prévisions combinées constitue une prévision probabiliste identifiable au fil du temps. Afin d'utiliser les

méthodes de combinaison, nous avons dû d'abord généraliser certaines formules de Baudin (2015), valides pour des CDFs en escalier avec une marche, au cas de CDFs en escalier avec un nombre quelconque de marches. Finalement, avec quatre ensembles bruts et six versions post-traitées de chaque ensemble, 28 CDFs en escalier sont combinées avec des poids convexes pour produire une CDF en escalier valide. Les poids de la combinaison sont calculés selon cinq méthodes. Certaines de ces méthodes sont empiriques. D'autres sont basées sur la théorie de la prévision avec avis d'experts présentée dans ce chapitre, et possèdent d'intéressantes propriétés théoriques en termes de performance. L'utilisation de cette théorie pour la combinaison de prévisions probabilistes est relativement innovante dans le cadre de la Météorologie.

Deux critères de performance sont employés : le CRPS moyenné sur tous les points de grille et tous les instants, et la proportion de points de grille avec un histogramme des rangs plat. L'histogramme des rangs est simplement l'histogramme du rang de l'observation lorsqu'elle est ajoutée à l'ensemble des valeurs prévues associées. La platitude de l'histogramme des rangs, condition nécessaire pour une prévision fiable, est évaluée avec les tests de platitude de Jolliffe-Primo, introduits dans Jolliffe and Primo (2008). Ces tests estiment l'existence de types de déviation particulière à la platitude de l'histogramme des rangs. Ils ne sont pas encore très utilisés pour la vérification de prévision, bien qu'ils éclairent l'origine des défauts des prévisions probabilistes, telles qu'un biais ou un manque de dispersion. Ces tests sont employés ici comme diagnostic automatique, ce qui est rendu nécessaire par le nombre important de points de grille à traiter opérationnellement. Ces deux critères de performance ne conduisent pas au même classement des 28 prévisions individuelles et des méthodes de combinaison. Minimiser le CRPS peut s'obtenir avec très peu d'histogrammes des rangs plats, tandis que la maximisation du nombre d'histogrammes des rangs plats s'accompagne d'une augmentation du CRPS très faible. Ce résultat est nouveau car l'approche habituelle est de minimiser dans un premier temps le CRPS puis de tester si l'histogramme des rangs peut être considéré plat. Il est donc proposé de choisir la méthode de combinaison selon la platitude de l'histogramme des rangs testée avec les tests de platitude de Jolliffe-Primo. Selon ce critère, la meilleure méthode de combinaison est la pondération par poids exponentiels (EWA), basée sur la théorie de prévision avec avis d'experts. EWA produit une importante proportion d'histogrammes des rangs plats sur la France (environ 85%), tout en obtenant un CRPS moyen similaire à celui de la méthode de combinaison minimisant le CRPS. La meilleure prévision avec EWA, bien qu'ajustée séparément pour chaque point de grille et échéance de prévision, présente une corrélation spatio-temporelle. Également, les poids de la combinaison peuvent varier rapidement comme attendu, ce qui autorise une adaptation rapide de la combinaison aux modifications des ensembles de PNT. Finalement, le choix optimal des paramètres de EWA peut être déterminé avec une seule année d'observations et de prévisions, ce qui permet de mettre à jour fréquemment le post-traitement par AS d'ensemble.

0.5 Conclusion

Cette thèse étudie plusieurs aspects du post-traitement de la prévision météorologique de la vitesse du vent sur la France. L'objectif est de produire des prévisions de vitesse de vent améliorées sur une grille, pour des prévisions déterministes et probabilistes. La stratégie adoptée procède en interpolant d'abord les mesures aux stations météorologiques vers la grille d'un modèle de haute résolution, puis en utilisant ces mesures en points de grille pour entraîner des méthodes d'AS et d'AS d'ensemble.

Pour les prévisions déterministes, des AS par blocs sont introduites, ainsi qu'une optimisation méticuleuse de la taille et des objets R associés. Ces AS par blocs allégées présentent de bonnes performances tout en autorisant une importante accélération des traitements en conditions opérationnelles. Ces AS sont en cours d'implémentation.

En ce qui concerne les AS d'ensemble, des méthodes empiriques de combinaison de prévisions et des méthodes de combinaisons basées sur la théorie de la prévision avec avis d'experts sont comparées. Puisque des CDFs en escalier sont combinées, cette partie de la thèse requiert d'étudier les propriétés des estimateurs du CRPS lorsque l'information sur la distribution prévue est limitée. Cette étude conduit à des recommandations pour estimer précisément le CRPS. Également, à cause de différences entre la platitude des histogrammes de rangs et de la valeur du CRPS, il est proposé de choisir parmi les prévisions probabilistes en imposant dans un premier temps la platitude de l'histogramme des rangs selon les tests de Jolliffe-Primo. La meilleure méthode de combinaison choisie selon ce critère obtient un CRPS similaire à celui de la prévision individuelle minimisant le CRPS, tout en présentant bien plus d'histogrammes des rangs jugées plats. Ces méthodes doivent également être mises en opérationnel dans le courant de l'année 2017.

Chapter 1

Introduction and Summary

Abstract This PhD thesis deals with the improvement of weather forecasts by statistical post-processing methods in an operational context. Improved wind speed forecasts are built on a grid, which poses issues due to the potentially high number of grid points in current weather forecast models. This introduction begins with a brief description of the post-processing procedure of weather forecasts and the motivation of the chosen topic. Then each part of the work is successively presented, with the motivations to the specific addressed issues, the state of the art and the main results.

Contents

1.1 Introduction 18
1.2 Chapter 2: Improved Gridded Wind Speed Forecasts with Block MOS 20
1.2.1 Motivation 20
1.2.2 State of the art of wind speed MOS and gridded MOS 20
1.2.3 Results 22
1.3 Chapter 3: Estimation of the CRPS with Limited Information 24
1.3.1 Motivation 24
1.3.2 State of the art of the estimation of the CRPS 25
1.3.3 Results 26
1.4 Chapter 4: Aggregation of Probabilistic Wind Speed Forecasts 27
1.4.1 Motivation 27
1.4.2 State of the art of EMOS and combination thereof 28
1.4.3 Results 29

1.5 Conclusion 30

1.1 Introduction

Most national weather services and some private firms use numerical weather prediction (NWP) models to produce weather forecasts and derived predictions in weather-sensitive fields, such as weather-related economic risks (Alexandridis and Zapranis 2012; Bertrand et al. 2015), renewable energy production (Costa et al. 2008; Dubus et al. 2014; Zamo et al. 2014a,b), electricity consumption (Taylor and Buizza 2002), air quality (Mallet and Sportisse 2006; Besse et al. 2007), agriculture (Cantelaube and Terres 2005; Baker and Kirk 2007; Trnka et al. 2011), to name but a few. Every NWP model is prone to errors that are not completely random and, as such, may be partly reduced thanks to statistical post-processing techniques that have been used since the early times of modern weather forecasting (Glahn and Lowry 1972).

The most frequent approach to post-process weather forecasts is to automatically build a statistical relationship between past observations and associated NWP forecast parameters thanks to the many statistical tools developed in the fields of data mining or machine learning (Hastie et al. 2009; Kuhn and Johnson 2013; Alpaydin 2014). The relationship is then applied to future forecasts to improve their performance, or to forecast observed parameters not directly predicted by the NWP model. This approach is called model output statistics (MOS) for a deterministic, or point, NWP forecast, and ensemble MOS (EMOS, Gneiting et al. 2005) when the post-processed forecast uses the several values generated by an ensemble NWP model (Leutbecher and Palmer 2008). MOS methods forecast a scalar value that may be the expected mean of the observation (Jacks et al. 1990; Zamo et al. 2014a) or any quantile (Friederichs and Hense 2008), see also Gneiting (2011a). EMOS methods forecast all or several aspects of the predictive distribution of the observation (Oesting et al. 2013; Zamo et al. 2014b). Although increasingly sophisticated and skillful NWP models have been developed during the last decades (Lynch 2008; Inness and Dorling 2013; Bauer et al. 2015), (E)MOS methods can still improve forecast performances, e.g. (Ruth et al. 2009; Hemri et al. 2014; Taillardat et al. 2016).

Since no theoretical result exists to choose *a priori* the best (E)MOS methods for a problem at hand, several (E)MOS methods, and sophistications thereof, are put to the test, in order to choose the best one according to some performance criterion. However, several studies (Bates and Granger 1969; Clemen 1989) have shown that combining different (post-processed or not) forecasts from one or several NWP models (Fritsch et al. 2000) may improve forecast performance of deterministic (Vislocky and Fritsch 1995) and ensemble forecasts (Baudin 2015). Again, many combination approaches exist (Cesa-Bianchi et al.

2006; Allard et al. 2012; Gneiting et al. 2013), among which one chooses after an empirical comparison (Palm and Zellner 1992; Mallet et al. 2007; Gerchinovitz et al. 2008).

Being usually fitted to get good forecast performances over a year or so, (E)MOS remove the long-term biases in the forecast errors. Filtering techniques or online bias corrections may thus further improve the post-processed forecasts. These dynamical corrections may remove biases occurring on short time intervals or resulting from possible changes in the implementation of the post-processed NWP model since the (E)MOS training (Fritsch et al. 2000; Woodcock and Engel 2005; Glahn 2014).

Météo-France, the French national weather service, is in charge of surveying and forecasting the evolution of the atmosphere, and to issue warnings in case of forthcoming weather hazards. Efficient warnings require accurate weather forecasts, to avoid misses and false alarms as much as possible. Since (E)MOS requires measurements, they are mainly built for the locations of meteorological stations, over continental areas. Currently, (E)MOS forecasts are available for several parameters and several models at meteorological station locations only, and with a time-step of 3 h. End-users from the industry and service sectors (electricity, aeronautics,...), such as the "Conseil supérieur de la météorologie" (CSM, High Council of Meteorology), an official gathering of important end-users of weather forecasts, require improved forecasts at locations that may not correspond to meteorological stations and with a finer temporal resolution. Thus a natural evolution is to build (E)MOS on the grid of NWP models, with a temporal resolution of 1 h. This requires to revise current methods due to a change of scale, from about 600 locations for (E)MOS at meteorological station locations to thousands of locations on a typical model grid. The objective of this work is thus to build improved 10 m hourly wind speed forecasts at the grid points of different NWP models over France. Wind speed has been chosen due to its possible severe impact on the safety of people and goods, and because the complexity of surface wind speed fields constitutes a good test for post-processing methods. The present work takes place at DOP ("Données de prévisions", or forecast data), the operational team of Météo-France in charge of the statistical post-processing of weather forecasts. Due to the operational goals of DOP, a constant concern is also to use methods compatible with operational constraints, such as time delays, available computer resources and memory storage. The work is done with the R statistical language (R Core Team 2015), for compatibility with the team's internally-developed tool for statistical post-processing.

This introduction proceeds by presenting, successively for each chapter of this study, our motivation to study the specific issues of the concerned chapter, then the state of the art and finally the results of our work. Section 1.2 deals with the issue of building MOS methods for wind speed forecasts on a grid. Section 1.3 gives details on the estimation of the continuous ranked probability score (CRPS), a performance measure much used for probabilistic forecasts of a scalar observation. The results obtained in this second part are required in the following part of the work. Section 1.4 presents the last topic, namely the

comparison of several ways to combine different EMOS-post-processed ensembles. Last, Section 1.5 concludes and summarizes the main topics and results.

1.2 Chapter 2: Improved Gridded Wind Speed Forecasts with Block MOS

1.2.1 Motivation

Over continental areas, observations are often available only at locations of meteorological stations. Therefore, (E)MOS are usually produced at these locations only, which may not coincide with end-users' desired locations. Thus a natural goal is to build post-processed forecasts at the grid points of a NWP model. This may be achieved by interpolating (E)MOS forecasts from station locations to grid point locations, or apply (E)MOS methods on a grid with previously gridded measurements as the observation. This work adopts the second approach, because it allows to build an archive of gridded observations that may be used for climatic studies and for training (E)MOS methods for several NWP models, which is done in the following. Furthermore, as stated in Bosart (2003) and Novak et al. (2014), a real-time accurate analysis of the atmospheric state is required for human forecasters to keep a sufficient knowledge of MOS biases to correct them and add a value to automatic forecasts. Large domains and/or fine grids may contain thousands of grid points, which requires fast (E)MOS methods or a careful optimization of the code to abide by the time delays allotted to produce and diffuse the forecasts to the end-users. This motivates a study of means to speed up operations in this part of the work.

Two goals are achieved: gridded measurements of wind speed over France; and quick MOS methods to post-process wind speed forecasts on thousands of grid points over France, as required by end-users and members of the CSM.

1.2.2 State of the art of wind speed MOS and gridded MOS

MOS methods for wind speed forecasts have been applied and studied deeply in the field of wind power generation, due to the important economic impacts of wind speed forecast errors for wind power firms and grid regulators. These methods are usually applied to the locations of wind farms only. Based on the reviews of Lei et al. (2009) and Jung and Broadwater (2014), two main approaches are used in the wind power field: time series analysis methods or data mining techniques.

As for the time series methods, autoregressive (AR), moving average (MA), autoregressive-moving average (ARMA) and so on are used to forecast wind speed or wind power for very

short term (less than 6 h ahead). These models build the forecast as a linear combination of past forecasts and/or forecast errors. For instance, Liu et al. (2010) decompose the wind speed time series in wavelet components, model each wavelet component with a time series method, then add the forecasted component wavelets.

For farther lead times, the main approach is to use support vector machines (SVM) or neural network (NN). Colak et al. (2012) review data mining techniques for wind speed and wind power forecasting. Some applications build an hybrid model with a modeling of the linear evolution of wind speed with time series methods, and a correction with SVM or NN to take account of the nonlinear evolution. Bhaskar and Singh (2012) decompose the wind speed time series on wavelets, then uses each wavelet component to regress on with an adaptive wavelet neural network. Guo et al. (2012) decompose the wind speed time series on a small number of mode functions, use NN to forecast each sub-series (with a variable selection step) then add the different forecast components. Douak et al. (2013) select weather predictors with an active learning technique, then use kernel ridge regression to build the MOS. Shi et al. (2012) and Cadenas and Rivera (2010) use ARIMA to model the linear evolution of the wind speed time series, and NN or SVM to model the nonlinear part. With ARIMA, Liu et al. (2012) choose the structure of an NN MOS or initialize a Kalman filter that serves to build their MOS. Haque et al. (2012) build different versions of NN, and add as inputs the average observed values on "similar days". Adding the similar days' average observation improves the performance over the NN without the similar days. Qin et al. (2011) choose dynamically between an NN method and a persistence method according to the wind conditions. In the meteorological community, Sfanos and Hirschberg (2000) from the National Oceanographic and Atmospheric Administration (NOAA) use multiple linear regression. Kusiak et al. (2009) compare the performance of different data mining techniques (SVM, NN, regression trees and random forest) to forecast wind speed and wind power up to a forecast horizon of 84 h. Some authors build several MOS forecasts and combine them to further improve their performance. Bouzgou and Benoudjit (2011) combine four MOS forecasts (linear regression, two neural networks and SVM), with three combination strategies (a simple average, a weighted average and NN). Li et al. (2011) use the Bayesian framework to combine several neural network MOS, for short term forecasts. Zhou et al. (2011) compare several SVM-based MOS with different kernels, and conclude that the choice of the kernel is not important but that forecast performances are sensitive to the value of the fitting parameters. Cheng and Steenburgh (2007) compare traditional MOS to bias removal with a Kalman filter or a 7-day running mean. They conclude that traditional MOS work better when the weather changes, the others perform the best during quiescent cool regimes, and both approaches are equivalent during quiescent warm regimes. They do not propose to combine the forecasts to try and get the best of each.

As for the gridded MOS, no consensus exists in the literature on whether one should grid MOS previously built at measurement locations, or grid measurements before building MOS on a grid. To the best of our knowledge, no systematic comparison of the two

approaches has been achieved. The NOAA uses gridded MOS operationally, by gridding MOS built at station locations. Glahn et al. (2009) and Gilbert et al. (2009) detail how their method iteratively corrects grid-point forecasts after comparing them to MOS built at nearby station locations. Based on the same methodology, Im et al. (2010) detail the NOAA analysis to grid hourly measurement of surface parameters, in order to verify the gridded MOS and produce very-short term forecasts. Mass et al. (2008) grid MOS by estimating the bias at station locations, associating each grid point to a station with similar elevation and/or land-use characteristics and estimating the bias correction at the grid point. Solari et al. (2012) use simulations with a very fine grid of 270 m on port areas to build a linear relationship between wind speed measurements and simulated values over the area. The statistical relationship built at measurement locations is used to correct wind speed forecasts all over the grid. Burlando et al. (2010) adopt a similar approach to forecast wind speed along a railway line. Charba and Samplatsky (2009) and Charba and G. Samplatsky (2011) build their MOS on a grid by using as observation an analysis of rainfall accumulation. Thorey et al. (2015) use previously gridded radiation measurements as observation.

1.2.3 Results

Several statistical techniques are tested to obtain gridded measurements, by interpolating wind speed measured at station locations towards a regular grid with a grid size of about 0.025° (about 2.5 km). These new wind speed analyses are built because existing analyses at Météo-France were not deemed accurate enough due to a lack of high wind speed or, until April 2015, were available for a time-step of 3 hours, which is not sufficient to build hourly MOS as required. The main difficulty here is that surface wind speed is a very unstationary and inhomogeneous phenomenon, in space and time. Local effects, such as tunneling, surface friction or sheltering by obstacles interact in very intricate ways, along with additional variations due to specific meteorological conditions of the day. For instance, in the tail end of a low, convective cells are accompanied with so-called density currents, that is, areas with descending air masses that increase surface wind speed over a few kilometers. On the other hand, along the front of a storm, strong wind speeds exist in areas of hundreds of kilometers in length and of only a few tens of kilometers in width. But density currents can also exist in some areas of the front, increasing wind speed locally. These, and other phenomena such as coastal breezes, orographic winds, and so on, make it very difficult to model the structure of surface wind fields. In our study, some interpolation techniques explicitly model the dependence between locations, others do not. In order to catch the right correlation length of the wind speed field, the interpolation is done over domains of varying sizes. The best statistical interpolation technique, among 48 techniques or variations tested, is based on thin plate regression splines (TPRS), trained on wind speed measured at all meteorological stations in France at the same time. The interpolation

function is very parsimonious with only two additive components: a first spline with the most recent wind speed forecast of Météo-France's fine grid, local area model AROME as the only input, and a second spline with a correction based on the 3-dimensional coordinates of the points. This is not surprising: the modeling of the dependence between grid-points is made unnecessary by the use of AROME forecasts, that contain information about the spatial structure of the wind field. On the other hand, due to the complexity of the wind speed field, classical spatialisation methods such as kriging cannot efficiently model the spatial dependence of wind speed.

By cross-validation, it is shown that this new analysis performs consistently better than the available AROME analysis. The verification of AROME wind speed analysis had never been done with cross-validation. This study shows that verifying AROME analysis against non assimilated wind speed measurements gives very different performances than when it is verified against assimilated data. A flexible code implementing this test-bed of interpolation methods has been written, to be applied to other parameters (gust wind speed has also been interpolated). The obtained gridded measurements are produced for a period of 4 years, starting on 01 January 2011, every hour. This long archive is used as the observation for training the (E)MOS tested in the remaining of this work.

Using these gridded measurements as the observation, several MOS methods are then compared to post-process ARPEGE, Météo-France's global model with a grid size of 10 km over France. Since forecasts up to several days ahead are required, only data mining techniques have been retained, the literature showing that time series methods perform well only up to a few hours ahead. SVM and NN have not been tried since, as noticed in Zhou et al. (2011) and during previous studies at DOP (unpublished), these methods obtain good forecast performances if their parameters are tuned with a great accuracy, which may not be achieved quickly enough in an automatic way on thousands of grid points. One tested technique is based on functional regression and forecasts the whole curve of hourly wind speed for the next 24 h by averaging previously observed curves. The goal is to take advantage of the diurnal cycle of wind speed, and to produce a realistic evolution of the forecast wind speed. Such functional MOS have been rarely studied in the literature about wind speed forecasts. Although this functional MOS improves over ARPEGE forecasts, it does not perform as well as the two other MOS trained for each lead time separately, probably because the diurnal cycle of wind speed is not regular enough. Multiple linear regression serves as a simple reference MOS. Random forests are used as a more complex MOS method. However, whereas the information to predict with multiple linear regression can be stored on a disk as a few kilobytes and uploaded in memory in a few milliseconds, the information necessary to use random forests can require several megabytes and a loading time of about half a second. Since France contains thousands of ARPEGE grid points, ways to treat such a large amount of grid points within operational delays have to be found. First, "block MOS" is used, meaning each random forest MOS is trained on nearby grid points pooled together. For operations, France will be divided into

contiguous blocks, each of which will be post-processed with its own MOS relationship. This pooling allows to reduce the number of R objects used to store the random forest without reducing the forecast performance compared to MOS trained grid-point-wise. Second, random forests can be optimized by reducing the number and depth of its constituent trees, again with no reduction in forecast performances. These two improvements reduce the number and storage size of R objects to load in memory when it comes to effectively produce the forecasts in operations. A reduction of loading time by a factor 10 is achieved, for an estimated total loading time of half a minute for the whole France. Random forests outperform multiple linear regression. This block MOS with random forest is expected to be implemented on Fall 2016, for wind speed and gust wind speed. This will produce full post-processed wind speed fields over France, which was not done during the work. Before going operational, it will be necessary to check that this block MOS does not produce unrealistic wind speed discontinuities at the boundaries of the blocks.

This chapter reproduces an article accepted in *Weather and Forecasting*.

1.3 Chapter 3: Estimation of the CRPS with Limited Information

1.3.1 Motivation

In Chapter 4, several probabilistic forecasts are linearly combined to get a more skillful forecast. For some of the tested combination methods, the combination weights depend on a performance measure for probabilistic forecasts, the continuous ranked probability score (CRPS). These probabilistic forecasts are either parametric and fully known, in which case the CRPS can be computed from a closed-form expression, or nonparametric and partly known, in which case the CRPS has to be estimated. To compute accurate combination weights, the CRPS of the nonparametric probabilistic forecasts have to be precisely estimated, which motivates the study in this chapter.

When the forecast distribution is partly known through a set of M values (called an "ensemble" in Meteorology), as it is the case in Chapter 4, both the precision of the CRPS estimation and the computation time increase with M. A motivation of this chapter is to assess the precision of the CRPS estimation with M, in order to find a good compromise between a high estimation precision and a low size M, which speeds up computation during operations.

We established error bounds of the CRPS estimation, that were of no practical use since they require to know the forecast distribution. Therefore, the approach used in this part of the work is to simulate cases where the forecast distribution is partly known and com-

pute the associated CRPS estimations. The simulations are produced by supposing the forecast distribution is some parametric distribution for which the CRPS has a closed form expression. This allows an empirical study of the estimation bias and variance.

This work was done under the precious guidance of Philippe Naveau.

1.3.2 State of the art of the estimation of the CRPS

The effect of the ensemble size M on the estimation of different performance measures for probabilistic forecasts has been studied in several articles, whose results are gathered and compared in Ferro et al. (2008). As for the CRPS, Ferro et al. (2008) demonstrate that, for an exchangeable and stationary ensemble and a stationary observation, the expectation of the usual CRPS estimator with an ensemble of size M, noted $\mathbb{E}[CRPS](M)$ is biased with

$$\mathbb{E}[CRPS](M) = \mathbb{E}[CRPS](\infty) + \frac{1}{M}\frac{\mathbb{E}|X_1 - X_2|}{2}, \tag{1.1}$$

where $\mathbb{E}[CRPS](\infty)$ is the expectation of the CRPS of an infinite ensemble, that is, the true CRPS, and X_1, X_2 are any two of the M available values for the same forecast/observation pair. This result holds for an ensemble of values randomly drawn from the unknown forecast distribution. Based on these remarks, Ferro et al. (2008) propose an unbiased version of the CRPS for a random ensemble. Ferro (2014) introduces the notion of fair scores, a measure of performance that is minimized for an ensemble of random values drawn from the same distribution as the observation. The unbiased estimator of the CRPS is shown to be fair. Specific structures of dependence between the random values may be taken account of to develop a fair CRPS.

The CRPS can be decomposed in the sum of three terms: the reliability term that quantifies the agreement between a forecast distribution and the distribution of the associated observation; the resolution term that quantifies the ability of the ensemble model to associate different distributions of the observation to different forecast distributions; and the uncertainty term that quantifies the variability of the observation (Bröcker 2009). Among other topics, Candille (2003) investigates the effect of a finite ensemble size and of a finite number of forecast/observation pairs on the reliability and resolution terms. The impact of the ensemble size on the usual estimation of the CRPS is also studied, leading to Equation (1.1). As Ferro et al. (2008), Candille (2003) also supposes that the ensemble is a random sample from an unknown forecast distribution.

Bröcker (2012) investigates the nature of the ensemble of M *fixed* values that minimizes the estimator of the CRPS averaged over the distribution of the observation. It demonstrates that these fixed values are the quantiles of orders $\frac{0.5}{M}, \ldots, \frac{M-0.5}{M}$ of the forecast distribution.

1.3.3 Results

Four estimators of the CRPS proposed in the literature are reviewed and shown to reduce to only two, and a relationship between these two is demonstrated. These results do not require any hypothesis. Specifically, the ensemble of M values is not required to be a random sample from the forecast distribution, and no stationarity or exchangeability assumption is required. The relationship between the two estimators relaxes all the hypotheses leading to Equation (1.1) and is valid for one forecast/observation pair.

Based on these findings, simulations are used to study the accuracy of the two estimators with an ensemble of random values from the forecast distribution or an ensemble of quantiles of known orders from the forecast distribution. Other studies focused only on random ensembles. This focus on ensembles of quantiles is motivated by Chapter 4, where such ensembles are used. The simulations show that the most accurate estimator depends on the kind of forecast ensemble: a random sample or a set of quantiles of known orders. For a random ensemble, the simulations confirm that the fair CRPS of Ferro et al. (2008) is, as proved, unbiased, and also add some new empirical indications about the variability of this estimator for small samples of forecast/observation pairs. In the case of an ensemble of quantiles, the most accurate estimation is obtained with the optimal set of orders found in Bröcker (2012), but the fair CRPS is not a good choice of estimator. The presence of ties in the ensemble of quantiles, possible with some EMOS methods used in Chapter 4, may dramatically decrease the accuracy of the CRPS estimation. An interpolation method is proposed to recover a better accuracy. To the best of our knowledge, this possible existence of ties in some EMOS forecasts has not been highlighted in previous studies, and its impact on the CRPS estimation has not been assessed. Based on these simulations, recommendations are issued to accurately estimate the CRPS of a probabilistic forecast when the forecast distribution is not fully known. Whereas the unbiasedness of the fair CRPS for a random ensemble can be demonstrated, it has not been possible to find theoretical justification of the validity of these recommendations in the case of an ensemble of quantiles of known orders. But a large variety of forecast distributions has been used in the simulations, from simple ones to complex mixtures, which leads to believe that the recommendations are quite general. Furthermore, a qualitative but sensible explanation of the good estimation accuracy in the case of an ensemble of quantiles is advanced. Since this explanation does not depend on the forecast distribution, it also hints at the generality of our results.

1.4 Chapter 4: Aggregation of Probabilistic Wind Speed Forecasts

1.4.1 Motivation

End-users of weather forecasts would be interested in using probabilistic forecasts, but may not be satisfied by their current performances (personal communications). Reliable forecasts would help them to assess the forecast uncertainty and use them as tools for decision making under uncertainty. This part of the work compares in a test-bed several methods to linearly combine several wind speed ensemble forecasts operationally available at Météo-France and the same ensemble forecasts post-processed with EMOS. The combination of EMOS is intended to improve forecast performance over a single EMOS. The focus is on the combination methods rather than on EMOS, whose performance has been investigated in another thesis at Météo-France and published by Taillardat et al. (2016). Here, three means are used to improve the performance of the combination. First, training several EMOS methods over different sliding or fixed periods produces different forecasts with a large range of performance patterns. Second, using several EMOS methods and ensembles should also produce forecasts with different error patterns. Third, the combination weights are computed over a sliding period so that the combination method may adapt quickly to performance changes and weight heavily forecasts with better performances.

Due to the high variability of the atmospheric state over space and time (Slingo and Palmer 2011; Holton and Hakim 2012; Rohli et al. 2013), most (E)MOS methods require an archive of past observations and associated forecasts covering several years and a sufficient number of different meteorological situations. The NWP model code must be constant over these several training years, so that patterns in the forecast errors can be modeled by the (E)MOS methods. The code of NWP models may be updated once or twice yearly, so that small performance changes accumulate over time or large performance breakthrough are achieved. These changes may eventually render the previously built relationship between the observations and the NWP forecasts useless for, or even detrimental to, the post-processed forecasts' performance. For instance, Erickson et al. (1991) investigate the impact of changes in a NWP model on the performance of MOS for different parameters, and show that the performance of MOS of cloud cover is adversely affected. The combination scheme proposed above intends to address this issue. An alternative would be to build reforecasts, that is, forecasting the past with the new version of the model (Wilks and Hamill 2007; Hagedorn et al. 2008; Hamill et al. 2008), and to recompute the (E)MOS relationship. This is very costly, and combining forecasts may probably render it unnecessary, as hinted in Miller et al. (1992).

The post-processing is done on a grid, with the gridded wind speed measurements built in Chapter 2 as observation. The four ensembles come from the TIGGE data set (Bougeault

et al. 2010; Swinbank et al. 2016), containing several ensemble forecasts from different weather services.

1.4.2 State of the art of EMOS and combination thereof

EMOS are much studied and used in the weather forecast community, as reviewed in Gneiting (2014). The issue is addressed with parametric or nonparametric approaches. In parametric EMOS, a forecast distribution is taken from a family of probability distribution, whose parameters depend on the ensemble. For wind speed, several distributions have been tried, such as the gamma distribution (Sloughter et al. 2010; Möller and Scheuerer 2013), the truncated normal distribution for the wind speed (Thorarinsdottir and Gneiting 2010) or transformation thereof (Hemri et al. 2014) or the log-normal distribution (Baran and Lerch 2015). The parameters of the distribution are modeled as a linear regression with the forecast values or statistics thereof as inputs. This kind of approach is called non homogeneous regression (NR, Gneiting et al. 2005). The more flexible Bayesian model averaging (BMA), introduced by Raftery et al. (2005), builds a mixture of several parametric distributions, whose weights are dynamically computed in a Bayesian framework (Baran et al. 2014). As noted in Gneiting (2014), other EMOS methods use kernel functions to estimate the forecast density or to fit density functions on the ensemble, but can be interpreted in the framework of BMA. As for nonparametric EMOS methods, quantile-to-quantile transformation, introduced by Bremnes (2007), is used operationally at the Hungarian Meteorological Service (HMS) to post-process ensemble forecasts (Ihász et al. 2010). This method establishes a bijection between quantiles of the same order in the climatology of the observation and in the climatology of the forecast. Each forecast value is then transformed to the associated observed value thanks to this bijection. Several variations exist to compute the required forecast climatology, such as using a rolling period as in Flowerdew (2012), or running the ensemble NPW model over past years as is done at the HMS. Hamill and Whitaker (2006) advocate the use of analog methods as an EMOS methods, by forecasting observations from past days whose ensemble forecast are much alike as the current ensemble forecast. In Hamill and Colucci (1998), the rank histogram is used to estimate the forecast distribution. The rank histogram is just the histogram of the rank of the observation in each forecast/observation pair. The proportion of observations of each possible rank in a training sample is used to associate an order to the future sorted forecast values. Between two consecutive sorted values, the distribution is supposed uniform. Extrapolating the probability function outside the range of the forecast values may require some parametric assumption. Taillardat et al. (2016) compare NR to EMOS based on random forests for several parameters.

The combination of EMOS does not seem to be a common practice. In some ways, BMA can be considered as a combination method. Baran (2014) uses non homogeneous regression for wind speed ensemble forecasts with a truncated normal and a log-normal distribution

then chooses among them according to the value of the ensemble mean. Baran and Lerch (2016) use as their final forecast a weighted average of the same two previous NR EMOS. The combination weight and the parameters of the two combined distributions are fitted by optimizing the CRPS or the likelihood. Baudin (2015) combines the pooled and sorted values of several ensembles with combination methods that have theoretical guarantees that the combination cannot perform much worse than some reference forecast. The used combination methods are an adaptation to probabilistic forecasts of methods to combine point forecasts (Cesa-Bianchi et al. 2006; Mallet et al. 2007). The weights are computed with functions of the CRPS as a measure of performance.

1.4.3 Results

The choice of the EMOS methods used in this part of the work is motivated as follows. BMA has been discarded due to its long computing time, incompatible with operational constraints. The quantile-to-quantile transformation and the rank histograms, tested in previous internal studies, did not seem to be efficient enough. The two retained EMOS methods are nonhomogeneous regression and quantile random forest (Meinshausen 2006). QRF was chosen because it performs well for wind speed, as proven in a previous study (Taillardat et al. 2016). These good performances stem from the nonparametric nature of QRF, which lets the method fit the wind speed distribution more freely than NR. Since QRF requires a long training period, it is combined with NR, which is less flexible but can be trained on short periods of time thanks to its few parameters. Following Hemri et al. (2014), who advocated it as the best choice for wind speed, a truncated normal distribution for the square-root of wind speed is chosen for NR. Several versions are trained with different sizes of sliding training window to build distributions that evolve more or less rapidly. This may allow the combination to adapt its weighting to more or less quiescent weather regimes.

Since QRF allows to produce only step-wise cumulative distribution functions (CDF), the different versions of NR are also discretized as step-wise CDF so that the combination can be achieved. This is much alike the work of Baudin (2015), who combines in a similar way the *sorted* forecast values of several pooled ensembles. This means that his individual probabilistic forecasts are step-wise CDF with one step. The individual forecasts from Baudin (2015) are not identifiable over time as required by the philosophy behind the theory of prediction with expert advice used in this work and ours. For instance, the lowest forecast value at different times can come from a different member of a different ensemble. Being a whole (post-processed or raw) ensemble, each of our combined forecasts is a truly probabilistic forecast identifiable over time. To use the combination methods, we had first to generalize some new formulae of Baudin (2015), valid for step-wise CDFs with one step, to the case of step-wise CDFs with any number of steps. Finally, with 4 raw ensembles and 6 post-processed versions of each, 28 step-wise CDFs are combined

as a convex combination to produce a valid step-wise CDF. The combination weights are computed with five methods. Some of these combination methods are empirical. Others are based on the theory of prediction with expert advice presented in this chapter, and exhibit interesting theoretical properties in terms of performance. The use of this theory for the combination of probabilistic forecasts is quite new in the field of Meteorology.

Two performance criteria are used: the CRPS averaged over all grid points and times, and the proportion of grid-points with a flat rank histogram. The flatness of the rank histogram, a necessary condition for a reliable forecast, is assessed with the Jolliffe-Primo flatness tests introduced in Jolliffe and Primo (2008). These tests assess the existence of particular kind of deviations from flatness in the rank histogram. They are yet not much used in forecast verification, despite the insight they give of the origin of forecasts issues, such as bias or lack of dispersion. These tests are used here as an automatic diagnosis, which is necessary due to the huge number of grid-points to be handled during operations. Both performance criteria do not give the same ranking of the 28 single forecasts and the combination methods. Minimizing the CRPS may be achieved with very few flat rank histograms, whereas maximizing the number of flat rank histograms goes with only a slight increase in CRPS. This result is new since the usual procedure is to first minimize the CRPS then to check whether the rank histogram can be deemed flat. It is therefore proposed to choose a combination method based on the flatness of the rank histogram as tested with the Jolliffe-Primo flatness tests. According to this criterion, the best combination method is the exponentiated weighted average (EWA) forecaster, based on the theory of prediction with expert advice. EWA produces a large proportion of flat rank histograms over France (about 85%), while getting an average CRPS similar to the one obtained by the combination method with the lowest average CRPS. This best EWA forecaster, although trained separately for each grid-point and at each lead time, exhibits some spatio-temporal correlation. Also, the combination weights can change quickly as expected, which may allow a fast adaptation of the combination to updates of the NWP ensembles. Finally, the optimal setting of the EWA forecaster's parameter can be chosen with only one year of forecast/observation pairs, a good asset to update frequently the EMOS post-processing.

1.5 Conclusion

This work investigates several aspects of weather forecast post-processing for wind speed over France. The aim is to build improved wind speed forecasts on a grid, for deterministic and probabilistic predictions. The adopted strategy is to first grid measurements, then to use these gridded measurements to train (E)MOS methods.

For deterministic forecasts, block MOS is introduced along with a careful optimization of the size and number of the associated R objects. This lightweight block MOS shows good

performance while allowing an important speeding up of operations. It will be implemented during Fall 2016 for operations.

As for the EMOS, empirical combination methods and combination methods based on the theory of prediction with expert advice are compared. Since step-wise CDF are combined, this part of the work required to study the properties of the estimators of the CRPS with limited information about the forecast distribution. This led to recommendations to accurately estimate the CRPS. Also, due to some discrepancy of the flatness of the rank histograms and the value of the CRPS in this study, it is proposed to choose among probabilistic forecasts by first imposing to have a flat histogram according to the Jolliffe-Primo tests. The best combination method chosen with this criterion obtains a similar CRPS as the one minimizing the CRPS, while exhibiting much more flat rank histograms. It is also planned to make it operational by the end of next year.

Chapter 2

Improved Gridded Wind Speed Forecasts by Statistical Post-Processing of Numerical Models with Block Regression

This chapter reproduces an article accepted in *Weather and Forecasting*, and written by Michaël Zamo (Météo-France), Liliane Bel (AgroParisTech), Olivier Mestre (Météo-France) and Joël Stein (Météo-France). This article is licensed under CC-BY.

Abstract Numerical weather forecast errors are routinely corrected through statistical post-processing by several national weather services. These statistical post-processing methods build a regression function called "model output statistics" (MOS) between observations and forecasts that is based on an archive of past forecasts and associated observations. Because of limited spatial coverage of most near-surface parameters' measurements, MOS have been historically produced only at meteorological station locations. Nevertheless, forecasters and forecast users increasingly ask for improved gridded forecasts. The present work aims at building improved hourly wind speed forecasts over the grid of a numerical weather prediction model. First, a new observational analysis, which performs better in terms of statistical scores than those operationally used at Météo-France, is described as gridded pseudo-observations. This analysis, which is obtained by using an interpolation strategy that was selected among other alternative strategies after an intercomparison study conducted internally at Météo-France, is very parsimonious since it requires only two additive components, and it requires little computation resources. Then, several scalar regression methods are built and compared, using the new analysis as ob-

servation. The most efficient MOS is based on random forests trained on blocks of nearby grid points. This method greatly improves forecasts compared to raw output of numerical weather prediction model. Furthermore, building each random forest on blocks and limiting those forests to shallow trees do not impair performances compared to unpruned and point-wise random forests. This alleviates the storage burden of the objects and speeds operations up.

Contents

2.1 Introduction 34
2.2 Gridding 10-m windspeed measurements 36
2.2.1 Methodology 37
2.2.2 Data description 37
2.2.3 Verification strategy 38
2.2.4 Results about the best interpolation strategy 39
Comparison to reference and cross-validated AROME analyses . . 40
Spatial structures of gridded measurements 43
Why a global training domain? 44
Further post-processing of TPRS interpolation 45
2.3 Improving wind speed forecasts on a grid by block regression 46
2.3.1 Data 47
2.3.2 Block MOS 49
2.3.3 Results 50
Best block MOS 50
Performance at station locations 54
Speeding up operations 56
2.4 Conclusion 58

2.1 Introduction

Numerical weather prediction (NWP) models, although essential for forecasting the dynamics of the atmosphere, are not perfect and may be consistently biased. This is particularly true near the surface (Haiden et al. 2015) because processes such as stress and surface heating are not well modeled and because model topography may not be accurate. Furthermore sources of errors, such as initial condition errors, model errors, parametrisation errors, accumulate in a very intricate way (e.g, initial condition errors are a mixture of model and

assimilation errors). These errors may not be easily or quickly corrected through improvement in the knowledge of the atmospheric behaviour or in the performance of computers or computing science. A cheap, quick and efficient means of correcting systematic errors is the so-called model output statistics (MOS, Glahn and Lowry 1972) method, which is used by many national weather services (Wilson and Vallée 2002; Baars and Mass 2005; Schmeits et al. 2005; European Center for Medium-Range Weather Forecasts 2006; Kang et al. 2011; Zamo et al. 2014a). MOS is a statistical post-processing technique consisting of building a statistical regression function between a predictand or response (what is to be predicted) and predictors or explanatory variables (what is used to make the prediction). Predictors are usually outputs of some NWP model, thus the term MOS. The chosen statistical regression function is then applied to future forecasts to improve their performance in terms of objective scores, such as the root mean squared error (RMSE) or the mean error.

The predictand in MOS is usually variable measured at meteorological stations. As a consequence, MOS is mainly applied to station locations and its performance is evaluated against measurements at those stations. However, forecast users need improved forecasts at arbitrary locations where measurements are not always available. For a national weather service, a most interesting goal is to have MOS available over the grid of some NWP model. To achieve this goal, two possible strategies are (1) to build MOS at station locations and then to grid them, as the National Oceanographic and Atmospheric Administration (NOAA) does (Glahn et al. 2009; Gilbert et al. 2009), or (2) to grid measurements and then to build MOS using this gridded field as predictand. In this study the second strategy is preferred and described.

Specifically, the aim is to build over France gridded MOS fields for hourly 10-m wind speed forecasts. Wind forecast fields have been selected due to their importance in warning systems and potential damages (damages to building roofs, fallen tower cranes, and injuries or death caused by fallen objects are just some examples). Furthermore, due to local phenomenon (e.g. slope wind, tunneling), surface wind speed is not the easiest field to interpolate or improve and, as such, it is a good candidate to test the efficiency of MOS methods. The same methodology will be applied to other fields, such as gusts and temperature. The first step is to build a new wind speed analysis and to demonstrate that it performs better in terms of statistical scores than those operational analyses at Météo-France. The necessity to use a different wind analysis comes from the insufficient availability of operational analyses (every 3 hours until April 2015), the requirement to have at least 3 years of hourly gridded wind speed to train MOS methods, and the opportunity to get more accurate analyses. The interpolation strategy described in this study has been selected among 48 strategies after an intercomparison led at Météo-France (not shown here). The 48 interpolation strategies varied with the interpolation function, the information used and the modelling of the spatial dependence. The second step is to build the best MOS using the new analysis. For that aim, two regression methods are com-

pared. Both are trained by pooling together the data at nearby grid points (or "blocks") and getting the most parsimonious regression functions while keeping the same forecast performance. By reducing the number of regression functions, this so-called "block MOS" is useful to speed up operations when using MOS over a whole country such as France.

The manuscript is organized as follow. In Section 2.2 the more efficient gridded analysis is introduced and compared against the analysis operational at Météo-France. Section 2.3 is devoted to build gridded MOS of wind forecasts using the more efficient analysis strategy as observation. Section 2.4 sums all the results up.

2.2 Gridding 10-m windspeed measurements

In order to get gridded fields of 10-m wind speed measurements even where actual measurements are not available, several interpolation strategies exist. The most straightforward one is to use as the predictand an existing analysis of some NWP model. However classical variational assimilation schemes such as 3D-Var (Courtier et al. 1991) or 4D-Var (Courtier et al. 1994) assimilate station measurements. Therefore, an objective verification of such an analysis versus those measurements is not straightforward and may lead to overconfidence in the forecasts' performances, as will be shown later. Furthermore, since assimilation schemes mix in some way forecasts and observations, the obtained analysis could be affected by the forecast bias. As presented in Schaefer and Doswell (1979), it is also possible to work on the two dimensional wind field, interpolating divergence and vorticity instead of the wind vector itself. This may allow imposing physical constraints, such as mass conservation, and working on the wind vector instead of the wind speed only. But while working on a limited domain, this solution requires boundary conditions which may not be trivial. A third efficient method to interpolate measurements is to run a very high resolution model and find a statistical relationship between measurements and short lead-time forecasts at the same (or nearby) locations. This interpolation function is built for locations where the predictand and the predictors are available and applied to points where only the predictors are known, as presented in Burlando et al. (2013). This approach typically uses NWP model with a resolution of the order of a few tens of meters. This is not feasible for a whole country as wide as France, but a good compromise could be using a model over the entirety of France with a grid size of a few kilometers. This statistical interpolation is the approach chosen here and compared to an analysis existing at Météo-France, which is a kilometer-scale analysis based on 4D-var assimilation. The methodology is presented in more detail hereafter.

2.2.1 Methodology

Let us suppose we have at our disposal past predictand and predictor values, at time $t = 1, \ldots, T$ for N^s stations located at sites s_i where $i = 1, \ldots, N^s$. Let us note $\mathcal{S}$ a fine (model) grid covering the region of interest and $\mathcal{T}$ be a fine temporal grid covering $(1; T)$. Then, for a generic spatio-temporal point (s, t), with $s \in \mathcal{S}$ and $t \in \mathcal{T}$, let us note $y(s, t)$ and $\mathbf{x}(s, t)$ the value of the predictand and the vector of predictors, respectively.

Interpolating the predictand consists in building some function f such that $y(s, t) = f(\mathbf{x}(s, t)) + \epsilon(s, t)$, with ϵ an interpolation error. The function f is built to have the best generalization capability, that is the lowest possible errors ϵ over the sites in $\mathcal{S}$. It is fitted locally, that is, for a given spatio-temporal point (s_i, t) the training set $\mathcal{D}(s_i, t)$ is made of a subset of $\{s_1, \ldots, s_{N^S}\} \times \mathcal{T}$ depending on (s_i, t).

Many interpolation strategies can be tried by varying the training set, the family of functions to which f belongs, the choice of the predictors $\mathbf{x}$, and the optional modelling of the error ϵ. The error can be supposed deterministic (Hengl 2007) with no modelling at all. Alternatively, the error can be modelled with statistical models either without spatio-temporal dependence (Hastie et al. 2009; Kuhn and Johnson 2013) or with spatial dependence explicitly modelled (Hengl 2007; Cressie and Wikle 2011).

2.2.2 Data description

The predictand is the hourly 10-m wind speed defined as the average of the instantaneous wind speed measurements taken during the ten minutes before each hour. These measurements are available at 436 meteorological stations over mainland France (named above s_i, with $i = 1, \ldots, Ns$), which are managed by Météo-France. In order to balance quantity and quality of measurements, retained data are actually measured at heights between 8 and 13 m for stations of environmental class lower than or equal to 3 according to the World Meteorological Organization's Guide to Meteorological Instruments and Methods of Observation (World Meteorological Organisation 2008, Chapter 1, Annex 1.B). For wind speed measurements, environmental class 3 requires that "the mast should be located at a distance of at least 5 times the height of surrounding obstacles" and that "sensors should be situated at a minimum distance of 10 times the width of narrow obstacles (mast, thin tree) higher than 8 m". Mean distance between couples of nearest stations is 21 km. The study period goes from January 2011 to March 2015.

For the best interpolation strategy described hereafter, the vector of predictors $\mathbf{x}$ at a site s is composed of the position of the site and the most recent wind speed forecast from an NWP model.

- Position: the position of each site $s \in \mathcal{S}$ is specified by its horizontal coordinates (s_x

and s_y) in the extended Lambert-93 georeferencing system and its altitude (s_z). The value of s_z is obtained by considering the altitude of the nearest point in the digital elevation model (DEM), called BDAlti[1], of the French geographical institute (IGN, Institut national de l'information géographique et forestière). The freely available version of this DEM, which is used in this study, has a resolution of 75 m and covers France only.

- Most recent wind speed forecast from an NWP model: the NWP model used is AROME (Applications de la Recherche à l'Opérationnel à Méso-Echelle, or mesoscale applications of research for operational use), Météo-France's high resolution NWP model. It is a limited area and non-hydrostatic model. During the study period, it had a 2.5 km grid size over France (Seity et al. 2011). For one specific site, date and time, the wind speed forecast comes from the most recent run, excluding the analysis, and it is noted $W_{AROME}(s,t)$. Since AROME runs four times per day, the used lead-times range from 1 to 6 hours. As an example, for an interpolation at 1600 UTC, the predictors come from the run of 1200 UTC with a lead-time of 4 hours. The wind speed forecast used at station locations is AROME's forecast bilinearly interpolated from AROME's grid towards these locations.

2.2.3 Verification strategy

Since no wind speed measurement is available at grid points, assessment of the interpolation strategy is achieved through cross-validated interpolation towards some test stations. Cross-validation consists in splitting the available archive into two subsets: one training set is used to fit the interpolation functions, one test set is used to assess the interpolation performance.

Since cross validation is time consuming, a subset of 150 test stations were chosen, representative of the French topography and hourly wind speed climatology. Ten lists of fifteen stations were built as test sets, so that each list gathers stations far enough from one another to ensure that results are close to those of leave-one-out cross validation. The closest test stations in each list are separated by at least 80 km. Interpolation is done toward each of these ten test lists separately and performance is assessed. Consequently, the training is always done with 421 stations (up to missing data).

Comparing this new analysis against the existing AROME analysis provides an assessment of its usefulness for operational purposes. However, AROME assimilation scheme already assimilates station measurements, which biases its scores toward better performances. Thus, in order to get an accurate assessment of the analysis performance as an interpolator, 10 AROME assimilations were rerun without assimilating one test set of 15

[1] http://professionnels.ign.fr/bdalti

stations each. Since this reanalysis takes time, it was only run for 120 dates between July 2013 and July 2014, at 1500 UTC corresponding to the maximum of the diurnal cycle of wind speed. This reanalysis is referred hereinafter as AROMEcv, since it is computed with cross-validation.

Last, until April 2015, AROME analysis was available only every 3 hours, whereas MOS is required at hourly rate. Consequently, a simple reference hourly interpolation method is built by a bilinear interpolation of AROME most recent wind speed forecast, with a lead-time of 1 to 6 hours. At some site s with geographical coordinates (s_x, s_y), bilinear interpolation takes as interpolation function $f(\mathbf{x}(s,t)) = a + bs_x + cs_y + ds_x s_y$. The parameters a, b, c and d are fitted on the 4 nearest AROME grid points from the interpolation point s. If this bilinear interpolation performs better, the retained analysis is simply the most recent wind speed AROME forecast.

For each of these analyses, the interpolation performance is assessed by pooling together the interpolated values in the 150 test stations at the 120 test dates. Classical performance measures are used such as:

- Bias:
$$BIAS = \overline{-\epsilon(s,t)}$$
- Root mean square error:
$$RMSE = \sqrt{\overline{\epsilon^2(s,t)}}$$
- Mean absolute error:
$$MAE = \overline{|\epsilon(s,t)|}$$
where $\epsilon(s,t)$ is the aforementioned interpolation error and $\bar{\cdot}$ signifies the mean over all test stations and test dates.

Since RMSE and MAE values alone do not give information about the distribution of errors, specifically about large errors, measures of error dispersion are also computed:

- Percentage of absolute errors lower than or equal to w, with $w = 1$ or $4\ \mathrm{m\,s^{-1}}$, noted $\%_{\leq 1}$ and $\%_{\leq 4}$, respectively.
- Quantile of order τ of absolute errors with $\tau = 0.5$ (median) or 0.9, noted $Q(0.5)$ and $Q(0.9)$, respectively.

2.2.4 Results about the best interpolation strategy

The best interpolation strategy among the 48 interpolation strategies previously tested is presented.

First, the training set $\mathcal{D}(s,t)$ is global and for a fixed time. This means that whatever the interpolation point (s,t) is, the training domain pools all the stations over France but it takes into account only the measurements at time t.

Second, the interpolation function is a mixture of two thin plate regression splines (TPRS, Wood 2003). This is a special case of Generalized additive models (GAM, Wood 2006). In GAM the actual predictand is some link function g of the expectation of y, taken as the sum of p functions: $g(\mathbb{E}(y|\mathbf{x})) = \sum_{j=1}^{p} f_j(x_j)$, with x_j one or several components of the predictors vector. Here the link function is the identity, and the functions f_i are two TPRS. Indeed, our best interpolation function is simply $f(\mathbf{x}(s,t;\mathcal{D}(s,t))) = tps(W_{AROME}(s,t)) + tps'(s_x, s_y, s_z)$, where tps and tps' are two TPRS, whose parameters are fitted for each date and time in an automatic way by means of the function *gam* in the R package *mgcv* (R Core Team 2015).

Third, the spatial dependence between the errors is not explicitly modelled in this strategy. It appears to be unnecessary since using AROME wind speed forecast implicitly imposes some structure to the interpolated field.

Unless otherwise stated, the following results are computed for the 150 test stations, the 120 test dates and at 1500 UTC.

Comparison to reference and cross-validated AROME analyses

The two first columns of Table 2.1 present the measures of performance for the TPRS analysis and the reference. For the whole sample, both analyses are unbiased. However, TPRS performs better than bilinear interpolation for the other measures of performance. The RMSE is improved by 16% and most of the errors are less than 4 $\mathrm{m\,s^{-1}}$in absolute value.

This table also shows the same measures of performance but for classes defined by the terciles of the wind speed distribution over France during the study period: weak (below 2.9 $\mathrm{m\,s^{-1}}$), average (between 2.9 $\mathrm{m\,s^{-1}}$and 4.8 $\mathrm{m\,s^{-1}}$) and strong (above 4.8 $\mathrm{m\,s^{-1}}$). For the lowest measured wind speeds, TPRS and reference tend to yield slightly too strong wind (positive bias) and the converse for the strongest measured wind speeds (negative bias). However, the bias remains low. Whatever the wind speed regime, TPRS outperforms bilinear interpolation whatever other performance measure is considered.

Figures 2.1 and 2.2 show the evolution of RMSE and BIAS over time of the day for TPRS and reference analyses, computed over the 150 test stations and all the dates in the study period. The curves may show abrupt changes every 6 hours, when the predictors are taken from a different run. This is due to the better performance of the underlying forecast thanks to the proximity of AROME assimilation. Anyway, TPRS is consistently better

Table 2.1 – Measures of performance for TPRS, bilinear reference interpolation (ref.), operational AROME analysis (AROME) and AROME reanalysis computed with cross-validation (AROMEcv). These figures concern 150 test stations and 120 dates at 1500 UTC, for all wind speed values and three different intervals of wind speed measurements. Bold figures indicate best performances among TPRS, reference and AROMEcv.

	TPRS	ref.	AROME	AROMEcv
All windspeed values				
BIAS	**0.0**	0.3	-0.1	**0.0**
MAE	**1.0**	1.2	0.6	1.1
RMSE	**1.4**	1.6	0.8	1.5
Q(0.5)	**0.8**	0.9	0.4	**0.8**
Q(0.9)	**2.1**	2.5	1.2	2.3
$\%_{\leq 1}$	**63.1**	53.8	86.3	58.3
$\%_{\leq 4}$	**98.8**	97.6	99.6	98.3
Weak wind (below 2.9 $\mathrm{m\,s^{-1}}$)				
BIAS	0.7	0.8	0.1	**0.5**
MAE	**0.9**	1.1	0.5	**0.9**
RMSE	**1.2**	1.5	0.7	1.3
Q(0.5)	**0.7**	0.9	0.3	**0.7**
Q(0.9)	**2.0**	2.4	1.0	**2.0**
$\%_{\leq 1}$	**66.9**	56.2	90.1	64.9
$\%_{\leq 4}$	**99.4**	97.7	99.8	98.9
Medium wind (between 2.9 $\mathrm{m\,s^{-1}}$and 4.8 $\mathrm{m\,s^{-1}}$)				
BIAS	0.1	0.4	-0.1	**0.0**
MAE	**0.8**	1.1	0.5	0.9
RMSE	**1.1**	1.4	0.7	1.2
Q(0.5)	**0.7**	0.8	0.3	**0.7**
Q(0.9)	**1.7**	2.2	1.0	2.0
$\%_{\leq 1}$	**70.6**	56.8	89.3	63.1
$\%_{\leq 4}$	**99.8**	99.1	100.0	99.5
Strong wind (above 4.8 $\mathrm{m\,s^{-1}}$)				
BIAS	-0.7	**-0.3**	-0.4	-0.7
MAE	**1.3**	**1.3**	0.7	1.4
RMSE	**1.7**	1.8	1.1	1.8
Q(0.5)	**1.0**	**1.0**	0.5	1.1
Q(0.9)	**2.7**	2.8	1.6	2.9
$\%_{\leq 1}$	**51.9**	48.8	79.5	46.8
$\%_{\leq 4}$	**97.4**	96.1	99.0	96.5

than the reference. And its performance shows less variability. This is also true for other performance measures (not shown here).

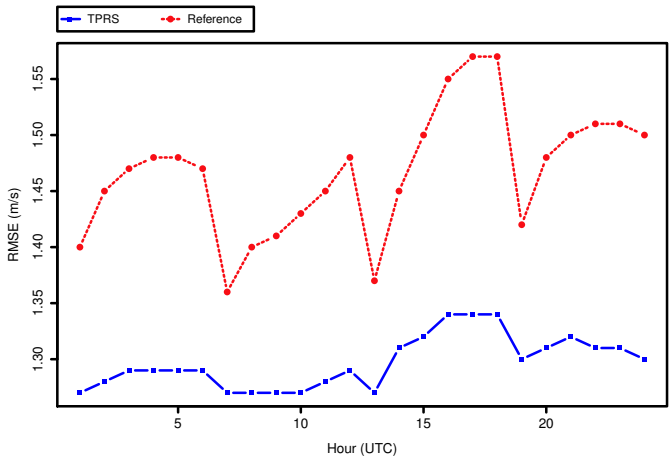

Figure 2.1 – Evolution of RMSE over time for TPRS and reference interpolation strategies, computed over the 150 test stations.

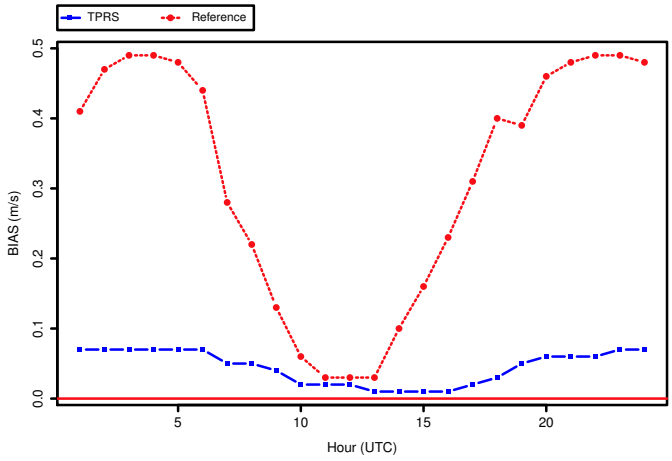

Figure 2.2 – Same as in Figure 2.1, but for the bias.

Table 2.1 also shows the performance measures of the operational AROME analysis with all stations assimilated and of the $AROME_{cv}$ reanalysis. As an example of the usefulness of this cross-validated reanalysis for assessing the performance of the TPRS analysis, let us note that without blacklisting some stations the operational AROME analysis gets an RMSE of about 0.8 $\mathrm{m\,s^{-1}}$over the test stations, a significantly better score compared to

the actual cross-validated RMSE of 1.5 $\mathrm{m\,s^{-1}}$(87.5% higher). This shows the strong local impact of the observations in the assimilation fields.

As for the new analysis, it appears that actually TPRS performs better than the $AROME_{cv}$ reanalysis, whatever the interval of measured wind speeds and the performance measure. Moreover, TPRS is computed very quickly: the complete hourly interpolated wind speed grid from January 2011 to March 2015 required only four days of computation at the resolution of AROME (2.5 km) on a standard workstation. This may allow a real-time computation of wind speed analysis to be used routinely, and gives a long enough archive to train MOS methods.

Spatial structures of gridded measurements

The used performance measures quantify the quality of interpolation strategies but say nothing about the likeliness of structures represented in the gridded wind field. Figure 2.3 allows a subjective evaluation of these structures. It concerns the storm Joachim that hit Western Europe in December 2011.

First, TPRS may increase or decrease wind speed compared to AROME forecast. As an example, in Figure 2.3 gridded wind speeds with TPRS are lower than forecasts in the south-west of France but more variable and stronger on the Pyrénées. The wind speed in the new analysis is also increased at the tip of Brittany and decreased on a large area to the East and South-East of Brittany. This high-impact event has been appraised by meteorologists thanks to Météo-France's internal reports of this event. The structures in TPRS have been judged more in agreement with the reality.

Moreover, gridded wind speeds, although usually smoother than AROME because of the use of smooth functions such as TPRS, still exhibit realistic physical structures. This may not be systematic for every interpolation strategy. Indeed, as an example, ordinary kriging led to unrealistic smooth wind speed fields (not shown). In Figure 2.3, the wind speed field is more variable on the Pyrénées for TPRS than for AROME forecast. Because AROME only includes a 2.5 km resolution topography whereas the new analysis includes the 75 m resolution BDAlti topography, this increased spatial variability of gridded wind speed over the mountains seems to be a positive feature.

Similar results hold for other dates and hours that have been subjectively appraised by Météo-France (not shown).

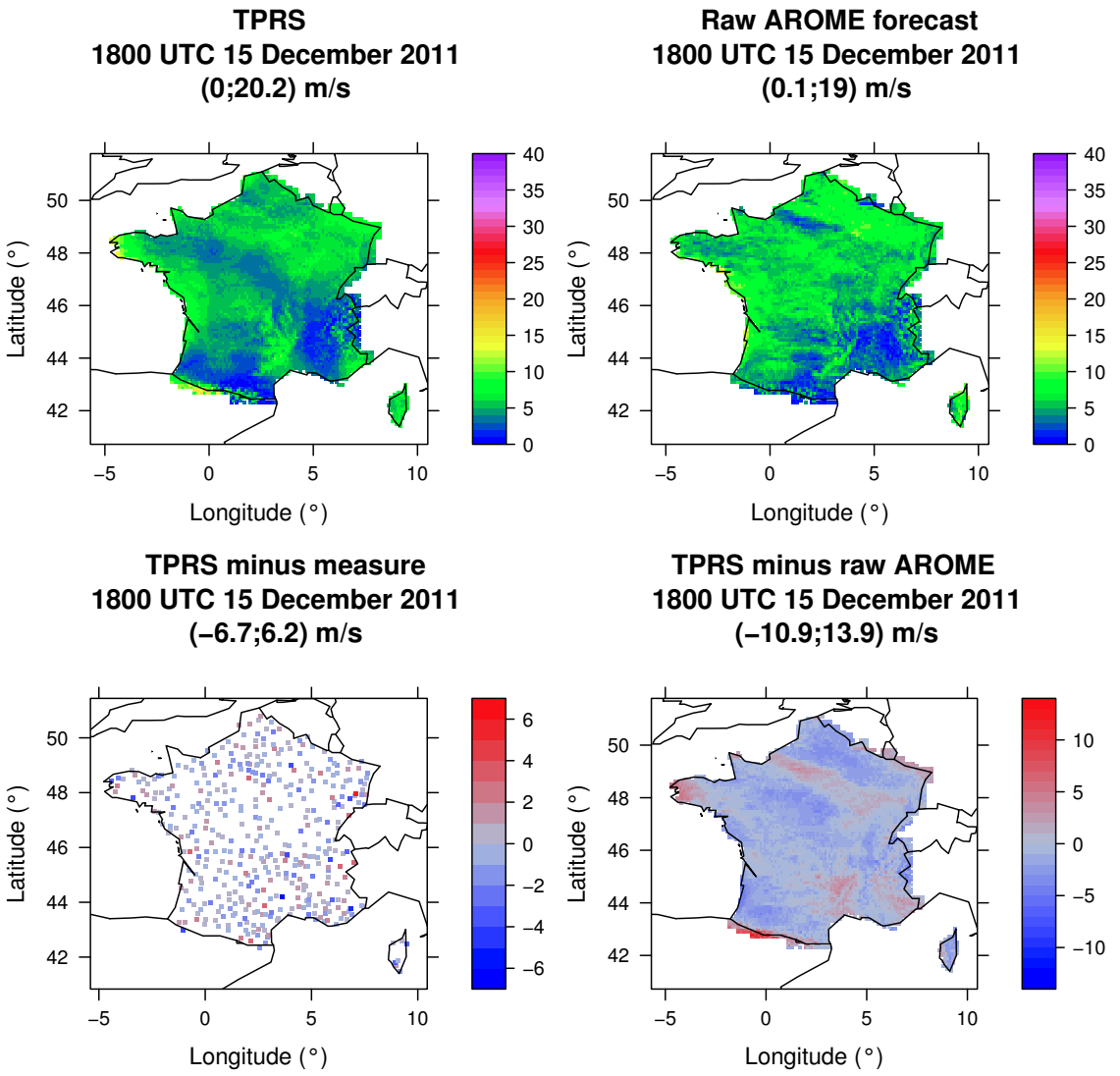

Figure 2.3 – Results of interpolation at 1800 UTC 15 December 2011. Top left: map of gridded wind speeds with TPRS. Top right: map of AROME wind speed forecast (run: 1200 UTC, lead time: +6 h). Bottom left: residuals of interpolation at station locations. Bottom right: differences between TPRS and AROME forecast. Under each map title is the interval of the corresponding quantity.

Why a global training domain?

A local training domain, containing only stations within a certain radius around each site s, was used for a sensitivity experiment. This training radius was varied between 20 km and

2000 km. Indeed, a variographic study (not shown here) showed that the correlation length of wind speed measurements is about 50 km, albeit with large differences according to the kind of meteorological and geographical zone (warm sector, tail end of a low, neighborhood of a front, mountains, slopy areas, etc.). It could be expected that, with smaller training domains, the wind speed measurements would be more correlated and the performances improved. But it turns out that these local training domains gave worse performance than a global training domain. It happens that the smaller the training domain, the less numerous the data and the less precise the estimation of the interpolation function, whence the worse interpolation performance (not shown). Inversely, by taking a global training domain, the interpolation method takes the best of all available data at one specific time. To improve performance by reducing the size of the training domain would require a much denser measurement network. In hilly or mountainous areas, with very local topographic effects, this requirement would become unrealistic.

Further post-processing of TPRS interpolation

By construction, TPRS linearly extrapolates as soon as there is a predictor exceeding the values in the training data. Due to this linear extrapolation, interpolated wind speeds may reach unrealistic values. Contrary to other tried interpolation strategies, TPRS nearly never exhibits such excessive wind speeds. In order to filter out and prevent these rare occurrences, a post-processing of the gridded wind speed fields illustrated before is applied (see example in Figure 2.4). The meteorological spline in TPRS, $tps(W_{AROME}(s,t))$, is constrained at each grid point to be less than $tps(max_{training}(W_{AROME}))$, where the maximum AROME forecast in the training data set is noted $max_{training}(W_{AROME})$. Since this filtering rarely changes the gridded measurements, performance measures of TPRS are not modified.

Finally, a visual comparison of measured values and TPRS interpolation at the station locations showed that for 12 stations, although the new analysis gets better performance than AROME analysis, very high errors (up to 80% below the measured value) remain. These stations are situated in hilly areas and exhibit very high values. These features make it very unlikely to get a good interpolation at these locations. To keep high wind speeds in the new analysis, measurements at all stations are simply copied out to the nearest grid point.

To conclude this section, TPRS is a quick and more efficient alternative to the usual data assimilation scheme to create a long archive of hourly gridded wind speed measurements. It also requires less computational resources and it can be run on a standard workstation.

The following section describes how to improve wind speed NWP forecasts using MOS with the new analysis based on TPRS as predictand (or response).

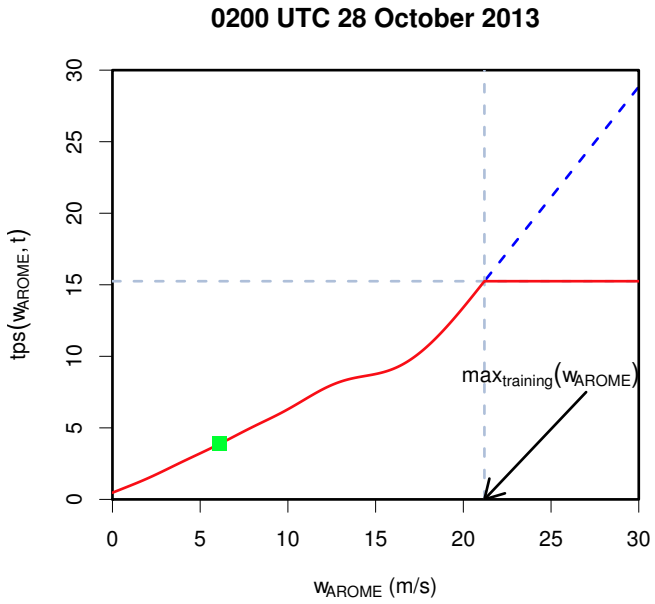

Figure 2.4 – Post-processing of gridded wind speed to prevent excessive extrapolation in TPRS. W_{AROME} is AROME wind speed forecast. The blue dashed line is the original meteorological spline component at the chosen grid point. The red continuous line is the post-processed meteorological spline. The green square is the actual gridded wind speed at the chosen grid point (unchanged in this case).

2.3 Improving wind speed forecasts on a grid by block regression

MOS aims at correcting forecasts by means of a regression function r between the variable Y to be predicted and some explanatory variable(s) (or predictors) X that can be NWP model output(s) or any other source of information. This regression function is estimated on an archive of past forecasts and associated observations and it is then applied to future forecasts to increase their accuracy. It is much similar to what has been done in the previous section when building an interpolation function, provided that the regression function is applied at future times ($t > T$) instead at non-monitored locations ($s \notin \{s_i, \quad i = 1, \ldots, N^s\}$).

In this study, two classical regression methods, namely linear model and random forest, are compared. The functional kernel regression (Ferraty and Vieu 2006; Ferraty et al. 2012) was also tested but results are not presented since this method is largely outperformed by

Table 2.2 – List of ARPEGE's explanatory variables, available for wind speed regression.

Abbreviations	Description
ffH10, ddH10f	10 m wind speed and discretised direction (North, South, East and West)
lat, lon, elevation	Latitude, longitude and elevation
month	Month as a qualitative variable with 12 categories
capeins	Convective available potential energy
nc, nt, nb, nm, nh	Nebulosities (c: convective, t: total, b: low altitude clouds, m: medium altitude clouds, h: high altitude clouds)
SLP_Adv, SLP_Trend	Advection and 3 h trend of sea level pressure
tpwHPA850, tpwHPA850_Adv, tpwHPA850_HVar	Potential wet-bulb air temperature at 850 hPa, and its advection (Adv) and horizontal variance (HVar)
tH_PCi, $i = 1 \dots 3$	First three components of a principal component analysis of temperature vertical profile (up to 1500 m)
ffH_PC, $i = 1 \dots 3$	First three components of a principal component analysis of wind speed vertical profile (up to 1500 m)

the two classical regression methods (not shown).

2.3.1 Data

The explanatory variables X come from Météo-France global NWP model ARPEGE (Action de Recherche Petite Échelle Grande Échelle or Small Scale Large Scale Research Project, Courtier et al. 1991). ARPEGE is a stretched-grid, hydrostatic NWP model, with an horizontal grid size of 0.1° (about 10 km) over France. It runs every six hours with hourly lead-times up to 60 or 102 hours depending on the run. Table 2.2 lists the 24 explanatory variables selected for building regression functions of the analyzed wind speed on forecasts. SLP_Adv, SLP_Trend, tpwHPA850 and tpwHPA850_Adv have been chosen as proxies of the synoptic dynamics of the atmosphere. capeins, tH_PCs, ffH_PCs and tpwHPA850_HVar aim at quantifying the instability of the boundary layer.

The response Y is ARPEGE wind speed forecast errors relative to the new TPRS wind speed analyses post-processed presented in Section 2.2.4). Several attempts showed that the performances were slightly improved when predicting the wind speed error instead of the wind speed itself. Performances are computed for the corrected wind speed forecasts. The regression function is build only on ARPEGE grid points and not for all AROME grid points due to computation time constraints for operational purposes and because of

ARPEGE's larger lead time range. Since the new analysis is available on the AROME 2.5 km grid, block MOS for AROME at its full resolution is planned for future applications and should probably accelerate operations a lot.

The study period covers three years, from 1 September 2011 to 31 August 2014.

Due to long computation times, regression methods have been trained only over ten spatial domains noted D01 to D10 (see Figure 2.5). Each domain contains a grid of 9x9 ARPEGE grid points (about 90x90 km^2). These domains have been chosen so that they represent a large range of conditions of winds and topography over France. Domains D06, D09, D07 and D08 cover increasingly ragged topography. Domains D01, D02 and D03 can be subject to strong local winds, namely Marin, Cers for the first two domains, Mistral for the third one.

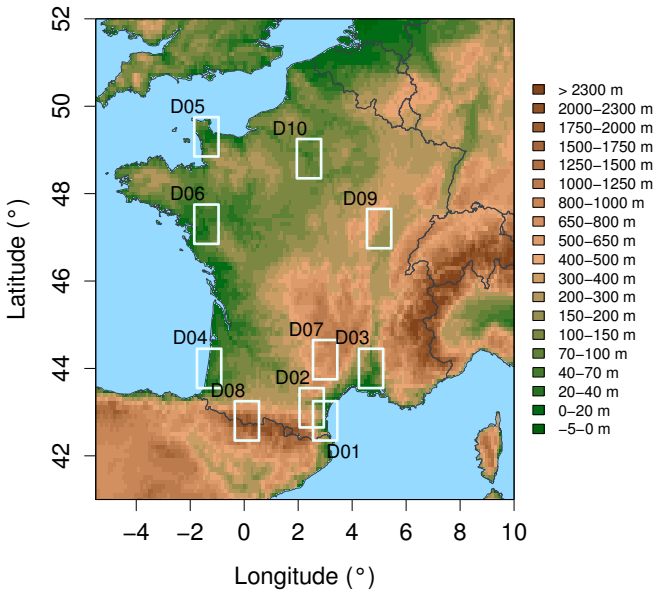

Figure 2.5 – The ten training domains used in this study.

Each regression method is trained separately for lead times 3 h, 15 h and 48 h of ARPEGE run of 0000 UTC. The lead times have been chosen to cover short and long lead times and, for 15 h, as representative of the hours in the day with usually the strongest winds.

2.3.2 Block MOS

The following regression methods (Hastie et al. 2009; Kuhn and Johnson 2013) are tested:

- Linear model (Azaïs and Bardet 2006; Weisberg and Fox 2010): the regression function is a second-order polynomial relationship of the explanatory variables: $\hat{r}(X) = \beta_0 + \boldsymbol{\beta} \cdot X^{1,2}$ where β_0 is a real, $\boldsymbol{\beta}$ is a vector of reals and $X^{1,2}$ is the vector containing every possible combination of product of explanatory variables of order 1 and 2 (called interactions). The parameters β_0 and $\boldsymbol{\beta}$ are fitted on the training data with an ascending selection of predictors based on the Bayesian information criterion (BIC, Schwarz 1978; Lebarbier and Mary-Huard 2006).

- Random forest (Breiman 2001): this is an average of several regression trees (Breiman et al. 1984). For a single regression tree, the regression function is built through an iterative splitting of available training data into two subsets. Splitting is done according to some threshold of a quantitative explanatory variable or some subset of modalities of a qualitative explanatory variable. The best split is chosen so that the two subsets of response values are the most homogeneous inside each subset and the most dissimilar between one another. The (dis)similarity criterion is the intra- or between- group variance. Splitting is stopped for some criterion, such as a maximum number of groups, called leaves. The predicted value is then the average of the response values in the leaf. A regression tree has usually a low bias but strongly depends on the training data.

  In random forest, each tree is similar to a regression tree but with two further randomizations. The first randomization is to start each tree from a bootstrapped sample of the training data (Diaconis and Efron 1983). Then each split, or node, of each tree is built from a random subset of the available explanatory variables. The final predicted value is the average of all leaves reached by the value of the vector of explanatory variables. This double randomization makes the trees of the forest more independent and thus decreases the variance of the errors without increasing the bias of each tree. The regression function $\hat{r}$ is an average of block-wise constant functions over the space of the explanatory variables. In this study, the fitted parameters are the number of trees in the forest and the number of tried predictors at each node.

In Statistics, more data sometimes imply better inference. Therefore, in the aim of further improving MOS performance, block regression is used. This means that inside each domain, the regression methods described above are trained by pooling data across several grid points. The area containing these pooled grid points is called a block. Consequently, one regression function $\hat{r}$ is built for a block and applied to all the grid points inside the block. However, the position of each grid point is available as a predictor through its latitude, longitude and elevation. If these predictors are selected during the training step, the regression function may actually depend on the grid point location. Another advantage

expected from block regression is to have fewer models which may speed up operations.

The size of the block is varied to assess its impact on the MOS performance. The sizes are 1x1 (or point-wise training), 3x3, 5x5, 7x7 and 9x9 grid points. The blocks of size 3x3, 5x5 and 7x7 contain the central square with respectively 9, 25 and 49 grid points of the domain. If MOS performs better for non point-wise block, it is planned to map France with contiguous blocks for using MOS in operations.

In order to assess forecast performances, the same measures of performance as in Section 2.2.3. are computed. In order to compare on the same data the training with different block sizes, performance measures are computed for the central 3x3 grid points in each domain. The so-called skill scores are also used: if S_A, S_B and S_∞ are the measures of performance for forecasts A, B and a perfect forecast, the associated skill score is $SS_{A/B} = \frac{S_A - S_B}{S_\infty - S_B} \in (-\infty; 1]$. For $RMSE$, MAE, $Q(0.5)$ and $Q(0.9)$, $S_\infty = 0$ whereas for $\%_{\leq 1}$ and $\%_{\leq 4}$, $S_\infty = 100$. A positive skill score implies the forecast A gets better performance than forecast B.

Furthermore, the variability of the performances are assessed thanks to 3-fold cross-validation: two years serve as training data, the remaining one being used as a test sample. All three possible combinations of two training years/one test year are tried.

2.3.3 Results

Best block MOS

Figure 2.6 presents the RMSE of raw ARPEGE forecasts and MOS forecasts built with the two regression methods and different combinations of parameters. The scores are computed for the three test years and with a training domain of 1x1, 3x3, 5x5, 7x7 and 9x9 grid points. The figure is for domain D03 and lead time 15 h.

Whatever the chosen training settings, both MOS methods improve performance over raw ARPEGE forecasts.

Performances of random forests are sensitive to the number of trees and number of tried predictors at each node. For a given number of trees and tried predictors, performance are slightly decreased by increasing the block size, but this effect is marginal. The best tuning is therefore to take about 6 to 8 tried predictors and at least 50 trees trained in 3x3 blocks. This optimal setting remains true for other domains and lead times (not shown here). In order to speed up operations, shallower trees may be used if this does not reduce forecast performance. With default settings, the complete random forests have 2,300 leaves for each tree, for a 3x3 grid points training domain. Constraining trees to a maximum number of leaves have been tested. Best performances are achieved by random forests even with no

more than 200 leaves as shown in Figure 2.7 for domain D03 and lead time 15 h. However, for some rare other domains and lead times, the minimum optimal number of leaves may be around 500 (not shown). To sum up, the best random forest MOS is obtained by building 200 trees with 8 tried predictors at each node and 500 leaves.

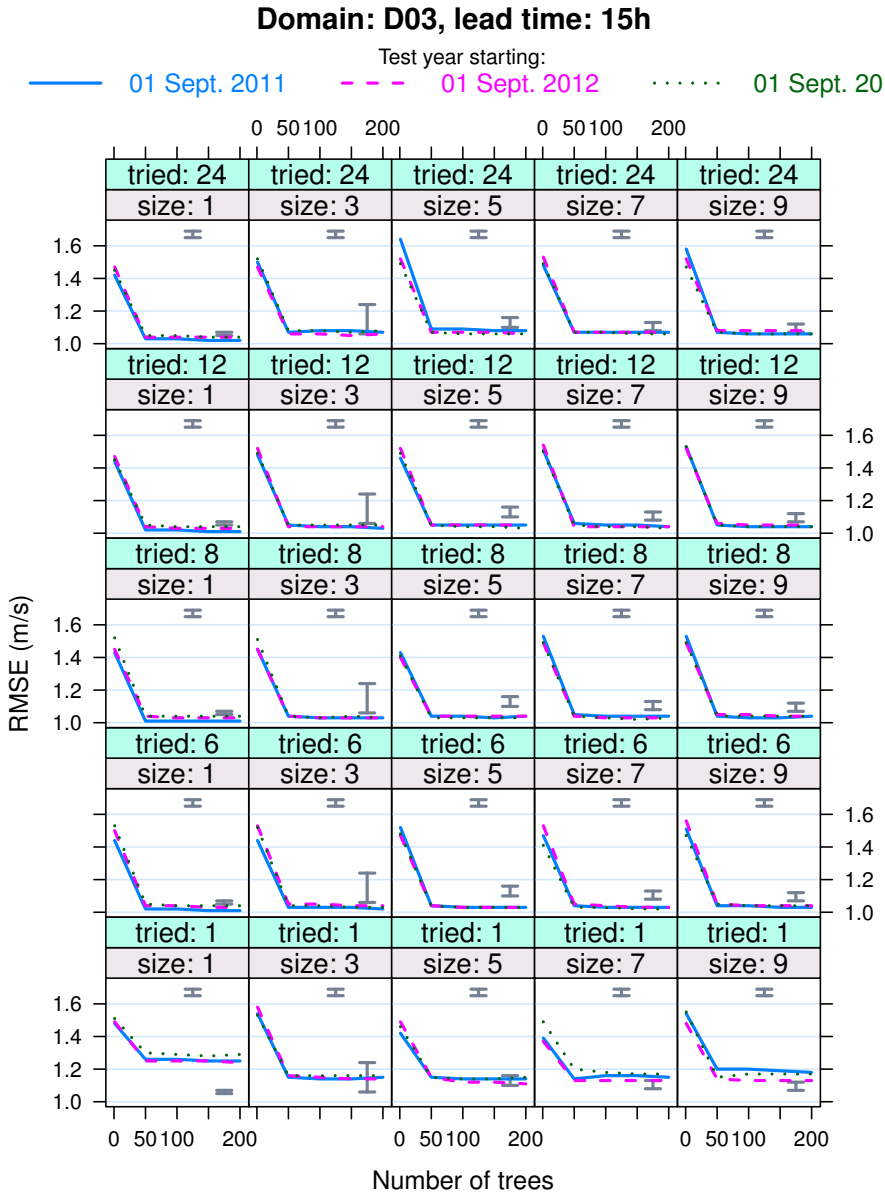

Figure 2.6 – RMSE for several MOS methods and settings, along with raw ARPEGE forecasts. In each panel, lines show the evolution of performance of random forest with the number of trees, for a specific training block size and number of tried predictors at each node and the three test years. In each panel, vertical bars indicate the interval of variation over the three test years of ARPEGE performance (left bars) and block MOS with a linear regression (right bar). For linear model and random forest, the training domain can be of size 1x1, 3x3, 5x5, 7x7 and 9x9 grid points from left to right. Scores are computed over 1 year of test (starting on 1 September), over the 3x3 central points of domain D03 and for lead time 15 h.

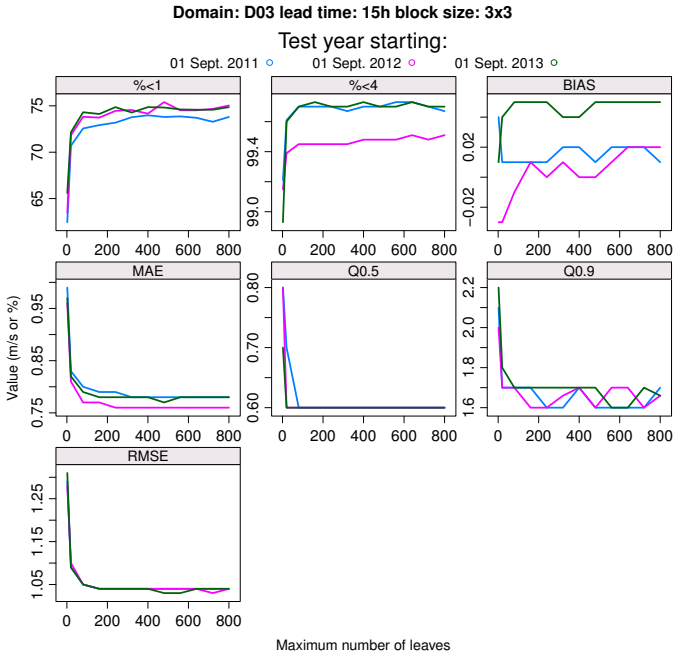

Figure 2.7 – Variation of measures of performance for random forests, with varying number of maximum allowed nodes. Random forests are built with 200 trees and 8 tried predictors at each node. This figure is for domain D03, lead time 3 h and a 3x3 training block.

As for the linear regression MOS specifically, Figure 2.6 shows that its performance varies a lot with the training block size. However, the best performance is achieved with a point-wise training in this case, and also whatever the domain, lead time or performance measure (not shown).

On Figure 2.6, the best linear model (trained point wise) and the best random forests apparently get similar performances. By showing skill scores for random forests (model A) versus the point-wise trained linear MOS (model B), Figure 2.8 shows that forecast performances are improved by several percents with random forests. The only exception is for the percentage of absolute errors lower than 4 $\mathrm{m\,s^{-1}}$($\%_{\leq 4}$) where performances may be decreased when using random forest compared to using linear regression. This figure also confirms that random forests trained on a 3x3 block get similar performances than point-wise random forests. Thus, even though training random forests on blocks does not improve forecast performance as could be hoped, it does not decrease it either. Skill scores computed for other domains and/or lead times confirm that random forest is usually a better choice than linear regression by a few percent, except for $\%_{\leq 4}$ where the best MOS is not always the same (see, e.g., Figure 2.8). In conclusion, the best MOS method is random forest with 6 to 8 tried predictors, 200 trees, 500 leaves and a 3x3 block training.

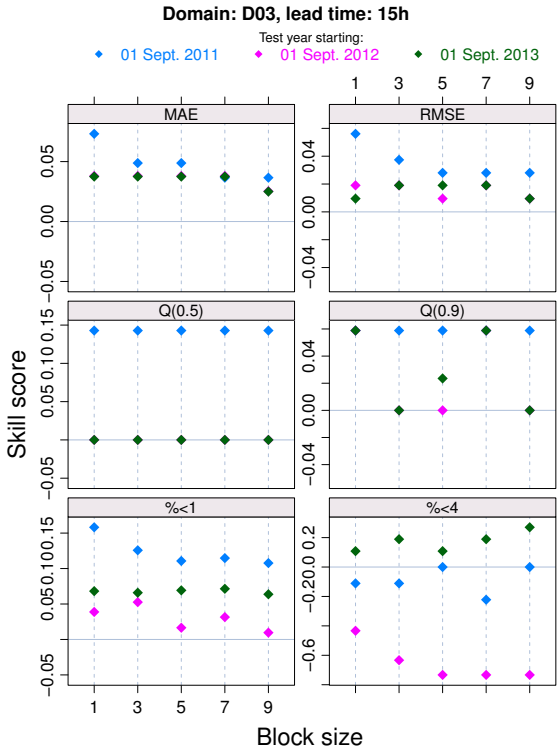

Figure 2.8 – Evolution of skill scores with the size of training block for random forest MOS, with point-wise trained linear MOS as a reference. 200 trees and 8 tried predictors are used to train the random forest with several sizes of training block.

On a more qualitative side, MOS successfully corrects the tendency of the raw model to overestimate wind speed, as illustrated in Figure 2.9 with a smoothed scatter plot. As can be seen, the MOS scatter plot is much more concentrated along the first bisecting line than the raw forecast. This line corresponds to the point set of perfect forecasts. This improvement is obvious whatever the strength of the gridded wind speed. These results hold for every other domain or lead time (not shown).

Performance at station locations

Table 2.3 shows performance measures of forecasts bilinearly interpolated at the locations of the meteorological stations inside the ten training domains. Scores are computed relative to measurements at those stations by pooling forecasts for all three test years and for all three lead times. Only 5 stations were included in any of the ten training domains. For the raw AROME forecasts, lead time 48 h is actually lead time 6 h for the run of 1800 UTC,

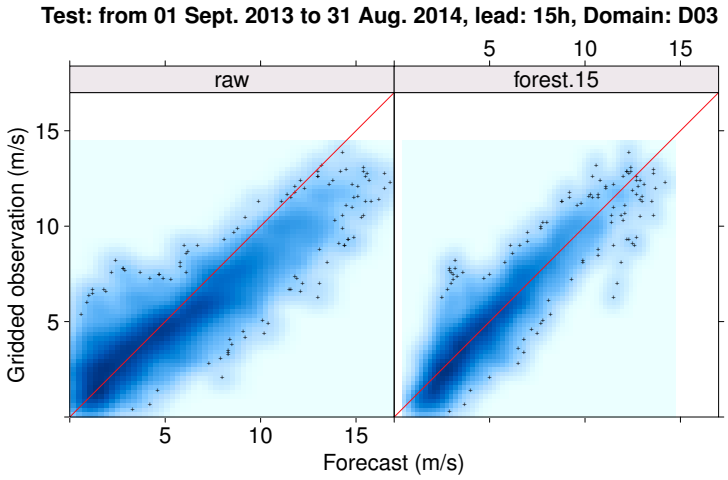

Figure 2.9 – Smoothed scatter plot of gridded observation against raw forecast (left) or random forest based MOS forecast (right). The darker the blue, the denser the points. The red oblique line is the first bisecting line. These scatter plots are for test grid-points in domain D03, the test year starting on 1 September 2013, lead time 15 h and a training on 3x3 grid points. The random forest is built with 200 trees and 8 input variables randomly drawn at each node.

since AROME does not yield forecasts beyond lead time 36 h. However, valid dates are the same for MOS at lead times 48 h and this lagged raw AROME.

The scores show that random forest gets better overall performance than interpolated raw ARPEGE forecasts, for a training domain of point-wise or 3x3 blocks. Furthermore, random forests get similar or better performance than interpolated forecasts from Météo-France's high resolution model AROME. Concerning the bias, whereas random forests have a negative bias and AROME is unbiased, the bias of random forests remains low (only -0.3 $\mathrm{m\,s^{-1}}$). For lead times 48 h, MOS are as good as raw AROME at a lead time of 6 h, an improved anticipation of 42 h.

However, the results vary at the scale of single stations. Table 2.4 shows the scores obtained for a station picked at random for different lead times. For this station situated in domain D03, random forests achieve much better performance than ARPEGE or AROME for a lead time of 3 h. At a lead time of 15 h, spatially interpolated random forests still get the upper hand over raw AROME but differences are slightly reduced. At a lead time of 48 h (6 h for raw AROME), random forests and AROME yield similar results. Over the 5 stations in the training domains, results are that variable, even though usually random forests get at least as good performances as interpolated AROME forecasts. Whatever,

Table 2.3 – Measures of overall performance of bilinearly interpolated forecasts at station locations. The forecasts are MOS based on random forest, with a training domain of size 1x1 or 3x3, raw ARPEGE forecast and raw AROME forecast. Scores are computed by pooling together forecasts over the three test years, every stations inside any study domain and the three lead times (3, 15 and 48 h). For AROME, lead-time 48 h is actually lead-time 6 h, since AROME forecasts do not extend up to 48 h. Bold scores indicate best performance.

	Random forest		ARPEGE	AROME
	1x1	3x3		
BIAS	-0.3	-0.3	0.2	**0.0**
MAE	**1.2**	1.3	1.7	1.4
RMSE	**1.7**	1.8	2.3	1.9
Q(0.5)	**0.9**	1.0	1.2	1.0
Q(0.9)	**2.7**	2.8	3.9	3.1
$\%_{\leq 1}$	**54.0**	52.8	45.1	50.8
$\%_{\leq 4}$	**96.5**	96.1	90.7	94.9

ARPEGE never prevails. Since the sample is small (only 5 stations), further investigations would be necessary to assess the best choice of interpolated forecasts. This will first require to build MOS forecasts for all the chosen grids over whole France. This will be done for future applications at Météo-France, but such a training will require weeks. Nevertheless, those first results point at interpolating MOS forecasts trained on a 3x3 blocks as a good solution to get improved forecasts at station locations.

Speeding up operations

Running MOS on a grid with thousands of grid points may be time consuming, at the training stage and during operations. One purpose of block regression is to build less regression models to accelerate memory loading during operations. Indeed, since the prediction with random forest is very quick, a limiting factor for operational purpose is the loading time of models in memory. For the linear models, only the regression coefficients β have to be saved on disk, with a disk occupation of a few kB. A random forest object can be much bigger if not optimized. In our case a random forest trained on one grid point amounts to 2 MB (for a total of 18 MB for a 9 grid point domain), whereas a random forest trained on a 3x3 domain requires 12 MB, one third less. Additionally, a shallow forest with 200 trees, 8 tried predictors and only 500 leaves, trained over 3x3 grid points requires only about 5 MB for each domain, a further reduction of 60 %. Removing components of R random forest objects unnecessary for prediction leads to a final storage size of 1.7 MB on disk.

Table 2.4 – Same as in Table 2.3, but for one station in domain D03 and for each lead time.

	Random forest		ARPEGE	AROME
	1x1	3x3		
Lead time: 3 h				
	Lead time: 3 h			
BIAS	**0.2**	**0.2**	1.5	0.9
MAE	**1.0**	**1.0**	1.7	1.3
RMSE	**1.3**	**1.3**	2.1	1.6
Q(0.5)	**0.9**	**0.9**	1.5	1.0
Q(0.9)	**2.0**	**2.0**	3.4	2.6
$\%_{\leq 1}$	60.1	**60.4**	33.8	48.2
$\%_{\leq 4}$	**99.5**	**99.5**	95.3	99.0
	Lead time: 15 h			
BIAS	**-0.1**	0.2	0.3	**-0.1**
MAE	**1.2**	**1.2**	1.4	**1.2**
RMSE	**1.5**	**1.5**	1.8	1.6
Q(0.5)	**1.0**	**1.0**	1.1	**1.0**
Q(0.9)	**2.3**	**2.3**	2.9	2.5
$\%_{\leq 1}$	51.5	**52.2**	45.1	49.3
$\%_{\leq 4}$	98.8	**99.1**	96.3	97.4
	Lead time: 48 h			(6 h)
BIAS	**0.2**	**0.2**	1.4	0.8
MAE	**1.2**	**1.2**	1.8	**1.2**
RMSE	**1.5**	**1.5**	2.2	**1.5**
Q(0.5)	**1.0**	**1.0**	1.5	1.1
Q(0.9)	**2.4**	**2.4**	3.6	2.5
$\%_{\leq 1}$	49.8	**50.1**	35.5	47.6
$\%_{\leq 4}$	98.5	98.5	92.7	**99.0**

In order to compare loading times for several MOS models as stored in R, the above objects have been loaded from disk 300 times for the ten studied domains. Figure 2.10 shows that the linear model objects load much more quickly (about 15 ms for the ten domains) than point-wise trained random forest objects (a few seconds accumulated over the ten domains). However, combining block regression, shallow trees and removal of unnecessary components allows dividing loading times of random forests by a factor 10. Since the complete mapping of France requires about 830 domains, the loading time would be about half a minute for the whole country. This still makes random forest longer to load than linear models, but it is compatible with operational constraints. As seen above, this acceleration is achieved without reducing the overall forecast performance.

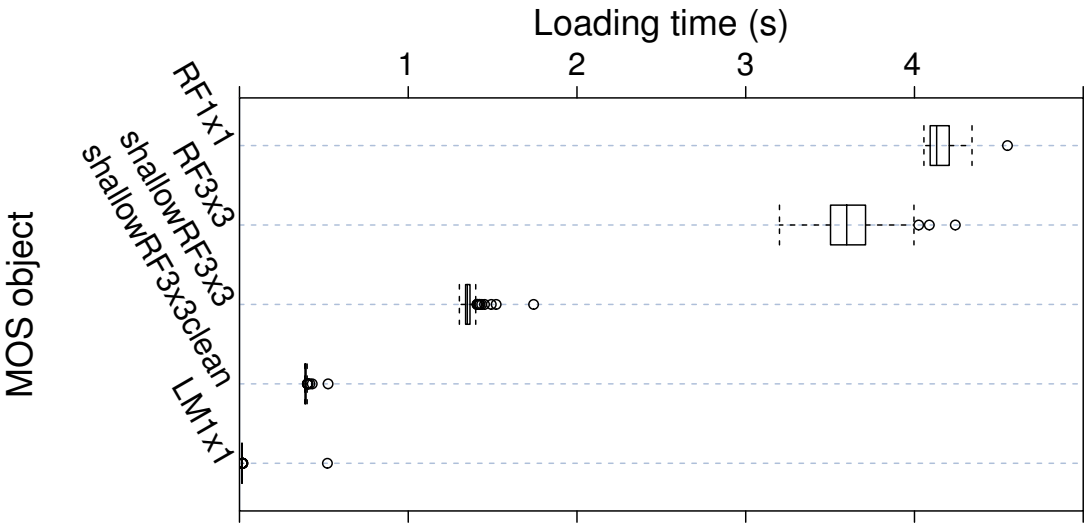

Figure 2.10 – Box plots of 300 loading times for the whole R objects over the 10 training domains for several MOS models: complete random forest trained point wise (RF1x1), complete random forest trained over a 3x3 block (RF3x3), shallow random forest trained over a 3x3 block (shallowRF3x3), shallow random forest trained over a 3x3 block and with removal of unnecessary elements for prediction in objects (shallowRF3x3clean), and point-wise linear model (LM1x1).

2.4 Conclusion

Accurate wind speed forecasts are crucial for decision making in weather-related activities or for weather warnings by national and regional weather services. NWP models provide forecasts that are not exempt of errors. Since these errors are not completely random, statistical post-processing methods, known as MOS, can be used to improve future forecasts by using regression functions fitted on past forecasts and associated observations. In order to apply those methods to wind speed forecasts at grid point locations, a new gridded analysis of wind speed measured at meteorological stations is built. An internal comparison of 48 interpolation strategies led at Météo-France showed the best hourly analysis is based on thin plate regression splines. This regression is very parsimonious with only two additive components: a first one with the most recent wind speed forecast of the high resolution model AROME as the only input and the second one with a correction based on the 3-dimensional coordinates of the points. By cross-validation, it is shown that this new analysis performs consistently better than available AROME analysis while keeping realistic structures of wind speed fields thanks to the use of AROME forecast in the interpolation

function. This allows to build an archive of gridded wind speed over France with a 2.5 km grid size starting in January 2011 and ending in March 2015.

This new analysis is used to build improved wind speed forecasts of Météo-France 10 km NWP model, ARPEGE, over France. The use of classical regression methods shows that ARPEGE forecasts are easily and greatly improved by all regression methods. The best MOS is based on random forests. The best combination of parameters for this model is shown to be not very sensitive: taking more than 200 trees with 6 to 8 tried predictors at each node is sufficient. Furthermore, random forests can be trained by pooling together data from nearby grid points without degrading performances. Also, the trees in the optimal random forests need not be very deep in order to achieve the best performances. These last remarks lead to building less numerous and shallower random forests. After removing unnecessary components in R random forest objects, the storage resources and loading times of the random forests is reduced by a factor 10. The time to produce MOS forecasts is mainly determined by the loading time of all the random forests into memory. Thanks to their reduced size and number, this operation can be done in a reasonable timing (about half a minute) that enables its application in every day operations. By Fall 2016, this MOS method with random forests trained over blocks will be made operational at Météo-France by covering France with contiguous blocks. Block MOS for other parameters of interest, such as gusts, will also be made operational.

Acknowledgments The authors are grateful to two anonymous reviewers whose comments contributed to improve the readability of this article.

Chapter 3

Estimation of the Continuous Ranked Probability Score with Limited Information - Applications to Ensemble Weather Forecasts

This work was done under the precious guidance of Philippe Naveau.

Abstract The continuous ranked probability score (CRPS) is a much used measure of performance for probabilistic forecasts of a scalar observation. It is a quadratic measure of the difference between the forecast cumulative distribution function (CDF) and the empirical CDF of the observation. Analytic formulations of the CRPS can be derived for most of the classical parametric distributions, and be used to assess the efficiency of different CRPS estimators. When the true forecast CDF is not fully known but represented as an ensemble of values, the CRPS is estimated with some error. Thus, using the CRPS to compare parametric probabilistic forecasts with ensemble forecasts may be misleading due to the unknown error of the estimated CRPS for the ensemble. With simulated data, the impact of the type of the verified ensemble (a random sample or a set of quantiles) on the CRPS estimation is studied. Based on these simulations, recommendations are issued to choose the most accurate CRPS estimator according to the type of ensemble. The interest of these recommendations is illustrated with real ensemble weather forecasts. Also, relationships between several estimators of the CRPS are demonstrated and used to explain the differences of accuracy between the estimators.

Contents

3.1 Introduction 62
3.2 Review of available estimators of the CRPS 65
3.3 Study with simulated data 66
3.3.1 CRPS estimation with a random ensemble 67
Methodology 67
Results 67
3.3.2 CRPS estimation with an ensemble of quantiles 70
Methodology 70
Results 71
Influence of ties in an ensemble of quantiles 74
3.4 Real data examples 77
3.4.1 Raw and calibrated ensemble forecast data sets 80
3.4.2 Issues estimating the CRPS of real data 80
3.4.3 Issues on the choice between QRF and NR 83
3.5 Conclusion and discussion 84
Appendix 86
3.A What is elicited when the 1-norm CRPS of an ensemble is minimized? 86
3.B Relationships between the estimators of the CRPS 87
3.B.1 Equality of $\widehat{crps}_{Fair}$ and $\widehat{crps}_{PWM}$ 87
3.B.2 Equality of $\widehat{crps}_{NRG}$ and $\widehat{crps}_{INT}$ 88
3.B.3 Relationship between $\widehat{crps}_{PWM}$ and $\widehat{crps}_{NRG}$ 88

3.1 Introduction

Verifying the quality of forecasts expressed in a probabilistic form requires specific graphical or numerical tools (Jolliffe and Stephenson 2011), among them some numerical measures of performance such as the Brier score (Brier 1950), the Kullback-Leibler divergence (Weijs et al. 2010) and many others (Winkler et al. 1996; Gneiting and Raftery 2007). When the probabilistic forecast is a cumulative distribution function (CDF) and the observation is a scalar, the continuous ranked probability score (CRPS) is often used as a quantitative

measure of performance. Classically (Matheson and Winkler 1976; Hersbach 2000), the instantaneous CRPS is defined as the quadratic measure of discrepancy between the forecast CDF, noted F, and $\mathbb{1}(x \geq y)$, the empirical CDF of the scalar observation y,

$$crps(F, y) = \int_{\mathbb{R}} \left[F(x) - \mathbb{1}(x \geq y)\right]^2 dx, \tag{INT}$$

where $\mathbb{1}$ is the indicator function.

Analytic formulations of $crps(F, y)$ can be derived for most of the classical parametric distributions, some of which are listed in Table 3.1. In some situations, the forecast CDF may not be fully known, such as for ensemble numerical weather prediction (NWP) or other kinds of Monte Carlo simulations, or the forecast CDF may be known but an analytic formulation of the CRPS may not be derivable. In the latter case, one may be able to sample values from F. Anyway, in these two situations, the forecast CDF is summarized with a set of M values $x_{i=1,\dots,M}$. Following the convention in Meteorology, such a set will be called here an "ensemble" and each value x_i will be called a "member". The instantaneous CRPS must then be estimated with this ensemble. This may be problematic when using the CRPS to compare parametric forecasts, whose CRPS may be computed exactly, and forecasts whose CRPS is estimated based on the limited information about F contained in the ensemble. The unknown error in the CRPS estimation may lead to the wrong choice of the best forecast.

Usually, the instantaneous CRPS is averaged in space and/or time, over several pairs of forecast/observation. Candille (2003) and Ferro et al. (2008) showed that, when the ensemble is a random sample from F, the usual estimator of the instantaneous CRPS based on (INT), introduced later, is biased: its expectation over an infinite number of forecast/observation pairs does not give the right theoretical value. This bias stems from the limited information about F contained in an ensemble with finite size M. Several solutions have been proposed to remove this bias. Ferro (2014) introduced the notion of fair score and a formula to correct the bias in the estimation of the averaged CRPS. Müller et al. (2005) proposed two solutions to the same problem of biased estimation of the ranked probability score (RPS), the version of the CRPS for ordinal random variables. Adapted to the CRPS, their first solution would be to use an absolute value instead of a square inside the integral (INT). As demonstrated in Appendix 3.A, this score for an ensemble is minimized if all the members x_i equal the median of F, which is obviously not the purpose of an ensemble. Their second solution is to compute the RPS skill score against some ensemble of size M whose RPS is estimated by bootstrapping past observations. Although interesting, this solution does not allow to assess the absolute performance of the ensemble but only the performance relative to this bootstrapped ensemble.

This study aims at improving heuristically the estimation of the average CRPS of a forecast CDF under limited information. The information is limited in two ways: (1) the CDF is

Table 3.1 – List of distributions whose closed-form CRPS exists and were used in this study. The reference of the original article where to find the formula is also given. Taillardat et al. (2016) gathers the closed form expression of the CRPS for these and other distributions.

Distribution	Original reference
Beta: $Y \sim Beta(\alpha, \beta)$	Taillardat et al. (2016)
Gamma: $Y \sim Gamma(\alpha, \beta)$	Möller and Scheuerer (2013)
Gaussian mixture: $Y \sim \sum_{i=1}^{p} \omega_i \mathcal{N}(\mu_i, \sigma_i)$, with $\sum_{i=1}^{p} \omega_i = 1, \quad \omega_{i=1,\dots,p} > 0$	Grimit et al. (2006)
Generalized Extreme Value: $Y \sim GEV(\mu, \sigma, \xi)$	Friederichs and Thorarinsdottir (2012)
Generalized Pareto: $Y \sim GPD(\mu, \sigma, \xi)$	Friederichs and Thorarinsdottir (2012)
Log-normal: $\ln(Y) \sim \mathcal{N}(\mu, \sigma)$	Baran and Lerch (2015)
Normal: $Y \sim \mathcal{N}(\mu, \sigma)$	Gneiting et al. (2005)
Square-root truncated normal: $\sqrt{Y} \sim \mathcal{N}^0(\mu, \sigma)$	Hemri et al. (2014)
Truncated normal: $Y \sim \mathcal{N}^0(\mu, \sigma)$	Thorarinsdottir and Gneiting (2010)

known only through an ensemble as defined above; and (2) the average CRPS is computed over a finite number of forecast/observation pairs. The problem is not to estimate the unknown forecast CDF F, but to estimate the CRPS of F under limited information about F. To improve the estimation with this limited information, the usual strategy is to correct the empirical mean score, as in Ferro (2014) or Müller et al. (2005). Here the approach is to improve the estimation of each term of the average, that is, the estimation of the instantaneous CRPS $crps(F, y)$.

The rest of this paper is organized as follows. Section 3.2 reviews several estimators of the instantaneous CRPS proposed in the literature, and demonstrates relationships among them. In particular, it is shown that the four proposed estimators reduce to two only. In Section 3.3, synthetic data are used to study the variations in accuracy of these two CRPS estimators, with the size M of the ensemble and the way this ensemble is built. These simulations lead to recommendations on the best estimation of the CRPS. Section 3.4 illustrates issues in CRPS estimation with two real meteorological data sets. Improvements in the inference obtained by following the recommendations from Section 3.3 are shown on these data. Section 3.5 gives a summary of the recommendations to get an accurate estimation of the instantaneous CRPS, concludes and discusses the results.

3.2 Review of available estimators of the CRPS

he instantaneous CRPS is defined as a quadratic discrepancy measure between the forecast CDF and the empirical CDF of the observation,

$$crps(F, y) = \int_{\mathbb{R}} [F(x) - \mathbb{1}(x \geq y)]^2 \, dx. \tag{INT}$$

Equation (INT) is called the integral form of the CRPS.

Gneiting and Raftery (2007) showed that, for forecast CDFs with a finite first moment, the CRPS can be written as

$$crps(F, y) = \mathbb{E}_X|X - y| - \frac{1}{2}\mathbb{E}_{X,X'}|X - X'|, \tag{NRG}$$

where X and X' are two independent random variables distributed according to F, and $\mathbb{E}_A$ is the expectation according to the law of the random variable(s) A. This form is called the energy form of the CRPS, since it is just the one-dimensional case of the energy score introduced by Gneiting and Raftery (2007), based on the energy distance of Székely and Rizzo (2013).

Taillardat et al. (2016) introduced a third expression of the CRPS, valid for continuous forecast CDFs,

$$crps(F, y) = \mathbb{E}_X|X - y| + \mathbb{E}_X X - 2\mathbb{E}_X XF(X), \tag{PWM}$$

which is called the probability weighted moment (PWM) form of the CRPS because its third term is a probability weighted moment (Greenwood et al. 1979; Rasmussen 2001; Furrer and Naveau 2007).

When F is known only through an M-ensemble $x_{i=1,\dots,M}$, the above definitions lead to the following estimators of the instantaneous CRPS,

$$\widehat{crps}_{INT}(M, y) = \int_{\mathbb{R}} \left[\frac{1}{M} \sum_{i=1}^{M} \mathbb{1}(x \geq x_i) - \mathbb{1}(x \geq y) \right]^2 dx, \tag{eINT}$$

$$\widehat{crps}_{NRG}(M, y) = \frac{1}{M} \sum_{i=1}^{M} |x_i - y| - \frac{1}{2M^2} \sum_{i,j=1}^{M} |x_i - x_j|, \tag{eNRG}$$

$$\widehat{crps}_{PWM}(M, y) = \frac{1}{M} \sum_{i=1}^{M} |x_i - y| + \hat{\beta}_0 - 2\hat{\beta}_1, \tag{ePWM}$$

respectively, where $\mathbb{E}_X X$ is estimated by $\hat{\beta}_0 = \frac{1}{M}\sum_{i=1}^{M} x_i$, and $\mathbb{E}_X XF(X)$ is estimated by $\hat{\beta}_1 = \frac{1}{M(M-1)}\sum_{i=1}^{M}(i-1)x_i$. Without loss of generality, the members x_i are supposed sorted in increasing order, and the size M of the ensemble is supposed greater than two.

Candille (2003) and Ferro et al. (2008) showed that the expectation of (eINT) over an infinite number of forecast/observation pairs is biased with, under conditions of stationarity of the observation and the ensemble, and exchangeability of the members,

$$\mathbb{E}_Y \widehat{crps}_{INT}(M, Y) = \mathbb{E}_Y crps(F, Y) + \frac{1}{M} \mathbb{E}_{X_1, X_2} \frac{|X_1 - X_2|}{2}, \tag{3.1}$$

where X_1 and X_2 are any two distinct members of one ensemble forecast. This relation holds only when the ensemble is a random sample from F. Ferro (2014) proposed the notion of fair score for an ensemble of random values, which leads to a fourth estimator of the instantaneous CRPS, the fair CRPS defined as

$$\widehat{crps}_{Fair}(M, y) = \frac{1}{M} \sum_{i=1}^{M} |x_i - y| - \hat{\lambda}_2, \tag{eFAIR}$$

where $\hat{\lambda}_2 = \frac{1}{2M(M-1)} \sum_{i,j=1}^{M} |x_i - x_j|$ estimates $\mathbb{E}_{X_1, X_2} \frac{|X_1 - X_2|}{2}$, and is unbiased when the members are independently sampled from F.

These four estimators reduce to only two since, as shown in Appendix 3.B,

$$\widehat{crps}_{INT}(M, y) = \widehat{crps}_{NRG}(M, y),$$
$$\widehat{crps}_{PWM}(M, y) = \widehat{crps}_{Fair}(M, y).$$

The properties of only two estimators have to be studied. In the light of the second equality, the fair CRPS can be interpreted as a PWM-based estimator of the instantaneous CRPS, which explains why it is an unbiased estimator of the average CRPS of a random ensemble as proven by Ferro (2014). Indeed, the unbiasedness property of the mean for the first term and of the PWMs for the second term, in the case of a random sample, immediately proves that the two terms in Equation (ePWM) are unbiased estimators of their population counterpart, if the members are randomly and independently drawn from F.

Moreover, the relationship

$$\widehat{crps}_{INT}(M, y) = \widehat{crps}_{PWM}(M, y) + \frac{\hat{\lambda}_2}{M} \tag{3.2}$$

holds for these two estimators, as shown in Appendix 3.B. Equation (3.2) holds for a single forecast/observation pair, and requires no assumption on the nature or statistical properties of the ensemble.

3.3 Study with simulated data

The accuracy of the two instantaneous CRPS estimators presented above, $\widehat{crps}_{PWM}(M, y)$ and $\widehat{crps}_{INT}(M, y)$, is studied with synthetic forecast/observation pairs. The forecast CDF

F is chosen such that the theoretical CRPS $crps(F, y)$ can be exactly computed with a closed-form expression (see Table 3.1 for a list of such distributions). To mimic actual situations when F is not fully known, two types of ensembles are built from this forecast CDF. The two types of ensembles successively used in the remaining of this section are random ensembles and ensembles of quantiles, defined later. The estimators are then computed and compared to the theoretical value.

3.3.1 CRPS estimation with a random ensemble

Methodology

A random ensemble is a sample of M independent draws from F. In actual applications, a random ensemble may be viewed as M members from an NWP ensemble model, or, more generally, as an M-sample from Monte Carlo simulations. Protocol 1 describes the simulation plan.

Protocol 1: Estimation of the CRPS with simulated random ensembles.

Input: M: number of members.
F: forecast CDF.
G: CDF of the observation.
N: number of ensemble forecast/observation pairs.

Output: N values of instantaneous CRPS for each estimator.

1 **for** $n \leftarrow 1$ ***to*** N **do**
2 Draw the observation y from G.
3 Compute the theoretical CRPS $crps_{th}(F, y)$ with its closed-form expression.
4 Draw $x_{i=1,\dots,M}$ from F.
5 Compute and store $\widehat{crps}_{INT}(M, y)$ and $\widehat{crps}_{PWM}(M, y)$ with this ensemble.

Results

The results are presented for a standard normal forecast CDF F. For the sake of simplicity the CDF of the observation is also standard normal ($G = F$).

Since the ensemble is random, the estimated CRPS is also a random variable that depends on the observation y and the members $x_{i=1,\dots,M}$. In order to study the variability of the estimated CRPS with the ensemble only, the observation is first held constant (with a value of -0.0841427, for each n in Protocol 1), while $N = 1000$ ensembles of M members are drawn from F. The impact of M on the accuracy of the estimated CRPS is assessed by observing Protocol 1 with different ensemble sizes M.

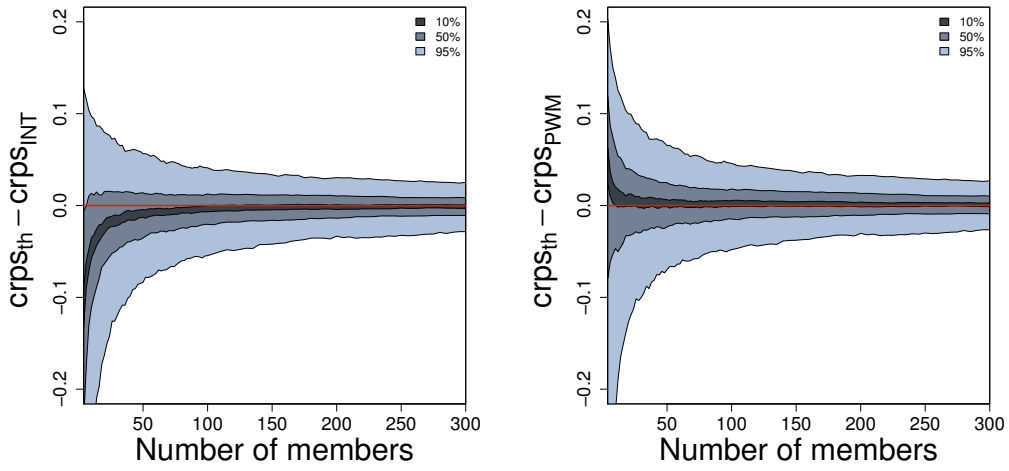

Figure 3.1 – Intervals of estimation error of $\widehat{crps}_{INT}$ (left) or $\widehat{crps}_{PWM}$ (right) for a random ensemble of varying size. Intervals are computed point-wise, with the 1000 CRPS of independently built random ensembles with the same observation. The observation and members come from a standard normal distribution.

The point-wise 10 %, 50 % and 95 % intervals of the estimation error $crps_{th} - \widehat{crps}$ (with $crps_{th} = 0.2365178$ here) are computed over these 1000 ensembles for each ensemble size M. The intervals contain the corresponding proportion of the 1000 computed CRPS errors for a given ensemble size. As shown in Figure 3.1 (left) for $\widehat{crps}_{INT}$, the error tends toward 0 when the ensemble size increases. However, important errors (as high as ± 10 % of $crps_{th}$) can still occur even for very large ensembles of several hundreds of members. As shown in Figure 3.1 (right), the estimator $\widehat{crps}_{PWM}$ exhibits a similar behaviour for large random ensembles, as deduced from Equation 3.2 if $M \to \infty$. But $\widehat{crps}_{PWM}$ becomes unbiased for much smaller ensemble sizes than $\widehat{crps}_{INT}$. The unbiasedness of $\widehat{crps}_{PWM}$ proven by Ferro (2014) holds only for ensembles with more than about 20 members. The variability of the estimation, as quantified by the half-width of the 50% central interval, may be important when the random ensemble contains less than 50 members (more than 10% of $crps_{th}$, in Figure 3.2). With increasing ensemble sizes, the variability of this estimation does not scale linearly with the number of members, as shown in Figure 3.2. Tripling the ensemble size from $M = 100$ to about $M = 300$ decreases the half-width of the 50% central interval of the relative estimation error by only about 2% (from 7% to 4%).

Common practice is to average instantaneous CRPSs over several locations and/or times. Here, this is mimicked by taking the average of N instantaneous CRPSs generated according to Protocol 1, without holding the observation constant any more. The number of forecast/observation pairs N is varied from 1 to 1000. The size M of the ensemble is

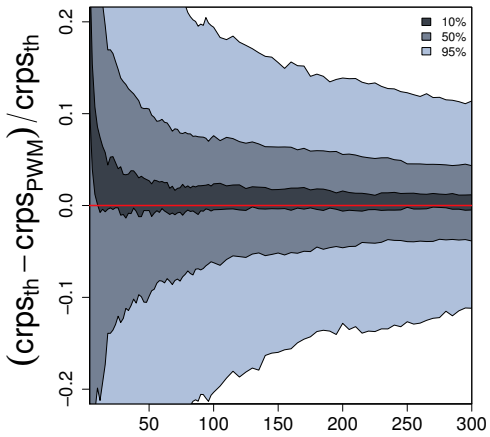

Figure 3.2 – Same as in Figure 3.1, but for the relative estimation error of $\widehat{crps}_{PWM}$.

also varied, with 10, 30, 50, 100 and 300 members. The average theoretical CRPS and average estimation are computed for each combination of N and M. As shown in the left of Figure 3.3 for $\widehat{crps}_{INT}$, a stable estimation of the average CRPS is reached if the number of averaged estimations is large enough (more than 300 for a random ensemble of 10 members). But a large ensemble is required to get an accurate estimation of the true average CRPS. As shown in the right of Figure 3.3, the averaged $\widehat{crps}_{PWM}$ shows a better estimate than the averaged $\widehat{crps}_{INT}$, even for small ensembles and small numbers of averaged estimations.

These behaviours for the instantaneous and the averaged estimates remain true for every distribution listed in Table 3.1, every parameters' value and even if the G and F are different (not shown).

The added value of these simulations to the results of Ferro (2014) is to show the behaviour of $\widehat{crps}_{PWM}$ for small ensemble sizes M and finite numbers of forecast/observation pairs. The poor scaling of this estimator's variability with the ensemble size has been empirically shown, which had never been done, to the best of our knowledge. Finding a formula for the variability of $\widehat{crps}_{PWM}$ would be interesting to quantify the estimation uncertainty for practical purposes. We demonstrated error bounds that were not usable in practice since they require to know the forecast distribution (not shown).

The conclusion of these simulations is that, for a random ensemble, the estimation of the instantaneous CRPS is not very accurate whatever the estimator is used, but the averaged CRPS can be estimated with a good accuracy. The unbiasedness of $\widehat{crps}_{PWM}$ for random ensembles stems from the use of estimators that are unbiased for independent samples

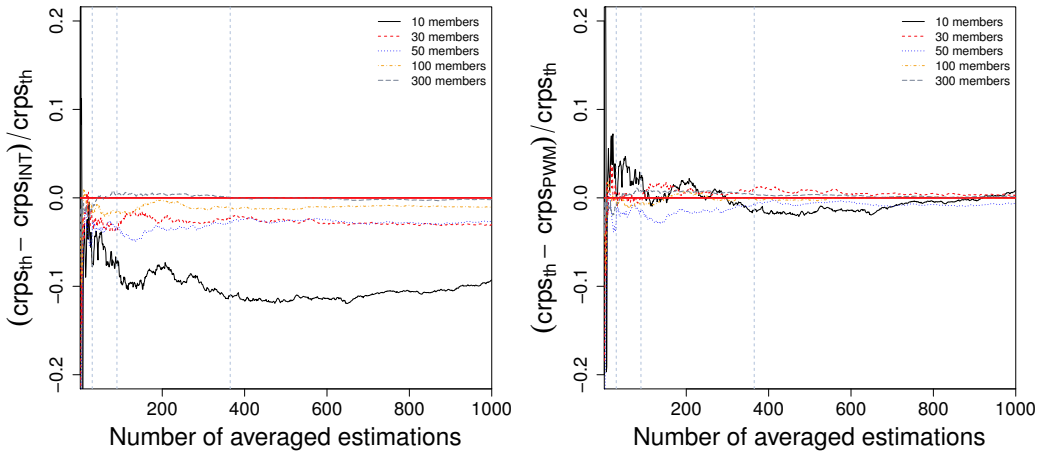

Figure 3.3 – Evolution of the relative estimation error of the averaged $\widehat{crps}_{INT}$ (left) or $\widehat{crps}_{PWM}$ (right) with the number of members for a random ensemble. The averaged CRPS is an arithmetic mean of the CRPS of several pairs of ensemble/observation among 1000. The vertical grey dashed lines correspond to an average computed with 30, 90 and 365 ensembles (to mimic a monthly, seasonal or yearly average CRPS).

from the underlying distribution F. In practice, if one seeks to estimate the potential performance of an ensemble with an infinite number of members, one should use the PWM estimator of the CRPS. The integral estimator of the CRPS assesses the global performance of the actual ensemble, and should be used for actual performance verification.

3.3.2 CRPS estimation with an ensemble of quantiles

Methodology

An ensemble of M quantiles of orders $\tau_{i=1,\dots,M} \in [0;1]$ is a set of M values $x_{i=1,\dots,M}$ such that: $x_i = F^{-1}(\tau_i) \quad \forall i \in \{1,\dots,M\}$. Contrasting with a random ensemble, the orders τ_i associated to the members x_i are known.

In this case, the data are simulated according to Protocol 2. The two built ensembles of quantiles are defined as:

- *regular* ensemble (reg): it is the ensemble of the M quantiles of orders τ_i, with $\tau_i \in \{\frac{1}{M}, \frac{2}{M}, \dots, \frac{M-1}{M}, \frac{M-0.1}{M}\}$ of F. The last order is not 1 to prevent infinite values.
- *optimal* ensemble (opt): it is the set of M quantiles of orders $\tau_i \in \{\frac{0.5}{M}, \frac{1.5}{M}, \dots, \frac{M-0.5}{M}\}$

of F. This ensemble is called "optimal" because Bröcker (2012) showed that this set of quantiles minimizes the expectation of the CRPS of an ensemble over an infinite number of forecast/observation pairs, when using Equation (eINT).

Protocol 2: Estimation of the CRPS with simulated ensembles of quantiles.

Input: M: number of quantiles.
F: forecast CDF.
G: CDF of the observation.
N: number of ensemble forecast/observation pairs.

Output: N values of the instantaneous CRPS for each estimator and kind of quantile ensemble.

Compute the ensemble of M regular quantiles of F.
Compute the ensemble of M optimal quantiles of F.
for $n \leftarrow 1$ ***to*** N **do**
    Draw y from G.
    Compute and store the theoretical CRPS $crps_{th}(F, y)$ with this observation.
    Compute and store $\widehat{crps}_{INT}(M, y)$ and $\widehat{crps}_{PWM}(M, y)$ with this observation for the ensemble of regular quantiles.
    Compute and store $\widehat{crps}_{INT}(M, y)$ and $\widehat{crps}_{PWM}(M, y)$ with this observation for the ensemble of optimal quantiles.

Results

Relative estimation errors of $\widehat{crps}_{INT}$ and $\widehat{crps}_{PWM}$ have been computed for a fixed observation ($N = 1$, $y = -0.0841427$) and regular and optimal ensembles, all built from a standard normal distribution ($G = F$ for the sake of simplicity). As shown in Figure 3.4, the CRPSs estimated with quantile ensembles clearly outperform the $\widehat{crps}_{PWM}$ estimation with one random ensemble whatever the number of members. Averaging the $\widehat{crps}_{PWM}$ estimations of 1000 random ensembles gives a similar estimation accuracy to the one of the best estimation with quantile ensembles, namely $\widehat{crps}_{INT}$ with optimal quantiles. This configuration is not feasible in most applications, since it requires 1000 forecast/observation pairs with the same observation. Anyway, computing one set of quantiles may be much simpler and quicker than creating 1000 random ensembles. Among the estimation with ensembles of quantiles, the combination of $\widehat{crps}_{INT}$ and optimal quantiles exhibits a dramatic improvement in accuracy over the other combination, even for ensembles with less than 10 quantiles. Whatever the distribution F is used, $\widehat{crps}_{INT}$ computed with the optimal quantiles gives a much more accurate estimation, for all ensemble sizes, than the other combinations of estimator and type of ensemble of quantiles (not shown).

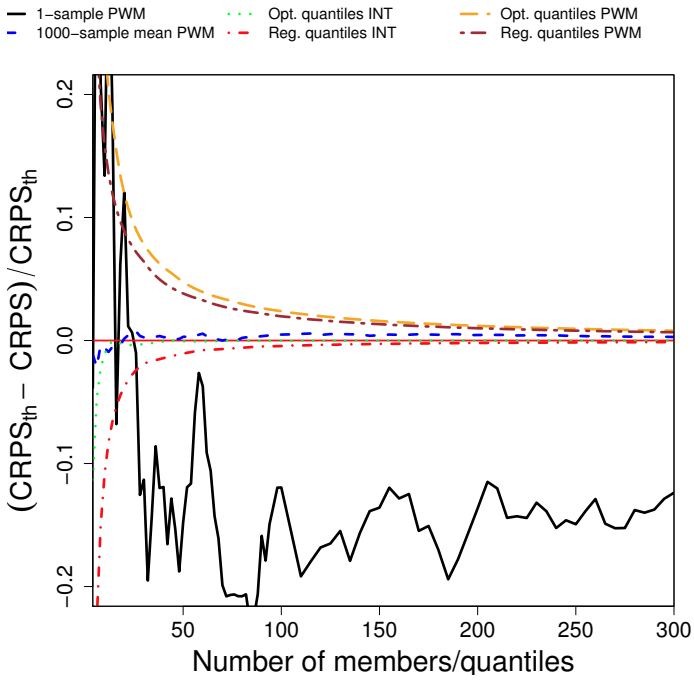

Figure 3.4 – Evolution with ensemble size of relative error of several estimations of CRPS, for different ensembles and different estimators of CRPS. All computation are done with the same observation for all forecasts. The ensembles and the observation come from a standard normal distribution.

In order to assess the robustness of the remarks in the last paragraph in regards to the observation, data are simulated with Protocol 2 for several ensemble sizes M, with $N = 1000$ ensemble forecast/observation pairs for each ensemble size. Note that, at M fixed, the ensemble of quantiles is the same for all the forecast/observation pairs. From the point-wise intervals of the relative estimation errors represented in Figure 3.5, it appears that computing $\widehat{crps}_{INT}$ with the optimal quantiles gives the most accurate estimation of $crps(F, y)$, whatever the number of quantiles is used. With only a few tens of quantiles, this estimation achieves a much higher precision than the other ones with several hundreds of quantiles. Figure 3.5 also shows that, for finite ensembles of quantiles, the PWM estimator is biased, being too low (positive relative errors). Indeed, according to Equation (3.2), since $\widehat{crps}_{INT}$ is an unbiased estimator of the average CRPS of an ensemble of quantiles as shown here, and since $\widehat{\lambda}_2$ is positive, $\widehat{crps}_{PWM}$ must be biased towards low values.

These conclusions hold for all the tried distributions and the set of parameters values for each distribution (not shown).

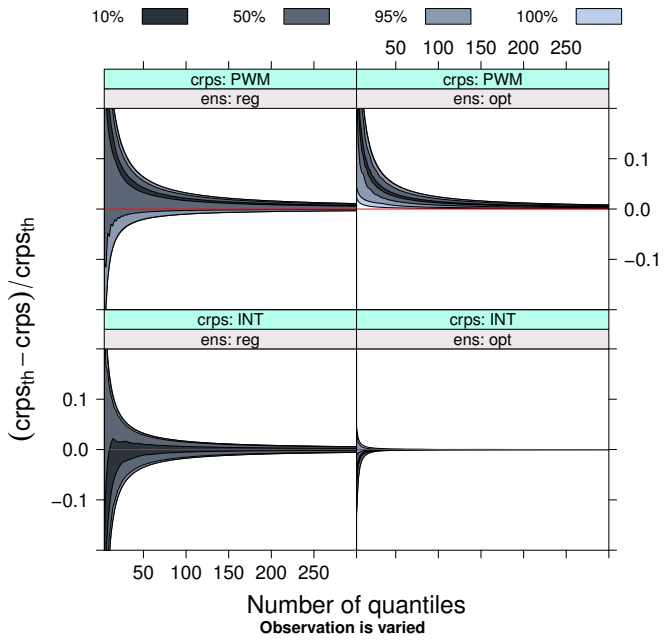

Figure 3.5 – 10%, 50%, 95% and 100% point-wise intervals of relative error for several combination of quantile ensembles and CRPS estimator. Intervals are computed by drawing 1000 observations from a standard normal distribution. Ensembles are regular (left column) or optimal (right column) quantiles of a standard normal distribution. The CRPS is estimated with the PWM (top) or integral (bottom) estimator.

As for the bad performance of $\widehat{crps}_{PWM}$ with an ensemble of quantiles, let us recall that $\widehat{crps}_{PWM}$ is a sum of terms that are unbiased estimators of their population counterpart when computed with a *random* sample, which is not the case of an ensemble of quantiles. The computation of $\widehat{crps}_{INT}$ uses the approximation of the forecast distribution as a step-wise CDF, with a fixed stair-step height $\frac{1}{M}$. The difference in estimation accuracy with the type of quantiles comes from the position of the stair steps. With regular quantiles, the step-wise CDF is always located below the forecast CDF. With optimal quantiles, the associated quantiles are shifted leftward, making the stair steps sometimes above F and sometimes below. This better approximates the forecast CDF F than with regular quantiles, thus improves the estimation of the CRPS.

Influence of ties in an ensemble of quantiles

An ensemble of quantiles may be produced by statistical methods called quantile regression (White 1992; Koenker 2005; Meinshausen 2006; Takeuchi et al. 2006). Some of these quantile regression methods can produce only a subset $\tau^{av}_{j=1,\dots,N_\tau} \in [0;1]$ of N_τ orders. The quantiles associated to these available orders are called "available quantiles" hereafter, and correspond to the abscissa of the black dots in Figure 3.6. If one requires a quantile with an order τ outside of the subset of available orders, the quantile regression will not return the associated quantile of the forecast CDF (abscissa of the blue circles in Figure 3.6), but the available quantile corresponding to the highest available order lower than τ (abscissa of the red triangles of Figure 3.6). The set of different values returned by the quantile regression method when certain orders are requested is called the "unique quantiles" hereafter. It is a subset of the available quantiles. The quantile regression methods with this feature will introduce many ties in the produced ensembles of quantiles, as shown in Figure 3.7 on real data. For the Canadian ensemble forecasts, although 1002 regular quantiles are required from a quantile regression method at one grid point and one lead time, the number of unique quantiles returned by the quantile regression function varies from a few tens of values to a few hundreds. On average, only about one hundred unique quantiles are produced in this example. Some implementations of quantile regression methods, such as the function `rq` in R package `quantreg`, have an option to produce the available orders τ^{av}_j and their associated quantiles. Other packages, such as `quantregForest`, have not yet implemented this possibility, and will return only forecast quantiles with (potentially many) ties.

In order to assess the impact of ties on the accuracy of the CRPS estimators for an ensemble of quantiles, ensembles of quantiles with ties are simulated with Protocol 3, with $N = 1000$ forecast/observation pairs. The left side of Figure 3.8 shows that with only $N_\tau = 30$ available orders, the four estimates become inaccurate. The distribution of the estimated CRPS becomes clearly biased whatever ensemble size is considered. This bias is pessimistic (negative estimation errors) for most ensemble sizes, but may be optimistic (positive estimation errors).

A way to address this issue of equal quantiles is to remove the ties by interpolation. The first considered case is when the implementation of the quantile regression method do not propose to know the available quantiles. Protocol 3 is modified as follow at lines 3 and 4: after computing the quantiles with ties, linear interpolation is done between unique values to recover the number of required regular or optimal quantiles, as explained in Figure 3.6. As shown in the right side of Figure 3.8, this interpolation results in a better estimation accuracy, even though the curves are less smooth than when all orders are available (compare with Figure 3.5). The best CRPS estimation is now obtained with $\widehat{crps}_{INT}$ and regular quantiles, with at least $M = 30$ regular quantiles to get a sufficient accuracy. This behavior barely depends on the chosen distribution and parameters' value, but requiring 100 regular quantiles seems to be the minimal number to get satisfactory

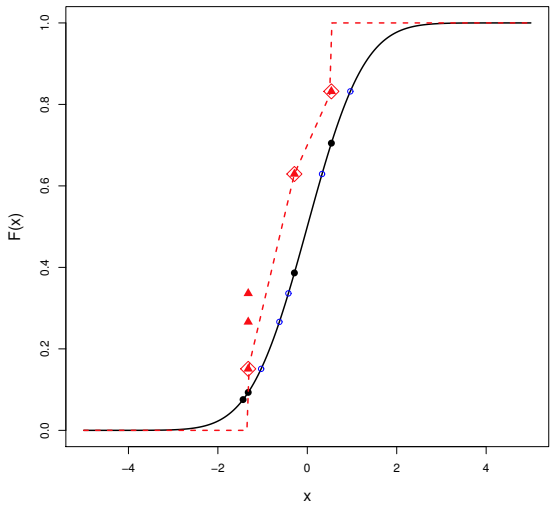

Figure 3.6 – Graphical illustration of the production of ties by quantile regression methods. The black continuous line is the forecast CDF. The abscissa (resp. ordinates) of the $N_\tau = 4$ black dots are the available quantiles (resp. orders), that can produce the quantile regression method. The empty blue dots are the 5 requested points. The red triangles are the 5 points actually obtained, due to the limited number of available quantiles and orders. Within each group of obtained points whose abscissa is the same, only the point with the lowest order is kept (3 red diamonds) for removing the ties by interpolation. The interpolation function (dashed red line) is a linear interpolation between the red diamonds, and a constant order of 0 or 1 outside (left and right, respectively).

accuracy, whatever the forecast distribution F is used (not shown). If the available quantiles and orders can be produced by the implementation of the quantile regression method, similar linear interpolation can be done relatively to the associated points, that is, the black dots in Figure 3.6. Figure 3.9 shows that this linear interpolation nearly fully reproduces the good accuracy obtained when all orders are available. The best estimation strategy is again to use $\widehat{crps}_{INT}$ with optimal quantiles, albeit with a slightly worst accuracy than the one reached without ties.

The influence of the number of available orders N_τ and the kind of post-processing on $\widehat{crps}_{INT}$ is crucial as shown in Figure 3.10. If the number of available quantiles is too low, no matter the post-processing of the quantile ensemble, the estimated CRPS will not converge to the true value due to insufficient information about F. The number of available quantiles necessary to achieve a good accuracy depends on the complexity of the forecast distribution: a gaussian mixture with many different modes requires more available quantiles to be accurately described (not shown here).

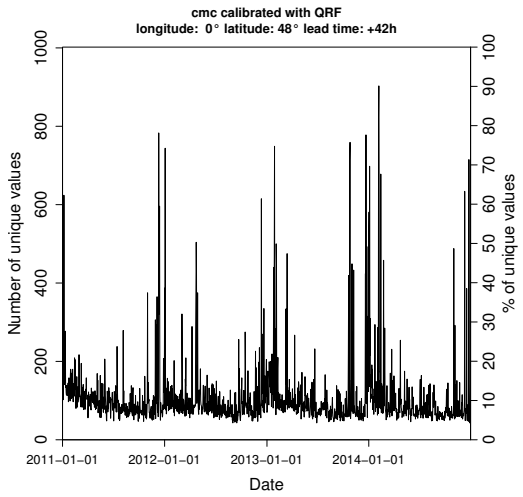

Figure 3.7 – Number and percentage of unique quantiles among 1002 regular quantiles requested from a quantile regression method applied to the Canadian ensemble model.

Protocol 3: Estimation of the CRPS with simulated ensembles of quantiles, with ties.

Input: M: number of quantiles.
F: forecast CDF.
G: CDF of the observation.
N: number of ensemble forecast/observation pairs.
N_τ: number of available quantiles.

Output: N values of instantaneous CRPS for each estimator and kind of quantile ensemble.

1 Draw uniformly in $[0;1]$ the N_τ available orders τ_j^{av}.
2 Compute the available quantiles of F: $F^{-1}(\tau_j^{av}) \forall j \in \{1, \ldots, N_\tau\}$.
3 Compute the ensemble of M regular quantiles of F. Make each regular quantile x_i equal to the available quantile with order τ_j^{av} immediately inferior to τ_i.
4 Compute the ensemble of M optimal quantiles of F. Make each optimal quantile x_i equal to the available quantile with order τ_j^{av} immediately inferior to τ_i.
5 **for** $n \leftarrow 1$ ***to*** N **do**
6 Draw y from G.
7 Compute and store the theoretical CRPS $crps_{th}(F, y)$ with this observation.
8 Compute and store $\widehat{crps}_{INT}(M, y)$ and $\widehat{crps}_{PWM}(M, y)$ with this observation for the ensemble of rounded regular quantiles.
9 Compute and store $\widehat{crps}_{INT}(M, y)$ and $\widehat{crps}_{PWM}(M, y)$ with this observation for the ensemble of rounded optimal quantiles.

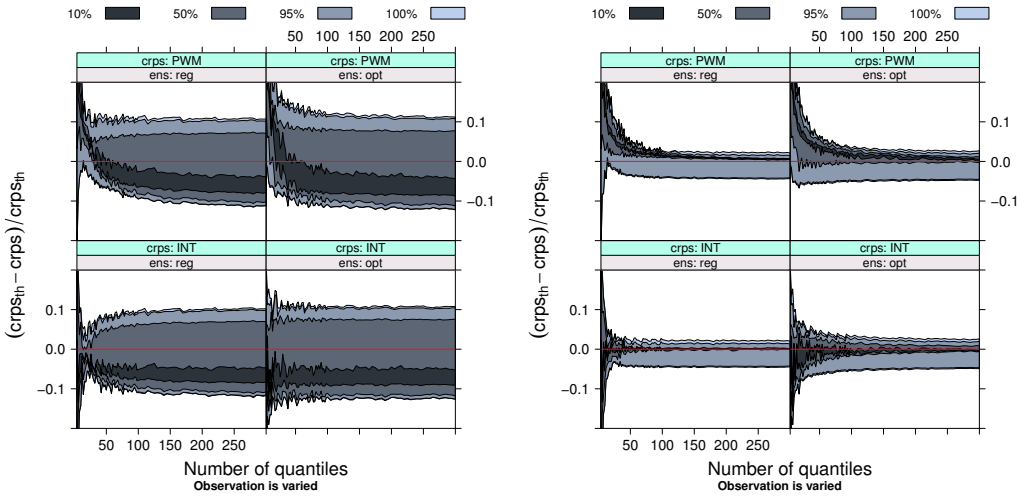

Figure 3.8 – Same as in Figure 3.5 but with ties in the ensembles and only $N_\tau = 30$ available orders (left), and after removing ties by linear interpolation of the unique quantiles in the forecasts (right).

Based on these simulations, several recommendations can be drawn to estimate the instantaneous CRPS of an ensemble of quantiles. First, if the quantile regression cannot yield enough available quantiles (less than about $N_\tau = 30$), the instantaneous CRPS should not be used whatsoever. Even the average CRPS should be used with care due to a (possibly large) estimation bias. However, if the number of available unique quantiles is sufficient (more than 30), the estimation of the instantaneous CRPS can be much improved by interpolating the quantiles and using of $\widehat{crps}_{INT}$. The best interpolation depends on the available information: if the whole set of available quantiles in the quantile regression method is not accessible, linear interpolation between the unique quantiles and their associated order toward regular quantiles should be preferred. However, if the available quantiles and orders can be known, linear interpolation of those quantiles and orders toward optimal quantiles is the best approach.

Table 3.2 sums up the recommendations to estimate the instantaneous CRPS for a random ensemble or an ensemble of quantiles.

3.4 Real data examples

With two real data sets, issues resulting from the uncertainty in the estimation of the instantaneous CRPS are illustrated. The practical benefits of following the recommendations

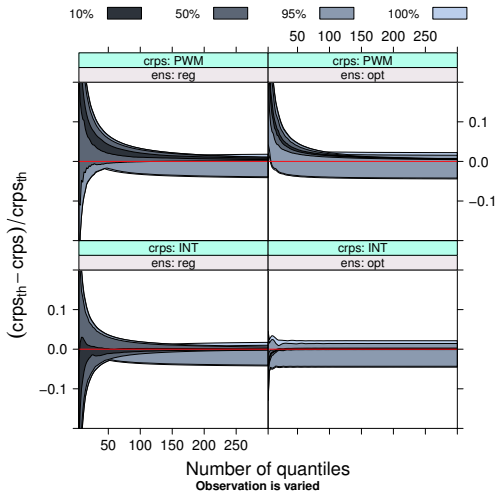

Figure 3.9 – Same as in Figure 3.5 but with ties in the ensembles, only $N_\tau = 30$ available orders, and linear interpolation of the N_τ available quantiles.

Table 3.2 – Summary of recommendations to estimate the CRPS.

Type of ensemble	Condition	Recommendation
Random	The purpose is to assess the performance of an infinite ensemble.	Use average $\widehat{crps}_{PWM}$.
	tThe purpose is to assess the performance of the actual ensemble.	Use average $\widehat{crps}_{INT}$.
Quantiles	All orders available.	Use average $\widehat{crps}_{INT}$ with optimal quantiles.
	$N_\tau \lesssim 30$.	Use average $\widehat{crps}_{INT}$ with care.
	$N_\tau \gtrsim 30$ and available quantiles unknown.	Use average $\widehat{crps}_{INT}$ with linearly interpolated regular quantiles between unique quantiles.
	$N_\tau \gtrsim 30$ and available quantiles known.	Use average $\widehat{crps}_{INT}$ with linearly interpolated optimal quantiles between available quantiles.

listed in Table 3.2 are highlighted.

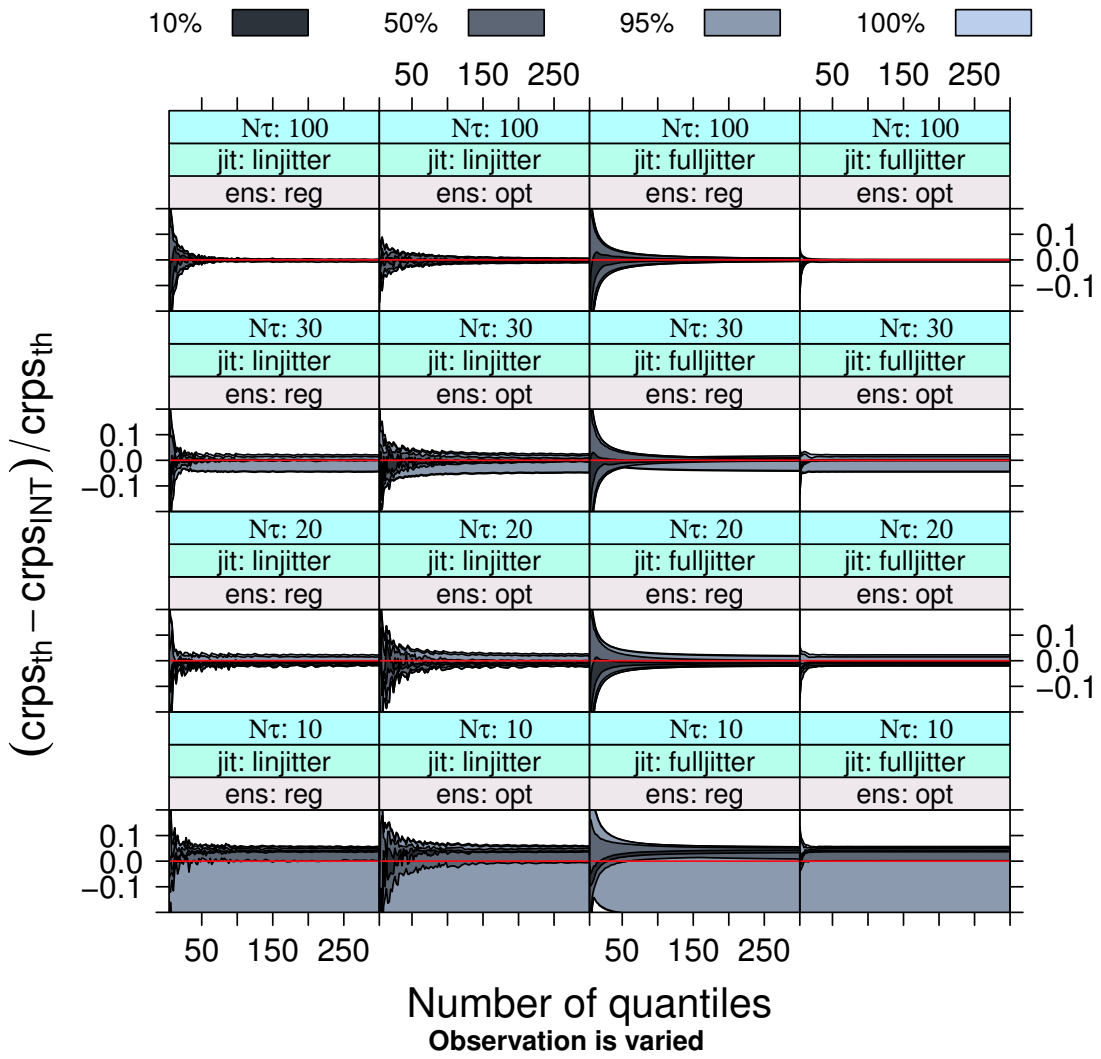

Figure 3.10 – Influence of post-processing. The ensemble quantiles are post-processed by linear interpolation between unique quantiles (*linjitter*) or between the N_τ available quantiles (*fulljitter*). Each panel represents the same intervals as in figure 3.5 for $\widehat{crps}_{INT}$ computed from post-processed quantile ensembles with a varying number N_τ of available quantiles.

3.4.1 Raw and calibrated ensemble forecast data sets

The first forecast data set consists in four NWP ensembles from the TIGGE project (Bougeault et al. 2010). 10 m height wind speed forecasts have been extracted from four operational ensemble models issued by meteorological forecast services: the U.S. National Centers for Environmental Prediction (NCEP), the Canadian Meteorological Center (CMC), the European Center for Medium-Range Weather Forecasts (ECMWF) and Météo-France (MF). Those ensembles have respectively 21, 21, 51 and 35 members. The study domain is France with a grid size of 0.5° (about 50 km), for a total of 267 grid points. Available forecast lead-times are every six hours. The period goes from 2011 to 2014.

The second forecast data set is composed of two versions of each ensemble calibrated with statistical post-processing methods. In order to improve the forecast performance, each ensemble has been post-processed thanks to two statistical methods: non homogeneous regression (NR, Gneiting et al. 2005) and quantile regression forests (QRF, Meinshausen 2006). In NR, the forecast probability distribution F is supposed to be some known distribution: here the square root of forecast wind speed follows a truncated normal distribution whose mean and variance depend on the ensemble forecast. This is similar to the work of Hemri et al. (2014), who also gives the closed form expression of the instantaneous CRPS for this case. QRF is non parametric and yields a set of quantiles x_i with chosen orders τ_i. We use here a simplified version of the model proposed in Taillardat et al. (2016). Since QRF is non parametric, the CRPS has to be estimated with limited information. Furthermore, QRF cannot yield every order and may lead to many ties among predicted quantiles, as seen in Figure 3.7. To the best of our knowledge, no implementation of QRF in R allows to know the available quantiles. Post-processing was done separately for each of the 267 grid points, each ensemble and each lead time. The regression was trained with cross-validation: three years were used as training data, the fourth one being used as test data. The four possible combinations of three training years and one test year were tested. The raw ensembles can be seen as random ensembles whereas the ensembles calibrated with QRF are ensembles of quantiles as defined above.

The observation comes from a wind speed analysis made at Météo-France, presented in Zamo et al. (2016).

3.4.2 Issues estimating the CRPS of real data

In figures 3.11 and 3.12, the CRPS is estimated with the first M members of the raw CMC ensemble at one grid point and for lead time +42 h. First, as shown in Figure 3.11, for very small ensemble sizes, differences between $\widehat{crps}_{INT}$ and $\widehat{crps}_{PWM}$ may be huge. With an increased ensemble size, both estimators get very similar values. Even for the largest number of members, $\widehat{crps}_{INT}$ is systematically higher than $\widehat{crps}_{PWM}$, in agreement

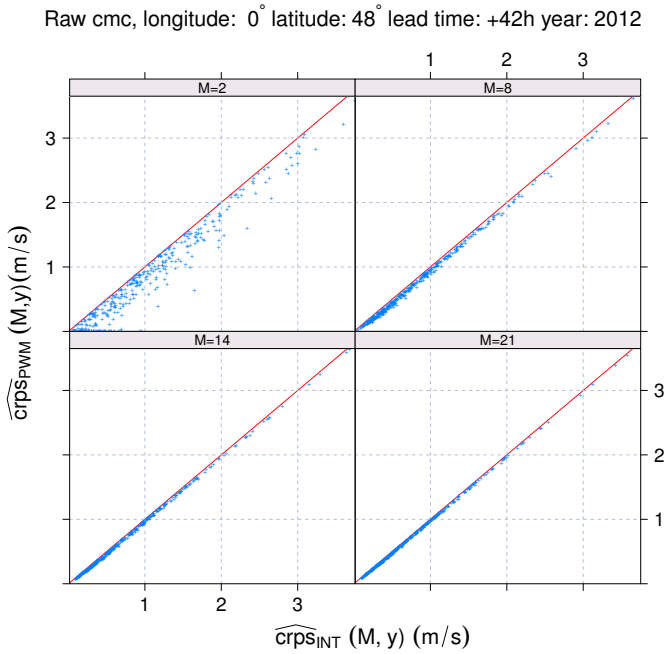

Figure 3.11 – Scatter plots of instantaneous CRPS computed for raw ensemble forecasts, with the two CRPS estimators. The forecasts are for one grid point of the Canadian ensemble forecast model. The number of members goes from 2 to 21 (the actual size of the ensemble). Each point corresponds to one forecast (one date and valid time).

with Equation (3.2). These differences result in important differences on the averaged CRPS, as shown in Figure 3.12, representing the evolution with M of the yearly-averaged $\widehat{crps}_{INT}$ and $\widehat{crps}_{PWM}$. Whereas the yearly-averaged $\widehat{crps}_{PWM}$ is nearly independent of M, the average $\widehat{crps}_{INT}$ requires a minimum ensemble size to yield a stable value. But even then, the two estimators do not yield the same average CRPS: for the year 2011, on average $\widehat{crps}_{INT}(M = 21) \simeq 0.75m/s$ whereas $\widehat{crps}_{PWM}(M = 21) \simeq 0.7m/s$, a difference of 7 %. These conclusions from Figure 3.12 are in agreement with those from Figure 3.3, that shows that the average $\widehat{crps}_{PWM}$ attains the true value with much smaller ensembles than $\widehat{crps}_{INT}$. The left side of Figure 3.3 exhibits negative estimation errors which is in agreement with the averaged $\widehat{crps}_{INT}$ being higher than the averaged $\widehat{crps}_{PWM}$ in Figure 3.12 and in agreement with Equation (3.2).

Figure 3.13 uses the version of the CMC ensemble calibrated with QRF. For each of the two sets of quantiles, $\widehat{crps}_{INT}$ and $\widehat{crps}_{PWM}$ are computed for each forecast date and averaged over each test year. The number M of requested quantiles is varied from 2 to 50 and are either of regular or optimal orders. Figure 3.13 shows the evolution of the four estimated

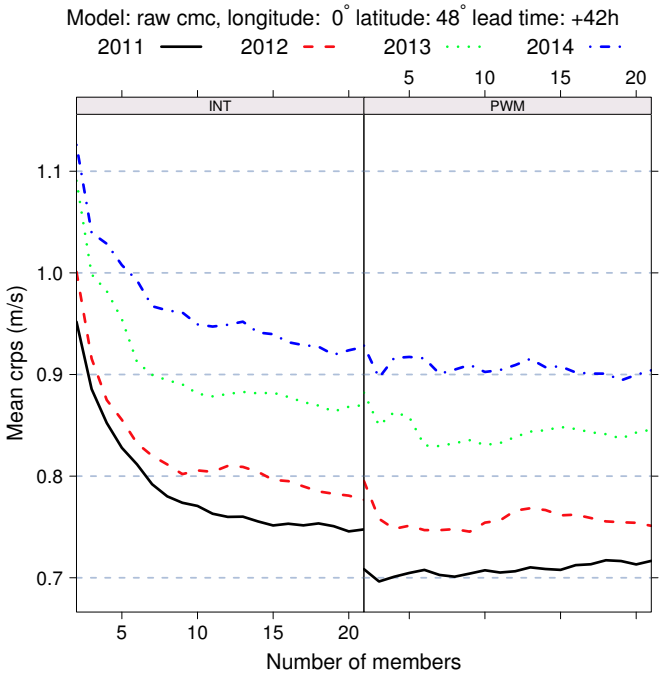

Figure 3.12 – Evolution of the yearly-averaged CRPS with the number of members for the raw CMC ensemble. Each panel contains the average CRPS computed by averaging the instantaneous CRPS estimator, $\widehat{crps}_{INT}$ (left) or $\widehat{crps}_{PWM}$ (right). Each curve is computed by averaging the estimated instantaneous CRPS over one test year, for forecasts at one grid point and for one lead time.

average CRPS with the number of quantiles, for the same grid point and lead time as above. First, the average $\widehat{crps}_{INT}$ decreases rapidly toward some value, whatever the kind of quantiles. Second, the yearly-averaged $\widehat{crps}_{PWM}$ is not independent of the number of quantiles, as it was independent of the number of members in Figure 3.12. Here, it slowly increases toward some value for a fixed kind of quantiles. Third, the limit values are on average $\widehat{crps}_{PWM}(50) \simeq 0.48m/s$, $\widehat{crps}_{INT}(50) \simeq 0.47m/s$ a difference of only 2 %. Last, the rate of evolution of the average CRPS with the ensemble size strongly depends on the choice of the CRPS estimator and of the type of required quantiles. For these data, removing ties in the forecast quantiles do not change the conclusions (not shown). In agreement with the recommendations from the simulated data, the fastest converging estimate is the average $\widehat{crps}_{INT}$ computed with optimal quantiles.

Other ensembles, grid points and lead-times give similar results (not shown).

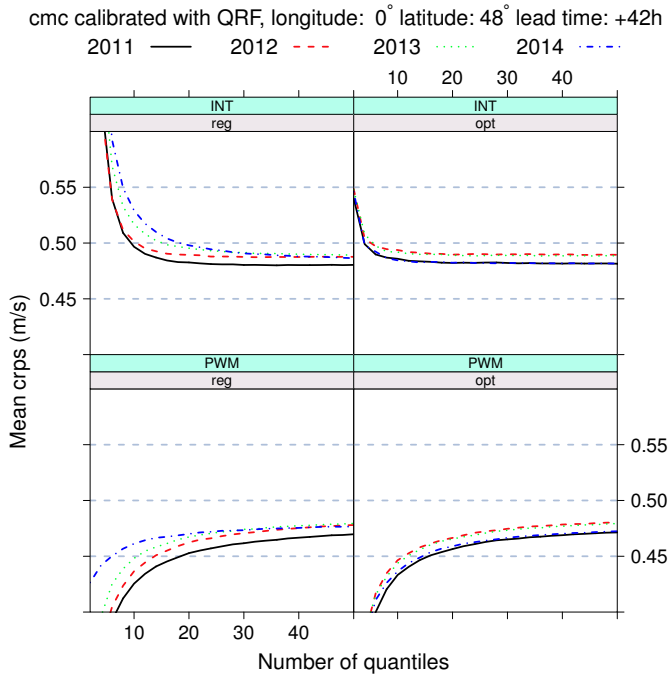

Figure 3.13 – Evolution with the number of members of the estimated CRPS averaged over one year for CMC ensemble forecast calibrated with quantile regression forests. Two sets of quantiles are requested: regular (left) and optimal (right). Ties between quantiles are not removed. The two formulae (eINT) (top) and (ePWM) (bottom) are used to estimate the instantaneous CRPS for each quantile sets. Each curve is then computed by averaging the estimated instantaneous CRPS over one year, for forecasts at one grid point and for one lead time.

3.4.3 Issues on the choice between QRF and NR

For the real data set, the CRPS of QRF has been estimated with $\widehat{crps}_{INT}$ and $\widehat{crps}_{PWM}$ computed with optimal quantiles, and ties have been kept or removed by interpolation. Figure 3.14 shows the proportion of times QRF gets a lower CRPS than NR, out of the 365 forecasts during test year 2012, for one grid point and one lead time with calibrated CMC data. The proportion of times QRF outperforms NR strongly depends on the number of quantiles but stabilizes at similar values when $\widehat{crps}_{INT}$ or $\widehat{crps}_{PWM}$ is used. In agreement with the conclusions on simulated data, the proportion stabilizes with less quantiles when $\widehat{crps}_{INT}$ is used. With too few quantiles (less than about 20), the difference of performance between QRF and NR may be deemed significant depending on the estimator. But in this specific case, after the curves have stabilized, the performance of QRF and NR are not

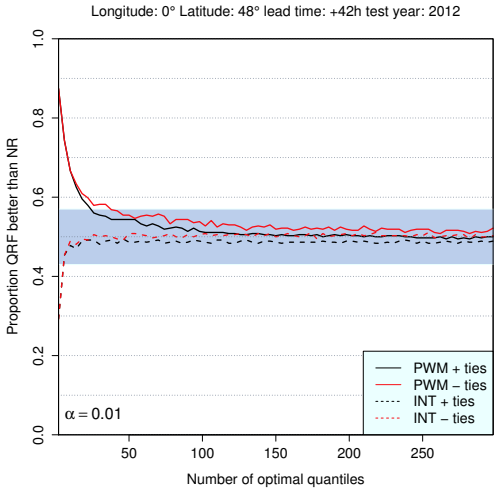

Figure 3.14 – Proportion of forecasts when QRF gets a lower CRPS than NR, for calibrated CMC ensemble, at one grid point, for one lead time and one test year. QRF yields M optimal quantiles and its CRPS is estimated with $\widehat{crps}_{PWM}$ (continuous line) or $\widehat{crps}_{INT}$ (dashed line), without removing ties (black curves) or after removing ties with linear interpolation between unique quantiles (red curves). NR's CRPS is computed with the closed form expression available in Hemri et al. (2014). The grey zone is the 0.01-confidence interval that the proportion is not significantly different from 0.5 (quantiles 0.995 and 0.005 of a binomial distribution with 365 tries).

statistically different to the level 0.01 for all the estimations. This shows that the choice of the best post-processed forecast may be misguided by poor performance estimates if the wrong estimator is used and/or not enough quantiles are required. The number of available quantiles is unknown but has been estimated to be at least 52 for this test year. Based on the recommendations in Table 3.2, the best method to estimate the CRPS of QRF would be to use $\widehat{crps}_{INT}$ and at least 30 optimal quantiles, which is in agreement with the previous remarks.

3.5 Conclusion and discussion

A review of four estimators of the instantaneous CRPS when the forecast CDF is known through a set of values have been done. Among these four estimators proposed in the literature, only two, called the integral estimator and the probability weighted moment estimator, are not equal. Furthermore, a relationship between these two estimators have been demonstrated, and generalizes to the instantaneous CRPS of any ensemble, a relation-

ship established by Ferro et al. (2008) for the average CRPS of a random ensemble. With simulated data, the accuracy of the two estimators has been studied, when the forecast CDF is known with a limited information and the number of forecast/observation pairs is finite. The study leads to recommendations on the best CRPS estimator depending on the type of ensemble, whether random or a set of quantiles. For a random ensemble, the best estimator of the CRPS is the PWM estimator $\widehat{crps}_{PWM}$ if one wants to assess the performance of the ensemble of infinite size, whereas the integral estimator $\widehat{crps}_{INT}$ must be used to assess the performance of the ensemble with its current size. For an ensemble of quantiles, ties introduced by quantile regression methods strongly affect the estimation accuracy, and removing these ties by an interpolation step is paramount to allow a good estimation accuracy. If the number of available quantiles is too low (say, $N_\tau \leq 30$) all the studied estimators exhibit a strong bias. But if the number of available quantiles is larger, the best estimation is obtained by computing the integral estimator $\widehat{crps}_{INT}$ with linearly interpolated quantiles, between the available quantiles if they are known or between the unique quantiles otherwise.

The established relationships between the estimators proposed in the literature have been linked to previous results. These relationships also explain why an estimator is more accurate for one type of ensemble and not for the other. The PWM estimator performs better on random ensembles because it is based on estimators that are unbiased for independent samples from the true underlying distribution. On the other hand, the integral estimator gives a good estimate when computed with optimal quantiles. This is because regular weights are associated to the members in the estimator formula but, when using optimal quantiles, the associated quantiles are shifted to better approximate the underlying forecast CDF.

The important consequences on the choice of method of estimation of the CRPS has also been illustrated on real meteorological data with raw ensembles and calibrated ensembles. As an example, the comparison of several calibrated ensembles may be mislead by a poor estimate of the average CRPS of ensembles of quantiles.

Acknowledgements: Part of the work of Philippe Naveau has been supported by the ANR-DADA, LEFE-INSU-Multirisk, AMERISKA, A2C2, CHAVANA and Extremoscope projects. Part of the work was done when Philippe Naveau was visiting the IMAGE-NCAR group in Boulder, Colorado, USA. Both authors are grateful to Liliane Bel (AgroParisTech, France) and Tilmann Gneiting (Heidelberg Institute for Theoretical Studies, Germany) for their useful comments on this paper.

Appendix

3.A What is elicited when the 1-norm CRPS of an ensemble is minimized?

Let $\{x_i\}_{i=1,\dots,M}$ be an ensemble of M values. Let $F_e(x) = \sum_{i=1}^{M} \omega_i \mathbb{1}(x \geq x_i)$ be the associated empirical CDF, with weights ω_i, such that $\omega_i \geq 0 \quad \forall i \in \{1, \dots, M\}$ and $\sum_{i=1}^{M} \omega_i = 1$. Let y be the observation.

Following Müller et al. (2005), the 1-norm CRPS of this ensemble relative to this observation is defined as

$$crps_1(F_e, y) = \int_{\mathbb{R}} |F_e(x) - \mathbb{1}(x \geq y)| dx.$$

This can be rewritten in a more interpretable form.

$$\begin{aligned}
crps_1(F_e, y) = &\int_{-\infty}^{y} \left| \sum_{i=1}^{M} \omega_i \mathbb{1}(x \geq x_i) \right| dx \\
&+ \int_{y}^{+\infty} \left| \sum_{i=1}^{M} \omega_i \left(\mathbb{1}(x \geq x_i) - 1 \right) \right| dx \\
= &\int_{-\infty}^{y} \sum_{i=1}^{M} \omega_i \mathbb{1}(x \geq x_i) dx \\
&+ \int_{y}^{+\infty} \sum_{i=1}^{M} \omega_i \left(1 - \mathbb{1}(x \geq x_i) \right) dx \\
= &\sum_{i=1}^{M} \omega_i \Bigg[\int_{-\infty}^{y} \mathbb{1}(x \geq x_i) dx \\
&+ \int_{y}^{+\infty} 1 - \mathbb{1}(x \geq x_i) dx \Bigg].
\end{aligned}$$

If $y \geq x_i$,

$$\int_{-\infty}^{y} \mathbb{1}(x \geq x_i) dx = \int_{-\infty}^{x_i} 0 dx + \int_{x_i}^{y} 1 dx = y - x_i,$$

and

$$\int_{y}^{+\infty} 1 - \mathbb{1}(x \geq x_i) dx = \int_{y}^{+\infty} 0 dx = 0.$$

If $y \leq x_i$,

$$\int_{-\infty}^{y} \mathbb{1}(x \geq x_i)dx = \int_{-\infty}^{y} 0dx = 0,$$

and

$$\int_{y}^{\infty} 1 - \mathbb{1}(x \geq x_i)dx = \int_{y}^{x_i} 1dx + \int_{x_i}^{+\infty} 0dx = x_i - y.$$

Therefore, $\forall y$ and $\forall i$

$$\int_{-\infty}^{y} \mathbb{1}(x \geq x_i)dx + \int_{y}^{+\infty} 1 - \mathbb{1}(x \geq x_i)dx = |y - x_i|.$$

Finally,

$$crps_1(F_e, y) = \sum_{i=1}^{M} \omega_i |y - x_i|.$$

The 1-norm CRPS is just the weighted mean of the absolute error of each member. The average 1-norm CRPS is thus minimized if all the members are equal to the median of the observation CDF (Gneiting 2011b).

3.B Relationships between the estimators of the CRPS

Without loss of generality, the forecast is an ensemble of M values $x_{i=1,\dots,M}$ sorted in increasing order.

3.B.1 Equality of $\widehat{crps}_{Fair}$ and $\widehat{crps}_{PWM}$

Following the definition of L-moments and their relationship with PWMs (Wang 1996; Hosking 1990), one can rewrite

$$\begin{aligned}
\hat{\lambda}_2 =& \frac{1}{2M(M-1)} \sum_{i,j=1}^{M} |x_i - x_j| \\
=& 2\hat{\beta}_1 - \hat{\beta}_0 \\
=& \frac{1}{M(M-1)} \sum_{i,j=1}^{M} (2i - M - 1)x_i,
\end{aligned} \tag{3.3}$$

where $\hat{\lambda}_2$, $\hat{\beta}_1$ and $\hat{\beta}_0$ are estimators of the second linear moment, the PWM of order 1 and the PWM of order 0 (i.e. the average), respectively. These estimators are unbiased if the ensemble is a random sample.

Introducing these notations in Equation (eFAIR) leads to

$$\begin{aligned}\widehat{crps}_{Fair}(M,y) =& \frac{1}{M}\sum_{i=1}^{M}|x_i - y| + \hat{\beta}_0 - 2\hat{\beta}_1 \\ =& \widehat{crps}_{PWM}(M,y).\end{aligned}$$

3.B.2 Equality of $\widehat{crps}_{NRG}$ and $\widehat{crps}_{INT}$

As Gneiting and Raftery (2007) showed, the representations (INT) and (NRG) are equivalent for forecast CDFs with a finite first moment. Since empirical distributions have a finite first moment, and since (INT) and (NRG) reduce to (eINT) and (eNRG) respectively, equality of $\widehat{crps}_{INT}$ and $\widehat{crps}_{NRG}$ follows immediately.

Thanks to Pr. Tilmann Gneiting for this proof, much more straightforward than the one initially proposed.

3.B.3 Relationship between $\widehat{crps}_{PWM}$ and $\widehat{crps}_{NRG}$

Using (3.3) leads to

$$\begin{aligned}\widehat{crps}_{NRG}(M,y) =& \frac{1}{M}\sum_{i=1}^{M}|x_i - y| - \frac{2M(M-1)}{2M^2}(2\hat{\beta}_1 - \hat{\beta}_0) \\ =& \frac{1}{M}\sum_{i=1}^{M}|x_i - y| + \hat{\beta}_0 - 2\hat{\beta}_1 + \frac{\hat{\lambda}_2}{M} \\ =& \widehat{crps}_{PWM}(M,y) + \frac{\hat{\lambda}_2}{M}.\end{aligned}$$

Chapter 4

Sequential Aggregation of Probabilistic Wind Speed Forecasts

Contents

4.1 Introduction 90
4.2 Theoretical framework and verification strategy 91
4.2.1 The individual sequence prediction framework 91
4.2.2 Sequential aggregation of step-wise CDFs 92
4.2.3 Verification strategy 94
4.3 Aggregation methods 97
4.3.1 Inverse CRPS weighting 98
4.3.2 Sharpness-calibration paradigm 98
4.3.3 Minimum CRPS 99
4.3.4 Exponential weighting 99
4.3.5 Exponentiated gradient 99
4.4 The experts and the observation 100
4.4.1 The TIGGE data set and experts 100
4.4.2 The calibrated experts 101
4.4.3 The observation 102
4.5 Results 103
4.5.1 CRPS, reliability and sharpness 103
4.5.2 Spatio-temporal characteristics of the most reliable aggregated forecast 107

4.6 Discussion about calibration and aggregation procedures . . . 108
4.7 Conclusion and perspectives . . . 108
Appendix . . . 119
4.A Formula for the gradient of the CRPS . . . 119
4.B Bounds for the EWA aggregation . . . 121
4.C Time series of the regret at each lead time . . . 123
4.D Maps of rank histograms of raw ensembles . . . 127
4.E Time series of the aggregation weights . . . 135

4.1 Introduction

As a chaotic dynamical system, the atmosphere has an evolution that is intrinsically uncertain (Malardel 2005; Holton and Hakim 2012). In the field of numerical weather prediction (NWP), assessing the forecast uncertainty is the primary goal of ensemble forecasts (Leutbecher and Palmer 2008). An ensemble forecast consists in a set of deterministic forecasts, called members. The most common ensemble NWP forecasts consist in several members obtained by running the same NWP model with different initial conditions and/or different parametrizations of the model physics (Descamps et al. 2011). From the distribution of the members, the uncertainty may be derived in a probabilistic way. Nowadays, several ensemble forecast systems are available routinely (Bougeault et al. 2010; Descamps et al. 2014). Being often biased and under-dispersed (Hamill and Colucci 1998; Buizza et al. 2005), these ensemble forecast systems are sometimes post-processed with statistical methods, called ensemble model output statistics (EMOS) to improve the forecast performances (Wilson et al. 2007; Thorarinsdottir and Gneiting 2010; Möller and Scheuerer 2013; Baran and Lerch 2015; Taillardat et al. 2016).

In the following, we investigate several ways to aggregate several raw or post-processed ensembles, called "experts", in order to create a more skillful "meta ensemble". Loosely speaking, the aggregation is simply a linear combination of the ensembles. Since we are dealing with probability distributions, only convex aggregation strategies will be investigated: the individual experts' weights are constrained to be positive and to sum up to one.

The desired properties of the aggregation are two-fold. The first one is to get a better forecast performance. It has been shown for deterministic forecasts (Fritsch et al. 2000; Baars and Mass 2005; Woodcock and Engel 2005) and, less commonly, for probabilistic forecasts (Allard et al. 2012; Gneiting et al. 2013; Baudin 2015; Baran and Lerch 2016) that aggregating several forecasts may improve the forecast performance compared to the most skillful post-processed forecast. Some aggregation methods even have theoretical

guarantees that the aggregated forecast will not perform much worse than some skillful reference forecast, called the oracle (Cesa-Bianchi et al. 2006; Stoltz 2010). In practice, the aggregated forecast may even outperform the oracle. The second goal is to dynamically tackle changes in the ensemble models, that may strongly affect the performance of the raw or post-processed ensembles. A good aggregation method should quickly detect changes in the performance of the experts and adapt the aggregation weights to discard the bad ones and favor the good ones. In this work, we will focus on forecasting the 10 m wind speed over France.

This work is inspired by the work of Baudin (2015) who aggregated in a similar way several ensembles, the experts being the *sorted* forecast values of the pooled ensembles. Thus, the experts from Baudin (2015) are not identifiable over time as required by the theory used in this work and ours. As an example, at different times, the lowest forecast value could come from a different member of a different ensemble. Furthermore, the experts are weighted as deterministic forecasts. Being a whole ensemble, each expert used in the present work is a truly probabilistic forecast identifiable over time.

In Section 4.2, the theoretical framework of sequential aggregation of step-wise cumulative distribution functions (CDF), used in this study, is presented, along with notations. The tools used to assess the performances of the forecasts are also introduced. Section 4.3 presents the different aggregation methods investigated. Some are rather empirical, while others exhibit interesting theoretical properties. Section 4.4 describes the ensemble forecasts, the EMOS methods used to post-process the ensembles, and the wind speed observation. The results of the comparison of the aggregation methods are presented in Section 4.5. These results motivate a discussion of the methodology of the post-processing and aggregation of ensemble forecasts, in Section 4.6. Finally, Section 4.7 concludes with a summary of the results.

4.2 Theoretical framework and verification strategy

4.2.1 The individual sequence prediction framework

The theoretical framework underlying the present study requires no assumption about the properties of the sequences of observations and forecasts, whose generating process can be deterministic or statistical, stationary or not, or anything else. Cesa-Bianchi et al. (2006) and Stoltz (2010) describe this very general framework, called the prediction of individual sequences.

An individual sequence is any sequence of values $y_t \in \mathcal{Y}$, at times $t = 1, 2, \ldots$, of a parameter of interest, or observation. The set of possible values of y_t, $\mathcal{Y}$, is usually but not necessarily the set of real numbers. At each time t, and before the value of y_t is revealed, a forecaster

produces a forecast $\widehat{y}_t \in \widehat{\mathcal{Y}}$, based on some information $\mathcal{I}_t$, possibly with $\widehat{\mathcal{Y}} \neq \mathcal{Y}$. The information $\mathcal{I}_t$ may contain the past observations of y_t, and any other piece of information such as outputs of numerical models, experts' advice, measurements of parameters related to the observation, and so on.

When the observation y_t is revealed, the forecaster suffers a loss, quantified with a loss function $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to \mathbb{R}$. The most general goal of the forecaster would be to build the best possible forecast, that is, to minimize its cumulative loss over a period of time $t = 1, \ldots, T$. This minimization is not possible for all possible individual sequences, since one can always build a sequence of observations that makes the forecaster's cumulative loss arbitrarily high. Therefore, a more realistic goal is to build the best possible forecast relatively to the best element from some class of reference forecasts. Let us note $\mathcal{C}$ a class of functions $\widetilde{y} : \mathcal{I}'_t \to \widetilde{y}_t \in \widehat{\mathcal{Y}}$, defined as a class of forecast algorithms based on information $\mathcal{I}'_t$, with possibly $\mathcal{I}'_t \neq \mathcal{I}_t$. Common classes $\mathcal{C}$ are the set of experts, or the set of fixed convex combinations of experts. Let us note

$$R_T^{\mathcal{C}} = \sum_{t=1}^{T} \ell(\widehat{y}_t, y_t) - \inf_{\widetilde{y} \in \mathcal{C}} \sum_{t=1}^{T} \ell(\widetilde{y}_t, y_t) \tag{4.1}$$

the *regret* of the forecaster relatively to the class $\mathcal{C}$. The regret is the cumulative additional loss suffered by the forecaster who used its own forecast algorithm instead of the best forecast algorithm from class $\mathcal{C}$. The computation of this best forecast algorithm from class $\mathcal{C}$, called the oracle, requires all the information for the whole period. Thus, the oracle cannot be used for real-time applications. It can be shown that, for some specific forecast algorithms and specific classes $\mathcal{C}$, the regret is sub-linear in T, that is, if the loss function ℓ is convex in its first argument, then

$$\sup R_T^{\mathcal{C}} = o(T), \tag{4.2}$$

where the supremum is taken over all possible individual sequences of the observation, and over all the sources of information $\mathcal{I}_t$ available to the forecaster. The class $\mathcal{C}$ depends on the algorithm used by the forecaster. Since the bounds hold for *any* individual sequence, the individual sequence prediction framework is very interesting to ensure to the forecaster good forecast performances.

4.2.2 Sequential aggregation of step-wise CDFs

The prediction with expert advice is the special situation of individual sequence prediction considered in this study. The information $\mathcal{I}_t$ available to the forecaster is composed of the current forecasts of $E \in \mathbb{N}^*$ so-called "experts" and of the past observations $y_{j=1,\ldots,t-1}$. An expert is any means, in a very general sense, to produce a forecast at each time t, and

before the observation y_t is revealed. The forecast of expert e at time t is noted $\widehat{y}_{e;t} \in \widehat{\mathcal{Y}}$, with $e \in \{1, \dots, E\}$. The forecaster produces an aggregated forecast as a combination of the experts' current forecasts $\widehat{y}_{e=1,\dots,E;t}$,

$$\widehat{y}_t = \sum_{e=1}^{E} \omega_{e;t} \widehat{y}_{e;t},$$

where $\omega_{e;t} \in \Omega$ is the aggregation weight of expert e at time t, and Ω is $\mathbb{R}$ or a subset thereof. The aggregation weights are computed using only past information $(\mathcal{I}_1, \dots, \mathcal{I}_{t-1})$, namely the past experts' forecasts $\widehat{y}_{e;j}$ and the past observations y_j, with $e = 1, \dots, E$ and $j = 1, \dots, t-1$. The aggregation is initialized with equal weights, that is, $\omega_{e;1} = \frac{1}{E}, \quad \forall e = 1, \dots, E$. Finally, in prediction with expert advice, the class $\mathcal{C}$ mentioned above is defined on a subset Ω^E of the aggregation weights' space, so that

$$\widetilde{y}_t = \sum_{e=1}^{E} \widetilde{\omega}_{e;t} \widehat{y}_{e;t},$$

with the weights $\widetilde{\omega}_{e;t} \in \Omega^E \subseteq \Omega$. The specific subset depends on the sequential aggregation algorithm used by the forecaster, some of which are described in Section 4.3. Mallet et al. (2007) and Gerchinovitz et al. (2008) contain concise reviews of many aggregation methods with expert advice, and numerical algorithms thereof. These two papers concern the case of real experts ($\widehat{\mathcal{Y}} \subseteq \mathbb{R}$) and a real observation ($\mathcal{Y} \subseteq \mathbb{R}$), along with theoretical bounds for the L_2-loss ($\widehat{\ell}(\widehat{y}_t, y_t) = (\widehat{y}_t - y_t)^2$), when they exist.

This study is more specifically concerned with sequential aggregation of step-wise CDFs, that is, the forecast space $\widehat{\mathcal{Y}}$ is the set of piece-wise constant, non-decreasing functions taking their values in $[0; 1]$. The experts are supposed to produce forecasts in the form of a discrete set of values from some (possibly unknown) CDF $F_{e;t}(x)$. For instance, these sets of values can be members from an NWP ensemble model, a set of quantiles from a statistical quantile regression method, or a sample from Monte Carlo simulations In mathematical notations, each expert forecast $\widehat{y}_{e;t}$ is a step function with jumps of heights $p_e^{m_e}$ (called weights) at the M_e values $x_{e;t}^{m_e}$. The weights are such that $p_e^{m_e} > 0$, for $m_e = 1, \dots, M_e$, and $\sum_{m_e=1}^{M_e} p_e^{m_e} = 1$. Then each expert's forecast CDF is approximated by the step-wise CDF $\widehat{y}_{e;t} = \widehat{F}_{e;t}(x) = \sum_{m_e=1}^{M_e} p_e^{m_e} H(x - x_{e;t}^{m_e})$, H being the Heaviside function, with $x \in \mathcal{Y}$. Without loss of generality, the $x_{e;t}^{m_e}$ are supposed sorted in ascending order for each expert and at each time t, so that $x_{e;t}^{m_e}$ is the quantile of order $\tau_e^{m_e} = \sum_{m'_e=1}^{m_e} p_e^{m'_e}$ of $F_{e;t}(x)$.

The aggregated forecast CDF, $\widehat{y}_t = \widehat{F}_t(x)$, consists in the pooled values $\{x_{e;t}^{m_e}; m_e = 1, \dots, M_e, e = 1, \dots, E\}$ with a jump of height $\omega_{e;t} p_e^{m_e}$ associated to the value $x_{e;t}^{m_e}$, thus

$$\widehat{F}_t(x_{e;t}^{m_e}) \;\; = \sum_{e'=1}^{E} \omega_{e';t} \left(\sum_{m'_{e'}=1}^{M_{e'}} p_{e'}^{m'_{e'}} H(x_{e;t}^{m_e} - x_{e';t}^{m'_{e'}}) \right) \;\; = \tau_{e;t}^{m_e}. \tag{4.3}$$

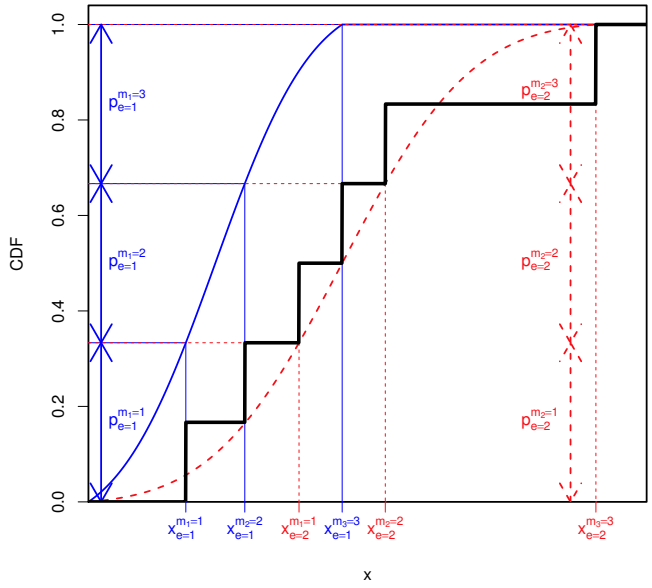

Figure 4.1 – Example of aggregation of $E = 2$ CDFs. The forecast CDFs (blue continuous line for expert $e = 1$, red dashed line for expert $e = 2$) are known only through a set of $M_1 = M_2 = 3$ values $x_e^{m_e}$ with associated jumps $p_e^{m_e} = \frac{1}{3}$ (blue for expert 1, red for expert 2). Following Equation (4.3), since x_2^1 is greater than x_1^1 and x_1^2, it is the quantile of order $\tau_{2;t}^1 = \omega_{1;t}(p_1^1 + p_1^2) + \omega_{2;t}p_2^1 = \frac{2}{3}\omega_{1;t} + \frac{1}{3}\omega_{2;t}$ of F_t. If $\omega_{1;t} = \omega_{2;t} = \frac{1}{2}$, then $\tau_{2;t}^1 = \frac{1}{2}$ and the aggregated step-wise CDF is the black continuous line.

In other words, $x_{e;t}^{m_e}$ is the quantile of order $\tau_{e;t}^{m_e}$ of F_t, whose computation is illustrated in Figure 4.1.

To produce a valid step-wise CDF, the aggregation algorithm must ensure that the aggregation weights are such that $\omega_{e;t} \geq 0$, $\forall e \in \{1, \ldots, E\}$ and $\forall t \in \{1, \ldots, T\}$, and that $\sum_{e=1}^{E} \omega_{e;t} = 1, \forall t \in \{1, \ldots, T\}$. Thus, the aggregated forecast is a convex combination of the expert forecasts.

4.2.3 Verification strategy

Three graphical or numerical tools are used to assess the forecast performance: the Continuous Ranked Probability Score, the rank histogram and the sharpness diagram.

For probabilistic forecasts expressed as a CDF F and a scalar observation $y \in \mathcal{Y} \subseteq \mathbb{R}$, a

natural and much used loss function ℓ is the Continuous Ranked Probability Score (CRPS, Matheson and Winkler 1976), defined as

$$\ell(F, y) = \int_{x \in \mathcal{Y}} (F(x) - H(x - y))^2 \, dx.$$

As all average score, the average CRPS can be decomposed into three terms that quantify specific properties of a forecast system or an observation: the reliability, the resolution and the uncertainty. A forecast system is reliable if the observations associated with a specific forecast distribution are distributed according to this distribution. As an example, for a stationary forecast system and observation, always forecasting the climatological distribution of the observation gives a perfectly reliable forecast system, although being not very informative. A forecast system has a high resolution if he can issue forecast distributions very different from the climatological distribution. By definition, the climarological distribution, although reliable, has no resolution at all. Uncertainty is a property of the observation only, and is defined as the variability of the observation. The interested reader is refered to Bröcker (2009) for the formulae defining the corresponding three terms in an average score, and to Hersbach (2000) for the method to compute them in the case of the CRPS of an ensemble forecast. The existence of theoretical bounds for the regret requires that the loss function is convex in its first argument, which is the case of the CRPS. Since the experts are step-wise CDFs, the information about the underlying forecast CDF is incomplete, which creates issues about the accuracy of the CRPS estimation as investigated in Chapter 3. The recommendations from this previous study have been followed to estimate accurately the CRPS of the experts and aggregated forecasts.

The rank histogram of an ensemble forecast, simultaneously introduced by Anderson (1996), Hamill and Colucci (1996) and Talagrand et al. (1997), is the histogram of the rank of the observation when it is pooled with its corresponding forecast members. For a reliable ensemble, the observation and the members must have the same statistical properties, resulting in a flat rank histogram. The deviations from flatness gives indications about the flaws of an ensemble. For instance, an L-shape histogram means the forecasts are consistently too high, while a J-shape histogram indicates consistently too low forecasts. But Hamill (2001) showed on synthetic data that a flat rank histogram can be obtained with an observation and members differently distributed, that is, for an unreliable ensemble. A flat rank histogram is thus a necessary but not sufficient condition for a forecast to be reliable. Since a necessary quality of a probabilistic forecast useful for decision making is its reliability, the flatness of the rank histogram is required to retain an expert or an aggregation method as a possible candidate for operational purposes. The flatness of a rank histogram can be statistically tested thanks to the chi-square test as follows. Consider the vector of normalized deviation from flatness in each rank,

$$\boldsymbol{\delta} = \left(\frac{n_1 - n_{th}}{\sqrt{n_{th}}}, \ldots, \frac{n_k - n_{th}}{\sqrt{n_{th}}} \right)',$$

where k is the number of possible ranks, n_i is the count of rank i and $n_{th} = \frac{\sum_{i=1}^{k} n_i}{k}$ is the theoretical count in each rank for a flat histogram. Under the null hypothesis that the rank histogram is compatible with a flat histogram up to sampling noise, the squared norm $||\boldsymbol{\delta}||^2$ is the chi-square test statistic. The chi-square test statistic is insensitive to the shape of the deviations to a flat histogram, as shown in Figure 4.2. To build this figure, as in Elmore (2005), 60 integer values from 1 to 16 have been drawn from a uniform distribution. Four histograms are shown: the histogram computed with the raw sample (top left), with the same counts sorted in ascending order (top right), with the counts reassigned to have a peak-shaped histogram (bottom left), and with the counts reassigned in a wave shape (bottom right). The p-value of the chi-square test and three other flatness tests presented below is reproduced under the histograms. Although the counts of each rank are reorganized, the p-value of the χ^2-test of the four histograms is the same. Because of this, in this study, the flatness of each rank histogram is assessed with the decomposition of the chi-square test statistic, as detailed in Jolliffe and Primo (2008). Any projection of $\boldsymbol{\delta}$ onto an orthonormal basis of $\mathbb{R}^k$ has $k-1$ components whose squares are asymptotically independent χ^2 random variables, each with 1 degree of freedom. If the basis vectors are chosen to describe a sloped histogram, a convex histogram, or any other shape of interest, the existence of the shape in the rank histogram can be tested. The existence of a shape is not rejected if the projection of $\boldsymbol{\delta}$ onto the corresponding basis vector has a component statistically different from 0. Jolliffe and Primo (2008) gives formulae to compute the basis vectors for deviations from flatness commonly encountered on real data. As an example, if $k = 2p + 1$, the basis vector for the slope (resp. convexity) test is proportional to $(-p, -p+1, \ldots, -p+(k-1))$ (resp. $(p^2 - p\frac{(p+1)}{3}, (p-1)^2 - p\frac{(p+1)}{3}, \ldots, -p\frac{(p+1)}{3}, 1 - p\frac{(p+1)}{3}, \ldots, p^2 - p\frac{(p+1)}{3})$). In this study, three tests are used, the slope and convexity tests, and the "wave" test not described in Jolliffe and Primo (2008). This last test assesses the presence of a deviation from flatness in the shape of a tilde, that was frequently observed in the literature (Scheuerer et al. 2015; Baran and Lerch 2016; Taillardat et al. 2016) and in internal studies at Météo-France. The corresponding basis vector is built thanks to the Grahm-Schmidt process as follows: the vector $(0, \sin(\frac{2\pi}{k-1}), \sin(2\pi\frac{2}{k-1}), \ldots, \sin(2\pi\frac{k-2}{k-1}), 0)$ is made orthogonal to the slope basis vector, and the resulting vector is normalized to get the basis vector for testing the presence of a wave shape. In Figure 4.2, the p-values for the test of existence of a slope, a convexity or a wave are in agreement with the shape of the histograms[1]. For instance, the slope test gets the lowest p-value for the sloped rank-histogram (top right), and it does not reject the existence of a slope as expected. In the results, a histogram is deemed flat at the 0.01 significance level if its p-values for the slope, convexity and wave tests under the null hypothesis of flatness are all higher than $\frac{0.01}{3}$, the factor $\frac{1}{3}$ being the Bonferroni correction for multiple testing.

[1]The p-values of the Jolliffe-Primo test for slope and convexity have been computed with the function `TestRankhist` in the R package `SpecsVerification` (Siegert 2015). The function has been modified to compute also the p-value for the Jolliffe-Primo test for a wave shape.

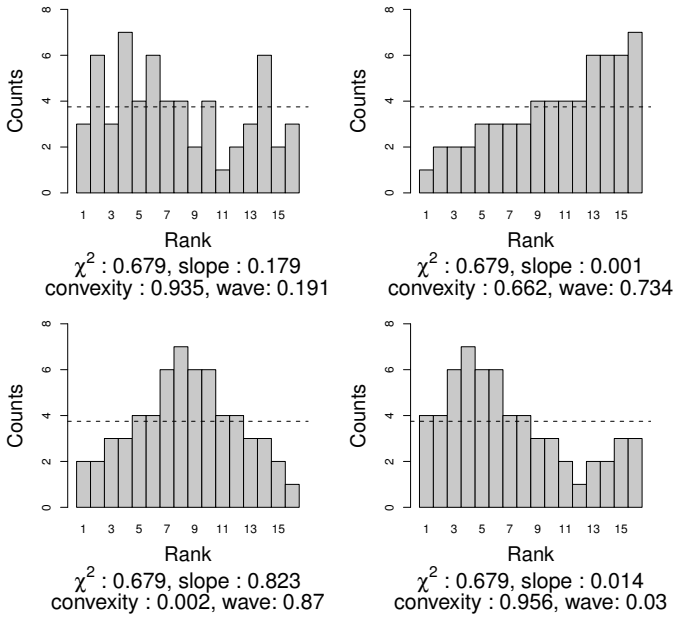

Figure 4.2 – Illustration of the Jolliffe-Primo flatness tests of a rank histogram.

The sharpness diagram proposed in Gneiting et al. (2007) is a box plot of the range of the central 50% and 90% forecast intervals (respectively noted $IQ50$ and $IQ90$) of each forecast distribution. As its name suggests, it graphically assesses the forecast concentration. Among reliable forecasts, more concentrated forecasts are preferred.

4.3 Aggregation methods

Several aggregation methods are introduced, from simple empirical ones to more sophisticated ones derived from the theory of prediction with expert advice. For all but the sharpness-calibration aggregation method, the loss function ℓ is the CRPS, so that each expert has a weight depending on its past forecast performance, in terms of CRPS.

4.3.1 Inverse CRPS weighting

The *inverse CRPS weighting method* (INV) weights each expert inversely proportional to its average CRPS over the last W days. In mathematical notations,

$$\omega_{e;t}^{INV} = \frac{\overline{CRPS}_{e;t}^{-1}(W)}{\sum_{e=1}^{E} \overline{CRPS}_{e;t}^{-1}(W)} \tag{4.4}$$

where $\overline{CRPS}_{e;t}(W)$ is the average CRPS of expert e during the W days before time t.

4.3.2 Sharpness-calibration paradigm

Gneiting et al. (2007) proposed that probabilistic forecast performance should be evaluated according to the paradigm of *maximizing the sharpness of the predictive distributions subject to calibration*. This means that a calibration method should aim at providing reliable probabilistic forecasts that are the less dispersed possible. Ideally, this would lead to a deterministic forecast (the highest sharpness possible) that would always be equal to the associated observation (*i.e.* that is reliable): the perfect forecast indeed. A more realistic and practical motivation of this sharpness-calibration paradigm is that decisions taken with this calibrated forecast would be optimal due to the reliability of the forecast and less uncertain due to the forecast's low dispersion (Zhu et al. 2002; Mylne 2002).

Here, this *sharpness-calibration paradigm* (SHARP) is used for the aggregation of experts. In words, the aggregated forecast is the forecast of the expert whose rane of the central 90% interval $IQ90$, averaged over the last W days is the lowest, among the experts whose reliability term, as computed in Hersbach (2000), is lower than a chosen threshold $Reli_{th}$ over the last W days. Roughly speaking, the aggregation weight is 1 for the sharpest reliable expert and 0 for the others. This defines an aggregated forecast computed with the weights

$$\omega_{e;t}^{SHARP} = \mathbb{1}\left(e = \underset{\{e|Reli_{e;t}(W)<Reli_{th}\}}{\arg\min} \overline{IQ90}_{e;t}(W)\right), \tag{4.5}$$

where $Reli_{e;t}(W)$ is the reliability term of expert e over the W days before time t, and $\overline{IQ90}_{e;t}(W)$ is the average range of the interval between quantiles of orders 0.95 and 0.05, forecasted by the expert e over the W days before time t. If no expert has a reliability term over the last W days lower than the reliability threshold $Reli_{th}$, the aggregated forecast is just the expert with the lowest mean CRPS over the last W days.

The three following methods have a theoretically bounded regret, whose bound are derived from the theory of prediction with expert advice.

4.3.3 Minimum CRPS

The *minimum CRPS method* (MIN) chooses the best recent expert in terms of CRPS, that is, the aggregation weight is 1 for the expert with the lowest average CRPS over the last W days, and 0 for all the other experts. The mathematical formulation is

$$\omega_{e;t}^{MIN} = \mathbb{1}(e = e_t^\star(W)), \tag{4.6}$$

where $e_t^\star(W)$ is the index of the expert with the minimum average CRPS during the last W days. The reference class $\mathcal{C}$ is the set of the E available experts, so that the oracle for this method is the expert with the lowest CRPS averaged over the period $\{1, \dots, T\}$. This aggregation method is called "follow-the-best-expert" in Cesa-Bianchi et al. (2006), which proves that, under several assumptions on the loss function, the regret of the aggregated forecast relatively to the oracle is $o(\ln(T))$.

4.3.4 Exponential weighting

The *exponentially weighted average forecaster* (EWA) computes the aggregation weights as

$$\omega_{e;t}^{EWA} = \frac{exp\{-\eta CRPS_{e;t}(W)\}}{\sum_{e=1}^{E} exp\{-\eta CRPS_{e;t}(W)\}}, \tag{4.7}$$

where $\eta \in \mathbb{R}^+$ is called the learning rate and $CRPS_{e;t}(W)$ is the cumulative CRPS of expert e over the last W days. The reference class $\mathcal{C}$ is the set of the E available experts. Thus, the oracle is the best expert, in terms of average CRPS over the whole period $\{1, \dots, T\}$. If W spans the whole period before t, that is, if $W = t - 1$ days, the EWA forecaster's regret relatively to the oracle is bounded. In Appendix 4.B, a proof, similar to the one in Cesa-Bianchi et al. (2006), is given of the following theoretical bound

$$\sup R_T^{\mathcal{C}} \leq \frac{\ln E}{\eta} + \frac{\eta T}{8} B^2, \tag{4.8}$$

where B is the upper bound of the loss function. In practice, for an unbounded loss function such as the CRPS, B is the maximum of observations and expert forecasts over the whole period $t = 1, \dots, T$.

4.3.5 Exponentiated gradient

The *exponentiated gradient forecaster* (GRAD) weights the experts with

$$\omega_{e;t}^{GRAD} = \frac{exp\{-\eta \partial_e CRPS_t^{GRAD}(W)\}}{\sum_{e=1}^{E} exp\{-\eta \partial_e CRPS_t^{GRAD}(W)\}}, \tag{4.9}$$

where $CRPS_t^{GRAD}(W)$ is the cumulative CRPS over the last W days of the GRAD forecast, and $\partial_e CRPS_t^{GRAD}(W) = \frac{\partial CRPS_t^{GRAD}(W)}{\partial \omega_{e;t}}$. Using Equation (4.13) from Appendix 4.A,

$$\begin{aligned}\partial_e CRPS_t^{GRAD}(W) &= \sum_{s=t-W}^{t-1} \frac{\partial CRPS^{GRAD}}{\partial \omega_e}(\widehat{y}_s, y_s) \\ &= \sum_{s=t-W}^{t-1} \left\{ \sum_{m_e=1}^{M_e} p_e^{m_e} |x_{e;s}^{m_e} - y_s| - \sum_{e'=1}^{E} \omega_{e';s}^{GRAD} \sum_{m_{e'}=1}^{M_{e'}} p_{e'}^{m_{e'}} x_{e';s}^{m_{e'}} \right\} \\ &\quad - \sum_{s=t-W}^{t-1} \left\{ \sum_{e'=1}^{E} \omega_{e';s}^{GRAD} \left(\sum_{m_e=1}^{M_e} \sum_{m_{e'}=1}^{M_{e'}} p_e^{m_e} p_{e'}^{m_{e'}} |x_{e;s}^{m_e} - x_{e';s}^{m_{e'}}| \right) \right\},\end{aligned} \tag{4.10}$$

which generalizes Equation (5.13) of Baudin (2015) to the case of the aggregation of ensembles with any number of members.

For this aggregation method, the reference class $\mathcal{C}$ is the set of convex combinations with constant weights over the whole period $\{1, \ldots, T\}$, that is, such that $\omega_{e;t} = \omega_e, \forall e \in \{1, \ldots, E\}$ and $t \in \{1, \ldots, T\}$. The oracle is the best, in terms of cumulative CRPS, constant convex combination of experts. This is usually a better oracle than the best expert. If $W = t - 1$ days, the following bound immediately follows from Mallet et al. (2007) or Baudin (2015),

$$\sup R_T^{\mathcal{C}} \leq \frac{\ln E}{\eta} + \frac{\eta T}{2} C^2, \tag{4.11}$$

where $C = \max_{t \in \{1,\ldots,T\}, e \in \{1,\ldots,E\}} |\frac{\partial CRPS^{GRAD}}{\partial \omega_e}(\widehat{y}_t, y_t)|$.

4.4 The experts and the observation

This section presents the $E = 28$ experts aggregated in this study, and the wind speed observation used to verify the forecasts.

4.4.1 The TIGGE data set and experts

The International Grand Global Ensemble, formerly the THORPEX Interactive Grand Global Ensemble (TIGGE) was an international project aiming, among other things, to provide ensemble prediction data from leading operational forecast centers (Bougeault et al. 2010; Swinbank et al. 2016). Although the TIGGE data set[2] includes 10 ensemble NWP

[2]The data set can be retrieved from the ECMWF at http://apps.ecmwf.int/datasets or from the Chinese Meteorological Administration at http://wisportal.cma.gov.cn/wis/

Table 4.1 – Ensembles from TIGGE used in this study, with some of their characteristics.

Weather service	Members	Used run (UTC)	Lead times
Canadian Meteorological Center (CMC)	21	1200	12h to 54h
European Center for Medium-Range Weather Forecasts (ECMWF)	51	1200	12h to 54h
Météo-France (MF)	35	1800	6h to 48h
US National Centers for Environmental Prediction (NCEP)	21	1800	6h to 48h

models, only the four ensemble models operationally available at Météo-France have been retained, due to the operational constraint of this study (see Table 4.1).

The TIGGE grid size is 0.5° over France, for a total of 267 grid points. The period goes from the 1st January, 2011 to the 31st December, 2014 (so $T = 1461$ in the notations of Section 4.2). The lead times go from 12 h to 54 h depending on the ensemble, with an interval of 6 h.

Each ensemble is an expert whose forecast CDF $\widehat{F}_{e;t}$ is the empirical CDF of the members associated with the same weight $p_e^{m_e} = \frac{1}{M}$, where M is the number of members in the ensemble.

4.4.2 The calibrated experts

Each ensemble is calibrated with two kinds of EMOS: quantile random forest (QRF, Meinshausen 2006, Taillardat et al. 2016) and non-homogeneous regression (NR, Gneiting et al. 2005, Hemri et al. 2014).

In QRF, the regression equation is built in a similar way as with random forest (see Zamo et al. 2016). A forecast CDF is then produced by going down the forest with the vector of explanatory variables, computing the step-wise CDF of the observations associated to each leave and averaging those CDFs. In practice, one requires a set of quantile orders and get the corresponding quantiles[3]. The forecast CDF $F_{e;t}$ is thus only partly known. The obtained quantiles may contain many ties, that must be suppressed with a proper interpolation, as explained in Chapter 3. QRF are here trained by 4-fold cross-validation and the chosen explanatory variables are the most important ones according to Taillardat et al. (2016).

With NR, the forecast CDF F_e is parametric. Following Hemri et al. (2014), the square root of the forecast wind speed $\widehat{f}_t$ is supposed to follow a normal distribution truncated at

[3]In R package `quantregForest` (R Core Team 2015; Meinshausen and Schiesser 2015).

Table 4.2 – Summary description of the EMOS methods used to calibrate each ensemble.

Quantile Regression Forest (QRF)	
Distribution	Non-parametric (set of quantiles).
Explanatory variables	Control member, ensemble mean, ensemble 0.1 and 0.9 quantiles, month.
Training method	4-fold cross-validation (3 training years, 1 test year).
Orders of the forecast quantiles	$0, \frac{1}{100}, \ldots, \frac{99}{100}, 1$.

Non-homogeneous Regression (NR)	
Distribution	Parametric (truncated normal distribution for the square-root of wind speed).
Explanatory variables	Mean and standard deviation of the raw ensemble.
Training method	Sliding window over the W_{tr} previous days, with $W_{tr} = 7, 30, 90, 365, t-1$ days.
Orders of the forecast quantiles	$0, \frac{1}{100}, \ldots, \frac{99}{100}, 0.999$.

0

$$\sqrt{\widehat{f}_t} \sim \mathcal{N}^0(a + b\overline{x}_t, c^2 + d^2 sd_t)$$

where $\overline{x}_t$ and sd_t are the mean and standard deviation of the associated ensemble, forecasted at time t. The real parameters a, b, c and d are optimized by maximizing the log-likelihood[4] over the last W_{tr} forecast days. To produce a step-wise CDF $\widehat{F}_{e;t}$ that may be aggregated in the framework of step-wise CDF aggregation, the quantiles of orders $\{0, \frac{1}{100}, \ldots, \frac{99}{100}, 0.999\}$ are computed from the parametric CDFs produced by NR, and squared. The last order is not 1 to avoid infinite values.

Table 4.2 gives the different values of the tuning parameters for both calibration methods.

4.4.3 The observation

The observation is the 10 m average wind speed analysis built in Zamo et al. (2016), for the 267 TIGGE grid points over France. Since each of these grid points is located at the same coordinates as one AROME grid point, no interpolation from the analysis grid to the TIGGE grid is required.

The calibration and the aggregation are produced and verified separately for eight lead times h (from 6 h to 48 h, with a time step of 6 h), and at each grid point. For computation of the aggregation weights, the supposed time of the day t is 1800 UTC, which implies that for experts based on CMC and ECMWF, whose runtime is 1200 UTC, the actual lead time is $h + 6$.

[4]With the function `optim` in R (R Core Team 2015).

Table 4.3 – Tried values for the parameters of the aggregation methods.

Aggregation method	Parameters
Minimum CRPS or Inverse CRPS	$W = 7, 15, 30, 90, 365, t-1$ days
Sharpness-calibration	$W = 7, 15, 30, 90, 365, t-1$ days $Reli_{th} = 0.1$ m/s
Exponentiated weighting or Exponentiated gradient weighting	$W = 7, 15, 30, 90, 365, t-1$ days $\eta = 10^{-1.5}, 10^{-1}, \ldots, 10^{2}$

4.5 Results

The five aggregation methods presented in Section 4.3 have been investigated, with the values of the tuning parameters listed in Table 4.3. When two parameters exist, all combinations have been tested.

The best forecast method can be chosen according to two criteria: the minimization of the average CRPS, as is classically done, or the maximization of the proportion of rank histograms who passed the three flatness tests. Hereafter, the best calibration or aggregation method in terms of minimum CRPS (resp. maximum proportion of simultaneously passed flatness tests) is called the most skillful (resp. reliable) method. Both choices are successively analyzed in the following section.

4.5.1 CRPS, reliability and sharpness

The time series of the regret of each most skillful or reliable aggregation method of each type relatively to the most skillful expert over the four years (QRF-calibrated ECMWF ensemble) is drawn in Fig. 4.3, for lead time 24 h. Whereas the best SHARP aggregation gets a consistently higher CRPS than the most skillful expert, the other aggregation methods manage to outperform the latter at least for some part of the four years. As could be hoped due to the existence of the theoretical bound on the regret, the most skillful EWA and GRAD settings get a negative regret. But for these two aggregation methods, the most reliable setting gets a positive final regret at the end of the period, slightly different than the negative regret of the most skillful settings. In terms of averaged CRPS over all lead times, this difference is about 4 %, as shown in Table 4.4. The time series of the regret exhibits an increasing trend and a diurnal cycle with the lead times (see figures 4.14 to 4.20 in Appendix 4.C), and so does the averaged CRPS (see Table 4.4). This is a frequent evolution of performance measures with lead time. The performance gets less good with farther lead times, and decreases during the late afternoon when the wind strengthens. According to the minimization of the CRPS, the chosen forecast method would be the most skillful GRAD settings, that is $\log_{10}(\eta) = -1$ and $W = 2000$ days.

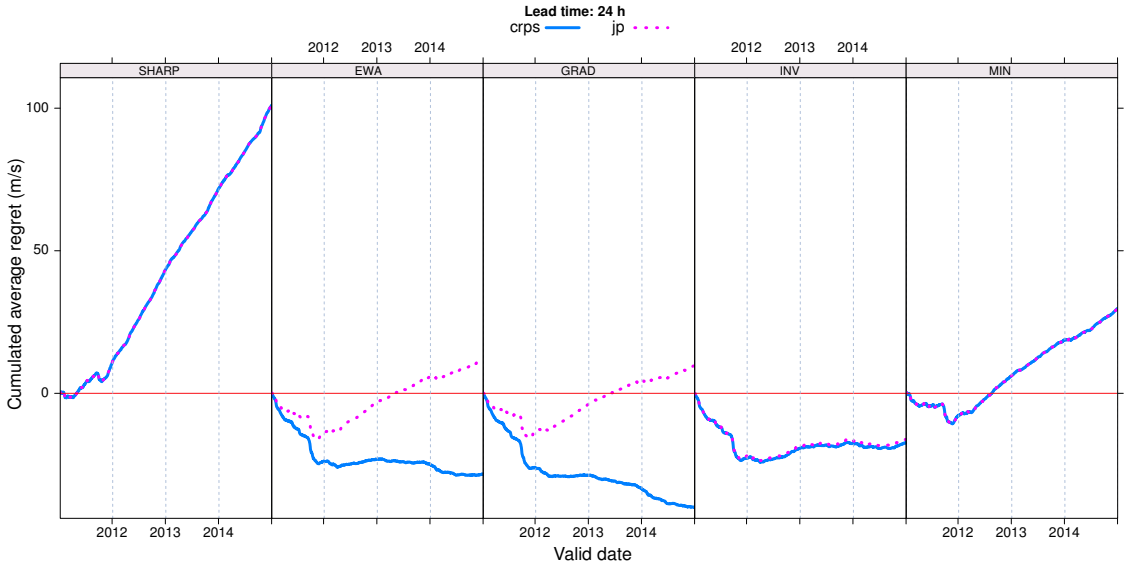

Figure 4.3 – Time series of the cumulative average regret, at lead time 24 h, for each aggregation method. At each valid date, the regret relatively to the most skillful expert (QRF-calibrated ECMWF ensemble) is computed at each grid-point, then averaged over the 267 grid-points. For each aggregation method, two settings are used to compute the regret: the most skillful one (blue continuous line) and the most reliable one (pink dashed line).

Table 4.4 – Comparison of the CRPS averaged over the four years and the 267 grid points, for the best (over all lead times) expert and aggregated forecasts. The average CRPS of the most skillful raw ensemble (ECMWF) is indicated. For each calibration or aggregation method, the average CRPS of the most skillful or the most reliable setting is indicated.

Method	Parameters		Lead time (h)								
	$\log_{10}(\eta)$	W/W_{tr}	all	6	12	18	24	30	36	42	48
RAW ECMWF			0.76	0.79	0.79	0.73	0.73	0.78	0.78	0.73	0.74
Most skillful settings.											
QRF ECMWF			0.49	0.47	0.46	0.48	0.50	0.49	0.49	0.52	0.53
SHARP		1095	0.55	0.52	0.52	0.55	0.56	0.55	0.55	0.59	0.60
GRAD	-1	2000	0.47	0.44	0.44	0.46	0.47	0.47	0.47	0.51	0.51
EWA	-1	365	0.48	0.44	0.44	0.47	0.48	0.47	0.48	0.52	0.52
INV		7	0.49	0.46	0.46	0.47	0.48	0.49	0.49	0.52	0.52
MIN		365	0.51	0.47	0.47	0.51	0.52	0.50	0.51	0.56	0.56
Most reliable settings.											
NR CMC		90	0.56	0.52	0.52	0.56	0.56	0.55	0.56	0.62	0.61
SHARP		1095	0.55	0.52	0.52	0.55	0.56	0.55	0.55	0.59	0.60
GRAD	0.5	2000	0.50	0.46	0.46	0.50	0.50	0.49	0.50	0.54	0.54
EWA	0.5	30	0.50	0.46	0.46	0.50	0.50	0.50	0.50	0.55	0.54
INV		30	0.49	0.46	0.46	0.47	0.48	0.49	0.50	0.53	0.53
MIN		365	0.51	0.47	0.47	0.51	0.52	0.50	0.51	0.56	0.56

In terms of rank histograms, as shown in Figure 4.4 (a) for lead time $h = 6$ h, the raw CMC ensemble is consistently biased with too strong forecast wind speeds in the north-west of France and too weak forecast wind speeds over the Alps and the Pyrénées. Elsewhere, although the ensemble is much less biased, the rank histogram is not deemed flat due to an obvious U-shape. The rank histograms at the other lead times and for the other raw ensembles exhibit the same features, albeit with variations in the importance of the bias and/or the convexity, as shown in Appendix 4.D. For the calibrated and aggregated forecasts, the rank histograms are computed with the nine forecast deciles. As illustrated in figures 4.4 (b) and (c), the QRF- and NR-calibrated versions of the CMC ensemble yield a higher number of flat rank histograms than the raw ensemble. The slope, convexity and wave tests simultaneously do not reject the null hypothesis of a flat rank histogram at many of the grid points. When this is not true, the rank histogram does not usually exhibit obviously rugged histograms. This result holds for the other lead times and calibrated ensembles, except when calibration is done with NR and a sliding training windows of 7 days. This last calibration produces rank histograms with a statistically significant U-shape for most of grid points and at all lead times (not shown). Finally, Figure 4.4 (d) shows that, the Jolliffe-Primo tests do not reject the flatness hypothesis at many more grid-points for

Table 4.5 – Same as Table 4.4 but for the proportion of grid points where the three flatness tests do not reject the hypothesis of a flat rank histogram.

Method	Parameters		Lead time (h)								
	$\log_{10}(\eta)$	W/W_{tr}	all	6	12	18	24	30	36	42	48
Most skillful settings.											
QRF ECMWF			0.50	0.28	0.21	0.57	0.72	0.39	0.34	0.76	0.74
SHARP		1095	0.38	0.29	0.25	0.39	0.36	0.50	0.43	0.43	0.40
GRAD	-1	2000	0.05	0.06	0.06	0.03	0.04	0.07	0.08	0.03	0.05
EWA	-1	365	0.05	0.04	0.04	0.04	0.04	0.06	0.06	0.05	0.04
INV		7	0.02	0.03	0.02	0.00	0.01	0.03	0.02	0.00	0.01
MIN		365	0.70	0.67	0.67	0.72	0.61	0.73	0.74	0.81	0.62
Most reliable settings.											
NR CMC		90	0.78	0.81	0.71	0.83	0.72	0.84	0.78	0.83	0.72
SHARP		1095	0.38	0.29	0.25	0.39	0.36	0.50	0.43	0.43	0.40
GRAD	0.5	2000	0.57	0.40	0.36	0.82	0.62	0.42	0.43	0.90	0.64
EWA	0.5	30	0.86	0.83	0.78	0.87	0.92	0.85	0.79	0.94	0.93
INV		30	0.02	0.03	0.02	0.00	0.01	0.02	0.02	0.00	0.01
MIN		365	0.70	0.67	0.67	0.72	0.61	0.73	0.74	0.81	0.62

the most reliable EWA forecast. Table 4.5 confirms quantitatively that the most reliable EWA outperforms the most reliable expert (NR-calibrated CMC with $W_{tr} = 90$ days) in terms of flatness of the rank histogram, at each lead time. However this may not be true depending on the chosen values for η and W, as illustrated in Figure 4.5. A bad choice of η and W may actually decrease the number of flat histograms compared to calibrated ensembles. For the EWA method, the most reliable settings is $\eta = 10^{0.5}$ and $W = 30$ days. All the other aggregation methods produce fewer flat rank histograms than the most reliable (and sometimes the most skillful) expert, for their respective most skillful settings and for all lead times, as shown in Table 4.5. So, in terms of flatness of the rank histograms, the best forecasts would be the most reliable EWA setting, with $\log_{10}(\eta) = 0.5$ and $W = 30$ days. This discrepancy between the two criteria for choosing the best forecasts is discussed more deeply in Section 4.6. Since a reliable forecast is a necessary condition for optimal decision making, the best retained forecast is EWA with $\log_{10}(\eta) = 0.5$ and $W = 30$ days.

The difference in sharpness of two forecasts xxx and yyy are represented in Figure 4.6 with the difference in average interquantile ranges $\overline{IQ}_\tau^{xxx} - \overline{IQ}_\tau^{yyy}$ at each grid point, with $\tau = 50, 90$. If $\overline{IQ}_\tau^{xxx} - \overline{IQ}_\tau^{yyy} > 0$ then forecast xxx is on average more dispersed than forecast yyy. The sharpness of the most reliable calibrated or aggregated forecast at each grid point is similar to or better than the average spread of the raw calibrated ensemble. Whatever the lead time is considered, the most reliable aggregated forecast is on average sharper than the most reliable calibrated expert, for more than 75% of the grid points.

Although the differences in sharpness remain low, getting sharper forecasts than the raw ensemble is noticeable. Indeed, as illustrated in Taillardat et al. (2016) or Baran and Lerch (2016), calibration tends to increase the spread of the forecast distribution, when the CRPS is minimized. Similarly, in Baran and Lerch (2016) the aggregation of two NR-calibrated versions of an ensemble resulted in a further increased dispersion compared to each calibrated version. So, getting aggregated forecasts as sharp as the calibrated experts, or as the raw ensembles, is not guaranteed by the optimization of the CRPS.

4.5.2 Spatio-temporal characteristics of the most reliable aggregated forecast

The most reliable aggregation method (EWA, $\log_{10}(\eta) = 0.5$, $W = 30$ days) exhibits a proportion of flat rank histograms very stable over the lead times and seasons, as illustrated in Figure 4.7. Only for autumn forecasts, at lead times 18 h, 24 h, 42 h and 48 h (that is, at 1200UTC and 1800UTC), the reliability is decreased. This reflects an increase of rank histograms biased toward too high forecasts, over north-western France.

Despite the point-wise calibration and aggregation, some spatio-temporal structures exists in the best aggregated forecast (EWA with $\log_{10}(\eta) = 0.5$ and $W = 30$ days). As an example, a storm that hit the north-west of France on 08 January 2011 in the morning is forecast by EWA as the green area of high medians moving along the north-west of France in the top three panels of Figure 4.8. It is in qualitative agreement with a short range deterministic forecast for the same valid dates (bottom three panels). Spatially extended spells of strong regional winds, such as Cers or Mistral in southern France, can also be forecasted as shown in figures 4.9 and 4.10 respectively.

The most reliable EWA aggregation is able to quickly redistribute the aggregation weights between the experts, as illustrated in Figure 4.11. For instance, during the middle of the year 2011, nearly all the aggregation weight shifts from the QRF-calibrated MF ensemble to the QRF-calibrated ECMWF ensemble, in a few days. The aggregation weights can also remain stable for long periods of time, such as during mi-2014, when the QRF-calibrated MF ensemble keeps a high weight for about two months. Moreover, although the raw ensembles do not perform well on their own, the EWA method may find periods where the raw ensembles can significantly contribute to the aggregated forecasts, such as in late 2012 in this example, for the raw CMC ensemble. Last, the time series of the aggregation weights is very different from one lead time to another, as shown in Appendix 4.E. These features make this aggregation method very adaptive. This may be very useful for operations when an ensemble undergoes important changes: the aggregation method will quickly detect a modification in performances and adjust its weighting accordingly.

4.6 Discussion about calibration and aggregation procedures

We now discuss more deeply the discrepancy already noticed between the choice of the best forecast according to the CRPS or the flatness of rank histograms. Although the CRPS is a natural measure of performances of forecast CDF, minimizing it does not ensure to get the highest number of grid-points with a flat histogram, as illustrated in figures 4.12 and 4.13. For instance, at lead time 36 h, the QRF-calibrated ECMWF ensemble (green dot) minimizes the CRPS of the experts, but exhibits less than 40% of flat histograms, whereas the NR-calibrated CMC ensemble (with $W = 90$ days, red downward triangle) has nearly 80% of flat histograms for a slightly higher CRPS. As for the aggregation methods, for each lead time, SHARP and MIN minimize their average CRPS while maximizing their proportion of flat rank histograms, whereas EWA and GRAD shows a minimum average CRPS for very low proportions of rank histograms. For these aggregation methods, when η gets lower, the experts get more and more similar weights. If the window W is well chosen, this allows to minimize the CRPS but reduce a lot the proportion of flat rank histograms. Actually, the point corresponding to this optimal CRPS is indicated by the graph of INV in Figure 4.13. Therefore, the average CRPS may be misleading as a main criterion to calibrate or aggregate ensemble forecasts. Instead of choosing the most skillful calibration or aggregation method, retaining the most reliable one comes with a very moderate increase in CRPS (compare the most skillful and most reliable EWA in Table 4.4), while greatly increasing the number of flat rank histograms (see Table 4.5).

The best aggregation methods and its parameters have been chosen *a posteriori* over a period of four years, and not with an on-line choice as proposed in Gerchinovitz et al. (2008); Devaine et al. (2009). In order to assess the variability of this choice, the proportion of flat rank histograms of the experts and the aggregated forecasts have been computed for each year separately. Whereas the most reliable expert varies from year to year, the most reliable aggregation is always EWA with $\eta = 10^{0.5}$ and $W = 30$ days, which is also the best set of values over the four years. The proportion of flat rank histograms is higher when computed over one year (about 90%) than over four years (about 80%). Although the actual proportion of flat rank histograms is sensitive to the sample used to compute it, it gives a stable ranking of the aggregation methods. This allows to have a fair confidence in the future performance of the aggregated forecast, even after fitting its parameters over only one year.

4.7 Conclusion and perspectives

The goal of the present study was to aggregate several ensemble forecasts (or experts) of 10 m wind speed over France. Contrary to the work of Baudin (2015) who aggregated unidentifiable experts constituted of the pooled and sorted members of several ensembles,

each expert used in the present work is a whole ensemble. The aggregation weight for the members of the same expert are also constrained to be equal. Therefore, as required by the theoretical framework of prediction with expert advice, each expert is thus identifiable over time. Some formulae of Baudin (2015) valid for step-wise CDFs with one step have been generalized to the case of step-wise CDFs with any number of steps.

Several aggregation methods to combine step-wise forecast CDFs have been presented and compared in terms of reliability, sharpness and CRPS. The reliability has been assessed by using the Jolliffe-Primo tests, who assess the presence in the rank histogram of typical deviations from a flat histogram. The systematic use of the Jolliffe-Primo flatness test highlighted that the minimization of the CRPS may not produce the maximum number of flat rank histograms, whether for the calibrated ensembles or the aggregated forecasts. It was also shown that choosing the best forecast by maximizing the proportion of rank histograms ensures reliable forecasts, without significantly increasing the CRPS.

On our data set, the best aggregation method, in terms of proportion of flat rank histograms, is the exponentially weighted average forecaster, with a learning rate $\eta = 10^{0.5}$ and an aggregation window $W = 30$ days. This aggregated forecast has a similar CRPS as the best expert in terms of CRPS, and produces many more flat rank histograms than the most reliable expert. Concerning sharpness, the best aggregated forecasts yields CDF with the same average spread as the raw ensembles, which was not expected since previous studies showed that calibration and aggregation both increase the forecast dispersion. Furthermore, the parameter value for this most reliable aggregation method is the same for all grid points and lead times, and may be found with a 1-year sample of forecast/observation pairs. It is also a very flexible method, that can produce weights with very different temporal patterns: rapidly evolving weighting of the experts, long period with constant weighting, short period with large weights for the raw ensembles. Added to the use of experts fitted over sliding windows of different size, this flexibility may help to solve a recurrent problem of calibrated experts: important changes in the NWP models that sometimes destroy the calibration property. Although the calibration and aggregation were conducted for each lead time and grid point separately, the aggregated forecast exhibits realistic spatio-temporal structures.

As for the perspectives, it is planned to study the same and other aggregation methods by pooling data in blocks of nearby grid-points. This may improve the fit or, at the very least, speed up operations on finer grids with thousands of points. Since calibration of other meteorological parameters, such as temperature and rainfall, has been already tested internally at Météo-France, aggregation methods will be tried on these parameters also. A deeper study of the discrepancy between the CRPS and the proportion of flat rank histograms as a performance criterion and its implication on calibration and aggregation constitutes a more theoretical perspective. At last, making these methods operational is a main perspective for the upcoming months.

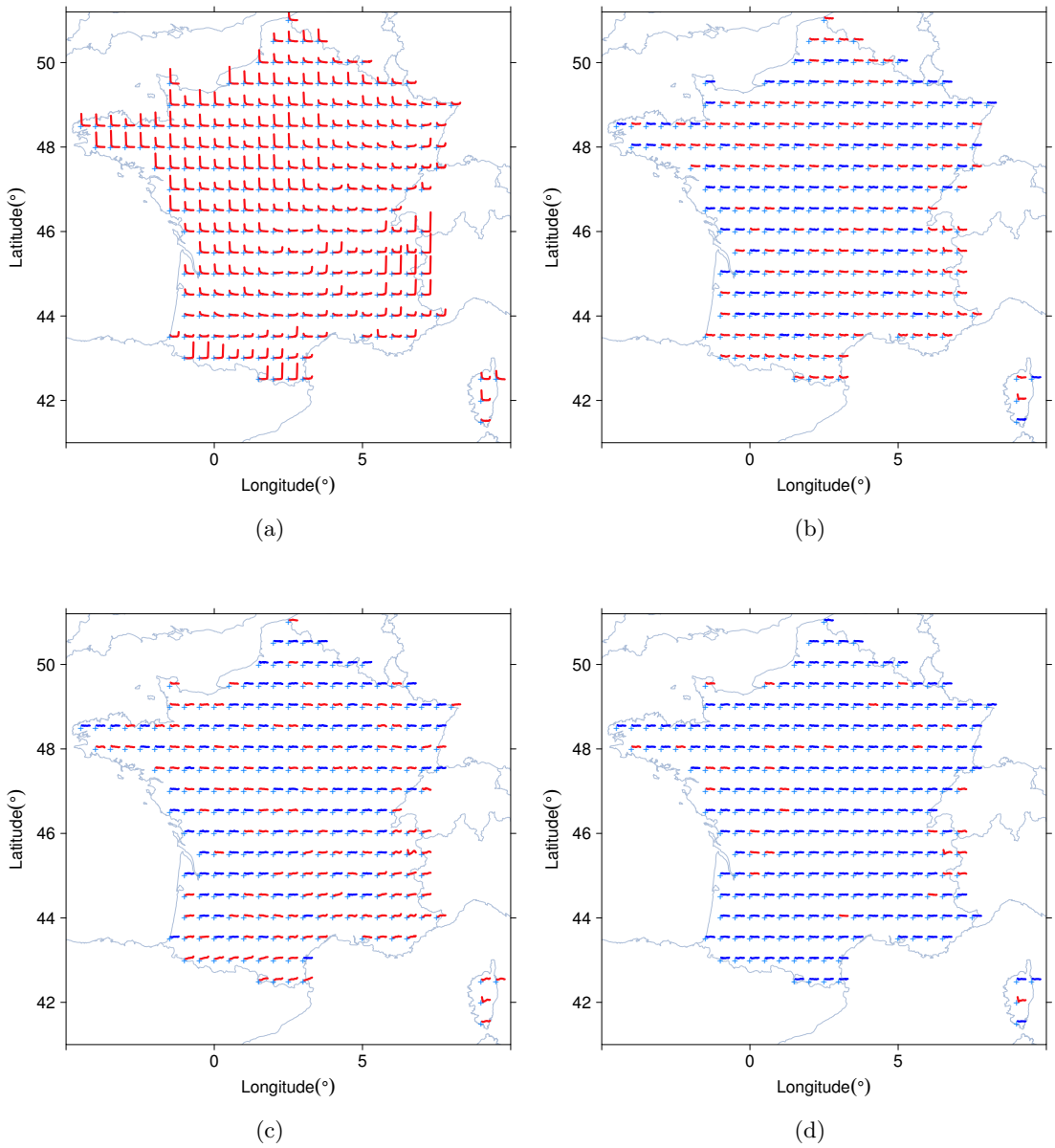

Figure 4.4 – Maps of sketches of the rank histograms computed over the 4 years, for (a) the raw CMC ensemble, the CMC ensembles calibrated with (b) QRF and (c) NR ($W_{tr} = t-1$ days), and (d) EWA (with $\eta = 10^{0.5}$ and $W = 90$ days). At each grid-point, the rank histogram is represented as a line, with the same vertical scale at all grid-points and for all maps. The lead time is $h = 6$ h. A blue line means that none of the slope, convexity and wave tests rejects the flatness hypothesis at a significance level of $\frac{0.01}{3}$, whereas a red line indicates at least one of the tests rejects the flatness hypothesis.

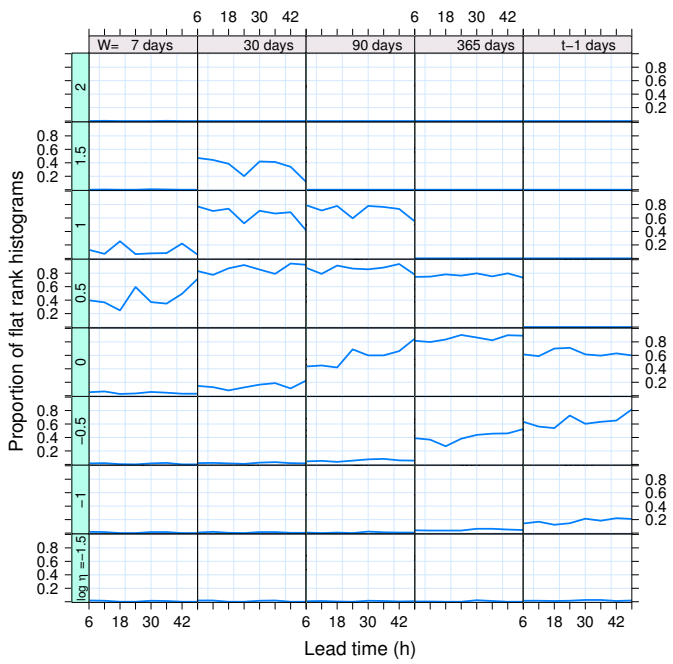

Figure 4.5 – Evolution with the lead time of the proportion of Jolliffe-Primo tests that do not reject the hypothesis of a flat rank histogram. The proportion is computed over the 267 grid points, for the different values of the parameters $(\log_{10}(\eta), W)$ in the EWA aggregation method.

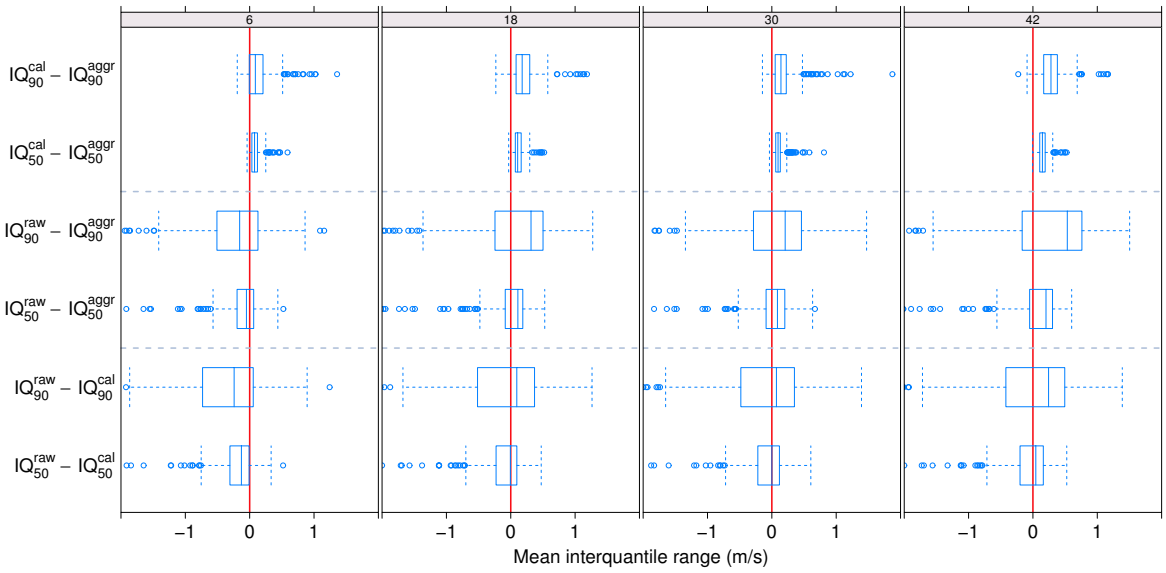

Figure 4.6 – Box plots of differences of average forecast interquantile ranges (IQ) at each grid point. The box plots are built from the interquantile ranges forecasted at each grid point, averaged over the four years of the study. IQ_{50}^{xxx} (resp. IQ_{90}^{xxx}) signifies the range between forecast quantiles of order 0.75 and 0.25 (resp. 0.95 and 0.05) of forecast xxx. The forecasts are the raw CMC ensemble (xxx = raw), the most reliable calibrated expert (xxx = cal, NR-calibrated ECMWF ensemble with $W_{tr} = 90$ days) and the most reliable aggregation method (xxx = aggr, EWA with $\log_{10}(\eta) = 0.5$ and $W = 30$ days).

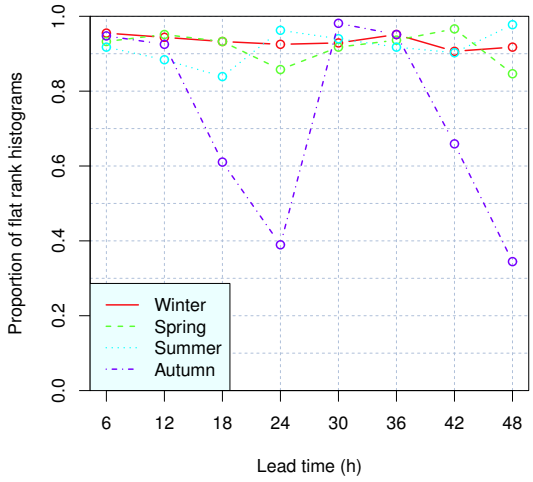

Figure 4.7 – Evolution of the proportion of flat rank histograms with the lead time and season, for the most reliable aggregation method (EWA, $\log_{10}(\eta) = 0.5$, $W = 30$ days).

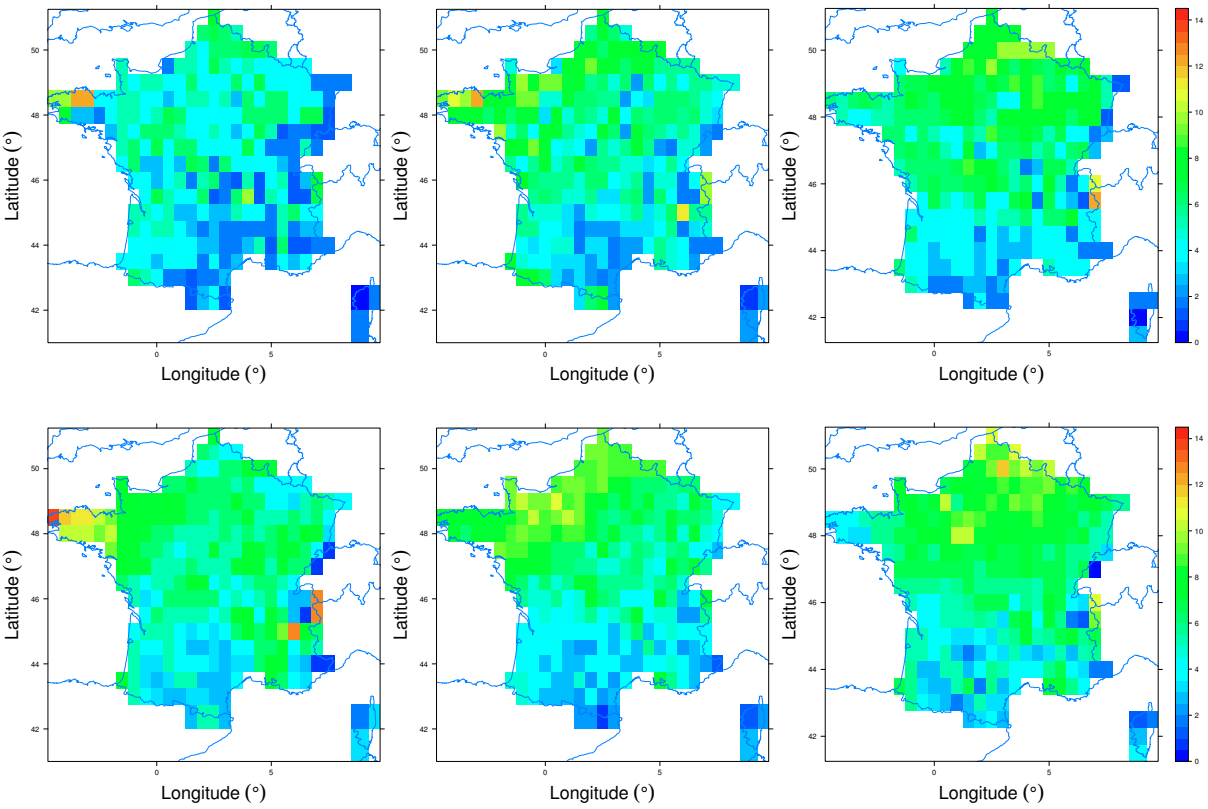

Figure 4.8 – Maps of forecast and observed wind speed for 08 January 2011, at 0000, 0600 and 1200 UTC from left to right. Top: median of wind speed forecasted on 06 January 2011 for lead times 30h, 36h and 42h by EWA with $\eta = 10^{0.5}$ and $W = 30$ days. Bottom: Corresponding observations. Wind speeds are in $\mathrm{m\,s^{-1}}$.

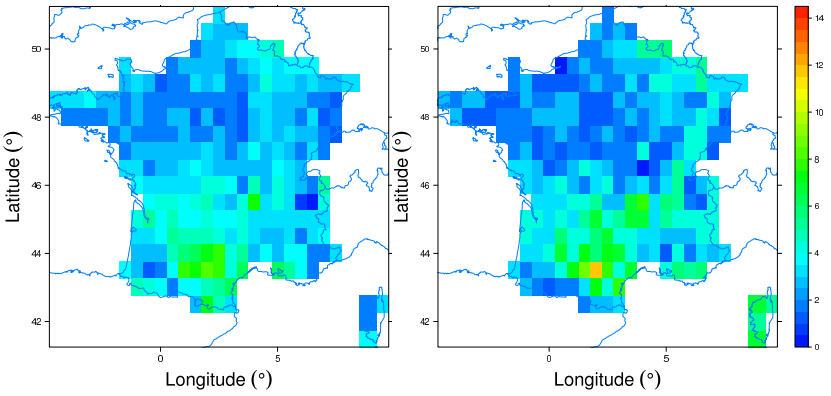

Figure 4.9 – Maps of forecast and observed wind speed for 08 March 2011, at 0000UTC. Left: median of wind speed forecasted on 06 March 2011 for lead time 30h by EWA with $\eta = 10^{0.5}$ and $W = 30$ days. Right: Corresponding observations. The Cers appears as the green zone in southern France. Wind speeds are in $\mathrm{m\,s^{-1}}$.

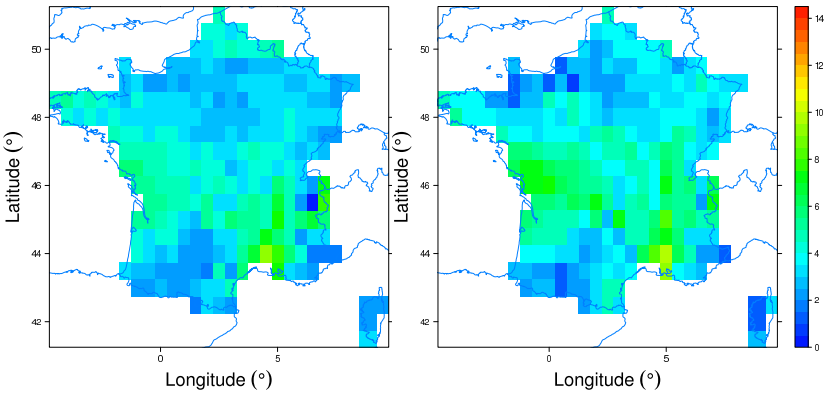

Figure 4.10 – Maps of forecast and observed wind speed for 19 January 2011, at 1200UTC. Left: median of wind speed forecasted on 17 January 2011 for lead time 18 h by EWA with $\eta = 10^{0.5}$ and $W = 30$ days. Right: Corresponding observations. The Mistral appears as the green zone in southern France. Wind speeds are in $\mathrm{m\,s^{-1}}$.

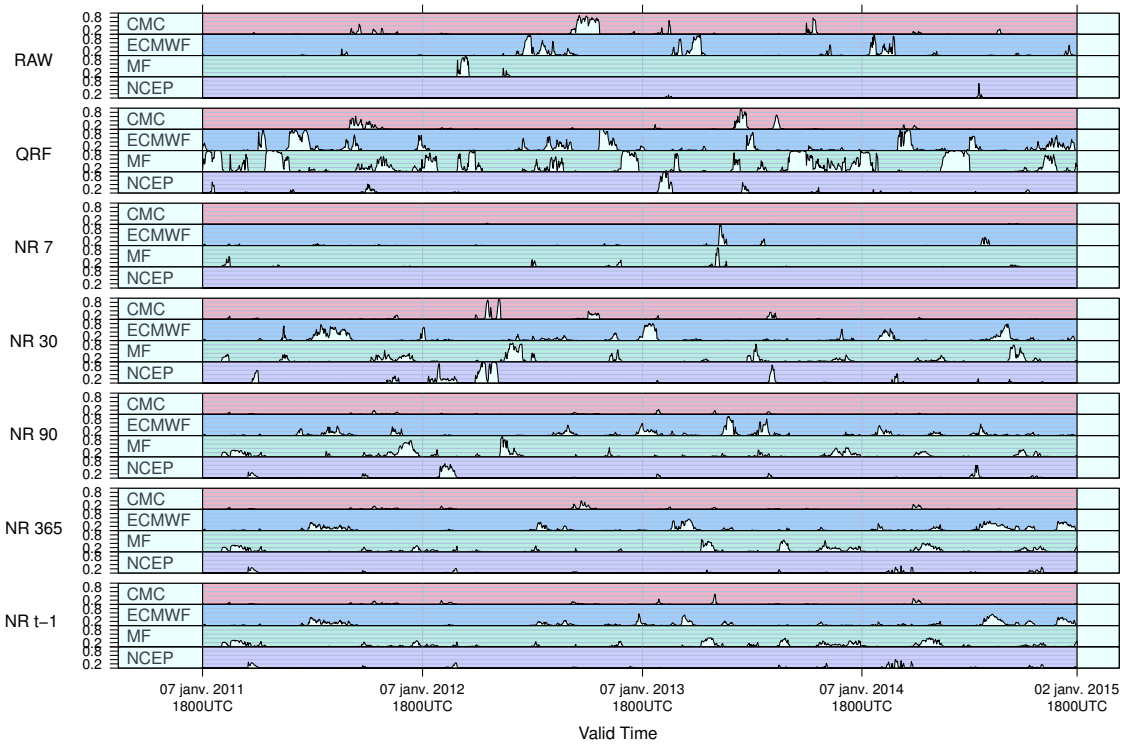

Figure 4.11 – Evolution of the aggregation weights with the valid date, for lead time 42 h. The aggregation method is the EWA forecaster with $\eta = 10^{0.5}$ and $W = 30$ days.

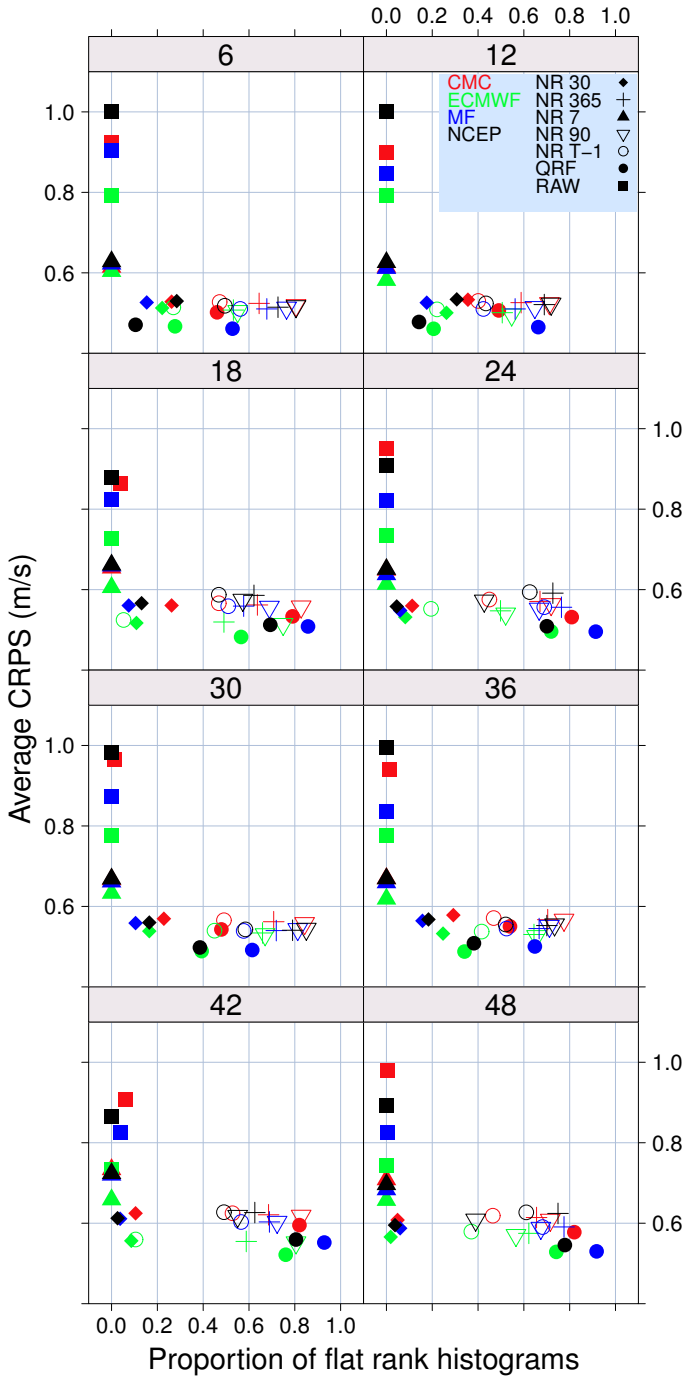

Figure 4.12 – CRPS averaged over space and time, versus the proportion of rank histograms deemed flat by the slope, convexity and wave tests, for each expert, by lead time.

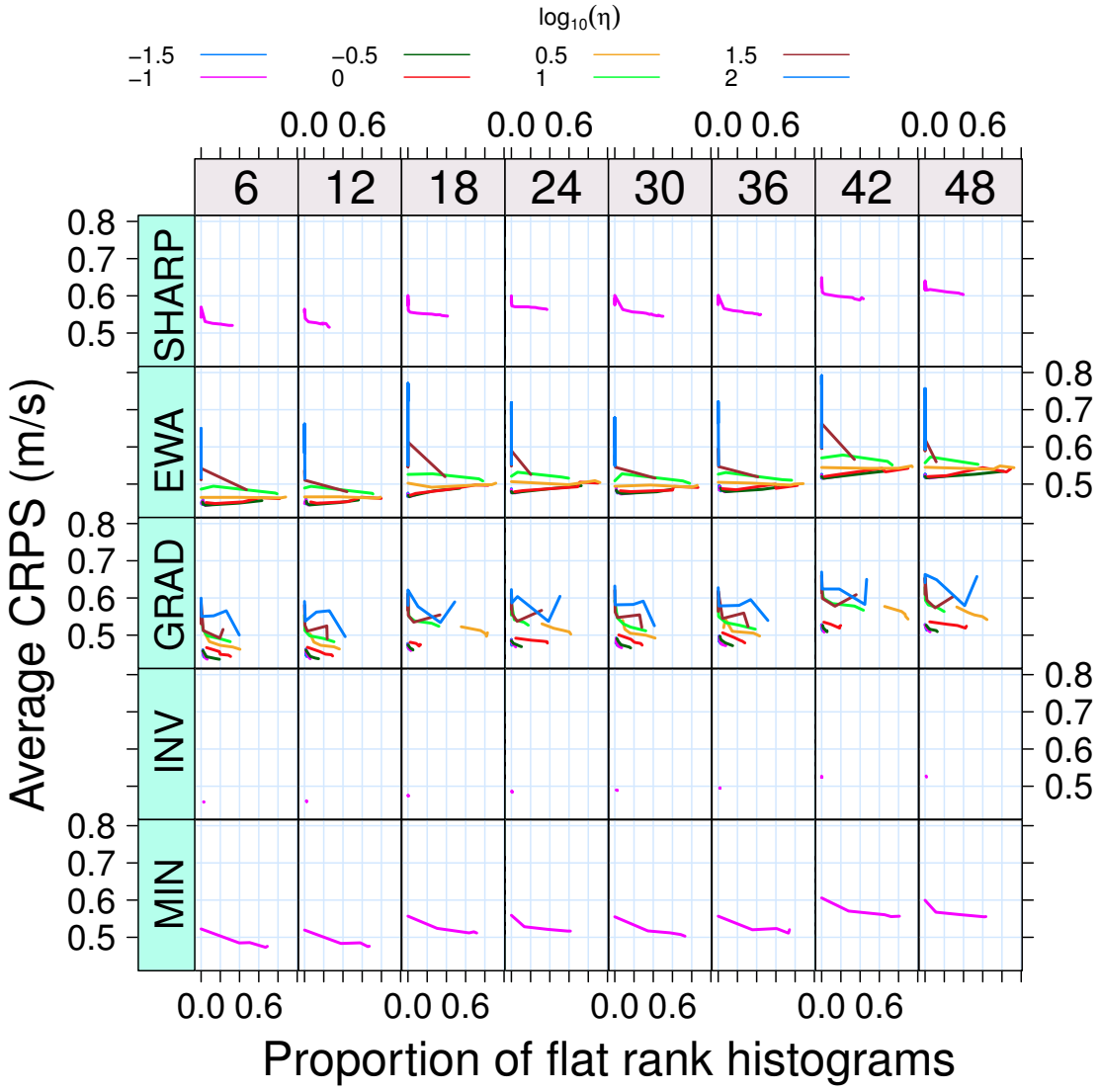

Figure 4.13 – CRPS averaged over space and time, versus the proportion of rank histograms deemed flat by the slope, convexity and wave tests, for each aggregation method, by lead time. For EWA and GRAD, each colored line corresponds to a fixed value of η.

Appendix

4.A Formula for the gradient of the CRPS

Baudin (2015) considers the aggregation of step-wise CDFs with one single step ($M_e = 1 \quad \forall e \in \{1, \dots, E\}$). This appendix generalizes equations (5.10) and (5.13), of Baudin (2015) for, respectively, the CRPS, and gradient thereof, to an aggregation of step-wise CDFs with any number of steps.

Dropping the time index t in the notations, the aggregated CDF at time t is

$$\widehat{y}(x) = \sum_{e=1}^{E} \omega_e \left[\sum_{m_e=1}^{M_e} p_e^{m_e} H_{x_e^{m_e}}(x) \right],$$

with the notation $H_a(x) = H(x - a)$.

Therefore, the CRPS of the aggregated CDF at time t is

$$\begin{aligned} CRPS(\widehat{y}, y) &= \int_{\mathbb{R}} \left\{ H_y(x) - \sum_{e=1}^{E} \omega_e \left[\sum_{m_e=1}^{M_e} p_e^{m_e} H_{x_e^{m_e}}(x) \right] \right\}^2 dx \\ &= \int_{\gamma}^{\Gamma} \left\{ H_y(x) - \sum_{e=1}^{E} \omega_e \left[\sum_{m_e=1}^{M_e} p_e^{m_e} H_{x_e^{m_e}}(x) \right] \right\}^2 dx, \end{aligned}$$

where $\gamma = \min(y, x_1^1, \dots, x_E^{M_E})$ and $\Gamma = \max(y, x_1^1, \dots, x_E^{M_E})$.

By developing the square inside the integral,

$$\begin{aligned} CRPS(\widehat{y}, y) = &\int_{\gamma}^{\Gamma} H_y(x) dx \\ &- 2 \int_{\gamma}^{\Gamma} \sum_{e=1}^{E} \omega_e \left[\sum_{m_e=1}^{M_e} p_e^{m_e} H_{x_e^{m_e}}(x) H_y(x) \right] dx \\ &+ \int_{\gamma}^{\Gamma} \left\{ \sum_{e=1}^{E} \omega_e \left[\sum_{m_e=1}^{M_e} p_e^{m_e} H_{x_e^{m_e}}(x) \right] \right\} \left\{ \sum_{e'=1}^{E'} \omega_{e'} \left[\sum_{m_{e'}=1}^{M_{e'}} p_e^{m_e} H_{x_{e'}^{m_{e'}}}(x) \right] \right\} dx. \end{aligned}$$

Noting that $H_a(x)H_b(x) = H_{\max(a,b)}(x)$, and $\int_\gamma^\Gamma H_a(x)dx = \Gamma - a \quad \forall a \in [\gamma;\Gamma]$, then

$$\begin{aligned}
CRPS(\widehat{y}, y) =& \Gamma - y \\
&- 2\sum_{e=1}^{E} \omega_e \left\{ \sum_{m_e=1}^{M_e} p_e^{m_e} [\Gamma - \max(x_e^{m_e}, y)] \right\} \\
&+ \sum_{e,e'=1}^{E} \omega_e \omega_{e'} \left\{ \sum_{m_e=1}^{M_e} \sum_{m_{e'}=1}^{M_{e'}} p_e^{m_e} p_{e'}^{m_{e'}} \left[\Gamma - \max(x_e^{m_e}, x_{e'}^{m_{e'}})\right] \right\} \\
=& - y \\
&+ 2\sum_{e=1}^{E} \omega_e \left[\sum_{m_e=1}^{M_e} p_e^{m_e} \max(x_e^{m_e}, y) \right] \\
&- \sum_{e,e'=1}^{E} \omega_e \omega_{e'} \left\{ \sum_{m_e=1}^{M_e} \sum_{m_{e'}=1}^{M_{e'}} p_e^{m_e} p_{e'}^{m_{e'}} \max(x_e^{m_e}, x_{e'}^{m_{e'}}) \right\},
\end{aligned}$$

because $\sum_{e=1}^{E} \omega_e = 1$, and $\sum_{m_e=1}^{M_e} p_e^{m_e} = 1 \quad \forall e \in \{1, \ldots, E\}$.

Since $\max(a, b) = \frac{1}{2}(a + b + |a - b|)$,

$$\begin{aligned}
CRPS(\widehat{y}, y) =& - y \\
&+ \sum_{e=1}^{E} \omega_e \left[\sum_{m_e=1}^{M_e} p_e^{m_e} (x_e^{m_e} + y + |x_e^{m_e} - y|) \right] \\
&- \frac{1}{2} \sum_{e,e'=1}^{E} \omega_e \omega_{e'} \left\{ \sum_{m_e=1}^{M_e} \sum_{m_{e'}=1}^{M_{e'}} p_e^{m_e} p_{e'}^{m_{e'}} (x_e^{m_e} + x_{e'}^{m_{e'}} + |x_e^{m_e} - x_{e'}^{m_{e'}}|) \right\} \\
=& \sum_{e=1}^{E} \omega_e \left[\sum_{m_e=1}^{M_e} p_e^{m_e} (x_e^{m_e} + |x_e^{m_e} - y|) \right] \\
&- \frac{1}{2} \sum_{e,e'=1}^{E} \omega_e \omega_{e'} \left\{ \sum_{m_e=1}^{M_e} \sum_{m_{e'}=1}^{M_{e'}} p_e^{m_e} p_{e'}^{m_{e'}} (x_e^{m_e} + x_{e'}^{m_{e'}} + |x_e^{m_e} - x_{e'}^{m_{e'}}|) \right\}.
\end{aligned} \tag{4.12}$$

The derivation with respect to ω_e results in

$$\frac{\partial CRPS}{\partial \omega_e}(\widehat{y}, y) = \sum_{m_e=1}^{M_e} p_e^{m_e}(x_e^{m_e} + |x_e^{m_e} - y|) - \sum_{e'=1}^{E} \omega_{e'} \left\{ \sum_{m_e=1}^{M_e} \sum_{m_{e'}=1}^{M_{e'}} p_e^{m_e} p_{e'}^{m_{e'}} (x_e^{m_e} + x_{e'}^{m_{e'}} + |x_e^{m_e} - x_{e'}^{m_{e'}}|) \right\}.$$

Finally, recalling that $\sum_{e=1}^{E} \omega_e = 1$, and $\sum_{m_e=1}^{M_e} p_e^{m_e} = 1 \quad \forall e \in \{1, \dots, E\}$

$$\frac{\partial CRPS}{\partial \omega_e}(\widehat{y}, y) = \sum_{m_e=1}^{M_e} p_e^{m_e} |x_e^{m_e} - y| - \sum_{e'=1}^{E} \omega_{e'} \sum_{m_{e'}=1}^{M_{e'}} p_{e'}^{m_{e'}} x_{e'}^{m_{e'}} - \sum_{e'=1}^{E} \omega_{e'} \left\{ \sum_{m_e=1}^{M_e} \sum_{m_{e'}=1}^{M_{e'}} p_e^{m_e} p_{e'}^{m_{e'}} |x_e^{m_e} - x_{e'}^{m_{e'}}| \right\}. \tag{4.13}$$

Formulae (4.12) and (4.13) generalize equations (5.10) and (5.13), respectively, of Baudin (2015).

4.B Proof of the theoretical bounds for the regret of the exponentially weighted average forecaster

The proof of Equation (4.8) closely follows the proof of theorem 2.2 in Cesa-Bianchi et al. (2006).

Let $\ell : \widehat{\mathcal{Y}} \times \mathcal{Y} \to [a; b]$ be a real-valued, bounded loss function. ℓ is supposed convex in its first argument.

The EWA weights at time t are computed as

$$\omega_{e;t}^{EWA} = \frac{exp\{-\eta L_{e;t}\}}{\sum_{e=1}^{E} exp\{-\eta L_{e;t}\}},$$

with $L_{e;t} = \sum_{s=1}^{t-1} \ell(\widehat{y}_{e;s}, y_s)$ the cumulative loss of expert e at time t, and with the convention $L_{e;1} = 0$ so that $\omega_{e,1}^{EWA} = \frac{1}{E} \quad \forall e$.

Let us define $W_t = \sum_{e=1}^{E} exp\{-\eta L_{e;t}\} \forall t \geq 1$ and $W_0 = E$. At all times $t = 1, \dots, T$, and

using the convention that a sum over 0 elements is 0 (for $t = 1$, such that $\omega_{e,0}^{EWA} = \frac{1}{E} \quad \forall e$),

$$\begin{aligned}\ln \frac{W_t}{W_{t-1}} &= \ln \frac{\sum_{e=1}^{E} exp\{-\eta \ell(\widehat{y}_{e;t}, y_t)\} exp\{-\eta L_{e;t-1}\}}{\sum_{e'=1}^{E} exp\{-\eta L_{e';t-1}\}} \\ &= \ln \frac{\sum_{e=1}^{E} \omega_{e;t-1}^{EWA} exp\{-\eta \ell(\widehat{y}_{e;t}, y_t)\}}{\sum_{e'=1}^{E} \omega_{e';t-1}^{EWA}}. \end{aligned} \tag{4.14}$$

The proof now needs Hoeffding's inequality (Hoeffding 1963). Let $a, b \in \mathbb{R}$ with $a < b$. Let Z be a bounded random variable with values in $[a; b]$, then, $\forall s \in \mathbb{R}$, Hoeffding's inequality states that

$$\ln \mathbb{E}\left[e^{sZ}\right] \leq s\mathbb{E}[Z] + \frac{s^2}{8}(b-a)^2.$$

Using Equation (4.14) and Hoeffding's inequality for the random variable Z taking the values $\ell(\widehat{y}_{e;t}, y_t)$ with discrete probability $\omega_{e;t-1}^{EWA}$, taking $s = -\eta$ and summing over $t = 1, \ldots, T$ leads to

$$\begin{aligned}\ln \frac{W_T}{W_0} &\leq -\eta \sum_{t=1}^{T} \sum_{e=1}^{E} \omega_{e;t}^{EWA} \ell(\widehat{y}_{e;t}, y_t) + \frac{\eta^2}{8}(b-a)^2 T \\ &\leq -\eta \sum_{t=1}^{T} \ell\left(\sum_{e=1}^{E} \omega_{e;t}^{EWA} \widehat{y}_{e;t}, y_t\right) + \frac{\eta^2}{8}(b-a)^2 T \\ &= -\eta \sum_{t=1}^{T} \ell\left(\widehat{y}_t, y_t\right) + \frac{\eta^2}{8}(b-a)^2 T, \end{aligned}$$

after using the convexity of the loss function ℓ in its first argument, and the definition of the EWA forecast.

Noting that the following relationship also holds

$$\begin{aligned}\ln \frac{W_T}{W_0} &= \ln \left(\sum_{e=1}^{E} exp\{-\eta L_{e;T}\}\right) - \ln E \\ &\geq \ln \left(\max_{e=1,\ldots,E} exp\{-\eta L_{e;T}\}\right) - \ln E \\ &= -\eta \min_{e=1,\ldots,E} L_{e;T} - \ln E, \end{aligned}$$

and combining it with the previous relationship leads to

$$-\eta \min_{e=1,\ldots,E} L_{e;T} - \ln E \leq -\eta \sum_{t=1}^{T} \ell\left(\widehat{y}_t, y_t\right) + \frac{\eta^2}{8}(b-a)^2 T.$$

Finally, dividing by $-\eta$ results in the following bound of the regret of the aggregated forecast relatively to the best expert

$$\sum_{t=1}^{T} \ell\left(\widehat{y}_t, y_t\right) - \min_{e=1,\dots,E} L_{e;T} \leq \frac{\ln E}{\eta} + \frac{\eta}{8}(b-a)^2 T. \tag{4.15}$$

Noting that this bound for the regret holds for any bounded loss function ℓ convex in its first argument, which are properties of the CRPS, concludes the demonstration.

4.C Time series of the regret at each lead time

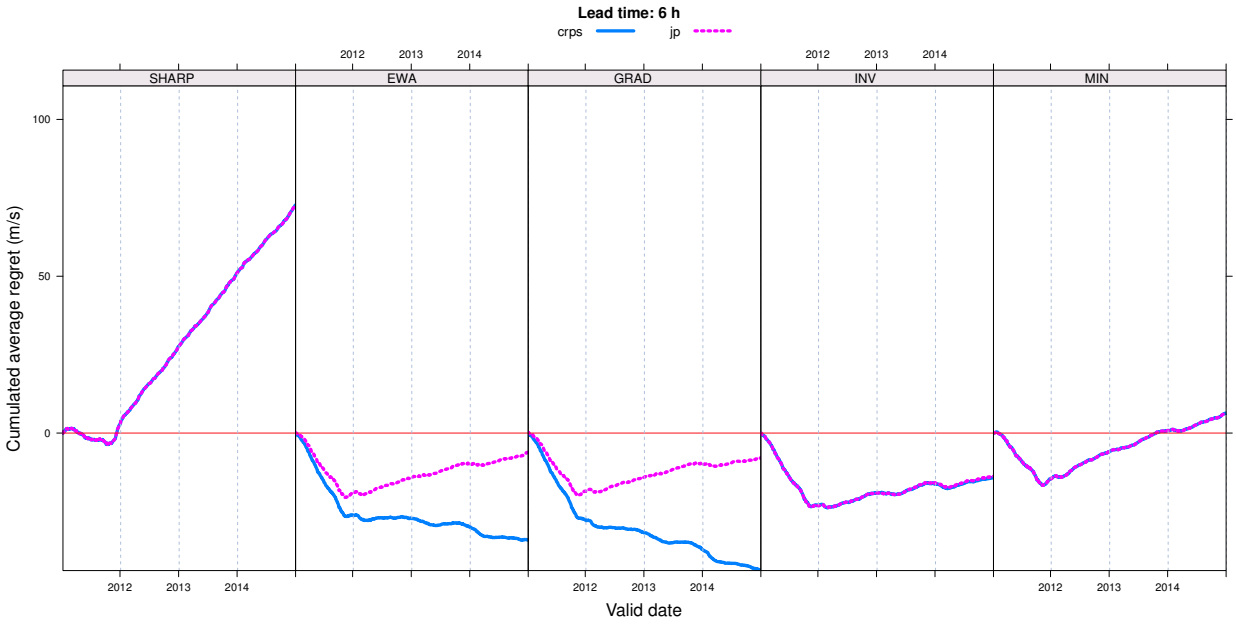

Figure 4.14 – Same as Figure 4.3, for lead time 6 h.

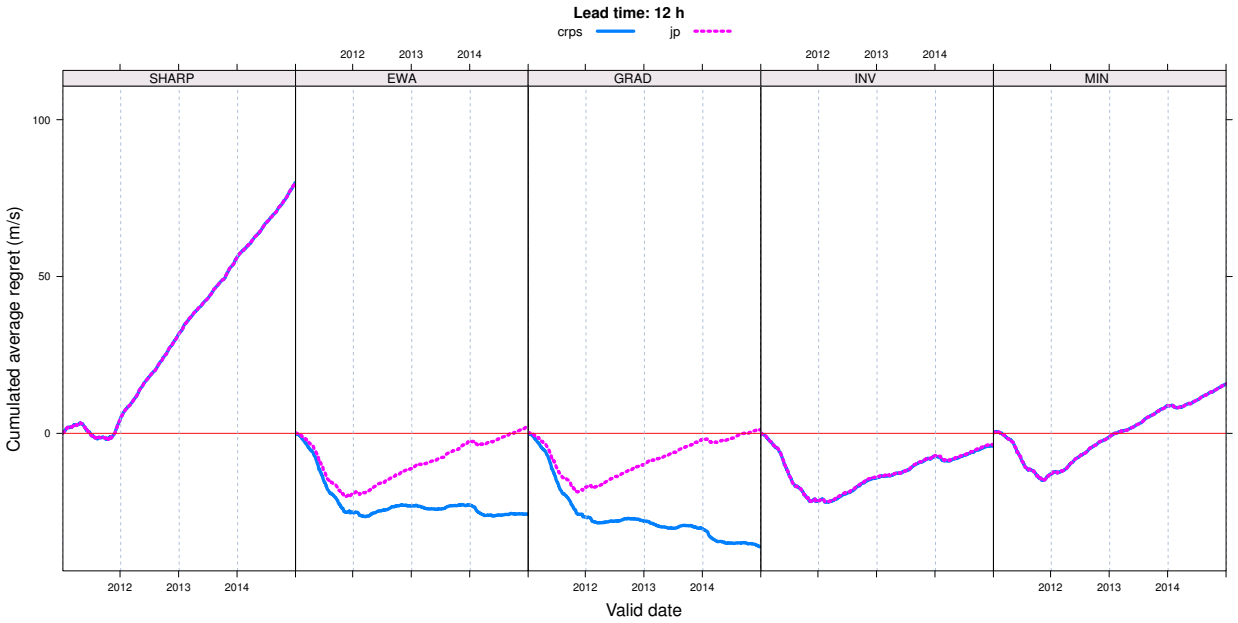

Figure 4.15 – Same as Figure 4.3, for lead time 12 h.

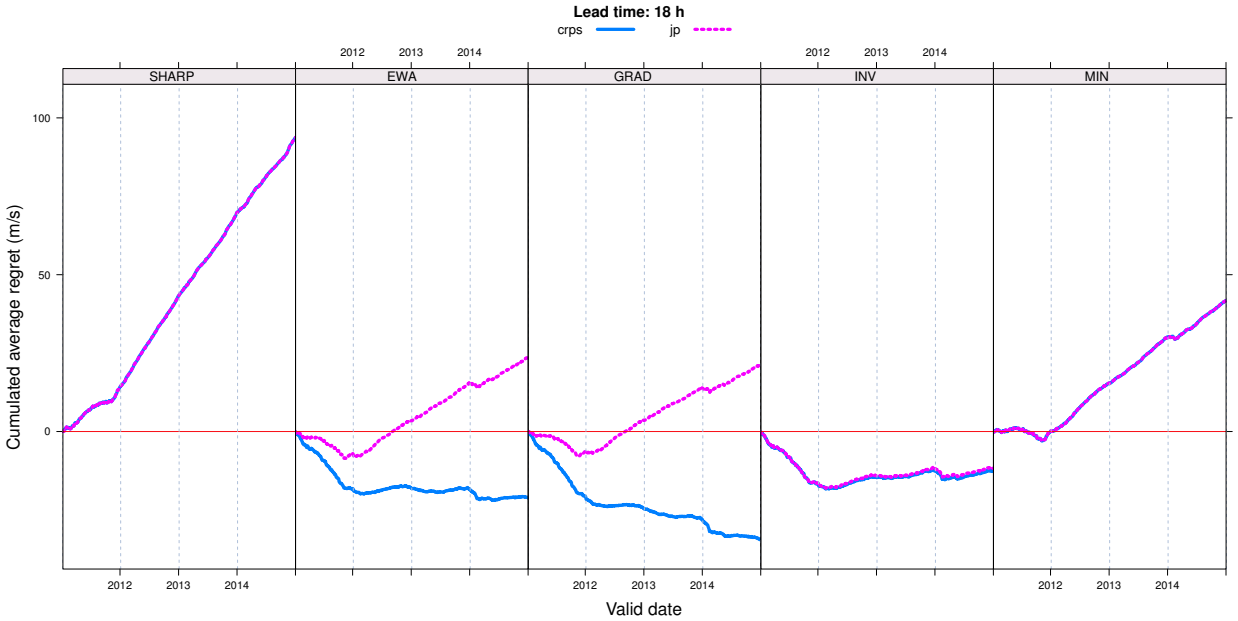

Figure 4.16 – Same as Figure 4.3, for lead time 18 h.

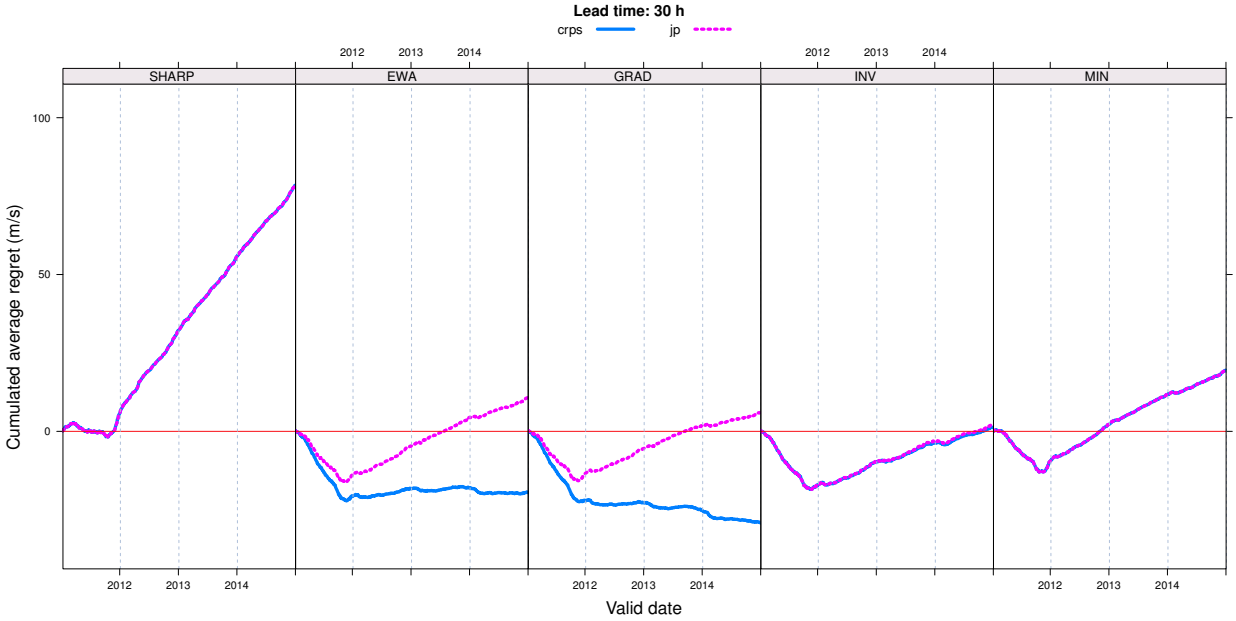

Figure 4.17 – Same as Figure 4.3, for lead time 30 h.

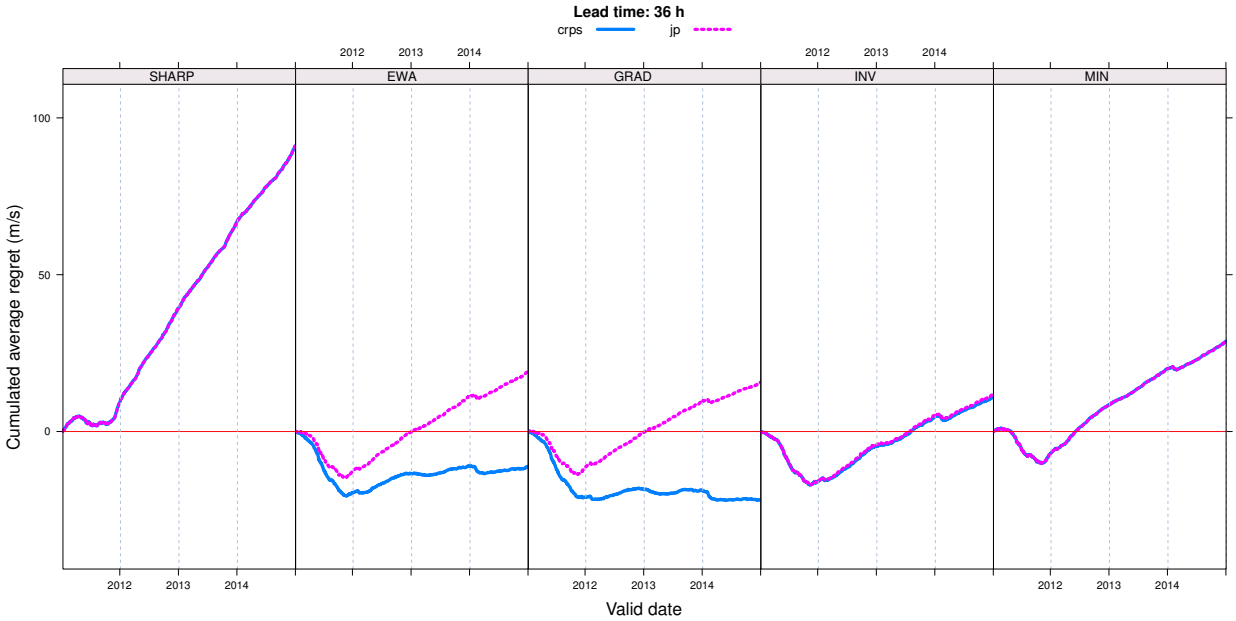

Figure 4.18 – Same as Figure 4.3, for lead time 36 h.

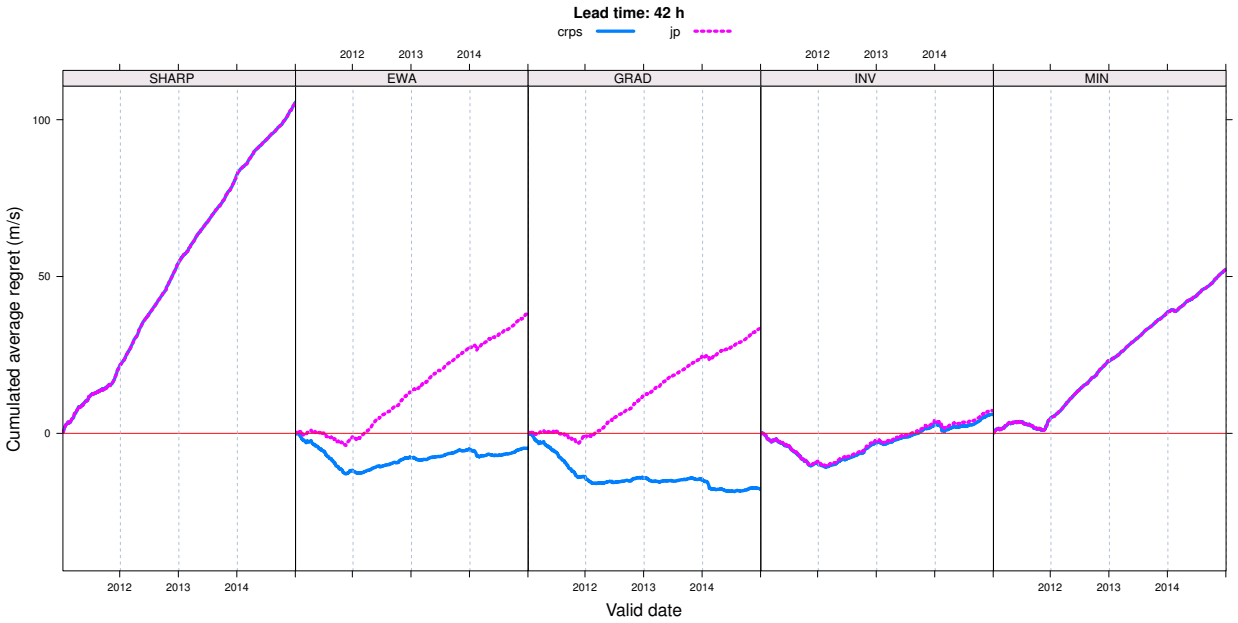

Figure 4.19 – Same as Figure 4.3, for lead time 42 h.

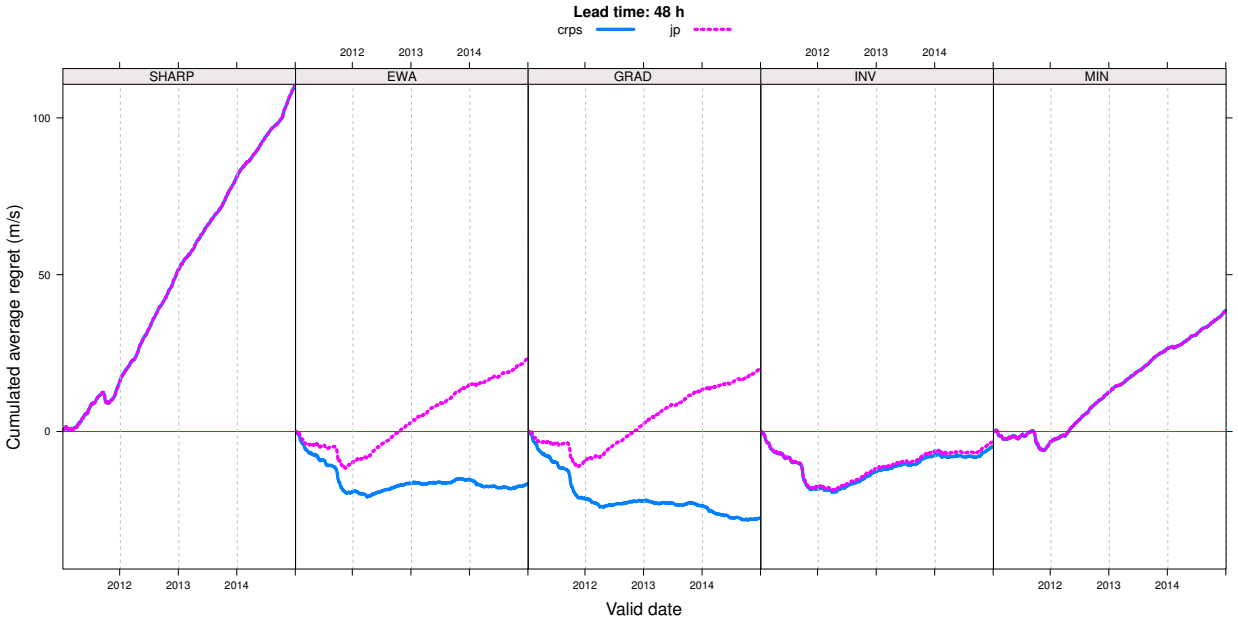

Figure 4.20 – Same as Figure 4.3, for lead time 48 h.

4.D Maps of rank histograms of raw ensembles

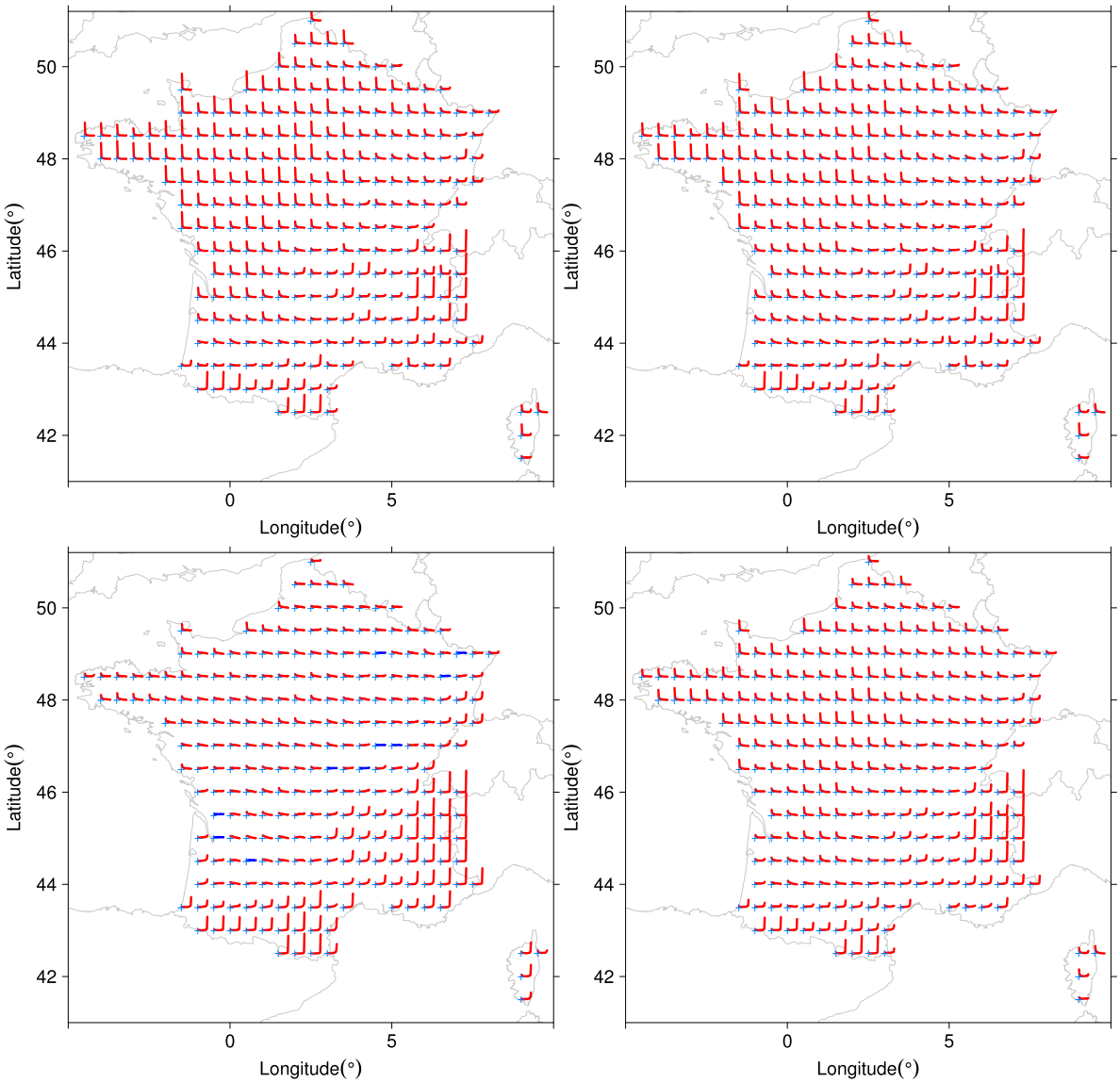

Figure 4.21 – Same as Figure 4.4 (a), but for lead times 6 h (top left), 12 h (top right), 18 h (bottom left) and 24 h (bottom right).

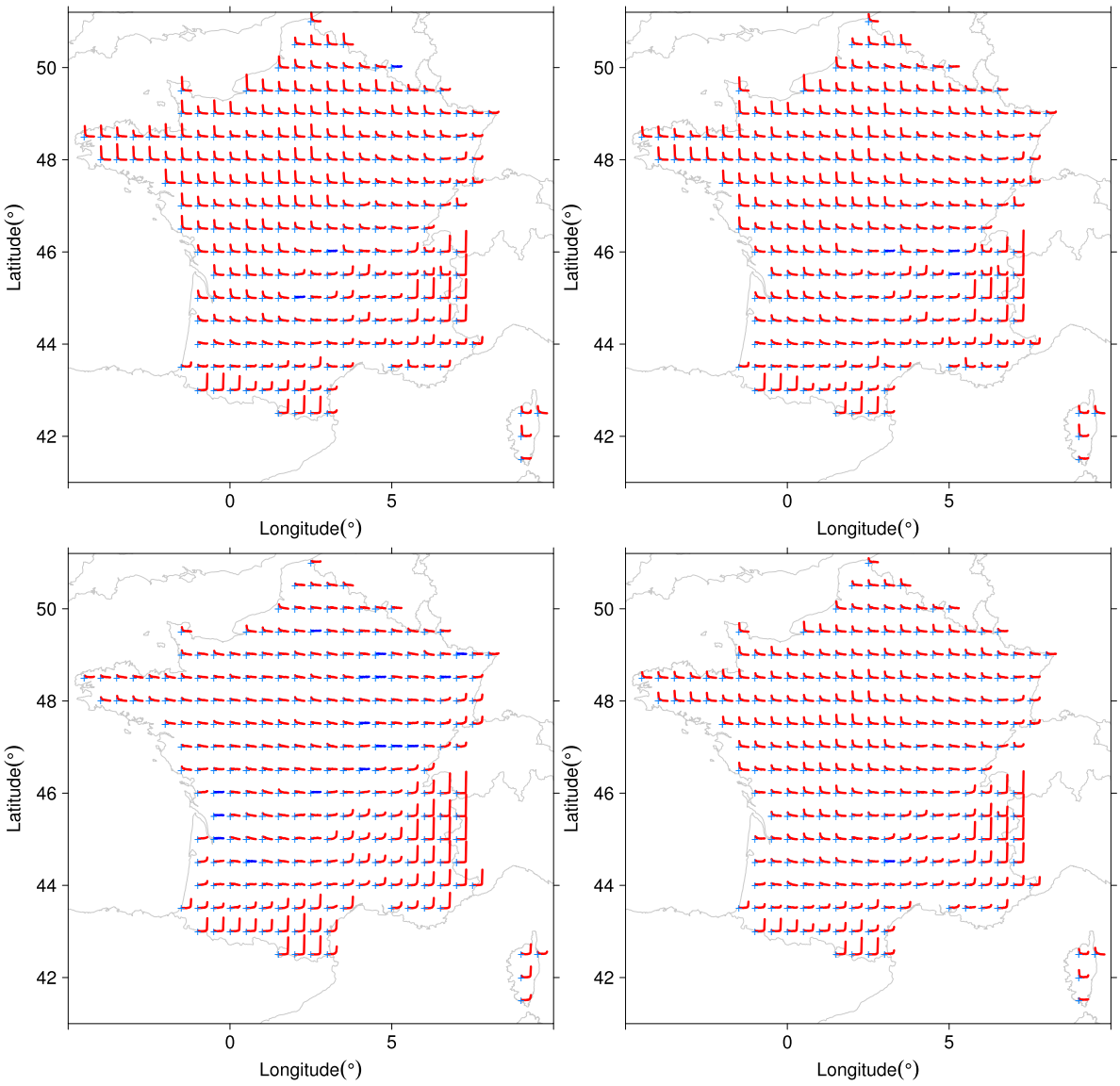

Figure 4.22 – Same as Figure 4.4 (a), but for lead times 30 h (top left), 36 h (top right), 42 h (bottom left) and 48 h (bottom right).

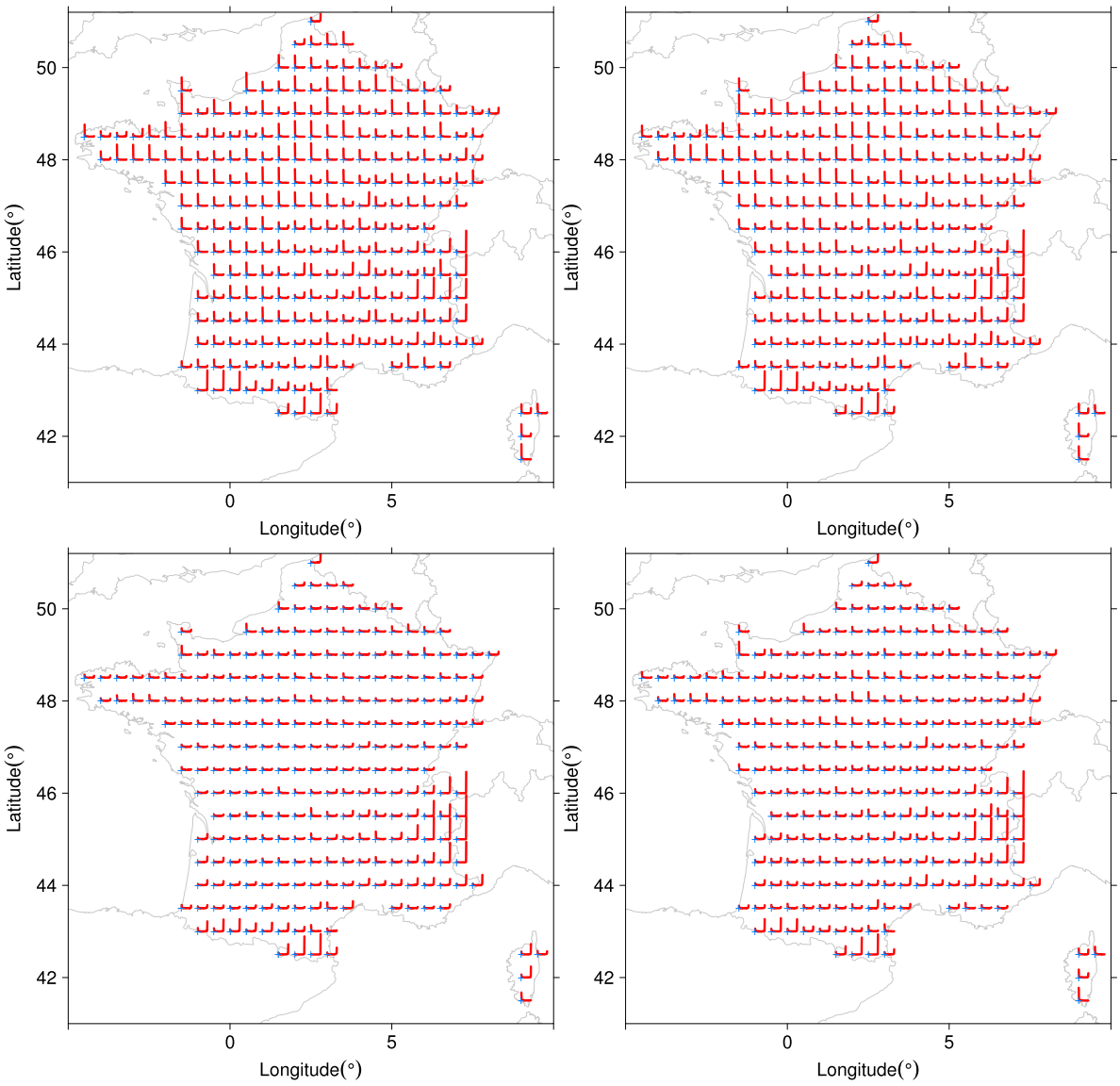

Figure 4.23 – Same as Figure 4.4 (a), but for raw ECMWF ensemble, and for lead times 6 h (top left), 12 h (top right), 18 h (bottom left) and 24 h (bottom right).

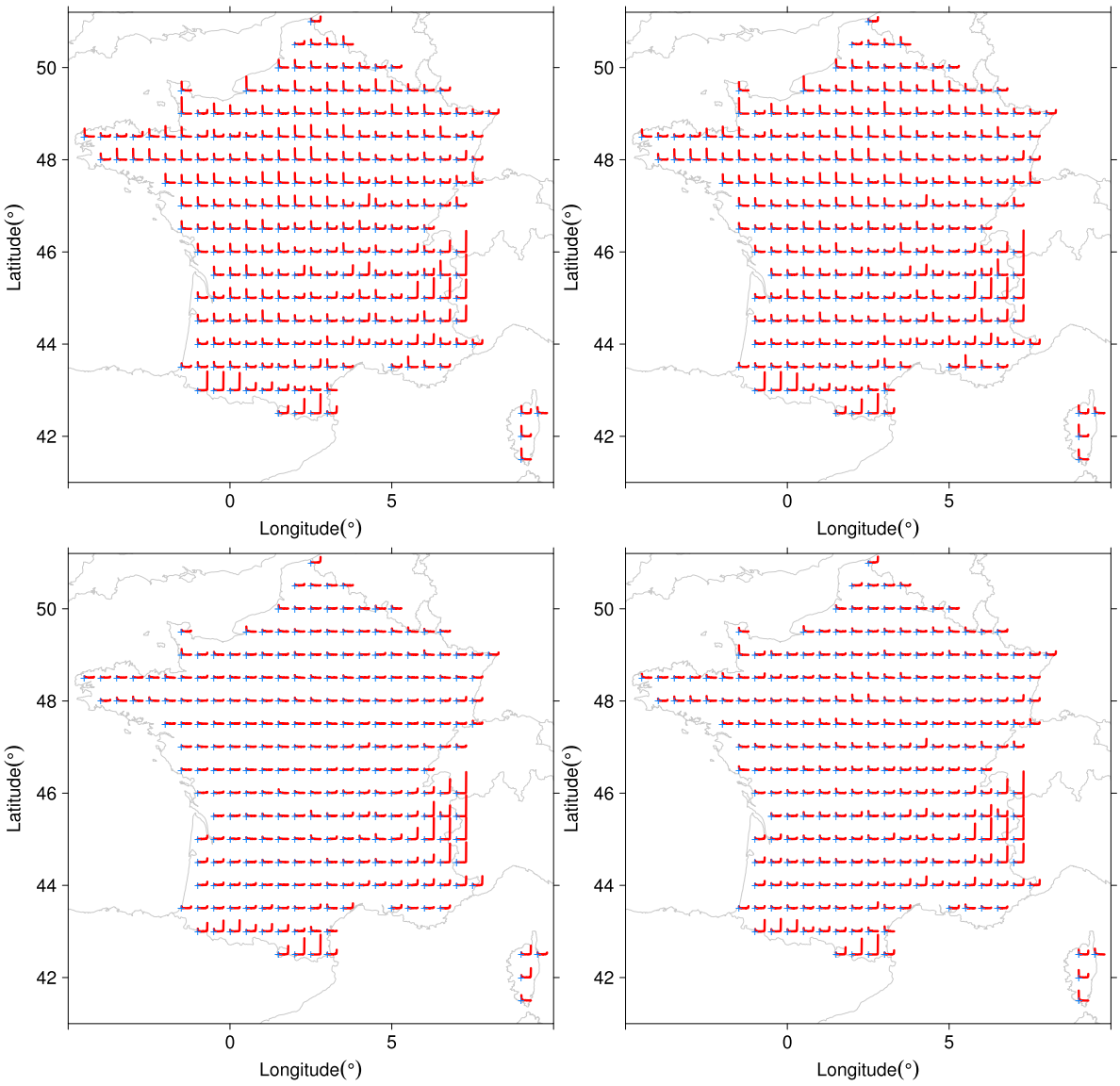

Figure 4.24 – Same as Figure 4.4 (a), but for raw ECMWF ensemble, and for lead times 30 h (top left), 36 h (top right), 42 h (bottom left) and 48 h (bottom right).

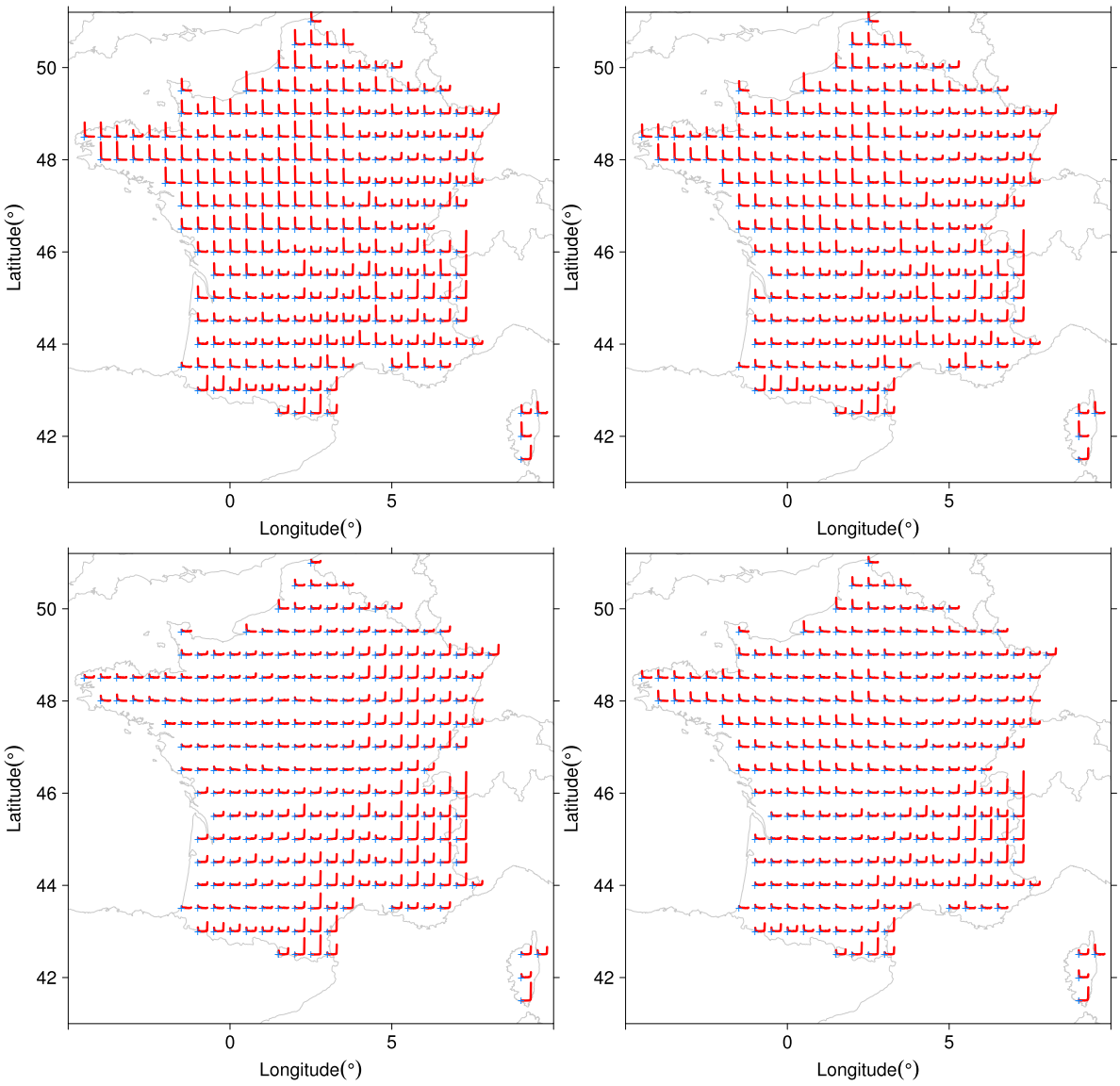

Figure 4.25 – Same as Figure 4.4 (a), but for raw MF ensemble, and for lead times 6 h (top left), 12 h (top right), 18 h (bottom left) and 24 h (bottom right).

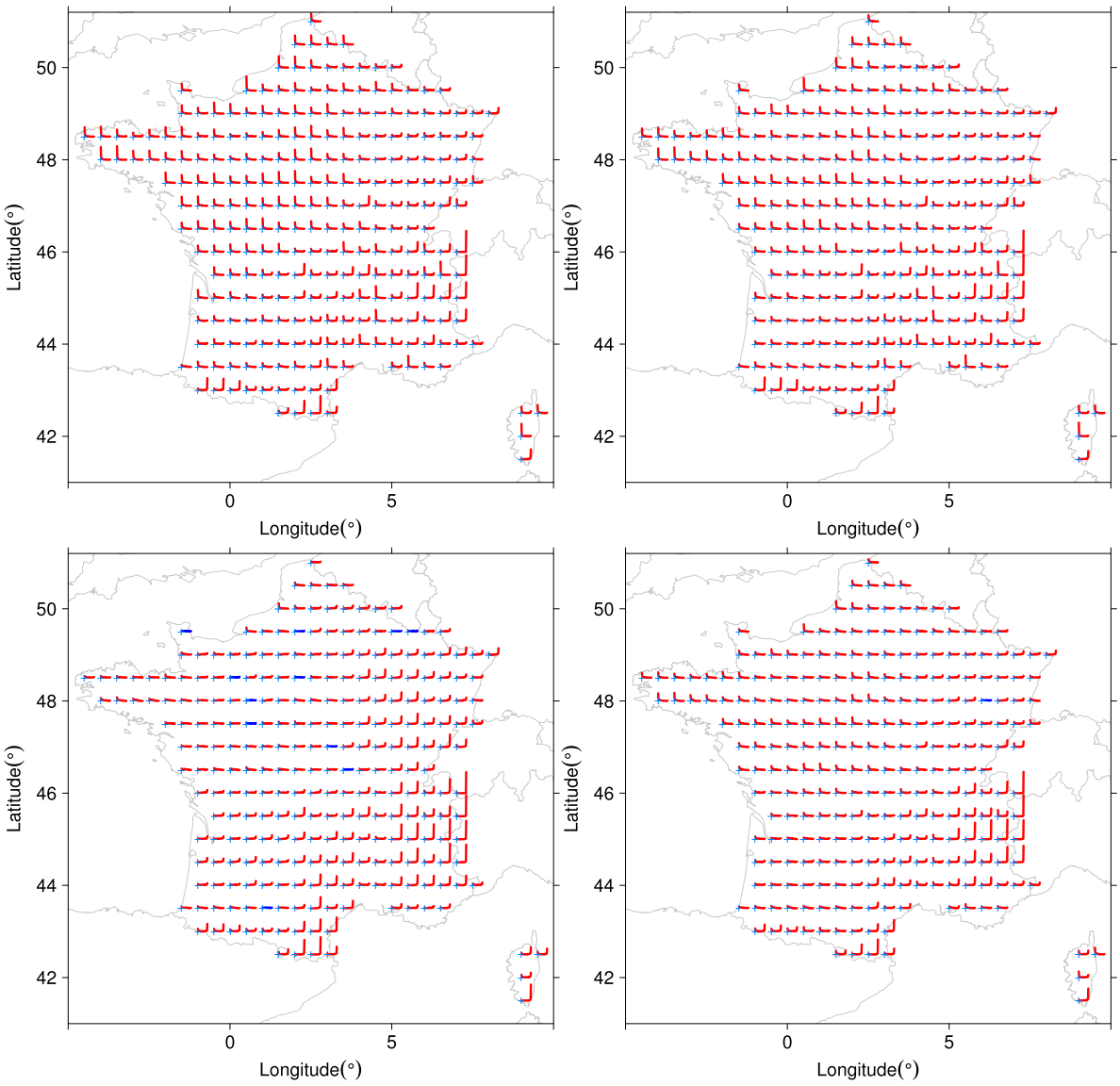

Figure 4.26 – Same as Figure 4.4 (a), but for raw MF ensemble, and for lead times 30 h (top left), 36 h (top right), 42 h (bottom left) and 48 h (bottom right).

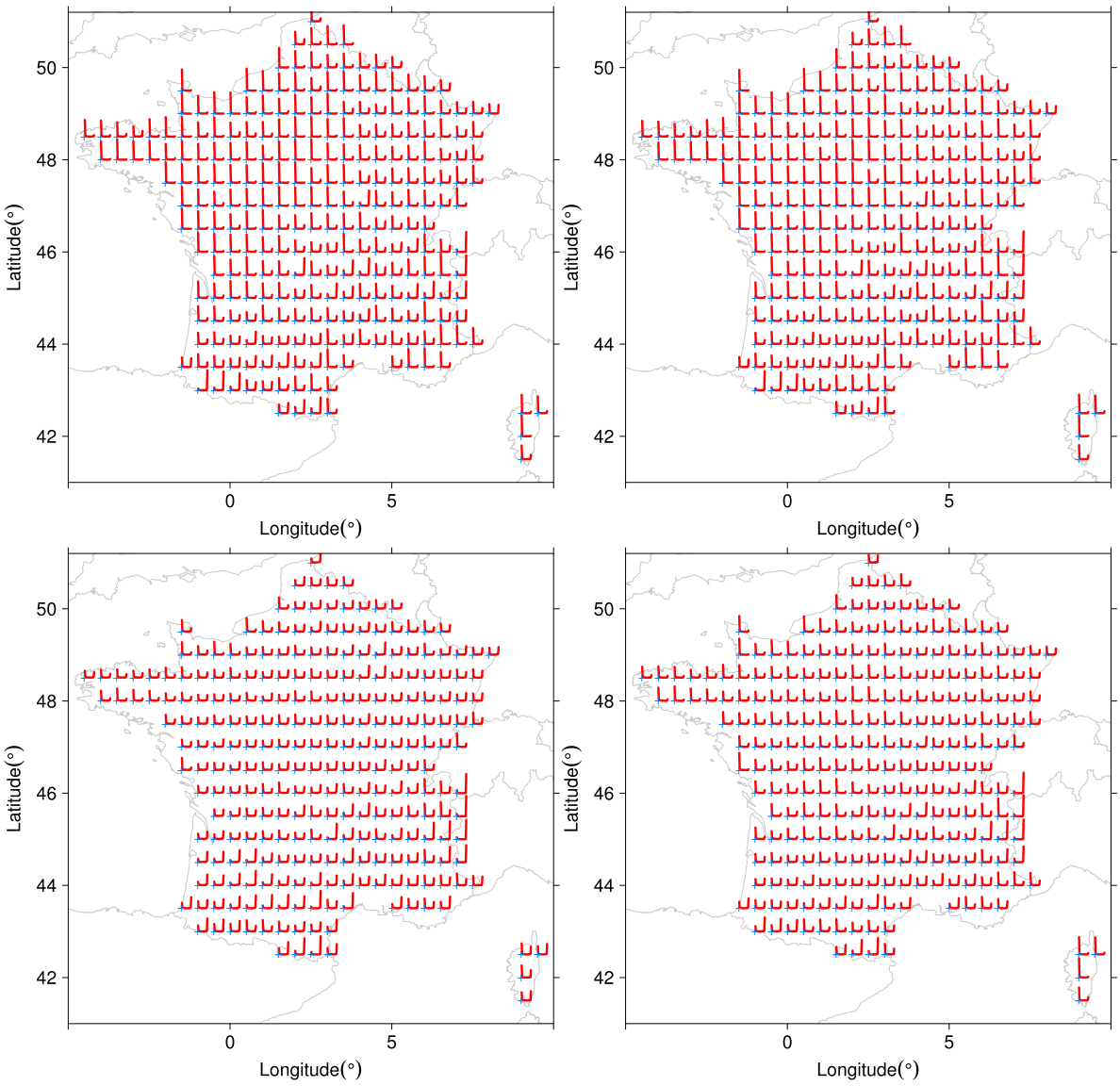

Figure 4.27 – Same as Figure 4.4 (a), but for raw NCEP ensemble, and for lead times 6 h (top left), 12 h (top right), 18 h (bottom left) and 24 h (bottom right).

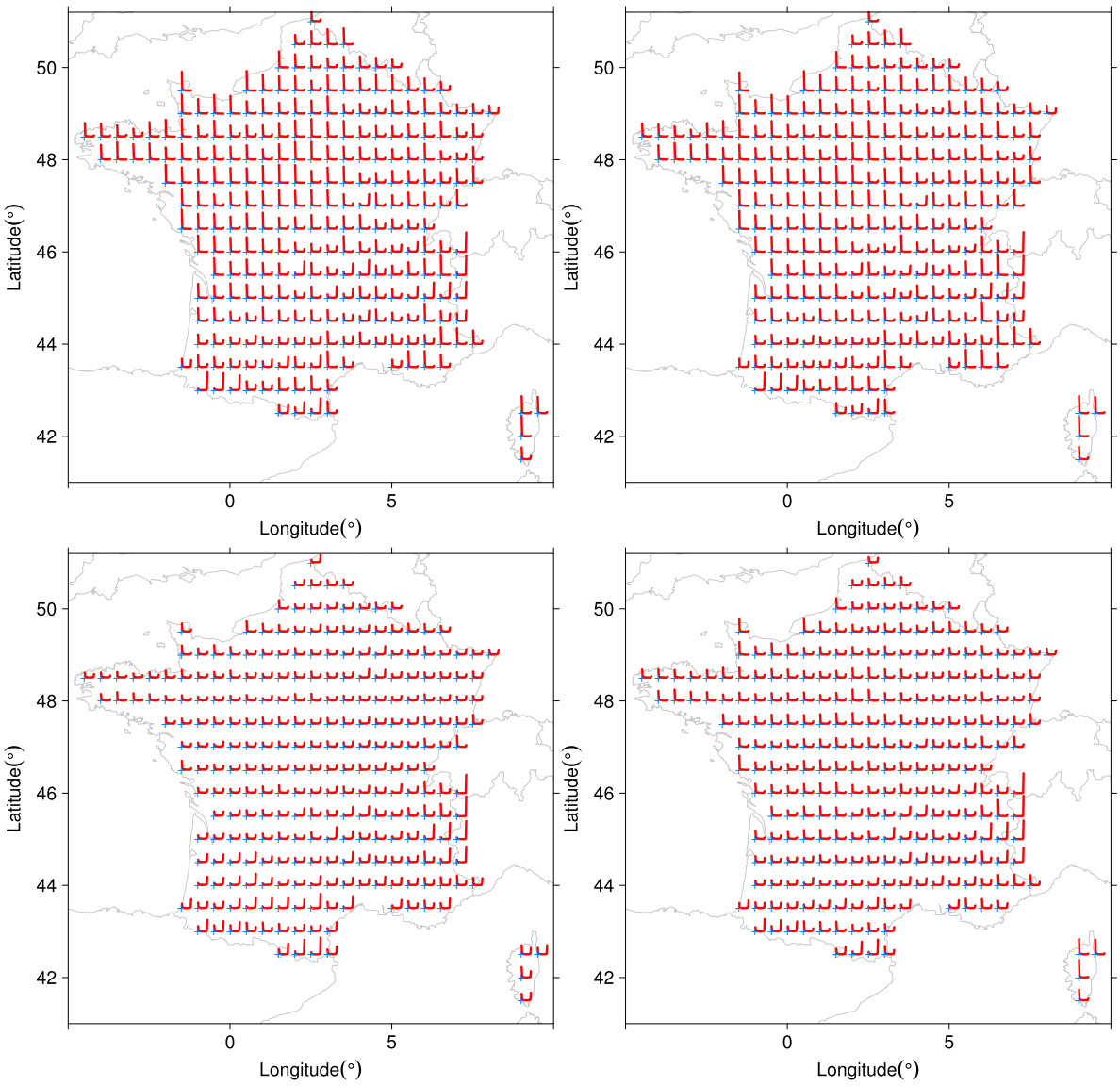

Figure 4.28 – Same as Figure 4.4 (a), but for raw NCEP ensemble, and for lead times 30 h (top left), 36 h (top right), 42 h (bottom left) and 48 h (bottom right).

4.E Time series of the aggregation weights

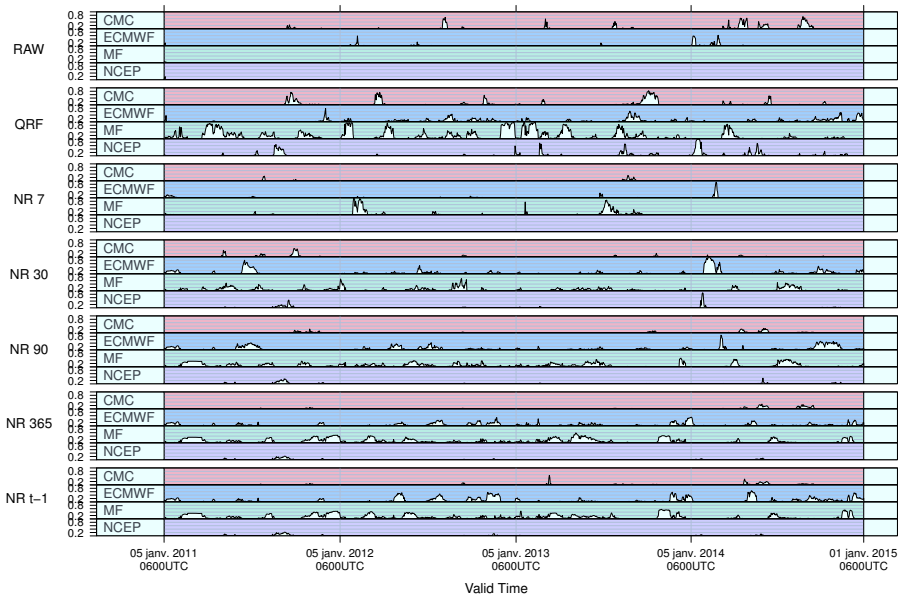

Figure 4.29 – Same as figure 4.11 but for lead time 6 h.

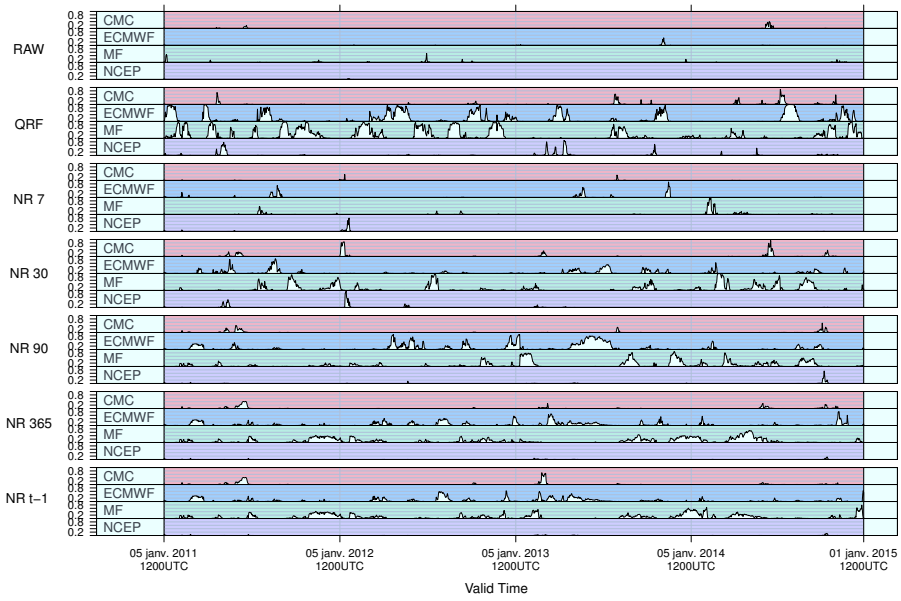

Figure 4.30 – Same as figure 4.11 but for lead time 12 h.

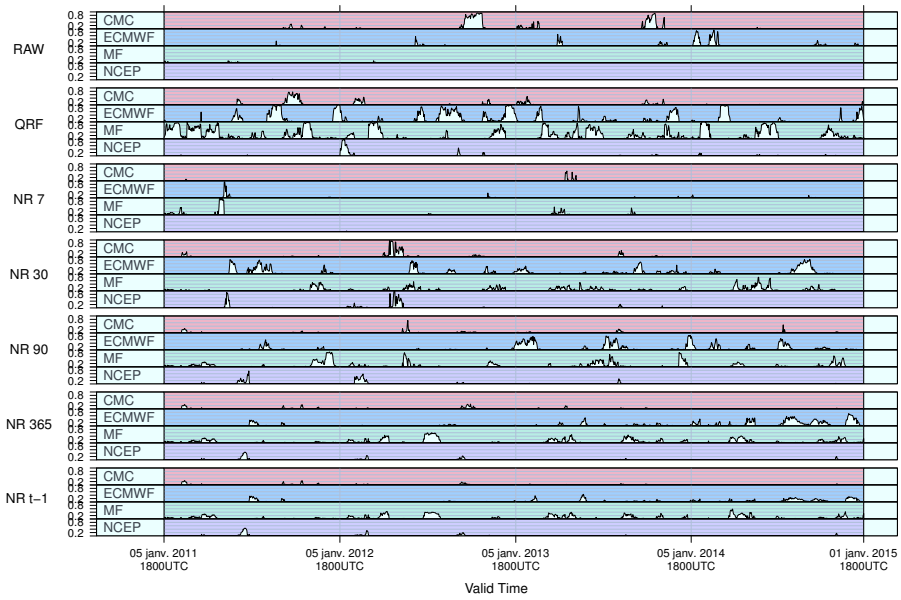

Figure 4.31 – Same as figure 4.11 but for lead time 18 h.

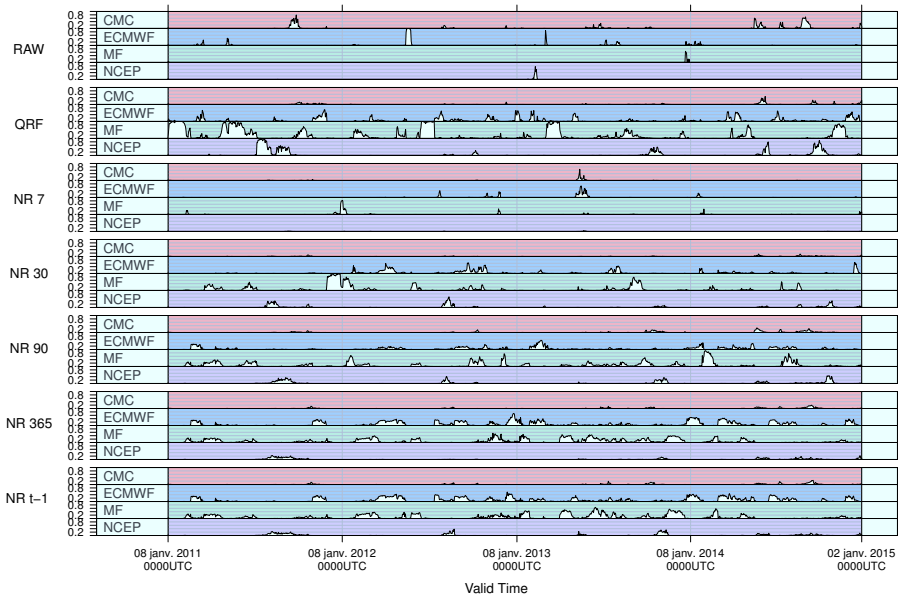

Figure 4.32 – Same as figure 4.11 but for lead time 24 h.

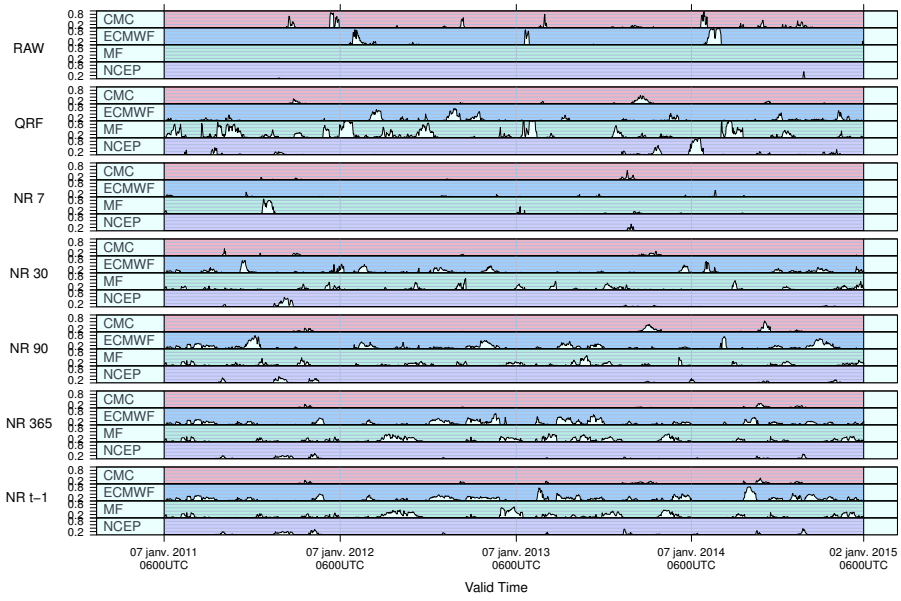

Figure 4.33 – Same as figure 4.11 but for lead time 30 h.

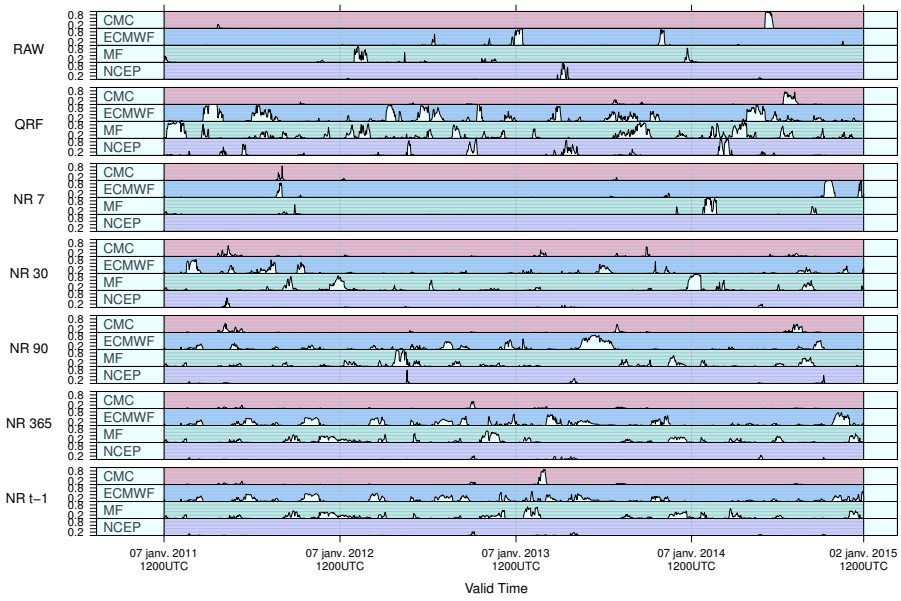

Figure 4.34 – Same as figure 4.11 but for lead time 36 h.

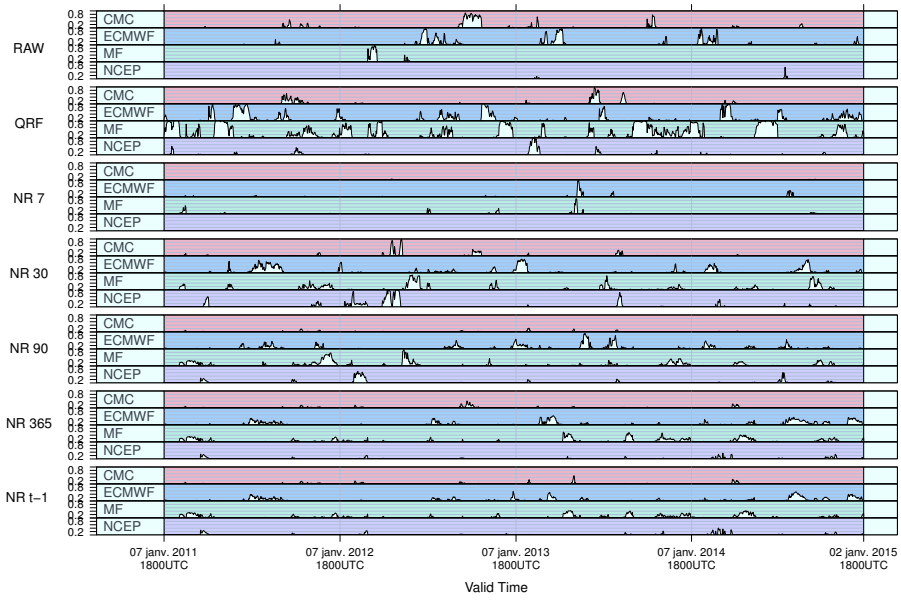

Figure 4.35 – Same as figure 4.11 but for lead time 42 h.

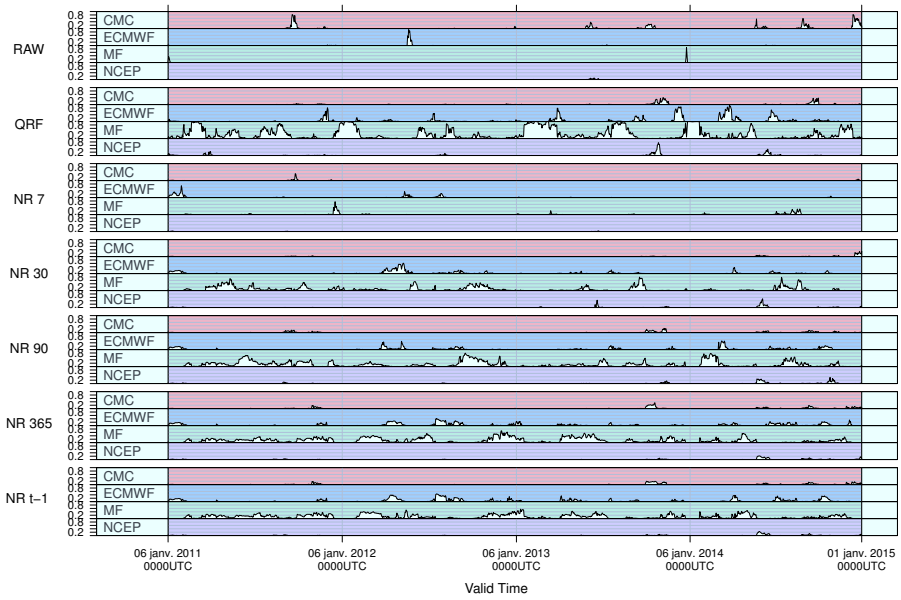

Figure 4.36 – Same as figure 4.11.

Chapter 5

Conclusion and Perspectives

This work investigated several aspects of weather forecast post-processing for wind speed forecasts over France. The aim was to build improved wind speed forecasts on a grid, for deterministic and probabilistic predictions. The adopted strategy is to first grid measurements, then to use these gridded measurements to train (E)MOS methods.

For deterministic forecasts, block MOS has been introduced along with a careful optimization of the size and number of the associated R objects. This lightweight block MOS shows good performances while allowing an important speeding up of operations. It will be implemented during Fall 2016 for operations.

As for the EMOS, empirical combination methods and combination methods based on the theory of prediction with expert advice have been compared. Since step-wise CDF are combined, this part of the work required to study the properties of the estimators of the CRPS with limited information about the forecast distribution. This led to recommendations to accurately estimate the CRPS. Also, due to some discrepancy of the flatness of the rank histograms and the value of the CRPS in this study, it is proposed to choose among probabilistic forecasts by first imposing to have a flat histogram according to the Jolliffe-Primo tests. The best combination method chosen with this criterion obtains a similar CRPS as the one minimizing the CRPS, while exhibiting much more flat rank histograms. It is also planned to make it operational by the end of next year.

The discrepancy between the forecast selection with the minimization of the CRPS or with the maximization of the proportion of flat rank histograms deserves more investigation on what this implies on the post-processing of ensemble forecasts.

Concerning the perspectives, when block MOS built on other models than ARPEGE are available for wind speed at Météo-France, aggregation methods will be tested. Aggregated MOS for temperature already exist at Météo-France, with much improved performances

compared to single MOS. A similar improvement in aggregated MOS for wind speed is expected. When they are chosen to minimize the RMSE, MOS forecasts estimate the conditional mean of the observation, and so do the mean of the aggregated EMOS forecasts. Therefore, an interesting study would be to compare the forecast performance of these two estimations.

The dynamic filtering step that usually ends forecast post-processing has not been implemented. Thus the performance of our aggregated wind speed EMOS forecasts may be further improved by the use of filtering techniques such as the ensemble Kalman filter.

For the longer term, implementing the same (E)MOS and aggregation methods for other parameters of interest, such as wind gusts or temperature, is aimed.

As for the probabilistic forecasts, due to the low number of grid-points for ensemble forecasts, the speeding up of the EMOS and aggregation methods was not a concern. But Météo-France high resolution ensemble forecast is expected to become operational by the end of 2017, with a grid size of 2.5 km and thousands of grid-points. A study of spatialised calibration and aggregation method is planned, to try to manage efficiently and quickly such an increase of the workload. The most obvious approach is to try block EMOS and block aggregation, since this pooling strategy worked well for MOS. An alternative would be to build EMOS and use aggregation methods on a manageable sub-sample of the grid points, before using geostatistical techniques to interpolate the parameters of the methods as in Scheuerer et al. (2015) and Dabernig et al. (2016), or even the forecast distributions themselves with functional kriging. An alternative and recent technique, multiple-point statistics, detailed in Mariethoz and Caers (2014), may allow to take into account the spatial dependence beyond the simple covariance, what kriging cannot do. Multiple-point statistics proceeds with some measured or simulated spatial representation of the spatial structures of the parameter of interest, called a "training image". This training image is then cut and paste over the new field of measurements, with some consistency conditions, to interpolate these measurements. Some "functional multiple-point statistics" may be used to interpolate the forecast distributions, but, to the best of our knowledge, this has never been developed.

The purpose of ensemble forecasts is to propose several alternative scenarios for the evolution of the atmosphere, in order to quantify the forecast uncertainty. Because EMOS methods are usually applied separately at each point or lead time of interest, the forecast spatio-temporal structures are not preserved after post-processing. For instance, we showed that the maps of forecast medians do contain information about incoming storm trajectories or occurrences of local winds, but a map of medians has no real physical interpretation as a possible scenario. In other words, the post-processing destroys the notion of alternative scenarios described by each member of the ensemble. Statistically speaking, we thus get calibrated forecasts of the marginal distribution at each point and time, but all information about the joint distribution over several points and/or times is lost. This

joint distribution is important for many applications. For instance, in France, warnings are issued if wind speed exceeds some threshold over half of a county, the probability of which cannot be assessed without knowing the joint distribution of the wind speed. Forecasting the probability of flooding may require a forecast of the joint distribution of rainfall over a catchment as input to hydrological models. A further post-processing step is usually applied to reconstruct calibrated members from the EMOS forecast, that is, a sample from the forecast joint distribution. Due to a lack of time, this step has not been achieved during the thesis. A most promising approach is the use of empirical copula. A copula is a multivariate function that links the marginal distributions of several random variables to their joint distribution. Empirical copula are estimates of the unknown copula based on a multivariate sample from the joint distribution. Two empirical copula methods are used in Meteorology: the ensemble copula coupling (ECC) and the Schaake shuffle (SS). ECC estimates the copula based from the forecast members of the raw ensemble (Schefzik 2011; Bouallegue et al. 2015). In our case, since several ensembles are combined, there is no notion of "raw ensemble" associated to the aggregated forecasts, and ECC cannot be used. The Schaake shuffle estimates the copula from past observations (Clark et al. 2004). It could thus be used for our aggregated forecasts. Since the copula is better estimated if a large sample is available, we should extend our gridded wind speed measurements to the longest period possible, that is from early 2009, when AROME became operational. Parametric copula could also be tried, but the spatial structure of the wind speed may be too complex to be easily modeled. An interesting alternative would be a Bayesian hierarchical modeling approach, as in Milliff et al. (2011). A Bayesian hierarchical model describes the probability distribution of a random variable with several nested models. Cressie and Wikle (2011) describe the approach and show how physical equations can be introduced in this hierarchy of nested models, allowing to mix Physics and Statistics. The physical equations are used as an implicit description of the spatio-temporal structure of the random variable of interest. The statistical part of the model describes the uncertainty sources, such as measurement errors, uncertain parameters, and so on. Due to a lack of time, a Bayesian hierarchical model was not investigated. An idea would be to use the governing equations of the atmosphere to model the joint distribution of the EMOS forecasts as the marginal distribution of the wind speed. Discussions with experts of this kind of approach showed that it may be a work of a thesis. Whatever approach is chosen, the verification of the reconstructed spatial structures would require the use of specific tools described in a special collection of the American Meteorological Society[1].

Besides the reconstruction of the joint distribution over space and time, the joint distribution of several parameters (e.g. temperature and wind speed) may be of interest for some applications. Vrac and Friederichs (2015) tackle this problem with the Schaake shuffle, while Wilks (2015) shows that the Schaake shuffle outperforms ECC. Another approach used in Pinson (2012) is to translate and dilate the members within a parametric and

[1] http://journals.ametsoc.org/topic/verification_icp

multivariate framework. The forecasts is thus not the joint distribution but a multivariate sample which supposedly reproduces the inter-parameter dependence. Here again, specific tools for verification of multivariate probabilistic forecasts the topic of ongoing studies, and may be interesting to investigate. In the case of multivariate probabilistic forecasts, the energy score can be used as a measure of performance. Since it is a generalization of the CRPS to the multivariate case, it is likely that the estimation of the energy score suffers similar issues as the one studied in Chapter 3 for the estimation of the CRPS. However, to the best of our knowledge, a closed-form expression of the energy score exists only for special cases of a multivariate normal random vector (Pinson and Tastu 2013). Furthermore, assessing the impact of the estimation of the copula would probably not be a straightforward task. For instance, the presence of ties would have to be taken into account, as suggested in Kojadinovic (2016). Furthermore, the CRPS is not sensitive to changes in the tail of the forecast distribution. Therefore, it cannot distinguish between two forecasts that predicts different probabilities of occurrence of rare, and usually hazardous, events. Developing a measure of performance for probabilistic forecasts that would not suffer this problem would be an interesting, but daunting, task. The SEDI is a measure of performance that was developed to solve the same kind of problems for point forecasts. Following the same reasoning that led to this score is maybe a lead to a solution for probabilistic forecasts.

Bibliography

Alexandridis, A., and A. D. Zapranis, 2012: *Weather derivatives: modeling and pricing weather-related risk*. Springer Science & Business Media.

Allard, D., A. Comunian, and P. Renard, 2012: Probability aggregation methods in geoscience. *Mathematical Geosciences*, **44 (5)**, 545–581.

Alpaydin, E., 2014: *Introduction to machine learning*. MIT press.

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *Journal of Climate*, **9 (7)**, 1518–1530.

Azaïs, J., and J. Bardet, 2006: *Le modèle linéaire par l'exemple: régression, analyse de la variance et plans d'expériences illustrés avec R, SAS et Splus*. Dunod, 326 pp.

Baars, J. A., and C. F. Mass, 2005: Performance of National Weather Service forecasts compared to operational, consensus, and weighted model output statistics. *Weather and Forecasting*, **20 (6)**, 1034–1047.

Baker, K. M., and W. W. Kirk, 2007: Comparative analysis of models integrating synoptic forecast data into potato late blight risk estimate systems. *Computers and electronics in agriculture*, **57 (1)**, 23–32.

Baran, S., 2014: Probabilistic wind speed forecasting using Bayesian model averaging with truncated normal components. *Computational Statistics & Data Analysis*, **75**, 227–238.

Baran, S., A. Horanyi, and D. Nemoda, 2014: Comparison of the BMA and EMOS statistical methods in calibrating temperature and wind speed forecast ensembles. *Idojárás*, **118 (3)**, 217–241.

Baran, S., and S. Lerch, 2015: Log-normal distribution based Ensemble Model Output Statistics models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological Society*.

Baran, S., and S. Lerch, 2016: Mixture EMOS model for calibrating ensemble forecasts of wind speed. *Environmetrics*, **27**, 116–130.

Bates, J. M., and C. W. Granger, 1969: The combination of forecasts. *Journal of the Operational Research Society*, **20 (4)**, 451–468.

Baudin, P., 2015: Prévision séquentielle par agrégation d'ensemble: application à des prévisions météorologiques assorties d'incertitudes. Ph.D. thesis, Université Paris-Saclay.

Bauer, P., A. Thorpe, and G. Brunet, 2015: The quiet revolution of numerical weather prediction. *Nature*, **525 (7567)**, 47–55.

Bertrand, J.-L., X. Brusset, and M. Fortin, 2015: Assessing and hedging the cost of unseasonal weather: Case of the apparel sector. *European Journal of Operational Research*, **244 (1)**, 261–276.

Besse, P., H. Milhem, O. Mestre, A. Dufour, and V.-H. Peuch, 2007: Comparaison de techniques de "Data Mining" pour l'adaptation statistique des prévisions d'ozone du modèle de chimie-transport MOCAGE. *Pollution Atmosphérique*, **49 (195)**, 285–292.

Bhaskar, K., and S. Singh, 2012: AWNN-assisted wind power forecasting using feed-forward neural network. *IEEE transactions on sustainable energy*, **3 (2)**, 306–315.

Bosart, L. F., 2003: Whither the weather analysis and forecasting process? *Weather and forecasting*, **18 (3)**, 520–529.

Bouallegue, Z. B., T. Heppelmann, S. E. Theis, and P. Pinson, 2015: Generation of scenarios from calibrated ensemble forecasts with a dynamic ensemble copula coupling approach. *arXiv preprint arXiv:1511.05877*.

Bougeault, P., and Coauthors, 2010: The THORPEX interactive grand global ensemble. *Bulletin of the American Meteorological Society*, **91 (8)**, 1059.

Bouzgou, H., and N. Benoudjit, 2011: Multiple architecture system for wind speed prediction. *Applied Energy*, **88 (7)**, 2463–2471.

Breiman, L., 2001: Random forests. *Machine learning*, **45 (1)**, 5–32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone, 1984: *Classification and regression trees, new edition*. Chapman and Hall, CRC, 368 pp.

Bremnes, J. B., 2007: Improved calibration of precipitation forecasts using ensemble techniques. *In practice*, **10**, 5.

Brier, G., 1950: Verification of forecasts expressed in terms of probability. *Monthly weather review*, **78 (1)**, 1–3.

Bröcker, J., 2009: Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, **135 (643)**, 1512–1519.

Bröcker, J., 2012: Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, **138 (667)**, 1611–1617.

Buizza, R., P. Houtekamer, G. Pellerin, Z. Toth, Y. Zhu, and M. Wei, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Monthly Weather Review*, **133 (5)**, 1076–1097.

Burlando, M., P. De Gaetano, M. Pizzo, M. P. Repetto, G. Solari, and M. Tizzi, 2013: Wind climate analysis in complex terrains. *Journal of Wind Engineering and Industrial Aerodynamics*, **123**, 349–362.

Burlando, M., A. Freda, C. Ratto, and G. Solari, 2010: A pilot study of the wind speed along the Rome–Naples HS/HC railway line. Part 1—Numerical modelling and wind simulations. *Journal of Wind Engineering and Industrial Aerodynamics*, **98 (8)**, 392–403.

Cadenas, E., and W. Rivera, 2010: Wind speed forecasting in three different regions of Mexico, using a hybrid ARIMA–ANN model. *Renewable Energy*, **35 (12)**, 2732–2738.

Candille, G., 2003: Validation des systèmes de prévisions météorologiques probabilistes. Ph.D. thesis, Paris 6.

Cantelaube, P., and J.-M. Terres, 2005: Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A*, **57 (3)**, 476–487.

Cesa-Bianchi, N., G. Lugosi, and Coauthors, 2006: *Prediction, learning, and games*, Vol. 1. Cambridge University Press Cambridge.

Charba, J. P., and F. G. Samplatsky, 2011: High-resolution GFS-based MOS quantitative precipitation forecasts on a 4-km grid. *Monthly Weather Review*, **139 (1)**, 39–68.

Charba, J. P., and F. Samplatsky, 2009: Hi-res gridded MOS 6-h QPF guidance. *23rd Conf. on Weather Analysis and Forecasting/19th Conf. on Numerical Weather Prediction*, URL https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154176.htm.

Cheng, W. Y., and W. J. Steenburgh, 2007: Strengths and weaknesses of MOS, running-mean bias removal, and Kalman filter techniques for improving model forecasts over the western United States. *Weather and Forecasting*, **22 (6)**, 1304–1318.

Clark, M., S. Gangopadhyay, L. Hay, B. Rajagopalan, and R. Wilby, 2004: The Schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *Journal of Hydrometeorology*, **5 (1)**, 243–262.

Clemen, R. T., 1989: Combining forecasts: A review and annotated bibliography. *International journal of forecasting*, **5 (4)**, 559–583.

Colak, I., S. Sagiroglu, and M. Yesilbudak, 2012: Data mining and wind power prediction: A literature review. *Renewable Energy*, **46**, 241–247.

Costa, A., A. Crespo, J. Navarro, G. Lizcano, H. Madsen, and E. Feitosa, 2008: A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews*, **12 (6)**, 1725–1744.

Courtier, P., C. Freydier, J. Geleyn, F. Rabier, and M. Rochas, 1991: The ARPEGE project at Météo-France. *Workshop on numerical methods in atmospheric models*, Vol. 2, 193–231.

Courtier, P., J.-N. Thépaut, and A. Hollingsworth, 1994: A strategy for operational implementation of 4D-Var, using an incremental approach. *Quarterly Journal of the Royal Meteorological Society*, **120 (519)**, 1367–1387.

Cressie, N. A. C., and C. K. Wikle, 2011: *Statistics for spatio-temporal data*, Vol. 465. Wiley, 624 pp.

Dabernig, M., G. J. Mayr, J. W. Messner, A. Zeileis, and Coauthors, 2016: Spatial ensemble post-processing with standardized anomalies. Tech. rep.

Descamps, L., C. Labadie, and E. Bazile, 2011: Representing model uncertainty using the multiparametrization method. *Proceedings of ECMWF Workshop on Representing Model Uncertainty and Error in Numerical Weather and Climate Prediction Models, 20-24 June 2011*, 175–182.

Descamps, L., C. Labadie, A. Joly, E. Bazile, P. Arbogast, and P. Cébron, 2014: PEARP, the Météo-France short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **141 (690)**, 1671–1685.

Devaine, M., Y. Goude, and G. Stoltz, 2009: Aggregation of sleeping predictors to forecast electricity consumption. *Rapport technique, EDF R&D et École normale supérieure, Paris*.

Diaconis, P., and B. Efron, 1983: Computer intensive methods in statistics. *Scientific American*, **248 (5)**.

Douak, F., F. Melgani, and N. Benoudjit, 2013: Kernel ridge regression with active learning for wind speed prediction. *Applied energy*, **103**, 328–340.

Dubus, L., S. E. Haupt, and Coauthors, 2014: Weather Matters for Energy.

Elmore, K. L., 2005: Alternatives to the chi-square test for evaluating rank histograms from ensemble forecasts. *Weather and forecasting*, **20 (5)**, 789–795.

Erickson, M. C., J. B. Bower, V. J. Dagostaro, J. P. Dallavalle, E. Jacks, J. S. Jensenius Jr, and J. C. Su, 1991: Evaluating the impact of RAFS changes on the NGM-based MOS guidance. *Weather and Forecasting*, **6 (1)**, 142–147.

European Center for Medium-Range Weather Forecasts, 2006: Application and Verification of ECMWF products in Member States and Co-operating States. URL http://www.ecmwf.int/sites/default/files/elibrary/2006/9218-application-and-verification-ecmwf-products-member-states-and-co-operating-states.pdf.

Ferraty, F., I. Van Keilegom, and P. Vieu, 2012: Regression when both response and predictor are functions. *Journal of Multivariate Analysis*, **109**, 10–28.

Ferraty, F., and P. Vieu, 2006: *Nonparametric functional data analysis: theory and practice*. Springer Science & Business Media, 260 pp.

Ferro, C. A., D. S. Richardson, and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, **15 (1)**, 19–24.

Ferro, C. A. T., 2014: Fair scores for ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **140 (683)**, 1917–1923.

Flowerdew, J., 2012: Calibration and combination of medium-range ensemble precipitation forecasts. *Met Office Forecasting Research Technical Report*, **567**.

Friederichs, P., and A. Hense, 2008: A probabilistic forecast approach for daily precipitation totals. *Weather and Forecasting*, **23 (4)**, 659–673.

Friederichs, P., and T. L. Thorarinsdottir, 2012: Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, **23 (7)**, 579–594.

Fritsch, J., J. Hilliker, J. Ross, and R. Vislocky, 2000: Model consensus. *Weather and forecasting*, **15 (5)**, 571–582.

Furrer, R., and P. Naveau, 2007: Probability weighted moments properties for small samples. *Statistics & probability letters*, **77 (2)**, 190–195.

Gerchinovitz, S., V. Mallet, and G. Stoltz, 2008: A further look at sequential aggregation rules for ozone ensemble forecasting. *Rapport technique, INRIA Paris-Rocquencourt et École normale supérieure, Paris*.

Gilbert, K. K., B. Glahn, R. Cosgrove, K. Sheets, and G. Wagner, 2009: Gridded model output statistics: Improving and expanding. *Preprints, 23rd Conf. Weather Analysis and Forecasting and 19th Conf. Numerical Prediction, Omaha, NE, Amer. Meteor. Soc*, URL https://ams.confex.com/ams/23WAF19NWP/techprogram/paper_154285.htm.

Glahn, B., 2014: Determining an Optimal Decay Factor for Bias-Correcting MOS Temperature and Dewpoint Forecasts. *Weather and Forecasting*, **29 (4)**, 1076–1090.

Glahn, B., K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009: The gridding of MOS. *Weather and Forecasting*, **24 (2)**, 520–529.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *Journal of applied meteorology*, **11 (8)**, 1203–1211.

Gneiting, T., 2011a: Making and evaluating point forecasts. *Journal of the American Statistical Association*, **106 (494)**, 746–762.

Gneiting, T., 2011b: Quantiles as optimal point forecasts. *International Journal of Forecasting*, **27 (2)**, 197–207.

Gneiting, T., 2014: Calibration of medium-range weather forecasts. *Technical Memoranda 719*, ECMWF.

Gneiting, T., F. Balabdaoui, and A. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **69 (2)**, 243–268.

Gneiting, T., and A. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, **102 (477)**, 359–378.

Gneiting, T., A. Raftery, A. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Monthly Weather Review*, **133 (5)**, 1098–1118.

Gneiting, T., R. Ranjan, and Coauthors, 2013: Combining predictive distributions. *Electronic Journal of Statistics*, **7**, 1747–1782.

Greenwood, J. A., J. M. Landwehr, N. C. Matalas, and J. R. Wallis, 1979: Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, **15 (5)**, 1049–1054.

Grimit, E., T. Gneiting, V. Berrocal, and N. Johnson, 2006: The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, **132 (621C)**, 2925–2942.

Guo, Z., W. Zhao, H. Lu, and J. Wang, 2012: Multi-step forecasting for wind speed using a modified EMD-based artificial neural network model. *Renewable Energy*, **37 (1)**, 241–249.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Monthly Weather Review*, **136 (7)**, 2608–2619.

Haiden, T., and Coauthors, 2015: Evaluation of ECMWF forecasts, including 2014-2015 upgrades. Tech. Rep. 765, ECMWF, 53 pp.

Hamill, T., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review*, **129 (3)**, 550–560.

Hamill, T., and S. Colucci, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Monthly Weather Review*, **126 (3)**, 711–724.

Hamill, T. M., and S. J. Colucci, 1996: Random and systematic error in NMC's short-range Eta ensembles. *Preprints, 13th Conf. on Probability and Statistics in the Atmospheric Sciences, San Francisco, CA, Amer. Meteor. Soc*, 51–56.

Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Monthly weather review*, **136 (7)**, 2620–2632.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Monthly Weather Review*, **134 (11)**, 3209–3229.

Haque, A. U., P. Mandal, M. E. Kaye, J. Meng, L. Chang, and T. Senjyu, 2012: A new strategy for predicting short-term wind speed using soft computing models. *Renewable and sustainable energy reviews*, **16 (7)**, 4563–4573.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: *The elements of statistical learning*. Springer, 745 pp.

Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden, 2014: Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, **41 (24)**, 9197–9205.

Hengl, T., 2007: A practical guide to geostatistical mapping of environmental variables. *JRC Scientific and Technichal Reports. Office for Official Publication of the European Communities, Luxembourg*.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, **15 (5)**, 559–570.

Hoeffding, W., 1963: Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, **58 (301)**, 13–30.

Holton, J. R., and G. J. Hakim, 2012: *An introduction to dynamic meteorology*, Vol. 88. Academic press.

Hosking, J., 1990: L-Moments: Analysis and Estimation of Distributions Using Linear Combinations of Order Statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, **52 (1)**, 105–124.

Ihász, I., Z. Üveges, M. Mile, and C. Németh, 2010: Ensemble calibration of ECMWF's medium-range forecasts. *Idojárás*, **114 (4)**, 275–286.

Im, J.-S., B. Glahn, and J. Ghirardelli, 2010: Real-time objective analysis of surface data at the Meteorological Development Laboratory. *Preprints, 20th Conf. on Probability and Statistics in the Atmospheric Sciences, Atlanta, GA, Amer. Meteor. Soc*, Vol. 219, URL https://ams.confex.com/ams/90annual/techprogram/paper_159910.htm.

Inness, P. M., and S. Dorling, 2013: *Operational weather forecasting*. John Wiley & Sons.

Jacks, E., J. B. Bower, V. J. Dagostaro, J. P. Dallavalle, M. C. Erickson, and J. C. Su, 1990: New NGM-based MOS guidance for maximum/minimum temperature, probability of precipitation, cloud amount, and surface wind. *Weather and Forecasting*, **5 (1)**, 128–138.

Jolliffe, I., and D. Stephenson, 2011: *Forecast verification: a practioner's guide in atmospheric science, second edition*. Wiley-Blackwell.

Jolliffe, I. T., and C. Primo, 2008: Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review*, **136 (6)**, 2133–2139.

Jung, J., and R. P. Broadwater, 2014: Current status and future advances for wind speed and power forecasting. *Renewable and Sustainable Energy Reviews*, **31**, 762–777.

Kang, J.-H., M.-S. Suh, K.-O. Hong, and C. Kim, 2011: Development of updateable model output statistics (UMOS) system for air temperature over South Korea. *Asia-Pacific Journal of Atmospheric Sciences*, **47 (2)**, 199–211.

Koenker, R., 2005: *Quantile regression*. 38, Cambridge university press.

Kojadinovic, I., 2016: Some copula inference procedures adapted to the presence of ties. *arXiv preprint arXiv:1609.05519*.

Kuhn, M., and K. Johnson, 2013: *Applied predictive modeling*. Springer, 600 pp.

Kusiak, A., H. Zheng, and Z. Song, 2009: Wind farm power prediction: a data-mining approach. *Wind Energy*, **12 (3)**, 275–293.

Lebarbier, É., and T. Mary-Huard, 2006: Une introduction au critère BIC: fondements théoriques et interprétation. *Journal de la SFdS*, **147 (1)**, 39–57.

Lei, M., L. Shiyan, J. Chuanwen, L. Hongling, and Z. Yan, 2009: A review on the forecasting of wind speed and generated power. *Renewable and Sustainable Energy Reviews*, **13 (4)**, 915–920.

Leutbecher, M., and T. N. Palmer, 2008: Ensemble forecasting. *Journal of Computational Physics*, **227 (7)**, 3515–3539.

Li, G., J. Shi, and J. Zhou, 2011: Bayesian adaptive combination of short-term wind speed forecasts from neural network models. *Renewable Energy*, **36 (1)**, 352–359.

Liu, H., H.-Q. Tian, C. Chen, and Y.-f. Li, 2010: A hybrid statistical method to predict wind speed and wind power. *Renewable Energy*, **35 (8)**, 1857–1861.

Liu, H., H.-q. Tian, and Y.-f. Li, 2012: Comparison of two new ARIMA-ANN and ARIMA-Kalman hybrid methods for wind speed prediction. *Applied Energy*, **98**, 415–424.

Lynch, P., 2008: The origins of computer weather prediction and climate modeling. *Journal of Computational Physics*, **227 (7)**, 3431–3444.

Malardel, S., 2005: *Fondamentaux de météorologie: à l'école du temps*, Vol. 45. Cepadues.

Mallet, V., B. Mauricette, and G. Stoltz, 2007: Description of Sequential Aggregation of Methods and their Performances for Ozone Ensemble Forecasting. *Technical report, École normale supérieure, DMA and CEREA*.

Mallet, V., and B. Sportisse, 2006: Ensemble-based air quality forecasts: A multimodel approach applied to ozone. *Journal of Geophysical Research: Atmospheres*, **111 (D18)**.

Mariethoz, G., and J. Caers, 2014: *Multiple-point geostatistics: stochastic modeling with training images*. John Wiley & Sons.

Mass, C. F., J. Baars, G. Wedam, E. Grimit, and R. Steed, 2008: Removal of systematic model bias on a model grid. *Weather and Forecasting*, **23 (3)**, 438–459.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Management science*, **22 (10)**, 1087–1096.

Meinshausen, N., 2006: Quantile regression forests. *The Journal of Machine Learning Research*, **7**, 983–999.

Meinshausen, N., and L. Schiesser, 2015: *quantregForest: Quantile Regression Forests*. URL http://CRAN.R-project.org/package=quantregForest, r package version 1.0.

Miller, C. M., R. T. Clemen, and R. L. Winkler, 1992: The effect of nonstationarity on combined forecasts. *International Journal of Forecasting*, **7 (4)**, 515–529.

Milliff, R. F., A. Bonazzi, C. K. Wikle, N. Pinardi, and L. M. Berliner, 2011: Ocean ensemble forecasting. Part I: Ensemble Mediterranean winds from a Bayesian hierarchical model. *Quarterly Journal of the Royal Meteorological Society*, **137 (657)**, 858–878.

Möller, D., and M. Scheuerer, 2013: Postprocessing of Ensemble Forecasts for Wind Speed over Germany. Ph.D. thesis, Diploma thesis, Faculty of Mathematics and Computer Science, Heidelberg University. Available online at http://www.rzuser.uni-heidelberg.de/˜kd4/files/Moeller2013.pdf.

Müller, W., C. Appenzeller, F. Doblas-Reyes, and M. Liniger, 2005: A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. *Journal of Climate*, **18 (10)**, 1513–1523.

Mylne, K. R., 2002: Decision-making from probability forecasts based on forecast value. *Meteorological Applications*, **9 (3)**, 307–315.

Novak, D. R., C. Bailey, K. F. Brill, P. Burke, W. A. Hogsett, R. Rausch, and M. Schichtel, 2014: Precipitation and temperature forecast performance at the Weather Prediction Center. *Weather and Forecasting*, **29 (3)**, 489–504.

Oesting, M., M. Schlather, and P. Friederichs, 2013: Conditional modelling of extreme wind gusts by bivariate brown-resnick processes. *arXiv preprint arXiv:1312.4584*.

Palm, F. C., and A. Zellner, 1992: To combine or not to combine? Issues of combining forecasts. *Journal of Forecasting*, **11 (8)**, 687–701.

Pinson, P., 2012: Adaptive calibration of (u, v)-wind ensemble forecasts. *Quarterly Journal of the Royal Meteorological Society*, **138 (666)**, 1273–1284.

Pinson, P., and J. Tastu, 2013: Discrimination ability of the energy score. Tech. rep., Technical University of Denmark.

Qin, X., C. Jiang, and J. Wang, 2011: Online clustering for wind speed forecasting based on combination of RBF neural network and persistence method. *2011 Chinese Control and Decision Conference (CCDC)*.

R Core Team, 2015: *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing, URL https://www.R-project.org.

Raftery, A., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review*, **133 (5)**, 1155–1174.

Rasmussen, P. F., 2001: Generalized probability weighted moments: application to the generalized pareto distribution. *Water Resources Research*, **37 (6)**, 1745–1751.

Rohli, R. V., A. J. Vega, and Coauthors, 2013: *Climatology*. Jones & Bartlett Publishers.

Ruth, D. P., B. Glahn, V. Dagostaro, and K. Gilbert, 2009: The performance of MOS in the digital age. *Weather and Forecasting*, **24 (2)**, 504–519.

Schaefer, J. T., and C. A. Doswell, 1979: On the interpolation of a vector field. *Monthly Weather Review*, **107 (4)**, 458–476.

Schefzik, R., 2011: Ensemble copula coupling. *Master's thesis, Faculty of Mathematics and Informatics, University of Heidelberg, Germany*.

Scheuerer, M., D. Möller, and Coauthors, 2015: Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *The Annals of Applied Statistics*, **9 (3)**, 1328–1349.

Schmeits, M. J., K. J. Kok, and D. H. Vogelezang, 2005: Probabilistic forecasting of (severe) thunderstorms in the Netherlands using model output statistics. *Weather and forecasting*, **20 (2)**, 134–148.

Schwarz, G., 1978: Estimating the dimension of a model. *The annals of statistics*, **6 (2)**, 461–464.

Seity, Y., P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, and V. Masson, 2011: The AROME-France convective-scale operational model. *Monthly Weather Review*, **139 (3)**, 976–991.

Sfanos, B., and P. Hirschberg, 2000: *AVN-based MOS wind guidance for the United States and Puerto Rico*. US Department of Commerce, National Oceanic and Atmospheric Administration, National Weather Service, Office of Meteorology, Office of Science and Technology.

Shi, J., J. Guo, and S. Zheng, 2012: Evaluation of hybrid forecasting approaches for wind speed and power generation time series. *Renewable and Sustainable Energy Reviews*, **16 (5)**, 3471–3480.

Siegert, S., 2015: *SpecsVerification: Forecast Verification Routines for the SPECS FP7 Project*. URL http://CRAN.R-project.org/package=SpecsVerification, r package version 0.4-1.

Slingo, J., and T. Palmer, 2011: Uncertainty in weather and climate prediction. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, **369 (1956)**, 4751–4767.

Sloughter, J. M., T. Gneiting, and A. E. Raftery, 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *Journal of the American Statistical Association*, **105 (489)**, 25–35.

Solari, G., M. P. Repetto, M. Burlando, P. De Gaetano, M. Pizzo, M. Tizzi, and M. Parodi, 2012: The wind forecast for safety management of port areas. *Journal of Wind Engineering and Industrial Aerodynamics*, **104**, 266–277.

Stoltz, G., 2010: Agrégation séquentielle de prédicteurs: méthodologie générale et applications à la prévision de la qualité de l'air et à celle de la consommation électrique. *Journal de la Société Française de Statistique*, **151 (2)**, 66–106.

Swinbank, R., and Coauthors, 2016: The TIGGE project and its achievements. *Bulletin of the American Meteorological Society*, **97 (1)**, 49–67.

Székely, G. J., and M. L. Rizzo, 2013: Energy statistics: A class of statistics based on distances. *Journal of statistical planning and inference*, **143 (8)**, 1249–1272.

Taillardat, M., O. Mestre, M. Zamo, and P. Naveau, 2016: Calibrated Ensemble Forecasts using Quantile Regression Forests and Ensemble Model Output Statistics. *Monthly Weather Review*, **144 (6)**, 2375–2393.

Takeuchi, I., Q. Le, T. Sears, and A. Smola, 2006: Nonparametric quantile estimation. *The Journal of Machine Learning Research*, **7**, 1231–1264.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability*, 1–25.

Taylor, J. W., and R. Buizza, 2002: Neural network load forecasting with weather ensemble predictions. *IEEE Transactions on Power systems*, **17 (3)**, 626–632.

Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **173 (2)**, 371–388.

Thorey, J., V. Mallet, C. Chaussin, L. Descamps, and P. Blanc, 2015: Ensemble forecast of solar radiation using TIGGE weather forecasts and HelioClim database. *Solar Energy*, **120**, 232–243.

Trnka, M., and Coauthors, 2011: Agroclimatic conditions in europe under climate change. *Global Change Biology*, **17 (7)**, 2298–2318.

Vislocky, R. L., and J. M. Fritsch, 1995: Improved model output statistics forecasts through model consensus. *Bulletin of the American Meteorological Society*, **76 (7)**, 1157–1164.

Vrac, M., and P. Friederichs, 2015: Multivariate—Intervariable, Spatial, and Temporal—Bias Correction. *Journal of Climate*, **28 (1)**, 218–237.

Wang, Q., 1996: Direct sample estimators of L moments. *Water resources research*, **32 (12)**, 3617–3619.

Weijs, S. V., R. Van Nooijen, and N. Van De Giesen, 2010: Kullback-Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition. *Monthly Weather Review*, **138 (9)**, 3387–3399.

Weisberg, S., and J. Fox, 2010: *An R companion to applied regression*. Sage Publications, Inc, 472 pp.

White, H., 1992: Nonparametric estimation of conditional quantiles using neural networks. *Computing Science and Statistics*, Springer, 190–199.

Wilks, D. S., 2015: Multivariate ensemble Model Output Statistics using empirical copulas. *Quarterly Journal of the Royal Meteorological Society*, **141 (688)**, 945–952.

Wilks, D. S., and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Monthly Weather Review*, **135 (6)**, 2379–2390.

Wilson, L. J., S. Beauregard, A. E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian model averaging. *Monthly Weather Review*, **135 (4)**, 1364–1385.

Wilson, L. J., and M. Vallée, 2002: The Canadian updateable model output statistics (UMOS) system: Design and development tests. *Weather and forecasting*, **17 (2)**, 206–222.

Winkler, R., and Coauthors, 1996: Scoring rules and the evaluation of probabilities. *Test*, **5 (1)**, 1–60.

Wood, S., 2006: *Generalized additive models: an introduction with R*, Vol. 66. CRC Press, 410 pp.

Wood, S. N., 2003: Thin plate regression splines. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **65 (1)**, 95–114.

Woodcock, F., and C. Engel, 2005: Operational consensus forecasts. *Weather and forecasting*, **20 (1)**, 101–111.

World Meteorological Organisation, 2008: WMO Guide to meteorological instruments and methods of observation. Tech. Rep. WMO-No. 8, World Meteorological Organisation. Updated in 2010.

Zamo, M., L. Bel, O. Mestre, and J. Stein, 2016: Improved gridded windspeed forecasts by statistical post-processing of numerical models with block regression. *Weather and Forecasting*, **31 (6)**, 1929–1945.

Zamo, M., O. Mestre, P. Arbogast, and O. Pannekoucke, 2014a: A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production, part I: Deterministic forecast of hourly production. *Solar Energy*, **105**, 792–803.

Zamo, M., O. Mestre, P. Arbogast, and O. Pannekoucke, 2014b: A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. *Solar Energy*, **105**, 804–816.

Zhou, J., J. Shi, and G. Li, 2011: Fine tuning support vector machines for short-term wind speed forecasting. *Energy Conversion and Management*, **52 (4)**, 1990–1998.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society*, **83 (1)**, 73.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Post-traitements statistiques de prévisions de vent déterministes et d'ensemble sur une grille

Mots clefs : prévision météorologique; post-traitement statistique; grille; vérification de prévision; vent de surface; agrégation séquentielle

Résumé : Les erreurs des modèles de prévision numérique du temps (PNT) peuvent être réduites par des méthodes de post-traitement (dites d'adaptation statistique ou AS) construisant une relation statistique entre les observations et les prévisions. L'objectif de cette thèse est de construire des AS de prévisions de vent pour la France sur la grille de plusieurs modèles de PNT, pour les applications opérationnelles de Météo-France en traitant deux problèmes principaux. Construire des AS sur la grille de modèles de PNT, soit plusieurs milliers de points de grille sur la France, demande de développer des méthodes rapides pour un traitement en conditions opérationnelles. Deuxièmement, les modifications fréquentes des modèles de PNT nécessitent de mettre à jour les AS, mais l'apprentissage des AS requiert un modèle de PNT inchangé sur plusieurs années, ce qui n'est pas possible dans la majorité des cas.
Une nouvelle analyse du vent moyen à 10 m a été construite sur la grille du modèle local de haute résolution (2,5 km) de Météo-France, AROME. Cette analyse se compose de deux termes: une spline fonction de la prévision la plus récente d'AROME plus une correction par une spline fonction des coordonnées du point considéré. La nouvelle analyse obtient de meilleurs scores que l'analyse existante, et présente des structures spatio-temporelles réalistes. Cette nouvelle analyse, disponible au pas horaire sur 4 ans, sert ensuite d'observation en points de grille pour construire des AS.
Des AS de vent sur la France ont été construites pour ARPEGE, le modèle global de Météo-France. Un banc d'essai comparatif désigne les forêts aléatoires comme meilleure méthode. Cette AS requiert un long temps de chargement en mémoire de l'information nécessaire pour effectuer une prévision. Ce temps de chargement est divisé par 10 en entraînant les AS sur des points de grille contigü et en les élaguant au maximum. Cette optimisation ne déteriore pas les performances de prévision. Cette approche d'AS par blocs est en cours de mise en opérationnel.
Une étude préalable de l'estimation du « continuous ranked probability score » (CRPS) conduit à des recommandations pour son estimation et généralise des résultats théoriques existants. Ensuite, 6 AS de 4 modèles d'ensemble de PNT de la base TIGGE sont combinées avec les modèles bruts selon plusieurs méthodes statistiques. La meilleure combinaison s'appuie sur la théorie de la prévision avec avis d'experts, qui assure de bonnes performances par rapport à une prévision de référence. Elle ajuste rapidement les poids de la combinaison, un avantage lors du changement de performance des prévisions combinées. Cette étude a soulevé des contradictions entre deux critères de choix de la meilleure méthode de combinaison : la minimisation du CRPS et la platitude des histogrammes de rang selon les tests de Jolliffe-Primo. Il est proposé de choisir un modèle en imposant d'abord la platitude des histogrammes des rangs.

Title : Statistical Post-processing of Deterministic and Ensemble Wind Speed Forecasts on a Grid

Keywords : weather forecast; statistical post-processing; grid; forecast verification; surface wind; sequential aggregation

Abstract : Errors of numerical weather prediction (NWP) models can be reduced thanks to post-processing methods (model output statistics, MOS) that build a statistical relationship between the observations and associated forecasts. The objective of the present thesis is to build MOS for windspeed forecasts over France on the grid of several NWP models, to be applied on operations at Météo-France, while addressing the two main issues. First, building MOS on the grid of some NWP model, with thousands of grid points over France, requires to develop methods fast enough for operational delays. Second, requent updates of NWP models require updating MOS, but training MOS requires an NWP model unchanged for years, which is usually not possible.
A new windspeed analysis for the 10 m windspeed has been built over the grid of Météo-France's local area, high resolution (2,5km) NWP model, AROME. The new analysis is the sum of two terms: a spline with AROME most recent forecast as input plus a correction with a spline with the location coordinates as input. The new analysis outperforms the existing analysis, while displaying realistic spatio-temporal patterns. This new analysis, now available at an hourly rate over 4, is used as a gridded observation to build MOS in the remaining of this thesis.
MOS for windspeed over France have been built for ARPEGE, Météo-France's global NWP model. A test-bed designs random forests as the most efficient MOS. The loading times is reduced by a factor 10 by training random forests over block of nearby grid points and pruning them as much as possible. This time optimisation goes without reducing the forecast performances. This block MOS approach is currently being made operational.
A preliminary study about the estimation of the continuous ranked probability score (CRPS) leads to recommendations to efficiently estimate it and to generalizations of existing theoretical results. Then 4 ensemble NWP models from the TIGGE database are post-processed with 6 methods and combined with the corresponding raw ensembles thanks to several statistical methods. The best combination method is based on the theory of prediction with expert advice, which ensures good forecast performances relatively to some reference forecast. This method quickly adapts its combination weighs, which constitutes an asset in case of performances changes of the combined forecasts. This part of the work highlighted contradictions between two criteria to select the best combination methods: the minimization of the CRPS and the flatness of the rank histogram according to the Jolliffe-Primo tests. It is proposed to choose a model by first imposing the flatness of the rank histogram.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France