Non-parametric Methods of post-processing for Ensemble Forecasting

Maxime Taillardat

université PARIS-SACLAY

NNT : 2017SACLV072

Thèse de doctorat

UNIVERSITÉ DE VERSAILLES ST-QUENTIN-EN-YVELINES

université PARIS-SACLAY

Méthodes Non-Paramétriques de Post-Traitement des Prévisions d'Ensemble

Thèse de doctorat de l'Université Paris-Saclay
préparée à l'Université Versailles Saint-Quentin en Yvelines

École doctorale n°129 Sciences de l'environnement d'Île-de-France (SEIF)

Spécialité de doctorat : Météorologie, océanographie, physique de l'environnement

Thèse présentée et soutenue à Toulouse, le 11 décembre 2017, par

M. Maxime Taillardat

Composition du Jury :

M. Jean-Michel Poggi Professeur, Université Paris Descartes, Orsay	Président
M. Luc Perreault Professeur, IREQ, Québec	Rapporteur
M. Mathieu Ribatet Maître de conférences HDR, IMAG UMR 5149, Montpellier	Rapporteur
Mme. Juliette Blanchet Chargée de recherche, IGE UMR 5001, Grenoble	Examinatrice
Mme. Petra Friederichs Professeure, Université de Bonn, Bonn	Examinatrice
M. Philippe Naveau Directeur de recherche, LSCE UMR 8212, Gif-sur-Yvette	Directeur de thèse
Mme. Anne-Laure Fougères Professeure, Institut Camille Jordan, Lyon I	Co-Directrice de thèse
M. Olivier Mestre IDTM, Météo-France, Toulouse	Co-Directeur de thèse

Ce matin de juin, j'écris dans un transat au fond du jardin anglais que le soleil levant caresse voluptueusement pour en essuyer la rosée. À portée de main, sur un guéridon de paille tressée, le thé aux herbes tiédit à la brise. Le bouvreuil effronté, qui m'espionnait hier déjà, sautille et pirouette à trois pas en stridulant des joliesses absconses dont j'appréhende cependant qu'elles veuillent dire : " Tire-toi de là bonhomme, que je finisse les miettes de ton croissant qui sont tombées dans l'herbe."

Eh bon, comme l'oiseau, j'ai la plume frivole et baladeuse et tendance à papillonner autour du sujet sans m'y soumettre, voire même à m'en écarter carrément. Ce qui est pénible, avec les livres, je veux dire quand on les écrit, c'est qu'on est plus ou moins poussé à s'en tenir au sujet qu'on prétend traiter. Il faut savoir que cette contrainte est parfois très pénible quand elle s'abat sur un auteur velléitaire par nature, incohérent par goût, et facilement déconnectable par l'imprévu, en l'occurrence ce petit pédé de bouvreuil qui fait rien que frétiller de la queue pour m'empêcher d'aller plus loin. Dieu merci, quand on se contente de penser au lieu d'écrire, on a parfaitement le droit de sauter du coq à l'âne, sans s'attirer des remarques désobligeantes.

J'aurais dû être dérouleur de pensées plutôt qu'écriveur de bouquins.

Pierre Desproges

Avant-propos

Cette fois ça y est. A vrai dire, cela fait trois années que j'ai en tête ce passage de ma thèse, et au cours de ces années qui m'ont amené à sillonner New York, Washington, le Colorado, Berlin ou encore Vienne je griffonnais des bribes, quelques mots ou quelques phrases surgissant d'un moment ou d'une idée que je voulais faire figurer ici. Avant tout pour moi-même, avant tout pour garder quelque souvenir de ce qui a été pour moi beaucoup plus que le condensat scientifique que vous trouverez dans les pages suivantes. Mais il faut se rendre à l'évidence. Face à l'échelle de l'espace et du temps, ces grandeurs incommensurables ; nous ne sommes finalement que des êtres quantiques. Aussi, nos décisions, nos opinions, nos vies ne peuvent se comprendre que dans l'instant. Je veux dire par là que la recherche de la perfection que je voudrais insuffler à ces quelques lignes est vaine. Je m'apprête donc à écrire dans l'instant. Viendra ce qui viendra. Les bribes, je les garde pour moi. Cela ne sera pas de la grande littérature, mais je pense que mon naturel taiseux et mon vocabulaire parfois "fleuri" pourront être mis de côté quelques instants.

Je voudrais tout d'abord remercier le CNRM, sa direction et son équipe administrative, qui ont permis que mon travail se passe dans les meilleures conditions possibles pour moi. Je remercie l'ICJ pour m'avoir accueilli dans leurs locaux durant la seconde partie de ma thèse. Je tiens à adresser de sincères et chaleureux remerciements à MM. François Lalaurette, Olivier Rouzaud et Joël Stein tant pour l'intérêt qu'ils ont eu pour mon travail que pour la direction professionnelle qu'ils m'offrent. Je remercie bien évidemment le service COMPAS et plus particulièrement l'équipe DOP pour leur gentillesse, leur disponibilité ainsi que leur bienveillance. Avec une mention toute spéciale à Michaël, mon co-padawan devenu jeune Jedi avant moi.

Je salue aussi toutes les personnes avec qui j'ai pris plaisir d'échanger au cours des trois dernières années sur mon sujet de thèse. La liste de ces personnes est bien trop longue et, par conséquent, si vous êtes concerné et que vous lisez ces lignes : cela s'adresse à vous.

Je veux remercier chaleureusement MM. Luc Perreault et Mathieu Ribatet d'avoir accepté de rapporter cette thèse. J'en profite aussi pour remercier les membres du jury et plus généralement toutes les personnes qui ont œuvré à mes différents comités de thèse et de FCPLR.

A titre plus personnel, je tiens à saluer tous mes amis, ils se reconnaîtront. Certains sont

de ma promotion à l'ENM, certains ont toléré mes montées de voix au sein des Cons Sonnants, certains jouent de la vuvuzela tous les midi, certains sont Toulousains, Toulonnais, Corses, Parisiens, Californiens, Lyonnais, Stéphanois. Certains sont Roannais, et ont beaucoup participé à ma vie durant mes retours "à la source", entre une petite balle blanche, un gros ballon de foot ou un tout autre ballon anisé quand ce n'était pas un alcool de gentiane pour les plus aguerris. Je salue le GROUIK, le CCA mais tout particulièrement l'Homoursporc ; dont la grandeur n'a d'égale que sa multiplicité.

Il fallait faire un paragraphe à part maintenant pour parler des personnes suivantes. J'ai une pensée à travers ces lignes pour Pierre-Jean Hormière, mon professeur de MP* à Saint-Étienne, qui pendant deux ans a su "voir" (je n'ai pas d'autre mot ici), et me passionner pour la recherche, par les moyens les plus imaginatifs. Je pense lui devoir d'avoir insufflé un caractère très "galoisien" au chercheur que je suis.

Je tiens enfin à remercier ma triplette de directeurs de thèse. Merci Philippe, d'avoir été enthousiaste sur ce sujet dès les premiers instants. Merci pour l'invitation au voyage, dans tous les sens du terme. Merci Anne-Laure, d'avoir élargi ton champ de recherche pour m'accueillir. Merci pour avoir su apaiser et écouter mes comportements et mes valeurs les plus "extrêmes", au propre comme au figuré. Merci enfin Olivier, mon Mestre Jedi. Nous nous connaissons depuis quelques années maintenant, merci pour m'avoir mis le pied à l'étrier sur cette thématique et pour m'avoir laissé une grande liberté dans le travail. Merci pour avoir été là en toute occasion. Et merci pour toujours y être, je l'espère pour longtemps.

Vous fûtes tous les trois un encadrement exceptionnel de compétence, de disponibilité mais surtout pour moi d'humanité. Je suis très honoré de compter parmi vos élèves, vos collègues, vos amis.

Enfin, merci à toute ma famille. Les ch'tits comme les grands. Quelle ironie du sort, j'ai toujours été passionné, entre autres, par la conquête spatiale, avec un intérêt pour l'abstraction, l'élévation. Au contraire de mon éducation, de mes idées, de mes valeurs, qui sont très terriennes. Finalement, il faut croire que la météorologie et ses nuages représente un compromis pas si dénué de sens.

Table des matières

Chapitre 1

Résumé

Abstract Cette thèse concerne l'amélioration des prévisions d'ensemble par l'utilisation de méthodes de post-traitement statistique. Les méthodes les plus couramment utilisées dans le domaine de la météorologie sont des méthodes paramétriques. On montre ici comment des méthodes non-paramétriques de régression quantile basées sur l'utilisation de techniques propres aux forêts aléatoires peuvent supplanter les méthodes existantes. En effet, elles ne nécessitent pas d'hypothèse sur la distribution de la variable à calibrer. Elles permettent aussi la prise en compte de phénomènes non-linéaires, la possibilité d'ajout de covariables d'intérêt et notamment catégorielles ou encore une sélection automatique des variables utiles à la régression. On s'attache aussi à montrer le bon comportement de ces prévisions et leur valeur ajoutée pour des variables météorologiques complexes et des événements extrêmes, aussi difficiles à réaliser qu'à évaluer avec les scores conventionnels propres à la prévision d'ensemble. Ce résumé vise à exposer le fil rouge ayant guidé la thèse. Chaque partie de ce travail est donc résumée en terme de résultats et d'apports face aux problématiques énoncées.

1.1 Introduction

Dans son *Essai philosophique sur les probabilités*, Laplace (1814) introduit le concept de déterminisme. Tout comme Descartes et Newton en leur temps, il pense que tout phénomène est prévisible dès lors que l'on connaît les conditions initiales et les lois qui régissent le phénomène considéré. C'est en substance ce qui fait dire à Bjerknes (1904) que la météorologie est un problème déterministe à conditions initiales.

Peu de temps après, Cooke (1906) déclare qu'une prévision n'est complète que si le degré de confiance relatif à celle-ci est fourni. En 1908, Henri Poincaré défend le caractère instable de la prévision météorologique, dans le sens où d'infimes variations sur la connaissance de l'état initial de l'atmosphère peuvent aboutir à des prévisions totalement différentes.

Ce n'est que bien plus tard que Lorenz (Lorenz, 1963, 1967) met en evidence le caractère chaotique et la sensibilité aux conditions initiales des modèles de prévision météorologique.

Le concept de prévision d'ensemble en météorologie est mis en place par Epstein (1969b). Partant de la constatation que les prévisions météorologiques souffrent de certaines lacunes (incertitude et incomplétude du réseau d'observations, simplification forcée des équations d'évolution de l'atmosphère etc.) (Bauer et al., 2015), Epstein propose une méthode probabiliste permettant d'échantillonner de façon représentative l'état de l'atmosphère. En effet, une résolution complète d'une solution probabiliste (par les équations de Liouville et de Fokker-Planck) n'est pas envisageable d'un point de vue informatique. L'idée va être de perturber légèrement les conditions initiales et/ou les paramétrisations physiques du modèle. Dans ce cas, nous ne disposons plus alors d'une seule prévision déterministe mais d'un ensemble de prévisions (les membres de la prévision d'ensemble). Pour des raisons de temps de calcul, les membres sont produits à une résolution spatiale généralement moins fine que le modèle déterministe correspondant.

L'avénement du calcul haute performance va permettre la mise en place progressive des premiers systèmes de prévisions d'ensemble à travers le monde (Toth and Kalnay, 1993; Mureau et al., 1993). La prévision d'ensemble devient petit à petit un outil majeur des services météorologiques nationaux par l'aide à la décision qu'elle peut fournir aux prévisionnistes et le calcul de probabilité d'événements d'intérêt qu'elle permet d'effectuer (Buizza et al., 1999; Palmer, 2002).

Des alertes efficaces nécessitent des prévisions météorologiques précises et informatives afin de réduire tant les non détections que les fausses alarmes. Une bonne représentation de l'incertitude est tout aussi nécessaire pour disposer de prévisions fiables d'événements météorologiques. Cela facilite les prises de décision des usagers, lesdits usagers ayant souvent des difficultés à utiliser une information probabiliste (Hagedorn, 2017). La prévision d'ensemble est d'ailleurs de plus en plus utilisée dans des domaines dits météo-sensibles comme la production d'énergie (Taylor and Buizza, 2003; Pinson et al., 2009), l'hydrologie (Krzysztofowicz, 2001; Schaake et al., 2007), l'agriculture (Calanca et al., 2011), l'écologie (Poulos et al., 2012), la qualité de l'air (Mallet and Sportisse, 2006) ou encore l'économie (Ravazzolo and Vahey, 2014).

Chaque modèle de prévision d'ensemble présente des erreurs et des biais qui ne sont pas entièrement aléatoires (on retrouve souvent des faiblesses dues au modèle déterministe qui sert de base à la génération des ensembles ainsi qu'une quantification de l'incertitude erronée). A ce titre, des méthodes de post-traitement statistique (ou en français *calibration*[1]) ont été utilisées dès l'essor de la prévision numérique du temps (Glahn and Lowry, 1972).

Les méthodes de post-traitement ont pour objet de construire automatiquement une relation statistique entre les observations et les variables météorologiques correspondantes prévues par le modèle numérique. De nombreuses techniques de fouille de données et d'apprentissage automatique peuvent être mises à profit (Hastie et al., 2009; Wu et al., 2014). Ces modèles statistiques sont ensuite appliqués aux nouvelles prévisions dans le but de les améliorer, ou de prévoir des variables observées mais non prévues par le modèle numérique.

1. Attention : le mot anglais *calibration* signifie autre chose, voir Table 1.1

Cette approche appelée *Adaptation Statistique d'Ensemble* prévoit tout ou partie de la distribution de l'observation ; voir par exemple Gneiting et al. (2005). Un fait particulièrement intéressant est que, quelque soit la performance initiale du modèle de prévision d'ensemble, un post-traitement statistique bien conçu parvient à améliorer la performance des prévisions (Ruth et al., 2009; Hemri et al., 2014; Taillardat et al., 2016).

Dans le cadre de ce travail de thèse, on étudie plus particulièrement le système de prévsion d'ensemble PEARP, produit par Météo-France (Descamps et al., 2015). Les méthodes mises au point sur PEARP seront à terme appliquées au système de prévision d'ensemble haute résolution PEAROME (Raynaud and Bouttier, 2016; Bouttier et al., 2016). Le Centre Européen de Prévisions Météorologiques à Moyen Terme (ECMWF) a établi un rapport faisant l'inventaire de différentes méthodes de post-traitement statistique existantes (Gneiting, 2014). L'objectif de ce travail de thèse est de confronter ces techniques avec une méthode non-paramétrique que nous avons développée. Notre approche, basée sur les forêts aléatoires (Breiman, 2001), apporte de nouvelles fonctionnalités par rapport aux techniques existantes, comme une prise en compte de phénomènes non-linéaires par exemple. Elle a donc été étudiée sur la vitesse du vent à 10 mètres et sur la température à 2 mètres (chapitre 2). Nous nous sommes aussi attachés à travailler avec la délicate variable que sont les cumuls de précipitations (chapitre 3). Nous avons pour cela développé des extensions aux méthodes existantes et considéré avec un intérêt tout particulier les précipitations extrêmes. Nous avons ensuite étudié plus généralement cette problématique d'extrêmes dans la prévision d'ensemble en proposant de nouveaux moyens de vérifier la performance de tels systèmes dans des cas extrêmes (chapitre 4).

Ce résumé se poursuit en présentant, pour chaque chapitre de la thèse, les motivations relatives aux problèmes étudiés, les innovations apportées et les résultats de notre travail. Ainsi, la Section 1.2 présente la comparaison sur la température et le vent en surface d'une technique non-paramétrique de post-traitement face aux méthodes constituant la référence dans le domaine. La Section 1.3 aborde la question du post-traitement du cumul de précipitations sexti-horaire. On présente pour cela des méthodes, paramétriques et non-paramétriques, spécifiquement adaptées au post-traitement de tels phénomènes. La Section 1.4 introduit une mesure de qualité d'une prévision d'ensemble pour les événements extrêmes, basée sur la théorie des valeurs extrêmes. La Section 1.5 conclut et résume les principaux sujets abordés et résultats obtenus dans la thèse.

1.1.1 Mémento

Comme le vocabulaire météorologique est parfois différent du langage utilisé par la communauté statistique, le Tableau 1.1 ci-après résume les principales notations, abréviations et définitions que vous pourrez rencontrer au cours de votre lecture.

TABLE 1.1 – Table des définitions et acronymes les plus importants.

Terme	Définition
PEARP	système de prévision d'ensemble de Météo-France
PEAROME	système de prévision d'ensemble haute résolution de Météo-France
ECMWF	Centre Européen de Prévisions Météorologiques à Moyen Terme
calibration (fr.)	synonyme de post-traitement
paramètre météorologique	variable météorologique (abus de langage)
ensemble brut	ensemble sorti directement du modèle, non calibré
membres échangeables	membres équiprobables
BMA	Bayesian Model Averaging
EMOS	Ensemble Model Output Statistics
QRF	Quantile Regression Forests
GF	Gradient Forests
GEV	distribution des valeurs extrêmes généralisées
EGP	extension de la distribution Pareto généralisée
EVT	théorie des valeurs extrêmes
CRPS	Score de Probabilité des Rangs Continu. Généralisation de l'erreur absolue moyenne aux distributions
CRPS(F, y)	CRPS instantané observé : valeur scalaire
$\mathbb{E}_{Y\sim G}(\text{CRPS}(F, Y))$	espérance du CRPS par rapport à la distribution G des observations
CRPS(F, Y)	variable aléatoire, correspondant au CRPS instantané pour lequel l'observation est aléatoire (F connue)
score propre	score résumant les caractéristiques d'une prévision, l'espérance d'un score propre est minimisée si et seulement si $F = G$
fiabilité/reliability	accord entre probabilité prévue et fréquence observée d'un événement
résolution/resolution	capacité de la prévision à se différencier d'une prévision climatologique
calibration (mot anglais)	fiabilité + résolution
acuité/sharpness	attribut de la prévision à être la moins dispersée possible (sous couvert de fiabilité)
valeur économique/value	capacité de la prévision à aider à prendre une décision (une prévision climatologique n'a aucune valeur)
FF10m	vitesse du vent à 10 mètres
T2m	température à 2 mètres
RR24	cumul journalier de précipitations
RR6	cumul sexti-horaire de précipitations

1.2 Calibrated Ensemble Forecasts using Quantile Regression Forests and Ensemble Model Output Statistics

1.2.1 Motivations

Actuellement, le modèle de prévision d'ensemble le plus abouti à Météo-France est la PEARP (Prévision d'Ensemble du modèle ARPège) (Descamps et al., 2015). Ce système n'est actuellement calibré que pour deux variables (vitesse du vent et cumul de précipitations) par une méthode déjà ancienne (Hamill and Colucci, 1997). Dans son rapport pour l'ECMWF, Gneiting (2014) discute de la potentielle application opérationnelle de techniques faisant office d'état de l'art dans le domaine. Des méthodes présentées, les plus utilisées sont la BMA (Bayesian Model Averaging) (Raftery et al., 2005) et l'EMOS/NR (Ensemble Model Output Statistics/Non-homogeneous Regression) (Gneiting et al., 2005). Notre intérêt pour la BMA est moindre dans la mesure où l'on considère la PEARP comme un ensemble dont les membres sont échangeables (équiprobables) et que cette technique est plus adaptée aux ensembles multi-modèles. L'EMOS fournit en sortie une distribution prédictive paramétrique de la forme

$$y|x_1, \ldots, x_N \sim f(y|x_1, \ldots, x_N),$$

où à gauche nous avons la distribution de la variable météorologique d'intérêt y conditionnellement aux prédicteurs (le plus communément les membres de l'ensemble) et à droite la densité de probabilité f dont les paramètres sont classiquement estimés par une régression linéaire.

A titre d'exemple, pour la température et la pression, Gneiting et al. (2005) utilisent des lois normales. En notant $\mathcal{N}(\mu, \sigma^2)$ une loi normale d'espérance μ et de variance σ^2, la loi produite par l'EMOS pour la température ou la pression est

$$y|x_1, \ldots, x_N \sim \mathcal{N}(a_0 + a_1 x_1 + \cdots + a_N x_N, b_0 + b_1 s^2),$$

où s^2 représente par exemple la variance des membres de l'ensemble. Les coefficients $a_0 \in \mathbb{R}$ et $a_1, \ldots, a_N, b_0, b_1 \geq 0$ sont généralement estimés sur une période d'apprentissage glissante, par la minimisation d'un critère de performance (généralement la vraisemblance ou le CRPS, généralisation de l'erreur absolue moyenne aux distributions).

Ces deux dernières méthodes, et tout particulièrement l'EMOS, sont déjà performantes. Elles peuvent néanmoins être améliorées :

– dans Gneiting (2014), on envisage des ajustements de paramètres utilisant des fenêtres glissantes. L'avantage est de pouvoir se passer d'un historique de données conséquent mais peut entraîner une certaine inertie des prévisions, ce qui peut poser problème lors de changements de temps. Par ailleurs, cela requiert le stockage des données online, ce qui malgré la parcimonie de ces méthodes peut s'avérer lourd en terme d'entrées/sorties ;

– elles reposent sur des hypothèses criticables sur la distribution de la variable à calibrer, laquelle est généralement déduite de considérations climatologiques. Si l'on trace des histogrammes de température ou de vitesse du vent observés, l'ajustement par un modèle gaussien (dans le cas des températures) ou par un modèle gamma (pour le vent) est tout à fait admis. En revanche, rien ne permet de justifier que la distribution prévue des températures un jour donné doive suivre une loi normale. C'est même à l'encontre de l'un des buts initiaux de la prévision d'ensemble qui est d'isoler et de probabiliser différents types de scenarii possibles. Enfin, l'estimation des paramètres des distributions prévues est faite via des algorithmes qui peuvent ne pas converger numériquement ou au contraire être trop simplifiés.

Dans cette perspective et afin d'améliorer la calibration de la PEARP, nous souhaitons utiliser ici des méthodes non-paramétriques. Par opposition aux méthodes paramétriques, les lois ne sont pas spécifiées : on fonctionne alors avec un minimum d'hypothèses.

Ce chapitre présente une nouvelle méthode de post-traitement non-paramétrique basée sur les forêts aléatoires, appelée Quantile Regression Forests (QRF). Cette méthode mise au point par Meinshausen (2006) est une régression non-linéaire adaptée à la prévision de quantiles. La plupart des méthodes citées précédemment se contentent généralement de prédicteurs issus de prévisions brutes de la variable à calibrer. QRF est construite à partir d'arbres sur tout un panel de prédicteurs envisageables : cela peut être évidemment des membres d'une prévision d'ensemble non calibrée mais aussi une donnée de vent, de pression ou encore un mois de l'année ; ce qui fait sa grande force. Cette méthode a de plus l'avantage de ne pas être dégradée par l'ajout de prédicteurs peu ou pas informatifs, au contraire des méthodes classiques.

QRF construit des forêts aléatoires où toutes les valeurs des feuilles sont cette fois conservées (au lieu de la seule moyenne comme dans la régression par forêts "classique"). Cela permet non plus de prédire une valeur scalaire pour une nouvelle observation mais bien la fonction de répartition associée. Il suffit ensuite d'inverser la fonction de répartition pour obtenir les quantiles prévus désirés. Prenons un exemple, illustré par la Figure 1.1. Nous nous plaçons dans $[0, 1]^2$, nous avons donc deux prédicteurs. Nous avons trois arbres et la forêt associée est à droite.

Chaque arbre est construit à partir d'un nombre fixé de couples observations/prédicteurs. Pour chaque arbre, les couples sont tirés aléatoirement avec remise parmi les couples disponibles. Ensuite, pour chacun des arbres (dans la Figure 1.1 ils sont au nombre de trois), nous divisons de façon itérative l'espace selon une règle (classiquement la règle est de minimiser la variance des espaces issus de la division). Pour chaque division, la séparation s'effectue par la valeur optimale d'un prédicteur parmi un ensemble de prédicteurs choisis au hasard (d'où le nom de forêt aléatoire). Passé un critère d'arrêt (un nombre minimal d'observations par région de l'espace par exemple), notre espace est entièrement découpé en sous-parties, appelées feuilles. Si l'on veut prévoir la distribution conditionnelle des observations selon un nouveau vecteur de prédicteurs (symbolisé dans la Figure 1.1 par la croix bleue), nous allons pour chaque arbre sélectionner les observations de la feuille qui accueille la croix bleue.

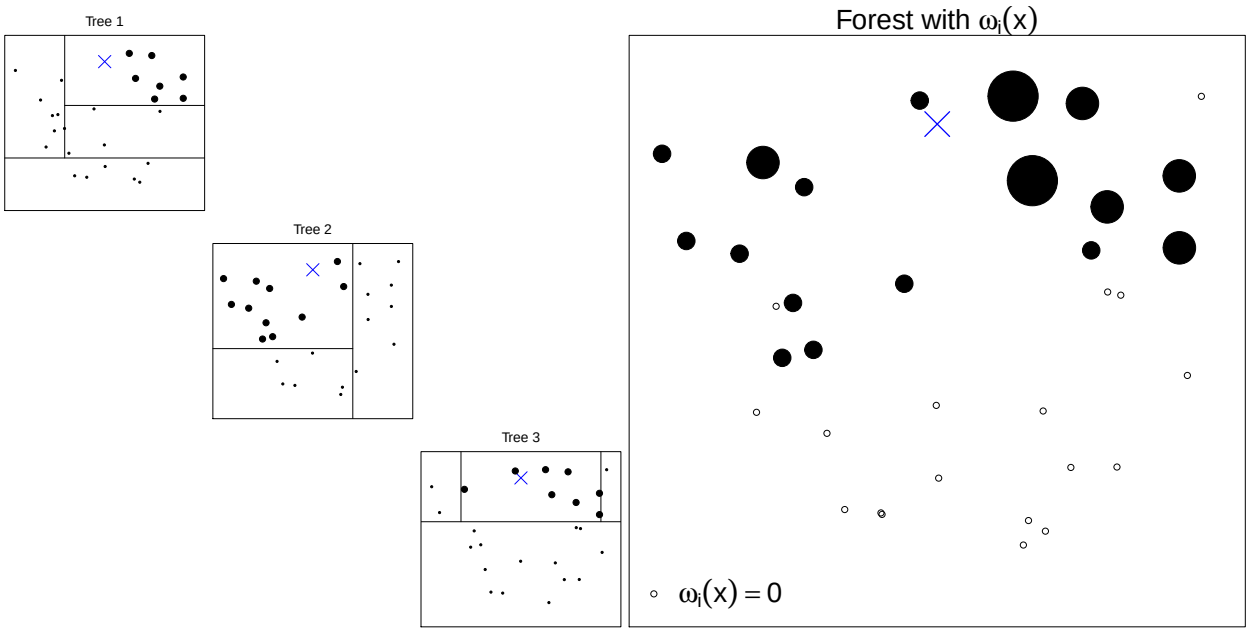

FIGURE 1.1 – Exemple en dimension 2 d'une forêt aléatoire (à droite) composée de 3 arbres (à gauche). On veut prédire une distribution conditionnellement à un nouveau vecteur (croix bleue).

Ces observations sont en gras sur les trois arbres de la Figure 1.1. L'étape finale consiste à agréger les observations de chaque arbre et à les ordonner pour construire une distribution empirique. Bien entendu, les observations étant apparues dans plusieurs feuilles sont donc affublées d'un poids plus gros (les plus gros disques noirs dans la Figure 1.1). A contrario, les observations n'ayant jamais partagé une feuille avec la croix bleue ont un poids nul (les disques vides dans la Figure 1.1).

1.2.2 Résultats

Notre jeu de données s'étend sur 4 années de prévisions PEARP sur 87 stations météorologiques en France (environ une par département) pour des échéances allant de 3 à 54 heures. Nous disposons d'observations de température à 2 mètres (T2m) et de vitesse du vent à 10 mètres (FF10m). Les membres de la PEARP étant échangeables, l'intérêt de la BMA est moindre et nous comparons donc deux méthodes QRF munies de prédicteurs différents avec l'EMOS. La première méthode (QRF_O) n'utilise que des prédicteurs issus du même paramètre météorologique que la variable à calibrer. La seconde méthode (QRF_M) se voit ajouter des covariables issues d'autres paramètres météorologiques. Elle peut ainsi prendre en compte des phénomènes atmosphériquement couplés. Ces trois méthodes sont comparées par validation croisée pour chaque échéance et chaque station. Plusieurs distributions ont été considérées pour l'EMOS, tant pour la T2m que la FF10m.

Cette étude, d'une part, introduit la technique QRF dans le domaine des prévisions météorologiques et, d'autre part, elle procède à une étude comparative avec des méthodes éprouvées [2].

A cette occasion, de nombreuses fonctions analytiques du CRPS ont été établies, voire redécouvertes dans certains cas. Certaines sont désormais disponibles dans le package R "scoringRules" (Jordan et al., 2017).

Une originalité de ce travail est tout d'abord d'introduire des mesures de fiabilité fournissant un diagnostic précis du comportement de l'ensemble (dans quel sens est-il biaisé ? Comment est-il dispersé ?). Nous présentons aussi une mesure plus générale basée sur l'entropie, qui se substitue à des mesures plus classiques utilisant des normes vectorielles. Sur la base de ces mesures, nous pouvons affirmer que la performance en terme de fiabilité est très bonne pour les ensembles calibrés. Cette performance est constante dans les échéances de prévision (si l'on écarte l'EMOS pour la FF10m). Elle est aussi plus homogène d'une station à l'autre par rapport à l'ensemble brut. Pour la T2m, nous remarquons que l'EMOS est meilleure qu'une méthode QRF triviale (QRF_O) mais est moins bonne que QRF_M, une méthode où des covariables portant sur d'autres paramètres météorologiques ont été ajoutées. Pour ce paramètre T2m, l'ajout de covariables est donc prépondérant sur la méthode choisie. Ce n'est en revanche pas le cas pour la FF10m, où les deux méthodes QRF surpassent l'EMOS. Cela s'explique en partie par l'ajout de covariables pour QRF_M mais aussi par une prise en compte des non-linéarités propres au vent par les QRF. Pour ce paramètre, il peut être aussi plus difficile de trouver une distribution paramétrique qui sied convenablement aux distributions de vent, rendant de fait l'EMOS moins performante.

Un point essentiel à conserver de ce travail est aussi la "valeur ajoutée" (c'est-à-dire ce qu'apporte la prévision comme information d'un point de vue décisionnel) des prévisions issues de QRF_M. En effet, l'ajout de covariables cruciales pour le système de prévision d'ensemble peut s'avérer bénéfique pour détecter et corriger des erreurs caractéristiques provenant de phénomènes non-linéaires. Un exemple est donné avec un refroidissement radiatif en hiver avec présence de neige au sol et absence de nébulosité.

1.3 Forest-based Methods and Ensemble Model Output Statistics for Rainfall Ensemble Forecasting

1.3.1 Motivations

Dans un article traitant du gain apporté par le post-traitement des prévisions d'ensemble, Hemri et al. (2014) indiquent que la variable météorologique de surface la plus difficile à calibrer est sans conteste le cumul journalier de précipitations (RR24). A juste titre, cette variable possède en effet une composante à modalités (pluie ou non-pluie), associée à une composante non-nulle qui peut prendre des valeurs extrêmes. De nombreuses méthodes ont

2. Les études menant des comparaisons de techniques de post-traitement sont malheureusement bien peu nombreuses au regard de la littérature disponible sur ces dites techniques.

été étudiées comme la régression logistique étendue (Roulin and Vannitsem, 2012; Ben Bouallègue, 2013), la méthode des Analogues (Hamill and Whitaker, 2006), ou encore la BMA (Sloughter et al., 2007). L'EMOS a quant à elle été testée avec de nombreuses distributions (Scheuerer (2014); Scheuerer and Hamill (2015) entre autres). Mais il est peu évident de trouver des distributions paramétriques donnant de bonnes performances en terme de calibration pour tous les événements possibles. À notre connaissance, la seule étude comparative a été menée par Schmeits and Kok (2010). Pour des raisons liées aux attentes d'utilisateurs spécifiques (hydrologie, prévision hydro-électrique etc.), nous nous concentrons dans ce chapitre sur un cumul sexti-horaire de précipitations (RR6) ; exacerbant de fait la difficulté et posant la question du transfert de toutes les études existantes d'un cumul journalier à un cumul de précipitations sur un pas de temps plus court.

La technique QRF peut présenter plusieurs problèmes inhérents à sa construction pour la calibration d'une telle variable :

- la fonction de répartition construite par les QRF est bornée par les valeurs extrêmes des observations de l'échantillon d'apprentissage. Les QRF ne peuvent donc restituer une valeur qu'elles n'ont pas appris ;
- la dichotomie employée propre aux QRF est basée sur une réduction de variance. Il peut exister d'autres règles ou fonctions de perte plus adéquates.

Nous proposons dans ce chapitre de confronter la technique QRF originelle à une méthode appelée Gradient Forests (GF) (Athey et al., 2016). Cette dernière utilise comme règle de dichotomie la fonction de perte associée à la régression quantile (Koenker and Bassett Jr, 1978). De plus, nous utilisons aussi les travaux de Naveau et al. (2016) pour procéder à une extension paramétrique des queues de distributions issues de QRF et GF. Nous créons ainsi une approche "hybride" où un ajustement paramétrique est fait sur un échantillon issu d'une méthode non-paramétrique. Une utilisation de la méthode des Analogues est également testée, dans l'optique de comparer plusieurs techniques non-paramétriques. Nous tentons aussi de l'améliorer en utilisant différents jeux de prédicteurs et différentes métriques pour choisir les analogues. Enfin, nous associons à l'EMOS un travail préliminaire en utilisant à la fois plusieurs distributions mais aussi un algorithme de sélection de variables basé sur les forêts aléatoires (Genuer et al., 2010).

Les apports de ce chapitre sont les suivants. Nous procédons toujours à une étude comparative de méthodes mais cette fois-ci sur un paramètre jamais étudié dans le contexte du post-processing. Nous introduisons une nouvelle distribution dans les méthodes EMOS avec une étude approfondie sur la sélection des prédicteurs. Les méthodes non-paramétriques sont aussi présentes : la méthode des analogues est ici essayée avec différentes métriques et jeux de prédicteurs. La méthode QRF est accompagnée d'une méthode très proche (GF) dont c'est, à notre connaissance, la première confrontation depuis les récents travaux de Athey et al. (2016). Enfin, une approche "hybride" de la problématique est envisagée en associant une distribution paramétrique astucieuse (elle peut en effet représenter tout type de précipitations) à un premier travail issu des algorithmes basés sur les forêts aléatoires.

L'objectif de ce travail est d'apporter une valeur ajoutée à la prévision des événements extrêmes, couplé à une bonne performance globale des ensembles calibrés.

1.3.2 Résultats

Notre jeu de données s'étend sur 4 années de prévisions PEARP sur 87 stations météorologiques en France pour l'échéance 51 heures au réseau de 18 heures. Cela correspond à un cumul portant sur la fin d'après-midi, propice aux phénomènes convectifs. Nous disposons d'observations de RR6 et pour chaque station toutes les méthodes sont comparées par validation croisée. La première chose à signaler est le manque de résolution des analogues, probablement imputable à une profondeur d'archive insuffisante. Néanmoins, on montre dans ce contexte que l'algorithme de sélection de variable par forêts aléatoires constitue une véritable alternative à des méthodes couramment employées dans la recherche d'analogues. L'algorithme de sélection de variables n'est pas efficace pour l'EMOS. Nous pensons que le problème réside surtout dans la façon d'estimer les paramètres des distributions choisies. En effet, les 3 distributions choisies (à savoir la loi de valeurs extrêmes généralisée (GEV) censurée, la loi gamma censurée et la loi Pareto étendue III décrite dans Papastathopoulos and Tawn (2013)) possèdent des paramètres de forme toujours difficiles à estimer. Un travail préliminaire a été effectué pour trouver les configurations optimales pour estimer les paramètres.

Concernant la performance globale, les méthodes "hybrides" surpassent les méthodes QRF et GF mais aussi l'EMOS. On voit ici la difficulté d'ajuster une distribution paramétrique pour un tel paramètre. En terme de pourcentage, le gain de QRF et GF par rapport à l'EMOS en terme de CRPS est comparable à ce que l'on retrouve pour FF10m. Cela est primordial à la lecture de notre constatation de départ sur la complexité du post-traitement des précipitations et la significativité de telles techniques pour ce paramètre météorologique. En outre, nous remarquons bien que la principale amélioration apportée par la calibration a lieu essentiellement sur la fiabilité. Concernant les extrêmes, au lieu de se focaliser sur le caractère discriminant des distributions prédictives (mesurable par l'aire sous une courbe ROC paramétrique), nous nous attachons à regarder la valeur ajoutée des ensembles calibrés par rapport à l'ensemble brut, par le biais du score de Peirce (c'est-à-dire en regardant la position du coin supérieur gauche de la courbe ROC empirique). Non seulement les méthodes avec extension de queues montrent ici leur supériorité, mais il est surprenant que QRF soit meilleur que l'EMOS pour ce critère là.

Au cours de cette étude, nous avons aussi pu nous apercevoir que le CRPS moyen en lui-même était difficile à améliorer dans la mesure où celui-ci se comporte plus ou moins comme l'espérance de la variable à calibrer. Dans le cas de distributions fortement asymétriques comme les précipitations, cette espérance est une mesure très peu informative sur la distribution (compte-tenu de la forte proportion de cumuls nuls). Nous avons donc apporté une explication à la difficulté de calibrer ce paramètre, indépendamment de sa prévisibilité.

Ce lien entre le CRPS et la variable d'intérêt est au cœur des questions que nous nous sommes posées dans le chapitre suivant.

1.4 CRPS-based Verification Tools for Extreme Events

1.4.1 Motivations

La vérification d'une prévision d'ensemble requiert des moyens différents par rapport à une prévision déterministe (Jolliffe and Stephenson, 2012). Le score le plus utilisé dans ce contexte est le CRPS (Epstein, 1969a; Hersbach, 2000; Bröcker, 2012). Ce score propre mesure simultanément la fiabilité, la résolution et l'acuité d'une prévision. La vérification des événements extrêmes ne peut se faire en utilisant le score sur de tels événements uniquement. En effet, ce processus a tendance à favoriser une prévision artificiellement erronée (par exemple qui possède un biais fort, favorisant les fausses alarmes) au détriment d'une prévision moyenne parfaite. C'est ce que Lerch et al. (2017) appellent le "dilemme du prévisionniste". Pour conserver la propreté des scores, indispensable si l'on veut faire une évaluation objective, il convient alors de pondérer les scores traditionnels pour mettre l'accent sur une région spécifique de la prévision. Cette méthode pose néanmoins plusieurs questions. En effet, le choix de la fonction de pondération oriente fatalement le classement potentiel entre différentes prévisions. Cette fonction se détermine généralement par une logique de rapports "coûts sur pertes" (Richardson, 2000; Gneiting and Ranjan, 2011; Patton, 2015). En météorologie, cette approche quantitative est beaucoup plus délicate, il n'existe pas de pondération miracle ou absolue. Et l'information probabiliste est de fait sous-utilisée (Hagedorn, 2017). De plus, la dégénerescence des scores vers des valeurs non informatives est un phénomène classique pour la vérification d'événements extrêmes (Brier, 1950).

Nous voulons donc trouver une alternative à tout ceci. Notre souhait est d'évaluer la performance des prévisions d'ensemble sans avoir à choisir une fonction de pondération, tout en gardant la notion de propreté qui demeure primordiale. L'idée est de s'appuyer sur la théorie des valeurs extrêmes (EVT). Cette approche mêlant post-traitement et EVT a été explorée dans Friederichs and Hense (2007); Friederichs and Thorarinsdottir (2012) entre autres. En outre, l'évaluation de la performance déterministe pour les événements rares utilisant l'EVT rencontre un certain succès (Ferro, 2007; Stephenson et al., 2008; Ferro and Stephenson, 2011). L'objectif est ici de reprendre et développer ces idées, en cherchant plus spécifiquement à définir une mesure compréhensible et répondant de façon pertinente à nos attentes.

Cette partie s'attache donc à comprendre sur un cas simple comment l'approche par scores pondérés (voire même l'utilisation elle-même du CRPS) peut être gênante en météorologie. Nous démontrons diverses propriétés du CRPS pour ensuite les utiliser dans un cadre théorique que nous avons fixé, en utilisant l'EVT. Nous vérifions enfin la cohérence de notre "nouvel indice" sur un cas expérimental et sur des données réelles.

1.4.2 Résultats

Dans un premier temps, plusieurs propriétés du CRPS sont démontrées. Nous montrons en premier lieu que ce score n'est pas naturellement adapté aux valeurs extrêmes, dans la mesure où des valeurs moyennes de CRPS très proches sont réalisables avec des distributions

prédictives aux comportements extrêmes totalement différents. Nous prenons le contre-pied de ce qui se fait actuellement dans les calculs du score “instantané", c'est-à-dire à partir d'une réalisation. Nous nous intéressons ici au comportement du score en le considérant comme une variable aléatoire. Ceci généralise l'idée de "propreté" qui est seulement basée sur l'espérance de cette variable aléatoire. Nous mettons en évidence les liens entre la variable aléatoire CRPS et la variable aléatoire des observations. L'usage de l'EVT et des approximations usuelles nous permet d'affirmer que la loi conditionnelle du score pour une observation extrême est asymptotiquement une loi de Pareto dont nous connaissons les paramètres. Ceux-ci dépendent d'ailleurs à la fois de la distribution climatologique et de certaines quantités moyennes de la prévision (considérée connue ici).

Cette découverte montre en quoi pour certaines variables météorologiques telles que les précipitations, l'amélioration des modèles de prévision d'ensemble ne doit pas se concentrer sur le biais (un meilleur modèle déterministe à l'origine des ensembles) mais plutôt la variance (les schémas de représentation de l'incertitude du modèle ensembliste). C'est un argument fondamental dans la relative difficulté de calibrer certains paramètres mais aussi sur la manière de les calibrer [3].

Le résultat principal de ce chapitre est que nous avons trouvé la loi asymptotique suivie par le CRPS pour des événements extrêmes. Ce résultat demeure asymptotique : il n'est vérifié que si la prévision est mauvaise pour les extrêmes. En sélection de modèles "classique", nous devons choisir, sur la base d'un critère adapté, une distribution en adéquation avec notre échantillon. Ici, c'est donc tout l'inverse, nous avons un échantillon qui ne doit surtout pas suivre une loi fixée. Nous calculons le critère en espérant qu'il soit le plus mauvais possible.

Nous mesurons donc l'adéquation (plutôt la non-adéquation) d'une distribution empirique de CRPS avec une distribution connue. L'utilisation de la statistique de Cramér-von Mises s'impose naturellement (Cramér, 1928; Von Mises, 1928). Nous déduisons une p-valeur de cette statistique qui devient ainsi notre "indice". Il a l'avantage d'être borné dans $[0, 1]$ (c'est pour cela que nous utilisons la p-valeur), et est négativement orienté. En effet, plus la p-valeur est basse et plus l'hypothèse d'adéquation du score et des observations est discutable, et donc meilleure est notre prévision pour ces observations.

Cet indice présente de nombreux avantages : il est basé sur le CRPS non pondéré, qui est une mesure simple et bien connue. De plus, nos simulations et le cadre expérimental du chapitre précédent confirment que cet indice conserve de nombreuses propriétés du CRPS, notamment la propreté et la sensibilité au biais et à la variance. De plus, l'approximation EVT nous permet de disposer de cet indice pour tous les seuils possibles, du moment qu'ils sont élevés. On rejoint ici les propriétés des "diagrammes de Murphy" dans Ehm et al. (2016) pour les prévisions déterministes et notamment la notion de domaine de dominance des prévisions selon le seuil. Notre indice presente des similitudes avec les scores de dépendance extrême dans Stephenson et al. (2008); Ferro and Stephenson (2011). Dans la même veine, il est important de ne pas oublier que notre indice peut être altéré par l'erreur de type II (à savoir que le CRPS peut être grand avec de petites observations) ; ce problème est réglé de la

3. A la lecture du chapitre précédent, nous comprenons comment les méthodes “hybrides" permettent de mieux représenter certains phénomènes, tout en restant sans biais.

même façon que dans Ferro and Stephenson (2011) où il convient de comparer des prévisions ayant un minimum de qualité intrinsèque.

1.5 Epilogue

Pour conclure, cette thèse présente de nouvelles méthodes de post-traitement statistique des prévisions d'ensemble. Celles-ci sont non-paramétriques. En effet, elles se basent sur la technique des forêts aléatoires bien connue en apprentissage statistique. Pour des paramètres météorologiques classiquement étudiés dans ce domaine, comme la température ou la vitesse du vent, l'approche proposée supplante les méthodes existantes en terme de résultats quantitatifs. L'amélioration se fait aussi sur la "valeur économique" des prévisions puisque notre méthode est à même de restituer des phénomènes qui peuvent être gommés par les algorithmes de calibration existants.

Nous avons aussi determiné dans quelle mesure notre méthode (et les méthodes existantes) étaient à même de calibrer correctement le paramètre de précipitations sexti-horaire. Pour cette variable, une nouvelle approche non-paramétrique a été testée, ainsi qu'une méthode hybride permettant l'extension de queues de distribution. Les techniques existantes ont aussi été améliorées, par l'essai de nouvelles fonctionnalités. Notre étude montre encore le bénéfice des méthodes non-paramétriques et aussi des méthodes hybrides, tout particulièrement pour améliorer la valeur économique sur les pluies extrêmes.

La question mathématique de la vérification ensembliste pour les événements extrêmes a aussi été abordée. En effet, le but était d'offrir une alternative mathématiquement adaptée aux scores pondérés. Nous avons pour cela utilisé des propriétés du CRPS qui, combiné à la théorie des valeurs extrêmes, fournit un indice de qualité des prévisions d'ensemble pour les valeurs extrêmes. Cet indice présente l'avantage d'être un indice "propre" au sens de la théorie des scores et d'être basé sur une mesure non-pondérée du CRPS, plus facile d'interprétation.

Cette thèse ouvre la porte à de nouvelles possibilités pour l'utilisation de l'apprentissage statistique en météorologie, par la prise en compte de potentiels phénomènes non-linéaires ainsi qu'à l'interaction entre plusieurs covariables, pas forcément météorologiques (elle peuvent aussi être temporelles voire géographiques). Ce travail sur le post-traitement (pour l'instant en points stations) vise à être généralisé prochainement à une grille pour être opérationnel sur les modèles ensemblistes de Météo-France (PEARP/PEAROME), pour avoir des prévisions fiables et informatives. Nous pouvons aussi discuter des techniques de reconstruction de champs multivariés par l'utilisation de copules empiriques. Enfin, notre indice présenté a d'ores et déjà montré son utilité sur des cas d'études et pourrait être rapidement mis en service pour les prévisionnistes de Météo-France.

À l'intérieur, le soleil cogne comme un boxeur devenu fou.
La pluie viendra laver les hommes et fera pousser les cajous.
En attendant le vent du large, je vais dans le bal du faubourg
Boire de la bière et de la cachaça, danser la nuit, dormir le jour.

Bernard Lavilliers

Chapitre 2

Calibrated Ensemble Forecasts using Quantile Regression Forests and Ensemble Model Output Statistics

This chapter reproduces an article accepted in *Monthly Weather Review*, and written by Maxime Taillardat (CNRM-Météo-France), Olivier Mestre (CNRM-Météo-France), Michaël Zamo (Météo-France) and Philippe Naveau (LSCE-CNRS).

Abstract Ensembles used for probabilistic weather forecasting tend to be biased and underdispersive. This paper proposes a statistical method for postprocessing ensembles based on Quantile Regression Forests (QRF), a generalization of random forests for quantile regression. This method does not fit a parametric Probability Density Function (PDF) like in Ensemble Model Output Statistics (EMOS) but provides an estimation of desired quantiles. This is a non-parametric approach which eliminates any assumption on the variable subject to calibration. This method can estimate quantiles using not only members of the ensemble but any predictor available including statistics on other variables for example.

The method is applied to the Météo-France 35-members ensemble forecast (PEARP) for surface temperature and wind speed for available lead times from 3 up to 54 hours and compared to EMOS. All postprocessed ensembles are much better calibrated than the PEARP raw ensemble and experiments on real data also show that QRF performs better than EMOS, and can bring a real gain for human forecasters compared to EMOS. QRF provides sharp and reliable probabilistic forecasts. At last, classical scoring rules to verify predictive forecasts are completed by the introduction of entropy as a general measure of reliability.

Contents

2.1 Introduction

In recent years, meteorologists have seen the rise of ensemble forecast in numerical weather prediction and its development in national meteorological services. Ensemble forecast is clearly a necessary tool that complements deterministic forecast. Ensemble forecasts seek to represent and quantify different uncertainty sources in the forecast : observation errors or a mathematical representation of the atmosphere still incomplete. In practice ensemble forecasts tend to be biased and underdispersed (Hamill and Colucci, 1997; Hamill and Whitaker, 2006).

Several techniques for the statistical postprocessing of ensemble model output have been developed to square up to these shortcomings. Local quantile regression and probit regression were used for probabilistic forecasts of precipitation by Bremnes (2004). Other techniques of regression like censored quantile regression have been applied to extreme precipitation (Friederichs and Hense, 2007) and logistic regression was employed for probabilistic forecasts of precipitation (Hamill et al., 2008; Wilks, 2009; Ben Bouallègue, 2013). Two approaches are baseline in postprocessing techniques namely the Bayesian Model Averaging (BMA) (Raftery et al., 2005) and the Ensemble Model Output Statistics (EMOS) (Gneiting et al., 2005). Whereas the BMA predictive distribution is a mixture of PDF depending on the variable to calibrate the EMOS technique fits a single PDF from raw ensemble. All parameters of theses PDFs are generally fitted on a sliding training period. In meteorology, BMA has been studied for many variables like for example surface temperature (Raftery et al., 2005), quantitative precipitation (Sloughter et al., 2007), surface wind speed (Sloughter et al., 2010) or surface wind direction (Bao et al., 2010). Meanwhile EMOS techniques have been used for surface temperature (Gneiting et al., 2005; Hagedorn et al., 2008), quantitative precipitation (Scheuerer, 2014), surface wind speed (Thorarinsdottir and Gneiting, 2010; Baran and Lerch, 2015), wind vectors (Pinson, 2012; Schuhen et al., 2012) or peak wind (Friederichs and

Thorarinsdottir, 2012). More recently, Hemri et al. (2014) have applied EMOS to many variables.

In this paper we define a new non-parametric postprocessing method based on Quantile Regression Forests (QRF) developed by Meinshausen (2006). Our QRF method will be compared to EMOS, which is efficient and simple to implement in an operational context by national meteorological services. QRF technique has already been used by Juban et al. (2007) for wind energy and by Zamo et al. (2014b) for photovoltaic electricity production.

The paper is organized as follows : in Section 2.2 we describe the QRF technique in detail and we do a quick reminder about EMOS technique. We explain how we verify ensemble forecasts. Guided by Gneiting et al. (2007) we apply tools like rank histograms and indices to quantify their behavior, in particular we introduce entropy for verification of reliability. Scoring rules like the Continuous Ranked Probability Score (CRPS) is also presented to assess both reliability and sharpness. Section 2.3 presents a case study comparing postprocessing techniques for surface temperature and surface wind speed over 87 French locations at 18 lead times using observations and the French ensemble forecast system of Météo-France called PEARP (Descamps et al., 2015). Data consists in 4 years between 1 January 2011 and 31 December 2014 using initializations at 1800 UTC. Section 2.4 shows general results of postprocessing techniques for studied variables. The QRF forecast and more particularly QRF forecasts based on multi-variable predictors are better calibrated than EMOS forecasts and bring a real gain in comparison to this technique. The paper closes with a discussion in Section 2.5.

2.2 Methods

2.2.1 Quantile Regression Forests (QRF)

For a calibration purpose the QRF method can be linked with the method of analogs (Hamill and Whitaker, 2006; Delle Monache et al., 2013) : its goal is to aggregate meteorological situations according to their forecasts, assuming that close forecasts lead to close observations. So, our QRF method aggregates observations according to their forecasts by iterative binary splitting on predictors. At the end we have for every meteorological situation restored a group of observations which creates an empirical Cumulative Distribution Function (CDF). This method requires a large learning sample but has the advantages to be non-linear and to potentially use others predictors than the raw ensemble forecast only.

We now describe the QRF method and explain the different means used to verify our ensemble forecasts. Let us remember that a quantile of order α is a value x_α such that the probability that the random variable will be less than x_α is α. Thus α is the value of the CDF for x_α.

$$\Pr[X \leq x_\alpha] = \alpha \tag{2.1}$$

While classical regression techniques allow to estimate the conditional mean of a response variable, quantile regression allows to estimate the conditional median or any other quantile of the response variable given a set of predictors (Koenker and Bassett Jr, 1978). Quantile

regression such as QRF consists in building random forests from binary decision trees called *classification and regression trees* (CART) which are presented below. This is a non-linear approach.

Decision trees (CART)

This technique (Breiman et al., 1984) consists in building binary decision trees whose interpretation is very easy. Zamo et al. (2014a) explain this technique in details. The binary decision tree method consists in an iterative split of the data into two groups. This split is done according to some threshold of one of the predictors for quantitative predictors or according to some groups of modalities for qualitative predictors. The predictor and the threshold or grouping are chosen in order to maximize the homogeneity of the corresponding values of the response variable in each of the resulting groups. Homogeneity is defined as the sum of variances of the response variable within each groups : Let $\mathcal{D}_0$ be a group to split and $\mathcal{D}_1$ and $\mathcal{D}_2$ the two resulting groups. The variance of a group is :

$$v(\mathcal{D}_i) = \sum_{y \in \mathcal{D}_i} \left(y - \overline{y}(\mathcal{D}_i)\right)^2 \tag{2.2}$$

With t the threshold or grouping for a predictor in the predictors' space $\mathcal{E}$, we define the homogeneity as :

$$H(t, \mathcal{D}_0) = v(\mathcal{D}_0) - [v(\mathcal{D}_1) + v(\mathcal{D}_2)] \geq 0 \tag{2.3}$$

And we choose t such as :

$$H(t, \mathcal{D}_0) = \max_{t \in \mathcal{E}}(H(t, \mathcal{D}_0)) \tag{2.4}$$

Each resulting group is itself split into two, and so on until some stropping criterion is reached, which can be a minimum number of data or an insufficient decrease in resulting groups' variance. Finally, for each final group (called leaves), the predicted value is the mean of observed values of the variable response belonging to the leaf. In order to avoid overfitting, binary trees are pruned at the splitting level that minimizes the squared error loss function estimated by cross-validation. When one is faced with a new prediction situation, one follows the path in the tree with the value of the situation's predictors until he reaches a final leaf. The forecast value is the mean of the predictand's values grouped in this leaf. Binary regression trees are easily interpretable because they can be represented by a decision tree, each node being the criterion used to split the data and each final leaf giving the predicted value. The interested reader can refer to Hastie et al. (2009) for detailed explanations.

Bootstrap aggregating (bagging)

According to the previous scheme, a tree can be a very unstable model, i.e. very dependent on the learning sample used for estimation. Breiman (1996) proposed to grow several trees and to average thier predicted values to yield a more stable final prediction. This would require a lot of data in order to build enough independent trees. Since such a big amount of data is usually not available, bootstrap samples are usually used to build the trees. This

means that artificial samples of data are simulated by randomly drawing with replacement among the original data. The complexity of the model is tuned with the number of bagged trees, and each individual tree is not pruned. The principle of bagging can be applied to other regression methods than binary trees.

Random forests

Since the binary trees used in bagging are built from the same data, they are not statistically independent and the variance of their mean cannot be indefinitely decreased. In order to make the bagged trees more independent, Breiman (2001) proposed to add another randomization step to bagging. Each split of each bagged tree is built on a random subset of the predictors. Hence, this method is called random forest. As in bagging, the overfitting problem is solved by tuning the number of trees.

Quantile regression forests

Quantile regression forests (Meinshausen, 2006) are a generalization of random forests and give a robust, non-linear and non-parametric way of estimating conditional quantiles. Whereas random forests approximate the conditional mean, quantile regression forests deliver an approximation of the full conditional distribution. In the same way as random forests, a quantile regression forest is a set of binary regression trees. But for each final leaf of each tree, one does not compute the mean of the predictand's values but instead their empirical CDF. Once the random forest is built, one determines for a new vector of predictors its associated leaf in each tree by following the binary splitting. Then the final forecast is the CDF computed by averaging the CDF from all the trees. Thus predictive quantiles are directly obtained from the CDF. By construction, the final CDF is bounded between the lowest and the highest value of the learning sample. For example, it is not possible to forecast a negative quantile of wind speed and QRF is unable to forecast a quantile higher than the maximum measured in the training sample.

Model fitting

QRF method is used with different inputs here. The first, called QRT_O, uses as predictors Only statistics on the variable to calibrate. The second, called QRT_M, contains not only statistics on the variable to calibrate but also on other meteorological variables issued from the ensemble : this is a Multi-variable approach. The lists of predictors are given in 2.6.1 . For these variants, one must fit the number of trees and the size of the leaves. For temperature, the final leaf size is set to 10 and the number of tree is set to 300, a good compromise between quality and computation speed. For wind speed, the final leaf size is 20 and the number of tree is set to 400. Note that these parameters are set empirically by means of cross validation (not shown here).

2.2.2 Ensemble model output statistics (EMOS)

A description of EMOS technique is given in Gneiting and Katzfuss (2014). The EMOS predictive distribution is a single parametric PDF whose parameters depends on the ensemble values. For example, it could be a normal density, where the mean is a bias corrected affine function of the ensemble members and the variance is a dispersion-corrected affine function of the ensemble variance.

Model fitting

EMOS technique was used considering the high resolution forecast called ARPEGE (Courtier et al., 1991), the control member of the raw ensemble and the mean of the raw ensemble as predictors as in Hemri et al. (2014). The parameter vector is estimated by means of a CRPS minimization over the moving training period. Following Scheuerer (2014) we use as initialization vector for a day the vector issued from the optimization at the precedent day. The optimization process is stopped after few iterations to avoid overfitting.

For surface temperature, distributions tried in EMOS are the normal distribution and the logistic distribution. We finally keep the normal distribution which is classical for temperatures. For wind speed, distributions tested are the truncated normal, gamma, truncated logistic and square root-transformed truncated normal following Hemri et al. (2014). This last model performs best and is kept throughout the study. The correct formula for the corresponding CRPS is given in 2.6.2 and we use it for our study.

2.2.3 Assessing sharpness and calibration

Gneiting et al. (2007) proposes to evaluate predictive performance based on *the paradigm of maximizing the sharpness of the predictive distributions subject to calibration*. Calibration refers to the statistical consistency between forecasts and observations. Also called reliability this is a joint property of predictions and events that materialize. Sharpness refers to the spread of predictive distributions and is a property of the forecasts only. For example, a climatological forecast would be reliable, but would have a poor sharpness.

Sharpness

To assess sharpness, we use summaries of the width of prediction intervals as in Gneiting et al. (2007). For example, we can introduce the average width of the central 50% prediction interval, the 90% prediction interval or both. In this study we check the width of the central 50% prediction interval only, we denote it IQR (for interquartile range) in the following results.

The Rank histogram and the PIT histogram

Rank Histograms (RH), also called *Talagrand diagrams* were developed independently by Anderson (1996); Talagrand et al. (1997); Hamill and Colucci (1997). We employ RH to

check the reliability of an ensemble forecast or a set of quantiles. A RH is built by ranking observations according to associated forecasts. Reliability implies that each rank should be filled with the same probability. Calibrated ensemble prediction systems should result in a flat RH. The opposite is not true : a flat RH may not refer to a calibrated system (Hamill, 2001). In a general way, a U-shaped histogram refers to underdispersion or conditional bias, a dome-shaped generally refers to overdispersion while a non-symmetric histogram refers to bias. PIT histogram is the continuous version of the RH and permits to check reliability between observations and a predictive distribution by calculating $Z' = F(Y)$ where Y is the observation and F the CDF of the associated predictive distribution. Subject to calibration the random variable Z' has a standard uniform distribution (Gneiting and Katzfuss, 2014) and we can check ensemble bias by comparing $\mathbb{E}(Z')$ to $\frac{1}{2}$ and ensemble dispersion by comparing the variance $var(Z')$ to $\frac{1}{12}$. We apply this approach to a RH with $K+1$ ranks using the discrete random variable $Z = \frac{rank(y)-1}{K}$. Subject to calibration Z has a discrete standard uniform distribution with $\mathbb{E}(Z) = \frac{1}{2}$ and a normalized variance $\mathbb{V}(Z) = 12\frac{K}{K+2}var(Z) = 1$.

Moreover, Delle Monache et al. (2006) introduces the reliability or discrepancy index for a RH with K+1 ranks :

$$\Delta = \sum_{i=1}^{K+1} \left| f_i - \frac{1}{K+1} \right| = \sum_{i=1}^{K+1} |\epsilon_i| = \|\epsilon\|_1 \tag{2.5}$$

where f_i is the frequency of observations in the ith rank.

We can complete this tool by checking $\|\epsilon\|_2$ (Quadratic index) or $\|\epsilon\|_\infty$ (Max index) which are more sensitive to bigger errors than Δ.

Another tool that we will use to assess calibration is the entropy :

$$\Omega = \frac{-1}{\log(K+1)} \sum_{i=1}^{K+1} f_i \log(f_i) \tag{2.6}$$

For a calibrated system the entropy is maximum and equals 1. Tribus (1969) showed that entropy is a tool for estimating reliability and it is linked with the Bayesian psi-test. Entropy is also a proper measure of reliability used in the Divergence Score described in Weijs et al. (2010).

Reliability diagram

The reliability diagram (Wilks, 1995) is a common graphical tool to evaluate and summarize probability forecasts of a binary event. We use the term *probability* because this tool evaluates a prediction based on a threshold exceedance for a given parameter (the frost probability for example). It consists in plotting observed frequencies against predicted probabilities. Subject to calibration, the resulting plot should be close to the first bisecting line. Nevertheless this tool should be computed with a sufficient number of observations (which is the case in our study) as recalled by Bröcker and Smith (2007a).

2.2.4 Scoring rules

Following Gneiting et al. (2007); Gneiting and Raftery (2007); Gneiting and Katzfuss (2014), scoring rules assign numerical scores to probabilistic forecasts and form attractive summary measures of predictive performance, since they address calibration and sharpness simultaneously. These scores are usually taken to be negatively oriented and we wish to minimize them. A *proper* scoring rule is designed such that the expected value of the score is minimized when the observation is drawn from the same distribution than the predictive distribution.

Following Ferro et al. (2008), if F represents an ensemble forecast with members $x_1, \ldots, x_K \in \mathbb{R}$, a so-called fair estimator of the CRPS (Ferro, 2014) is given by :

$$\widehat{CRPS}(F, y) = \frac{1}{K}\sum_{i=1}^{K}|x_i - y| - \frac{1}{2K(K-1)}\sum_{i=1}^{K}\sum_{j=1}^{K}|x_i - x_j| \tag{2.7}$$

We can also define the skill score in term of CRPS between two ensemble prediction systems, in order to compare them directly :

$$CRPSS(A, B) = 1 - \frac{CRPS_A}{CRPS_B} \tag{2.8}$$

The value of the CRPSS will be positive if and only if the system A is better than the system B for the CRPS scoring rule.

Some theoretical and analytic formulas for CRPS for several distributions are available in 2.6.2.

2.3 Analysis of the French operational ensemble forecast system (PEARP)

We now compare QRF and EMOS techniques for lead times from 3 up to 54-h for forecasts of surface temperature and wind speed over 87 French stations using observations and the French ensemble forecast system of Météo-France called PEARP (Descamps et al., 2015). Data consists in 4 years between 1 January 2011 and 31 December 2014 using initializations at 1800 UTC. Verification and results are made over the years 2013 and 2014. The aim of our study is to compare both techniques according to their specificities and advantages : on the one hand QRF method is non-parametric so it needs a large data sample for learning that is why we employed a cross-validation method (each month of years 2013 and 2014 are retained as validation data for testing the model while the all four years of data without the forecasted month is used for learning). On the other hand, a sliding period of the 40 last days prior the forecast output as in Gneiting et al. (2005); Schuhen et al. (2012); Thorarinsdottir and Gneiting (2010) gives good results for EMOS. But EMOS has to be tuned optimally for a fair comparison that is why for temperature all the data available for each day (4 years

less the forecast day) with a seasonal dependance like in Hemri et al. (2014) is taken. For wind speed, a sliding period of one year gives the best results for EMOS.

For verification, we choose for all methods to form a K-member ensemble from predictive CDFs by taking forecast quantiles at level $i/(K+1)$ for $i = 1, \ldots, K$ respectively, to conciliate with PEARP raw ensemble here $K = 35$. So all scores are computed with 35 quantiles and rank histograms have 36 classes, but for graphical reasons we show RH computed on 12 ranks only (each group of 3 consecutive ranks are gathered as a single rank).

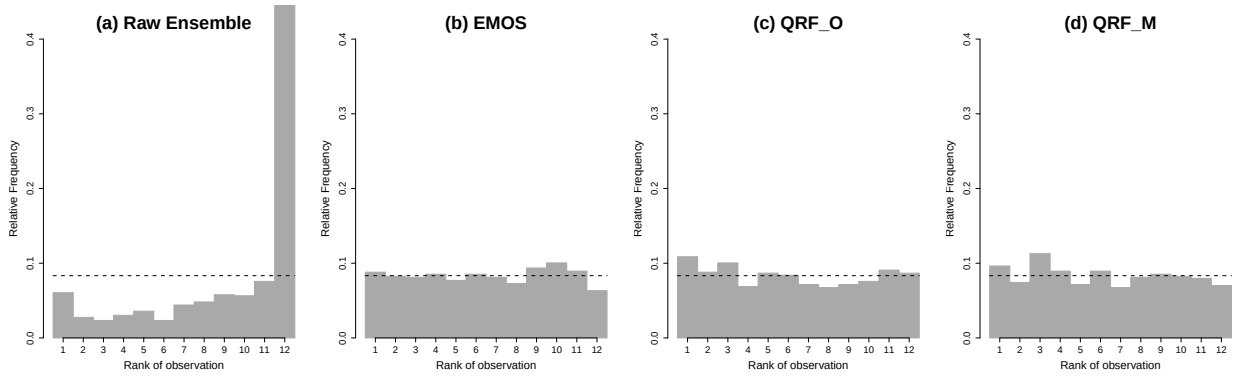

FIGURE 2.1 – Rank Histograms for Lyon airport for 36-h forecast of surface temperature. Raw ensemble is clearly biased and underdispersed. QRF techniques are very efficient.

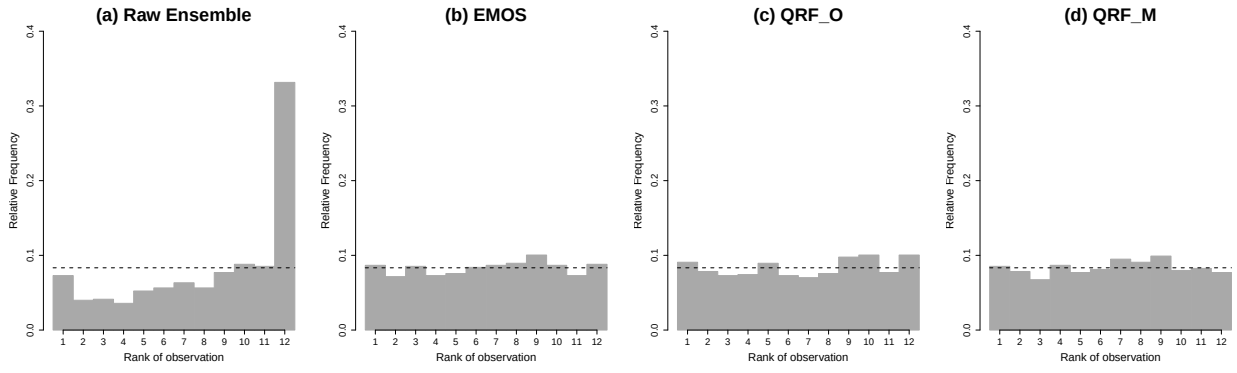

FIGURE 2.2 – Rank Histograms for Paris-Orly airport for 36-h forecast of surface temperature. Raw ensemble is clearly biased and underdispersed. QRF techniques are very efficient.

2.4 Results

2.4.1 Surface temperature

We now give results for surface temperature. We show an example for 36-h lead time (corresponding to 0600 UTC) at two locations which are Lyon and Paris-Orly airports in France. Figures 2.1 and 2.2 show RH for all presented methods. For both examples, the raw ensemble is biased and underdispersive whereas EMOS and QRF techniques show graphically good calibration. Table 2.1 confirms these first results. We can see that the raw ensemble is not reliable and has the worst CRPS. EMOS and QRF techniques are unbiased and dispersion is satisfying. In a general way, lowest CRPS are for QRF_M. It is very interesting to notice that most of the time all indices of reliability (discrepancy index, quadratic index, max index and entropy) exhibit same rankings for the different models. Reliability for EMOS and QRF_O focuses only on the example of Paris-Orly. The discrepancy index shows a better reliability for QRF whereas other indexes penalize this. Thus it is sometimes interesting to assess calibration with several tools.

TABLE 2.1 – Results for surface temperature at two locations for a 36-h forecast. QRF_M performs better than other techniques and gives sharp ensembles.

	CRPS	Δ	$\|\epsilon\|_2$	$\|\epsilon\|_\infty$	Ω	$\mathbb{E}(Z)$	$\mathbb{V}(Z)$	IQR
				Lyon				
Raw ensemble	1.221	0.891	0.38	0.37	0.752	0.762	1.12	1.232
EMOS	0.804	0.175	0.036	0.013	0.992	0.496	0.991	1.874
QRF_O	0.828	0.224	0.048	0.020	0.988	0.482	1.07	1.783
QRF_M	0.790	0.190	0.040	0.019	0.992	0.481	1.00	1.825
				Paris-Orly				
Raw ensemble	0.851	0.578	0.21	0.19	0.895	0.669	1.19	1.278
EMOS	0.694	0.156	0.031	0.010	0.995	0.509	0.996	1.548
QRF_O	0.703	0.150	0.032	0.013	0.995	0.513	1.05	1.450
QRF_M	0.671	0.147	0.032	0.013	0.995	0.507	0.957	1.531

Now let us focus on all stations for 36-h lead time. Figure 2.3 shows RH for the three techniques where a boxplot represents the distribution of a rank for all stations. Results are satisfying, all the RHs are unbiased but we have a "wavy" RH for EMOS whereas the RH for QRF techniques seems to be better. Nevertheless we can assume a slightly U-shaped RH for QRF_O and a slightly dome-shaped for QRF_M, signs of an unperfect dispersion. These first remarks are strengthened by Figure 2.4 where we see that the three calibration techniques are unbiased and QRF techniques are a little more reliable than EMOS technique for discrepancy index (we only show this index of reliability here according to our previous remarks on indices of reliability) but we can assume that results are quite mixed now. The diagnosis of spread ensembles exhibits a slight underdispersion for QRF_O and little overdispersion

for QRF_M even if the boxplot is close to 1. There are three main remarks when we are looking at Figure 2.4. First, we can assume that contrary to the raw ensemble, all boxplots concerning reliability are quite small for the three techniques of calibration : we can say that performances of techniques of calibration for reliability do not depend on location nor time. In addition, we can see that the IQR boxplots for calibrated ensembles are taller than raw ensemble. And last but not least, when we focus on CRPS Skill Score computed with regard to QRF_M for each station we see that almost all the values of the different boxplots are under 0 : not only QRF_M has a better CRPS in general but QRF_M is better in CRPS than all other ensembles and this for almost all stations in this study.

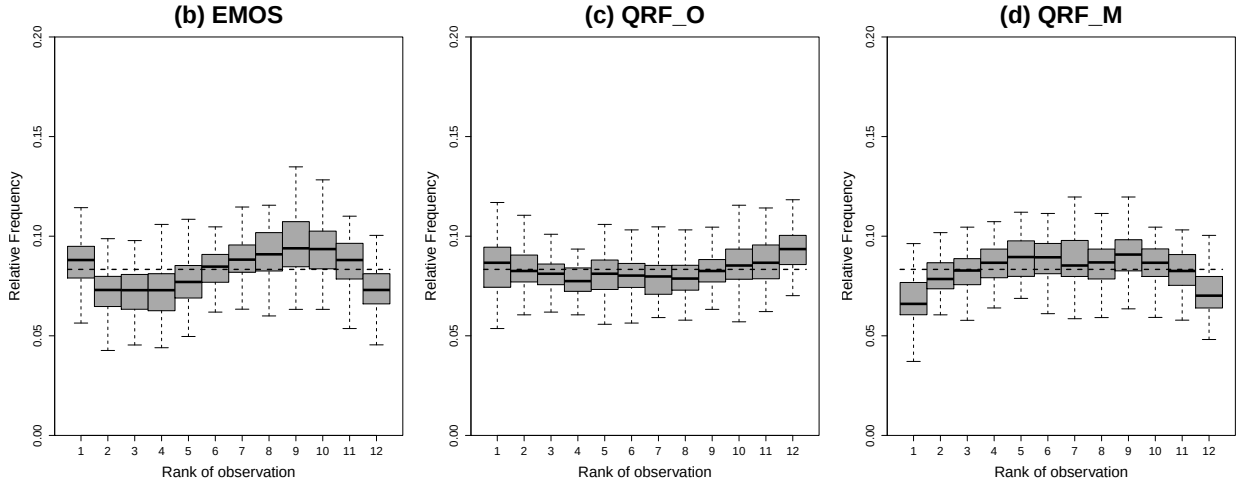

FIGURE 2.3 – Boxplot of Rank Histograms for all locations for 36-h forecast of surface temperature. QRF_M tends to be a little overdispersed. There is a little overdispersion for EMOS

We also investigate performances of probabilistic forecasts of frost for all stations for 36-h lead time. Figure 2.5 shows reliability diagrams for all ensembles. We can see very good performances of calibrated ensembles whereas raw ensemble tends to overpredict frost. This is not surprising since in Figure 2.4 we see that raw ensemble is essentially cold-biased.

We continue this study on surface temperature by showing results across lead times in Figure 2.6. We note that raw ensemble follows a diurnal cycle for all scores. This phenomenon is not shared by calibration techniques concerning reliability but just for CRPS and IQR : we conclude that reliability is not influenced by lead time for calibrated ensembles, only IQR is concerned and thus the CRPS. In addition, the very good entropy of calibrated ensembles (the raw ensemble entropy is around 0.75) let us think that the gain is mainly in reliability. It is interesting to see that raw ensemble does not manage to conciliate good dispersion with small bias. Moreover, reliability of raw ensemble tends to increase among lead times : indeed predictions are less sharp so they can manage to catch the observation in. Besides, we can

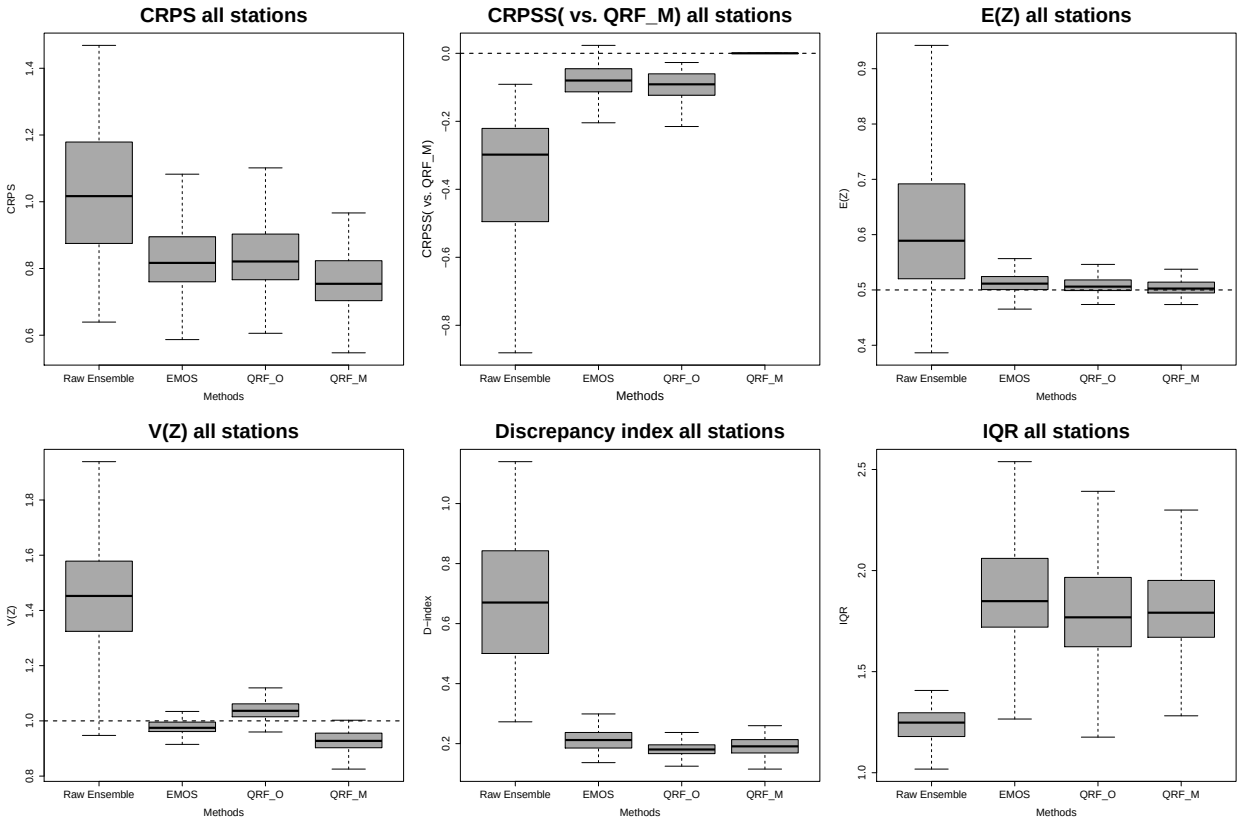

FIGURE 2.4 – Boxplot of different scores for all locations for 36-h forecast of surface temperature. QRF_M technique has better CRPS for almost all stations according to CRPS Skill Score. All calibrated ensembles are unbiased, reliable and quite well dispersed.

note that calibrated ensembles still remain unbiased and reliable with a preference for QRF techniques concerning entropy and quite well dispersed. QRF_O technique is a little bit underdispersed and QRF_M a little bit overdispersed but both are quite close to 1. Last but not least, we see for CRPS that QRF_O and EMOS are very similar and the gap with QRF_M tends to remain the same across lead times. We can explain the gain in CRPS by the introduction of predictors from other variables than surface temperature and shows us all the interest of QRF_M method regarding QRF_O.

Now let us conclude by showing the interest of QRF techniques and in particular QRF_M technique for forecasters. In our opinion, the main issue of the EMOS technique is that it loses one of main aim of ensemble forecasting which is to assess different scenarios from different initial conditions ie. to build different trajectories that can converge or diverge in order to create meteorological scenarios. Indeed, EMOS technique fits a single and unimodal PDF and does not permit to make alternative scenarios. In Figure 2.7 we have four examples of meteorological situations where QRF_M can show all its interest : for Melun in Figure 2.7 we have a situation with snowy ground and clear skies during night causing a rapid

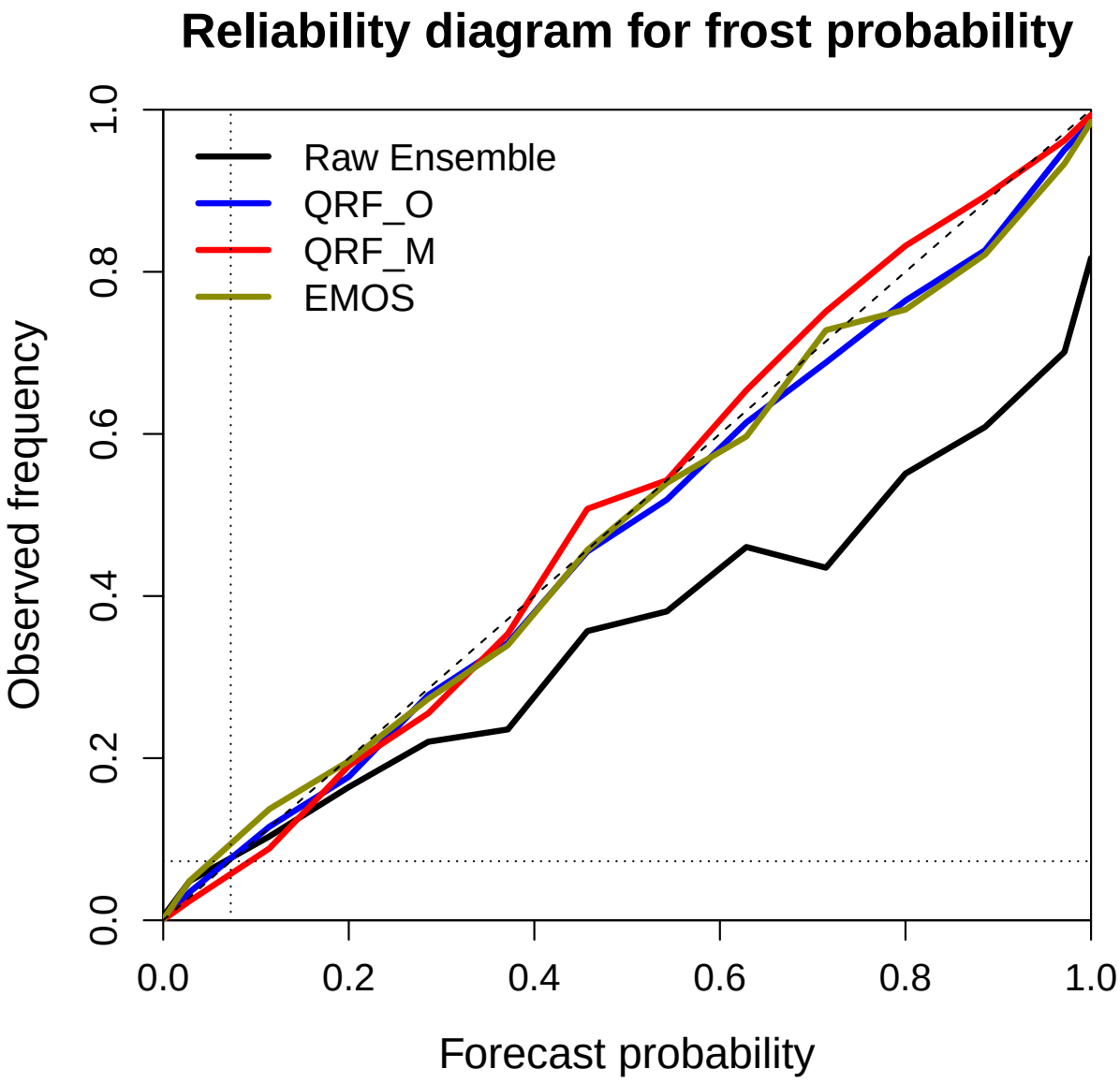

FIGURE 2.5 – Reliability Diagram for probabilistic 36-h forecast of frost for all locations. Dotted lines represent climatology. Calibrated ensembles are almost perfect here.

cooling. Here, even if all calibrated ensembles give a mode around -4 degrees we can see that QRF_M proposes cooler scenarios. The forecaster knowing this phenomenon of rapid cooling would choose this scenario to make a deterministic forecast for example. We can assume here that the combined predictors *snowfall amount* and *surface irradiation* permit to detect a non-linear phenomenon. For the forecast at Carcassonne we see that raw ensemble

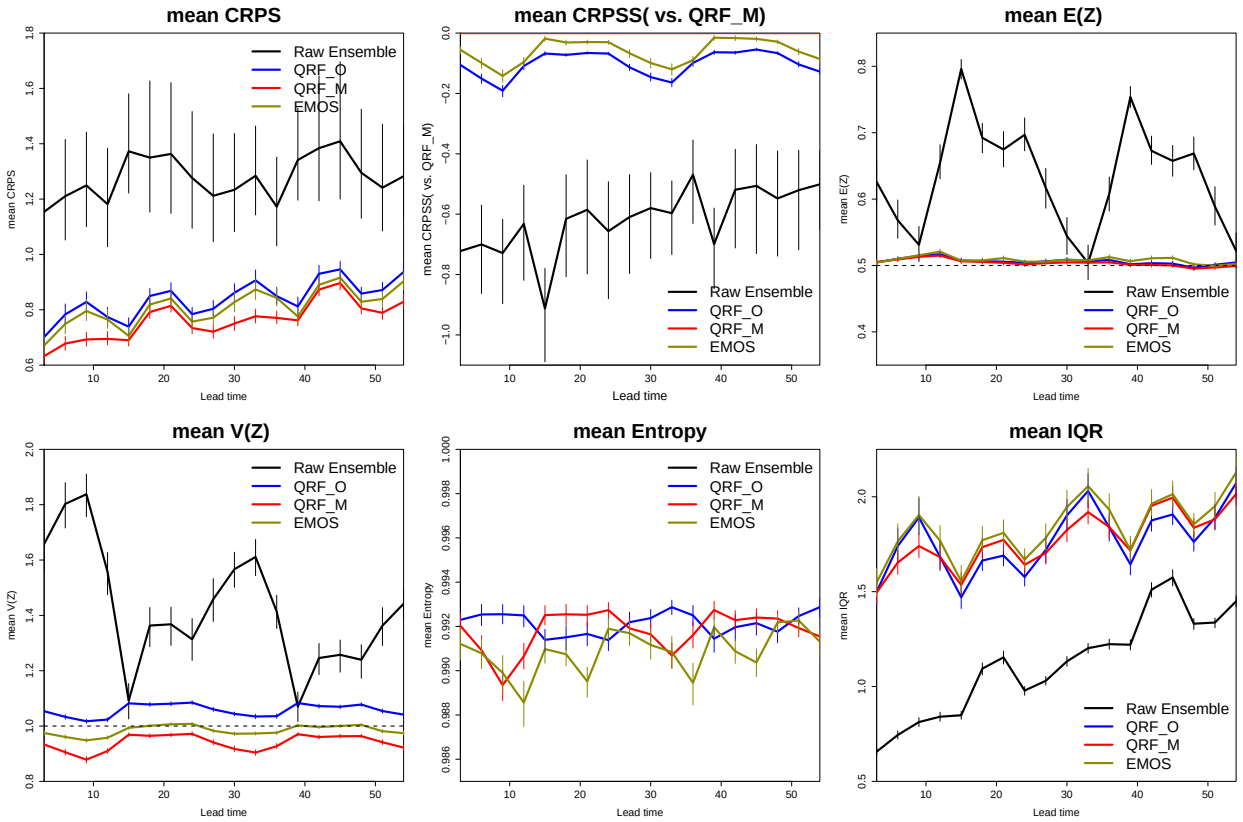

FIGURE 2.6 – Mean scores with 95% bootstrap confidence intervals for all locations across lead times for surface temperature. QRF_M is the best technique for CRPS and CRPSS. Calibrated ensembles are unbiased and in general better dispersed than raw ensemble. QRF techniques tend to provide more reliable forecasts than EMOS (the raw ensemble entropy is around 0.75). Raw ensemble is the sharpest, but it is not reliable anyway.

is bimodal. QRF_M technique is able to detect a situation conducting to a bimodality and so fits a bimodal PDF (and if this bimodality is just an artifact it is an artifact now shared by the raw ensemble and the QRF_M ensemble). Moreover, observation corresponds to the first mode of QRF_M PDF whereas other calibrated ensembles are unimodal. It is the same case for Boulogne-sur-Mer : the bimodal raw ensemble leads to bimodal PDFs for QRF techniques (second modes are between 18 and 19 degrees) and the first mode is preferred and almost corresponds to observation. EMOS technique here fits the PDF in order to avoid mistakes and put its mean between the two raw ensemble modes. It can happen that meteorological situations detected by QRF_M technique lead to a unimodal PDF whereas raw ensemble sees two different scenarios. It is the case of the forecast at Paris-Le Bourget airport where QRF_M does not take into account of (misleading) raw ensemble bimodality and fits its mode between these modalities, and is consistent with observation. Nevertheless we remember that we cannot evaluate ensemble forecasts on single cases. In Figure 2.7, a BMA calibration was

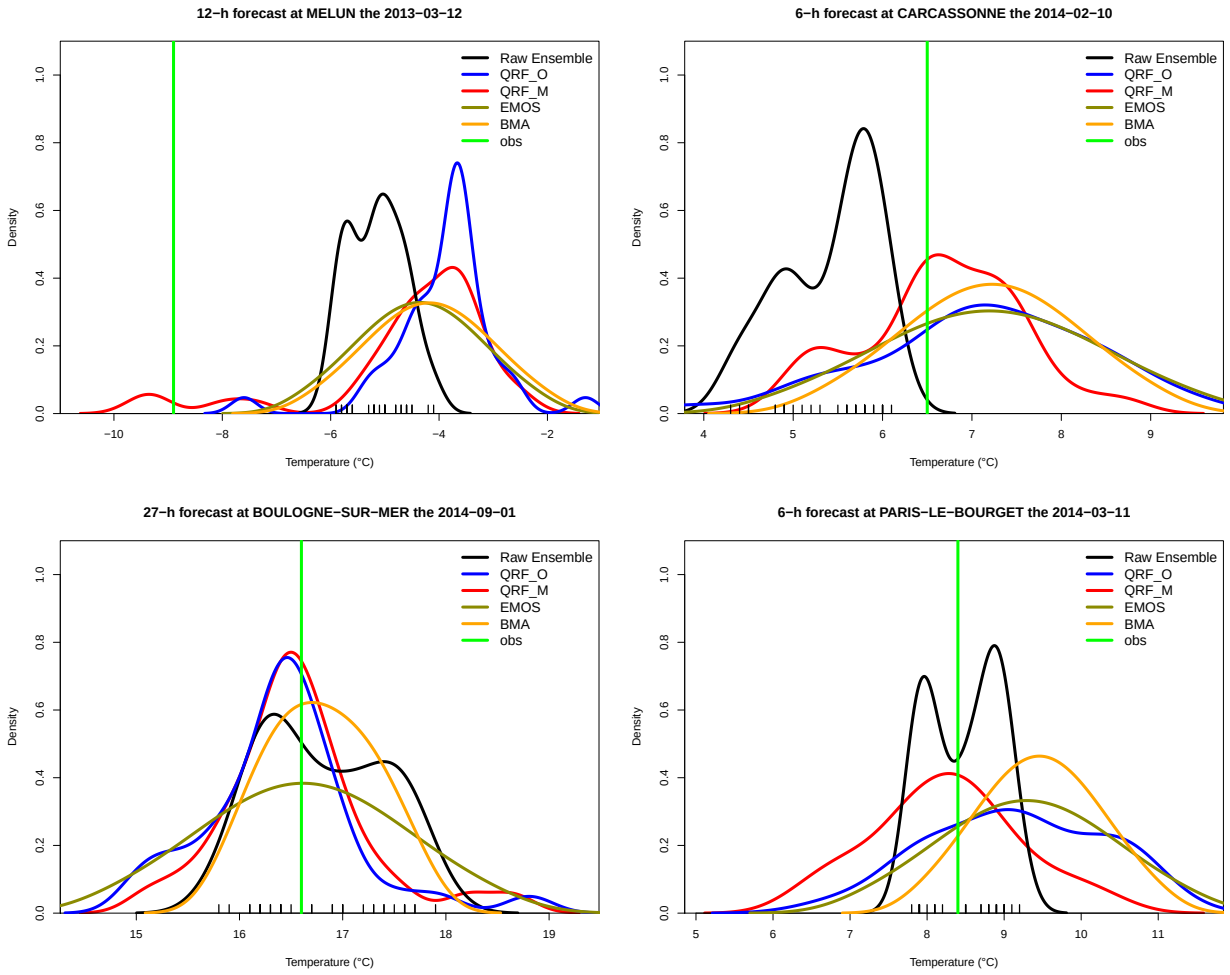

FIGURE 2.7 – Some forecasts for different meteorological situations where QRF_M technique is useful for forecasters. Upper left QRF_M technique proposes cooler scenarios. Upper right the bimodality of raw ensemble is preserved. Lower left bimodality is still conserved but a mode is preferred to the other. Lower right QRF_M technique proposes a unimodal PDF contrary to raw ensemble. Little segments on the x-axis represent the 35 raw members : there are several members associated to the same temperature.

also made with the same learning sample than EMOS. If BMA technique permits bimodalities this is not the case here : we think that the deterministic forecast, the control member and the mean of the raw ensemble are too much close in order to have bimodalities. BMA should be more convenient with ensembles made of several deterministic forecasts.

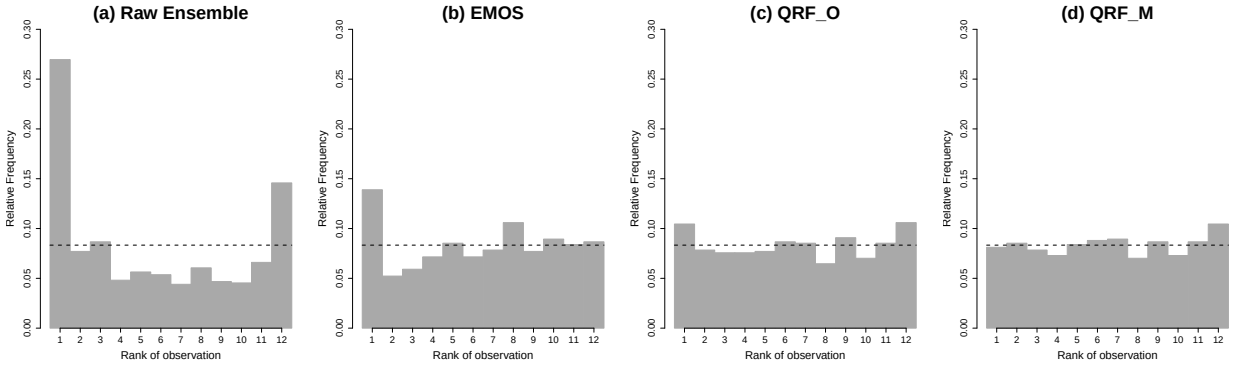

FIGURE 2.8 – Rank Histograms for Lyon airport for 24-h forecast of surface wind speed. Raw ensemble is clearly biased and underdispersed. QRF techniques are very efficient.

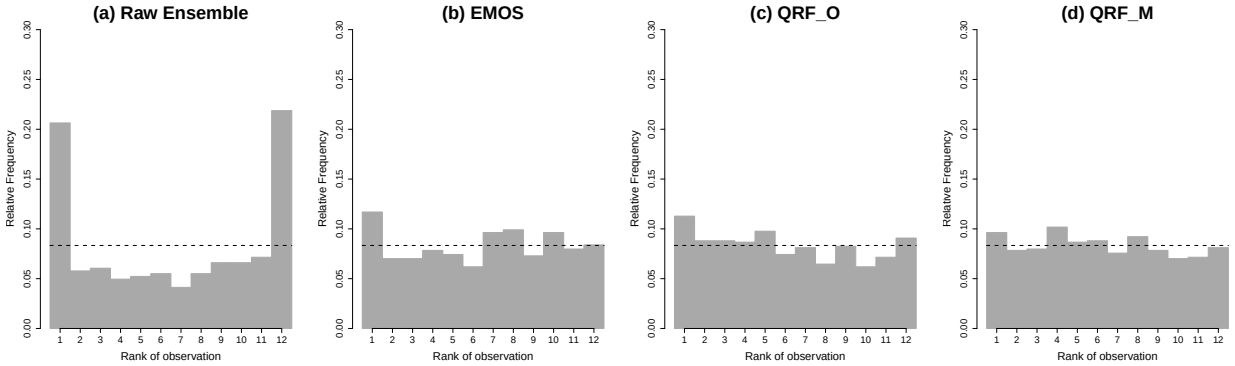

FIGURE 2.9 – Rank Histograms for Paris-Orly airport for 24-h forecast of surface wind speed. This time, raw ensemble is not biased but still underdispersed.

2.4.2 Surface wind speed

We now give results for surface wind speed. Like for surface temperature we choose to begin with an example for 24-h lead time (corresponding to 1800 UTC) at the same locations. Figures 2.8 and 2.9 and Table 2.2 show RH and scores for all presented methods. Mainly the commentaries are the same as for surface temperature. EMOS tends to be a little underdispersed.

Figure 2.10 showing RH for all stations confirms that there is still a little issue with the first rank for EMOS : this is likely due to a sub-optimally chosen distribution type. The square-root truncated normal distribution used here minimizes the average CRPS on whole stations. The form of this distribution may not be optimal for calibrated ensemble forecasting little wind speed. This behavior is similar to the PIT histogram in the middle of Figure 5 of Scheuerer et al. (2015). In the same time we can note that QRF_M dispersion is almost perfect. Figure 2.11 confirms the good dispersion of QRF_M. We also note that calibrated

TABLE 2.2 – Results for surface wind speed at two locations for a 24-h forecast. QRF_M performs better than other techniques and gives sharp ensembles

	CRPS	Δ	$\|\epsilon\|_2$	$\|\epsilon\|_\infty$	Ω	$\mathbb{E}(Z)$	$\mathbb{V}(Z)$	IQR
				Lyon				
Raw ensemble	0.858	0.538	0.19	0.17	0.906	0.422	1.51	1.090
EMOS	0.765	0.241	0.060	0.045	0.984	0.501	1.09	1.595
QRF_O	0.759	0.212	0.045	0.016	0.990	0.504	1.07	1.492
QRF_M	0.735	0.184	0.039	0.019	0.992	0.510	1.03	1.523
				Paris-Orly				
Raw ensemble	0.739	0.526	0.27	0.12	0.917	0.517	1.58	0.9487
EMOS	0.630	0.202	0.042	0.019	0.991	0.498	1.05	1.454
QRF_O	0.656	0.204	0.043	0.019	0.991	0.470	1.06	1.352
QRF_M	0.613	0.176	0.036	0.015	0.993	0.483	0.998	1.318

ensembles seem unbiased and QRF techniques provide reliable ensembles. Last but not least and as for temperatures CRPS Skill Score shows that QRF_M method is the best in terms of CRPS for almost all locations.

We can look at the performance of probabilistic forecast of threshold $5ms^{-1}$ for all stations and 24-h lead time. Figure 2.12 reveals an overprediction of threshold exceedances by raw ensemble, and this feature is corrected by calibrated ensembles. It is not shown here but the results for $10ms^{-1}$ are as good as for $5ms^{-1}$. We have examined the threshold $15ms^{-1}$ but there are not enough observations and the reliability diagram is too noisy to be meaningful. Figure 2.13 shows results across lead times for surface wind speed. If conclusions are strictly the same that for surface temperature, we can add here that sharpness and entropy of QRF ensembles are better than EMOS. Last, QRF techniques are very well dispersed and reliable and thus QRF_O (and QRF_M of course) has much better CRPS than EMOS. We can explain these differences with surface temperature by the fact that finding a good parametric distribution is a little bit more tricky for wind speed than for temperatures and so EMOS performs less well than QRF techniques in that case.

2.4.3 Importance of the QRF predictors

One of the peculiarities of QRF method is that we can see the most useful predictors for the model by watching the *importance* of predictors : the importance shows how much the mean-squared error of a whole forest increases when a predictor is randomly permuted. "Randomly permuted" means that the values of the given predictor are a random sample (without replacement) of the original values. Indeed, if randomly permuting a predictor does not result in a much larger mean-squared error, it means that this particular predictor is of little importance. Whereas important predictors will change the quality of predictions by quite a bit if randomly permuted.

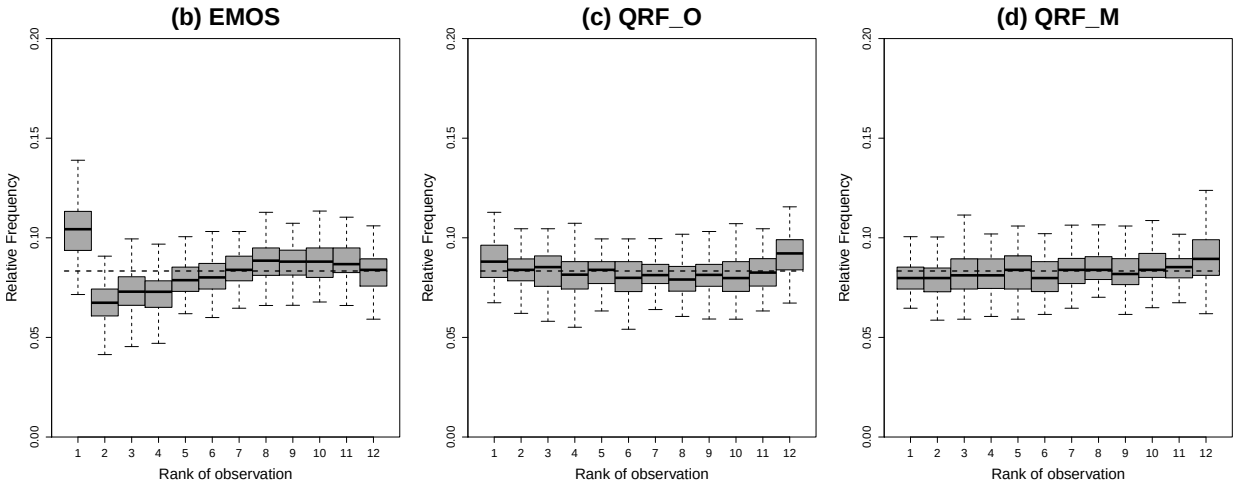

FIGURE 2.10 – Boxplot of Rank Histograms for all locations for 24-h forecast of surface wind speed. QRF RH are almost flat whereas EMOS RH has a high first rank.

Figure 2.14 shows importance of QRF_O predictors for 24-h forecast of surface wind speed. As expected, the most important predictors are those who give an information on the center of the distribution. Next, we have the month (a seasonal information) and the first and the ninth decile. It is interesting to see that informations on spread or other moments are quite useless, they even have same importance that artificially generated random variables (not shown here). We can explain this by the fact that spread information is already contained in decile predictors (in addition to a value on the variable of interest), and it is easier for the model to split meteorological situations by their extreme quantiles rather than their predictability summarized by a statistic such as standard deviation. It is not shown here, but Figure 2.14 also applies to another lead time and the other variable which is surface temperature (with a slightly higher seasonal importance however).

For QRF_M method we focus on surface temperature to show that we can detect a meteorological consistency in QRF model : figures 2.15 and 2.16 show importances for two different lead times (33-h for 0300 UTC and 42-h for 1200UTC). We can assume that both figures have quite the same shape. Indeed, we find that central parameters, deciles and the month are important. In addition, the predictor TPW850 is also important : there is a clear link between surface temperature and potential wet-bulb temperature. If most of other predictors have same importance than noisy predictors, let us focus on FLIR3 and FLVIS3 : for 33-h forecast (day) FLIR3 is higher than FLVIS3 in importance but this is the contrary for 42-h forecast (night). These differences show that QRF model takes into account diurnal and nocturnal radiation (in terms of wavelengths). Last but not least, we note that RR3 and SN3 have small importance : these predictors are often zero and thus permuting zeros does not change anything, explaining small importance. We can understand this phenomenon when we are looking at RR3_q90 and SN3_q90. Higher quantiles are less frequently zero

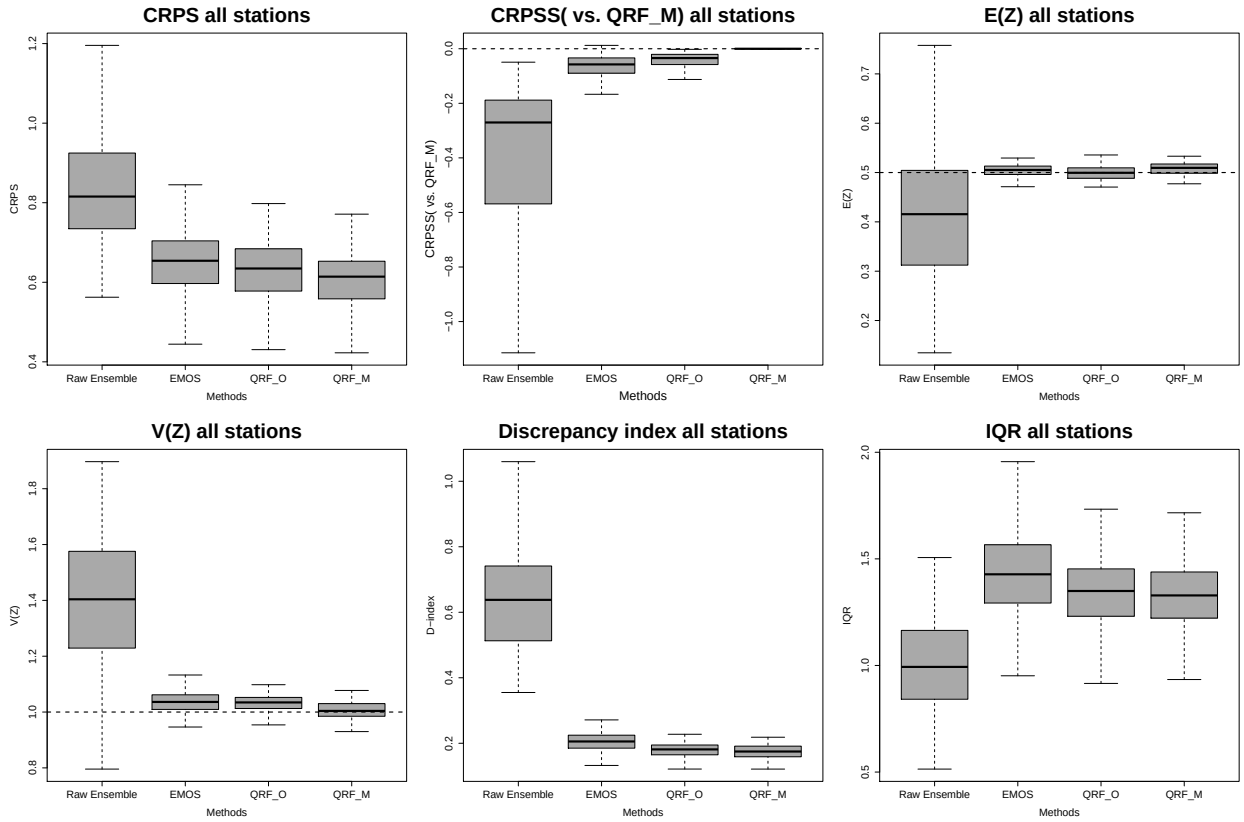

FIGURE 2.11 – Boxplot of different scores for all locations for 24-h forecast of surface wind speed. QRF_M technique has better CRPS for almost all stations according to CRPS Skill Score. All calibrated ensembles are unbiased, reliable and well dispersed even if there is still a little bit of underdispersion for EMOS.

and they have higher importance. Nevertheless we can keep them in the model since we remember that random forests do not choose these predictors during node splitting anyway.

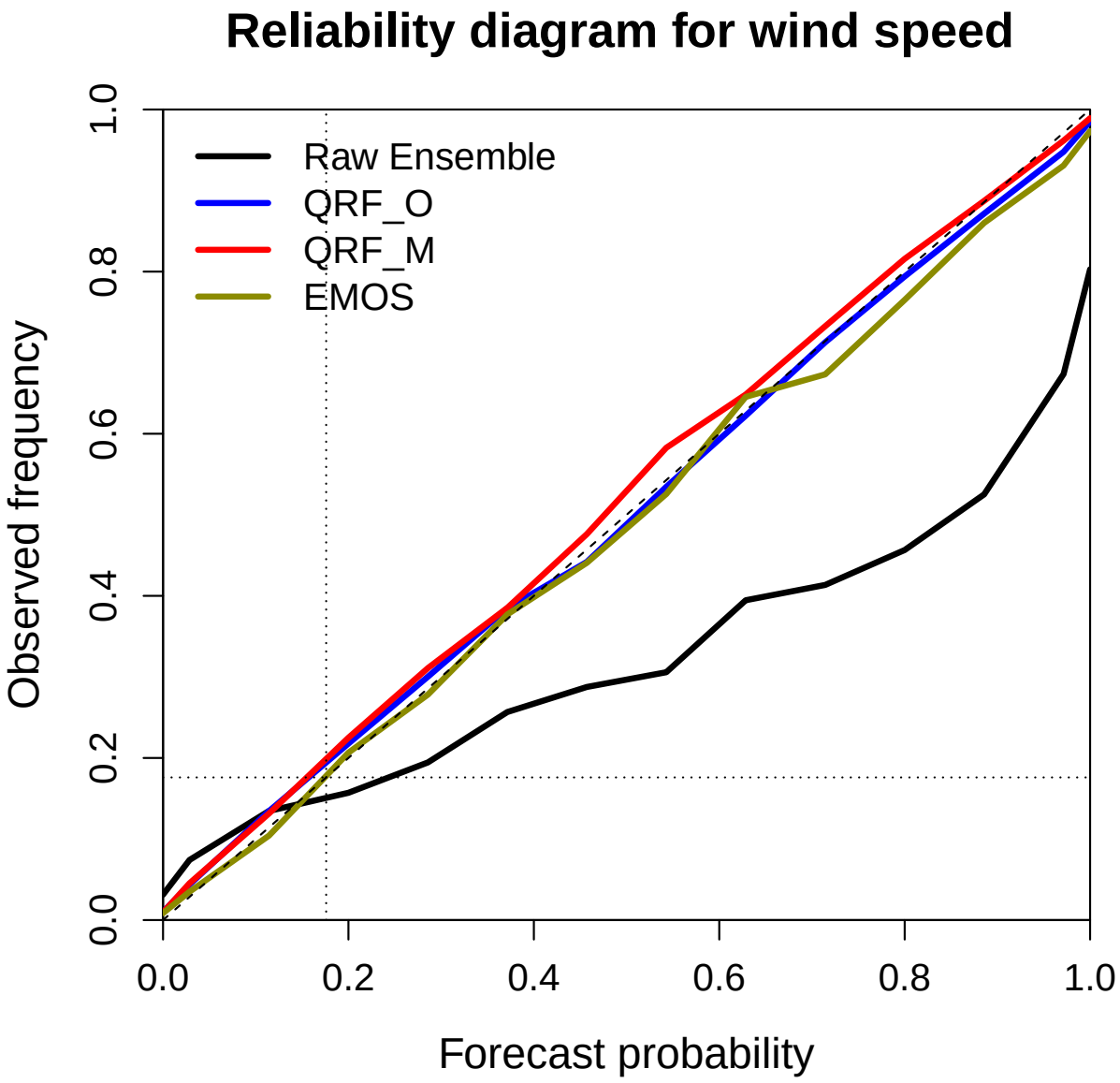

FIGURE 2.12 – Reliability Diagram for probabilistic 24-h forecast of exceedance of threshold $5ms^{-1}$ in all locations. Dotted lines represent climatology. Calibrated ensembles give reliable probabilistic forecasts for this threshold.

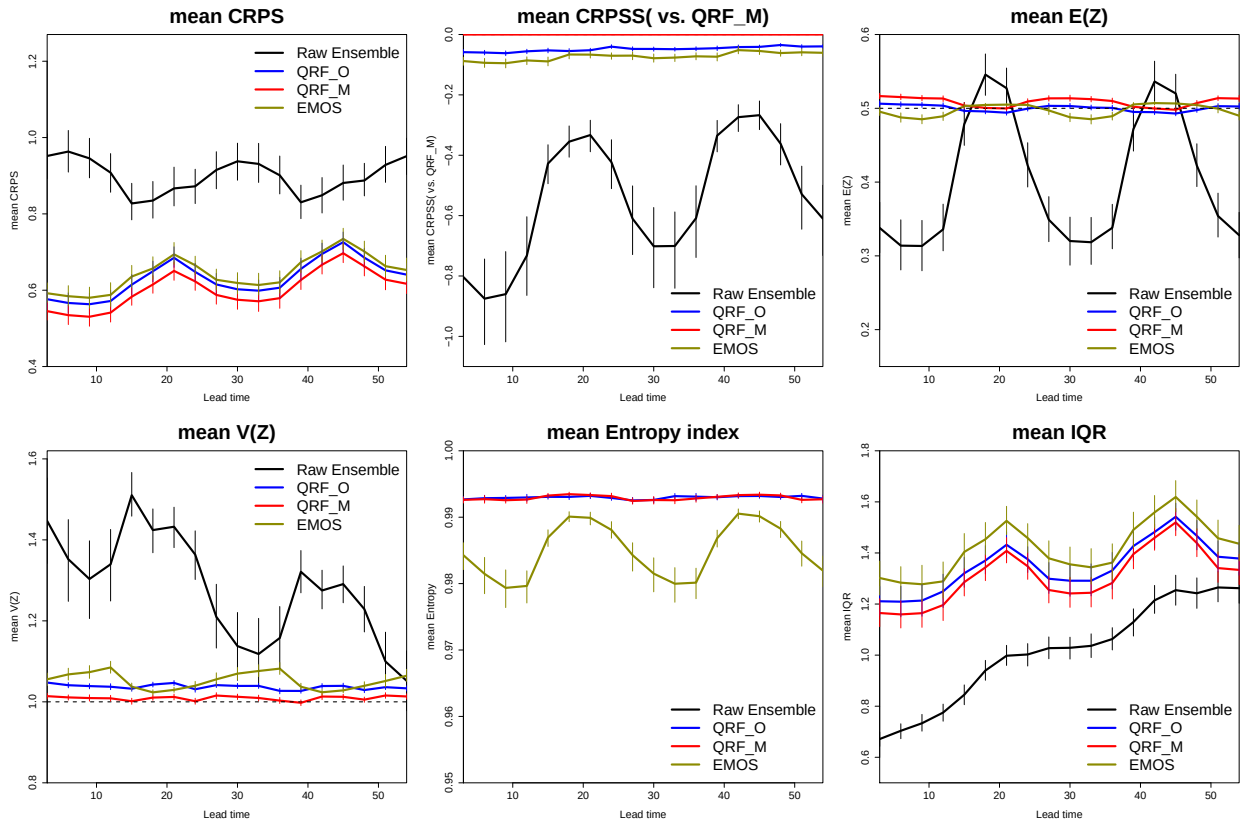

FIGURE 2.13 – Mean scores with 95% bootstrap confidence intervals for all locations across lead times for surface wind speed. QRF_M is the best technique for CRPS and CRPSS. Calibrated ensembles are unbiased and in general better dispersed than raw ensemble (the raw ensemble entropy is around 0.85). QRF techniques tend to provide sharper, more reliable and better dispersed forecasts than EMOS.

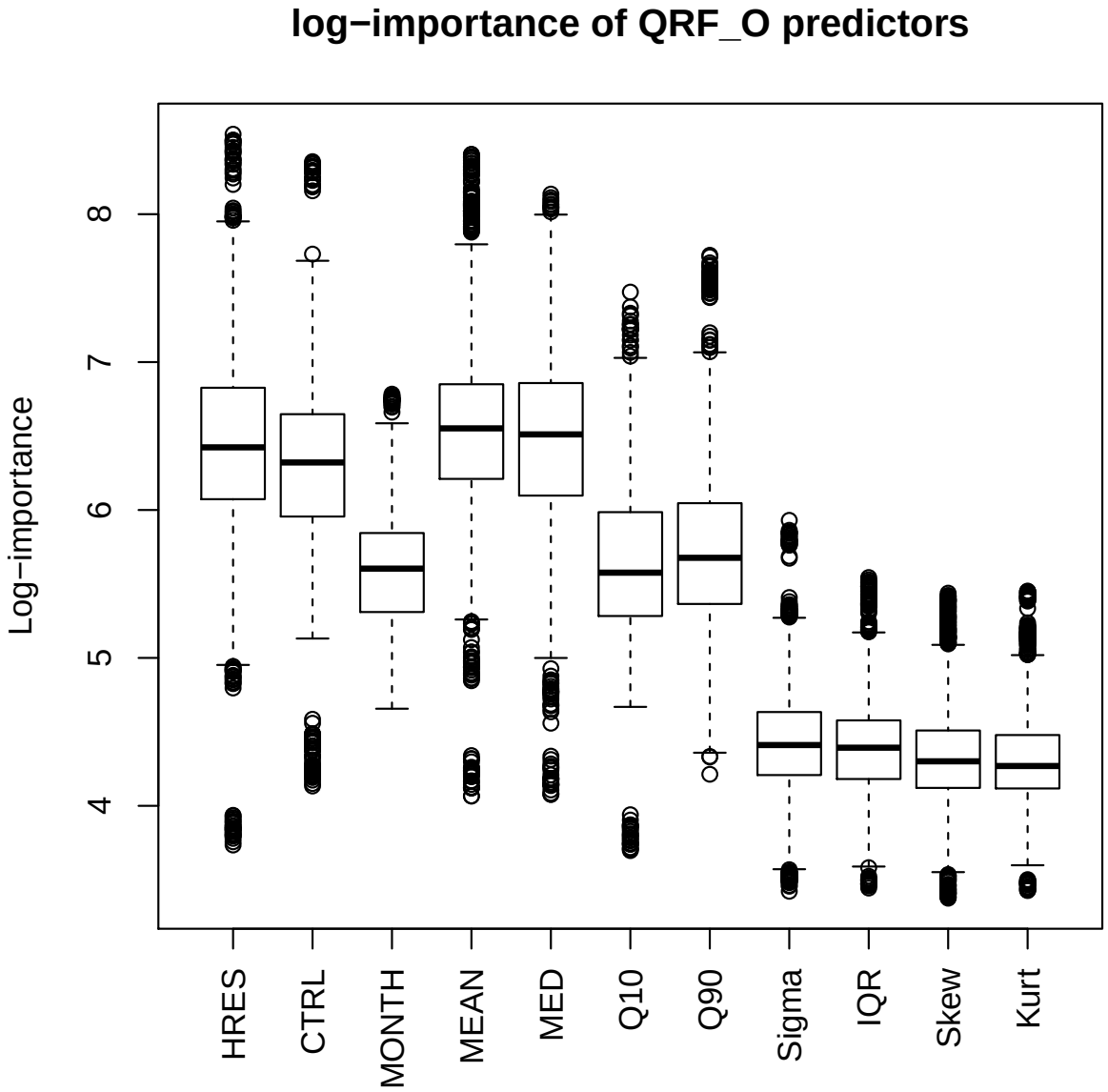

FIGURE 2.14 – Log-importance of QRF_O predictors for 24-h forecast of surface wind speed. A boxplot is composed of measures of log-importance of all the forests and all the stations (so 24 forests x 87 stations = 2088 measures of log-importance per predictor). Most important predictors are those who represent central and extreme locations of the ensemble.

2.5 Discussion

Through this article, we see that the QRF techniques and QRF_M technique which yields on multi-variable predictors give reliable and sharp ensembles compared to EMOS techniques. Moreover, we have noticed that the improvement is more consequent for a non-Gaussian variable like surface wind speed than for surface temperature. This improvement is quite the same among lead times showing that non-parametric calibration methods do not lose predictive performance compared to EMOS and can improve over this method. We also believe that non-parametric calibration are more useful for forecasters since output PDF is not constrained by QRF technique. It allows to keep the notion of scenario for our calibrated ensembles and it can detect non-linear phenomenons. It is not just a correction of bias and dispersion for a given distribution. This non-parametric method is a data-driven technique. This may be viewed as a drawback but the advent of big data and reforecast techniques let us think that non-parametric methods will be frequently used in order to calibrate forecast ensembles and more generally for ensemble output statistics in meteorology. The QRF technique is linked to the method of analogs (Hamill and Whitaker, 2006; Delle Monache et al., 2013) in the sense that QRF is another way to find the closest observations given a set of predictors. The method of analogs consists in finding the closest past forecasts (the analogs) according to a given metric of the predictors' space to build an analog-based ensemble. The QRF technique proceeds by iterative dichotomies on the predictors' space to find the closest past forecasts. So both methods share many advantages (no parametric assumption, easily applicable to multi-predictor settings for example) and drawbacks (large datasets). Moreover, Delle Monache et al. (2013) applied the method of analogs for surface temperature and wind speed on much smaller datasets (and with only three or four predictors) than in Hamill and Whitaker (2006) for rainfall : the size of the dataset is an issue depending on the weather variable under consideration, it will be interesting to check the performances of the analogs technique and QRF with smaller datasets than in Hamill and Whitaker (2006) for rainfall but with many more predictors (we remember that our QRF_M technique uses more than 40 predictors).

In addition, we show out in passing that it is always better to have several methods for assessing performance. Moreover, we have presented some alternatives to interpretate rank histograms other than in a graphic way, by the use of entropy in particular.

As a perspective we will apply QRF techniques to other parameters (rainfall as said above) and try regional calibration : we could add for example predictors like longitude, latitude and altitude to make regional QRF to regroup some stations/grid points in order to have fewer (but bigger) forests and model some spatial interactions. Some works in the same vein have been published recently for EMOS (Feldmann et al., 2014). We will also try techniques for trajectory recovery in ensemble forecasts by using the non-parametric technique of Ensemble Copula Coupling (Bremnes, 2007; Krzysztofowicz and Toth, 2008; Schefzik et al., 2013). We are also interested in combining bias correction for deterministic forecasts to the correction of ensemble forecasts. Last but not least, we could try to use Multivariate regression forests (De'Ath, 2002) directly to make multivariate calibrated forecasts.

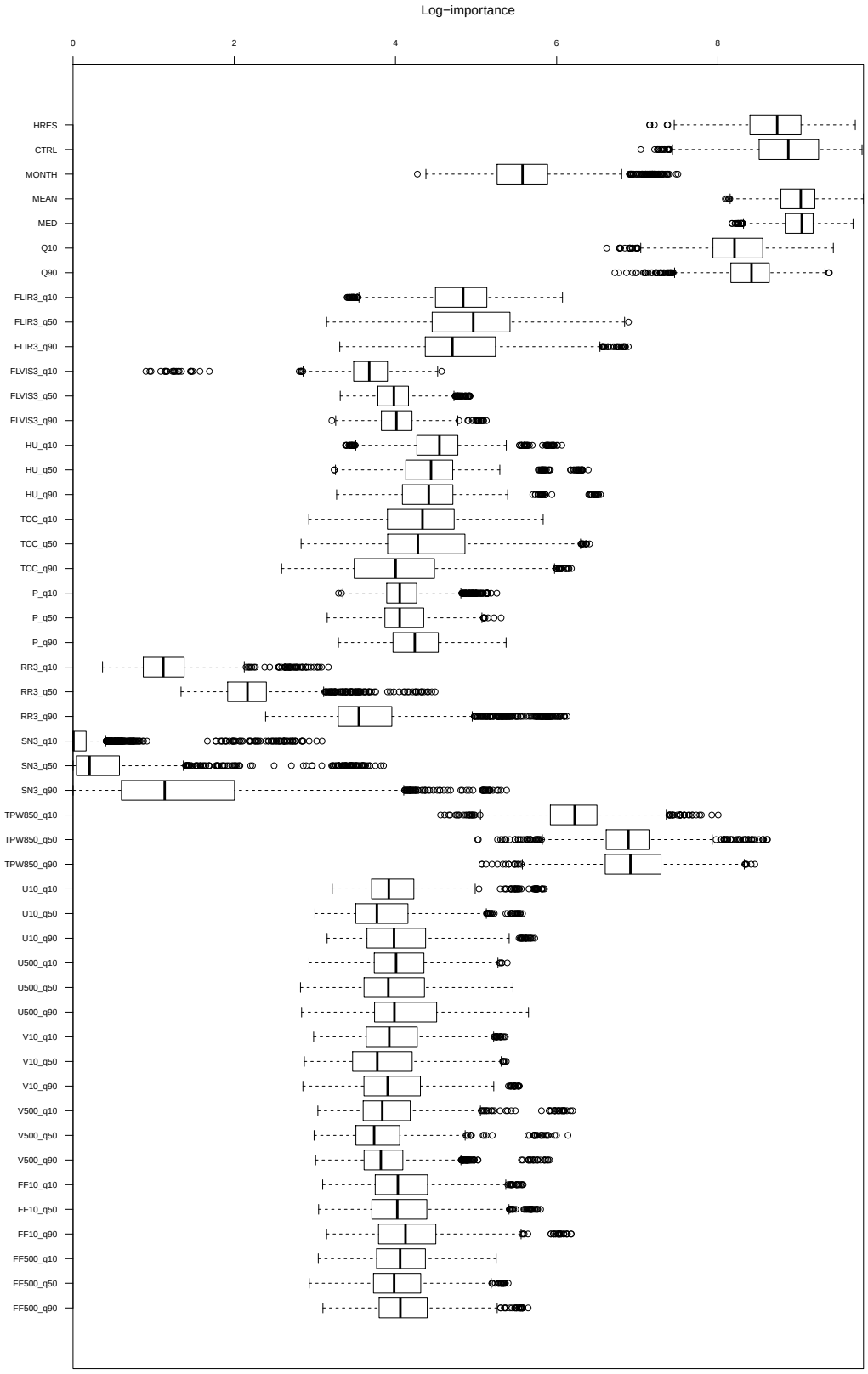

FIGURE 2.15 – Log-importance of QRF_M predictors for 33-h forecast of surface temperature. A boxplot is composed of measures of log-importance of all the forests and all the stations (so 24 forests x 87 stations = 2088 measures of log-importance per predictor). Temperature predictors are the most important with the month. Note the high importance of surface irradiation in infra-red wavelengths.

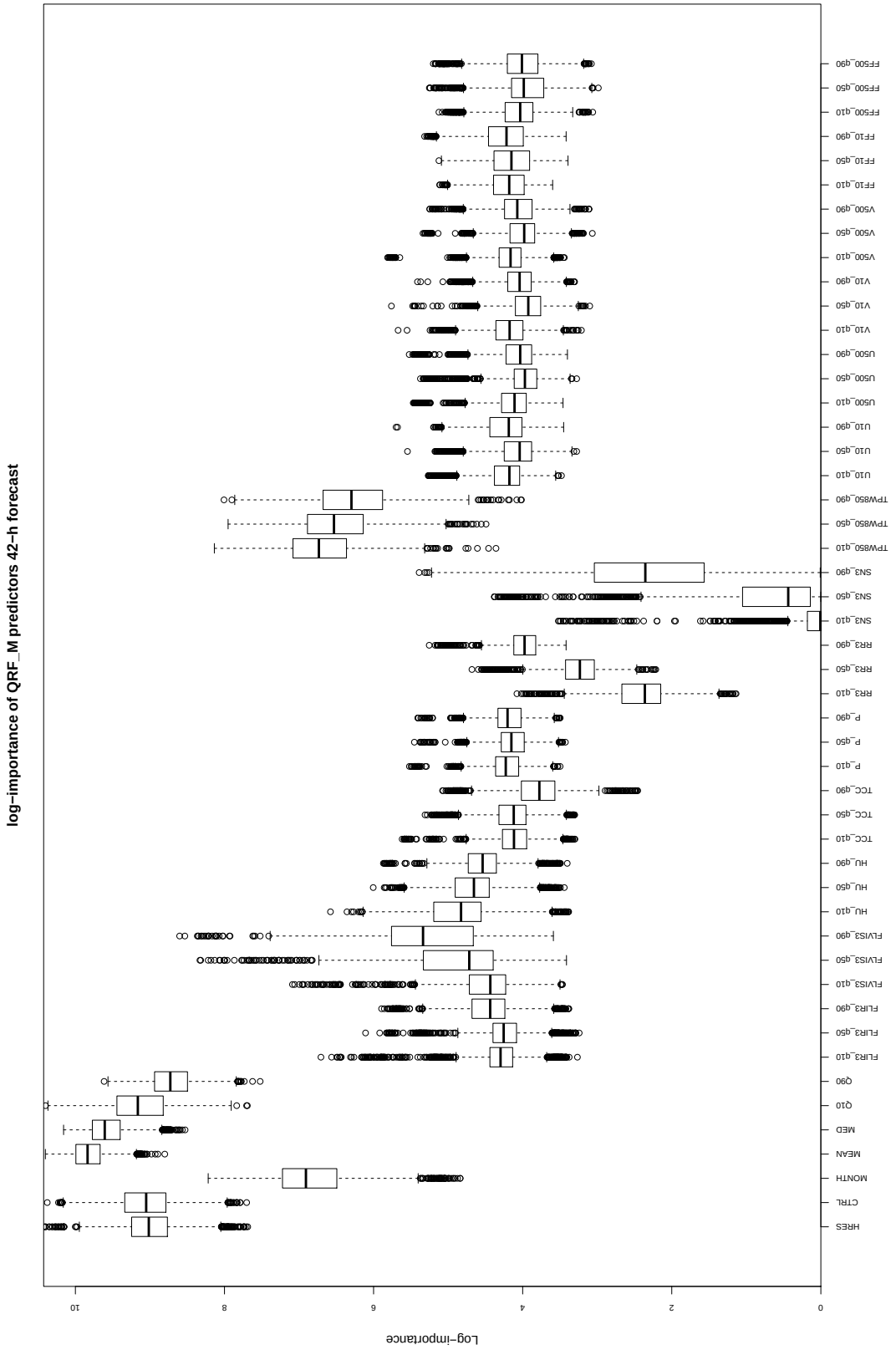

FIGURE 2.16 – Log-importance of QRF_M predictors for 42-h forecast of surface temperature. A boxplot is composed of measures of log-importance of all the forests and all the stations (so 24 forests x 87 stations = 2088 measures of log-importance per predictor). Temperature predictors are the most important together with month. Note the high importance of surface irradiation in visible wavelengths for this lead time.

Acknowledgments

The authors want to thank the three anonymous reviewers for their helpful advices and remarks on this paper. Part of the work of P. Naveau has been supported by the ANR-DADA, LEFE-INSU-Multirisk, AMERISKA, A2C2, CHAVANA and Extremoscope projects. Part of the work was done when P. Naveau was visiting the IMAGE-NCAR group in Boulder, Colorado.

2.6 Appendix

2.6.1 List of predictors for QRF_O and QRF_M

TABLE 2.3 – Lists of predictors for QRF_O.

For surface temperature and surface wind speed
high resolution member
control member
mean of raw ensemble
median of raw ensemble
first decile of raw ensemble
ninth decile of raw ensemble
standard deviation of raw ensemble
IQR of raw ensemble
skewness of raw ensemble
kurtosis of raw ensemble
month of the year

TABLE 2.4 – Lists of predictors for QRF_M.

	surface temperature	both variables	surface wind speed
HRES		high resolution member	
CTRL		control member	
MEAN		mean of raw ensemble	
MED		median of raw ensemble	
Q10		first decile of raw ensemble	
Q90		ninth decile of raw ensemble	
MONTH		month of the year	
Sigma	—		standard deviation of raw ensemble
IQR	—		IQR of raw ensemble
Skew	—		skewness of raw ensemble
Kurt	—		kurtosis of raw ensemble
q10,50,90 are respectively the first decile, the median and ninth decile of the raw ensemble for the following variables :			
HU_q10,50,90		surface humidity	
P_q10,50,90		sea level pressure	
TCC_q10,50,90		total cloud cover	
RR3_q10,50,90		3-h rainfall amount	
SN3_q10,50,90		3-h snowfall amount	
U10_q10,50,90		surface zonal wind	
V10_q10,50,90		surface meridional wind	
U500_q10,50,90		500m zonal wind	
V500_q10,50,90		500m meridional wind	
FF500_q10,50,90		500m wind speed	
TPW850_q10,50,90		850hPa potential wet-bulb temperature	
FLIR3_q10,50,90		3-h total surface irradiation in infra-red wavelengths	
FLVIS3_q10,50,90		3-h total surface irradiation in visible wavelengths	
T_q10,50,90	—		surface temperature
FF10_q10,50,90	surface wind speed		—

2.6.2 List of theoretical formulas and analytic formulas for the CRPS for several distributions

The *Continuous Ranked Probability Score* (CRPS) (Matheson and Winkler, 1976; Hersbach, 2000) is defined directly in terms of the predictive CDF, F, as :

$$CRPS(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}\{x \geq y\})^2 \, \mathrm{d}x$$

Another representation (Gneiting and Raftery, 2007) shows that :

$$CRPS(F, y) = \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'|$$

where X and X' are independent copies of a random variable with distribution F and finite first moment.

Another elegant representation that we found for continuous distributions using the L-moments (Hosking, 1989) is :

$$CRPS(F, y) = \mathbb{E}_F|X - y| - \mathbb{E}_F(X(2F(X) - 1))$$

Here we find some analytic formulas for the CRPS. Some of them are already known and a reference is mentioned (to the best of our knowledge) but the others have been computed. This list permit to sum up some formulas for further studies.

Normal distribution

For $X \sim \mathcal{N}(\mu, \sigma)$,

$$CRPS(X, y) = \sigma \left(\omega(2\Phi(\omega) - 1) + 2\phi(\omega) - \frac{1}{\sqrt{\pi}} \right)$$

where $\omega = \frac{y-\mu}{\sigma}$ and ϕ and Φ are the PDF and the CDF of the standard normal distribution respectively. You can find this formula in Gneiting et al. (2005).

Truncated normal distribution

For $X \sim \mathcal{N}^0(\mu, \sigma)$,

$$CRPS(X, y) = \frac{\sigma}{p^2} \left[\omega p \left(2\Phi(\omega) + p - 2\right) + 2p\phi(\omega) - \frac{1}{\sqrt{\pi}}\Phi\left(\frac{\mu\sqrt{2}}{\sigma}\right) \right]$$

where $\omega = \frac{y-\mu}{\sigma}$, $p = \Phi\left(\frac{\mu}{\sigma}\right)$ and ϕ and Φ are the PDF and the CDF of the standard normal distribution respectively. You can find this formula in Thorarinsdottir and Gneiting (2010).

Square root-transformed truncated normal distribution

For $\sqrt{X} \sim \mathcal{N}^0(\mu, \sigma)$,

$$\begin{aligned} CRPS(X, y) &= (\mu^2 + \sigma^2 - y)\left(1 - 2\frac{\Phi(\omega) - q}{p}\right) + 2\frac{\phi(\omega)}{p}(\omega\sigma^2 + 2\sigma\mu) \\ &\quad - \left(\frac{\sigma}{p}\phi\left(\frac{-\mu}{\sigma}\right)\right)^2 - 2\frac{\sigma\mu}{p^2\sqrt{\pi}}\left(1 - \Phi\left(\frac{-\mu\sqrt{2}}{\sigma}\right)\right) \end{aligned}$$

where $\omega = \frac{\sqrt{y}-\mu}{\sigma}$, $q = 1 - p = \Phi\left(\frac{-\mu}{\sigma}\right)$ and ϕ and Φ are the PDF and the CDF of the standard normal distribution respectively. Notice that this formula is equivalent to but more convenient than the formula proposed in Hemri et al. (2014).

Log-normal distribution

For $X \sim \log\mathcal{N}(\mu, \sigma)$,

$$CRPS(X, y) = 2\mathrm{e}^{\mu + \frac{\sigma^2}{2}}\left[1 - \Phi\left(\frac{\sigma}{\sqrt{2}}\right) - \Phi(\omega - \sigma)\right] + y(2\Phi(\omega) - 1)$$

where $\omega = \frac{\log(y)-\mu}{\sigma}$ and ϕ and Φ are the PDF and the CDF of the standard normal distribution respectively. You can find this formula in Baran and Lerch (2015).

Gamma distribution

For $X \sim \text{Gamma}(p, \lambda)$,

$$CRPS(X, y) = \left(\frac{p}{\lambda} - y\right)(1 - 2\Phi(y)) + 2\frac{y}{\lambda}\phi(y) - \frac{1}{\lambda\mathcal{B}(\frac{1}{2}, p)}$$

where $\mathcal{B}$ is the Beta function and ϕ and Φ are the PDF and the CDF of the Gamma(p, λ) distribution respectively. You can find this formula written to another form in Scheuerer et al. (2015).

Beta distribution

For $X \sim \mathcal{B}(p, q)$,

$$CRPS(X, y) = \frac{p}{p+q}\left[1 - 2\Phi(y; p+1, q)\right] - y\left[1 - 2\Phi(y; p, q)\right] - \frac{1}{p+q}\frac{\Gamma(p+q)\Gamma(p+\frac{1}{2})\Gamma(q+\frac{1}{2})}{\sqrt{\pi}\,\Gamma(p+q+\frac{1}{2})\Gamma(p)\Gamma(q)}$$

where Γ is the Gamma function and $\Phi(; p, q)$ is the CDF of the Beta(p, q) distribution.

Logistic distribution

For $X \sim \text{Logis}(\mu, s)$,

$$CRPS(X, y) = s(2\log(1 + \mathrm{e}^{\omega}) - 1 - \omega)$$

where $\omega = \frac{y-\mu}{s}$.

Truncated logistic distribution

For $X \sim \text{Logis}^0(\mu, s)$,

$$CRPS(X, y) = y - \left(\frac{2p-1}{p}\right)\left(\frac{\mu + s\log(1 + \mathrm{e}^{\frac{-\mu}{s}})}{p}\right) + \frac{s}{p}\left(2\log(1 + \mathrm{e}^{-\omega}) - 1\right)$$

where $\omega = \frac{y-\mu}{s}$ and $p = \frac{\mathrm{e}^{\frac{-\mu}{s}}}{1+\mathrm{e}^{\frac{-\mu}{s}}}$. You can find this formula written to another form in Scheuerer et al. (2015).

Log-logistic distribution

For $X \sim \log \text{Logis}(\alpha, \beta)$ and $\beta > 1$,

$$CRPS(X, y) = \left(\frac{\beta - 1}{\beta^2}\right)\frac{\pi\alpha}{\sin(\pi/\beta)} + y\left[1 - 2\frac{\alpha^\beta}{\alpha^\beta + y^\beta}\ {}_2F_1\left(1, 1; 1 + \frac{1}{\beta}; \frac{y^\beta}{\alpha^\beta + y^\beta}\right)\right]$$

where ${}_2F_1(a, b; c; z)$ is the ordinary hypergeometric function.

Truncated logistic distribution with a point mass in 0

X is a non-negative random variable whose CDF is : $F(x) = \frac{\mathrm{e}^{a+bx}}{1+\mathrm{e}^{a+bx}}$
where a is real and $b > 0$ (the PDF has a Dirac delta in 0 : $\delta(x)F(0)$)

$$CRPS(X, y) = \frac{1}{b}\left(2\log(1 + \mathrm{e}^{a+by}) - \log(1 + \mathrm{e}^{a}) - \frac{1}{1 + \mathrm{e}^{a}} - (a + by)\right)$$

Square-root transformed truncated logistic distribution with a point mass in 0

X is a non-negative random variable whose CDF is : $F(x) = \frac{\mathrm{e}^{a+b\sqrt{x}}}{1+\mathrm{e}^{a+b\sqrt{x}}}$
where a is real and $b > 0$ (the PDF has a Dirac delta in 0 : $\delta(x)F(0)$)

$$\begin{aligned} CRPS(X, y) &= \frac{1}{b^2}\left(4Li_2(-\mathrm{e}^{a+b\sqrt{y}}) - 2Li_2(-\mathrm{e}^{a}) - 2\log(1 + \mathrm{e}^{a}) - 2b\sqrt{y}\log(1 + \mathrm{e}^{a+b\sqrt{y}})\right) \\ &\quad + \frac{a(a+2)}{b^2} - y \end{aligned}$$

where $Li_2(z)$ is the dilogarithm function. These two last distributions are extracted from Wilks (2009).

Generalized Pareto Distribution (GPD) and Generalized Extreme Value (GEV) distribution

Formulas are quite long for these distributions used for extreme values. You can refer to Friederichs and Thorarinsdottir (2012) to get analytic formulas for these distributions.

Von Mises distribution

This distribution is used for circular variables. You can refer to Grimit et al. (2006) to get the analytic formula.

I've looked at clouds from both sides now
From up and down and still somehow
It's cloud's illusions I recall
I really don't know clouds at all

Joni Mitchell

Chapitre 3

Forest-based Methods and Ensemble Model Output Statistics for Rainfall Ensemble Forecasting

This chapter mostly reproduces an article submitted in *Applied Statistics*, and written by Maxime Taillardat (CNRM-Météo-France), Anne-Laure Fougères (Univ Lyon 1-ICJ), Philippe Naveau (LSCE-CNRS) and Olivier Mestre (CNRM-Météo-France).

AbstractRainfall ensemble forecasts have to be skillful for both low precipitation and extreme events. We present statistical post-processing methods based on Quantile Regression Forests (QRF) and Gradient Forests (GF) with a parametric extension for heavy-tailed distributions. Our goal is to improve ensemble quality for all types of precipitation events, heavy-tailed included, subject to a good overall performance.

Our hybrid proposed methods are applied to daily 51-h forecasts of 6-h accumulated precipitation from 2012 to 2015 over France using the Météo-France ensemble prediction system called PEARP. They provide calibrated predictive distributions and compete favourably with state-of-the-art methods like Analogs method or Ensemble Model Output Statistics. In particular, hybrid forest-based procedures appear to bring an added value to the forecast of heavy rainfall.

Contents

3.1 Introduction

3.1.1 Post-processing of ensemble forecasts

Accurately forecasting weather is paramount for a wide range of end-users, e.g. air traffic controllers, emergency managers and energy providers (see, e.g. Pinson et al., 2007; Zamo et al., 2014a). In meteorology, ensemble forecasts try to quantify forecast uncertainties due to observation errors and incomplete physical representation of the atmosphere. Despite its recent developments in national meteorological services, ensemble forecasts still suffer of bias and underdispersion (see, e.g. Hamill and Colucci, 1997). Consequently, they need to be post-processed. At least two types of statistical methods have emerged in the last decades : analogs method and ensemble model output statistics (EMOS) (see, e.g. Delle Monache et al., 2013; Gneiting et al., 2005, respectively). The first one is fully non-parametric and consists in finding similar atmospheric situations in the past and using them to improve the present forecast. In contrast, EMOS belongs to the family of parametric regression schemes. If y represents the weather variable of interest and $(x_1, \ldots, x_m)$ the corresponding m ensemble member forecasts, then the EMOS predictive distribution is simply a distribution whose parameters depend on the values of $(x_1, \ldots, x_m)$. Less conventional approaches have also been studied recently. For example, Van Schaeybroeck and Vannitsem (2015) investigated member-by-member post-processing techniques and Taillardat et al. (2016) found that quantile regression forests (QRF) techniques performed well for temperatures and wind speed data.

3.1.2 Forecasting and calibration of precipitation

Not all meteorological variables are equal in terms of forecast and calibration. In particular, Hemri et al. (2014) highlighted that rainfall forecasting represents a steep hill. In this study, we will focus on 6-h rainfall amounts in France because this is the unit of interest of the ensemble forecast system of Météo-France. For daily precipitation, extended logistic regression was frequently applied (see, e.g. Hamill et al., 2008; Roulin and Vannitsem, 2012; Ben Bouallègue, 2013). Bayesian Model Averaging techniques (Raftery et al., 2005; Sloughter et al., 2007) were also used in rainfall forecasting, but we will not cover them here because a gamma fit is often applied to cube root transformed precipitation accumulations and this complex transformation may not be adapted to 6h rainfall. Concerning analogs and EMOS techniques, they have been applied to calibrate daily rainfall (see Hamill and Whitaker, 2006; Scheuerer, 2014; Scheuerer and Hamill, 2015). As the QRF method in Taillardat et al. (2016) performed better than EMOS for temperatures and wind speeds, one may wonder if QRF could favourably compete with EMOS and analogs techniques for rainfall calibration. This question is particularly relevant because recent methodological advances have been made concerning random forests and quantile regressions. In particular, Athey et al. (2016) proposed an innovative way, called gradient forests (GF), of using forests to make quantile regression. In this context, we propose to implement and test this quantile regression GF method for rainfall calibration and compare it with other approaches, see Section 3.2.

3.1.3 Parametric probability density functions (pdf) of precipitation

Modeling precipitation distributions is a challenge by itself. It is a mixture of zeros (dry events) and positive intensities, i.e. rainfall amounts for wet events. The latter have a skewed distribution. One popular and flexible choice to model rainfall amounts is to use the gamma distribution or to built on it. For example, Scheuerer and Hamill (2015) and Baran and Nemoda (2016) in a rainfall calibration context employed the censored-shifted gamma (CSG) pdf defined by

$$f_{CSG}(y) \quad = \quad \begin{cases} (1-\pi) \cdot \frac{(y+\delta)^{\kappa-1}}{\Gamma(\kappa)} \exp(-(y+\delta)/\theta), & \text{if } y > 0 \\ \pi, & \text{if } y = 0, \end{cases} \tag{3.1}$$

where $y \geq 0$, the positive constants (κ, θ) are the two gamma law parameters and the probability $\pi \in [0, 1]$ represents the mass of the gamma cumulative distribution function (cdf) below the level of censoring $\delta \geq 0$. Hence, the probability of zero and positive precipitation are treated together. One possible drawback of the CSG is that heavy daily and subdaily rainfall may not always have a nice upper tail with an exponential decay like a gamma distribution, but rather a polynomial one, the latter point being a key element in any weather risk analysis (see, e.g. Katz et al., 2002; De Haan and Ferreira, 2007). To bring the necessary flexibility in modelling upper tail behavior in a rainfall EMOS context, Scheuerer (2014) worked with

a so-called censored generalized extreme value (CGEV) defined by

$$f_{CGEV}(y) = \begin{cases} (1-\pi)\cdot g(y;\mu,\sigma,\xi), & \text{if } y>0 \\ \pi, & \text{if } y=0, \end{cases} \tag{3.2}$$

where $\pi = G(0;\mu,\sigma,\xi)$ and the pdf $g(y;\mu,\sigma,\xi)$ which cumulative distribution function G is the classical GEV

$$G(y;\mu,\sigma,\xi) = \exp\left[-\left(1+\frac{\xi(y-\mu)}{\sigma}\right)_+^{-1/\xi}\right] \text{ for } \xi \neq 0.$$

Note that $a_+ = \max(0,a)$ and that, if $\xi = 0$, then $g(y;\mu,\sigma,0)$ represents the classical Gumbel pdf. To be in compliance with extreme value theory (EVT) not only for heavy rainfall but also for low precipitation amounts, Naveau et al. (2016) recently proposed a class of models referred as the extended generalized Pareto (EGP) that allows a smooth transition between generalized Pareto (GP) type tails and the middle part (bulk) of the distribution. It bypasses the complex thresholds selection step to define extremes. Low precipitation can be shown to be gamma distributed, while heavy rainfall are Pareto distributed. Mathematically, a cdf belonging to the EGP family has to be expressed as

$$T\left\{H_\xi(y/\sigma)\right\}, \text{ for all } y>0,$$

where $H_\xi(y) = 1-(1+\xi y)^{-1/\xi}$ represents the GP cdf, while T denotes a continuous cdf on the unit interval. To insure that the upper tail behavior of T is driven by the shape parameter ξ, the survival function $\bar{T} = 1-T$ has to satisfy that $\lim_{u\downarrow 0}\frac{\bar{T}(1-u)}{u}$ is finite. To force low rainfall to follow a GPD for small values near zero, we need that $\lim_{u\downarrow 0}\frac{T(u)}{u^s}$ is finite for some real $s>0$. Studies have already made this choice (see, e.g. Vrac and Naveau, 2007; Naveau et al., 2016). In Naveau et al. (2016), different parametric models of the cdf T satisfying the required constraints were compared. The special case where $T(u) = u^\kappa$ with $\kappa > 0$ obeys these constraints and also corresponds to a model studied by Papastathopoulos and Tawn (2013). In practice, this simple version of T appears to fit well daily and subdaily rainfall and consequently, we will only focus on this case in this paper. In other words, our third model for the precipitation pdf is

$$f_{EGP}(y) = \begin{cases} (1-\pi)\cdot\frac{\kappa}{\sigma}\cdot\left\{H_\xi(x/\sigma)\right\}^{\kappa-1}\cdot h_\xi(y/\sigma), & \text{if } y>0 \\ \pi, & \text{if } y=0, \end{cases} \tag{3.3}$$

where $h_\xi(.)$ is the pdf associated with $H_\xi(.)$. In contrast to (3.1) and (3.2), the probability weight π is not obtained by censoring, and it is just a parameter independent of $(\kappa,\sigma,\xi)^T$.

At this stage, we have three parametric pdfs, see (3.1) and (3.2) and (3.3), to implement a EMOS approach to 6-hour rainfall data, see Section 3.3. Besides comparing these three EMOS models, it is natural to wonder if QRF and GF methods could take advantage of these three parametric forms.

3.1.4 Coupling parametric pdfs with random forest approaches

A drawback of data driven approaches like QRF and GF is that their intrinsic non parametric nature make them useless to predict beyond the largest recorded rainfall. To circumvent this limit, we also propose to combine random forest techniques with a EGP pdf defined by (3.3), see Section 3.2.3. Hence, random forest-based post-processing techniques will be in compliance with EVT and this should be an interesting path to improve prediction behind the largest values of the sample at hand.

3.1.5 Outline

This article is organized as follows. In Section 3.2, we recall the basic ingredients to create quantile regression forests and gradient forests. In particular, we review the calibration process of the GF method recently introduced by Athey et al. (2016) for quantile regression. Then, we explain how these trees are combined with the EGP pdf defined by (3.3).

In Section 3.3, we propose to integrate the EGP pdf within a EMOS scheme.

The different approaches are implemented in Section 3.4 where the test bed dataset of 87 French weather stations and the French ensemble forecast system of Météo-France called PEARP (Descamps et al., 2015) is described. Then, we assess and compare each method with a special interest for heavy rainfall, see Section 3.5. The paper closes with a discussion in Section 3.6.

3.2 Quantile regression forests and gradient forests

3.2.1 Quantile regression forests

Given a sample of predictors-response pairs, say (X_i, Y_i) for $i = 1, \ldots, n$, classical regression techniques connect the conditional mean of a response variable Y to a given set of predictors X. The quantile regression forest (QRF) method introduced by Meinshausen (2006) also consists in building a link, but between an empirical cdf and the outputs of a tree. Before explaining this particular cdf, we need to recall how trees are constructed.

A random forest is an aggregation of randomized trees based on bootstrap aggregation on the one hand, and on classification and regression trees (CART) (Breiman, 1996; Breiman et al., 1984) on the other hand. These trees are built on a bootstrap copy of the samples by recursively maximizing a splitting rule. Let $\mathcal{D}_0$ denote the group of observations to be divided into two subgroups, say $\mathcal{D}_1$ and $\mathcal{D}_2$. For each group, we can infer its homogeneity defined by

$$v(\mathcal{D}_j) = \sum_{Y \in \mathcal{D}_j} [Y - \overline{Y}(\mathcal{D}_j)]^2,$$

where $\overline{Y}(\mathcal{D}_j)$ corresponds to the sample mean in $\mathcal{D}_j$. To determine if this splitting choice is optimal, the homogeneities $v(\mathcal{D}_1)$ and $v(\mathcal{D}_2)$ are compared to the one of $\mathcal{D}_0$. For example,

if wind speed is one predictor in X and dividing low and large winds could better explain rainfall, then the cutting value, say s, will be the one that maximizes

$$\mathcal{H}(\mathcal{D}_1, \mathcal{D}_2) = \max_{s \in \mathcal{E}^*} [v(\mathcal{D}_0) - v(\mathcal{D}_1) - v(\mathcal{D}_2)] \tag{3.4}$$

where $\mathcal{E}^*$ is a random subset of the predictors in the predictors' space $\mathcal{E}$. Each resulting group is itself split into two, and so on until some stopping criterion is reached. As each tree is built on a random subset of the predictors, the method is called "random forest" (Breiman, 2001). Binary regression trees can be viewed as decision trees, each node being the criterion used to split the data and each final leaf giving the predicted value. For example, if we observe a given wind speed x, we can find the final leaf that corresponds to this value of x and the associated observations y, then we can compute the conditional cumulative distribution function introduced by Meinshausen (2006)

$$\widehat{F}(y|x) = \sum_{i=1}^{n} \omega_i(x) \mathbf{1}(\{Y_i \leq y\}), \tag{3.5}$$

where the weights $\omega_i(x)$ are deduced from the presence of Y_i in a final leaf of each tree when one follows the path determined by x. The interested reader is referred to Taillardat et al. (2016) for an application of this approach to ensemble forecast of temperatures and winds.

3.2.2 Gradient forests

Meinshausen (2006) proposed splitting rule using CART regression splits. Arguing that this splitting rule is not tailored to the quantile regression context, Athey et al. (2016) proposed another optimisation scheme. Instead of maximizing the variance heterogeneity of the children nodes, one maximizes the criterion

$$\Delta(\mathcal{D}_1, \mathcal{D}_2) = \sum_{j=1}^{2} \frac{-1}{|\{i : Y_i \in \mathcal{D}_j\}|} \left(\sum_{\{i: Y_i \in \mathcal{D}_j\}} \rho_i \right)^2 \tag{3.6}$$

where the indicator function $\rho_i = \mathbf{1}(\{Y_i \geq \hat{\theta}_{q,\mathcal{D}_0}\})$ is equal to one when Y_i is greater than the q-th quantile $\hat{\theta}_{q,\mathcal{D}_0}$ of the observations of the parent node $\mathcal{D}_0$. The terminology of *gradient forests* was suggested because the choice of ρ_i is here linked with a gradient-based approximation of the quantile function

$$\Psi_{\hat{\theta}_{q,\mathcal{D}_0}}(Y_i) = q\mathbf{1}(\{Y_i > q\}) + (1-q)\mathbf{1}(\{Y_i \leq q\}).$$

This technique using gradients is computationally feasible, an issue not to be omitted when dealing with non-parametric techniques. Note here that for each split the order of the quantile is chosen among given orders $(0.1, 0.5, 0.9)$. In the special case of least-square regression, ρ_i becomes $Y_i - \overline{Y}(\mathcal{D}_0)$, and $\mathcal{H}(\mathcal{D}_1, \mathcal{D}_2)$ becomes equivalent to $\Delta(\mathcal{D}_1, \mathcal{D}_2)$. In this special case, gradient trees are equivalent to build a standard CART regression tree.

3.2.3 Fitting a parametric form to QRF and GF trees

As mentioned in Section 3.1.4, the predicted cdf defined by (3.5) cannot predict values which are not in the learning sample. This can be a strong limitation if the learning sample is small or rare events are of interest or both. The GF method has the same issue. To parametrically model rainfall, the EGP pdf defined by (3.3) appears to be a good candidate. It allows more flexibility in the fitting than CSG or CGEV. This distribution has four parameters, π, κ, σ and ξ, it is in compliance with EVT for low and heavy rainfalls and works well in practice (see, e.g. Naveau et al., 2016). In terms of inference, a simple and fast method-of-moment can be applied. Basically, probability weighted moments (PWM) of a given random variable, say Y, with survival function $\overline{F}(y) = \mathbb{P}(Y > y)$, can be expressed as (see, e.g. Hosking and Wallis, 1987)

$$\mu_r = \mathbb{E}([Y\overline{F}^r(Y)]) = \int_0^1 F^{-1}(q)(1-q)^r dq. \tag{3.7}$$

If Y follows a EGP pdf defined by (3.3), then we have

$$\begin{aligned} \frac{\xi}{\sigma}\mu_0 &= \kappa B(\kappa, 1-\xi) - 1 \text{ and } \frac{\xi}{\sigma}\mu_1 = \kappa\left(B(\kappa, 1-\xi) - B(2\kappa, 1-\xi)\right) - \frac{1}{2}, \\ \frac{\xi}{\sigma}\mu_2 &= \kappa\left(B(\kappa, 1-\xi) - 2B(2\kappa, 1-\xi) + B(3\kappa, 1-\xi)\right) - \frac{1}{3}, \end{aligned}$$

where $B(.,.)$ represents the beta function. Knowing the PWM triplet $(\mu_0, \mu_1, \mu_2)^T$ is equivalent to know the parameter vector $(\kappa, \sigma, \xi)^T$. Hence, we just need to estimate these three PWMs. For any given forest, it is possible to estimate the distribution of $[Y|X = x]$ by the empirical cdf $\widehat{F}(y|X = x)$ defined by (3.5). Then, we can plug it in (3.7) to get

$$\hat{\mu}_r(x) = \int_0^1 \widehat{F}^{-1}(q|X = x)(1-q)^r dq.$$

This leads to the estimates of $(\kappa(x), \sigma(x), \xi(x))^T$ and consequently of $f(y|X = x)$ via Equation (3.3). Note that the probability of no rain $\pi(x)$ is just inferred by counting the number of dry events in the corresponding trees. In the following, this technique is called "EGP TAIL", despite the fact that the whole distribution is fitted from QRF and GF trees.

3.3 Ensemble model output statistics and EGP

In Section 3.1.3, three definitions of parametric pdfs were recalled. By regressing their parameters on the ensemble values, different EMOS models have been proposed for the CSG and CGEV pdfs defined by (3.1) and by (3.2), respectively. More precisely, Baran and Nemoda (2016) used the CSG pdf by letting the mean $\mu = \kappa\theta$ and variance $\sigma^2 = \kappa\theta^2$ depend linearly as functions of the raw ensemble values and their mean, respectively. The coefficients of this regression were estimated by miminizing the continuous ranked probability

score (CRPS) (see, e.g. Scheuerer and Hamill, 2015; Hersbach, 2000). The same strategy can be applied to fit the CGEV pdf (see, e.g. Hemri et al., 2014). Scheuerer (2014) modelled the scale parameter σ in (3.2) as an affine function of the ensemble mean absolute deviation rather than of the raw ensemble mean or variance. Another point to emphasize is that the shape parameter ξ was considered invariant in space in Hemri et al. (2014).

In this section, we basically explain how an EMOS approach can be built with the EGP pdf defined by (3.3) and we now highlight common features and differences between the two EMOS with CSG and CGEV. The scale parameter σ^2 in (3.3) is estimated in the same way than for CGEV. The presence of the parameter κ allows an additional degree of freedom. The expectation of our EGP is mainly driven by the product $\kappa\sigma$. Consequently, we model κ as an affine function of the predictors divided by σ. As France has a diverse climate, it is not reasonable to assume a constant shape parameter among all locations, see the map in Figure 3.1. In addition, minimizing the CRPS to infer different shape parameters may be inefficient (see, e.g. Friederichs and Thorarinsdottir, 2012). To estimate ξ at each location, we simply use the PWM inference scheme described in Section 3.2.3. To complete the estimation of the parameters in (3.3), the probability π is modeled as an affine function on $[0, 1]$ of the raw ensemble probability of rain and affine function parameters are also estimated by CRPS minimization. The table 3.1 sums up the optimal estimation strategies that we have found for each distribution.

TABLE 3.1 – Optimal strategies for parameter estimation using CRPS minimization in the EMOS context.

Distribution	Parameter	Comments
CSG	δ	free in $\mathbb{R}$
	μ	affine function of covariates in C
	σ	affine function of raw ensemble mean
	κ	$\kappa = \mu^2/\sigma$
	θ	$\theta = \sigma/\mu$
CGEV	μ	affine function of covariates in C
	σ	affine function of the mean absolute deviation of the raw ensemble
	ξ	free in $(-\infty, 1)$
	θ	$\theta = \sigma/\mu$
EGP	σ	affine function of the mean absolute deviation of the raw ensemble
	μ	maximum between 0 and an affine function of covariates in C
	κ	$\kappa = \mu/\sigma$
	ξ	fixed, see Figure 3.1 for stations' values
	π	affine function of PR0 in C, bounded on $[0, 1]$

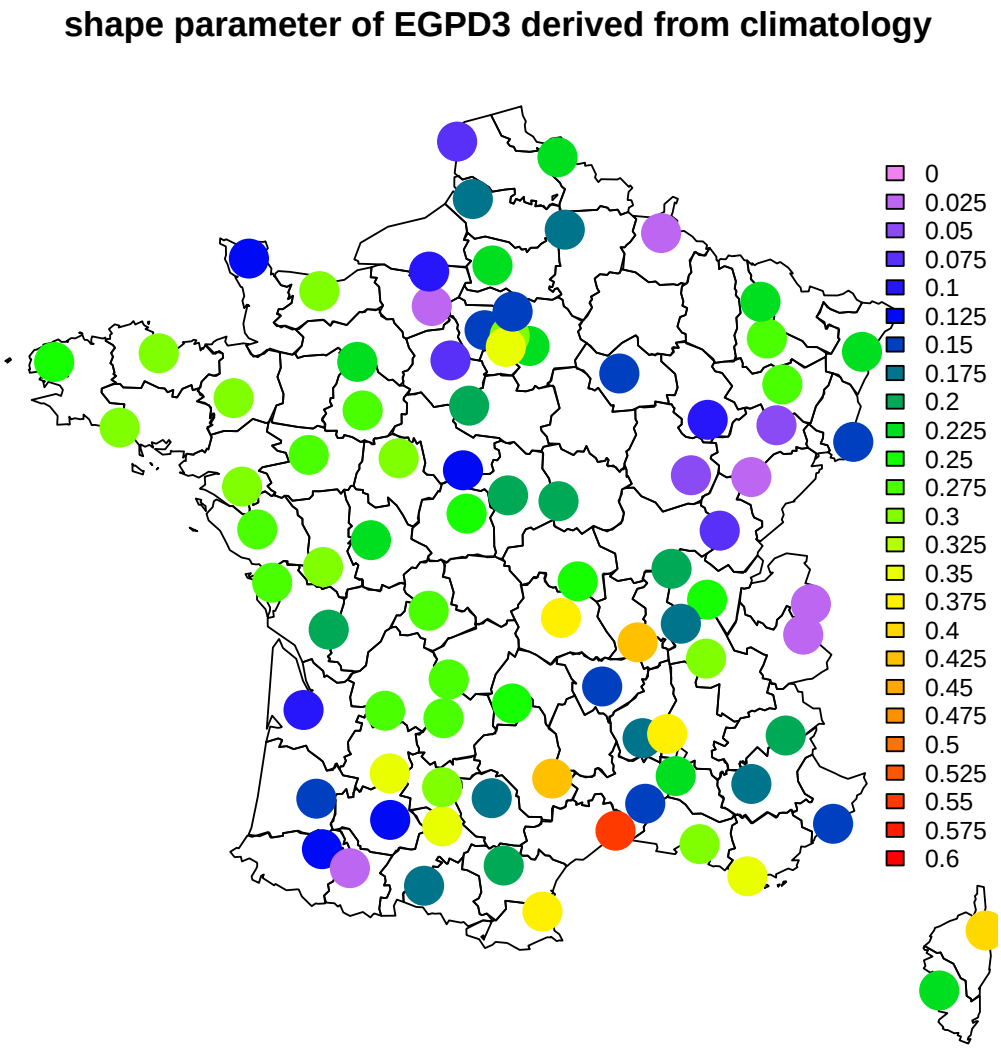

FIGURE 3.1 – Spatial values of ξ among locations.

3.4 Case study on the PEARP ensemble prediction system

3.4.1 Data description

Our rainfall dataset corresponds to 6-h rainfall amounts produced by 87 French weather stations and the 35-member ensemble forecast system called PEARP (Descamps et al., 2015) at a 51-h lead time forecast. Our period of interest spans four years from 1 January 2012 to 31 December 2015.

3.4.2 Inferential details for EMOS and analogs

Verification has been made on this entire period. For a fair comparison each method has to be tuned optimally. EMOS uses all the data available for each day (4 years less the forecast day as a training period). The same strategy is used to fit the analogs method, see Appendix 3.7.1 for details on this approach. QRF and GF employ a cross-validation method : each month of the 4 years is kept as validation data while the rest of the 4 years is used for learning. The tuning algorithm for EMOS is stopped after few iterations in order to avoid overfitting, as suggested in Scheuerer (2014) concerning the parameter estimations.

3.4.3 Sets of predictors used

We either use a subset of classical predictors (denoted by "C" in the rest of the paper) detailed in Table 3.2 or the whole set of available predictors as listed in Table 3.3.

TABLE 3.2 – Subset "C" representing the most classical predictors.

Name	Description
HRES	high resolution member
CTRL	control member
MEAN	mean of raw ensemble
PR0	raw probability of rain

Note that we also considered for EMOS a third type of predictors set based on a variable selection algorithm (see Appendix 3.7.3). But this did not improve the results and we removed them from the analysis (available upon request).

3.4.4 Zooming on extremes

Finding a way to assess the quality of ensembles for extreme and rare events is quite difficult, as seen in Williams et al. (2014) in a comparison of ensemble calibration methods

TABLE 3.3 – Set of all available predictors.

Name	Description
HRES	high resolution member
CTRL	control member
MEAN	mean of raw ensemble
MED	median of raw ensemble
Q10	first decile of raw ensemble
Q90	ninth decile of raw ensemble
PR0	raw probability of rain
PR1	raw probability of rain $>$ 1mm/6h
PR3	raw probability of rain $>$ 3mm/6h
PR5	raw probability of rain $>$ 5mm/6h
PR10	raw probability of rain $>$ 10mm/6h
PR20	raw probability of rain $>$ 20mm/6h
SIGMA	standard deviation of raw ensemble
IQR	IQR of raw ensemble
HU1500	deterministic forecast of 6-h mean 1500m humidity
UX	deterministic forecast of 6-h maximum of zonal wind gust
VX	deterministic forecast of 6-h maximum of meridional wind gust
FX	deterministic forecast of 6-h maximum of wind gust power
TCC	deterministic forecast of 6-h mean total cloud cover
RR6CV	deterministic forecast of 6-h convective rainfall amount
CAPE	deterministic forecast of 6-h mean convective available potential energy
q10,50,90 are the first decile, the median and ninth decile of the raw ensemble for these variables :	
HU_q10,50,90	6-h mean surface humidity
P_q10,50,90	6-h mean sea level pressure
TCC_q10,50,90	6-h mean total cloud cover
RR6CV_q10,50,90	6-h convective rainfall amount
U10_q10,50,90	6-h mean surface zonal wind
V10_q10,50,90	6-h mean surface meridional wind
U500_q10,50,90	6-h mean 500m zonal wind
V500_q10,50,90	6-h mean 500m meridional wind
FF500_q10,50,90	6-h mean 500m wind speed
TPW850_q10,50,90	6-h mean 850hPa potential wet-bulb temperature
FLIR6_q10,50,90	6-h mean surface irradiation in infra-red wavelengths
FLVIS6_q10,50,90	6-h mean surface irradiation in visible wavelengths
T_q10,50,90	6-h mean surface temperature
FF10_q10,50,90	6-h mean surface wind speed

for extreme events. Weighted scoring rules can be adopted as done in Gneiting and Ranjan (2011); Lerch et al. (2017) but there are here two main issues. The ranking of compared methods depends on the weight function used, as already suggested in Gneiting and Ranjan (2011). Besides, giving a weight to such rare events avoid discriminant power of scoring rules, the same issue than for the Brier score (Brier, 1950). Moreover, reliability is not sound here since there are not enough extreme cases (by definition) to measure it. We have finally decided to focus on two ideas here, matching with forecasters' desires : first, what is the discriminant power of our forecasts for extreme events in terms of binary decisions ? Second, what is the potential risk of our ensemble to mismatch an extreme event ? The choice done in our study is discussed in Section 3.5.

3.5 Results

Table 3.4 compares different metrics for all post-processing techniques which have been fitted to the 87 stations and averaged over 4 years of verification. Ten methods are competing : The raw ensemble, 4 analogs, 3 EMOS (3 different distributions using the set C), 2 forest-based methods (1 QRF and 1 GF) and 2 tail-extended forest-based methods (1 QRF and 1 GF). Scores used concern respectively (i) global performance (calibration and sharpness) measured by the CRPS ; (ii) reliability performance, measured by the mean, the normalized variance and the entropy of the PIT histograms, denoted by Ω in the sequel ; (iii) gain in CRPS compared to the raw ensemble, measured by the Skill of the CRPS using the raw ensemble as baseline. A brief summary about these measures is done in 3.7.4, where references are also provided. And the boxplots showing rank histograms are in 3.7.5.
According to Table 3.4, the raw ensemble is biased and underdispersive. The EMOS post-processed ensembles share with QRF and GF a good CRPS. Moreover, we can consider them as unbiased and mostly well-dispersed. The tail-extended methods get a lower CRPS, that can be explained by their skill for extreme events. Finally, the four analog methods show a quite poor CRPS compared to the raw ensemble, even if they exhibit reliability. Nevertheless we can notice that a weightning of the predictors, especially with a non-linear variable selection algorithm (Analogs_VSF), brings benefits to this method. This phenomenon can be explained by Figure 3.2, where the ROC curves are given for the event of rain. Consider a fixed threshold s and the contingency table associated to the predictor $\mathbf{1}\{rr6 > s\}$. Recall that the ROC curve then plots the probability of detection (or hit rate) as a function of the probability of false detection (or false alarm rate). A "good" prediction must maximize hit rates and minimize false alarms (see, e.g. Jolliffe and Stephenson, 2012). Figure 3.2 explicitely shows the lack of resolution of the analogs technique. Incidently, we can also notice that the rain event discrimination is not improved by post-processed ensembles. For information, more ROC curves are provided in Appendix 3.7.6.

To sum up, the best improvement with respect to the raw ensemble is for the forest-based methods, according to the CRPSS (which definition is in Appendix 3.7.4). This improvement is however less significant than for other weather variables (see Taillardat et al. (2016)). This corroborates Hemri et al. (2014)'s conclusion that rainfall amounts are tricky to calibrate. If the analogs method looks less performant, that might be imputable to the data depth of only 4 years. Indeed, this non-parametric technique is data-driven (such as QRF and GF) and needs more data to be effective (see e.g. Van den Dool (1994)).

Concerning extreme events, Figure 3.3 shows the benefit of the tail extension for forest-based methods. Note that we prefer to pay attention to the *value* of a forecast more than to its *quality*. According to Murphy (1993), the *value* can be defined as the ability of the forecast to help users to take better decisions. The *quality* of a forecast can be summarized by the area on the *modelled* ROC curve (classically denoted by AUC), with some potential drawbacks exhibited by Lobo et al. (2008); Hand (2009). Zhu et al. (2002) made a link between optimal decision thresholds, value and cost/loss ratios. In particular, they show that the value of a forecast is maximized for the "climatological" threshold and equals the hit rate minus the false alarm rate which is the maximum of the Peirce Skill Score (Manzato,

TABLE 3.4 – Comparing performance statistics for different post-processing methods for 6-h rainfall forecasts in France. The mean CRPS estimations come from bootstrap replicates, the estimation error is under 6.1×10^{-3} for all methods.

Types	Methods	pdf	CRPS	$\mathbb{E}(Z)$	$\mathbb{V}(Z)$	Ω	CRPSS
	Raw ensemble		0.4694	0.4164	1.0612	0.9809	0%
Non-parametric	Analogs		0.5277	0.5175	1.0190	0.9956	-12.4%
	Analogs_C		0.5376	0.5050	1.0051	0.9964	-14.5%
	Analogs_COR		0.5276	0.5062	1.0015	0.9964	-12.4%
	Analogs_VSF		0.5247	0.5060	0.9986	0.9961	-11.8%
	QRF		0.4212	0.5006	0.9995	0.9961	10.3%
	GF		0.4134	0.5070	0.9771	0.9957	11.9%
Parametric	EMOS	CSG	0.4224	0.4992	1.0363	0.9955	10.0%
with	EMOS	GEV	0.4228	0.5000	1.0073	0.9961	9.9%
covariates $\in$ C	EMOS	EGP	0.4292	0.4623	1.0723	0.9905	8.6%
Hybrid	QRF	EGP TAIL	0.4138	0.5095	0.9558	0.9957	11.8%
	GF	EGP TAIL	0.4127	0.5152	0.9425	0.9948	12.1%

2007). This value corresponds to the upper left corner of ROC curves, which is of main interest in terms of extremes verification, as explained in Section 3.4.4. Several features already seen on Figure 3.2 can be observed on Figure 3.3 : analogs lack resolution and the other post-processed methods compete more or less favourably with the raw ensemble. Nonetheless, the other post-processing techniques stay better than the raw ensemble even for methods that cannot extrapolate observed values such as QRF and GF. Note that QRF is rather surprisingly better than EMOS techniques. Tail extension methods show their gain in a binary decision context. For inofrmation, binormal-modelled ROC curve for the threshold 15mm is available in Appendix 3.7.7. Contrary to the Figure 3.3 we show here that the resolution of calibrated forecasts are very close. Thus, we conclude that the improvement lies in the increase of forecast value.

3.6 Discussion

Throughout this study, we see that forest-based techniques compete favourably with EMOS techniques. It is a good point to see that QRF and GF compared to EMOS exhibit nearly the same kind of improvement when focusing on rainfall amounts or on temperature and wind speed (see Taillardat et al. (2016) Figures 6 and 13). It could be interesting to check these methods (especially GF) on smoother variables.

Tail extension of these non-parametric techniques generates ensembles more tailored for extremes catchment. However, reliability as well as resolution remain quite stable when

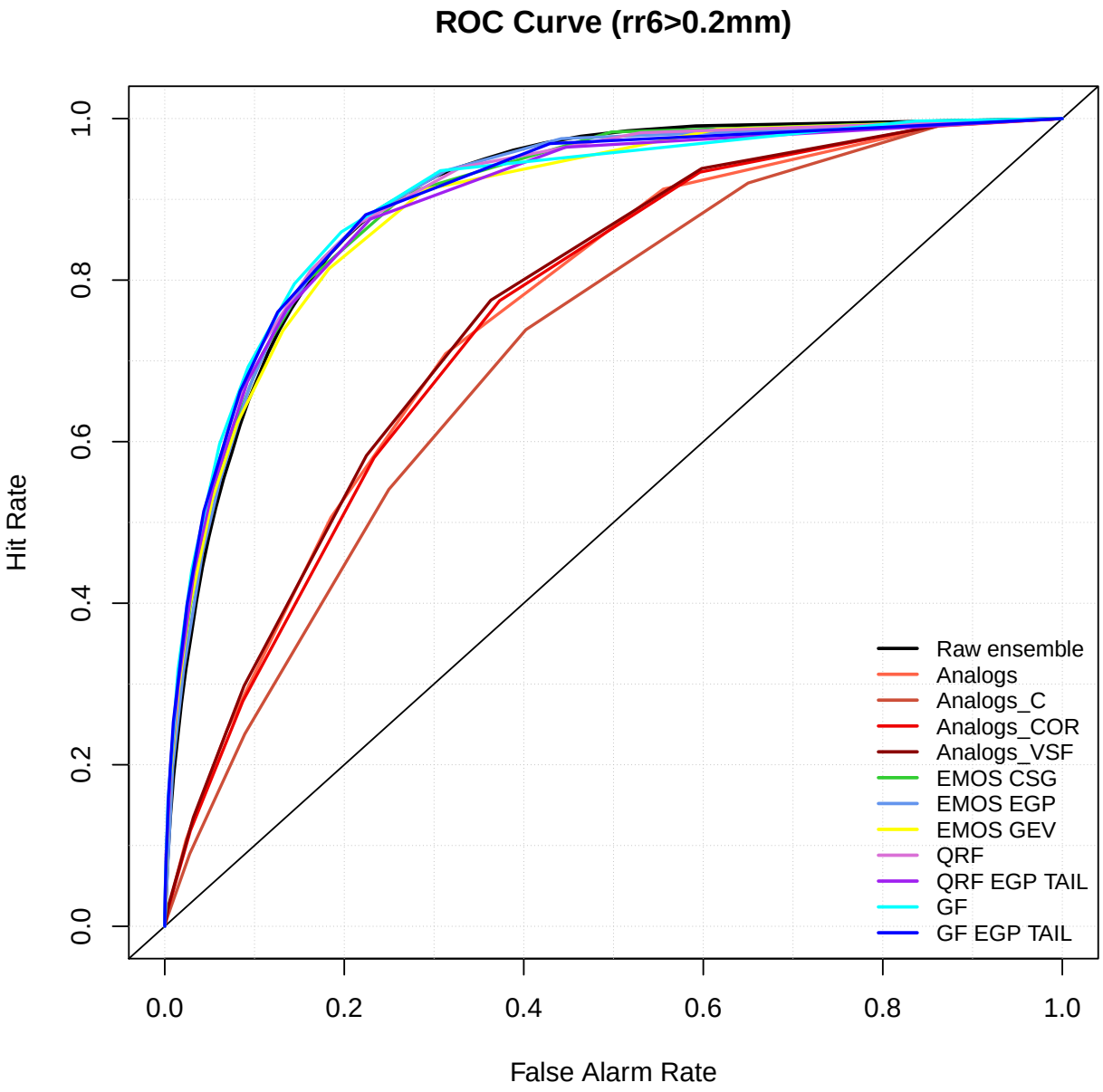

FIGURE 3.2 – ROC Curves for the event of rain. A "good" prediction must maximize hit rate and minimize false alarms. The analogs method lacks resolution. We can notice that there is no improvement of post-processed methods compared to the raw ensemble.

extending the tail, so that our paradigm about verification (good extreme discrimination subject to satisfying overall performance) remains.

One of the advantages of distribution-free calibration (analogs, QRF and GF) is that there is no assumption on the parameters to calibrate. This benefit is emphasized for rainfall

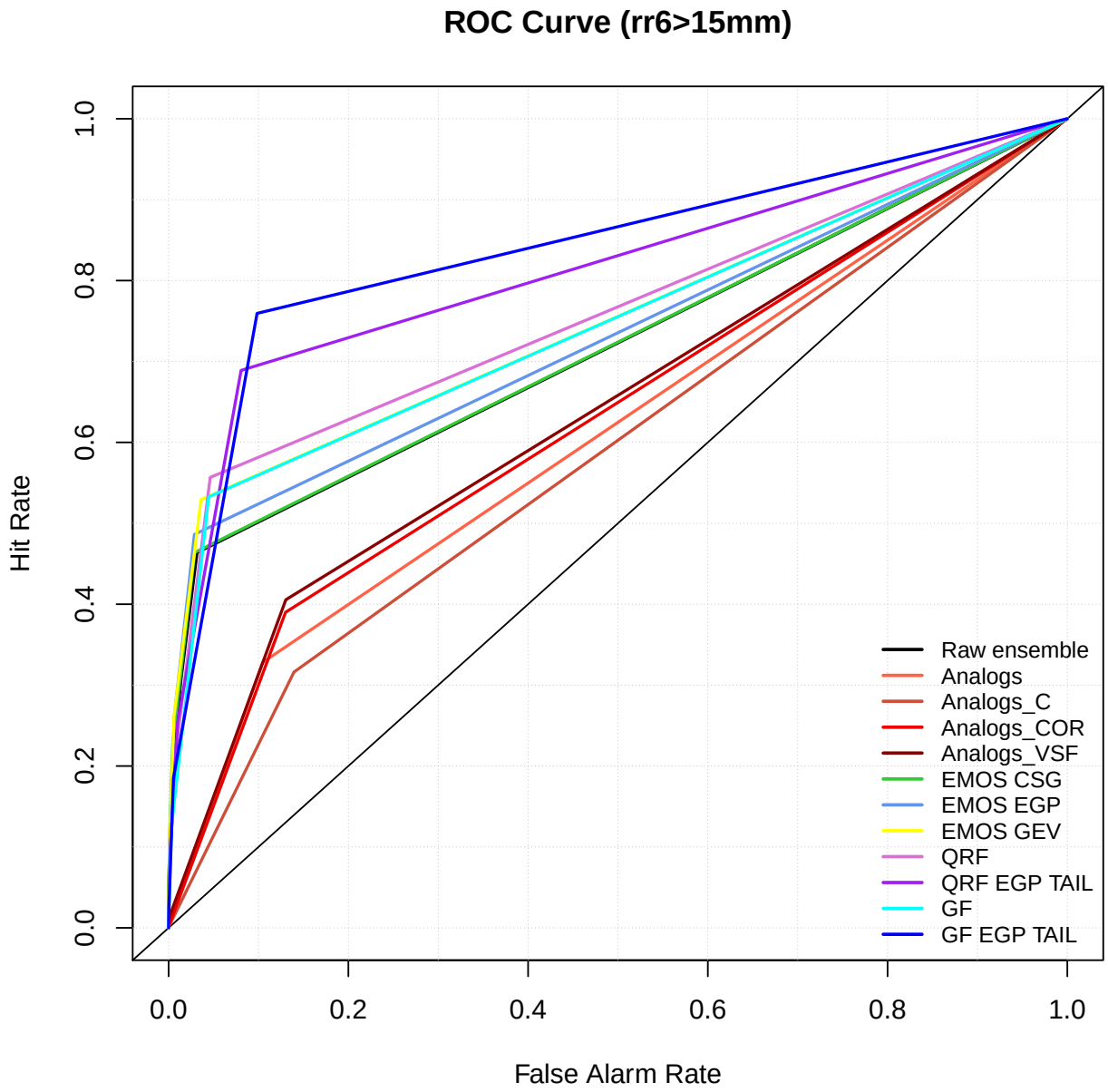

FIGURE 3.3 – ROC Curves for the event of rain above 15mm. A "good" prediction must maximize hit rate and minimize false alarms. The analogs method lacks resolution. Tail extension methods show their gain in a binary decision context.

amounts for which EMOS techniques have to be studied using different distributions. In this sense, the recent mixing method of Baran and Lerch (2016) looks appealing. A brand new alternative solution consists in working with (standardized) anomalies as done in Dabernig et al. (2016).

Another positive aspect of the forest-based methods is that there is no need of a predictor selection. Concerning the analogs method, our results suggest that the work of Genuer et al. (2010) could be a cheaper alternative to brute force algorithms like in Keller et al. (2017) for the weightning of predictors. For analogs techniques, we can notice that the complete set of predictors gives the best results. In contrast, the choice of the set of predictors is still an ongoing issue for EMOS techniques regarding precipitation. For easier variables to calibrate, Messner et al. (2017) shows that some variable selection can be effective.

The tail extension can be viewed as a semi-parametric technique where the result of forest-based methods is used to fit a distribution. This kind of procedure can be connected to the work of Junk et al. (2015) who uses analogs on EMOS inputs. An interesting prospect would be to bring forest-based methods in this context.

A natural perspective regarding spatial calibration and trajectory recovery could be to make use of block regression techniques as done in Zamo et al. (2016), or of ensemble copula coupling, as suggested by (Bremnes, 2007; Schefzik, 2016).

Finally, it appears that more and more weather services work on merging different forecasts from different sources (multi-model ensembles). In this context, an attractive procedure could be to combine raw ensembles and different methods of post-processing via sequential aggregation (Mallet, 2010; Thorey et al., 2016), in order to get the best forecast according to the weather situations.

Acknowledgements

Part of the work of P. Naveau has been supported by the ANR-DADA, LEFE-INSU-Multirisk, AMERISKA, A2C2, CHAVANA and Extremoscope projects. This work has been supported by Energy oriented Centre of Excellence (EoCoE), grant agreement number 676629, funded within the Horizon2020 framework of the European Union. This work has been supported by the LABEX MILYON (ANR-10-LABX-0070) of Université de Lyon, within the program "Investissements d'Avenir" (ANR-11-IDEX-0007) operated by the French National Research Agency (ANR). Thanks to Julie Tibshirani, Susan Athey and Stefan Wager for providing gradient-forest source package.

3.7 Appendix

3.7.1 Analogs method

Contrary to EMOS, this technique is data-driven. An analog for a given location and forecast lead time is defined as a past prediction, from the same model, that has similar values for selected features of the current model forecast. The method of analogs consists in finding these closest past forecasts according to a given metric of the predictors' space to build an analog-based ensemble (see e.g. Hamill and Whitaker (2006)). We assume here that close forecasts leads to close observations. Making use of analogs requires to choose both the set of predictors and the metric. Concerning the metric, several have been tried like

the Euclidean or the Mahalanobis distance but they have been outperformed by the metric provided in Delle Monache et al. (2013) :

$$\sum_{j=1}^{N_v} \frac{w_j}{\sigma_{f_j}} \sqrt{\sum_{i=-\tilde{t}}^{\tilde{t}} (F_{j,t+i} - A_{j,t'+i})^2}, \tag{3.8}$$

where F_t represents the current forecast at time t for a given location. The analog for another time t' at this same location is $A_{t'}$. The number of predictors is N_v and $\tilde{t}$ is half the time window used to search analogs. We standardize the distance by the standard deviation of each predictor σ_{f_j} calculated on the learning sample for the considered location. In this study we take $\tilde{t} = 1$ so the time window is ± 24 hours the forecast to calibrate. This distance has the advantages of being flow-dependent and thus defines a real weather regime associated with the research of the analogs. Note that one could weight the different predictors f_j with w_j and we fixed $w_j = 1$ for all predictors in a first method (Analogs). We have also tried two other weightning techniques using the absolute value of correlation coefficient between predictors and the response variable (Analogs_COR) like in Zhou and Zhai (2016), and a weightning based on the frequency of predictors' occurrences in variable selection algorithm described in Appendix 3.7.3 (Analogs_VSF). Note finally that other weightning techniques have been considered (Horton et al., 2017; Keller et al., 2017) but we did not use them in this study because of their computational cost.

3.7.2 CRPS formula for EGP

The CRPS for the distribution F detailed in 3.3 is :

$$\begin{aligned} CRPS(F,y) &= y(2F(y)-1) + \frac{\sigma}{\xi}(4\pi - 2F(y) - \pi^2 - 1) \\ &\quad + \frac{2\kappa\sigma(1-\pi)}{\xi}\left[B\left(\left[1+\frac{\xi y}{\sigma}\right]^{-\frac{1}{\xi}}; 1-\xi, \kappa\right) - (1-\pi)B(1-\xi, 2\kappa) - \pi B(1-\xi,\kappa)\right], \end{aligned}$$

where $0 < \xi < 1$ and $B(\,;\,,\,)$ and $B(\,,\,)$ denote respectively the incomplete beta and the beta functions.

3.7.3 Variable selection using random forests

We have seen that most parameters in EMOS and the distance used in analogs can be inferred using different sets of predictors. Contrary to the QRF and GF methods where the add of a useless predictor does not impact the predictive performance (since this predictor is never retained in the splitting rule), it can be misguiding for EMOS and analogs. We have therefore investigated some methods that keep the most informative meteorological variables and guarantee the best predictive performance. Our first choice was to use the well-known Akaike information criterion and the Bayesian information criterion (Akaike, 1998; Schwarz et al., 1978) but it resulted that the selection was not enough discriminant

(too many predictors kept in our initial set). The algorithm of Genuer et al. (2010) has then been considered. Such an algorithm is appealing since it uses random forests (and we already have these objects from the QRF method) and it permits to keep predictors without redundancy of information. For example this algorithm eliminates correlated predictors even if they are informative. A reduced set of predictors (mostly 3 or 4) is thus obtained, which avoids misestimation generated by multicolinearity. The method of variable selection used here is one among plenty others. The interested reader in variable selection using random forests can refer to Genuer et al. (2010) for detailed explanations.

The variable selection algorithm is used to keep the first predictors (max 4 of them) that form the set of predictors for each location. Figure 3.4 shows the ranked frequency of each chosen predictor. Predictors never retained are not on this figure. We can see here that only one third of the predictors in A are retained at least in 10% of the cases. Moreover, predictors representing central and extreme tendencies are preferred. Some predictors appear that differ from rainfall amounts ; see CAPE, FX or HU. It is not surprising since these parameters are correlated with storms. It is not shown here but when the MEAN variable is not selected, either MED or CTRL stands in the set. This shows that the algorithm mostly selects just one information concerning central tendency and avoid potential correlations. So the results concerning the variable algorithm selection seem to be sound. Last but not least, one notices that the predictors of the set C are often chosen. This remark confirms both the robustness of the algorithm and the relevance of previous studies on precipitation concerning the choice of the predictors.

3.7.4 Verification of ensembles

We recall here some facts about the scores used in this study.

Reliability

Reliability between observations and a predictive distribution can be checked by calculating $Z' = F(Y)$ where Y is the observation and F the cdf of the associated predictive distribution. Subject to calibration, the random variable Z' has a standard uniform distribution (Gneiting and Katzfuss, 2014) and we can check ensemble bias by comparing $\mathbb{E}(Z')$ to $\frac{1}{2}$ and ensemble dispersion by comparing the variance $\mathrm{Var}(Z')$ to $\frac{1}{12}$. This approach is applied to a $(K+1)$ ranked ensemble forecast using the discrete random variable $Z = \frac{\mathrm{rank}(y)-1}{K}$. Subject to calibration, Z has a discrete standard uniform distribution with $\mathbb{E}(Z) = \frac{1}{2}$ and a normalized variance $\mathbb{V}(Z) = 12\frac{K}{K+2}\mathrm{Var}(Z) = 1$.

Another tool used to assess calibration is the entropy :

$$\Omega = \frac{-1}{\log(K+1)} \sum_{i=1}^{K+1} f_i \log(f_i).$$

For a calibrated system the entropy is maximum and equals 1. Tribus (1969) showed that the entropy is an indicator of reliability linked to the Bayesian psi-test. It is also a proper

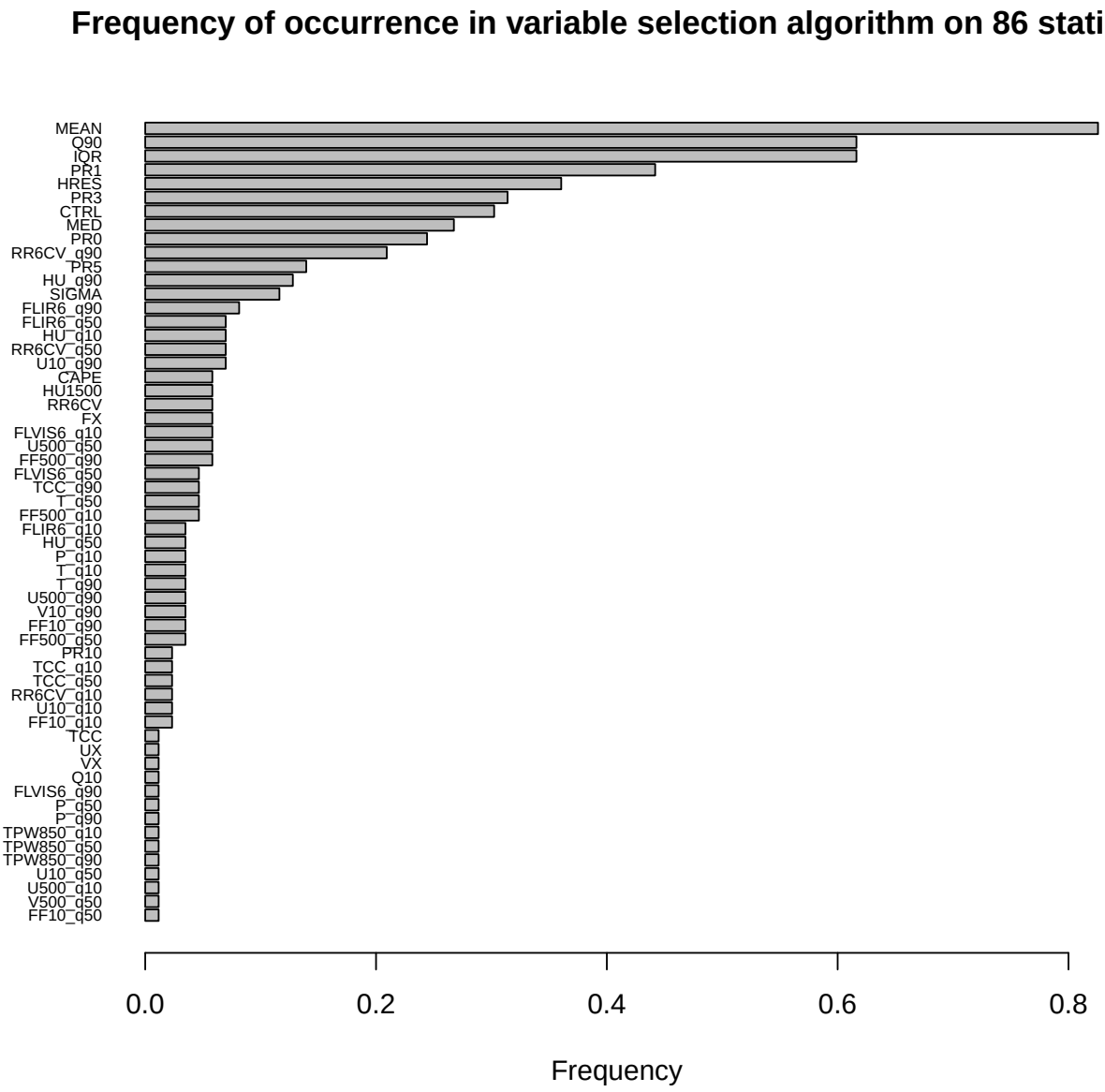

FIGURE 3.4 – Frequency of predictors' occurrence in variable selection algorithm. Variables representing central and extreme tendencies are preferred. Some covariables like CAPE, FX or HU can be retained. It is interesting to see that only one third of the predictors of the set is taken more than in 10% of the cases.

measure of reliability used in the divergence score described in Weijs et al. (2010); Roulston and Smith (2002).

These quantities are closely related to rank histograms which are discrete version of

Probability Integral Transform (PIT) histograms. However if one can assume the property of flatness of these histograms, Jolliffe and Primo (2008) exhibit a test accounting for the slope and the shape of rank histograms. In a recent work, Zamo (2016) extends this idea for accounting the presence of wave in histograms as seen in Scheuerer and Hamill (2015); Taillardat et al. (2016). A more complete test can thus be implemented that tests each histogram for flatness. Such a test is called the JPZ test (for Jolliffe-Primo-Zamo). The results of the JPZ test is provided for each method in the 3.7.5.

Scoring rules

Following Gneiting et al. (2007); Gneiting and Raftery (2007); Bröcker and Smith (2007b), scoring rules assign numerical scores to probabilistic forecasts and form attractive summary measures of predictive performance, since they address calibration and sharpness simultaneously. These scores are generally negatively oriented and we wish to minimize them. A *proper* scoring rule is designed such that the expected value of the score is minimized by the perfect forecast, ie. when the observation is drawn from the same distribution than the predictive distribution. The *Continuous Ranked Probability Score* (CRPS) (Matheson and Winkler, 1976; Hersbach, 2000) is defined directly in terms of the predictive cdf, F, as :

$$CRPS(F, y) = \int_{-\infty}^{\infty} (F(x) - \mathbf{1}\{x \geq y\})^2 \, \mathrm{d}x.$$

Another representation (Gneiting and Raftery, 2007) shows that :

$$CRPS(F, y) = \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'|,$$

where X and X' are independent copies of a random variable with distribution F and finite first moment.

An alternative representation for continuous distributions using L-moments (Hosking, 1989) is :

$$CRPS(F, y) = \mathbb{E}_F|X - y| + \mathbb{E}_F(X) - 2\mathbb{E}_F(XF(X)).$$

Throughout our study, if F is represented by an ensemble forecast with K members $x_1, \ldots, x_K \in \mathbb{R}$, we use a so-called fair estimator of the CRPS (Ferro, 2014) given by :

$$\widehat{CRPS}(F, y) = \frac{1}{K}\sum_{i=1}^{K}|x_i - y| - \frac{1}{2K(K-1)}\sum_{i=1}^{K}\sum_{j=1}^{K}|x_i - x_j|.$$

Notice that all CRPS have been computed following the recommendations of the Chapter 3 in Zamo (2016).

We can also define the skill score in term of CRPS between an ensemble prediction system A and a baseline B, in order to compare them directly :

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

The value of the CRPSS will be positive if and only if the system A is better than B for the CRPS scoring rule.

3.7.5 Rank histograms boxplots

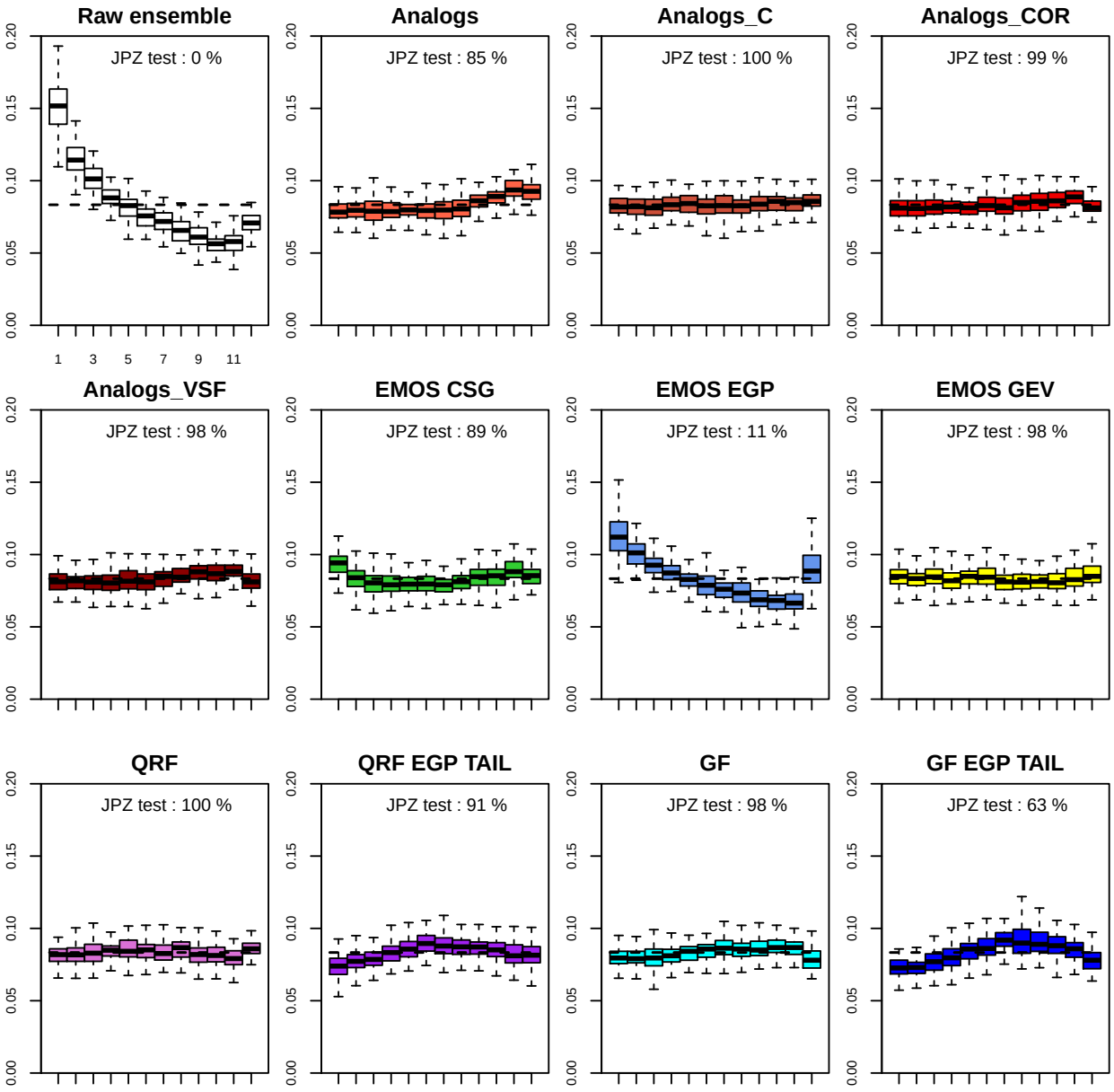

FIGURE 3.5 – Boxplots of rank histograms for each technique according to the locations. The proportion of rank histograms for which the JPZ test does not reject the flatness hypothesis is also provided. The results confirm the Table 3.4.

3.7.6 ROC curves for other thresholds

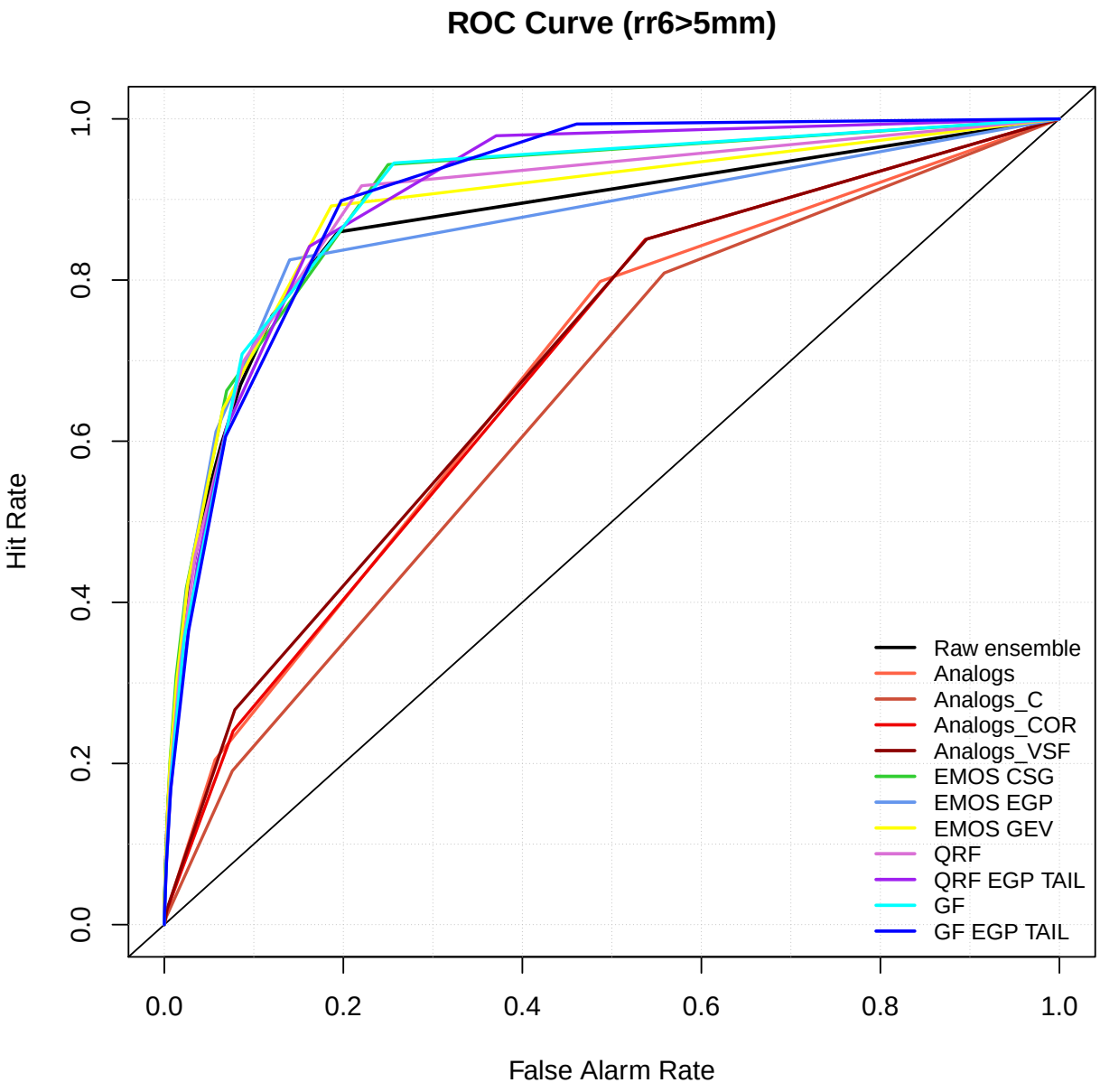

FIGURE 3.6 – ROC Curves for the event of rain above 5mm. A "good" prediction must maximize hit rate and minimize false alarms. The analogs method here lacks resolution.

3.7.7 Modelled ROC curve for high threshold

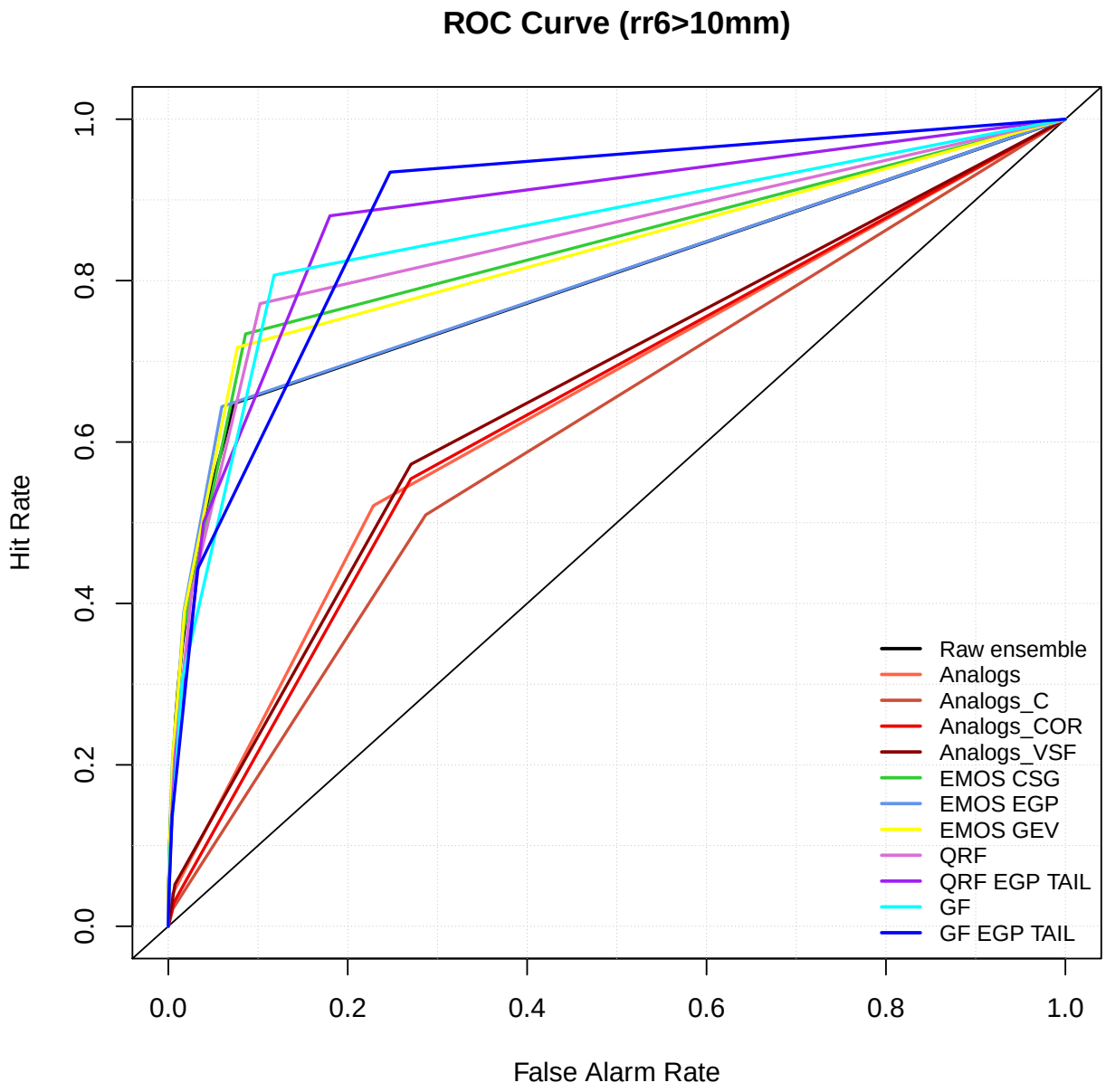

FIGURE 3.7 – ROC Curves for the event of rain above 10mm. A "good" prediction must maximize hit rate and minimize false alarms. The analogs method here lacks resolution. Here, the comments tend to be the same than for the 15mm event's ROC curve.

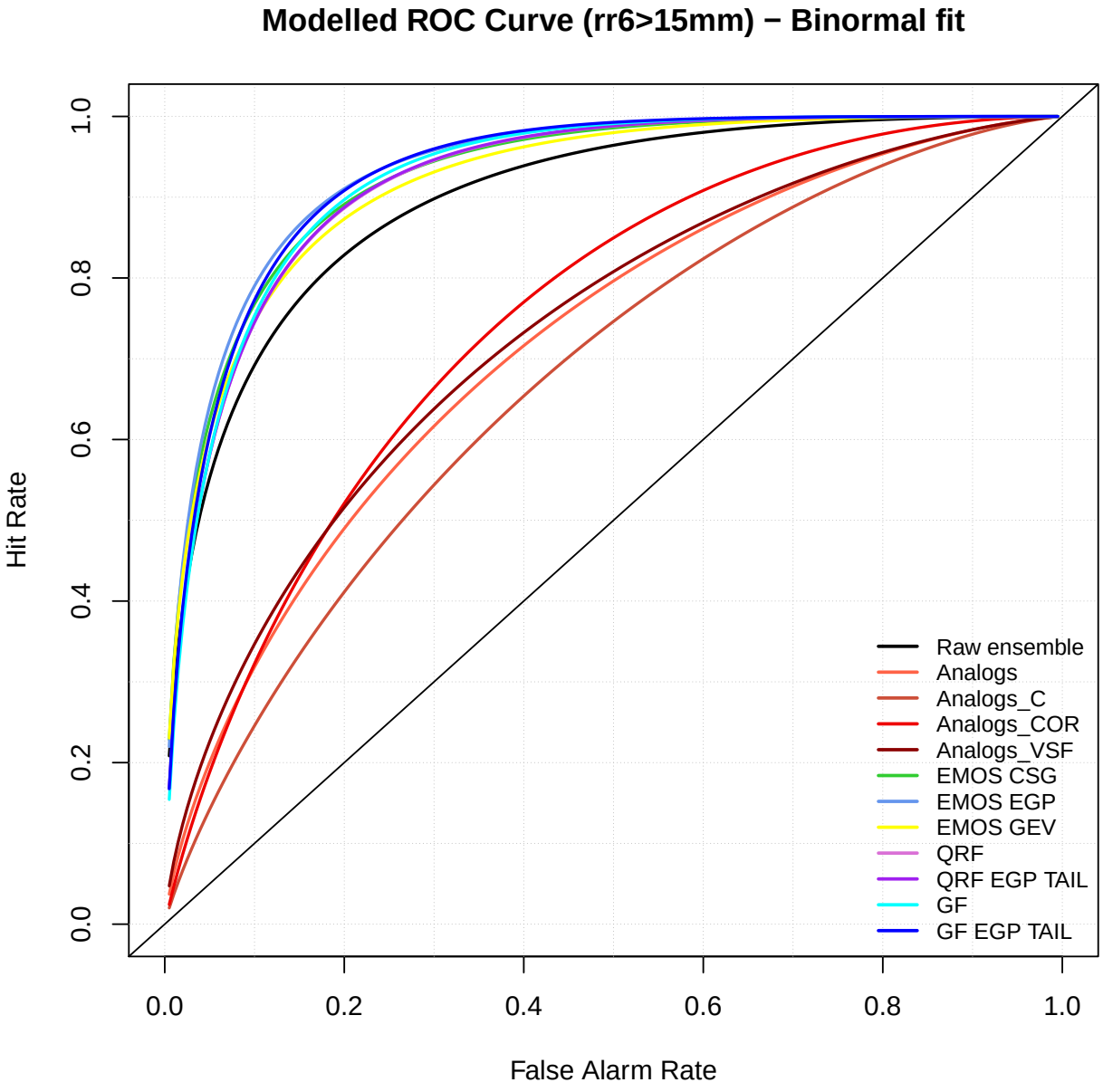

FIGURE 3.8 – Modelled ROC Curves with a binormal fit for the event of rain above 15mm. A "good" prediction must maximize hit rate and minimize false alarms. We show here that the discriminant capacity is nearly the same among calibrated forecasts other than the Analogs one.

Quand on vit avec les fous, il faut faire aussi son apprentissage d'insensé.

Edmond Dantès

Chapitre 4

CRPS-based Verification Tools for Extreme Events

Abstract Verification of ensemble forecasts for extreme events remains a challenging question. The general public as well as the media naturely pay particular attention on extreme events and conclude about the global predictive performance of ensembles, which are often unskillful when they are needed. Ashing classical verification tools to focus on such events can lead to unexpected behaviors. To square up these effects, thresholded and weighted scoring rules have been developed. Most of them use derivates of the Continuous Ranked Probability Score (CRPS). However, some properties of the CRPS for extreme events generate undesirable effects on the quality of verification. Using theoretical arguments and simulation examples, we illustrate some pitfalls of conventional verification tools and propose an original way to assess ensemble forecasts using extreme value theory.

Contents

4.1 Introduction

In a pioneering paper on forecast verification, Murphy (1993) distinguished 3 types of "goodness" :

- the quality : how the forecast corresponds to what actually happened ;
- the consistency : how the forecast corresponds to a forecaster's best judgment, based upon his knowledge base ;
- the value : how the forecast helps the decision maker to proceed efficiently.

The quality of a forecast can be decomposed in some attributes. In order to measure this quality, one uses proper scoring rules (Matheson and Winkler (1976); Gneiting and Raftery (2007); Schervish et al. (2009); Tsyplakov (2013), among others) in order to retrieve the best available forecast in average. Proper scoring rules can be decomposed in terms of reliability, sharpness and resolution. Bröcker (2015) showed that resolution is strongly linked with discrimination. In Gneiting et al. (2007), resolution and reliability are merged into the term *calibration*. Thus the aim of ensembles is to *maximize the sharpness subject to calibration*. In ensemble verification, the most popular scoring rule is the CRPS (Epstein, 1969a; Hersbach, 2000; Bröcker, 2012). Consider an absolutely continuous cdf F and two independent random variables X and X' with cdf F. The CRPS associated to the forecast distribution F and a observed value $y \in \mathbb{R}$ can be defined (among others) as :

$$
\begin{aligned}
CRPS(F,y) &= \int_{-\infty}^{\infty} (F(x) - \mathbf{1}\{x \geq y\})^2 \, dx, && (4.1)\\
&= \mathbb{E}_F|X - y| - \frac{1}{2}\mathbb{E}_F|X - X'|, && (4.2)\\
&= \mathbb{E}_F|X - y| + \mathbb{E}_F(X) - 2\mathbb{E}_F(XF(X)), && (4.3)\\
&= y + 2\overline{F}(y)\mathbb{E}_F(X - y|X > y) - 2\mathbb{E}_F(XF(X)). && (4.4)
\end{aligned}
$$

The expression (4.2) of Gneiting and Raftery (2007) aims at interpreting the terms of calibration and sharpness. The two last formulas are respectively issued from Taillardat et al. (2016); De Fondeville (2014). Regarding extremes verification, it is important to counteract some cognitive biases bounding to discredit skillful forecasters (examples of cognitive biases can be found in Kahneman and Tversky (1979); Morel (2014); Benamran (2017)). That it is called in Lerch et al. (2017) the "Forecaster's dilemma". Indeed, the only remedy is to consider all available cases when evaluating predictive performance. Proper weighted scoring rules (Gneiting and Ranjan, 2011; Diks et al., 2011) attempt to emphasize predefined regions of interest. In particular, keeping the previous notation, denote by w a positive weight function on $\mathbb{R}$, by W its primitive $W(x) = \int_{-\infty}^{x} w(t)dt$ and assume that $\mathbb{E}_F(W(X)) < +\infty$. Then the weighted CRPS can be defined as :

$$
\begin{aligned}
wCRPS(F,y) &= \int_{-\infty}^{\infty} (F(x) - \mathbf{1}\{x \geq y\})^2 w(x)\, dx, && (4.5)\\
&= \mathbb{E}_F|W(X) - W(y)| - \frac{1}{2}\mathbb{E}_F|W(X) - W(X')|, && (4.6)\\
&= \mathbb{E}_F|W(X) - W(y)| + \mathbb{E}_F(W(X)) - 2\mathbb{E}_F(W(X)F(X)), && (4.7)\\
&= W(y) + 2\overline{F}(y)\mathbb{E}_F(W(X) - W(y)|x > y) - 2\mathbb{E}_F(W(X)F(X)). && (4.8)
\end{aligned}
$$

Another aspect just as important as the forecast quality for extremes events is the forecast value. For example, severe weather warnings are still made by forecasters' and despite a possible inaccurate prediction quantitatively speaking, the forecaster has to retrieve some information in the forecast. The approach is completely linked with the economic value of the forecast. For deterministic forecasts, such tools are well-known, see e.g. Richardson (2000); Zhu et al. (2002). Other widely used scores based on the dependence between forecasts and observed events have been considered in Stephenson et al. (2008); Ferro and Stephenson (2011). Recently, Ehm et al. (2016) have introduced the so-called "Murphy diagrams" for deterministic forecasts. This original approach allows to appreciate dominance among different forecasts and anticipate their skill area.

In this chapter, we aim at improving ensemble forecasts value by using Extreme Value Theory (EVT), as was originally suggested in Friederichs et al. (2009); Friederichs (2010). The idea is to use EVT to link observed events and their corresponding CRPS. We do not focus only on score's expectation but on the score as a random variable.

To be more precise, a proper score like the wCRPS satisfies :

$$\mathbb{E}_{Y\sim G}(\text{wCRPS}(G,Y)) \leq \mathbb{E}_{Y\sim G}(\text{wCRPS}(F,Y)).$$

So, an "instantaneous" score like (4.5) can be viewed as a random variable, whenever the observed value y is replaced by the random variable Y itself. Besides, the forecast is considered as known, so that one refers to it via its cdf F only. The properties of the score are captured by its own distribution, in particular its mean for properness. In practice, we make an assumption of ergodicity [1] and the score's mean is averaged on the instantaneous scores available (DeGroot and Fienberg, 1983; Dawid, 1984; Gneiting and Raftery, 2007).

Our goal is to foresee the distributional behaviour of the CRPS subject to an extreme event. Using the CRPS, we ensure that the best forecast in average is never discredited.

This chapter is organized as follows : in Section 4.2 we pinpoint some undesirable properties of the CRPS and its weighted derivation. We expose the non-tail equivalence of the wCRPS and the potential difficulties of using the wCRPS for extreme weather evaluation. Section 4.3 links the tail behaviour of the observations with the tail behaviour of the CRPS. The chapter closes with a discussion in Section 4.4.

1. Such an assumption can be debated but is essential if one wants to estimate scores.

4.2 Tail equivalence, wCRPS and choice of a weighting function

4.2.1 Non tail equivalence of the CRPS

Definition 1 (Resnick (1971))**.** Two distribution functions F and G are *tail equivalent* iff they share the same endpoint $x_F = x_G$ and

$$\lim_{x \to x_F} \frac{1 - F(x)}{1 - G(x)} = c \in (0, +\infty).$$

The properness of the CRPS and its weighted derivations ensure that for tail equivalent cdfs F and G, $\mathbb{E}_G(\text{wCRPS}(G, Y)) \leq \mathbb{E}_G(\text{wCRPS}(F, Y))$. But the wCRPS is not a tail equivalent score, as the following theorem holds :

Theorem 1. *For any given $\epsilon > 0$, it is always possible to construct a cdf F that is not tail equivalent to G such that*

$$|\mathbb{E}_G(\text{wCRPS}(G, Y)) - \mathbb{E}_G(\text{wCRPS}(F, Y))| \leq \epsilon,$$

This proposition is proven in Appendix 4.5.1.

This non tail equivalence of the CRPS can explain the weakness for estimating the shape parameter in Friederichs and Thorarinsdottir (2012). In addition, Figure 1 in Friederichs and Thorarinsdottir (2012) showed that the ignorance score is more sensitive to outliers than the CRPS. In order to compare the performance between underlying distributions (not only ensemble forecasts) these scores may be unable to clearly distinguish different characteristic patterns concerning tail behaviors.

4.2.2 Weight functions and forecast comparison

We consider here the design of experiments introduced by Gneiting et al. (2007) for evaluating predictive performance. This scenario is described in Table 4.1. At times $t = 1, 2, ...$ one chooses a normal distribution $G_t = \mathcal{N}(\mu_t, 1)$ [2], where μ_t is a random draw from the standard normal distribution. We assume that the μ_t are independent and that the observed values x_t come from the cdf G_t. Several forecasters with distributions F_t are competing. The *perfect* forecaster issues a perfect probabilistic forecast ie. $F_t = G_t$. The *climatological* forecaster takes the unconditional distribution, $F_t = \mathcal{N}(0, 2)$, as probabilistic forecast. The *unfocused* forecaster adds a Rademacher-type bias in his forecast. The *sign-reversed* and the *extremist* forecasters are both biased (see Table 4.1). More informations on how predictive distributions need to be assessed can be found in Dawid (1984); Diebold et al. (1997).

We want to assess the predictive performance on extreme events. In this context we use the simple weighting function $W_q(x) = x\mathbf{1}\{x \geq q\}$ to compute the expected weighted

2. We write $\mathcal{N}(\mu, \sigma^2)$ for the normal distribution with mean μ and variance σ^2.

TABLE 4.1 – Design of the simulation study. Both the $(\mu_t)_t$ and the $(\tau_t)_t$ are independent and identically distributed. They are mutually independent of each other. Here $t = 10^7$.

Truth	$G_t = \mathcal{N}(\mu_t, 1)$ where $\mu_t \sim \mathcal{N}(0, 1)$
Forecasts :	
Perfect forecaster	$F_t = \mathcal{N}(\mu_t, 1)$
Climatological forecaster	$F_t = \mathcal{N}(0, 2)$
Unfocused forecaster	$F_t = \frac{1}{2}(\mathcal{N}(\mu_t, 1) + \mathcal{N}(\mu_t + \tau_t, 1))$ with $\tau_t = \pm 2$ with $1/2$ probability each
Extremist forecaster	$F_t = \mathcal{N}(\mu_t + \frac{5}{2}, 1)$
Sign-extremist forecaster	$F_t = \mathcal{N}(-\mu_t - \frac{5}{2}, 1)$

CRPS for each forecaster. The choice of this function is arbitrary here. The variation of q between the quantile thresholds 0.75 and 1 of the climatology distribution here provides as many weighted scores as we want. Consequently, for each quantile q there can be a different forecast ranking. Nevertheless, the scoring rule theory ensures that the perfect forecaster has the lowest score expectation in any case.

Figure 4.1 provides the weighted score expectations in terms of quantile thresholds. Several remarks can be formulated.

First, we (hopefully) expect that proper scoring rules reward the best forecast (ie. the lowest, the best). But the ranking of the forecasts can be different for two very close weight functions. We can observe the ranking in Figure 4.1 between the quantiles 0.85 and 0.9. Last, for very high threshold (ie. when the base rate is vanishing), weighted scores converge to zero. This loss of information is well-known by Brier score users. This issue is clearly pointed out by Stephenson et al. (2008).

From a more general point of view, and as forecast makers, we can wonder whether the use of weight scoring rules should be harmonized. The choice of weighting a score should not be a consideration of forecast makers but forecast users. This opinion is relayed in Ehm et al. (2016) by the purpose of Gneiting and Ranjan (2011) and Patton (2015) respectively arguing that "the scoring function [must] be specified ex ante" and "forecast consumers or survey designers should specify the single specific loss function that will be used to evaluate forecasts". This is an essential point for forecasts in economics. For weather forecast, end users are often quite unaware of the potential impact of poor forecasts. Moreover, it can happen that the cost/loss ratio cannot be easily quantified (for example, quantify the economic value of a human life). This can be the main argument for the underuse [3] of relative value in meteorology studied by Richardson (2000); Wilks (2001); Buizza (2001); Zhu et al. (2002).

This also raises the delicate question of the uniqueness of weather services' forecasts in

3. A recent survey (Hagedorn, 2017) concludes that 81% of probabilistic information users relies on heuristic/experience rather than cost/loss models in order to take decisions.

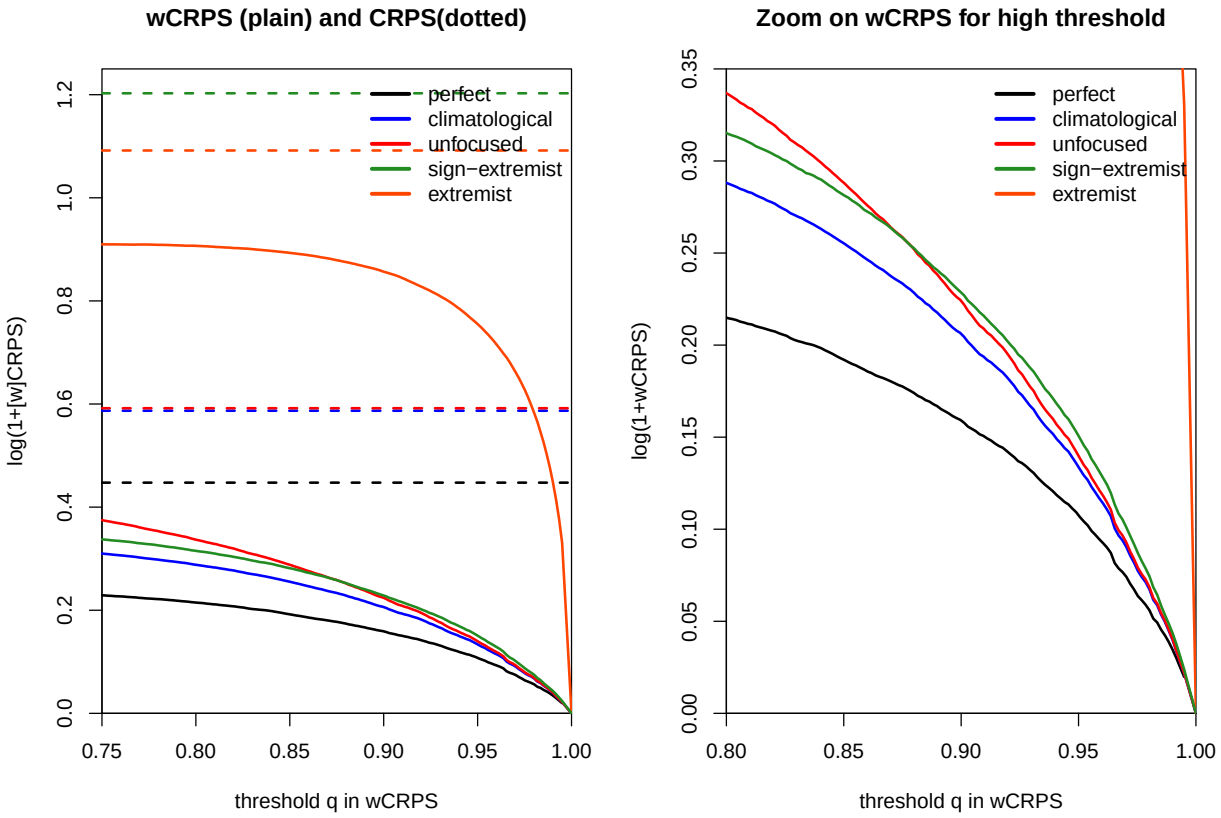

FIGURE 4.1 – Values of $\log(1 + \mathbb{E}_G[\mathrm{CRPS}(F, Y)])$ (left) and $\log(1 + \mathbb{E}_G[\mathrm{w_qCRPS}(F, Y)])$ (right) among with $W_q(x) = x\mathbf{1}\{x \geq q\}$ in (4.8) and quantile q represents the x-axis. The forecast ranking changes between very close weight functions. For high threshold, it is not possible to distinguish a clear forecast ranking among forecasters.

meteorology.

4.3 A CRPS-based tool using extreme value theory

In this Section, we aim at providing another way of assessing the performance of ensemble forecasts for extreme events. Ferro (2007) tried to link EVT with forecast verification of extreme events. He gave a theoretical framework to characterize the joint distribution of forecasts and observations. This concerns deterministic forecasts and Stephenson et al. (2008) states that "development of verification methods for probability forecasts of extremes events is an important area that clearly requires attention". Indeed, one drawback of non-deterministic forecasts for extreme verification is that the forecast has to be summarized in an informative way in order to be compared to the observation. One simple solution can

be to keep the maximum or the mode of the forecast distribution but this approach may penalize ideal forecast and leads to improper scores. Our idea is to summarize the relation between forecast and observation by the unweighted CPRS itself. Using theoretical results and EVT, we propose a new index for judging ensemble forecast quality for extreme events.

4.3.1 Behavior of the CRPS for extreme events

In the following, let X and Y be two random variables with finite means and absolutely continuous cdfs respectively denoted by F and G. Denote also x_F (respectively x_G) their right endpoints and f (respectively g) their densities.

Theorem 2 (Fréchet (1927); Fisher and Tippett (1928); Gnedenko (1943))**.** *Let $X_1, \cdots, X_n$ be independent and identically distributed random variables with cdf F, for which there exist appropriate constants $b_n > 0$ and $a_n \in \mathbb{R}$ such that*

$$\lim_{n\to+\infty} \mathbb{P}\left(\frac{\max[X_1, \cdots, X_n] - a_n}{b_n} \leq x\right) = H_\gamma(x), \quad x \in \mathbb{R},$$

where H_γ is a non-degenerate cdf (called "attractor of F". One says equivalently that F is "in the domain of attraction of H_γ", denoted by $F \in \mathcal{D}(H_\gamma)$).

Thus H_γ can be of the following type :

1. $H_0(x) = \exp(-e^{-x}), \quad x \in \mathbb{R}$, *Gumbel distribution ;*
2. $H_\gamma(x) = \exp(-x^{-1/\gamma}), \quad x \geq 0, \gamma > 0$, *Fréchet distribution ;*
3. $H_\gamma(x) = \exp(-(-x)^{-1/\gamma}), \quad x \leq 0, \gamma < 0$, *Weibull distribution.*

Theorem 3 (De Haan (1970))**.** *Let γ be a real number. A distribution F belongs to $\mathcal{D}(H_\gamma)$ iff for some positive auxiliary function b and $1 + \gamma t > 0$*

$$\frac{1 - F(u + tb(u))}{1 - F(u)} = \mathbb{P}\left(\frac{X - u}{b(u)} > t | X > u\right) \longrightarrow (1 + \gamma t)^{-1/\gamma} = 1 - GP_{\gamma,1}(t) \tag{4.9}$$

as $u \to x_F$, where the generalized Pareto distribution, $GP_{\gamma,\sigma}$ has the following form :

$$GP_{\gamma,\sigma}(x) \quad = \quad \begin{cases} 1 - \left(1 + \frac{\gamma x}{\sigma}\right)^{-\frac{1}{\gamma}}, & \text{if } \gamma \neq 0, \\ 1 - \exp(-\frac{x}{\sigma}), & \text{if } \gamma = 0, \end{cases} \tag{4.10}$$

with x positive when γ is positive and $0 \leq x \leq -\frac{\sigma}{\gamma}$ when γ is negative.

Lemma 1. *Consider a random variable X with finite mean that belongs to domain of attraction $\mathcal{D}(H_\gamma)$ for $\gamma < 1$. There exist real positive numbers α and β such as*

$$0 \leq 2\mathbb{E}_F(X - u | X > u) \leq \alpha u + \beta, \tag{4.11}$$

for large $u \to x_F$. We define by $M(F, u)$ the excess mean function of F : $M(F, u) = \mathbb{E}_F(X - u | X > u)$.

Proof. Let decompose the proof depending on the sign of γ :

1. First case : F belongs to $\mathcal{D}(H_\gamma)$ with $0 < \gamma < 1$:
   In this case, Embrechts et al. (1997) (Section 3.4) shew that $M(F,u) \sim \frac{\gamma u}{1-\gamma}$ as $u \to x_F$, and we can conclude directly.
2. Second case : F belongs to $\mathcal{D}(H_\gamma)$ with $\gamma < 0$:
   In this case, the result also follows easily from Embrechts et al. (1997) since when $u \to x_F$, $M(F,u) \sim \frac{\gamma(x_F-u)}{\gamma-1}$.
3. Third case : F belongs to $\mathcal{D}(H_0)$:
   When F is in the Gumbel domain of attraction, the Theorem 3.9 in Ghosh and Resnick (2010) ensures that $\frac{M(F,u)}{u} \longrightarrow 0$ as $u \to x_F$.

□

This lemma allows to derive the asymptotic conditional behavior of $\mathrm{CRPS}(F,Y)$. This is the purpose of the following theorem.

Theorem 4. *Let X and Y be two independent random variables with finite first moments and respectively absolutely continuous cdfs F and G such as $x_F = x_G$. If G belongs to $\mathcal{D}(H_\gamma)$ and $c_F = 2\mathbb{E}_F(XF(X))$, then*

$$\mathbb{P}\left(\frac{\mathrm{CRPS}(F,Y) + c_F - u}{b(u)} > t | Y > u\right) \longrightarrow (1+\gamma t)^{-1/\gamma} \tag{4.12}$$

as $u \to x_F$.

Proof. According to the formula (4.4) of the CRPS, we can write that

$$\mathrm{CRPS}(F,Y) \overset{a.s.}{=} Y - c_F + 2\overline{F}(Y)\mathbb{E}_F(X - Y | X > Y).$$

Fix a large (conditionnally to Y) u, one gets thanks to Lemma 1 :

$$Y \leq \mathrm{CRPS}(F,Y) + c_F \leq (1 + \alpha\overline{F}(Y))Y + \beta\overline{F}(Y) \quad a.s.$$

So that

$$\begin{aligned} 0 \leq \mathbb{P}\left(\frac{\mathrm{CRPS}(F,Y) + c_F - u}{b(u)} > t | Y > u\right) - \mathbb{P}\left(\frac{Y-u}{b(u)} > t | Y > u\right) \quad \leq \\ \mathbb{P}([1+\alpha\overline{F}(Y)]Y + \beta\overline{F}(Y) > tb(u) + u | Y > u) - \mathbb{P}(Y > tb(u) + u | Y > u) \quad \leq \\ \mathbb{P}\left(Y > \frac{tb(u) + u - \beta\overline{F}(u)}{1+\alpha\overline{F}(u)} | Y > u\right) - \mathbb{P}(Y > tb(u) + u | Y > u). \end{aligned}$$

We recognize the probability for Y to be in an interval denoted by $[\delta_u, \Delta_u]$:

$$\frac{1}{\overline{F}(u)}\mathbb{P}\left(Y \in \left[\frac{tb(u)+u-\beta\overline{F}(u)}{1+\alpha\overline{F}(u)}, tb(u)+u\right]\right) = \frac{\mathbb{P}(Y \in [\delta_u, \Delta_u])}{\overline{F}(u)}$$

Notice then that

$$\begin{aligned}\frac{\mathbb{P}(Y \in [\delta_u, \Delta_u])}{\overline{F}(u)} &\leq \sup_{v\in[\delta_u,\Delta_u]} g(v)\,\frac{\Delta_u-\delta_u}{\overline{F}(u)} \\ &= \sup_{v\in[\delta_u,\Delta_u]} g(v)\,\frac{\alpha(tb(u)+u)+\beta}{1+\alpha\overline{F}(u)} \\ &= O(ug(u)) \longrightarrow 0\end{aligned}$$

as $u \to x_F$. The dominance in $ug(u)$ is provided by the sublinear/linear behavior of b (Von Mises condition in Von Mises (1936)). Indeed, Von Mises (1936) noticed that a possible choice for $b(u)$ can be the mean excess function of Y which is (sub)linear. The limit to 0 is due to the finite first moment of the random variable Y, because in this case $ug(u) \sim 1-G(u) \to 0$ as $u \to x_F$.

□

The previous theorem shows that for large values of an observed event, the resulting CRPS value according to this event is mainly led by the climatology. The constant component c_F, only depending on the forecast, can be interpreted. We can write c_F as :

$$c_F = \mathbb{E}_F(X) + \frac{1}{2}\mathbb{E}_F(|X-X'|)$$

In this theoretical framework, there are only two ways to get a lower CRPS for extreme events (ie. to make c_F larger) :

1. improving the mean of our forecast conditionnally to extreme observations,
2. inflate the mean absolute difference of our forecast.

The first statement is the most natural to understand. First, increasing the mean of the forecast is not contradictory with keeping a high sharpness. Note also that making c_F larger for extreme events would reduce the bias of our forecast for these events. The paradigm about ensemble verification remains. To sum up, one can improve our forecast for extremes by reducing the bias for these events or by increasing the dispersion. The first solution is generally preferable but the easiest solution depends on the variable of interest. If we recall the study in Section 4.2.2, for normal distributed forecasts $\mathcal{N}(\mu, \sigma^2)$ if we want to add 1 to c_F it is easier to do $\mu' = \mu + 1$ than $\sigma' = \sigma(1+\sqrt{\pi}/\sigma)$. Since if $F \sim \mathcal{N}(\mu, \sigma^2)$ then

$$\frac{1}{2}\mathbb{E}_F(|X-X'|) = \frac{\sigma}{\sqrt{\pi}}.$$

For ensemble forecasts, the latter statement is equivalent to improve the deterministic forecast (ie. correcting the bias) priority to uncertainty quantification.
Nevertheless, for some variables (typically leptokurtic such as wind speed or rainfall), the mean is not a good summary of the variable behavior. Uncertainty quantification improvement should be the priority in this case. A tradeoff can be made in order to increase ensemble dispersion for these variables. This statement joins the conclusions of Williams et al. (2014) on Lorenz 1996's setting.

As a conclusion, we can say that the tradeoff between bias reduction and variance inflation must be the pinpoint of the choice among post-processing methods and their inherent possibilities. It also underlines the necessity of a simultaneous correction of bias and uncertainty.

4.3.2 How to use Pareto approximation of the CRPS ?

Asymptotic approximations

Using both the theorem 3 and the Balkema and De Haan (1974); Pickands III (1975)'s theorem, we can approximate the conditional distribution of $\text{CRPS}(F,Y)+c_F$ for large values of Y by the same Generalized Pareto (GP) distribution as the one approximating $Y|Y>u$.

$$\text{CRPS}(F,Y)+c_F|Y>u \;\sim\; Y|Y>u \;\sim\; GP_{\gamma,\sigma}. \tag{4.13}$$

If one wants to use the location parameter of the GP, and denoting by $\Upsilon_u(x)$ the cdf of the conditional distribution $\text{CRPS}(F,Y)|Y>u$ we have the following approximation :

$$\Upsilon_u(x) \;\approx\; \begin{cases} GP_{\sigma_u,\gamma,-c_F}(x) = 1-\left(1+\frac{\gamma(x+c_F)}{\sigma_u}\right)^{-\frac{1}{\gamma}}, & -c_F \le x \le -c_F-\frac{\sigma_u}{\gamma} \text{ if } \gamma<0, \\ GP_{\sigma_u,\gamma,-c_F}(x) = 1-\left(1+\frac{\gamma(x+c_F)}{\sigma_u}\right)^{-\frac{1}{\gamma}}, & x \ge -c_F \text{ if } \gamma=0, \\ GP_{\sigma_u,\gamma}(x) = 1-\left(1+\frac{\gamma x}{\sigma_u}\right)^{-\frac{1}{\gamma}}, & x \ge 0 \text{ if } \gamma>0. \end{cases} \tag{4.14}$$

The fact that c_F is vanishing in the Fréchet case is the result of the linear behavior of the auxiliary function in formula (4.9) (Von Mises, 1936; Embrechts et al., 1997).

For large enough quantile thresholds u (so that the Pareto approximation can be considered as acceptable), the idea is to compare the empirical cdf generated by the CRPS values (denoted $K_{u,n}$ here) to the theoretical Pareto cdf (denoted by Υ_u). The parameters of the Pareto cdf are estimated on the n observations above the chosen threshold u. In order to assess the goodness of fit we rely on the Cramér-von Mises criterion (Cramér, 1928; Von Mises, 1928) :

$$\omega_u^2 = \int_{-\infty}^{+\infty} [K_{u,n}(t)-\Upsilon_u(t)]^2 d\Upsilon_u(t).$$

Let $v_1, \cdots, v_n$ be the ordered CRPS values (in increasing order). We recall that the v_i are the CRPS values of the forecast F for observations above u. The test statistic can be rewritten as :

$$T_u = n\omega_u^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - \Upsilon_{u,i}(v_i) \right]^2 .$$

If the value of T_u is larger than some tabulated values, the hypothesis that the CRPS values come from Υ_u can be rejected. A description of the algorithm calculating T_u is provided in Table 4.2.

How could we get an index from it ?

Actually, we can easily admit that a latent variable Z (the state of the atmosphere) is inferred by the forecast X. The observation is a variable Y trying to realize a partial state of Z. If we stand that the Pareto framework suits with theoretical assumptions that we have made, taking into account the link between the forecast and the associated observations makes the result above collapse. Stated differently, we claim that the higher the statistic T_u is, the better the forecast is for extremes. We admit that a formal argument is missing so far, but we surmise that this is relevant.

Thus, we can calculate the p-value. Instead of defining a significance level, the p-value renders the non-association between high CRPS and high observations. Here, we are just using the test statistic in order to get a bounded index in $[0, 1]$. We do not use the test in order to accept or reject a hypothesis.

Of course, in practice CRPS values can be high without the observations being large (corresponding to the type II error). That is why it is important that forecasts should be calibrated before computing this p-value, as Ferro and Stephenson (2011) already recommended.

Our previous argument therefore suggests that we want the lowest p-value for our forecasts.

We propose to take the same design experiment than in Section 4.1. We have our 10^7 CRPS (and c_F) values for each forecast. For climatology thresholds u from 0.75 to 0.9995 we draw p-values representing the goodness of fit with the expected Pareto distribution. We are not in the Fréchet case but in the Gumbel one so c_F is taken into account. The Table 4.2 describes how to obtain the p-values and thus the Figure 4.2 (and 4.3 further). The results are available in Figure 4.2.

Figure 4.2 is the climax of this chapter and it is very informative.

The yellow curve in Figure 4.2 increases from the threshold 0.75 ,to the threshold 0.82, and then decreases. The switch in the curve's variation can be interpreted as the starting threshold for extremes verification. Another mean to assess the starting threshold can be to

TABLE 4.2 – Description of the algorithm used to compute test statistic and p-value for a forecast F on a dataset of N couples forecast/observation.

0. Computation of the CRPS values :	-For the N couples forecast/observation compute the N corresponding instantaneous CRPS.
1. Determination of the parameter γ and the range where the Pareto approximation is suitable :	-Using the N observations, find a threshold u_0 where the Pareto approximation is acceptable and estimate the Pareto shape parameter γ .
2. For a threshold $u \geq u_0$:	-Estimate the Pareto scale parameter σ_u using the n observations above the observations' quantile of order u.
If $\gamma \leq 0$:	-Compute the n values of c_F associated to the n CRPS values of the n highest observations (the CRPS values have been computed in step 0.).
3. Determination of the statistic T_u	-Order the n CRPS values (associated to the n highest observations) in increasing order $v_1, \ldots, v_n$.
For $i \in [1, n]$	-Compute for each CRPS value v_i, $\Upsilon_{u,i}(v_i)$.
	-Compute $\left[\frac{2i-1}{2n} - \Upsilon_{u,i}(v_i)\right]^2$.
	-After summation you have T_u.
End 3.	
	-You have T_u (and so the corresponding p-value) for each threshold above u_0.
End 2.	

look at the intersection between the climatological and the perfect CRPS ; but it is unfeasible in practice.

Next, we see that the p-value plot keeps the properness of the CRPS (the perfect remains unbeated). Moreover, the ranking of the forecasts is both quite logical for extremes and consistent with the classical CRPS prior ranking. Indeed, the perfect is first, the unfocused forecast is preferred to the climatological forecast and the gap is increasing with the threshold. One can think here that the gain of information between the unfocused forecast and the climatological forecast is rewarded. Last but not least, the two high-biased forecasts are highly penalized. This behavior shows that the p-value is also sensitive to bias. Thus the p-value shares many attributes with the CRPS.

We can complete this study with an illustration on a real dataset that involves the testbed of the previous chapter (see Section 3.4).

We are in a Fréchet case here. The starting quantile order is taken at 0.935 corresponding

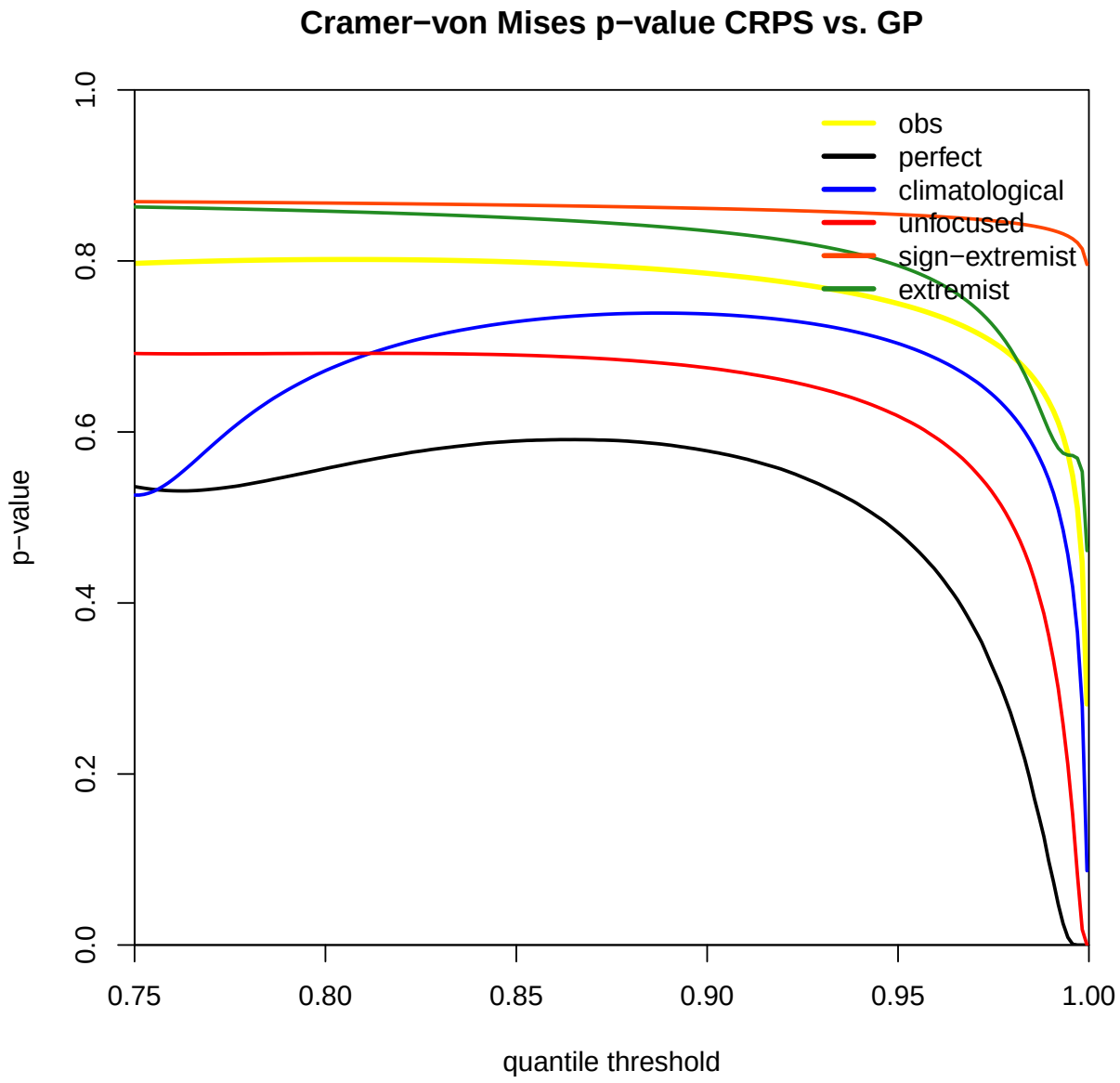

FIGURE 4.2 – Cramér-von Mises' p-values as a function of the threshold. The lower the better. The yellow distribution represents the goodness of fit with thresholded observations. The p-value seems to keep properness, is bias- and variance-sensitive.

to the 3mm rainfall amount. P-value plot is provided in the Figure 4.3. Figure 4.3 confirms the results of the previous chapter, especially the Figure 3.3. A major difference remains : the behavior of the GF method. Despite its good results in the previous chapter, this technique is ranked just before the raw ensemble here. After looking at the connexion between high

observations and high CRPS, we have seen that the CRPS of the GF technique is about 3mm higher than the CRPS for techniques ranked below [4]. This explains the a priori suspicious behavior of the GF technique here.

4.4 Discussion

In this chapter, we notice that the CRPS and its weighted version can exhibit some undesirable properties. This does not call into question its properness, simplicity of use and thus popularity in ensemble verification. We now discuss two main issues considered :

The first one is the choice of an appropriate weighting function. We show on a simple example how it could be tricky sometimes to assess verification, especially for forecast ranking. This work in meteorology comes in addition of Lerch et al. (2017), and we fully agree with the statements about weighting specification described in Patton (2015); Gneiting and Ranjan (2011).

The second concern is how to deal with ensemble forecast and extreme events. Inspired by Friederichs (2010), the choice is to apply extreme value theory on common verification measures itselves. Relying on some properties of the CRPS for large observed events we put a theoretical framework concerning the score's behavior for extremes. As a result, we obtain a bounded index in $[0, 1]$ to assess the nexus between forecasts and observations. In addition, this index seems to be suitable for extreme precipitation forecast assessment.

One can view this work as an additional step in bridging the gap in the field of ensemble verification and extreme events, see Ferro (2007); Friederichs and Thorarinsdottir (2012); Ferro and Stephenson (2011). The ensemble forecast information is kept by the use of the CRPS and this measure, as a dependence index, can be considered like the probabilistic alternative to the deterministic scores introduced by Ferro (2007); Ferro and Stephenson (2011). This index may help forecasters to take better decisions. This index is directly linked with the value of the ensemble. Relying on it, we are able to say that the paradigm of *maximizing the sharpness subject to calibration* can be associated with the paradigm of *maximizing the value for extreme events subject to a good overall performance*. In this way, and as a future work, it would be convenient to study the specific properties of this CRPS-based tool and its potential paths and pitfalls. Another potentially interesting investigation could be to extend this procedure to other scores like the ignorance [5] score (Diks et al., 2011) or the Dawid-Sebastiani score (Dawid and Sebastiani, 1999).

4. Please be very careful when understanding this sentence : what we said here lead to an improper score. It is just to be taken as a qualitative remark !

5. Indeed, the Pearson-Neyman lemma described in Lerch et al. (2017) let us think that this score could be a natural candidate.

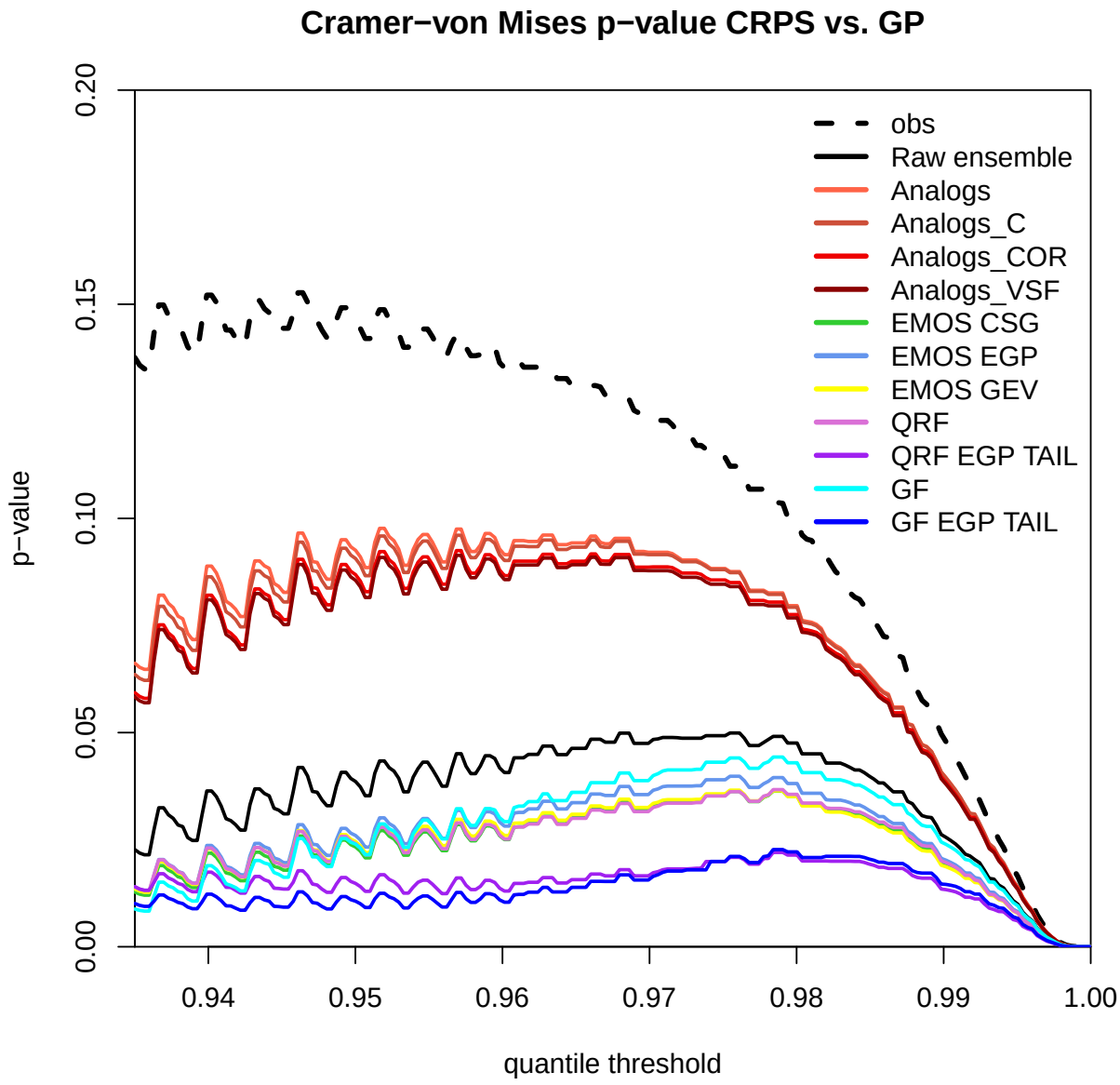

FIGURE 4.3 – Cramér-von Mises' p-values as a function of the threshold for the 6-h rainfall forecast. The lower the better. This plot confirms the results of the previous chapter, especially the Figure 3.3.

4.5 Appendix

4.5.1 Proof of proposition 1

This proposition was already proven by Raphaël de Fondeville, Philippe Naveau and Daniel S. Cooley. We expose here a proof for the unweighted CRPS.

Proof. Let u be a positive real. Denote Z a non-negative random variable with finite mean and cdf H.

We introduce the new random variable

$$Y = X\mathbf{1}\{u \geq X\} + (Z+u)\mathbf{1}\{X > u\} \tag{4.15}$$

with survival function $\overline{G}$ defined by

$$\overline{G}(x) \;=\; \begin{cases} \overline{F}(x), & \text{if } x \leq u \\ \overline{H}(x-u)\overline{F}(u), & \text{otherwise ,} \end{cases} \tag{4.16}$$

where Z has the same end point than X, $\overline{H}(0) = 1$ and

$$\overline{H}(x-u) \leq \overline{F}(x)/\overline{F}(u), \text{ for any } x \geq u. \tag{4.17}$$

This latter condition implies that

$$\overline{G}(x) \leq \overline{F}(x), \text{ for all } x. \tag{4.18}$$

Equation 4.16 allows to write almost surely that

$$\mathbb{E}(X\mathbf{1}\{X < x\}) = \mathbb{E}(Y\mathbf{1}\{Y < x\}) \text{ , for any } x \leq u. \tag{4.19}$$

Equality 4.19 combined with the expression of the CRPS implies that

$$\begin{aligned} \frac{1}{2}[\mathrm{CRPS}(G,x) - \mathrm{CRPS}(F,x)] &= \mathbb{E}_Y[(Y-x)\mathbf{1}\{Y > x\}] - \mathbb{E}_X[(X-x)\mathbf{1}\{X > x\}] \\ &\quad + \mathbb{E}_X(XF(X)) - \mathbb{E}_Y(YG(Y)), \\ &= \mathbb{E}_Y(Y\overline{G}(Y)) - \mathbb{E}_X(X\overline{F}(X)) \\ &\quad - \mathbb{E}_Y[(Y-x)\mathbf{1}\{Y \leq x\}] + \mathbb{E}_X[(X-x)\mathbf{1}\{X \leq x\}] \\ &= \mathbb{E}_Y(Y\overline{G}(Y)) - \mathbb{E}_X(X\overline{F}(X)) + \int_u^{x_F} \Delta(x)dF(x), \end{aligned}$$

where

$$\Delta(x) = \mathbb{E}_X[(X-x)\mathbf{1}\{X \leq x\}] - \mathbb{E}_Y[(Y-x)\mathbf{1}\{Y \leq x\}].$$

The stochastic ordering between Y and X implies that $\mathbb{E}_Y(Y\overline{G}(Y)) - \mathbb{E}_X(X\overline{F}(X)) \leq 0$. This leads to

$$\frac{1}{2}\left|\mathbb{E}_X[\text{CRPS}(G,X)] - \mathbb{E}_X[\text{CRPS}(F,X)]\right| \leq \int_u^{x_F} \Delta(x) dF(x).$$

For $x > u$ we can write that

$$\begin{aligned}
\Delta(x) &= \mathbb{E}_X[(X-x)\mathbf{1}\{u < X \leq x\}] - \mathbb{E}_Y[(Y-x)\mathbf{1}\{u < Y \leq x\}], \\
&\leq \mathbb{E}_Y[(x-u)\mathbf{1}\{u < Y \leq x\}], \text{ because } X - x \leq 0 \text{ and } 0 \leq x - Y \leq x - u \text{ here}, \\
&= (x-u)[G(x) - G(u)], \\
&\leq (x-u)\overline{G}(u), \\
&= (x-u)\overline{F}(u).
\end{aligned}$$

Hence, we can write that

$$\begin{aligned}
\left|\mathbb{E}_X[\text{CRPS}(G,X)] - \mathbb{E}_X[\text{CRPS}(F,X)]\right| &\leq 2\overline{F}(u)\int_u^{x_F}(x-u)dF(x), \\
&\leq 2\overline{F}^2(u)\mathbb{E}_X[X-u|X>u].
\end{aligned}$$

This inequality is true for any u and H. The right hand side of the last inequality does not depend on $\overline{H}(x)$. Thus, the tail behaviour of the random variables X and Z can be completely different, although the CRPS of F and G can be as closed as one wishes. The right hand side goes to 0 due to the finite mean of X.

□

I've seen things you people wouldn't believe.
Attack ships on fire off the shoulder of Orion.
I watched C-beams glitter in the dark near the Tannhäuser Gate.
All those moments will be lost in time,
like tears in rain.

Roy Batty

Epilogue

Les contributions de cette thèse concernent le post-traitement statistique des systèmes de prévisions d'ensemble météorologiques. L'objectif est d'introduire de nouvelles méthodes non-paramétriques de post-traitement et de les comparer aux techniques de référence dans le domaine. Par rapport à ces dernières, nos méthodes basées sur les forêts aléatoires, offrent de nombreuses fonctionnalités comme la prise en compte de potentiels phénomènes non-linéaires entre covariables. Ceci est primordial en météorologie de par le caractère même des équations qui régissent la dynamique atmosphérique. De plus, nos techniques possèdent l'avantage de gérer très facilement des prédicteurs d'une nature différente de la variable à calculer, et sélectionnent automatiquement les prédicteurs dont elles ont besoin.

Les approches développées dans le cadre de cette thèse nécessitent toutefois des historiques de données assez long (de l'ordre de l'année ici), et des ressources informatiques conséquentes. Cependant, à l'heure des "données massives" et d'une mise à jour régulière de la capacité des super-calculateurs, nous pensons que nos propositions constituent une réelle alternative aux techniques existantes. De récents travaux (Chen et al., 2017; Genuer et al., 2017) montrent néanmoins que nos méthodes s'inscrivent tout à fait dans les problématiques actuelles sur les "données massives".

Les résultats obtenus sur des paramètres classiques tels que la vitesse du vent et la température de surface sont très encourageants. En termes quantitatifs, nos techniques présentent un gain en performance quasi constant quelque soit l'échéance de la prévision. Ce gain est même plus conséquent pour la vitesse du vent, un phénomène bien moins linéaire que la température. De plus, nous montrons que le post-traitement non-paramétrique peut amener une réelle plus-value en terme de valeur économique de la prévision. En effet, un large choix de prédicteurs peut nous permettre de détecter des phénomènes jusqu'à maintenant non prévus par le modèle non post-traité. L'exemple le plus significatif est dans ce travail un cas de refroidissement radiatif par ciel clair et sol enneigé.

Il a été question dans un deuxième temps de traiter la délicate question du post-traitement ensembliste des précipitations sexti-horaires. Pour cela, de nombreux moyens pour traiter ce paramètre spécifique ont été étudiés. Un travail préliminaire conséquent a été réalisé sur les méthodes déjà existantes en ce qui concerne les distributions choisies pour modéliser l'ensemble de précipitation, les métriques utilisées pour estimer les paramètres de ces lois ainsi que les prédicteurs choisis pour de telles estimations.

Par ailleurs, une nouvelle façon de partitionner l'espace au sein même des forêts aléatoires a été utilisée. Nous avons aussi proposé des méthodes pour étendre nos distributions non-paramétriques pour améliorer la prévision des cumuls extrêmes de précipitation.

Il est intéressant de remarquer qu'en terme de pourcentage, le gain observé de toutes nos méthodes non-paramétriques (y compris les plus triviales) reste le même par rapport aux méthodes existantes et ceci aussi bien pour les précipitations que les paramètres étudiés précédemment. En particulier, nous améliorons de façon substantielle la valeur économique des prévisions d'ensemble pour les événements extrêmes de précipitation sexti-horaire.

Durant toute la durée de cette thèse, un soin particulier a été porté pour évaluer de façon la plus informative possible les systèmes de prévision d'ensemble. En effet, la première étude a consisté à introduire de nouvelles mesures quantitatives sur les caractéristiques des histogrammes de rang, notamment l'entropie de ces histogrammes. Cette mesure ainsi qu'un test statistique provenant de Jolliffe and Primo (2008) et étendu par Zamo (2016) ont été conservés dans l'étude sur le post-traitement des précipitations. Il a d'ailleurs été exposé dans cette partie comment discerner la valeur économique des prévisions d'attributs plus généraux comme la résolution en utilisant respectivement les versions empiriques et paramétriques de courbes ROC.

Concernant la vérification des prévisions d'ensemble pour les événements extrêmes, une alternative aux scores pondérés a été trouvée. Cet indice, mêlant le populaire CRPS à la théorie des valeurs extrêmes permet de juger la concordance entre une observation extrême et sa prévision d'ensemble associée. Il partage de nombreuses propriétés avec le CRPS. De plus, il complète les travaux de Ferro (2007); Stephenson et al. (2008); Ferro and Stephenson (2011) sur la verification déterministe des événements extrêmes, et répond à un besoin réel de la part des utilisateurs de la prévision d'ensemble (Zamo, 2016). Sa mise en place relativement simple peut par exemple servir d'aide à la décision dans des cas où l'expertise d'un prévisionniste humain est indispensable.

Les perspectives qu'offre ce travail sont multiples. La première et la plus naturelle est de travailler sur d'autres variables (humidité, nébulosité...) et passer d'un travail sur des stations à un travail en points de grille. Aussi, il convient d'étendre à la toute nouvelle PEAROME les résultats obtenus sur PEARP. Une étude parallèle à cette thèse et menée par nos soins confirme tout le bénéfice des méthodes non-paramétriques par rapport aux méthodes existantes sur ce nouvel ensemble (et bien sûr par rapport à l'ensemble brut). Dans la même veine, l'équipe de mathématiciens du service météorologique néerlandais (Royal Netherlands Meteorological Institute) est en train de tester nos techniques de calibration sur leur modèle baptisé HARMONIE. Les premiers résultats sont excellents et confirment le bien-fondé des méthodes que nous avons développées (Whan and Schmeits, 2017).

La seconde perspective serait, en utilisant des forêts aléatoires, de pouvoir calibrer plusieurs paramètres simultanément (De'Ath, 2002). Cependant, la tendance actuelle est de se concentrer sur des méthodes de post-traitement univarié pour ensuite utiliser des méthodes de reconstruction spatio-temporelle des champs météorologiques après calibration. Soit en

utilisant la structure de l'ensemble brut avant calibration ; c'est ce qu'on appelle l'Ensemble Copula Coupling (ECC) (Bremnes, 2007; Krzysztofowicz and Toth, 2008; Schefzik, 2016; Ben Bouallègue et al., 2016). Soit en utilisant la structure des observations passées, analogue à la structure post-traitée ; ce qu'on appelle le Schaake Shuffle (SS) (Clark et al., 2004). Ce travail d'étude sur l'ECC et le SS est indispensable pour mettre en valeur les champs post-traités et permet ainsi de pouvoir calculer des probabilités d'événements (spatiaux et/ou temporels) conjoints. Ceci présente un intérêt majeur, dans des secteurs météo-sensibles tels que la viabilité hivernale du transport routier (température au sol proche de zéro et précipitations), la prévision des épisodes cévennols (précipitations sur une région donnée), les épisodes caniculaires (températures minimales et maximales supérieures à un seuil pendant plusieurs jours consécutifs) ou encore les travaux publics, dont un exemple significatif est le goudronnage des routes. La problématique du goudronnage des routes est complexe : à partir du moment où l'enrobé bitumineux est préparé, il faut l'utiliser, sinon le chargement d'enrobé est perdu. Pour que l'enrobé puisse être répandu, il faut trois conditions : absence de précipitations, température positive (devant se situer dans un certain intervalle) et vent faible. Dans ce cas présent, l'exploitation optimale de l'information météorologique repose donc sur la probabilité de scenarii multi-paramètres cohérents.

A titre d'exemple, la Figure 4.4 représente une prévision probabiliste de gel issue de PEARP (à gauche), utilisant la méthode de calibration QRF sur 920 stations et une version d'un modèle spatial déjà utilisé à Météo-France (au centre) et l'observation correspondante de gel (à droite). En rouge la probabilité est nulle, elle vaut 1 pour les parties en bleu, et adopte un nuancier pour des valeurs intermédiaires. L'impact de la calibration est ici particulièrement significatif. Nous arrivons à prévoir de nouvelles zones de gel tout en affinant la prévision sur certaines autres zones, notamment à proximité des reliefs.

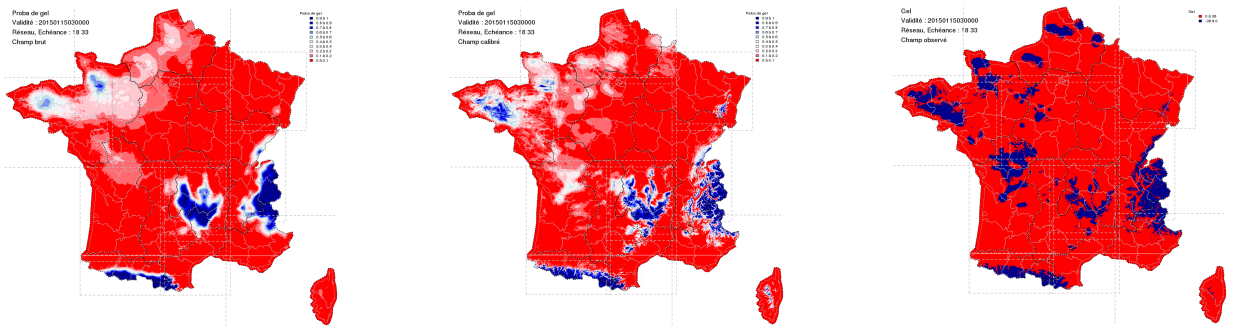

FIGURE 4.4 – Exemple de production spatialisée : prévision probabiliste de gel utilisant PEARP (à gauche), la calibration QRF sur 920 stations avec un modèle de spatialisation (au centre). L'observation correspondante se trouve à droite. On voit sur cet exemple l'impact visuel du post-traitement statistique sur les probabilités issues des membres calibrés.

Une perspective serait finalement de conjuguer différentes méthodes de calibration. Cela peut être fait au sein même d'une méthode "pilote" (Junk et al., 2015). Mais la voie la plus

intéressante semble se trouver du côté de l'agrégation d'experts (Mallet, 2010; Thorey et al., 2016; Zamo, 2016), permettant de prendre en compte toutes les prévisions (post-traitées ou non) afin qu'une pondération adaptative de ces prévisions fournisse une prévision toujours performante quantitativement et cohérente qualitativement.

Bibliographie

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*, pages 199–213. Springer.

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

Athey, S., Tibshirani, J., and Wager, S. (2016). Solving heterogeneous estimating equations with gradient forests. *arXiv preprint arXiv :1610.01271*.

Balkema, A. A. and De Haan, L. (1974). Residual life time at great age. *The Annals of probability*, pages 792–804.

Bao, L., Gneiting, T., Grimit, E. P., Guttorp, P., and Raftery, A. E. (2010). Bias correction and bayesian model averaging for ensemble forecasts of surface wind direction. *Monthly Weather Review*, 138(5) :1811–1821.

Baran, S. and Lerch, S. (2015). Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Quarterly Journal of the Royal Meteorological Society*, 141(691) :2289–2299.

Baran, S. and Lerch, S. (2016). Mixture emos model for calibrating ensemble forecasts of wind speed. *Environmetrics*.

Baran, S. and Nemoda, D. (2016). Censored and shifted gamma distribution based emos model for probabilistic quantitative precipitation forecasting. *Environmetrics*, 27(5) :280–292.

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

Ben Bouallègue, Z. (2013). Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather and Forecasting*, 28(2) :515–524.

Ben Bouallègue, Z., Heppelmann, T., Theis, S. E., and Pinson, P. (2016). Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach. *Monthly Weather Review*, 144(12) :4737–4750.

Benamran, B. (2017). Music was better in the old days - quickie 16 - e-penser. `https://www.youtube.com/watch?v=NiPSK4sOQ3s/`. [Online ; accessed 28-august-2017].

Bjerknes, V. (1904). *Das Problem der Wettervorhersage : betrachtet vom Standpunkte der Mechanik und der Physik*.

Bouttier, F., Raynaud, L., Nuissier, O., and Ménétrier, B. (2016). Sensitivity of the arome ensemble to initial and surface perturbations during hymex. *Quarterly Journal of the Royal Meteorological Society*, 142(S1) :390–403.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2) :123–140.

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

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Bremnes, J. (2004). Probabilistic forecasts of precipitation in terms of quantiles using nwp model output. *Monthly Weather Review*, 132(1) :338–347.

Bremnes, J. (2007). Improved calibration of precipitation forecasts using ensemble techniques. part 2 : Statistical calibration methods. met. Technical report, no research report 04.

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

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

Bröcker, J. (2015). Resolution and discrimination–two sides of the same coin. *Quarterly Journal of the Royal Meteorological Society*, 141(689) :1277–1282.

Bröcker, J. and Smith, L. A. (2007a). Increasing the reliability of reliability diagrams. *Weather and forecasting*, 22(3) :651–661.

Bröcker, J. and Smith, L. A. (2007b). Scoring probabilistic forecasts : The importance of being proper. *Weather and Forecasting*, 22(2) :382–388.

Buizza, R. (2001). Accuracy and potential economic value of categorical and probabilistic forecasts of discrete events. *Monthly Weather Review*, 129(9) :2329–2345.

Buizza, R., Milleer, M., and Palmer, T. (1999). Stochastic representation of model uncertainties in the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125(560) :2887–2908.

Calanca, P., Bolius, D., Weigel, A., and Liniger, M. (2011). Application of long-range weather forecasts to agricultural decision problems in europe. *The Journal of Agricultural Science*, 149(1) :15–22.

Chen, J., Li, K., Tang, Z., Bilal, K., Yu, S., Weng, C., and Li, K. (2017). A parallel random forest algorithm for big data in a spark cloud computing environment. *IEEE Transactions on Parallel and Distributed Systems*, 28(4) :919–933.

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

Cooke, E. (1906). Forecasts and verifications in western australia. *Monthly Weather Review*, 34(1) :23–24.

Courtier, P., Freydier, C., Geleyn, J., Rabier, F., and Rochas, M. (1991). The arpege project at meteo-france. In *ECMWF seminar proceedings*, volume 2, pages 193–231.

Cramér, H. (1928). On the composition of elementary errors : First paper : Mathematical deductions. *Scandinavian Actuarial Journal*, 1928(1) :13–74.

Dabernig, M., Mayr, G. J., Messner, J. W., and Zeileis, A. (2016). Spatial ensemble post-processing with standardized anomalies. *Quarterly Journal of the Royal Meteorological Society*.

Dawid, A. P. (1984). Present position and potential developments : Some personal views : Statistical theory : The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, pages 278–292.

Dawid, A. P. and Sebastiani, P. (1999). Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, pages 65–81.

De Fondeville, R. (2014). Scoring and multivariate extremes : Assessing climate forecast of extremes. Technical report, Colorado State University - LSCE/CNRS.

De Haan, L. and Ferreira, A. (2007). *Extreme value theory : an introduction*. Springer Science & Business Media.

De Haan, L. F. M. (1970). On regular variation and its application to the weak convergence of sample extremes.

De'Ath, G. (2002). Multivariate regression trees : a new technique for modeling species-environment relationships. *Ecology*, 83(4) :1105–1117.

DeGroot, M. H. and Fienberg, S. E. (1983). The comparison and evaluation of forecasters. *The statistician*, pages 12–22.

Delle Monache, L., Eckel, F. A., Rife, D. L., Nagarajan, B., and Searight, K. (2013). Probabilistic weather prediction with an analog ensemble. *Monthly Weather Review*, 141(10) :3498–3516.

Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X., and Stull, R. B. (2006). Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *Journal of Geophysical Research : Atmospheres (1984–2012)*, 111(D24).

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., and Cébron, P. (2015). Pearp, the météo-france short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 141(690) :1671–1685.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1997). Evaluating density forecasts.

Diks, C., Panchenko, V., and Van Dijk, D. (2011). Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics*, 163(2) :215–230.

Ehm, W., Gneiting, T., Jordan, A., and Krüger, F. (2016). Of quantiles and expectiles : consistent scoring functions, choquet representations and forecast rankings. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 78(3) :505–562.

Embrechts, P., Klüppelberg, C., and Mikosch, T. (1997). *Modelling extremal events, volume 33 of Applications of Mathematics*. New York. Springer-Verlag, Berlin.

Epstein, E. S. (1969a). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6) :985–987.

Epstein, E. S. (1969b). Stochastic dynamic prediction. *Tellus*, 21(6) :739–759.

Feldmann, K., Scheuerer, M., and Thorarinsdottir, T. L. (2014). Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous gaussian regression. *arXiv preprint arXiv :1407.0058*.

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

Ferro, C. A. (2007). A probability model for verifying deterministic forecasts of extreme events. *Weather and Forecasting*, 22(5) :1089–1100.

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

Ferro, C. A. and Stephenson, D. B. (2011). Extremal dependence indices : Improved verification measures for deterministic forecasts of rare binary events. *Weather and Forecasting*, 26(5) :699–713.

Fisher, R. A. and Tippett, L. H. C. (1928). Limiting forms of the frequency distribution of the largest or smallest member of a sample. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 24, pages 180–190. Cambridge University Press.

Fréchet, M. (1927). Sur la loi de probabilité de l'écart maximum. In *Annales de la societe Polonaise de Mathematique*. [t. VI, p. 93].

Friederichs, P. (2010). Statistical downscaling of extreme precipitation events using extreme value theory. *Extremes*, 13(2) :109–132.

Friederichs, P., Göber, M., Bentzien, S., Lenz, A., and Krampitz, R. (2009). A probabilistic analysis of wind gusts using extreme value statistics. *Meteorologische Zeitschrift*, 18(6) :615–629.

Friederichs, P. and Hense, A. (2007). Statistical downscaling of extreme precipitation events using censored quantile regression. *Monthly weather review*, 135(6) :2365–2378.

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

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14) :2225–2236.

Genuer, R., Poggi, J.-M., Tuleau-Malot, C., and Villa-Vialaneix, N. (2017). Random forests for big data. *Big Data Research*, 9 :28 – 46.

Ghosh, S. and Resnick, S. (2010). A discussion on mean excess plots. *Stochastic Processes and their Applications*, 120(8) :1492–1517.

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

Gnedenko, B. (1943). On the theory of domains of attraction of stable laws. *Uchenye Zapiski Moskov. Gos. Univ. Matematika*, 30 :61–81.

Gneiting, T. (2014). *Calibration of medium-range weather forecasts*. European Centre for Medium-Range Weather Forecasts.

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

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1 :125–151.

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

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

Gneiting, T. and Ranjan, R. (2011). Comparing density forecasts using threshold-and quantile-weighted scoring rules. *Journal of Business & Economic Statistics*, 29(3) :411–422.

Grimit, E. P., Gneiting, T., Berrocal, V., and Johnson, N. A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Technical report, DTIC Document.

Hagedorn, R. (2017). Slowly but surely : Observing and supporting the growing use of ensemble forecasts. In *Annual Seminar, ECMWF 2017, Reading, United Kingdom*, pages http ://www.ecmwf.int/sites/default/files/elibrary/2017/17625–slowly–surely–observing–and–supporting–growing–use–ensemble–products.pdf.

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

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

Hamill, T. M. and Colucci, S. J. (1997). Verification of eta-rsm short-range ensemble forecasts. *Monthly Weather Review*, 125(6) :1312–1327.

Hamill, T. M., Hagedorn, R., and Whitaker, J. S. (2008). Probabilistic forecast calibration using ecmwf and gfs ensemble reforecasts. part ii : Precipitation. *Monthly weather review*, 136(7) :2620–2632.

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

Hand, D. J. (2009). Measuring classifier performance : a coherent alternative to the area under the roc curve. *Machine learning*, 77(1) :103–123.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning : second edition*. Springer.

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

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

Horton, P., Jaboyedoff, M., and Obled, C. (2017). Global optimization of an analog method by means of genetic algorithms. *Monthly Weather Review*, 145(4) :1275–1294.

Hosking, J. R. and Wallis, J. R. (1987). Parameter and quantile estimation for the generalized pareto distribution. *Technometrics*, 29(3) :339–349.

Hosking, J. R. M. (1989). *Some theoretical results concerning L-moments*. IBM Thomas J. Watson Research Division.

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

Jolliffe, I. T. and Stephenson, D. B. (2012). *Forecast verification : a practitioner's guide in atmospheric science*. John Wiley & Sons.

Jordan, A., Krueger, F., and Lerch, S. (2017). *scoringRules : Scoring Rules for Parametric and Simulated Distribution Forecasts*. R package version 0.9.3.

Juban, J., Fugon, L., and Kariniotakis, G. (2007). Probabilistic short-term wind power forecasting based on kernel density estimators. In *European Wind Energy Conference and exhibition, EWEC 2007, MILAN, Italy*, pages http :/ewec2007proceedings.info hal–00526011.

Junk, C., Delle Monache, L., and Alessandrini, S. (2015). Analog-based ensemble model output statistics. *Monthly Weather Review*, 143(7) :2909–2917.

Kahneman, D. and Tversky, A. (1979). Prospect theory : An analysis of decision under risk. *Econometrica : Journal of the econometric society*, pages 263–291.

Katz, R. W., Parlange, M. B., and Naveau, P. (2002). Statistics of extremes in hydrology. *Advances in water resources*, 25(8) :1287–1304.

Keller, J. D., Delle Monache, L., and Alessandrini, S. (2017). Statistical downscaling of a high-resolution precipitation reanalysis using the analog ensemble method. *Journal of Applied Meteorology and Climatology*, (2017).

Koenker, R. and Bassett Jr, G. (1978). Regression quantiles. *Econometrica : journal of the Econometric Society*, pages 33–50.

Krzysztofowicz, R. (2001). The case for probabilistic forecasting in hydrology. *Journal of hydrology*, 249(1) :2–9.

Krzysztofowicz, R. and Toth, Z. (2008). Bayesian processor of ensemble (bpe) : Concept and implementation. In *Slides presented at the 4th NCEP/NWS Ensemble User Workshop, Laurel, MD*.

Laplace, P. S. (1814). *Essai philosophique sur les probabilités*. Bachelier.

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., Gneiting, T., et al. (2017). The forecaster's dilemma : extreme events and forecast evaluation. *Statistical Science*, 32(1) :106–127.

Lobo, J. M., Jiménez-Valverde, A., and Real, R. (2008). Auc : a misleading measure of the performance of predictive distribution models. *Global ecology and Biogeography*, 17(2) :145–151.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the atmospheric sciences*, 20(2) :130–141.

Lorenz, E. N. (1967). *The nature and theory of the general circulation of the atmosphere*, volume 218. World Meteorological Organization Geneva.

Mallet, V. (2010). Ensemble forecast of analyses : Coupling data assimilation and sequential aggregation. *Journal of Geophysical Research : Atmospheres*, 115(D24).

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

Manzato, A. (2007). A note on the maximum peirce skill score. *Weather and Forecasting*, 22(5) :1148–1154.

Matheson, J. E. and Winkler, R. L. (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.

Messner, J. W., Mayr, G. J., and Zeileis, A. (2017). Nonhomogeneous boosting for predictor selection in ensemble postprocessing. *Monthly Weather Review*, 145(1) :137–147.

Morel, C. (2014). *Les décisions absurdes*, volume 1. Editions Gallimard.

Mureau, R., Molteni, F., and Palmer, T. (1993). Ensemble prediction using dynamically conditioned perturbations. *Quarterly Journal of the Royal Meteorological Society*, 119(510) :299–323.

Murphy, A. H. (1993). What is a good forecast ? an essay on the nature of goodness in weather forecasting. *Weather and forecasting*, 8(2) :281–293.

Naveau, P., Huser, R., Ribereau, P., and Hannart, A. (2016). Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, 52(4) :2753–2769.

Palmer, T. (2002). The economic value of ensemble forecasts as a tool for risk assessment : From days to decades. *Quarterly Journal of the Royal Meteorological Society*, 128(581) :747–774.

Papastathopoulos, I. and Tawn, J. A. (2013). Extended generalised pareto models for tail estimation. *Journal of Statistical Planning and Inference*, 143(1) :131–143.

Patton, A. J. (2015). Comparing possibly misspecified forecasts. Working paper, Duke University. Available at `http://public.econ.duke.edu/~ap172/Patton_bregman_comparison_22dec16.pdf`.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *the Annals of Statistics*, pages 119–131.

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

Pinson, P., Chevallier, C., and Kariniotakis, G. N. (2007). Trading wind generation from short-term probabilistic forecasts of wind power. *IEEE Transactions on Power Systems*, 22(3) :1148–1156.

Pinson, P., Madsen, H., Nielsen, H. A., Papaefthymiou, G., and Klöckl, B. (2009). From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind energy*, 12(1) :51–62.

Poulos, H. M., Chernoff, B., Fuller, P. L., and Butman, D. (2012). Ensemble forecasting of potential habitat for three invasive fishes. *Aquatic Invasions*, 7(2).

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

Ravazzolo, F. and Vahey, S. P. (2014). Forecast densities for economic aggregates from disaggregate ensembles. *Studies in Nonlinear Dynamics & Econometrics*, 18(4) :367–381.

Raynaud, L. and Bouttier, F. (2016). Comparison of initial perturbation methods for ensemble prediction at convective scale. *Quarterly Journal of the Royal Meteorological Society*, 142(695) :854–866.

Resnick, S. I. (1971). Tail equivalence and its applications. *Journal of Applied Probability*, 8(1) :136–156.

Richardson, D. S. (2000). Skill and relative economic value of the ecmwf ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126(563) :649–667.

Roulin, E. and Vannitsem, S. (2012). Postprocessing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Monthly weather review*, 140(3) :874–888.

Roulston, M. S. and Smith, L. A. (2002). Evaluating probabilistic forecasts using information theory. *Monthly Weather Review*, 130(6) :1653–1660.

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

Schaake, J. C., Hamill, T. M., Buizza, R., and Clark, M. (2007). Hepex : the hydrological ensemble prediction experiment. *Bulletin of the American Meteorological Society*, 88(10) :1541–1547.

Schefzik, R. (2016). Combining parametric low-dimensional ensemble postprocessing with reordering methods. *Quarterly Journal of the Royal Meteorological Society*, 142(699) :2463–2477.

Schefzik, R., Thorarinsdottir, T. L., Gneiting, T., et al. (2013). Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science*, 28(4) :616–640.

Schervish, M. J., Seidenfeld, T., and Kadane, J. B. (2009). Proper scoring rules, dominated forecasts, and coherence. *Decision Analysis*, 6(4) :202–221.

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quarterly Journal of the Royal Meteorological Society*, 140(680) :1086–1096.

Scheuerer, M. and Hamill, T. M. (2015). Statistical postprocessing of ensemble precipitation forecasts by fitting censored, shifted gamma distributions. *Monthly Weather Review*, 143(11) :4578–4596.

Scheuerer, M., Möller, D., et al. (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. and Kok, K. J. (2010). A comparison between raw ensemble output,(modified) bayesian model averaging, and extended logistic regression using ecmwf ensemble precipitation reforecasts. *Monthly Weather Review*, 138(11) :4199–4211.

Schuhen, N., Thorarinsdottir, T. L., and Gneiting, T. (2012). Ensemble model output statistics for wind vectors. *Monthly Weather Review*, 140(10) :3204–3219.

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

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

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., and Fraley, C. (2007). Probabilistic quantitative precipitation forecasting using bayesian model averaging. *Monthly Weather Review*, 135(9) :3209–3220.

Stephenson, D., Casati, B., Ferro, C., and Wilson, C. (2008). The extreme dependency score : a non-vanishing measure for forecasts of rare events. *Meteorological Applications*, 15(1) :41–50.

Taillardat, M., Mestre, O., Zamo, M., and Naveau, P. (2016). Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Monthly Weather Review*, 144(6) :2375–2393.

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

Taylor, J. W. and Buizza, R. (2003). Using weather ensemble predictions in electricity demand forecasting. *International Journal of Forecasting*, 19(1) :57–70.

Thorarinsdottir, T. L. and Gneiting, T. (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., Mallet, V., and Baudin, P. (2016). Online learning with the continuous ranked probability score for ensemble forecasting. *Quarterly Journal of the Royal Meteorological Society*.

Toth, Z. and Kalnay, E. (1993). Ensemble forecasting at nmc : The generation of perturbations. *Bulletin of the american meteorological society*, 74(12) :2317–2330.

Tribus, M. (1969). *Rational Descriptions, Decisions and Designs*. Pergamon Press, Elmsford, New York.

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts : proper scoring rules and moments. [Online ; Available at SSRN : `http://ssrn.com/abstract=2236605`].

Van den Dool, H. (1994). Searching for analogues, how long must we wait ? *Tellus A*, 46(3) :314–324.

Van Schaeybroeck, B. and Vannitsem, S. (2015). Ensemble post-processing using member-by-member approaches : theoretical aspects. *Quarterly Journal of the Royal Meteorological Society*, 141(688) :807–818.

Von Mises, R. (1928). Statistik und wahrheit. *Julius Springer*.

Von Mises, R. (1936). La distribution de la plus grande de n valeurs. *Rev. math. Union interbalcanique*, 1(1).

Vrac, M. and Naveau, P. (2007). Stochastic downscaling of precipitation : From dry events to heavy rainfalls. *Water resources research*, 43(7).

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

Whan, K. and Schmeits, M. (2017). Probabilistic forecasts of extreme local precipitation using harmonie predictors and comparing 3 different post-processing methods. In *EGU General Assembly Conference Abstracts*, volume 19, page 5596.

Wilks, D. (2001). A skill score based on economic value for probability forecasts. *Meteorological Applications*, 8(2) :209–219.

Wilks, D. S. (1995). *Statistical methods in the atmospheric sciences*. Academic press.

Wilks, D. S. (2009). Extending logistic regression to provide full-probability-distribution mos forecasts. *Meteorological Applications*, 16(3) :361–368.

Williams, R., Ferro, C., and Kwasniok, F. (2014). A comparison of ensemble post-processing methods for extreme events. *Quarterly Journal of the Royal Meteorological Society*, 140(680) :1112–1120.

Wu, X., Zhu, X., Wu, G.-Q., and Ding, W. (2014). Data mining with big data. *IEEE transactions on knowledge and data engineering*, 26(1) :97–107.

Zamo, M. (2016). *Statistical Post-processing of Deterministic and Ensemble Windspeed Forecasts on a Grid*. PhD thesis, Université Paris-Saclay.

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

Zamo, M., Mestre, O., Arbogast, P., and Pannekoucke, O. (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., Mestre, O., Arbogast, P., and Pannekoucke, O. (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, B. and Zhai, P. (2016). A new forecast model based on the analog method for persistent extreme precipitation. *Weather and Forecasting*, 31(4) :1325–1341.

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

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences de l'environnement d'île-de-France (SEIF)

Titre : Méthodes Non-Paramétriques de Post-Traitement des Prévisions d'Ensemble

Mots clefs : Météorologie, régression quantile, forêts aléatoires, événements extrêmes, vérification.

Résumé : En prévision numérique du temps, les modèles de prévision d'ensemble sont devenus un outil incontournable pour quantifier l'incertitude des prévisions et fournir des prévisions probabilistes. Malheureusement, ces modèles ne sont pas parfaits et une correction simultanée de leur biais et de leur dispersion est nécessaire.

Cette thèse présente de nouvelles méthodes de post-traitement statistique des prévisions d'ensemble. Celles-ci ont pour particularité d'être basées sur les forêts aléatoires.
Contrairement à la plupart des techniques usuelles, ces méthodes non-paramétriques permettent de prendre en compte la dynamique non-linéaire de l'atmosphère. Elles permettent aussi d'ajouter des covariables (autres variables météorologiques, variables temporelles, géographiques...) facilement et sélectionnent elles-mêmes les prédicteurs les plus utiles dans la régression. De plus, nous ne faisons aucune hypothèse sur la distribution de la variable à traiter. Cette nouvelle approche surpasse les méthodes existantes pour des variables telles que la température et la vitesse du vent.

Pour des variables reconnues comme difficiles à calibrer, telles que les précipitations sexti-horaires, des versions hybrides de nos techniques ont été créées. Nous montrons que ces versions hybrides (ainsi que nos versions originales) sont meilleures que les méthodes existantes. Elles amènent notamment une véritable valeur ajoutée pour les pluies extrêmes.

La dernière partie de cette thèse concerne l'évaluation des prévisions d'ensemble pour les événements extrêmes. Nous avons montré quelques propriétés concernant le Continuous Ranked Probability Score (CRPS) pour les valeurs extrêmes. Nous avons aussi défini une nouvelle mesure combinant le CRPS et la théorie des valeurs extrêmes, dont nous examinons la cohérence sur une simulation ainsi que dans un cadre opérationnel.

Les résultats de ce travail sont destinés à être insérés au sein de la chaîne de prévision et de vérification à Météo-France.

Title : Non-parametric Methods of Post-processing for Ensemble Forecasting

Keywords : Meteorology, quantile regression, random forests, extreme events, verification.

Abstract : In numerical weather prediction, ensemble forecasts systems have become an essential tool to quantify forecast uncertainty and to provide probabilistic forecasts. Unfortunately, these models are not perfect and a simultaneous correction of their bias and their dispersion is needed.

This thesis presents new statistical post-processing methods for ensemble forecasting. These are based on random forests algorithms, which are non-parametric.
Contrary to state of the art procedures, random forests can take into account non-linear features of atmospheric states. They easily allow the addition of covariables (such as other weather variables, seasonal or geographic predictors) by a self-selection of the most useful predictors for the regression. Moreover, we do not make assumptions on the distribution of the variable of interest. This new approach outperforms the existing methods for variables such as surface temperature and wind speed.

For variables well-known to be tricky to calibrate, such as six-hours accumulated rainfall, hybrid versions of our techniques have been created. We show that these versions (and our original methods) are better than existing ones. Especially, they provide added value for extreme precipitations.

The last part of this thesis deals with the verification of ensemble forecasts for extreme events. We have shown several properties of the Continuous Ranked Probability Score (CRPS) for extreme values. We have also defined a new index combining the CRPS and the extreme value theory, whose consistency is investigated on both simulations and real cases.

The contributions of this work are intended to be inserted into the forecasting and verification chain at Météo-France.

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