

Statistical Post-Processing Methods And Their Implementation On The Ensemble Prediction Systems For Forecasting Temperature In The Use Of The French Electric Consumption

Adriana Geanina Cucu Gogonel

▶ **To cite this version:**

Adriana Geanina Cucu Gogonel. Statistical Post-Processing Methods And Their Implementation On The Ensemble Prediction Systems For Forecasting Temperature In The Use Of The French Electric Consumption. General Mathematics [math.GM]. Université René Descartes - Paris V, 2012. English. NNT : 2012PA05S014 . tel-00798576

HAL Id: tel-00798576

https://theses.hal.science/tel-00798576

Submitted on 8 Mar 2013

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

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

Université Paris Descartes

Laboratoire MAP 5

École Doctorale de Sciences Mathématiques de Paris Centre

THÈSE
pour obtenir le grade de

Docteur en Mathématiques Appliquées
de l'Université Paris Descartes

Spécialité : STATISTIQUE

Présentée par **Adriana CUCU GOGONEL**

Statistical Post-Processing Methods And Their Implementation On The Ensemble Prediction Systems For Forecasting Temperature In The Use Of The French Electric Consumption.

Thèse dirigée par
Avner BAR-HEN

soutenue publiquement le 27 novembre 2012

Jury

Rapporteurs	Jerome Saracco, Institut Polytechnique de Bordeaux
	Jocelyn Gaudet, Hydro-Québec
Examinatrice	Virginie Dordonnat, EDF R&D
Examinateur	Eric Parent, AgroParisTech/INRA
Directeurs de thèse	Avner Bar-Hen, Université Paris Descartes
	Jérôme Collet, EDF R&D

Remerciements

Je voudrais en premier lieu remercier les personnes qui ont rendu possible l'existence de cette thèse, je pense à mon directeur de thèse, Avner Bar-Hen et à mon encadrant EDF, Jérôme Collet. Je remercie Avner d'avoir bien voulu assumer la direction de cette thèse. Son soutien, sa bonne humeur, sa disponibilité et son exigence ont été des éléments indispensables à la réussite de ce travail. Je remercie Jérôme qui m'a fait découvrir la recherche et m'a donné le gout du SAS. Il a toujours su se rendre disponible et me faire bénéficier de son expérience tout en me laissant une grande liberté d'action.

Je tiens à exprimer ma gratitude aux personnes qui m'ont fait l'honneur de participer au jury de cette thèse. Je suis reconnaissante envers Jérôme Saracco et Jocelyn Gaudet pour l'intérêt qu'ils ont porté à cette thèse en acceptant d'en être les rapporteurs, leurs rapports m'ont permis d'améliorer ce manuscrit. Je suis particulièrement reconnaissante à Virginie Dordonnat, pour ses conseils précieux dans tous les domaines : technique, métier, valorisation du travail, écriture du rapport ainsi que pour son soutien continu qui a été d'autant plus précieux dans mes moments de doutes. Je tiens à remercier également Eric Parent de m'avoir fait part de ses idées sur mes travaux, ce qui m'a ouvert des perspectives pour la suite de la thèse.

Je remercie mes collègues Julien Najac, Laurent Dubus, Thi-Thu Huong-Hoang, Christophe Chaussin et Marc Voirin d'avoir participé aux réunions de suivi de ma thèse. Ces réunions ont toujours été d'une aide importante dans l'évolution des travaux de thèse. Je remercie Thi-Thu pour son aide sur la partie des valeurs extrêmes de la thèse ainsi que pour sa gentillesse et sa disponibilité.

Je tiens ensuite à remercier Sandrine Charousset et François Regis-Monclar qui m'ont permis de démarrer cette thèse à EDF R&D. Je remercie également Bertrand Vignal et plus particulièrement Marc Voirin qui ont supervisé le bon déroulement de ma thèse ainsi que mon intégration dans l'équipe R32 du département OSIRIS. Ils m'ont encouragée tout en mettant à ma disposition des conditions de travail idéales. Je remercie également Laetitia Hubert et Christelle Roger qui m'ont soutenue dans mes démarches administratives.

J'ai une pensée pour tous les membres du département OSIRIS que j'ai eu le plaisir de côtoyer durant ces quatre années de thèse ainsi qu'au cours de mon stage de Master 2. Une pensée plus particulière va à Xavier Brossat qui m'a fait découvrir la beauté des conférences et le charme de Prague et à qui j'ai fait découvrir le "talent de buveuse" des filles de l'est !

Je remercie mes collègues de bureau pendant les quatre années, par ordre chronologique : Laurence pour le plein de conseils sur le fonctionnement de la R&D et sur la vie d'une maman-chercheur, Codé pour sa bonne humeur permanente et pour ses chansonnettes, Nicolas pour les discussions sport et pas que et Kuon pour sa sagesse et son optimisme, à mon égard.

Je tiens à remercier mes collègues de la "tant aimée" équipe R32, pour leur bonne humeur, pour leur soutien. Ils m'ont appris tellement de choses sur des domaines différents, ma culture générale s'est améliorée à leur coté et ils m'ont sûrement donné envie d'être un bon ingénieur. Je vais avoir du mal à retrouver ailleurs un autre Dominique, toujours à l'écoute, discret mais toujours présent avec un bon conseil. Un remerciement particulier à Sabine pour son soutien, toujours prête à m'aider et à m'écouter, j'aurais aimé savoir en profiter plus. Je voudrais également remercier Bogdan, mon cher collègue roumain, un mélange parfait de professionnalisme et de joie. Je remercie Jérôme et Jérémy de m'inciter puis m'écouter parler des exploits sportifs de mon époux. Je tiens à remercier Sébastien, ses conseils sur le timming dans l'écriture de rapport m'ont été d'un réel aide, mais c'est quand même son humour noir qui l'emporte dans mes souvenirs. Je voudrais également remercier Aline, nos échanges lors de mon arrivée à EDF resterons toujours dans ma mémoire. J'ai une pensée également pour Muriel qui m'a apporté un vrai soutien à mon retour de congé de maternité. Je remercie Florent pour nos discussions détendues mais tout aussi intéressantes et puis pour les discussions "bébé". Plus particulièrement je voudrais remercier Pascale, un exemple d'élégance et de professionnalisme, et un mélange très réussi de sérieux et de folie. Je remercie également Mathilde, Thi-Thu et Virginie pour leurs encouragements et leurs conseils d'autant plus précieux qu'ils viennent de leurs récentes expériences de doctorantes. Je voudrais remercier Avner, Jérôme, Virginie, Sabine, Xavier, Olivier, Thi-Thu et Lili pour leur aide dans l'amélioration de mes transaparents de soutenance et du discours qui les accompagne, j'espère qu'ils seront contents du résultat!

Cette section ne serait pas complète si je ne remerciais pas mes amies de l'université, Lian, Nathalie et Souad à côté desquelles j'ai découvert les statistiques. Elles sont sûrement plus capables que moi de faire une thèse. Bien que ce soit moi qui l'ai faite finalement c'est comme si cette thèse était un peu la leur. Je profite pour remercier nos professeurs de Paris 5, pour les connaissances qu'ils ont su si bien nous transmettre et pour leur accueil chaleureux lors de mon arrivée en thèse.

Je tiens à exprimer mes remerciements aux amis du kyokushin qui m'ont aidée, par la folie de leur jeunesse, à m'échapper de temps en temps aux doutes de la thèse. Je remercie Djéma pour son humour contagieux, Guillaume pour ses compliments culinaires et Yacine pour son éternelle gentillesse !

Je voudrais remercier mes amis roumains, même s'ils sont loin, ils m'ont soutenue et encouragée tout au long de cette période de thèse. Je remercie Marian C pour ses conseils judicieux, toujours bien intentionnés, Aura S pour sa façon unique de me rassurer et Bianca C de m'avoir partagé son expérience de gestion d'une thèse en tant que maman. Je remercie également Mihaela B pour ses conseils avisés et son attitude sans retenue envers une ancienne étudiante. Je tiens à remercier mon professeur de mathématique au collège, Ion Lixandru, de m'avoir fait aimer les maths et de m'avoir donné confiance en moi. Je remercie Veronica -qui m'amène ici un peu de Roumanie- pour son énergie débordante, que j'aimerais tant lui piquer !

Mes derniers remerciements vont vers ma famille. Je remercie mes parents de la richesse de ce qu'ils ont su me transmettre, c'est sans doute la base de toutes les réussites de ma vie. Je remercie Lili, ma s œur et meilleure amie, de n'avoir jamais cessé de croire en moi et de m'avoir porté bon conseil dans toutes les décisions que j'ai dû prendre dans ma vie et dans mes études en particulier. Sans son aide ma vie aurait été sûrement moins simple et le chemin vers le bonheur beaucoup plus long. Je tiens à remercier Sam pour nos échanges enrichissants dans de si nombreux domaines. J'ai toujours pu m'appuyer sur son épaule et ses conseils m'ont souvent inspiré. Je remercie également Claude pour sa présence apaisante qui m'a toujours donné confiance. Une pensée émue va à mon neveu, Mathis et à ma nièce Julia, nés pendant mes années de thèse, qui j'espère liront un jour avec intérêt ce manuscrit !

Un remerciement particulier à Lucian, mon cher champion et le grand sage de la famille. Son soutien sans faille et sa patience pendant ces années de thèse m'ont été d'une aide précieuse. Il a su m'encourager, me donner envie d'avancer et me conseiller des breaks quand mon état le montrait. Je lui remercie aussi pour un cadeau très spécial en deuxième année de thèse, notre fils, Thomas, que je remercie de m'avoir fait découvrir la douce maternité en même temps que la recherche.

À Lili et à Lucian

Résumé

L'objectif des travaux de la thèse est d'étudier les propriétés statistiques de correction des prévisions de température et de les appliquer au système des prévisions d'ensemble (SPE) de Météo France. Ce SPE est utilisé dans la gestion du système électrique, à EDF R&D, il contient 51 membres (prévisions par pas de temps) et fournit des prévisions à 14 jours. La thèse comporte trois parties. Dans la première partie on présente les SPE, dont le principe est de faire tourner plusieurs scénarios du même modèle avec des données d'entrée légèrement différentes pour simuler l'incertitude. On propose après des méthodes statistiques (la méthode du meilleur membre et la méthode bayésienne) que l'on implémente pour améliorer la précision ou la fiabilité du SPE dont nous disposons et nous mettons en place des critères de comparaison des résultats. Dans la deuxième partie nous présentons la théorie des valeurs extrêmes et les modèles de mélange et nous proposons des modèles de mélange contenant le modèle présenté dans la première partie et des fonctions de distributions des extrêmes. Dans la troisième partie nous introduisons la régression quantile pour mieux estimer les queues de distribution.

Mots clés: Prévisions de température, systèmes de prévisions d'ensemble, méthode du meilleur membre, critères de validation des SPE, théorie des valeurs extremes, modèles de mélange, regréssion quantile.

The thesis has for objective to study new statistical methods to correct temperature predictions that may be implemented on the ensemble prediction system (EPS) of Meteo France so to improve its use for the electric system management, at EDF France. The EPS of Meteo France we are working on contains 51 members (forecasts by time-step) and gives the temperature predictions for 14 days. The thesis contains three parts: in the first one we present the EPS and we implement two statistical methods improving the accuracy or the spread of the EPS and we introduce criteria for comparing results. In the second part we introduce the extreme value theory and the mixture models we use to combine the model we build in the first part with models for fitting the distributions tails. In the third part we introduce the quantile regression as another way of studying the tails of the distribution.

Keywords: Temperature forecasts, ensemble prediction systems, best member method, EPS validation criteria, extreme value theory, mixture models, quantile regression.

Contents

Version abrégée 11

Summary 15

Publications and Conferences 19

1 Introduction 21
- 1.1 Context and Data Description . . . 22
  - 1.1.1 Context . . . 22
  - 1.1.2 Data Description . . . 24
- 1.2 Ensemble prediction systems (EPS) . . . 28
  - 1.2.1 History and building methods . . . 28
  - 1.2.2 Forecasts and uncertainty in meteorology . . . 29
    - Sources of forecast errors . . . 30
- 1.3 The verification methods for EPS . . . 31
  - 1.3.1 Standard Statistical Measure . . . 31
    - Bias . . . 31
    - Correlation coefficient . . . 31
    - Mean absolute error (MAE) . . . 32
    - The root mean square error (RMSE). . . . 32
  - 1.3.2 Reliability Criteria . . . 32
    - Talagrand diagram . . . 32
    - Probability integral transform (PIT) . . . 33
    - Reliability Diagram . . . 33
  - 1.3.3 Resolution (sharpness) Criteria . . . 34
    - Brier Score . . . 34
    - Continuous Rank Probability Score (CRPS) . . . 34
    - Ignorance Score . . . 35
    - ROC Curve . . . 35
- 1.4 Post-processing methods . . . 36
  - 1.4.1 The best member method . . . 36
    - The un-weighted members method . . . 36

The weighted members method . . . 37
1.4.2 Bayesian model averaging . . . 38

2 Implementation of two Statistic Methods of Ensemble Prediction Systems for Electric System Management (CSBIGS Article) 41

3 Mixture Models in Extreme Values Theory 61
3.1 Extreme Value Theory . . . 62
3.1.1 Peaks Over Thresholds . . . 64
Choice of the threshold . . . 65
Dependence above threshold . . . 66
3.2 Mixture models . . . 67
3.2.1 Parameter estimation . . . 68
Method of Moments . . . 68
Bayes Estimates . . . 68
Maximum Likelihood Estimation . . . 69
The Expectation Maximization algorithm . . . 69
3.3 Mixture models in the Extreme Value Theory . . . 71
3.4 The proposed extreme mixture model . . . 72
3.5 Implementation of the Extreme Value Theory on Temperature Forecasts Data . . . 74
3.5.1 Context and Data . . . 74
3.5.2 Choices of extreme parameter values . . . 75
3.5.3 Mixture model and criteria of comparison of the final distributions . . . 80
Mixture Models with the GEV parameters estimated by tail and by season 80
Mixture Models with the EVT parameters estimated by tail and for the right tail, by time-horizon. . . . 81
Mixture Models with the GEV parameters estimated by tail, by month and by package of time-horizon. . . . 95
Discussion extreme mixture models . . . 96

4 Improvement of Short-Term Extreme Temperature Density Forecasting using Best Member Method (NHESS Article) 99

Conclusion 105

Appendix 109
Case of the Extreme Values given by the 1st and 51st rank . . . 123

List of Figures 128

List of Tables 131

Version abrégée

Ces travaux sont réalisés dans le cadre d'une thèse CIFRE entre l'Université Paris Descartes (Laboratoire MAP5) et le département OSIRIS de la R&D d'EDF.

L'objectif des travaux de la thèse est d'étudier les propriétés statistiques de correction des prévisions de température et de les appliquer au système des prévisions d'ensemble (SPE) de Météo France pour améliorer son utilisation pour la gestion des systèmes électriques, à EDF R&D. Le SPE de Météo France que nous utilisons contient 51 membres (prévisions par pas de temps) et fournit des prévisions pour 14 horizons (un horizon est le pas de temps pour lequel une prévision est faite, et il correspond à 24 heures), pour une période de 4 ans: de mars 2007 à mars 2011.

C'est une étude univariée, dans le sens où tous les horizons ne seront pas traités en même temps et que les méthodes portent sur un seul horizon à la fois. Néanmoins nous implémentons les méthodes choisies pour les horizons de 5 à 14 indépendemment. Nous n'intégrons pas à notre étude les horizons de 1 à 4, car les prévisions de température détérministes sont très bonnes dans ce cas.

La thèse comporte trois parties: une première grande partie où on présente les SPE, les méthodes statistiques proposées (et leur implémentation) pour améliorer la précision ou la fiabilité du SPE et les critères de comparaison des résultats. Dans la deuxième partie nous proposons des modèles de mélange du modèle présenté dans la première partie et des fonctions de distributions des extrêmes et dans la troisième partie nous introduisons aussi la régression quantile.

En revenant à la première partie, nous commençons par présenter les SPE dont le principe est de faire tourner plusieurs scénarios du même modèle avec des données d'entrée légèrement différentes pour simuler l'incertitude. On obtient alors une distribution de probabilité qui donne la probabilité de réalisation d'un certain événement. Dans l'idéal les membres d'un SPE ont la même probabilité de donner la meilleure prévision.
Les méthodes que nous avons évaluées sont la méthode du meilleur membre et la méthode bayesienne. Les résultats obtenus par ces méthodes lors de l'application aux données de Météo France sont comparés entre eux et avec les prévisions initiales par l'intermédiaire des critères de précision et de fiabilité.
La variante la plus complexe de la méthode du meilleur membre est proposée par V. Fortin (voir [FFS06]). L'idée est de "habiller" chaque membre d'un SPE avec un modèle d'erreurs construit

sur une base de prévisions passées, en prenant en compte seulement les erreurs données par les meilleurs membres pour chaque pas de temps (on considère qu'un membre est le meilleur pour un certain pas de temps quand la prévision qu'il donne pour ce pas de temps fait la plus petite erreur, en valeur absolue, par rapport à la réalisation au pas de temps considéré). Cette approche ne donne pas de bons résultats dans le cas des SPE qui sont déjà sur dispersifs. Une deuxième approche a été mise au point pour permettre la correction des SPE de ce type. Cette approche propose d'"habiller" les membres de l'ensemble avec des poids différents par classes d'ordres statistiques ce qui revient à mettre de poids dans la simulation finale, sur les scénarios, en fonction de leurs performances observées dans la période d'étude.
La méthode Bayesienne a été proposée par A. Raftery (voir [RGBP04]). C'est une méthode statistique de traitement de sorties de modèles qui permet d'obtenir des distributions de probabilité calibrées même si les SPE eux-mêmes ne sont pas calibrés (on considère qu'une prévision est bien calibrée quand un événement ayant une probabilité d'apparition p se produit en moyenne à une fréquence p). Le traitement statistique proposé est d'inspiration bayesienne, où la densité de probabilité du SPE est calculée comme une moyenne pondérée des densités de prévision des modèles composants. Les poids sont les probabilités des modèles estimées à posteriori et reflètent la performance de chacun des modèles, performance prouvée dans la période de test (la période de test est une fenêtre glissante qui permet d'utiliser une base de données moins lourde pour estimer les nouveaux paramètres).

Les prévisions obtenues par les deux méthodes sont comparées par des critères de précision et/ou de calibrage des SPE comme: l'erreur absolue moyenne(MAE), la racine carrée de l'erreur quadratique moyenne (RMSE), l'indice continu de probabilité (CRPS), le diagramme de Talagrand, la courbe de fiabilité, le biais, la moyenne. La méthode bayesienne améliore le calibrage du SPE dans la partie centrale de la distribution mais elle perd en précision par rapport au SPE initial. La méthode du meilleur membre améliore aussi la distribution dans sa partie centrale et elle améliore la précision des températures de point de vue du CRPS, mais pas de point de vue RMSE. Par rapport à ces résultats nous avons continué les travaux en regardant plus en détail ce qui se passe dans les queues de distribution. Une autre suite possible aurait été celle des prévisions multidimensionnelles, ce qui revenait à traiter tous les horizons de temps simultanément mais la première piste est privilégiée par rapport aux besoins d'EDF dans la gestion du système électrique de mesurer et réduire les risques de défaillance.

Dans la deuxième partie on présente la théorie des valeurs extrêmes et les modèles du mélange pour introduire après les modèle de mélange d'extrêmes que nous utilisons pour construire un modèle qui nous permet de combiner la méthode du meilleur membre pour la partie centrale de la distribution et un modèle spécifique à la théorie des valeurs extrêmes pour les queues de distribution. Nous proposons d'abord le meilleur moyen, adéquat à notre cas, de séparer les queues de distribution de la partie centrale (trouver l'épaisseur des queues), puis nous construisons le modèle de mélange adéquat à nos besoins. Nous faisons également des tests pour voir quelle est la modélisation adéquate si nous avons besoin d'un seul modèle d'extrême, ou une combinaison des modèles extrêmes (en fonction de l'estimation des paramètres des fonctions d'extrême: forme,

location, échelle). Nous allons choisir trois différents modèles de mélanges (par trois différents critères) et nous les utilisons pour produire des nouvelles prévisions, que nous allons du nouveau comparer aux prévisions initiales. Tous les trois modèles améliorent la compétence globale du SPE (CRPS), mais donnent un effet étrange au dernier rang de la queue droite du diagramme des rangs, effet confirmé par le calculs des quantiles (0.99, 0.98, 0.95) qui ne sont pas bien estimés par les prévisions données par nos modèles de mélange. Nous proposons de mettre en oeuvre une dernière méthode, en s'intéressant cette fois aux quantiles et non plus aux moments.

La méthode proposée est la méthode de régression des quantiles et elle fait l'objet de la troisième et dernière partie de thèse. Puisque nous voulons modéliser les queues, il est important de tenir compte des erreurs relatives aux quantiles. C'est pourquoi nous allons utiliser une distance de χ^2 qui permet d'expliciter la sur-pondération des queues. Nous avons choisi les classes autour de la probabilité d'intérêt pour nous, soit 1 %: $[0; 0,01]$, $[0,01; 0,02]$ et $[0,02; 0,05]$ pour la partie inférieure de la queue et les classes symétriques pour la queue supérieure. Nous allons utiliser cette mesure pour estimer les améliorations apportées aux prévisions extrêmes. Les résultats sont positifs, même si il reste quelques biais dans la représentation de la queue.

Après avoir appliqué toutes ces méthodes, sur des données de températures de Météo France, sur une période de quatre ans nous sommes amenés à la conclusion que la meilleure méthode consiste à utiliser une méthode du type meilleur membre pour produire des simulations de température pour le coeur de la distribution et d'adapter une régression quantile pour les queues de la distribution.

Summary

This work is carried out under a CIFRE thesis, as a partnership between the University of Paris Descartes (Laboratory MAP5) and the OSIRIS department of EDF R&D, France.

The thesis has for objective to study new statistical methods to correct temperature predictions that may be implemented on the ensemble prediction system (EPS) of Meteo France so to improve its use for the electric system management, at EDF France. The EPS of Meteo France we are working on contains 51 members (forecasts by time-step) and gives the temperature predictions for 14 days (also called time-horizons, 1 time-horizon = 24 hours) for the period: March 2007 - March 2011.

It is an univariate study: the 14 time-horizons will not be processed at the same time, the methods will focus on one horizon at a time. Nevertheless we implement the chosen methods for the ten horizons: from 5 to 14 independently. The time-horizons from 1 to 4 are not being integrated in our study because the deterministic forecasts are very good in this case of short time forecast.

The thesis contains three parts: in the first one we present the EPS, then we present and implement two statistical methods supposed improving the accuracy or the spread of the EPS and we introduce criteria for comparing results. In the second part we introduce the Extreme Value Theory and the mixture models we use to combine the model we build in the first part with models for fitting the distributions tails. In the third part we introduce quantile regression as another way of studying the tails of the distribution.

Coming back to EPS, its principle is to run multiple scenarios of the same model with slightly different input data to simulate the uncertainty. This gives a probability distribution giving the probability of occurrence of a certain event. Ideally all the members of an EPS have the same probability to give the best prediction for a given time-step.

The methods we have evaluated are the method of the best member and the bayesian method. The forecasts obtained by the two methods when they are implemented to data from Meteo France are compared between them and then with the initial forecasts, through criteria of skill

and spread.

The method of the best member in its most complex variant is proposed by V. Fortin (see [FFS06]). The idea is to design for each time-step in the data set, the best member (we consider that a member is the best for a certain time-step when the prediction it gives has the smallest error, in absolute value, relative to the realization of the time-step.) among all scenarios (51 in our case) and to construct an error pattern using only the errors made by those best members and then to "dress" all members with this error pattern. This approach does not give good results in the case of the EPS that are already over dispersive. A second approach was developed to permit correction of this type of EPS. This approach proposes to "dress" ensemble members with different weights by class statistical ranks.

The bayesian method we introduce in the thesis is proposed by A. Raftery [RGBP04]. This is a statistical treatment of model output that provides calibrated probability distributions even if the initial EPS are not calibrated (we assume that a forecast is well calibrated when an event with a probability of occurrence of p occurs in average with a p frequency). With this method we compute the density function of the EPS as a weighted average of the densities of the components predictions scenarios. The weights are the probabilities of the scenarios of giving the best forecasts, estimated a posteriori and reflect the performance of each scenario, proven in a training period (the training period is a sliding window of a optimum length (found by certain criteria) that allows using a lighter database to estimate new parameters).

The predictions obtained by both methods are compared using criteria of skill and spread, specific of the EPS as: the mean absolute error (MAE), the root mean square error (RMSE), the continuous rank probability score (CRPS), the Talagrand diagram, the reliability diagram. The bayesian method slightly improves the spread of the EPS for the bulk of the distribution but losses in overall skill of the EPS. The best member method does also improve the spread in its main mode and it also improve the overall skill from the CRPS point of view. These results make us want to study more in detail what happens in the tails of the distribution. Another possible continuation would have been the multidimensional forecasting (treating all time horizons simultaneously). The need to manage the power system to measure and reduce the risk of failure makes us privilege the first track, the one concerning the study of the extreme values of the distribution.

In the second part of the thesis we build a mixture model which allows us to use the best member method for the bulk of the distribution and a model specific to the extreme value theory for the tails. We first find a way for separating the distribution (find the tails heaviness), then we build the mixture model adequate to our needs. We also make some tests to find out if what we need to fit for the distributions tails is only one extreme model, or a combination of extreme models (function of time, tail, time-horizon). We choose three different mixtures models and we use them to produce new forecasts, which are compared to the initial forecasts. All the three models improve the global skill of the forecasting system (CRPS) but give a strange effect to the last rank of the right tail of the rank diagram confirmed by the quantile computations (.99,

.98, .95) that are not well estimated by the forecasts given by the mixture model. We propose to implement a last method, and this time we are not interest in moments, we are interested in quantiles.

The method is the quantile regression method and it is the subject of the third and last part of the thesis. Since we want to model the tails, it is important to take account the relative errors on quantiles. That is why we will use a χ^2 distance which allows explicit over-weighting of the tails. We choose classes around the probability of interest for us, which is 1%: [0; 0.01], [0.01; 0.02] and [0.02; 0.05] for the lower tail, and the symmetric classes for the upper one. We will use this to measure all improvements and the results are positive even if there remains some biases in the tail representation.

In the end all the methods we implement on the ensemble prediction system for a four years period provided by Meteo France, when trying to improve its skill and/or spread bring us to the conclusion that the optimum method is to use a best member method type for the heart of the distribution and to adapt a quantile regression for the tails.

Publications and Conferences

Published Articles Gogonel A. Collet J. and Bar-Hen A., *Statistical Post-processing Methods Implemented on the Outputs of the Ensemble Prediction Systems of Temperature, for the French Electric System Management*, is to be published in the Volume 5(2) of the Case Studies In Business, Industry And Government Statistics Journal, of the Bentley University, Massachusetts, USA.

Submitted Articles Gogonel A. Collet J. and Bar-Hen A., *Improvement of Short-Term Extreme Temperature Density Forecasting using Best Member Method*, is submitted at the Natural Hazards and Earth System Sciences Journal, published by the Copernicus GmbH (Copernicus Publications) on behalf of the European Geosciences Union (EGU).

Research Report Research Report Gogonel A., Collet J. et Bar-Hen A., Implementation of two methods of Ensemble Prediction Systems for electric system management, H-R32-2011-00797-EN, EDF R&D, May 2011

Conferences

1. 20th International Conference on Computational Statistics (COMPSTAT) 2012 Limassol (see [GAA12a]);
2. Workshop on Stochastic Weather Generators (SWGEN), Roscoff 2012 (see [GAA12b]);
3. International Conference of Young Scientists and Students (ICYSS-CS) , Sevastopol April 2012 (see [oYSS12]);
4. The World Statistics Congress (ISI), Dublin, Poster, August 2011 (see [GAAer]);
5. International Symposium on Forecasting, Prague, June 2011 (see [GAA11]);
6. 2ème Congrès de la Sociéte Marocaine de Mathématiques Appliquées, Rabat, Juin 2012 (see [GAA10]).

Chapter 1

Introduction

This first part of the thesis contains the presentation of the context where the need of this thesis appeared. It also contains the data we use to develop the methods we are proposing.

In this first part we are also presenting the Ensemble Prediction Systems, how and when they appeared, how they work and in what case they might be used. We use this first part to dedicate a few words to the uncertainty as this is "the reason" that makes important to have and to improve forecasts in meteorology or other domains.

We continue by presenting the two statistical methods supposed to improve the accuracy or the spread of the EPS and we introduce criteria for comparing results. The methods we have evaluated are the method of the best member and the bayesian method. The forecasts obtained by the two methods when they are implemented to data from Meteo France are compared between them and then with the initial forecasts, through criteria of skill and spread of EPS as: the mean absolute error (MAE), the root mean square error (RMSE), the continuous rank probability score (CRPS), the Talagrand diagram, the reliability diagram.

The results we obtain will be presented in the next part of the thesis in the form of an article that is to be published in the Volume 5(2) of the Journal *Case Studies In Business, Industry And Government Statistics*, Bentley University, Massachusetts, USA.

1.1 Context and Data Description

1.1.1 Context

More and more users are interested in local weather predictions with uncertainty information, i.e. probabilistic forecasts. The energy sector is highly weather-dependent, hence it needs accurate forecasts to guarantee and optimize its activities. Predictions of the production are needed to optimize electricity trade and distribution. The needs in electricity depend on the meteorological conditions.

Electricité de France (EDF) is the most important electricity producer in France. The EDF R&D OSIRIS department is in charge of studying the management methods of the EDF-production system from a short time horizon (CT) to a medium time horizon (MT) (between 3 hours and 3 years, approximately). Short-term forecasting (that we will consider here) is important because the national grid requires a balance between the electricity produced and consumed at any moment in the day. In this department, the group Risk Factor, Price and Decisional Chain is in charge of studying and modeling the risk factors which can impact production system management. Among these risk factors, we shall be interested here in the physical risks [COL08].

Numerous specialists of the physics (for example the meteorologists) build sophisticated numerical models, with uncertainty on the input data. To take into account this uncertainty, they run the same model several times, with slight, but not random perturbation of the data. The probability distributions obtained from the model is not a perfect representation of the risk factor, thus we need to implement statistical methods before we use it but when modeling the results we had to take into account that the results of physical models contain an irreplaceable information.

The correlation between the temperature and the electricity consumption. Temperature is the main risk factor for an electricity producer such as EDF. Indeed, electric heating is well developed in France. If we take into account the variability of the temperature, the power consumed for the heating for a winter given-day may fluctuate about 20GW, that is 40 % of the average consumption. In what concerns the energy, the climatic risk factor is quantitatively less important, because the difference of energy consumed between the warmest and the coldest winters represents approximately 5 % of the energy over the year. Nevertheless, the climatic risk factor remains the first source of uncertainty for EDF. [CD10]

To explain the correlation between the temperature and the electricity consumption we can start by specifying that the French electrical load is very sensitive to temperature because the electrical heating development since the 70's. The influence of the temperature on the French load is mostly known, except for the impact of air conditioning whose trend remains difficult to estimate . The electric heating serves to maintain a temperature close to 20°C inside the buildings. Taking into account that the "free" contributions of heat (sun, human heat), it is considered that the electric heating turns on approximately below 18°C. Beyond that temperature, the heat loss being proportional with the heat difference between inside and outside, the consumption

increases approximatively linearly. Besides, the buildings putting certain time to warm up or to cool down, the reaction to the outside temperature variations is delayed. In Figure 1.1 the non-linear relationship between electricity load and average national temperature at 9 AM (see [DOR09]).

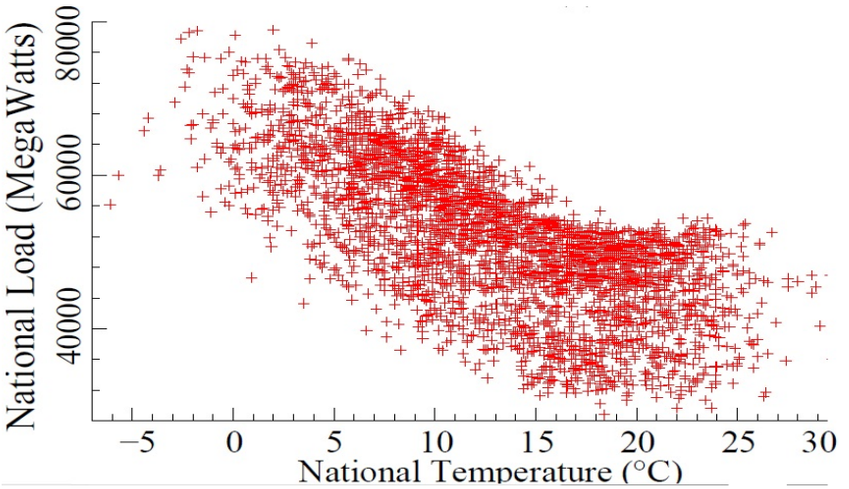


Figure 1.1: Daily electricity loads versus the national average temperature from September 1, 1995 to August 31, 2004, at 9 AM.

This paper has for objective to study the ensemble prediction systems (EPS) provided by Meteo-France. We implement statistical post-processing methods to improve its use for electric system management, at EDF France.

Models used for load Forecasting at EDF are regression models based on past values of load, temperature, date and calendar events. The relationship of load to these variables is estimated by nonlinear regression, using a specifically preconditioned variant of S.G. Nash's truncated Newton (see [NAS84]) developed by J.S. Roy. Load forecasting is performed by applying the estimated model to forecasted or simulated temperature values, date and calendar state. The short term forecasts are performed using an auto-regressive processes applied to the past two weeks residuals of the model (see [BDR05]). The model is based on a decomposition of the load into two components: the weather independent part of the load that embeds trend, seasonality and calendar effects and the weather dependent part of the load.

Up to the 4th time horizon the deterministic forecasts give high quality forecasts, this is why it continues to be used by Météo France for the wheatear forecasts up to four days ahead (see [ENM]). Starting from the 4th horizon, forecasts can be improved and/or the uncertainties in the forecasts better estimated. Currently the value used to predict the consumption is the mean

of the 51 forecasts.

Resorting to the EPS method allows on one hand to extend the horizon where we have good forecasts and on the other hand to give a measure of forecast uncertainty. Unlike the deterministic solution the probability forecast is better adapted to the analysis of risk and decision-making.

First, we study the Meteo-France temperature forecasts in retrospective mode and the temperature realizations in order to establish the statistical link between these two variables. Then, we examine two statistical processing methods of the pattern's outputs. From the state of the art of the existing methods and from the results obtained by the verification of the probability forecasts, a post-processing module will be developed and tested. The goal is to achieve a robust method of statistical forecasts calibration. This method should thus take into account the uncertainties of the inputs (represented by the 51 different initial conditions added to the pattern).

The first method is the best member method (BMM) and it is proposed by V. Fortin ([FFS06]) as an improvement of the one built by Roulston and Smith [RS02], and improved by Wang and Bishop [WB05]. The idea is to design for each lead time in the data set, the best forecast among all the k forecasts provided by the temperature prediction system, to construct an error pattern using only the errors made by those "best members" and then to "dress" all the members of the initial prediction system with this error pattern. This approach fails in cases where the initial prediction systems are already under (or over) dispersive because when an EPS is under dispersive, the outcome often lies outside the spread of the ensemble increasing the probability of the extreme forecast to give the best prediction. And the other way, when a ensemble is over dispersive the probability of the members close to the ensemble mean to be the best members is increasing. It is why a second method was created. It allows to "dress" and weight each member differently by classes of its statistical order.

The second method we implement is the bayesian method that has been proposed by A. Raftery ([RGBP04]). It is a statistical method for post processing model outputs which allows to provide calibrated and sharp predictive Probability Distribution Functions even if the output itself is not calibrated (forecasting are well calibrated if for a p probability forecasts, the predicted event is observed p times). The method allows to use a sliding-window training period (TP) to estimate new models parameters, instead of using all the database of past forecasts and observations.
Results will be compared using standard scores verifying the skill and/or the spread of the EPS: MAE, RMSE, ignorance ccore, CRPS, Talagrand diagram, reliability diagram, bias, mean.

1.1.2 Data Description

We are working on the daily average temperatures in France from march 2007 to march 2011 and the predictions given by Meteo France for the same period. The temperature forecasts is provided by Meteo-France as an ensemble of temperatures prediction system containing the daily average temperatures in France (the weighted mean of 26 values of daily means temperatures

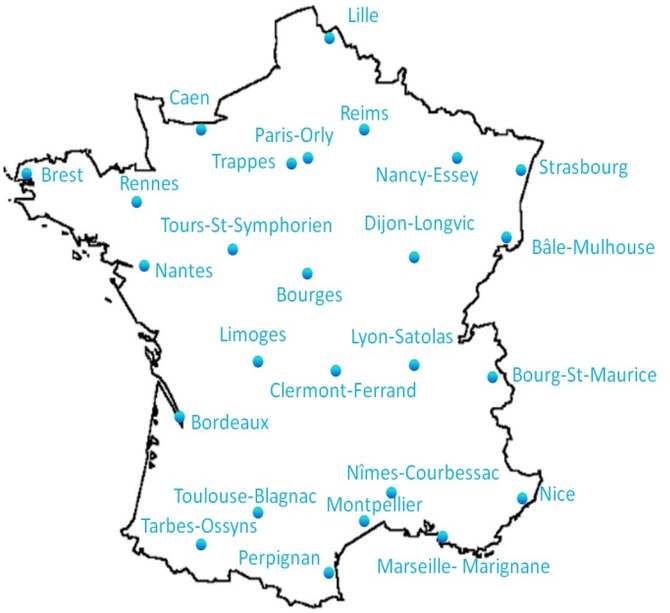


Figure 1.2: The map of the Meteo France stations.

observed by the Meteo stations in France, see Figure 1.1.2) from March 2007 to March 2011 and the predictions given by Meteo France for the same period as an Ensemble Prediction System (EPS) containing 51 forecasts by day, up to 14 time-horizons, corresponding to 14 days.

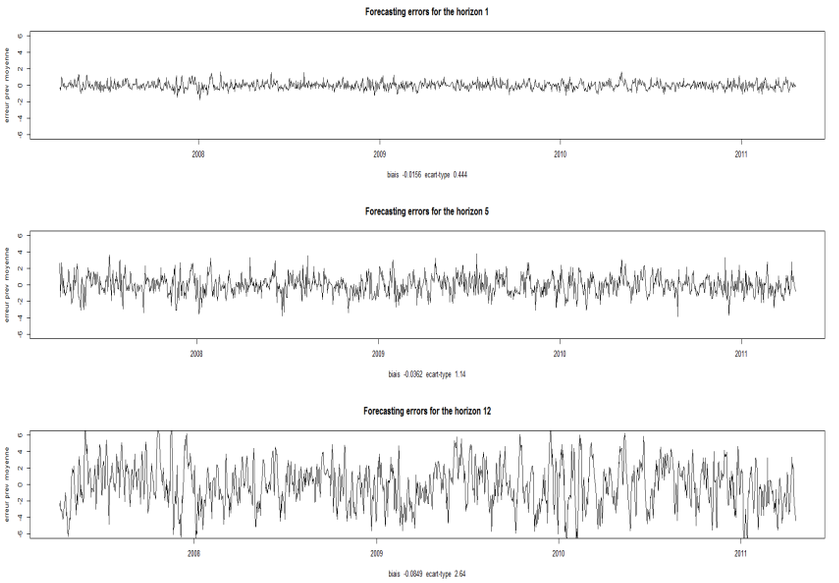
The 51 members of the EPS are 51 equiprobable scenarios obtained by running the same forecasting model with slightly different initial conditions. Figure 1.1.2 represents for one fixed time-step the curves of the evolutions of the 51 scenarios by time-horizon, from 1 to 14. We can notice a small bias ($\sim 0.1°$) starting with the first horizon. The bias is increasing with the horizon-time (normal) and that for all the scenarios in a resembling way.

We observe that the scenarios are randomly named by numbers from 0 to 50, which are not constant from one day to another. Hence, the uncertainty added to ensemble members is not related to the number of the ensemble member (the scenario 0 is the only one standing the same, as it is the one with no perturbation of the initial conditions added).

We start by making an univariate study i.e. we fix the horizon-time. Depending on the quality of the results we could consider a multivariate study.

éTherefore the horizon is fixed, we choose to study the forecasts starting with the 5th horizon because up to horizons 3-4 the determinist forecasts are very good (the Meteo-France pattern is

Average Forecast and Real Temperatures Curves



The evolution of the predictions by horizon from one to fourteen, for all scenarios

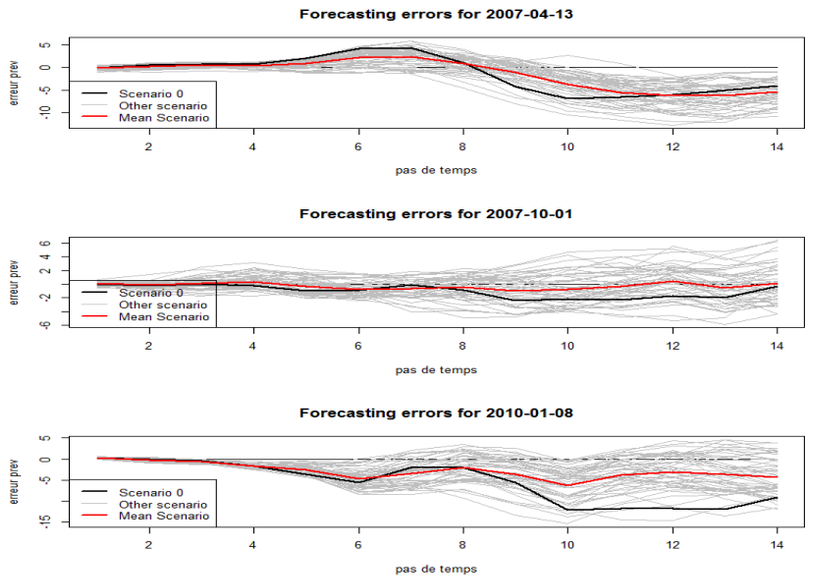


Figure 1.3: Figures corresponding to initial predictions. On top, the curve of realizations and the one of the average predicted temperature. At the bottom, the curves of temperatures for one day, for all the 14 horizons time.

build on purpose under dispersive up to 3 days).

1.2 Ensemble prediction systems (EPS)

The ensemble prediction systems are a rather new tool in operational forecast which allows faster and scientifically justified comparisons of several forecast models. The EPS are built with the aim of obtaining the probability of the meteorological events and the zone of inherent uncertainty in every planned situation. It is a technique to predict the probability distribution of forecast states, given a probability distribution of random analysis error and model error.

The principle of the kind of EPS we are interested in, is to run several scenarios of the same model with slightly different input data in order to simulate the uncertainty (another way to simulate it would be by varying the physical models and/or its parametrization). Then we obtain a probability distribution function informing us about the probability of realization of a forecast. Ideally, the members of a EPS are independent and have the same probability to give the best forecast. Nevertheless, this information is not completely suited to the use for electric system management. For example, PDFs are not smooth, it is a major issue for uncertainties management.

At present, the EPS are based on the notion that forecast uncertainty is dominated by error or uncertainty in the initial conditions. This is consistent with studies that show that, when two operational forecasts differ, it is usually differences in the analysis rather than differences in model formulation that are critical to explaining this difference, see [DOC02].

Among the points of interest in using EPS ([MAL08]) is that it allows to estimate the uncertainty, to have a representative spread of the uncertainty that is in practice, to have an empirical standard deviation of the forecasts comparable with the standard deviation of the observations. EPS may also be used to give an estimation of the probability of occurrence of an event, it helps finding the threshold above (or beyond) which the risk of occurrence of the event is important.

1.2.1 History and building methods

The stochastic approaches in predicting weather and climate was seriously reconsidered after Lorentz discovered the chaotic nature of atmospheric behavior in the 1960s (see [BS07]). According to him the atmosphere has a "sensitivity" in the initial conditions i.e. small differences in the initial state of the atmosphere could lead to large differences in the forecast.

In 1969, Epstein proposed a theoretical stochastic-dynamic approach to describe forecast error distributions in model equations. But the computing power at that time made his approach unrealistic [EPS69].

Instead, Leith proposed a more practical Monte-Carlo approach with limited forecast members. Each forecast member is initiated with randomly perturbed, slightly different initial condition

(IC). Leith showed that Monte-Carlo method is a practical approximation to Epstein's stochastic-dynamical approach, [LEI74]. Leith's Monte-Carlo approach is basically the traditional definition of ensemble forecasting although the content of this definition has been greatly expanded in the last 20 years [DU07].

As computing power increased, operational ensemble forecasting became a reality in early the 1990s. Both National Centers for Environmental Prediction in the USA (NCEP) and the European Center for Medium-Range Weather Forecast(ECMWF) operationally implemented their own global model-based, medium-range ensemble forecast systems. Hence the idea of this kind of prediction systems is to get an ensemble of forecasting models starting with just one model. There are different methods the prediction centers use to create an Ensemble Prediction System (EPS)(not to confound with the methods we will implement, which are post processing methods, hence they are applied on EPS which have already been created). Among the most known methods used to create EPS, there are:

- Method of crossing (used in the USA, NCEP), see [SWa06].
- Method of the singular vector (used in Europe, ECMWF, see [DOC06]).
- Kalman Filter (used in Canada), see [VER06].

The EPS we are using is built by the singular vector method, more exactly it is a ECMWF EPS which evolutions in time can be studied as in [MBPP96]. A description of the current operational version is given by Leutbecher and Palmer (see [MP07]). Since 12 September 2006, the ECMWF EPS has been running with 51 members one being the control forecast (starting from unperturbed initial conditions) and the other 50 are member to whom the initial conditions have been perturbed by adding small dynamically active perturbations (see [BBW+07]). These 51 runs are performed twice a day at initial time 00 and 12 UTC, up to 15 days-ahead, with a certain resolution from day 0 to 10 and a lower resolution for the days up to 15.

1.2.2 Forecasts and uncertainty in meteorology

Given the presence of errors at all levels of the forecasts building, the state of the atmosphere at some point is not perfectly known. This makes delicate to define a single "true" set of parameters. It is nevertheless possible to evaluate the probability of a set of parameters or data input. The uncertainty indicates the intrinsic difficulty in forecasting the event during a period. The best characterization of the uncertainty would be the probability density functions of the simulation errors. Computing a probability density function (PDF) for given model outputs (such as forecast error statistics) is in practice a difficult task primarily because of the computational costs [MAL05]. If we want to list some large classes of uncertainty that would be:

- Natural or fundamental uncertainty. It is the uncertainty of the probabilistic nature of the model. It is the uncertainty that remains when one has the right model and the right model parameters.

- The statistical uncertainty. The model parameters being estimated, there is certainly a bias between them and the "true" model parameters.
- The model uncertainty itself. Choosing a wrong law for as the distribution law. For example, suppose in a certain context that the population is normally distributed when in fact a gamma could be the "good" law to apply.

Sources of forecast errors

To be able to predict the forecast's error, it is necessary to understand the most important components of this error. In order to find the step in the building process of the forecasting, creating errors Houtemaker ([HOU07]) made a list of the most important components of forecast error. There are some of those errors:

Incomplete observations of the atmosphere it is not possible to observe all variables at all locations and at all times.

Error in input data. The precision of the observations is sometimes limited. It can be a random error or a systematic one. Different observations taken by identical platforms can present correlated errors. Because of the observation error among others, the weather forecast will never be established from an perfect initial state.

Weighting observations. In the procedure of data assimilation, a weighted average is calculated from the new observations. The precision of the chosen weights is function of the relative precision of the observations.

Error due to the pattern. Because of the lack of resolution, the model shows a behavior a little bit different from the real behavior of the atmosphere in identical initial conditions.

1.3 The verification methods for EPS

Meteorologists have used EPS for several years now and in the same time many methods of evaluating their performances were also developed [SWB89]. A proper scoring rule maximizes the expected reward (or minimizes the expected penalty) for forecasting one's true beliefs, thereby discouraging hedging or cheating (see [JS07]). One can distinguish two kind of methods: the ones permitting to evaluate the quality of the spread and the ones giving a score (a numerical result) permitting to evaluate the performance of the forecasts [PET08].
Hence, we need to verify *skill* or accuracy (how close the forecasts are to the observations) and *spread* or variability (how well the forecasts represent the uncertainty) (see [JS03]). If model errors played no role, and if initial uncertainties were fully included in the EPS initial perturbations, a small spread among the EPS members would be an indication of a very predictable situation i.e. whatever small errors there might be in the initial conditions, they would not seriously affect the deterministic forecast. By contrast, a large spread indicates a large uncertainty of the deterministic forecast (see [PER03]). As for the skill, it indicates the correspondence between a given probability, and the observed frequency of an event in the case this event is forecast with this probability. Statistical considerations suggest that even for a perfect ensemble (one in which all sources of forecast error are sampled correctly) there is no need to have a high correlation between spread and skill (see [WL98]).

1.3.1 Standard Statistical Measure

Let y be the vector of model outputs and let o be the vector of the corresponding observations. These vectors both have n components. Their means are $\overline{y}$ and $\overline{o}$.

Bias

To study the (multiplicative) bias of the EPS is to study if the forecast mean is equal to the observed mean.

$$\text{Bias}_m \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{o_i} \tag{1.1}$$

It is simple to use. It does not measure the magnitude of the errors and it does not measure the correspondence between forecasts and observations. The perfect score is 1 but it is possible to get a perfect score for a bad forecast if there are compensating errors.

Correlation coefficient

It measures the correspondence between forecasts and observations and is given by the variance between forecasts and observations (r^2), a perfect correlation coefficient is 1.

$$r = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(o_i - \overline{o})}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 \sum_{i=1}^{n}(o_i - \overline{o})^2}} \tag{1.2}$$

Mean absolute error (MAE)

It measures overall accuracy and is defined as:

$$MAE \ = \ \frac{1}{n}\sum_{i=1}^{n}|y_i - o_i| \tag{1.3}$$

It is a linear score which provides the average amplitude of the errors without showing in what direction the errors are.

The root mean square error (RMSE).

A related error measure is the mean square error (MSE), defined as $MSE \ = \ \frac{1}{N}\sum_{i=1}^{N}(y_i - o_i)^2$ but one uses more its square root, RMSE as it has the advantage of being recorded in the same unit as the verifications. Like the MAE, it does not provide the direction of the errors but compared to the MAE puts greater influence on large errors than small errors in the average. If the RMSE is much greater than the MAE this would show a high error variance, the case of equality appears when all errors have the same magnitude. It can never be smaller than the MAE.

1.3.2 Reliability Criteria

The reliability (or spread) measures how well the predicted probability of an event correspond to its observed probability of occurrence. For a p probability forecast, the predicted event should be observed p times.

Talagrand diagram

The Talagrand diagram is a type of bar chart in which categories are represented by bars of varying ranks rather than specific values - a histogram of ranks. It measures how well the spread of the ensemble forecast represents the true variability (uncertainty) of the observations. For each time instant (day) we consider the ensemble of the forecasts values (the observation value included). The values within this ensemble are ordered and the position of the observation is noted (the rank). For example the rank will be 0 if the observation is below all the forecasts and N if the observation is above all the forecasts. Repeating the procedure for all the forecasts we obtain a histogram of observations rank. By examining the shape of the Talagrand diagram, we can draw conclusions on the bias of the overall system and the adequacy of its dispersion (see Figure 1.3.2):

- A flat histogram - ensemble spread correctly represents forecast uncertainty. It does not necessarily indicate a skilled forecast, it only measures whether the observed probability distribution is well represented by the ensemble.
- A U-shaped histogram - ensemble spread too small, many observations falling outside the extremes of the ensemble

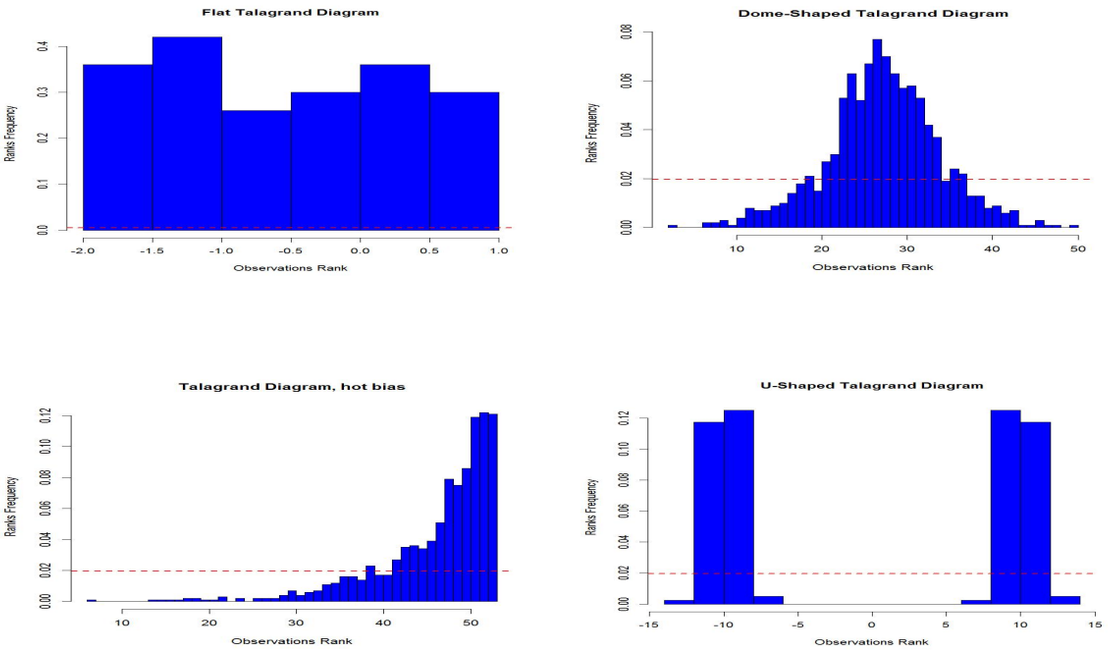


Figure 1.4: Possible situations for the Talagrand Diagram

- A Dome-shaped - ensemble spread too large, too many observations falling near the center of the ensemble
- Asymmetric - ensemble contains bias.

Probability integral transform (PIT)

This method is the equivalent of the diagram of Talagrand for an ensemble having the shape of a probability density function (PDF). Let F be the the cumulative PDF, the value for the observation $x(0)$ is simply $F(x(0))$, or a value between 0 and 1. By repeating the process for all the observations, we obtain a histogram having exactly the same characteristics as the diagram of Talagrand, as regards the performance of the ensemble [GRWG05].

Reliability Diagram

Reliability diagram is the plot of observed probability against forecast probability for all probability categories. A good reliability implies a curve close to diagonal. The deviation from diagonal shows a conditional bias, if it is below the diagonal then the probabilities are too high, if it is above diagonal then the probabilities are too low. A flatter curve shows a lower resolution. For a specific event the reliability diagram represents the frequency of occurrence of a probability

in the EPS and among the observations [POC10]. In this diagram, the forecast probability is plotted against the observed relative frequency.

1.3.3 Resolution (sharpness) Criteria

The resolution (or accuracy, or skill) is the measure of the ability of the forecasts.

Brier Score

The Brier score measures the mean squared probability error (Brier 1950). It is useful for exploring dependence of probability forecasts on ensemble characteristics and is applied to a situation of two possible outcomes (dichotomous variables). Let p_i the probability given by the EPS that the event occurs at each lead time i and o_i the probability observed (at every lead time i) that the event occurs so $o_i = 1$ or $o_i = 0$. The perfect BS is 0.

$$BS = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{1.4}$$

Continuous Rank Probability Score (CRPS)

The CRPS measures the difference between the forecast and observed cumulative distribution functions (CDFs). The CRPS compares the full distribution with the observation, where both are represented as CDFs. If F is the CDF of the forecast distribution and x is the observation, the CRPS is defined as:

$$CRPS(F, x) = \int_{-\infty}^{+\infty} [F(y) - \mathbb{1}_{y \geq x}]^2 \, dy \tag{1.5}$$

where $\mathbb{1}_{y \geq x}$ denotes a step function along the real line that attains the value 1 if $y \geq x$ and the value 0 otherwise. In the case of probabilistic forecasts the CRPS is a probability-weighted average of all possible absolute differences between forecasts and observations. The CRPS tends to be increased by forecast bias and reduced by the effects of correlation between forecasts and observations (see [SDH+07]). One of the advantages is that it has the same units as the predicted variable (so is comparable to the MAE) and does not depend on predefined classes. It is the generalization of the Brier score for the case of the continuous variables. The CRPS provides a diagnostic of the global skill of an EPS, the perfect CRPS is 0, a higher value of the CRPS indicates a lower skill of the EPS.
The CRPS can be decompose in: reliability and resolution terms obtained by the CRPS decomposition proposed by Hersbach ([GIL12]).
We can notice that if all the members of the EPS give all the same prediction, the CRPS is equal to the MAE (it is the case for deterministic forecasts).

Ignorance Score

The Ignorance Score is defined as the opposite of the logarithm of the probability density function f for the observation o [PET08]. Hence, for a single probability we have:

$$ign(f, o) = -\log(f(o))$$

For the PDF of a normal law $\mathcal{N}(\mu, \sigma^2)$ th ignorance score is :

$$ign[\mathcal{N}(\mu, \sigma^2), y] \;=\; \frac{1}{2}\ln(2\pi\sigma^2) \;+\; \frac{(y-\mu)^2}{2\sigma^2}$$

The average ignorance is given by

$$IGN = \frac{1}{n}\sum_{i=1}^{n} ign[\mathcal{N}(\mu, \sigma^2), o] \tag{1.6}$$

ROC Curve

Probabilistic forecasts can be transformed into a categorical yes/no forecasts defined by some probability threshold. Hit rates H and false alarm rate F can be computed and entered into a ROC diagram with H defining the y-axis, F the x-axis (see Figure 1.3.3). The closer the F, H is to the upper left corner (low value of F and high of H) the higher the skill. A perfect forecast system would have all its points on the top left corner, with $H = 100\%$ and $F = 0$.

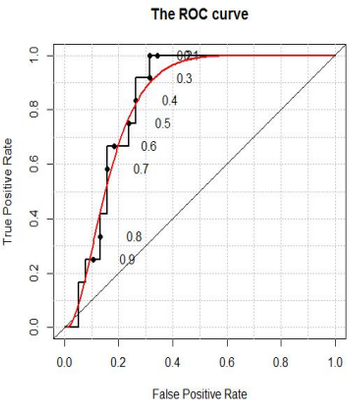


Figure 1.5: Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold.

1.4 Post-processing methods

1.4.1 The best member method

The best member method was proposed by V.Fortin [FFS06] and improves the studies previously led by Roulston and Smith [RS02] then by Wang and Bishop [WB05]. The idea is to design for each time-step in the data set, the best forecast among all scenarios (51 in our case) and to construct an error pattern using only the errors made by those "best members" and then to "dress" all members with this error pattern.

This approach fails in cases where the initial ensemble members are already under or over dispersive, because when an EPS is under dispersive, the outcome often lies outside the spread of the ensemble increasing the probability of the extreme forecast to give the best prediction (conversely for over dispersive EPS). A possible solution is to "dress" and weight each member differently, using a different error distribution for each order statistic of the ensemble. So we can distinguish two more specialized methods: **the one with constant dressing, or the un-weighted members method** and **the one with variable dressing, or the weighted members method**.

The un-weighted members method

The temperature prediction system provides the forecasts $\mathbf{x}_{t,k,j}$, where k is the scenario's number, t is the time when the forecast is made and j is the time-horizon. The method is presented in univariate case so from the start j is fixed, hence $\mathbf{x}_{t,k,j}$ becomes $\mathbf{x}_{t,k}$ where we rename t as the time for which the forecast is made.
Let $\mathbf{y}_t$ be the unknown variable which is forecasted at the moment t, and let $X_t = \{\mathbf{x}_{t,k}, k = 1, 2, ..., K\}$ be the set of all ensemble members of the forecasting system. Given X_t the purpose is to obtain a probabilistic forecasts i.e. $p(\mathbf{y}_t|\mathbf{X}_t)$ in order to provide many more predictive simulations $p(\mathbf{y}_t|\mathbf{X}_t)$ sampled from $p(\mathbf{y}_t|X_t)$ where $X_t = \{\mathbf{x}_{t,m}, m = 1, 2, ..., M\}$ with $M \gg K$.
The concept of conditional probability allows to take into account in a forecast additional information (in this case it will be the forecasts given by Meteo France).

The basic idea of the method is to "dress" each ensemble member $\mathbf{x}_{t,k}$ with a probability distribution being the error made by this member when it happened to give the best forecast. The best scenario is defined $\mathbf{x}_t^*$ as the one minimizing $\|\mathbf{y}_t - \mathbf{x}_{t,k}\|$ for a given norm $\|\cdot\|$. As we are working in an univariate space, the norm is the absolute value:

$$\mathbf{x}_t^* = \arg_{\mathbf{x}_{t,k}} \min |\mathbf{y}_t - \mathbf{x}_{t,k}|$$

For an archive of past forecasts a probability distribution (p_{ε^*}) is created from the realizations of $\varepsilon_t^* = \mathbf{y}_t - \mathbf{x}_t^*$:

$$p(\mathbf{y}_t|X_t) \approx \frac{1}{K} \sum_{k=1}^{K} p_{\varepsilon^*}(\mathbf{y}_t - \mathbf{x}_{t,k}) \tag{1.7}$$

To dress each ensemble members one resamples from the archive of all best-member errors. Hence, the simulated forecasts are obtained by:

$$\hat{\mathbf{y}}_{t,n} = \mathbf{x}_t + \varepsilon_{t,n} \tag{1.8}$$

where $\varepsilon_{t,n}$ is randomly drawn from the estimated probability distribution of ε_t^* and n is the number of simulations by time step.

This first sub-method where all the scenarios are 'dressed' using the same distribution of the error is a choice adapted to the EPS where the scenarios have, a priori, the same probability to give the best forecast. When the initial EPS is already over dispersive, adding noise to each member will increase the variance of the new system.

The weighted members method

Fortin applied this method on a synthetic EPS[1]. He observed that this method failed in the case of over dispersive or under dispersive EPS. The explanation is that when an EPS is under dispersive, the outcome often lies outside the spread of the ensemble. Hence, an extreme forecast has much more chances of giving the best prediction than a forecast close to the ensemble mean. Conversely, when a ensemble is over dispersive the members close to the ensemble mean have much more chances to be the best members than the extremes forecast. Hence, the probability that an ensemble member gives the best forecast as well as the error distribution of the best member are depending on the distance to the ensemble mean. For univariate forecasts we can sort the ensemble members by their distance to the best ensemble members and consider the rank of a member at the dressing sequence.

The only changes in second method, proposed for over (or under) dispersive systems, are the ones related to the use of the statistical rank:

Let

- $\mathbf{x}_{t,(k)}$ be the kth member of the ensemble $X_t = \{\mathbf{x}_{t,k}, k = 1, 2, ..., K\}$ ordered by statistic rank.
- $\varepsilon^*_{(k)} = \{\mathbf{y}_t - \mathbf{x}_t^* |\ \mathbf{x}_t^* = \mathbf{x}_{t,(k)}, t = 1, 2, , ..., T\}$ be the errors of the best ensemble members for every t moment in a database of past forecasts, when the best forecast has the rank k.

To dress each ensemble members differently, instead of resampling from the archive of all best member errors, one resamples from $\varepsilon^*_{(k)}$ to obtain dressed ensemble members. Hence, the simulated forecasts are obtained:

$$\hat{\mathbf{y}}_{t,k,n} = \mathbf{x}_{t,(k)} + \omega \cdot \varepsilon_{t,(k),n} \tag{1.9}$$

[1] A synthetic EPS is an EPS built under certain conditions (ensemble members are independent and identically distributed) and to whom we can vary the parameters we want in order to test different hypothesis or methods

where

- $\varepsilon_{t,(k),n}$ drawn at random from $\varepsilon^*_{(k)}$
- $\omega = \sqrt{\frac{s^2}{s^2_{\varepsilon^*}}}$ and $s^2_{\varepsilon^*} = \frac{1}{T-1}\sum_{t=1}^{T}(\varepsilon^*_{t,(k)})^2$ is the estimated variance of the best-member error. The parameter ω is greater than 1, in the case of the over dispersive forecasts and $\omega < 1$ in the case of under dispersives systems.

1.4.2 Bayesian model averaging

The Bayesian approach is based on the fact that computing the probability of realization of an event does not depend only on its frequency of appearance but also on the knowledge and experience of the researcher. His judgment should naturally be coherent and reasonable. The Bayesian approach implies that two different points of view are not necessarily false, as in the principle of equifinality : the same final state can be reached from different initial states by different ways [BER72].

We base our study of the bayesian method, more exactly on the work of A.Raftery and T.Gneiting proposing bayesian model averaging (BMA) as a standard method which combine predictive distributions from different sources (see [RGBP04]). It is a statistical method for postprocessing model outputs which allows to provide calibrated and sharp PDFs even if the output itself is not calibrated. Raftery's study is implemented on a 5 terms ensemble system coming from different, identifiable sources. It is specified that it is also applicable to the exchangeable situation, and the R-package ensembleBMA, by the same authors, is proving it by offering the choice of the exchangeability in its function definitions. But the second main difference between this study and ours is the significantly larger number of the ensemble members, so we expect to have much longer computation times.

The original idea in this approach is to use a moving training period (sliding-window) to estimate new models parameters, instead of using all the database of past forecasts and observations. This implies the choice of length for this sliding-window training period and the principle guiding this choice is that probabilistic forecasting methods should be designed to maximize sharpness subject to calibration. It is an advantage to use a short training period in order to be able to adapt rapidly to changes (as weather patterns and model specification change over time) but the longer the training period, the better the BMA parameters are estimated [RGBP04]. After comparing training period lengths (from 10 to 60) by measurements as the RMSE, the MAE, the CRPS Raftery concludes that there are substantial gains in increasing the training period up to 25 days, and that beyond that there is little gain. The main difference between their case and ours is that they have 5 models and we have 51 (scenarios).

Let y^T be the quantity to be forecasted and $M_1, ..., M_K$ K statistical models providing forecasts.

According to the law of total probability, the forecasts PDF, $p(y)$ is given by:

$$p(y) \;=\; \sum_{k=1}^{K} p(y|M_k)p(M_k) \tag{1.10}$$

where $p(y|M_k)$ is the forecast PDF based on M_k and $p(M_k)$ is the posterior probability of model M_k giving the best forecast computed on the training data and tells how the model is fitting the training data. The sum of all k posterior probabilities corresponding to the k models is 1: $\sum_{k=1}^{K} p(M_k) = 1$. This allows us to use them as weights, so to define the BMA PDF as a weighted average of the conditional PDFs.

This approach uses the idea that there is a best "model" for each prediction ensemble but it is unknown. Let f_k the bias-corrected forecast provided by M_k, giving the best prediction, to which a conditional PDF $g_k(y|f_k)$ is associated. The BMA predictive model will be:

$$p(y|f_1, .., f_k) \;=\; \sum_{k=1}^{K} w_k g_k(y|f_k) \tag{1.11}$$

where w_k is the posterior probability of forecast k being the best one and is based on forecast's k performance in the training period. $\sum_{k=1}^{K} w_k = 1$.
For temperature the conditional PDF can be fit reasonably well using a normal distribution centered at a bias-corrected forecast $a_k + b_k f_k$:, as shown by Raftery et al. (see [RGBP04]):

$$y|f_k \sim \mathcal{N}(a_k + b_k f_k, \sigma^2)$$

The parameters a_k, b_k as well as the w_k are to be estimated on the basis of the training data set: a_k and b_k by simple linear regression of y_t on f_{kt} for the training data and $w_k, k = 1, .., K$, and σ by maximum likelihood (see [AF97]) from the training data. For algebraic simplicity and numerical stability reasons it is more convenient to maximize the logarithm of the likelihood function rather than the likelihood function itself and the expectation-maximization (EM) algorithm [DLR77] is used.
Finally the BMA PDF is a weighted sum of normal PDFs, the weights, w_k, reflect the ensemble members overall performance over the training period, relative to the other members.

Chapter 2

Implementation of two Statistic Methods of Ensemble Prediction Systems for Electric System Management (CSBIGS Article)

This chapter is in the form of an article, which is to be published in the Volume 5(2) of the *Case Studies In Business, Industry And Government Statistics* Journal, of Bentley University, Massachusetts, USA.

It contains the detailed presentation of the implementation of the best member method and the bayesian method on the four years Meteo France temperature data. It also contains the computation of the specific criteria of skill and spread, so that we may compare the obtained forecasts to one another. We will have three EPS to compare: the initial one, the one obtained by the best member method and the one obtained by the bayesian method.

The conclusions of the study presented in the article are that the bayesian method slightly improves the spread of the EPS for the bulk of the distribution but losses in overall skill of the EPS, the best member method does also improve the spread in its main mode and it also improve the overall skill from the CRPS point of view. These results make us want to continue our works by improving the extreme parts of the distribution. The need to measure and reduce the risk of failure, in electrical consumption managing, encourages us to follow our feeling by continuing the study of the extreme values of the distribution.

Implementation of two Statistic Methods of Ensemble Prediction Systems for Electric System Management

Gogonel, Adriana
EDF R&D, OSIRIS Department and University of Paris Descartes, France

Collet, Jérôme
EDF R&D, OSIRIS Department, France

Bar-Hen, Avner
University of Paris Descartes, MAP5 Laboratory, France

Abstract

This paper presents a study of two statistical post-processing methods when implementing it on the forecasts provided by the ensemble prediction system (EPS) of temperature of Meteo-France. The results could be useful in the management of the electricity consumption at EDF France. Those methods are the best-member method (BMM), proposed by Fortin ([FFS06]), and the bayesian method (BMA), proposed by Raftery ([RGBP04]). The idea of the BMM is to design for each lead time in the data set, the best forecast among all the k forecasts provided by the temperature prediction system, to construct an error pattern using only the errors made by those "best members" and then to "dress" all the members of the initial prediction system with this error pattern. The BMA is a statistical method combining predictive distributions from different sources. The BMA predictive probability density function (PDF) of the quantity of interest is a weighted average of PDFs centered on the bias-corrected forecasts, where the weights are equal to posterior probabilities of the models generating the forecasts and reflecting the models skill over the training period. The resulting forecasts when implementing it on our data set are compared one with another and both compared to the initial forecasts, using scores verifying the skill and/or the spread of the EPS: the mean absolute error (MAE), the root mean square error (RMSE), ignorance score, the continuous rank probability score (CRPS), Talagrand diagram, bias, mean. The purpose is to improve the probability density function of the forecasts, preserving in the same time the quality of the mean forecasts.

Keywords: *forecasting; ensemble prediction systems; energy; bayesian analyse*

1 Introduction

1.1 Context

The energy sector is highly weather-dependent, hence it needs accurate forecasts to guarantee and optimize its activities. Predictions of the production are needed to optimize electricity trade and distribution. Of course needs in electricity depend on the meteorological conditions.

Numerous specialists of physics (for example meteorologists) build sophisticated deterministic numerical models, with uncertainty on the input data. To take into account this uncertainty, they run the same model several times, with slight but non random perturbation of the data. It seems obvious that the results of physical models contain an irreplaceable information. Nevertheless we notice that the probability distributions obtained from the model is not a perfect representation of the risk factor, thus we need to submit it to a statistical processing before we use it.

The correlation between temperature and electricity consumption. Temperature is the main risk factor for EDF as an electricity producer in France, country where electric heating is well developed. If we take into account the variability of the temperature, the power consumed for heating for a winter given-day can vary about 20GW, that is 40 % of the average consumption. In what concerns the energy, the climatic risk factor is quantitatively less important, because the difference of energy consumed between the warmest and the coldest winters represents approximately 5 % of the energy over the year. To explain the correlation between the temperature and the electricity consumption we can start by specifying that the French electrical load is very sensitive to temperature because of the electrical heating development since the 70's. The influence of the temperature on the French load is mostly known, except for the impact of air conditioning whose trend remains difficult to estimate. The electric heating serves to maintain a temperature close to 20°C inside the buildings. Taking into account the "free" contributions of heat (sun, human heat), it is considered that the electric heating turns on approximately below 18°C. Beyond that temperature, the heat loss being proportional with the heat difference between inside and outside, the consumption increases approximatively linearly. Besides, the buildings taking certain time to warm up or to cool down, the reaction to the outside temperature variations is delayed. To take into account this delay, one uses a smoothed temperature (based on a "average" temperature France) as a predictor of the consumption [BDR05]. The representation is similar for the air conditioning.

1.2 Purpose of the work

This study has for objective to improve the probabilistic distribution of forecasts provided by the ensemble prediction systems (EPS) of Meteo-France, preserving the skill of the mean forecasts. The initial EPS contains $k = 51$ members - scenarios of the same model - one starting with unperturbed initial weather conditions (the control forecasts) and 50 from perturbed initial conditions defined by adding small dynamically active perturbations to the operational analysis for the day. Each one of the 51 members of the studied EPS provides trajectories of temperature for 14 time-horizons (1 horizon corresponds to 1 day). We implement statistical post-processing methods to improve its use for electric system management, at EDF France.

The use of the EPS method allows on the one hand to extend the horizon where we have

good forecasts and on the other hand to give a measure of forecast uncertainty. Unlike the deterministic solution the probability forecast is better adapted to the analysis of risk and decision-making.

First, we study the Meteo-France temperature forecasts and the temperature realizations in retrospective mode in order to establish the statistical link between these two variables.
Then, we examine two statistical processing methods of the pattern's outputs. From the state of the art of the existing methods and from the results obtained by the verification of the probability forecasts, a post-processing module will be developed and tested. The goal is to achieve a robust method of statistical forecasts calibration. This method should thus take into account the uncertainties of the inputs (represented by the 51 different initial conditions added to the pattern).

The first method is the best member method (BMM) and it has been proposed by Fortin (see [FFS06]). The idea is to design for each lead time in the data set, the best forecast among all the k forecasts provided by the temperature prediction system, to construct an error pattern using only the errors made by those "best members" and then to "dress" all the members of the initial prediction system with this error pattern. This approach fails in cases where the initial prediction system are already over dispersive. It is why a second sub method was created. It allows to dress and weight each member differently by classes of its statistical order. We present in this paper the second sub method, that we will call the W-BMM.

The second method we implement is the bayesian method that has been proposed by Raftery (see [RGBP04]). It is a statistical method for post processing model outputs which allows to provide calibrated and sharp predictive probability distribution functions (PDFs) even if the output itself is not calibrated (forecasting are well calibrated if for a p probability forecasts, the predicted event is observed p times). The method allows to use a sliding-window training period to estimate new models parameters, instead of using all the database of past forecasts and observations.
Results will be compared using scores verifying the skill and/or the spread of the EPS: MAE, RMSE, ignorance score, CRPS, Talagrand diagram, reliability diagram, bias, mean.

2 Ensemble prediction systems (EPS)

The ensemble prediction systems are a rather new tool in operational forecast which allows faster and scientifically justified comparisons of several forecast models. The EPS are conceived in order to give the probability of the meteorological events and the zone of inherent uncertainty in every planned situation. It is a technique to predict the probability distribution of forecast states, given a probability distribution of random analysis error and model error.

The principle of the EPS is to run several scenarios of the same model with slightly different input data in order to simulate the uncertainty. In the current system Meteo-France is using, each EPS perturbation is a linear combination of singular vectors with maximum growth computed using a total energy norm. The assumption underlying the linear combination is that initial error is normally distributed in the space spanned by singular vectors. A Gaussian sampling technique is used to sample realizations from this distribution (see [DOC06]).

At present, the EPS are based on the notion that forecast uncertainty is dominated by error or uncertainty in the initial conditions. This is consistent with studies that show that, when two operational forecasts differ, it is usually differences in the analysis rather than differences in model formulation, see [DOC02].

We can notice some of the primary objectives to whom the EPS performances should respond ([MAL08]):

1. Allows to estimate the uncertainty, to have a representative spread of the uncertainty that is in practice, to have an empirical standard deviation of the forecasts comparable with the standard deviation of the observations;

2. Gives a good estimation of the probability of an event;

3. It is convenient to linear combinations of models in forecast.

3 The verification methods for EPS

Meteorologists have been using EPS for several years now and in the same time many methods of evaluating their performances were also developed (see [SWB89]). A proper scoring rule maximizes the expected reward (or minimizes the expected penalty) for forecasting one's true beliefs, thereby discouraging hedging or cheating (see [JS07]). One can distinguish two kind of methods: the ones permitting to evaluate the quality of the spread and the ones giving a score (a numerical result) permitting to evaluate the performance of the forecasts (see [PET08]).
Hence, we need to verify *skill* or accuracy (how close the forecasts are to the observations) and *spread* or variability (how well the forecasts represent the uncertainty). If model errors played no role, and if initial uncertainties were fully included in the EPS initial perturbations, a small spread among the EPS members would be an indication of a very predictable situation i.e. whatever small errors there might be in the initial conditions, they would not seriously affect the deterministic forecast. By contrast, a large spread indicates a large uncertainty of the deterministic forecast (see [PER03]). As for the skill, it indicates the correspondence between a given probability, and the observed frequency of an event. Statistical considerations suggest that even for a perfect ensemble (one in which all sources of forecast error are sampled correctly) it may not have a high correlation between spread and skill (see [WL98]).

3.1 Standard Statistical Measures

Let y be the vector of model outputs and let o be the vector of the corresponding observations. These vectors both have n components. Their means are respectively $\overline{y}$ and $\overline{o}$.

Bias given by:

$$\text{Bias}_m \;=\; \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{o_i} \tag{1}$$

Correlation Coefficient given by:

$$r = \frac{\sum_{i=1}^{n}(y_i \,-\, \overline{y})(o_i \,-\, \overline{o})}{\sqrt{\sum_{i=1}^{n}(y_i \,-\, \overline{y})^2 \sum_{i=1}^{n}(o_i \,-\, \overline{o})^2}} \tag{2}$$

Mean absolute error (MAE) measures overall accuracy and is defined as:

$$MAE \;=\; \frac{1}{n}\sum_{i=1}^{n}|y_i - o_i| \tag{3}$$

The root mean square error (RMSE) has the advantage of being recorded in the same unit as the verifications and it is the root square of the $MSE \;=\; \frac{1}{N}\sum_{i=1}^{N}(y_i - o_i)^2$

3.2 Reliability

The reliability (or spread) measures how well the predicted probability of an event correspond to its observed probability of occurrence. For a p probability forecast, the predicted event should be observed $round(p)$ times.

The Talagrand Diagram. It is a type of bar chart in which categories are represented by bars of varying ranks rather than specific values - a histogram of ranks. The Talagrand diagram has its origins in the PIT [DC10]. It measures how well the spread of the ensemble forecast represents the true variability (uncertainty) of the observations. For each time instant (day) we consider the ensemble of the forecasts values (the observation value included). The values within this ensemble are ordered and the position of the observation is noted (the rank). For example the rank will be 0 if the observation is bellow all the forecasts and N if the observation is above all the forecasts. Repeating the procedure for all the forecasts we obtain a histogram of observations rank. By examining the shape of the Talagrand diagram, we can draw conclusions on the bias of the overall system and the adequacy of its dispersion:

- A flat histogram - could show an ensemble spread correctly represents forecast uncertainty. It does not necessarily indicate a skilled forecast, it only measures whether the observed probability distribution is well represented by the ensemble.

- A U-shaped histogram - ensemble spread too small, many observations falling outside the extremes of the ensemble
- A dome-shaped - ensemble spread too large, too many observations falling near the center of the ensemble
- Asymmetric - ensemble contains bias.

3.3 Resolution (sharpness)

The resolution (or accuracy, or skill) is the measure of the ability of the forecasts.

Continuous rank probability score (CRPS) The CRPS measures the difference between the forecast and observed cumulative distribution functions (CDFs). The CRPS compares the full distribution with the observation, where both are represented as CDFs. If F is the CDF of the forecast distribution and x is the observation, the CRPS is defined as:

$$CRPS(F, x) = \int_{-\infty}^{+\infty} [F(y) - \mathbb{1}\{y \geq x\}]^2 \, dy \tag{4}$$

where $\mathbb{1}\{y \geq x\}$ denotes a step function along the real line that attains the value 1 if $y \geq x$ and the value 0 otherwise. In the case of probabilistic forecasts the CRPS is a probability-weighted average of all possible absolute differences between forecasts and observations. The CRPS tends to be increased by forecast bias and reduced by the effects of correlation between forecasts and observations (see [SDH+07]). One of the advantages is that it has the dimensions of the predicted variable (so is comparable to the MAE) and does not depend on predefined classes. It is the generalization of the Brier score for the case of the continuos variables. The CRPS provides a diagnostic of the global skill of an EPS, the perfect CRPS is 0, a higher value of the CRPS indicates a lower skill of the EPS.

4 Post-processing methods

4.1 The best member method

The best member method was proposed by V.Fortin [FFS06] and improves the studies previously led by Roulston and Smith [RS02] then by Wang and Bishop [WB05]. The idea is to design for each lead time in the data set, the best forecast among all 51s (in our case) and to construct an error pattern using only the errors made by those "best members" and then to "dress" all members with this error pattern. This approach doesn't work in cases where the undressed ensemble members are already over or under dispersive and the solution is to weight and dress each member differently, that is using a different error distribution for each order statistic of the ensemble. So we can distinguish two more specialized methods: the one with constant dressing, or the un-weighted members method and the one with variable

dressing, or the weighted members method. We will implement and present in this paper the weighted members method (W-BMM).

4.1.1 The weighted members method

Fortin applied this method on a synthetic EPS[1]. He observed that this method failed in the case of over dispersive or under dispersive EPS. The explanation is that when an EPS is under dispersive, the outcome often lies outside the spread of the ensemble. Hence, an extreme forecast has much more chances of giving the best prediction than a forecast close to the ensemble mean. Conversely, when a ensemble is over dispersive the members close to the ensemble mean have much more chances to be the best members than the extremes forecast. Hence, the probability that an ensemble member gives the best forecast as well as the error distribution of the best member are depending on the distance to the ensemble mean. For univariate forecasts we can sort the ensemble members from the smallest to the biggest, note theirs ranks and consider the rank of a member at the dressing sequence.

Let $\mathbf{x}_{t,k,j}$ be the temperature predictions provided by a given EPS, where k is the scenario's number, t is the time for when the forecast is made and j is the time-horizon. The method is presented in univariate case so from the start j is fixed, hence $\mathbf{x}_{t,k,j}$ becomes $\mathbf{x}_{t,k}$.
Let $\mathbf{y}_t$ be the unknown variable which is forecasted at the moment t, and let $X_t = \{\mathbf{x}_{t,k}, k = 1, 2, ..., K\}$ be the set of all ensemble members of the forecasting system. Given X_t the purpose is to obtain a probabilistic forecasts i.e. $p(\mathbf{y}_t|\mathbf{X}_t)$ in order to provide many more predictive simulations $p(\mathbf{y}_t|\mathbf{X}_t)$ sampled from $p(\mathbf{y}_t|X_t)$ where $X_t = \{\mathbf{x}_{t,m}, m = 1, 2, ..., M\}$ with $M \gg K$.
The concept of conditional probability allows to take into account in a forecast an additional information (in this case it will be the forecasts given by Meteo France).

The basic idea of the method is to "dress" each ensemble member $\mathbf{x}_{t,k}$ with a probability distribution being the error made by this member when it happened to give the best forecast. The best scenario is defined $\mathbf{x}_t^*$ as the one minimizing $\|\mathbf{y}_t - \mathbf{x}_{t,k}\|$ for a given norm $\|\cdot\|$. As we are working in an univariate space, the norm is the absolute value:

$$\mathbf{x}_t^* = \arg_{\mathbf{x}_{t,k}} \min |\mathbf{y}_t - \mathbf{x}_{t,k}|$$

Let

- $\mathbf{x}_{t,(k)}$ be the kth member of the ensemble $X_t = \{\mathbf{x}_{t,k}, k = 1, 2, ..., K\}$ ordered by statistic rank.
- $\varepsilon^*_{(k)} = \{\mathbf{y}_t - \mathbf{x}_t^* |\ \mathbf{x}_t^* = \mathbf{x}_{t,(k)}, t = 1, 2, , ..., T\}$ be the errors of the best ensemble members for every t moment in a database of past forecasts, when the best forecast has the rank k.

[1] A synthetic EPS is an EPS built under certain conditions (ensemble members are independent and identically distributed) and to whom we can vary the parameters we want in order to test different hypothesis or methods

- p_k be the probability that $\mathbf{x}_{t,(k)}$ be the best member, i.e. $p_k = Pr[\mathbf{x}_t^* = \mathbf{x}_{t,(k)}]$

To dress each ensemble members differently, instead of resampling from the archive of all best member errors, one resamples from $\varepsilon^*_{(k)}$ to obtain dressed ensemble members. Hence, the simulated forecasts are obtained:

$$\hat{\mathbf{y}}_{t,k,n} = \mathbf{x}_{t,(k)} + \omega \cdot \varepsilon_{t,(k),n} \tag{5}$$

where

- $\varepsilon_{t,(k),n}$ drawn at random from $\varepsilon^*_{(k)}$;
- n is the number of simulations by time step;
- $\omega = \sqrt{\frac{s^2}{s^2_{\varepsilon^*}}}$ and $s^2_{\varepsilon^*} = \frac{1}{T-1}\sum_{t=1}^{T}(\varepsilon^*_{t,(k)})^2$ is the estimated variance of the best member error. This way of computing ω fails in case of EPS where the uncertainty is already over estimated (s^2 negative).

4.2 Bayesian model averaging

The bayesian approach is based on the fact that the probability of realization of an event does not depend only on its frequency of appearance but also on the knowledge and experience of the researcher.

We base our study on the bayesian method (BMA) (see [RGBP04]). The BMA predictive probability density function (PDF) of the quantity of interest is a weighted average of PDFs centered on the bias-corrected forecasts, where the weights are equal to posterior probabilities of the models generating the forecasts, reflecting the models skill over the training period.

An original idea in this approach is to use a moving training period (sliding-window) to estimate new models parameters, instead of using all the database of past forecasts and observations. This implies the choice of length for this sliding-window training period and the principle guiding this choice is that probabilistic forecasting methods should be designed to maximize sharpness subject to calibration. It is an advantage to use a short training period in order to be able to adapt rapidly to changes (as weather patterns and model specification change over time) but the longer the training period, the better the BMA parameters are estimated [RGBP04]. After comparing training period lengths (from 10 to 60) by measurements as the RMSE, the MAE, the CRPS Raftery concludes that there are substantial gains in increasing the training period up to 30 days, and that beyond that there is little gain. The main difference between their case and ours is that they have 5 models and we have 51 (scenarios).

Let y^T be the quantity to be forecasted and $M_1, ..., M_K$ K statistical models providing forecasts. According to the law of total probability, the forecasts PDF, $p(y)$ is given by:

$$p(y) \;=\; \sum_{k=1}^{K} p(y|M_k)p(M_k|y^T) \tag{6}$$

where $p(y|M_k)$ is the forecast PDF based on M_k and $p(M_k|y^T)$ is the posterior probability of model M_k being correct given the training data and tells if the model is fitting the training data. The sum of all k posterior probabilities corresponding to the k models is 1: $\sum_{k=1}^{K} p(M_k|y^T) = 1$. This allows us to use them as weights, so to define the BMA PDF as a weighted average of the conditional PDFs.

This approach uses the idea that there is a best "model" for each prediction ensemble but it is unknown. Let f_k the bias-corrected forecast provided by M_k, giving the best prediction, to which a conditional PDF $g_k(y|f_k)$ is associated. The BMA predictive model will be:

$$p(y|f_1, .., f_k) \;=\; \sum_{k=1}^{K} w_k g_k(y|f_k) \tag{7}$$

where w_k is the posterior probability of forecast k being the best one and is based on forecast's k performance in the training period. $\sum_{k=1}^{K} w_k = 1$.
For temperature and sea level pressure, the conditional PDF can be fit reasonably well using a normal distribution centered at a bias-corrected forecast $a_k + b_k f_k$:, as shown by Raftery et al. (see [RGBP04]):

$$y|f_k \sim \mathcal{N}(a_k + b_k f_k, \sigma^2)$$

The parameters a_k, b_k as well as the w_k are to be estimated on the basis of the training data set: a_k and b_k by simple linear regression of y_t on f_{kt} for the training data and $w_k, k = 1, .., K$, and σ by maximum likelihood (see [AF97]) from the training data. For algebraic simplicity and numerical stability reasons it is more convenient to maximize the logarithm of the likelihood function rather than the likelihood function itself and the expectation-maximization (EM) algorithm [DLR77] is used.
Finally the BMA PDF is a weighted sum of normal PDFs, the weights, w_k, reflect the ensemble members overall performance over the training period, relative to the other members.

5 Application

5.1 Data description

We are working on temperature forecasts provided by Meteo-France as an ensemble of weather prediction system which contains 51 members, or 51 equiprobable scenarios obtained by running the same forecasting model with slightly different initial conditions.

The data set corresponds to the period between the 30 of March, 2007 and the 20 of April, 2011 and contains forecasts up to 14 time-horizons corresponding to 14 days (1 horizon corresponds to 24 hours). Currently the value used to predict the consumption is the mean of the 51 forecasts. In Figure 1, on top we represent for three fixed time-step the curves of

the prediction errors for the 51 scenarios function of time-horizon (from 1 to 14-days ahead). The errors are normally increasing with the time-horizon; there are particularly small up to the 4th time-horizon. Hence, we consider that up to the 4th time-horizon the deterministic forecasts give high quality forecasts and we implement our improvement methods starting with the 5th horizon. In the same figure, on the bottom we can see the prediction errors for all the period but for three different time-horizons; we notice the same (normal) correlation between the errors and the time-horizon.
As we said above every scenario, among the 51, gives forecasts up to 14-days ahead. The difference between the scenarios comes from the small dynamically active perturbation added to theirs initial conditions. Hence this perturbation is not related to the name of the scenario (numbers between 0 and 50) and is not the same from one day of forecasting start to another [2].
The temperature measurements are made by 26 different French stations, of which we make a weighted average to obtain a single temperature for France. The weights are defined so to explain best the electricity consumption for the different French regions.

We start by setting the time-horizon. Therefore the horizon is fixed, we study the forecast starting with the 5-days ahead horizons as up to 4-days ahead the determinist forecasts are very good (the pattern, of Meteo-France is build on purpose under dispersive up to 3 days). In this paper we present the 5-days ahead results. We can see in Figure 2 superposed on the same graph the curve of the realizations and the curve of the average predicted temperatures.

5.2 Application of the weighted best member method

Let $\mathbf{y}_t$ be the temperature variable we are forecasting at the moment t, and let $X_t = \{x_{t,k}, k = 1, 2, ..., K\}$ be the set of all ensemble members of the Meteo-France forecasting system. We would like to obtain a probabilistic forecasts i.e. $p(\mathbf{y}_t|X_t)$. The conditional probability allows to take into account in a forecast an additional information, in our case the forecasts given by Meteo-France. The best scenario $\mathbf{x}_t^*$ is the one minimizing $|\mathbf{y}_t - \mathbf{x}_{t,k}|$.
To compute the W-BMM method we use the SAS software. We use a cross-validation method to build and verify our models: we separate the four years in our data set in two equal parts: the first part serves for testing period to the model we will validate on the second part and vice versa.

As we mentioned in the presentation of the method, the statistical rank of the ensemble members is taken into account. Though we will have:

- $\mathbf{x}_{t,(k)}$ the kth forecast and $\varepsilon^*_{(k)} = \{y_t - \mathbf{x}_t^* |\ \mathbf{x}_t^* = \mathbf{x}_{t,(k)}, t = 1, 2, , ..., T\}$ as defined above (see 4.1.1).

[2]For example: forecasts given by the scenario 15 computed on July 1st for the period July 1st-July 7th take into account from the beginning a certain perturbation. That perturbation will not be the same as the one taken into account by the scenario 15 when on July 2nd it provides forecasts for the period July 2nd-July 8th

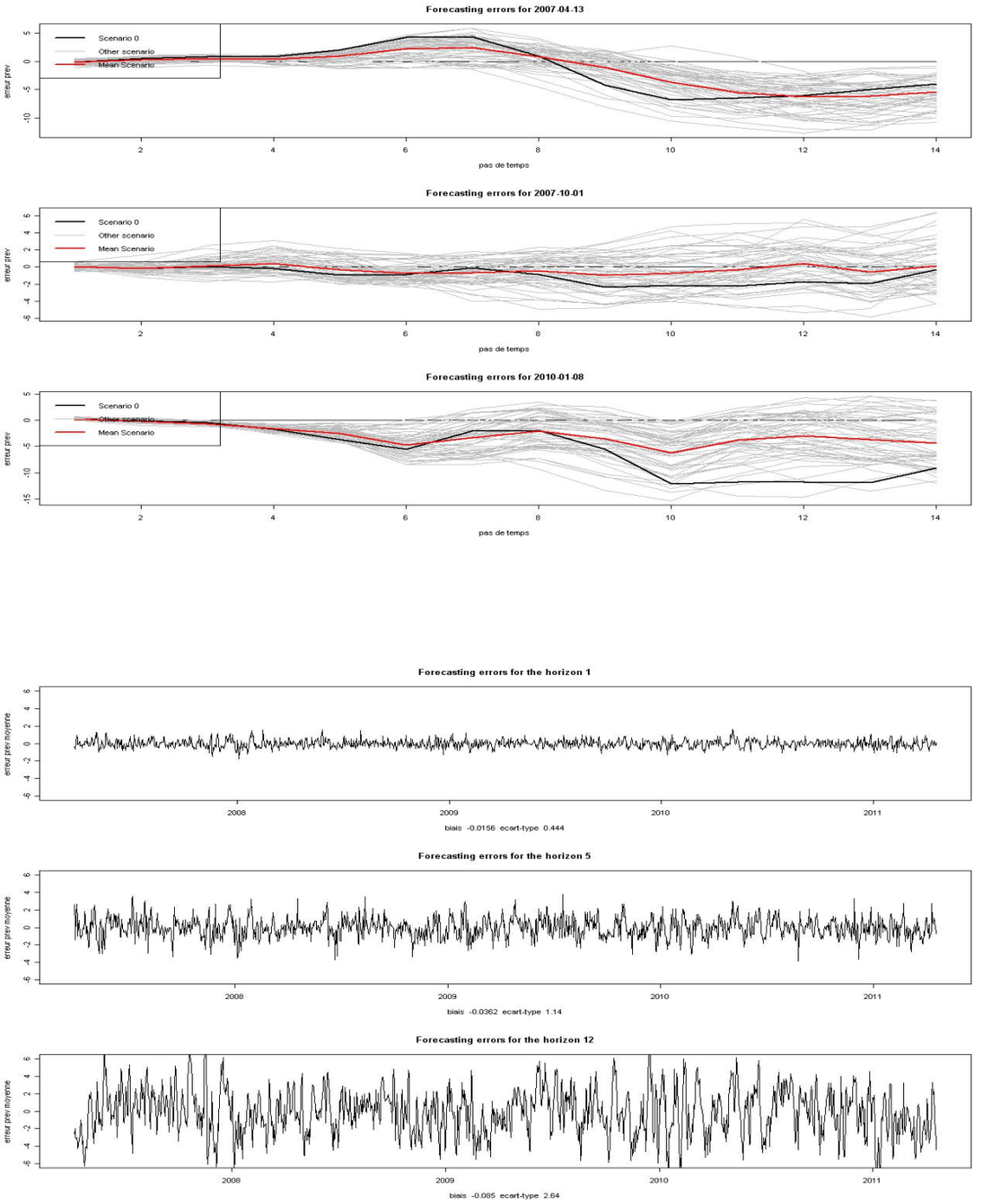


Figure 1: Figures corresponding to initial predictions. On top, the curves of prediction errors for the 51 scenarios for three fixed days, for all the 14 time-horizon (in gray there are scenarios errors from 1 to 51, in black the scenarios 0 -the one with no perturbed initial conditions - and in red the mean of the 51). At the bottom, there are the forecasting errors for three different time-horizons and we can see the errors becomes larger with the time-horizon.

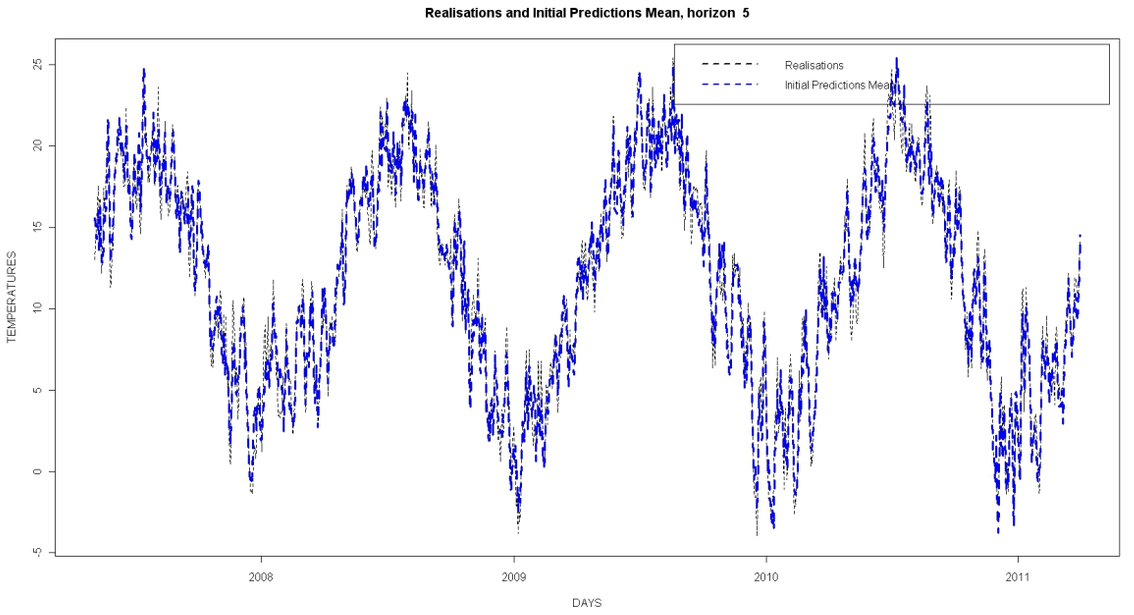


Figure 2: Figures corresponding to initial predictions for 5-days ahead. In black there are the observed temperatures curve, in blue the initial prediction means. We can notice a good precision of the mean forecasts except for the extreme temperatures.

For the archive of past forecasts and a given norm we create a probability distribution from the realizations of $\varepsilon^*_{(k)} = \mathbf{y}_t - \mathbf{x}^*_{t,(k)}$:

$$\varepsilon_{(k)}(t) = \mu_{prev}(t) + \exp(\nu_{prev}(t))\mathcal{N}(0,1) \tag{8}$$

where:

- t the time-step;
- μ_{prev} are the values of the errors, predicted by a linear regression model $M1$ as described below;
- ν_{prev} are the log of absolute values of the residues of the $M1$ pattern, predicted by a linear regression model $M2$ as described below;
- the statistical rank doesn't interfere directly at that moment of the study, it interferes indirectly in the creation of the $M1$ and $M2$ patterns .

The $M1$ pattern explains the prediction error by the initial forecast, day-position within the year **and the statistical rank,** τ_t:

$$\mu_{prev} = \alpha_1 \cdot \mathbf{x}_t + \sum_{i=1}^{3} [\alpha_{2,i} \cdot a(i) + \alpha_{3,i} \cdot b(i)] + \alpha_4 \cdot \tau_t \tag{9}$$

The $M2$ pattern explains the log of the absolute value of the residuals of the $M1$ pattern i.e. ν_{prev}, by the temperature, day-position within an year **and the statistical rank,** τ_t:

$$\nu_{prev} = \beta_1 \cdot \mathbf{x}_t + \sum_{i=1}^{3} [\beta_{2,i} \cdot a(i) + \beta_{3,i} \cdot b(i)] + \beta_4 \cdot \tau_t \tag{10}$$

Hence, we also have two parameters μ_{prev} (predicted by the $M1$ pattern) and ν_{prev} (predicted by the $M2$ pattern). Both of them are generated by 7 parameters: α_1, $\alpha_{2,i}$, $\alpha_{3,i}$ for μ_{prev} $(i = 1, 2, 3)$ and β_1, $\beta_{2,i}$, $\beta_{3,i}$ for ν_{prev} $(i = 1, 2, 3)$. Both of them have the same length as the studied period - 1459. As we said above we use them as parameters of the normal law simulating our new forecasts. We want to obtain $M = 10 \times K = 10\mathbf{x}51 = 510$ simulations so we will draw $N_k = p_k \times M$ dressed ensemble members from each $\mathbf{x}_{t,(k)}$. In this way rang classes having the posterior probability of giving better forecasts will be simulated more than the classes with a small such a probability:

$$\hat{y}_{t,k,n} = \mathbf{x}_{t,(k)} + \omega \times \varepsilon_{t,(k),n} \tag{11}$$

where $\omega = p_k = Pr[\mathbf{x}_t^* = \mathbf{x}_{t,(k)}]$ is the probability that $\mathbf{x}_{t,(k)}$ gives the best forecasts among the $K = 51$.
In the Figure 3 we can observe the median of simulated forecasts, the real temperatures curve and the probability interval 10% - 90% of the simulated forecasts. The curve of the forecasts we simulated is still not perfectly close to the curve of observations. The interesting thing to observe on that graphic is either yes or no the real temperatures curve is always in the 10% - 90% interval. Other results of the tests verifying skill and spread are presented in Chapter 6.

5.3 Application of the bayesian model averaging

Applying the bayesian method consists in constructing the BMA PDF as a weighted sum of normal PDFs, where the weights are reflecting the ensemble members overall performance over the training period. In the application of this method we use a R package for probabilistic forecasting, ensembleBMA created by Raftery's team [FRGS09], using ensemble postprocessing via bayesian model averaging to provide functions for modeling and forecasting data. When we construct the bayesian model we consider that forecasts ensembles members are interchangeable (because of the independence between the forecasts scenarios names, see 2) that is, their forecasts can be assumed to come from the same distribution.

The first and an important step of this method is to choose the length of the training period. We are looking for a good compromise. The advantage of a short training period is that it is able to adapt rapidly to changes (as weather patterns and model specification change over time). The advantage of a longer training period is that the BMA parameters are better estimated. We compare training period lengths (from 10 to 60 days, by 5 or 10 days step) by measurements as the mean absolute error (MAE) and the continuous ranked probability score (CRPS).

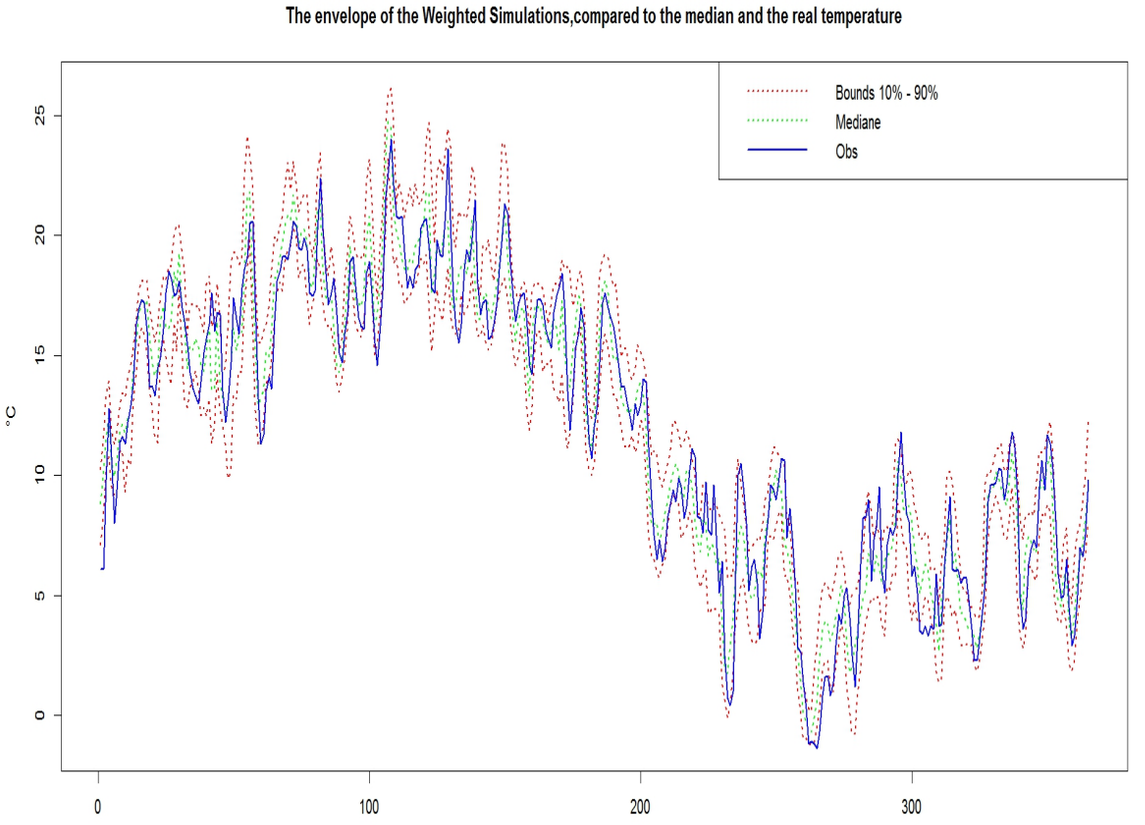


Figure 3: The 10% - 90% interval of the 510 daily simulations (in red) their median (in green) and the realizations curve (in blue) for one year period. To observe on this graph is how the median of the simulations stands in the [10%, 90%] interval.

The values of CRPS and the values of the MAE corresponding to different lengths of the training period are given by Fig 4. We can notice that the two curves are alike, the CRPS decreases from 1.0 (10 days) to 0.73 (60 days). As for the MAE it decreases from 4.2 (10 days) to 1.5 (50 days) and then increases again to 3.1 at 60 days. A 50-days training period is chosen.

Once we decided length period we construct the pattern that fit those data, so that we can obtain the new forecasts system and the corresponding probabilities. Scores are calculated in the section below to decide on the spread and skill of the BMA forecasts.

6 Comparison of the methods by means of the criteria

To compare the quality of the forecasts provided by that statistical post processing methods, we use some of the criteria presented earlier in this paper. We compare here below three kind

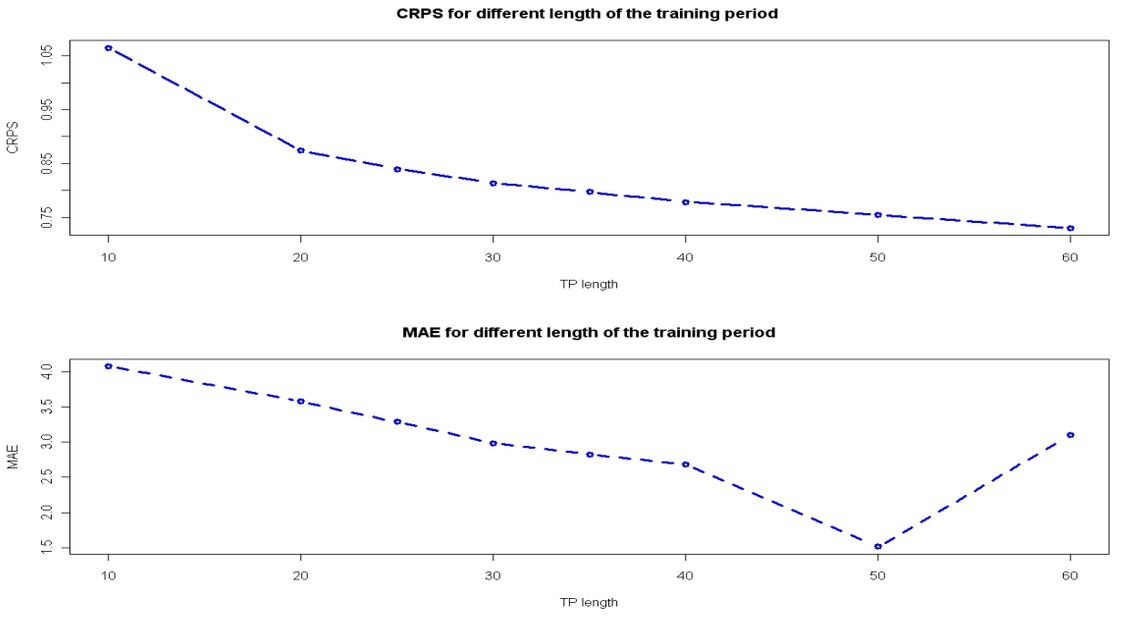


Figure 4: BMA method. The CRPS and the MAE for 5-days ahead for different length of the training period, from 10 to 50, by 5 days step. For the CRPS the values decrease from 1.07 (10 days) to 0.73 (60 days). For the MAE the values decrease from 4.08 (10 days) to 1.5 (50 days) and then increases again to 3.1 at 60 days. A 50-days training period is chosen.

of scores: standard measures, reliability scores and resolution scores for the initial forecasts, unweighted forecast from the best member method, weighted forecasts from the best member method and the forecasts obtained with the bayesian method.

6.1 Standard Measures

Bias We compare the bias for the initial forecasts and the bias for the forecasts obtained by the three methods (see Table 1). The perfect score is 1. The ones obtained for the three methods are 1, so that show good forecasts but it is possible to get a perfect score for a bad forecast if there are compensating errors.

Correlation coefficient The (R^2) we obtain for the two methods has values very closed to 1: 0.96 for the bayesian forecasts and 0.97 for the W-BMM forecasts (see Table 1). Knowing that that a perfect correlation coefficient is 1 our scores show a good correlation between observation and forecasts. The correlation coefficient for the initial predictions is 0.99 so the degree of correlation is not lost after post processing the forecasts.

The root mean square error (RMSE) The RMSE's values for the two methods show small errors of the models. Nevertheless the RMSE for the initial forecasts is smaller than

the RMSE for the forecasts we simulated by W-BMM (see Table 1), and the BMA RMSE is even larger. One possible explanation would be that the RMSE puts greater influence on large errors than smaller errors.

From the standard measures point of view, the forecasts we created have predictive qualities almost as good as the initial predictions.

The mean absolute error (MAE) The smaller the MAE, the better. When we compare the MAE for the initial forecasts and the MAE for the forecasts obtained by the two methods, we find a larger value for the W-BMM: 1.30 and even larger for the BMA 1.53 (see Table 1). Hence, post processing the forecasts by the two methods slowly increases the MAE, as for the RMSE. Nevertheless those are good values of the MAE.

Forecasts	**Bias**	$\mathbf{R^2}$	**RMSE**($^\circ C$)	**MAE**	**CRPS**
Initial forecasts	1	0.99	1.14	0.88	0.63
W-BMM forecasts	1	0.97	1.70	1.29	0.60
Bayesian Forecasts	1	0.96	1.90	1.52	0.75

Table 1: The values of the standard measures for the three applied methods.

6.2 Criteria of reliability

The Talagrand diagram For the initial system of forecast for the 5-days ahead forecasts, the rank histogram is given in the Fig 5a. We notice an asymmetric U-shaped histogram meaning that the ensembles spread is too small (under dispersive), many observations falling outside the extremes of the ensemble. The EPS is under dispersive, the uncertainty is under estimated. The rank histogram of the ensemble obtained by the best member weighted method has as well an U-shape, but it is more close to a plate diagram (see Fig 5b) than the first one. The rank histogram of the ensemble obtained by the BMA Method is given by the Fig 5c. We still notice a U-diagram, but more symmetrical than the BMM one.

6.3 Resolution Criteria

Continuous rank probability score (CRPS) The CRPS measures the difference between the forecast and observed (CDFs). The values of CRPS for the two methods calculated for the entire studied period are given in the Table 1. Those are good values, knowing that the perfect CRPS is 0, proving a high skill of the new created EPS. The better CRPS is the one of the W-BMM forecasts (better than the initial predictions).

a. Initial Predictions b. W-BMM weighted c. BMA

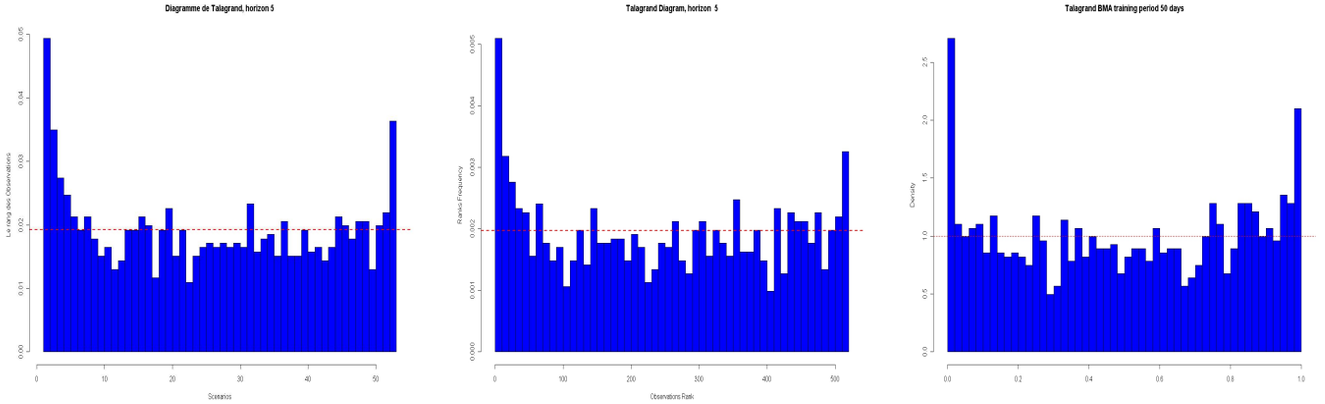


Figure 5: Comparison of the ranks diagrams of the two methods and with the initial predictions Talagrand.

7 Conclusion

The objective of this paper is to extend the number of simulated forecasts of temperature (the 51 per day) provided by Meteo-France and still have a forecasting system with a good quality (spread and skill) that will be useful for the electric system management, at EDF France. Up to the 4th time horizon (1 horizon corresponds to 1 day) the deterministic forecasts give high quality forecasts so we tried to improve the forecasts beyond this time-horizon.
Hence we examined two methods of statistical processing of the pattern's outputs which are taking into account the uncertainties of the inputs (represented by the 51 different initial conditions added to the pattern). We studied their implementation on the data-set provided by Meteo-France. It contains forecasts for the 30 March, 2007 - the 20 April, 2011 period. There are 51 values of forecasts for 14 time-horizons, we studied separately several horizons-time: starting with the 5-days ahead (the study of the 5th horizon presented in the current article). We want to improve the probability density function of the forecasts, preserving at the same time the quality of the mean forecasts.

The first method is the best member method proposed by [FFS06]. The idea is to design for each lead time in the data set, the best forecast among all the k forecasts provided by the temperature prediction system, to construct an error pattern using only the errors made by those "best members" and then to "dress" all the members of the initial prediction system with this error pattern. This method allows us to extend the number of simulated temperatures. We presented here the case where the ensemble members are dressed and weighed differently by classes of its statistical order.

The second method we have implemented is the bayesian method proposed by [RGBP04]. It is a statistical method for post processing model outputs which allows to provide calibrated

and sharp predictive PDFs even if the output itself is not calibrated. The method allows to use a sliding-window training period to estimate new models parameters, instead of using all the database of past forecasts and observations.

Reliability and resolution are the attributes that make the quality of a probabilistic prediction system. Hence, comparing EPS (three in our case, including the initial system) is comparing their scores verifying the skill and the spread.

From the spread point of view there is no significant improvement by any of the two methods: the ranks diagram of the initial EPS shows under dispersion (see Figure 5) and the ranks diagrams of the W-BMM and BMA keep the same shape, so we still see an under-dispersion phenomenon with a more symmetrical diagram for BMA simulations, so the bias is the same in both ways.

From the skill point of view, the bayesian method gives less good results than the initial predictions (see the CRPS, RMSE, MAE values in Table. 1). The W-BMM method has a better (smaller) CRPS but the RMSE end MAE are larger. So as from the spread point of view we can say that the quality of the initial prediction system is preserved but not improved.

The results we obtained are convenient, considering the objective. We increase the number of the forecasting values, which help us better represent the risk, improving in the same time the overall precision of the forecasts.

References

[AF97] J. ALDRICH and R.A. FISHER, *The making of maximum likelihood*, Statistical Science **12** (1997), 162 –176.

[BDR05] A. BRUHNS, G. DEURVEILHER, and J-S. ROY, *A non-linear regression model for mid-term load forecasting and improvements in seasonality*, 15th PSCC, Liege,, 2005.

[DC10] V. DORDONNAT and J. COLLET, *Méthodes de prévision en loi : état de l'art*, Tech. report, EDF R&D, 2010.

[DLR77] A. P. DEMPSTER, N. M. LAIRD, and D. B. RUBIN, *Maximum likelihood from incomplete data via the em algorithm*, Journal of the Royal Statistical Society **39** (1977), 1–38.

[DOC02] IFS DOCUMENTATION, *The ensemble prediction system*, Tech. report, ECMWF, 2002.

[DOC06] ———, *Part v: The ensemble prediction systems*, Tech. report, ECMWF, 2006.

[FFS06] V. FORTIN, A.C. FAVRE, and M. SAID, *Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member*, Q. J. R. Meteorol. Soc. **132** (2006), 1349 –1369.

[FRGS09] C. FRALEY, A. E. RAFTERY, T. GNEITING, and J. M. SLOUGHTER, *ensemblebma: An r package for probabilistic forecasting using ensembles and bayesian model averaging*, Tech. report, Department of Statistics University of Washington, 2009.

[JS07] I. T. JOLLIFFE and D. B. STEPHENSON, *Proper scores for probability forecasts can never be equitable*, American Meteorological Society (2007).

[MAL08] V. MALLET, *Prévision d'ensemble*, Tech. report, INRIA Roquencourt, 2008.

[PER03] A. PERSSON, *User guide to ecmwf forecast products*, Tech. report, ECMWF, 2003.

[PET08] T. PETIT, *Evaluation de la performance de prévisions hydrologiques logiques d'ensemble issues de prévisions météorologiques d'ensemble*, Ph.D. thesis, Faculté Des Sciences Et De Génie Université Laval Québec, 2008.

[RGBP04] A.E. RAFTERY, T. GNEITING, F. BALABDAOUI, and M. POLAKOWSKI, *Using bayesian model averaging to calibrate forecast ensembles*, Physical Review (2004), 20.

[RS02] ROULSTON and SMITH, *Combining dynamical and statistical ensembles*, Tellus **55A** (2002), 16–30.

[SDH+07] J. SCHAAKE, J. DEMARGNE, R. HARTMAN, M. MULLUSKY, E. WELLES, L. WU, H. HERR, X. FAN, and D. J. SEO, *Precipitation and temperature ensemble forecasts from single-value forecasts*, Hydrology and Earth System Sciences Discussions **4** (2007), 655 – 717.

[SWB89] H.R. STANSKI, L.H. WILSON, and W. R. BURROWS, *Survey of common verification methods in meteorology*, Tech. report, Environement Canada, Service de l'environement atmosphérique., 1989.

[WB05] X. WANG and C.H. BISHOP, *Improvement of ensemble reliability with a new dressing kernel.*, Q. J. R. Meteorol. Soc. **131** (2005), 965–986.

[WL98] J.S. WHITAKER and A. F. LOUGHE, *The relationship between ensemble spread and ensemble mean skill*, Monthly Weather Review **126** (1998), 3292 – 3302.

Chapter 3

Mixture Models in Extreme Values Theory

In this part of the thesis we propose a mixture model which allows us to use the best member method for the bulk of the distribution and a specific extreme model for the tails.

We first present the basic background of the extreme value theory and then we present what are mixture models and in what cases it may be interesting to use it. We continue by making a short recap of the recent work in the domain of mixture models including extreme distribution functions.

We then passe to our specific case: after finding the way of separating the distribution (find the tails heaviness), we built the mixture model adequate to our needs. We also make some tests to find out if what we need to fit for the distributions tails is only one extreme model, or a combination of extreme models (function of time, tail, time-horizon). We choose three different mixtures models and we use them to produce new forecasts, which we once more compare to the initial forecasts. All the three models improve the global skill of the forecasting system (CRPS) but give a strange effect to the last rank of the right tail of the rank diagram confirmed by the quantile computations (.99, .98, .95) that are not well estimated by the forecasts given by the mixture model. For correcting this, in the Chapter 4 we propose to implement a last method, and this time we are not interest in moments, we are interested in quantiles.

3.1 Extreme Value Theory

The Theory of the Extreme Values (EVT) estimates the probability of occurrence of the extreme events. This theory studies the behavior of the upper and lower tails for sequences of random variables when their distribution function is unknown. EVT is based on asymptotic arguments for sequences of observations; it provides information about the distribution of the maximum value as the sequences size increases (see [COL01]).

Let $X_1, .., X_n$ be a sample of size n drawn from the variables X. Its distribution function F is then given by:

$$F(x) \;=\; P(X_i \leq x) \quad \text{for} \;\; i = 1, .., n \tag{3.1}$$

Without knowing the general statistical behavior of the sequence we would like to study its extreme behavior. We consider then the maximum of the sequence M_n denoted by: $M_n = \max(X_1, X_2, .., X_n)$ (we can treat the minimum in the same way using the correspondence between $\min(X_1, X_2, .., X_n) = -\max(-X_1, -X_2, .., -X_n)$, so all the results for the maximum can be transposed for minimum). The observations in the sample being i.i.d. then the distribution function of the maximum is:

$$F_{M_n}(x) = P(M_n \leq x) = P(X_1 \leq x, X_2 \leq x, .., X_n \leq x) = P^n(X \leq x) = F^n(x) \tag{3.2}$$

It is not possible to find this distribution without knowing the distribution function of the random variable X. Nevertheless, under certain assumptions we may find the asymptotic behavior of M_n, for large values of n (the samples size). EVT theory provides information about the distribution of the maximum value of such an i.i.d. sample as n increases (see [GYH+11]).

Definition 1. (Distributions of the same type). The distributions F and F^* are of the **same type** if there are constants $a > 0$ and b such that $F^*(ax + b) = F(x)$ for all x. Two random variables are of the same type if theirs distributions are of the same type. In other words, the variables of the same type have the same law in a factor of location and scale near.

Similar to the central limit theorem (CLT), we can find normalization constants $a_n > 0$ and $b_n \in \mathbb{R}$ and a non-degenerate distribution H such as:

$$P\left\{\frac{M_n - b_n}{a_n} \leq x\right\} = (F(a_n x + b_n))^n \to H(x), \text{for} \;\; n \to \infty \tag{3.3}$$

The foundations of the theory of extreme values are set by Fisher and Tippett which propose a first solution to the problem associated to Equation 3.3. The following theorem presenting this solution is often called *the first EVT theorem.*

Theorem 1. ***(Fisher-Tippett or Extremal Types Theorem)*** *Let $X_1, X_2, .., X_n$ be independent random variables with the same probability distribution, and $M_n = max(X_1, X_2, .., X_n)$. If there exist sequences of constants $a_n > 0$ and b_n, such that, as $n \to \infty$, $Pr\left\{\frac{M_n - b_n}{a_n} \leq x\right\} \to G(x)$ for some non-degenerate distribution G, then G has the same type as one of the following distributions:*

Type I (Gumbel)

$$G(x) = \exp\Big\{ - \exp\Big(- \frac{x-b}{a}\Big)\Big\}, \quad -\infty < x < \infty;$$

Type II (Fréchet)

$$G(x) = \begin{cases} 0, & x \leq b, \\ \exp\Big(- \big(\frac{x-b}{a}\big)^{-\alpha}\Big), & x > b; \end{cases}$$

Type III (Weibull)

$$G(x) = \begin{cases} \exp\left\{ - \left(- \big(\frac{x-b}{a}\big)\right)^{\alpha}\right\}, & x < b \\ 1, & x \geq b. \end{cases}$$

for parameters $a > 0$, b and in case of families II and III, $\alpha > 0$.

Thus theorem 1 states that *if* the distribution of the rescaled maxima $\frac{M_n - b_n}{a_n}$ converges, then the limit $G(x)$ is one of the three types, whatever the distribution of the variable parent (e.g. [RAG09]).

Although the behavior of the three laws is completely different, they can be combined into a single parametrization containing one parameter ξ that controls the "heaviness" of the tail, called the shape parameter. This law is called the Generalized Extreme Value distribution (GEV) and it is obtained by introducing a location, μ and scale, σ parameters:

$$G(x) = \exp\left\{ - \left[1 + \xi\left(\frac{x-\mu}{\sigma}\right)\right]_{+}^{-\frac{1}{\xi}} \right\} \tag{3.4}$$

The location parameter, μ determines where the distribution is concentrated, the scale parameter, σ determines its width. The shape parameter ξ determines the rate of tail decay (the larger ξ, the heavier the tail), with:

- $\xi > 0$ indicating the heavy-tailed (Fréchet) case
- $\xi = 0$ indicating the light-tailed (Gumbel, limit as $\xi \longrightarrow 0$) case
- $\xi < 0$ indicating the truncated distribution (Weibull) case

According to the type corresponding to their domain of attraction, the most common distributions could be distributed as in the Table 3.1
If we take into account the GEV, then the extremal theorem may be reformulated as follows: the asymptotic behavior of the maximum of a sufficiently large sample is a GEV distribution. In the same way as for the CLT, a max-stability property makes possible the convergence of the maxima and it allows to find the distribution it converges to.

Attraction domain	**Gumbel** $\xi = 0$	**Fréchet** $\xi > 0$	**Weibull** $\xi < 0$
Law	Normal Lognormal Exponential Gamma	Cauchy Pareto Student	Uniform Beta

Table 3.1: The most common laws distributed by attraction domain

Definition 2. (Max-stability) A distribution G is said to be max-stable if a linear combination of two independent variables from the G distribution, has also a G distribution, up to affine transformations, i.e. up to location and scale parameters.

So in this case we would like to find what type of distributions are stable for the maxima M_n up to affinity, i.e. the distributions satisfying: $M_n = max(X_1, .., X_n) \overset{d}{=} a_n X_i + b_n$ ($X \overset{d}{=} Y$ means that the two random variables X and Y are equal in distribution) for the sample-size-dependent scale and location parameters $a_n > 0$ and b_n and X_i from the parent distribution.

The GEV theory is used for the block maxima approach i.e. only the maximum value of the data within a certain time interval (mostly a year) are considered to be extreme values. In practice this approach may have some restrictions. For example in the case of values representing daily temperature, we would like to describe the extreme-values behavior, using the GEV for annual maxima. But it might happen that a year has one or more values superior of the other years maxima. Taking only one of those values we might loose information that would have contributed to a better understanding of the extremes behavior. And this is the case for the short time series (it is our case). An alternative approach avoiding this inconvenient is the Peaks Over Thresholds (POT) that consists in modeling exceedances above a pre-chosen threshold.

3.1.1 Peaks Over Thresholds

The method of Peaks-over-Threshold (POT) studies the behavior of the values exceeding u, a pre-chosen threshold sufficiently large to assure the asymptotic ground of the analysis. This method was introduced by Pickands [1975] and its advantage is that there are more values to study, not only one by block.

Let $X_1, .., X_n$ be an i.i.d. n-sample drawn from the random variable X, with $X_1, .., X_n \sim F$ and $M_n = max(X_1, X_2, .., X_n)$. We suppose that F satisfies the GEV theorem i.e. for n sufficiently large

$$P(M_n < x) \approx G(x)$$

with $G(x)$ member of the GEV family having ξ, μ, σ, the shape, the location and the scale parameters.

Let $u \in \mathbb{R}$ be the chosen threshold with $N_u = card\{i : i = 1, .., n, X_i > u\}$ the number of exceedances above u among the $(X_i)_{i \leq n}$ and let $Y_i = X_i - u > 0$ be the corresponding exceedances. We define F_u the distribution of the values X_i exceeding u, conditional to the distribution F and the threshold u as follows:

$$F_u(y) \;=\; P(X - u \leq y | X > u) \;=\; \frac{F(y+u) - F(u)}{1 - F(u)}, \quad y \geq 0 \tag{3.5}$$

The Pickands-Balkema-de Haan theorem provides the asymptotic behavior of the distributions F_u their intensities are approximated by the Generalized Pareto Distribution (GPD) and their frequencies by a Poisson point process. The GPD is expressed as a two parameters distribution (shape and scale) by:

$$H_{\xi,\sigma}(y) = \begin{cases} 1 - \left[1 + \frac{\xi y}{\sigma}\right]^{\frac{1}{\xi}} & \text{if} \quad \xi \neq 0 \\ 1 - \exp\left[-\frac{y}{\sigma}\right] & \text{if} \quad \xi = 0 \end{cases} \tag{3.6}$$

where ξ and $\sigma = \sigma(u) > 0$ are the shape parameters and scaling function (depending on the threshold u) of this function. The survival function of the GPD is given by $\overline{H}_{\xi,\sigma}(x) \;=\; 1 - H_{\xi,\sigma}(y)$.

Theorem 2. ***(The Pickands-Balkema-de Haan)*** *For a large class of distribution functions $F(x) = P(X \leq x)$ the GPD is the limiting distribution for the distribution of the excesses, as the threshold tends to τ_F (the upper bound of the distribution function). Formally, we can find a positive measurable function $F(u)$ such that:*

$$\lim_{u \to \tau_F} \sup_{0 \leq y \leq \tau_F - u} |F_u(y) - H_{\xi,\sigma(u)}(y)| = 0 \tag{3.7}$$

if and only if F is in the maximum domain of attraction of the extreme value distribution H_ξ i.e. $F \in MDA(H_\xi)$.

Definition 3. (MDA) A distribution F is in the maximum domain of attraction of a distribution H, $F \in MDA(H)$, if for independent and identically distributed $X_1, X_2, ..X_n$ with distribution function F and $M_n = max(X_1, X_2, .., X_n)$ we can find sequences of real numbers $a_n > 0$ and b_n such that the normalized sequence $(M_n - b_n)/a_n$ converges in distribution to H, where $M_n = max(X_1, X_2, .., X_n)$:

$$\lim_{n \to \infty} \mathbb{P}\left(\frac{M_n - b_n}{a_n} \leq x\right) \;=\; \lim_{n \to \infty} F(a_n x + b_n)^n \;=\; H(x)$$

Choice of the threshold

According to the Pickands-Balkema-De Haan theorem, the distribution function F_u of the exceedances can be approximated by a GPD with the parameters ξ and $\tau = \tau(u)$ to be estimated.

In practice, the choice of the threshold u is difficult and the estimation of the parameters ξ and τ is a question of compromise between bias and variance. A lower u increases the sample

size N_u but the bias will grow since the tail satisfies less well the convergence criterion (equation 3.7) while if we increase the threshold, fewer observations are used and the variance increases.

The GPD has the following properties:

$$E(Y) = \frac{\tau}{1-\xi}, (\xi < 1) \quad \text{and} \quad Var(Y) = \frac{\tau^2}{(1-\xi)^2(1-2\xi)}, (\xi < 1/2) \tag{3.8}$$

Generally, u is chosen graphically using the linearity of the sample mean excess function by plotting $\{(u, e_n(u)), X_{n:n} < u < X_{1:n}\}$ where $X_{1:n}$ and $X_{n:n}$ are the first and nth order statistics of the studied sample and $e_n(u)$ is the sample mean excess function defined by:

$$e_n(u) = \frac{\sum_{i=1}^{n}(X_i - u)^+}{\sum_{i=1}^{n} \mathbb{1}_{X_i > u}}; \tag{3.9}$$

Thus $e_n(u)$ is the sum of the excesses over the threshold u divided by the number of data points which exceed the threshold u. It is an empirical estimate of the mean excess function which is defined as $e(u) = E[X - u | X > u]$ (e.g [MS97]). If the empirical plot seems to follow a reasonably straight line with positive gradient above a certain value of u, then this is an indication that the excesses over this threshold follow a GPD with positive shape parameter (more details of the interpretation of the plot in [EKM97]).
This method may be helpful but in practice it might not give u's right value, so several values should be tested. The choice of the threshold is the subject of many works in the EVT literature: Danielsson et al (see [DDHDV01]) introduced Bootstrap approaches to find the optimum threshold, other authors proposed methods using a random threshold (see [MF00] or [MRT09]) taking the largest k exceedances where k can be deducted by Monte-Carlo methods. Time-varying thresholds may also be appropriate, though there is little guidance on how to make such a choice (see [NJ11]). Alternative methods based on mixture models have also been studied in order to avoid the problem of choosing a threshold (see [FHR02] or [CB09]).

Dependence above threshold

Often, threshold excesses are not independent. For example, a cold day is likely to be followed by another cold day. Methods to handle dependence have been studied:

- Model all the exceedances using an extremal dependence structure (see [ROSG09]);
- De-clustering, that is deciding on a number of clusters and choosing only one values from every cluster (see [ROB]);
- Resampling to estimate standard errors (see [KAT08]).

3.2 Mixture models

The theory of the mixture models appeared from the need of studying populations of individuals naturally composed by several populations distributed according to a certain parametric form (see [PIC07]). Once the populations are identified each one of them is modeled differently. The first step in a mixture model study is to find k, the number of components in the mixture. The main approach for finding the right k is to choose some values of k and to propose different classification functions of k so the problem can be formulated as a model selection problem, that is solved by using criterion as: the Akaike information criterion (AIC), the bayesian information criterion (BIC) (introduced by [SCH78]) or the integrated classification criterion (ICL) that considers the clustering objective of mixture models (introduced by Biernacki, see [BCG00]). Compared to BIC which selects too large number of components, ICL selects a lower number of components which provides good clustering results in real situations (see [PIC07]). But BIC might be interesting to use, for its simplicity of implementation and for its statistical properties (see [GDC97]).

Let $Y = \{Y_1, .., Y_T\}$ be a random n-sample where Y_t is a random vector, y_t its realization and $f(y_t) \in \mathbb{R}^q$ is its density function. In a mixture model, data are supposed to come in different proportions from a mixture of initially specified populations, so the density of Y_t can be written as a combination of densities of the included populations:

$$f(y_t, \psi) \ = \ \sum_{k=1}^{K} \pi_k \ f_k(y_t, \theta_k) \tag{3.10}$$

where $f_k(y_t, \theta_k)$ is the density of the k-th component of the mixture and belongs to a parametric family and θ_k is the vector of the unknown parameters of the density function. π_k is the weight of the k component with $0 \leq \pi_k \leq 1$ and $\sum_{k=1}^{K} \pi_k = 1$ and $\psi = (\pi_1, .., \pi_{k-1}, \theta_1, .., \theta_k)$ is the vector of all the unknown parameters of the mixture. Since we are interested in modeling the distribution by different populations, we notice that one information is missing: the appartenance/or not of the data values to different populations. To quantify this information we introduce a new random variable Z_{tk} that equals 1 if y_t belongs to k population and 0 if not. We suppose that $Z_1, .., Z_T$ are independent ($Z_t = (Z_{t1}, .., Z_{tK})$) and Z_t is supposed to have a multinomial distribution (one draw on k categories with $\pi_1, .., \pi_k$ probabilities):

$$Z_{t1}, .., Z_{tK} \sim \mathcal{M}(1; \pi_1, .., \pi_k)$$

The weights π_k can be viewed as the prior probabilities that one data value belongs to the k population. The posterior probability of Z_{tp} given the realization y_t would be:

$$\tau_{tk} = Pr\{Z_{tk} = 1 | Y_t = y_t\} = \frac{\pi_k f(y_t; \theta_k)}{\sum_{i=1}^{K} \pi_i f(y_t; \theta_i)}$$

Adding the information given by Z_t we obtain a data set called "complete data": $x = (y, z) = \{x_1, .., x_T\} = \{(y_1, z_1), .., (y_T, z_T)\}$ having a distribution function given by:

$$g(x_t; \psi) = \prod_{k=1}^{K} [\pi_k f(y_t; \theta_k)]^{z_{tk}} \tag{3.11}$$

Once the belonging of the data points is established the parameters of each density components of the mixture can be estimated via the data points of k populations. There is a variety of techniques to estimate the parameters of a mixture, main approaches are: graphical method, the maximum likelihood method (see [DLR77]), the method of moments, the bayesian method or robust estimations (see [TL00]).

3.2.1 Parameter estimation

There are several methods to estimate parameters but the most popular are the maximum Likelihood, Bayes estimates and the method of moments. Comparing the three methods, means evaluating the properties of the estimators they build (bias, mean square error, variance, consistency, see [EAT08]).

Method of Moments

The basic idea of this method is to estimate an expectation, par example, by an empirical mean, a variance by a empirical variances etc. If $\theta = \mathbb{E}(Y)$ than the estimator of θ by the method of moments is $\widehat{\theta}_n = \overline{Y}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
More generally, if our data is drawn from $f(y|\theta)$ with the population moments given by $\mu'_k = \mathbb{E}[Y^k]$. From the data we can compute sample moments: $m'_k = \frac{1}{n}\sum_{i=1}^{n} Y_i^k$. The method of moments gives the estimator as the $\widehat{\theta}$ solving the equation:

$$\widehat{\mu}'_k = m'_k$$

for $k = 1, , t$ where t is the number of parameters of the family f.
It is an estimator easy to compute and it provides good estimates, whenever the empirical distribution of the samples converges in some sense to the probability distribution (see [MMP]).

Bayes Estimates

We are still in the general case of the data $y_1, ..y_n$ drawn from $f(y|\theta)$ and the aim is to find an estimator for θ. The first step of the method would be to select a prior distribution, $\pi(\theta)$, expressing ours beliefs and uncertainty about θ.
After gathering data we can compute posterior probability distribution that update the original beliefs. Posterior distribution, which is defined as the conditional distribution of θ:

$$\pi(\theta|Y_1, ..Y_n) = \frac{f(\theta, Y_1, ..Y_n)}{f(Y_1, ..Y_n)} = \frac{L(Y_1, ..Y_n|\theta)\pi(\theta)}{f(Y_1, ..Y_n)} \tag{3.12}$$

where $L(Y_1, ..Y_n|\theta)$ is the likelihood function. This equation is the implementation of the Bayes theorem for θ. Let $C = \frac{1}{f(Y_1,..Y_n)}$ then the posterior distribution becomes: $\pi(\theta|Y_1, ..Y_n) = CL(Y_1, ..Y_n|\theta)\pi(\theta)$. We continue by simplifying $L(Y_1, ..Y_n|\theta)\pi(\theta)$ then pull all the terms not depending of θ into C (normalizing constant). The terms that remained are forming the kernel of the distribution for θ. The last step is to find out the classical distribution θ belongs to, and to determine its parameters (see [JMW09]).

Maximum Likelihood Estimation

As mentioned before the mixture models can be seen as a missing data problem, as only the observed data (y_t) is available. We created z_t so that for each y_t there is an unknown value of z_t and we can interpret the weights of components as prior probabilities of belonging to a given component: $P\{Z_{tk} = 1\} = \pi_k$. Knowing that our aim is to find $g(x_t; \psi)$ given by Eq.3.11, the next step is to estimate all the parameters present in that definition of the distribution of the mixture model.
Given a n-sample of independent observations from a mixture defined in 3.10, the likelihood function is:

$$\mathcal{L}(y; \psi) = \prod_{t=1}^{T} \left\{ \sum_{k=1}^{K} \pi_k f_k(y_t; \theta_k) \right\} \tag{3.13}$$

The aim of the ML is to maximize this likelihood and so obtain the ML estimator. The specificity of the ML when applied to mixture models is that it maximizes the likelihood of the observed data, and not the complete data i.e. the ML can't be applied straightforward, it requires an iterative procedure. The expectation-maximization (EM) algorithm (developed by Dempster in 1977) is the most common method for doing the ML estimation for the parameter of a mixture distribution. The first theorem in Dempster's article (see [DLR77]) says that each iteration of EM either increases or holds constant the incomplete likelihood. The theorem implies that EM converges to a local maximum of the observed-data likelihood, even if it doesn't guarantee that that there is a global maximum of the incomplete likelihood (e.g. [PIC07]).
The important properties of the ML estimator are: it only depends of the likelihood (function of MLE), maybe simple to compute (for the classical cases), it is consistent and asymptotically unbiased.

The Expectation Maximization algorithm

The objective is to estimate the parameter vector $\psi = (\pi_1, .., \pi_{k-1}, \theta_1, .., \theta_k)$ describing each component density $f_k(y_t, \theta_k)$, where k are the components that generated each data point y_t and π_k are the mixing proportions of the components (see [DLR77]). **The procedure** for doing it is by maximizing the likelihood given by the eq. 3.13.
Returning to our data, we remember that the complete data-set is given by $x = (y, z)$ that the density of the observed data can be written as:

$$g(x; \psi) = f(y; \psi)k(z|y; \psi) \tag{3.14}$$

where $k(z|y;\psi)$ is the conditional density of the missing observations, giving the data. We can though write the likelihood of the complete data as:

$$\log \mathcal{L}^c(x;\psi) = \log \mathcal{L}(y;\psi) + \log k(z|y;\psi)$$

where

$$\log k(z|y;\psi) = \sum_{t=1}^{T}\sum_{k=1}^{K} z_{tk} \log \mathbb{E}\{Z_{tk}|Y_t = y_t\} \tag{3.15}$$

But in accordance to the equation 3.11 we have

$$\log \mathcal{L}^c(x;\psi) = \sum_{t=1}^{T} \log g(x_t;\psi) = \sum_{t=1}^{T}\sum_{k=1}^{K} z_{tk} \log\{\pi_k f(y_t;\theta_k)\} \tag{3.16}$$

.
Let $\psi^{(h)}$ the value of the ψ parameter on step h. The EM algorithm maximizes indirectly the likelihood of the incomplete-data by maximizing $\mathbb{E}\{Z_{tk}|Y_t = y_t\}$, the conditional expectation of the complete-data likelihood:

$$\log \mathcal{L}(y;\psi) = Q(\psi;\psi^{(h)}) \; - \; H(\psi;\psi^{(h)}) \tag{3.17}$$

where

$$\begin{aligned} Q(\psi;\psi^{(h)}) &= \mathbb{E}_{\psi^{(h)}}\{\log \mathcal{L}^c(X;\psi)|Y\} \\ H(\psi;\psi^{(h)}) &= \mathbb{E}_{\psi^{(h)}}\{\log k(Z|Y;\psi)|Y\} \end{aligned}$$

The EM algorithm for fitting a mixture model proceeds to repeat two important steps: estimation (E) and maximization (M) that are framed by two other steps: initialization and choice of the estimator:

1. Choose an initial value $\psi_{(h)}$ for the parameters ψ: any intuition values.

2. **(E)** Compute $Q(\psi;\psi^{(h)})$

3. **(M)** Find a new parametrization $\psi^{(h+1)}$ that maximizes
   $Q: \; \psi^{(h+1)} = \text{Argmax}_{\psi}\{Q(\psi;\psi^{(h)})\}$

4. If the difference $\psi^{(h+1)} - \psi^{(h)}$ barely changed, then stop. Otherwise continue at Step2.

This algorithm will improve the estimate of ψ, increasing the value of Q at every M step until it reaches a local maximum.

3.3 Mixture models in the Extreme Value Theory

Alternatively there are various mixture models fitting all the distribution: a certain model for the bulk of the distribution and a flexible extreme value model for the tails. These models may include the threshold among the parameters to be estimated (estimation parameters methods are the same as for classical mixture methods presented in the Section 3.2.1, see [EAT08]) or use smooth transition functions between the main mode of the distribution and the tails to avoid the choice of the threshold. Below there are some of the methods built in the last few years.

Frigessi and Haug proposed in 2002 (see [FHR02]) a dynamically weighted model, as a mixture of a GPD and a light-tailed density distribution, where the weight function varies over the range of support, in such way that for large values the GPD component is predominant and thus takes the role of threshold selection. In this way the estimation of the threshold is replaced with the estimation of the parameters of the transition function which is done by the maximum likelihood method. This approach may be useful in unsupervised heavy tail estimation and small percentiles.

In 2004 Mendes and Lopes (see [ML04]) propose a mixture model where the main mode of the distribution is assumed to be normal and tails are fitted by two separate GPD models. They use a combination between the maximum likelihood and L-moments methods to estimate the proportion in each tail as well as the threshold. Behrens and Lopes (see [BHG04]) presented in 2004 a mixture model that combines a parametric form (gamma distribution) for observations in the main mode up to some threshold and a GPD for the observations above this threshold (the observations in the tails). In this approach all observations are used to estimate the parameters of the model, including the threshold.

In 2006 Tancredi and Anderson proposed (see [TAO06]) to overcome the difficulties of the fixed-threshold approach by using a model combining piecewise uniform distributions from a known low threshold up to an unknown end point (u, the actual threshold) and a GPD with u as a threshold, for the rest of the distribution, the tails. The threshold is estimated in the same time with the parameter of the model, by bayesian inference.

Recently, Carreau and Bengio (see [CB09]) introduced a conditional mixture model with hybrid Pareto components to approximate distributions with support on the entire $\mathbb{R}$. The hybrid Pareto is a Gaussian whose upper tail has been replaced by a GPD. The heaviness of the upper tail is controlled by a new parameter which is to be estimated along with the location and spread parameters of the Gaussian distribution. A conditional density estimator is built, by modeling those parameters, as functions of some variables giving information over the observations of interest. This functions are implemented using a neural network. The hybrid Pareto can be adapt to multimodality, asymmetry, and heavy tails distributions and have important applications in domains such as finance and insurance.

In 2011 MacDonald and Scarrott (see [MSL+11]) propose a flexible mixture model combining

a non-parametric kernel density estimator below some threshold and a GPD model for the upper tail above the threshold. In this way they avoid choosing a parametric form for the main mode of the distribution and the mixture model has only one more parameter to estimate (kernel bandwidth) than the usual GPD function (plus the threshold) which potentially simplifies computational aspects of the parameter estimation compared to the uniform mixture model of Tancredi and Anderson and mixture of hybrid Pareto distributions of Carreau and Bengio. The uncertainties related to the threshold choice are considered and new perspectives on the impact of threshold choice on the density and quantile estimates are obtained. Bayesian inference with Markov chain Monte Carlo sampling is used to account for all uncertainties and enables inclusion of expert prior information.

Whether mixtures of distributions are employed as a flexible modeling device to estimate densities or are used to model data thought to arise from several populations, they provide an efficient tool to approximate a distribution. Indeed, mixtures of distributions can model multiple modes, different types of skewness, but they can also be employed to classify observations from heterogeneous data sets. In 2011 Evin et al (see [EMP11]) studied mixtures of distributions with normal, gamma, and Gumbel components. Moving away from the standard normal setting, gamma mixtures are developed in order to model strictly positive hydrological data and Gumbel mixtures for extreme variates. Since the data analyzed can exhibit dependency through time, they treat both the independent and dependent cases, where the last one is modeled through a Markov process. A fairly unified approach is adopted for the different distributions and the problem is treated from the bayesian perspective, which enables them to use marginal densities to automatically compare the adequacy of the different models for a given data set. This model-selection framework allows to formally test the relevance of using mixture models by computing the marginal likelihoods of single distribution models and to verify the presence of a persistence in the time series by comparing independent and identically distributed (IID) and Markovian mixture models.

Studying all this work we can see that there is no perfect method, each one of them has its advantages and its inconveniences but they give us indices of what it might fit better with our need when building the mixture model: choosing or not a threshold, fixing or not a parametrical form for the bulk of the distribution, estimating the threshold separately or once with the models parameter.

3.4 The proposed extreme mixture model

This section details the mixture model we propose which describes in the same time the bulk of the distribution and the tails. When we build the mixture model for fitting our forecasting temperature data we take into account its final use in the management of the electric consumption which is to keep the risk of using exceptional means to produce electricity, lower than 1%. For the main mode of the distribution, the forecasting method (see subsection 1.4.1) implemented in

the article CSBIGS, gives good results. We may thus keep this weighted combination of normal distributions, as the kernel density describing the bulk of the distribution.

The second step is now to establish the heaviness of the upper/lower tail. One way to do it, which is consistent with the first part of the study (where the errors are defined by the ranks of the forecasts) and it also provides a clear threshold (no need to estimate it), is to consider as being in the tails all the values given by the extreme ranks (1 and 51 or 1,2 and 50, 51) in the initial EPS. *We remember that a forecast has the rank 1 when it gives the smallest value among the 51 values of the EPS for the same time-step and the same time-horizon.* In the interests of having a larger number of extreme values we will start by taking into account also the the 1st and the 2nd ranks values for the left tail and the 50th and 51st ranks values for the right tail. Based on the results that will be obtained we will decide or not to see what happens if we consider as extreme values only those corresponding to 1 and 51. The observations selected as extreme are supposed to follow GEV functions $G(\xi, \sigma, \mu)$ which might be different for the left and the right tails.

Now that we established what are the values in the upper/lower tail, we can consider the mixture model for our distribution of independent observations, $X = \{x_{ij(k)} | i = 1, .., N,\ j = 5, .., 14,\ k = 1, 2, 50, 51\}$ where i is the time-step, j is the time-horizon, and k is the statistical rank of the forecast. So the mixture model may be defined by the distribution function F as it follows:

$$F(x_{ij(k)}, \xi, \sigma, \mu, X) = \begin{cases} G_1(x_{ij(k)}, \xi_1, \sigma_1, \mu_1) & \text{if} \quad (k) = 1 & \xi \in \{\xi_1, \xi_2\} \\ F_2(x_{ij(k)}, X) & \text{if} \quad (k) \in \{2, .., 50\} & \sigma \in \{\sigma_1, \sigma_2\} \\ G_3(x_{ij(k)}, \xi_2, \sigma_2, \mu_2) & \text{if} \quad (k) = 51 & \mu \in \{\mu_1, \mu_2\} \end{cases} \tag{3.18}$$

By developing the G_1, F_2, G_3 functions as we decide them above, F becomes:

$$F(x_{ij(k)}) = \begin{cases} \exp\left\{ -\left[1 + \xi_1 \left(\frac{x_{ij(k)} - \mu_1}{\sigma_1}\right)\right]^{-\frac{1}{\xi_1}} \right\} & \text{if} \quad (k) = 1 \\ x_{ij(k)} + \omega \times \left[\mu_{prev}(i) + \exp(\nu_{prev}(i))\ \varepsilon_{ij(k)}\right] & \text{if} \quad (k) \in \{2, .., 50\} \\ \exp\left\{ -\left[1 + \xi_2 \left(\frac{x_{ij(k)} - \mu_2}{\sigma_2}\right)\right]^{-\frac{1}{\xi_2}} \right\} & \text{if} \quad (k) = 51 \end{cases} \tag{3.19}$$

where:

- $\varepsilon_{ij(k)}$ is the normalized prediction error, $\varepsilon_{ij(k)} \sim\ \mathcal{N}(0,\ 1)$
- $\omega = p_k = Pr[\mathbf{x}_i^* = \mathbf{x}_{i,j,(k)}]$ is the probability that $\mathbf{x}_{i,j,(k)}$ gives the best forecasts among the $K = 51$;
- μ_{prev} are the values of the errors, predicted by a linear regression model $M1$ as described below;

- ν_{prev} are the log of absolute values of the residues of the $M1$ pattern, predicted by a linear regression model $M2$ as described below;
- the statistical rank doesn't intrude directly at that moment of the study, it intervenes indirectly in the creation of the $M1$ and $M2$ patterns .

The $M1$ pattern explains the prediction error by the initial forecast, day-position within the year **and the statistical rank,** τ_t:

$$\mu_{prev} = \alpha_1 \cdot \mathbf{x}_i + \sum_{p=1}^{3} [\alpha_{2,p} \cdot a(p) + \alpha_{3,p} \cdot b(p)] + \alpha_4 \cdot \tau_i \tag{3.20}$$

The $M2$ pattern explains the log of the absolute value of the residuals of the $M1$ pattern i.e. ν_{prev}, by the temperature, day-position within an year **and the statistical rank,** τ_t:

$$\nu_{prev} = \beta_1 \cdot \mathbf{x}_i + \sum_{p=1}^{3} [\beta_{2,p} \cdot a(i) + \beta_{3,p} \cdot b(i)] + \beta_4 \cdot \tau_i \tag{3.21}$$

where:
$a(i)$ and $b(i)$ are the coefficient of Fourier series used to decompose the function giving the day-position within an year: $a(i) = \cos\left(\frac{2\pi i \times \text{day}(i)}{365}\right)$, $b(i) = \sin\left(\frac{2\pi i \times \text{day}(i)}{365}\right)$;

Once the mixture model is built, the next step is the estimation of the parameters of the GEV functions (the parameters of the combinations of normals, are computed in the same time when we compute the forecasts, as function of initial predictions of temperature, day-position within an year and the statistical rank). For computing the parameters estimation we use the `evd` package from R software (see [STE12]). In this package the method used to estimate the GEV parameters is the maximum-likelihood (see 3.2.1) and its expectation-maximization algorithm.

In the next section we will implement the mixture model built as presented above. We will be confronted to choices of grouping data (by tail, by time-horizon, by season or by month), function of the shape parameter estimation we may obtain.

3.5 Implementation of the Extreme Value Theory on Temperature Forecasts Data

3.5.1 Context and Data

We are working on the same data as in the implementation of the BMM method presented in the CSBIGS article (see Chapter 2), which are the daily average temperatures in France (the weighted mean of 26 values of daily means temperatures observed by the Meteo stations in France hourly) from March 2007 to March 2011 and the predictions given by Meteo France for the same

period as an EPS containing 51 forecasts by day, up to 14 time-horizons The predictions are given as an Ensemble Prediction System (EPS) : 51 forecasts by day, up to 14 time-horizons corresponding to 14 days (built as described in the data description section).

We implemented for this period the statistical Best Member method in order to improve forecasts starting with 5 days ahead. Once we have created the model, we made simulations and we obtained 10 time more values by day ($51 * 10$). Comparing the scores obtained on the simulated forecasts with the scores for the initial EPS, we could notice an improvement of the quality of the spread forecasts for the same performance, for the central part of a distribution. This is not the case for the tails of the distribution, we are thus interested in modeling the extreme values separately.

One of the important questions when applying the Extreme Value Theory (EVT) is the number of the extreme values included for estimating the parameters of the extreme functions - a small number of values could worsen the quality of the estimators. According to the GEV theory the most current methods of selecting extreme values are the block maxima and the Peaks Over Thresholds (see the section 3.1). But we are in a special case where we need to find a criterion which will be consistent with the first part of the study where the errors are defined by the ranks of the forecasts. Following the same principle of statistical order of the EPS members we decide to consider as extreme values the forecasts having the smallest (1) and the greatest (51) statistical ranks among the forecasts. For the interests of having a larger number of extreme values we start by taking into account also the 2nd and the 50st ranks. Based on the results we obtain we decide or not to see what happens if we consider as extreme values only those corresponding to 1 and 51.

We take into account all the values corresponding to the time-horizons from 5 to 14 and we make adjustment tests (Kolmogorov-Smirnov) to decide if we can consider that it allows us to observe if there are significant changes in the parameter estimation when they are computed separately by board (1, 2, 50, 51), by time-horizon and/or by time (month, season).

3.5.2 Choices of extreme parameter values

We keep the same notation as in the implementation of the BMM so let $X = \{x_{ijk} | i = 1, .., n,\ j = 5, .., 14,\ k = 1, 2, 50, 51\}$ be the ensemble of forecasts, therefore $\mathbf{x}_{ijk}$ is the forecasts given by the scenario k for the time-step i and time-horizon j. Let $\mathbf{y}_n$ be the temperature we are forecasting. As in the first part of the thesis we work on the forecasting errors $\varepsilon_{ij(k)} = \mathbf{y}_t - \mathbf{x}_{ij(k)}$, that we will standardize to remove seasonality, so the new ε is computed from the old ones by $\frac{\varepsilon_{ij(k)} - \mu}{\sigma}$ where $\mu = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$ and $\sigma = \sqrt{\text{var}(y_{ij})}$ are obtained from the pattern use in the W-BMM (presented in the CSBIGS article) to modelize the errors for the forecast ranking from 3 to 49 and an GEV function to modelize the rest of it. As we mentioned before the first thing to study is by how many variables can we separate the extreme values in our database when estimating parameter of the GEV function. The cardinal of the ensemble of all the selected extreme values

is $N = \text{card}\{X\} = 3566$ or if we allocate those values by horizon and by rank, we obtain the Table 3.2. In the Appendix we can find the same kind of table giving the number of extreme values by time-horizon and/or by season and by month.

Horizon	**Rank 1+2**	**Rank 50+51**	**All 4 ranks**
5	190	286	476
6	182	212	394
7	162	204	366
8	142	184	326
9	170	176	346
10	168	164	332
11	162	168	330
12	170	176	346
13	160	166	326
14	144	180	324
All	1650	1916	3566

Table 3.2: The number of values included in our study of the extreme values.

We start by looking at what the density function looks like, if we take all the extreme values together, also if we separate them by left ranks (1,2) and right ranks (50,51) (Figure 3.1). On the same graphics we draw the GEV distribution. For the computation part we use the R software and for estimating parameters we use the *fgev* function from the *evd* package in R. This function estimates the maximum-likelihood parameters fitting for the GEV distribution: the location parameters μ, the scale parameter σ and the shape parameter ξ. We notice that among the four presented cases, only in the case of the right tails the computed GEV function does not fit well the data: taking more criteria into account is necessary. The KS tests with $H0$: *the observed distribution and the computed GEV distribution are the same* say the same thing: we reject only for the Rank (50+51). So from the "fitting GEV" point of view given by the first two graphics we could study all the extreme values together. But when we make directly KS tests with $H0$: *the left extreme values and the right extreme values belong to the same distribution* they are rejected for the ranks $(1+2)+(50+51)$ and for the ranks $(1+2)-(50+51)$. This gives us a first clue, we should follow the classical rule in the extreme literature: separate the hot extreme values from the cold ones.
We don't forget that among all the extreme values, right and left, we have 10 different time-horizons. We now separate those values by time-horizons (Figure 3.2) and then also by left and right ranks.
When separating them only by time-horizon we notice the densities shapes look alike but theirs mean looks divided between 0 and 1. When separating them also by hot and cold we notice that the densities of the left values, for different horizons are much more alike than the densities

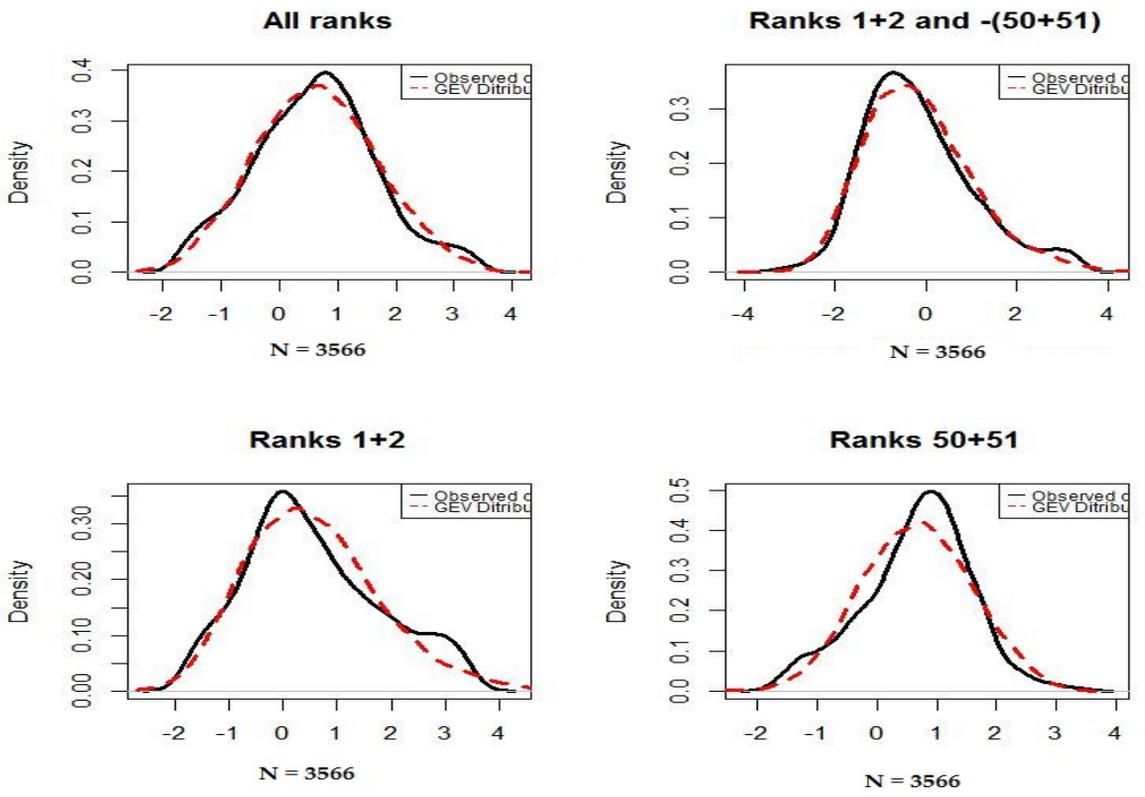


Figure 3.1: Errors densities and empirical distributions of all the extreme values included, all ranks included (rank 1 and rank 2 with rank 50 and rank 51, and also rank 1 and rank 2 with -rank 50 and -rank 51) and left ranks and right ranks separated.

of the different horizons for the right values. This tell us that for the left tail we can study all the time-horizon all-together but not for the right tails. That confirms our previous conclusions concerning the separation of the right and left tails. We group the left and the right ranks in two tails as follows:

$B = B1$ when the rank $(k) \in \{1, 2\}$,

$B = B51$ when the rank $(k) \in \{50, 51\}$.

We made other tests to see if mixing cold values and the opposite of the hot values makes sense, and we further see what final results such a mixture is giving. To verify the information given by the two density graphics concerning the possibility of studying all the time-horizons together we make adjustment (Kolmogorov-Smirnov (KS)) tests at the 5% significance level. The results tell us if rather yes or not we can consider the extreme values of all the horizons as coming from the

same distribution. One by one and one at the time we compare the distribution of one-horizon-values with the distribution of the other horizons values, by KS tests with the hypothesis:

$H0$: the extreme values corresponding to the time-horizon i and the values corresponding to the time-horizon j (with $i \neq j$) belong to the same distribution.
$H1$: the extreme values corresponding to the time-horizon i and the values corresponding to the to the time-horizon j (with $i \neq j$) not belong to the same distribution.

The tests are made for the tail $B1$ and $B51$ separately and take it all together: $B1 + B51$ and $B1 + (-B51)$. We study the $C_{10}^2 = 45$ tests results for the tail $B1$ and tail $B51$ separately than for the two tails together, and also for the $B1$ and $-B51$. In the Table 3.3 we can see for the $B1$ we reject 11 times (including 8 times that are very close to 0.05) the $H0$ hypothesis ($H7 - H11$, $H8 - H13$, $H11 - H13$), much more for the $B51$: 25 times. We can consider that from the adjustment tests point of view it certainly makes sense to have a single GEV function for the left tail, for all the time-horizons, as suggested the densities graphics in the Figure 3.2. This is not that obvious for the right extremes values: to make a decision we implement a different adjustment test, at the 5% significance level, with hypothesis:

$H0$: the extreme values corresponding to the time-horizon i and the values corresponding to all the time-horizons, other than i belong to the same distribution.
$H1$: the extreme values corresponding to the time-horizon i and the values corresponding to all the time-horizons, other than i not belong to the same distribution.

The results (given in table 3.4) support our decision to consider all the horizons together for the left tail $B1$ and separately for the right board $B51$.
Other way of making this KS-test is to compare the values coming from one time-horizon with the values coming from all the others time-horizons: $H0$: the extreme values from the time-horizon i are coming from the same distribution that the extreme values of $\sum_{j \neq i} j$. We can find the results in table 3.4 and is saying that for $B51$ $H0$ is rejected for all the horizons excepting horizons 6, 8 and 14. For $B1$ $H0$ is not rejected and that is for all the horizons excepting horizons 11 and 13. This results confirm the test horizon by horizon.
The conclusions of the tests - to see after what values we should separate our extreme data values to compute the GEV parameters - made until now are:

- separation by time-horizons only on the right tail
- by tails.

We also study what brings a separation by season, when seasons are defined from an electrical consumption point of view:

Spring when the month $\in$ April, May,

Summer when the time-horizon $\in$ June, July, August,

Autumn when the time-horizon $\in$ September, October,

$Winter$ when the time-horizon $\in$ November,December, January, February, March

When separating the left tail by season, and computing the corresponding GEV parameters we find the distributions presented in Figure 3.3. The GEV are fitting close enough the four distributions respectively, and the corresponding adjustment tests do not reject, at 5% significance level, the $H0$ hypothesis saying that the observed distribution and the GEV distribution are the same.
When separating the right tail by season, and computing the corresponding GEV parameters we find the distributions presented in Figure 3.4. The GEV are fitting close enough the summer and autumn distribution, less close the spring distribution and even less close the winter distribution. The corresponding adjustment tests reject the $H0$ hypothesis saying that the observed distribution and the GEV distribution are the same, only for the winter values.
When we compute the parameters of the GEV estimated by season, we get the values in the table. 3.5. An important thing to notice is that the shape parameter is in all the cases negative, corresponding to a GEV distribution of Weibull type.
We can also take a look at how is evolving the parameters of the GEV function, mainly the shape parameter, when computed by season (see Figure 3.5). We can see that the shape parameter computed to the right extreme values has a small variation from a season to another and it is always negative.
We do the same think for the left tails (see Figure 3.6). We can see that the shape parameter has a even smaller variation from a season to another, it is always negative and they are also very closed to the MLE value computed for all the seasons together.
We can see that the GEV parameters do not change significantly from a season to another but it removes any seasonality might be left after normalizing our series so it might be interesting to see what results this model may give.**We consider it as one of our final choice for the simulation and mixture model: the separation by season and by tail. Results are presented in the next Section**.

For the right tails we also study what brings a separation by groups of time horizons as follows (the results for the left tails are available in the Appendix, Figure 4.1, but wasn't necessary as we decided that the left values from all the time-horizons can be considered as coming from the same distribution, and may be considered together for estimating the GEV function) :

$Pack1H$ when the time-horizon $i \in \{5, 6, 7\}$,

$Pack2H$ when the time-horizon $i \in \{8, 9, 10\}$,

$Pack3H$ when the time-horizon $i \in \{11, 12, 13, 14\}$

We can see in Figure 3.7 the errors densities and their corresponding GEV distributions for the right extreme values for three packages of horizons. The first package, with the 5th, 6th and 7th horizon is the one fitting less good the distribution. For this package only we have a $p - value < 0.05$ of the KS test, so the $H0$ (the observed distribution and the computed GEV distribution are the same) is rejected. This means that for the right tail we can not consider

estimating together values coming from the same time-horizon neither, we have to estimate it for each horizon separately. **This is be the second choice of the final simulation and the results are presented in the next section**.

Another criterion for choosing the values after which we should separate the extreme values, is the Akaike Criterion (AIC) crossed to the asymptotic condition in the GEV, of the number of values by estimation class. The AIC makes a compromise between the complexity of a pattern (number of parameters) and its performance. In the Figure 3.8 we have all the possible combinations of parameters, and the corresponding AIC. From the AIC point of view the best models is the 15th but it is corresponding to a very large number of parameters as it separates the extreme values by day, this would make very few values for every computation of the parameter of the GEV so the asymptotic EVT condition is not respected. The next model is the 16th and it separates the extreme values by Package of time-horizon, by month and by rank, where we have the same non-respect of the number of extreme values not large enough. The first model respecting the two criteria is the model number 18 in fig 3.8 which supposes a **separation by month and by time-horizon package** of the extreme values. We select the 18th model as the third choice to make the simulations within a mixture model in the next section.

3.5.3 Mixture model and criteria of comparison of the final distributions

As mentioned in the section before we build three mixture models that we use to simulate other temperature forecasts, for the time-horizons for 5 to 10, for all the time period. The models are the one we have chosen in the section before:

1. a mixture model between the BMM for the central part of the distribution, and GEV functions, with parameters computed by tail and by season (the classical choice in the GEV, not rejected in our case by the tests we made).

2. a mixture model between the BMM for the central part of the distribution, and GEV functions, with parameters computed by tail, and for the right tail by time-horizon also (the choice to which bring us the tests).

3. a mixture model between the BMM for the central part of the distribution, and GEV functions, with parameters computed by tail, by package of time-horizons, by month (the choice indicated by the AIC).

Mixture Models with the GEV parameters estimated by tail and by season

Let us remember that the pattern for the central part of the distribution is the one build by the BMM method (that we can find in CSBIGS article) with the difference that we keep the same number of daily simulations as for the initial EPS. As for the pattern for the two tails, we built it as two separated GEV functions, one for left tail and one for the right tails, and for each of

them we estimate the parameters by season. Studied made before showed that those GEV are a Weibull type, which is rather classical for the temperature forecasts.

After making the simulations with this new mixture model we compute the criteria to compare it with the initial EPS of forecasts. In table 3.6 we have the scores for the mixture model and in table 3.7 we have the scores for the initial EPS. In the appendix we can find the corresponding table 4.2 for the BMM studied in the first part of the thesis.
We can see when comparing the two tables that we improve the CRPS component giving the overall precision of the forecasts, for all the horizons as for the other scores we loose a little bit in precision.
We also draw the Talagrand Diagram, we can see in the Figure 3.9 the ones for the horizons 5, 6, 7 with the corresponding diagrams of the initial MF forecasts. We notice that the diagrams for the mixture model are slightly more flat than the ones of the MF forecasts but we can notice that the extreme ranks are badly estimated. This is the case for all the horizons (figures are in the Appendix).

Mixture Models with the EVT parameters estimated by tail and for the right tail, by time-horizon.

We can see in table 3.8 the scores for the simulated forecasts obtained with this second mixture model. We notice that the scores are very close to the ones obtained with the first mixture model we built: we improve the CRPS component giving the overall precision of the forecasts, for all the horizons as for the RMSE, it is larger, meaning that we have more large errors (we remember that the main difference between CRPS and RMSE is that the second one puts greater influence on large errors).
We also draw the Talagrand Diagram, we can see in the Figure 3.10 the ones for the horizons 5, 6, 7 with the corresponding diagrams of the initial MF forecasts. As for the case before we notice that the diagrams for the mixture model are more slightly flat the ones of the MF forecasts.

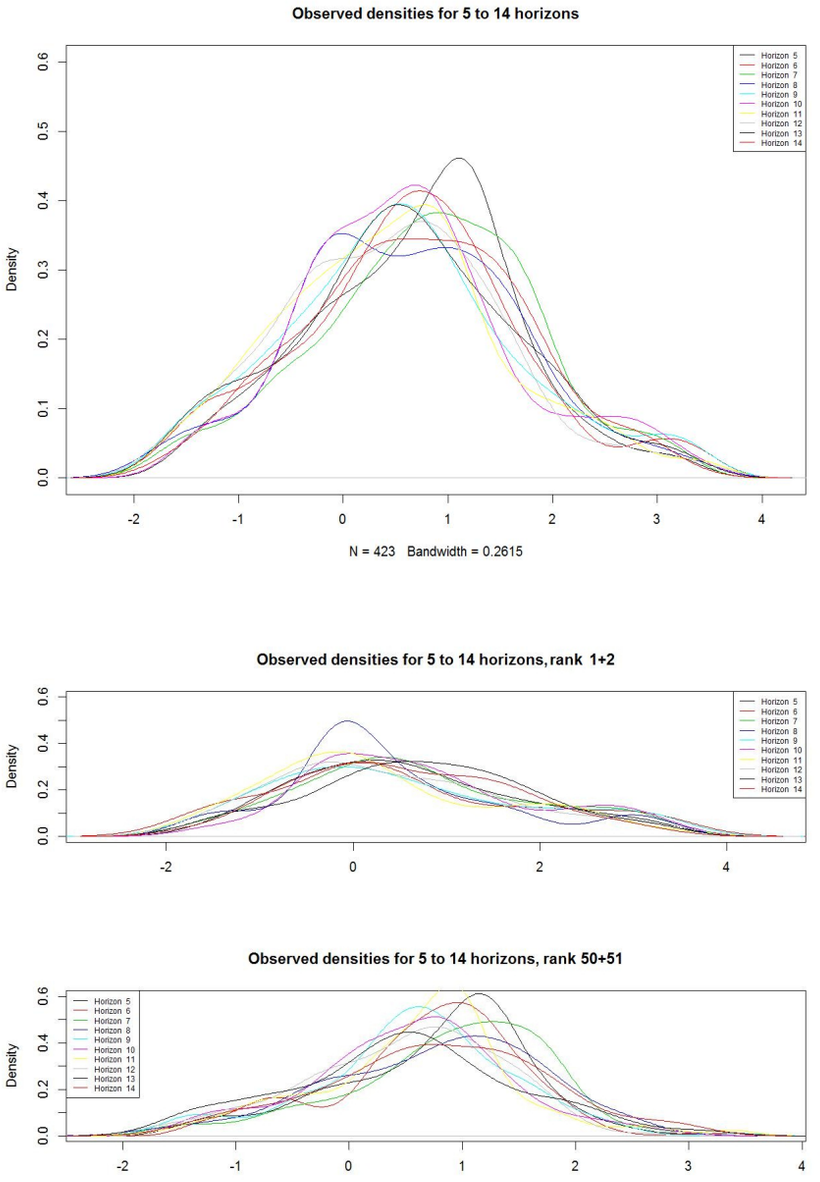


Figure 3.2: At the top: Errors densities by time-horizon, all ranks included. We notice the densities shapes look alike but theirs mean are divided between 0 and 1. At the bottom errors densities by time-horizon, by ranks, separating left and right. We notice that the densities of the left values, for different horizons are much more alike than the densities of the different horizons for the right values.

Horizon i	**Horizon j**	$p-value$ **B1**	$p-value$ **B51**	$p-value$ **B1+B51**	**B1+(-B51)**
5	6	0.3223	0.0431	0.1408	0.4795
5	7	0.8268	0.0362	0.0527	0.2005
5	8	0.0739	0.19	0.1593	0.3163
5	9	0.6529	**2e-04**	**0.0024**	**0.0024**
5	10	0.3208	**1e-04**	**0.0043**	**0**
5	11	0.026	**0.0001**	**0.0002**	**0.0017**
5	12	0.2246	**0.0001**	**0.0014**	**0.0057**
5	13	0.3495	**0**	0.0169	**0**
5	14	0.9916	0.2213	0.5523	0.3888
6	7	0.1703	0.0115	0.0774	0.1798
6	8	0.4905	0.2101	0.5556	0.4707
6	9	0.8207	0.082	0.2885	0.0438
6	10	0.1475	0.0524	0.4038	**4e-04**
6	11	0.2411	0.1264	0.1123	0.0541
6	12	0.6887	0.0405	0.0817	0.0403
6	13	0.0221	**0.0042**	0.7748	**0.0006**
6	14	0.3928	0.1337	0.383	0.6119
7	8	0.0223	0.3436	0.0144	0.2397
7	9	0.2008	**0**	**6e-04**	**0.0015**
7	10	0.5229	**0**	**2e-03**	**0**
7	11	**0.0076**	**0**	**0.0001**	**0.0011**
7	12	0.0774	**0**	**0.0008**	**0.0036**
7	13	0.5158	**0**	**0.0049**	**0.0001**
7	14	0.7988	0.0755	0.3836	0.4607
8	9	0.3728	0.0108	0.3737	0.0263
8	10	0.05	**0.0019**	0.2158	**0.0015**
8	11	0.0868	**0.001**	0.0345	0.0129
8	12	0.1949	**0.0072**	0.1822	0.0489
8	13	**0.0014**	**0.0006**	0.6699	**0.0006**
8	14	0.1684	0.8527	0.3478	0.6783
9	10	0.0417	0.875	0.3306	0.0505
9	11	0.4213	0.5543	0.7605	0.6616
9	12	0.8202	0.6059	0.5713	0.9659
9	13	0.0326	0.3374	0.957	0.0388
9	14	0.4363	0.0375	0.0772	0.094
10	11	0.0128	0.4764	0.1368	0.0124
10	12	0.0539	0.8594	0.1597	0.1114
10	13	0.678	0.6426	0.5151	0.4041
10	14	0.3931	**0.008**	0.0487	**0.0049**
11	12	0.5285	0.4818	0.9478	0.5245
11	13	**0.0004**	0.0953	0.2506	**0.003**
11	14	0.0262	**0.002**	**0.004**	0.0277
12	13	0.0158	0.5542	0.2747	0.0786
12	14	0.1846	0.0347	0.0517	0.1524
13	14	0.2451	0.0271	0.2543	**0.0053**

Table 3.3: The p-values for the KS test, at the 5% significance level with $H0$: the values of the time-horizon i and the ones from the time-horizon j (with $i \neq j$) belong to the same distribution.

horizon	$p-values$ $B51$	$p-value$ $B1$
5	0	0.725
6	0.283	0.301
7	0	0.16
8	0.073	0.096
9	0.039	0.357
10	0.012	0.069
11	0.009	0.013
12	0.022	0.21
13	0.001	0.006
14	0.118	0.618

Table 3.4: The KS $p-value$ tests, when taking out values from each horizon, one at the time and comparing the distribution of what is resting with the distribution of what we took out. For $B51$ $H0$ is rejected, at a 5% significance level for all the horizons excepting horizons 6, 8 and 14. For $B1$ $H0$ is not rejected and that is for all the horizons excepting horizons 11 and 13. This results confirm the test horizon by horizon.

Tail	season	location	scale	shape	σ(location)	σ(scale)	σ(shape)
all	spring	0.043	1.013	-0.166	0.052	0.037	0.032
all	summer	0.1	1.054	-0.177	0.042	0.03	0.026
all	autumn	0.25	1.05	-0.228	0.057	0.039	0.031
all	winter	0.355	0.994	-0.283	0.029	0.02	0.013
B1	spring	-0.008	1.127	-0.097	0.087	0.064	0.063
B1	summer	0.02	1.106	-0.137	0.055	0.04	0.037
B1	autumn	0.035	1.134	-0.143	0.083	0.06	0.053
B1	winter	0.119	1.072	-0.153	0.078	0.056	0.051
B51	spring	0.115	0.875	-0.288	0.06	0.041	0.026
B51	summer	0.256	0.885	-0.232	0.06	0.04	0.028
B51	autumn	0.566	0.765	-0.345	0.063	0.043	0.033
B51	winter	0.398	0.959	-0.298	0.03	0.021	0.011

Table 3.5: The estimations parameters by season, to notice the shape parameter that is always negative, corresponding to a GEV distribution of Weibull type.

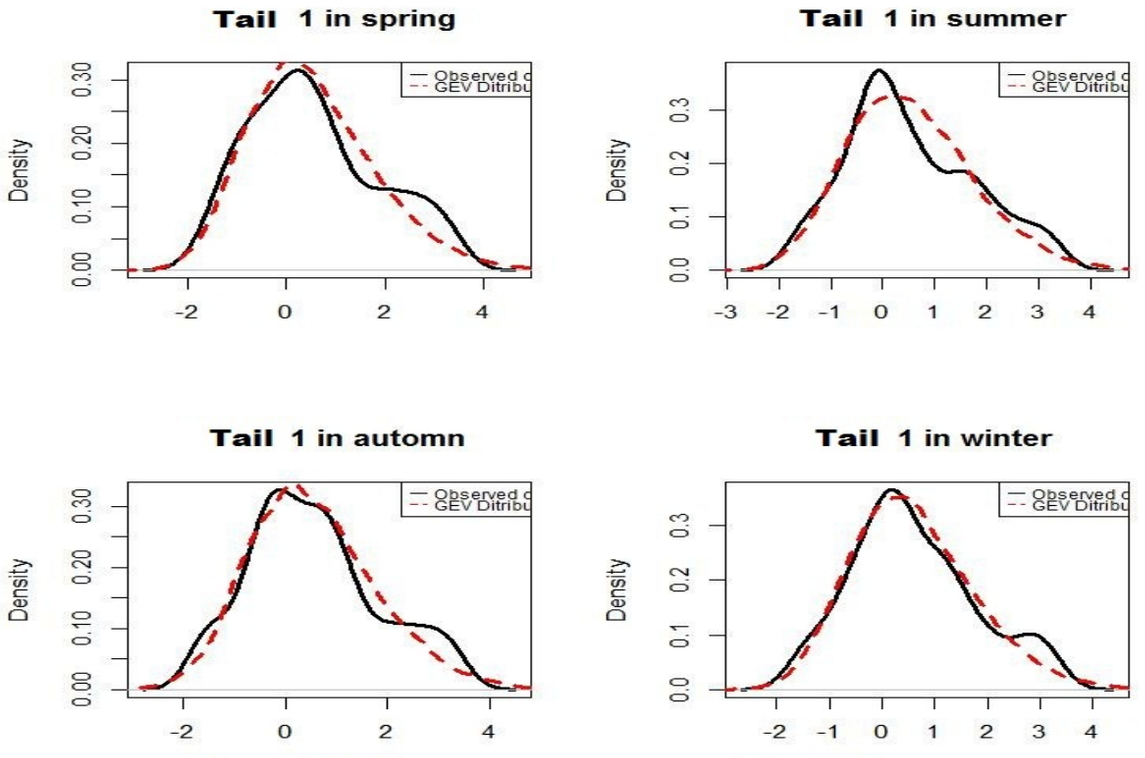


Figure 3.3: Errors densities and GEV distributions of the left extreme values for the four season.

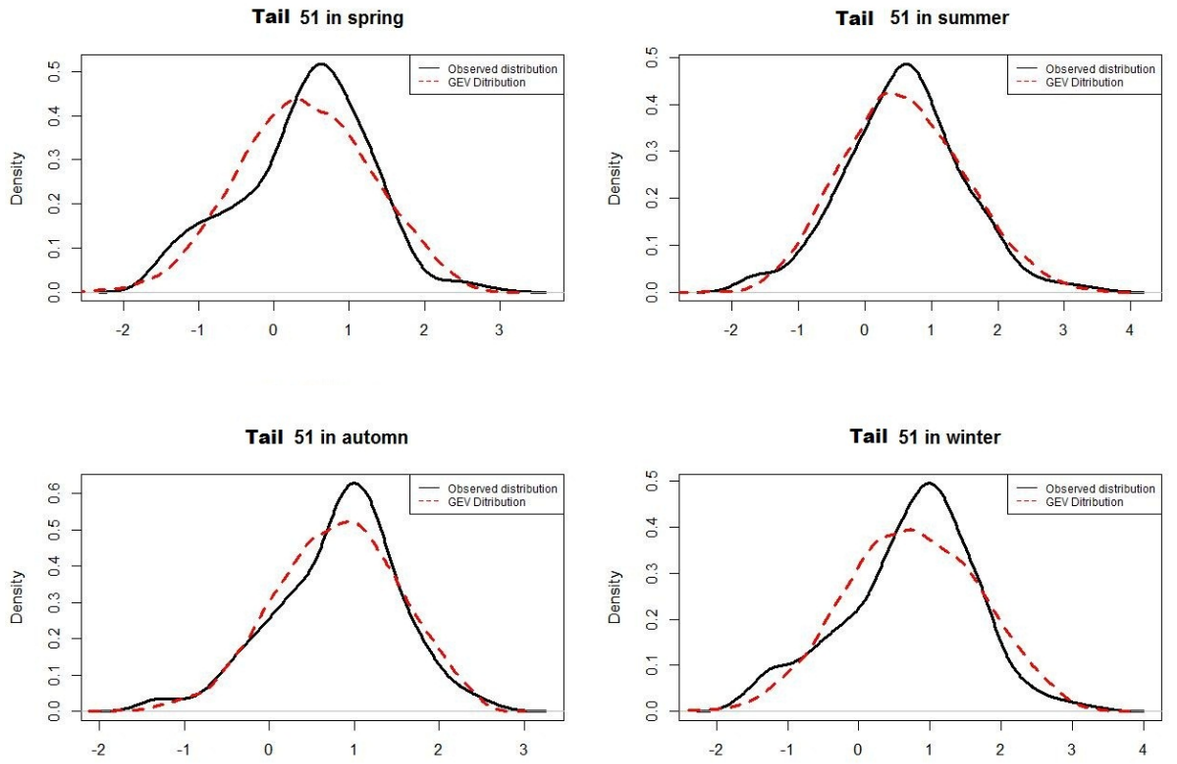


Figure 3.4: Errors densities and GEV distributions of the right extreme values for the four season.

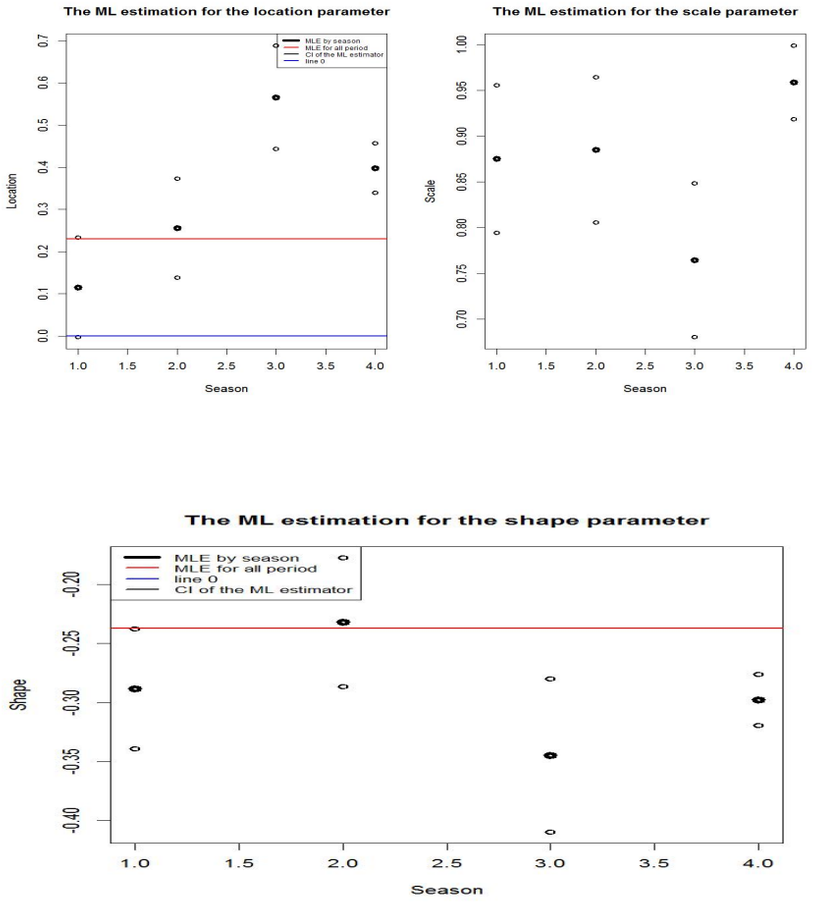


Figure 3.5: The GEV parameters for the values coming from the $B51$. We can see that the shape parameter (the most important as it decides of the case of the GEV function we are in) has a small variation from a season to another and it is always negative.

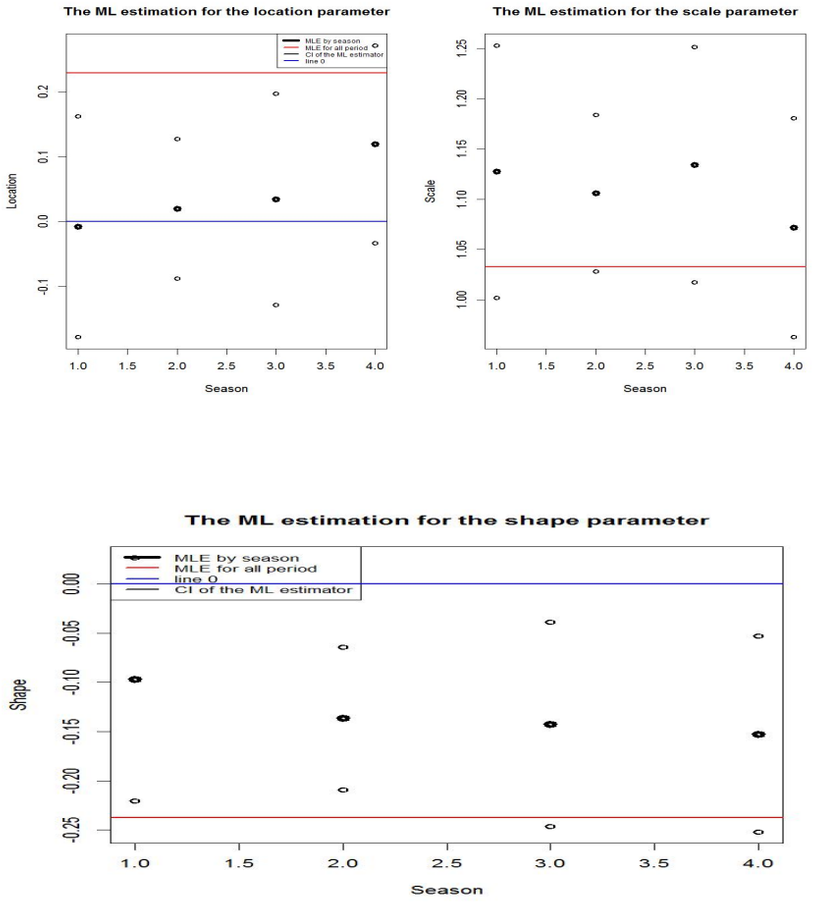


Figure 3.6: The GEV parameters for the values coming from the $B1$. We can see that the shape parameter (the most important as it decides of the case of the GEV function we are in) has a very small variation from a season to another and it is always negative.

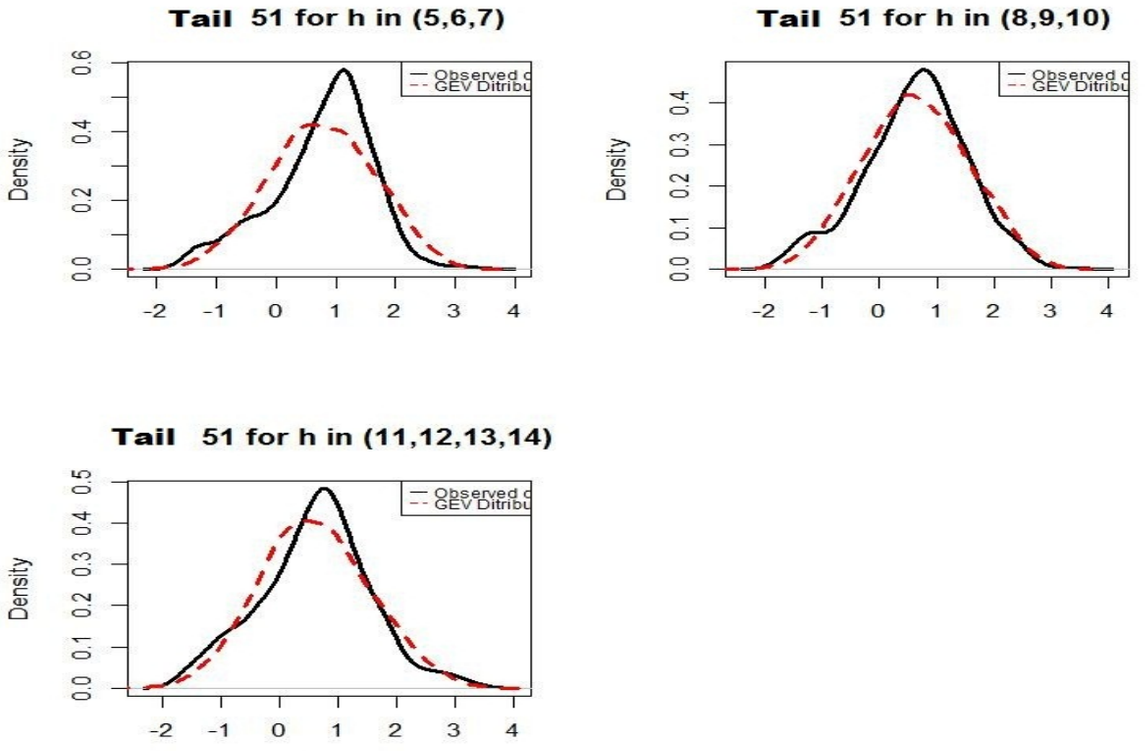


Figure 3.7: The errors densities and their corresponding GEV distributions for the right extreme values for three packages of horizons. The first package, with the 5th, 6th and 7th horizon is the one fitting less good the distribution.

	AIC	rank(AIC)	BIC	rank(BIC)	sumlogliks	npars	horizon	temps	bords	nparmax
1	4541.325	22.0	4701.217	10.0	-2240.6624	30	10	568	4	68160
2	9150.000	33.5	33533.605	33.5	0.0000	4575	10	568	2	34080
3	4332.298	7.0	4396.255	1.0	-2154.1492	12	10	568	1	17040
4	9150.000	33.5	33533.605	33.5	0.0000	4575	10	12	4	1440
5	4377.228	9.0	13459.121	26.0	-484.6140	1704	10	12	2	720
6	4549.969	23.0	4741.840	12.0	-2238.9844	36	10	12	1	360
7	4527.538	17.0	4591.495	8.0	-2251.7690	12	10	4	4	480
8	4461.949	13.0	4493.928	2.0	-2224.9745	6	10	4	2	240
9	9150.000	33.5	33533.605	33.5	0.0000	4575	10	4	1	120
10	4535.837	18.5	4583.805	6.5	-2258.9185	9	10	1	4	120
11	4539.057	20.0	4555.047	5.0	-2266.5287	3	10	1	2	60
12	9150.000	33.5	33533.605	33.5	0.0000	4575	10	1	1	30
13	9150.000	33.5	33533.605	33.5	0.0000	4575	3	568	4	20448
14	9150.000	33.5	33533.605	33.5	0.0000	4575	3	568	2	10224
15	3864.093	1.0	10355.728	25.0	-714.0463	1218	3	568	1	5112
16	4087.070	2.0	7652.672	24.0	-1374.5349	669	3	12	4	432
17	4328.398	26.0	6539.708	21.0	-1950.4989	360	3	12	2	216
18	4093.452	3.0	6635.743	23.0	-1569.7260	477	3	12	1	108
19	4473.671	15.0	5752.810	20.0	-1996.8353	240	3	4	4	144
20	4571.719	24.0	5211.289	18.0	-2165.8597	120	3	4	2	72
21	4337.857	8.0	4977.427	15.0	-2048.9283	120	3	4	1	36
22	4460.854	12.0	4780.639	13.0	-2170.4268	60	3	1	4	36
23	4541.325	21.0	4701.217	9.0	-2240.6624	30	3	1	2	18
24	6846.000	30.0	25089.733	30.0	0.0000	3423	3	1	1	9
25	5532.000	28.5	20274.088	28.5	0.0000	2766	1	568	4	6816
26	5532.000	28.5	20274.088	28.5	0.0000	2766	1	568	2	3408
27	4319.626	5.0	6542.132	22.0	-1742.8132	417	1	568	1	1704
28	4468.839	14.0	5588.087	19.0	-2024.4197	210	1	12	4	144
29	4593.005	25.0	5168.618	17.0	-2188.5025	108	1	12	2	72
30	4320.436	6.0	5087.920	16.0	-2016.2178	144	1	12	1	36
31	4431.061	10.0	4814.803	14.0	-2143.5305	72	1	4	4	48
32	4515.743	16.0	4707.614	11.0	-2221.8716	36	1	4	2	24
33	4313.859	4.0	4505.730	3.0	-2120.9295	36	1	4	1	12
34	4454.145	11.0	4550.080	4.0	-2209.0724	18	1	1	4	12
35	4535.837	18.5	4583.805	6.5	-2258.9185	9	1	1	2	6
36	5228.962	27.0	18228.222	27.0	-175.4812	2439	1	1	1	3

Figure 3.8: From the AIC point of view the best models is the 15th but it is corresponding to a very large number of parameters as it separates the extreme values by day. The estimation parameters classes does not respect the condition of the number of values large enough. The first model respecting it is the 18th and it separates the extreme values by Package of time-horizon, by month and by rank.

Horizon	CRPSpot	CRPSReli	CRPS	MAE	RMSE	R2
5	0.5841	0.0050	0.5891	1.257	1.691	0.967
6	0.7367	0.0038	0.7406	1.553	2.080	0.951
7	0.9044	0.0042	0.9086	1.890	2.502	0.931
8	1.0481	0.0037	1.0518	2.171	2.852	0.912
9	1.1760	0.0035	1.1795	2.420	3.156	0.895
10	1.2872	0.0040	1.2912	2.653	3.434	0.878
11	1.3730	0.0037	1.3767	2.819	3.644	0.866
12	1.4183	0.0047	1.4230	2.943	3.789	0.854
13	1.471	0.0045	1.4761	3.023	3.873	0.846
14	1.4789	0.0046	1.4835	3.075	3.949	0.842

Table 3.6: The different scores by time-horizon, for the forecasts obtained when creating a mixture model: BMM for the central part of the distribution and GEV functions for the tails, having their parameters estimated by tail and by season.

Horizon	CRPSpot	CRPSReli	CRPS	MAE	RMSE	R2
5	0.6286	0.0035	0.6321	1.178	1.555	0.972
6	0.7841	0.0036	0.7877	1.485	1.958	0.956
7	0.9608	0.0047	0.9655	1.814	2.379	0.937
8	1.1080	0.0055	1.1135	2.104	2.740	0.918
9	1.2393	0.0048	1.2440	2.362	3.048	0.901
10	1.3540	0.0048	1.3588	2.586	3.314	0.885
11	1.4343	0.0040	1.4383	2.756	3.521	0.872
12	1.4897	0.0042	1.4939	2.879	3.662	0.861
13	1.5386	0.0042	1.5428	2.980	3.779	0.852
14	1.5645	0.0037	1.5681	3.036	3.845	0.847

Table 3.7: The different scores by time-horizon, for the initial MF forecasts

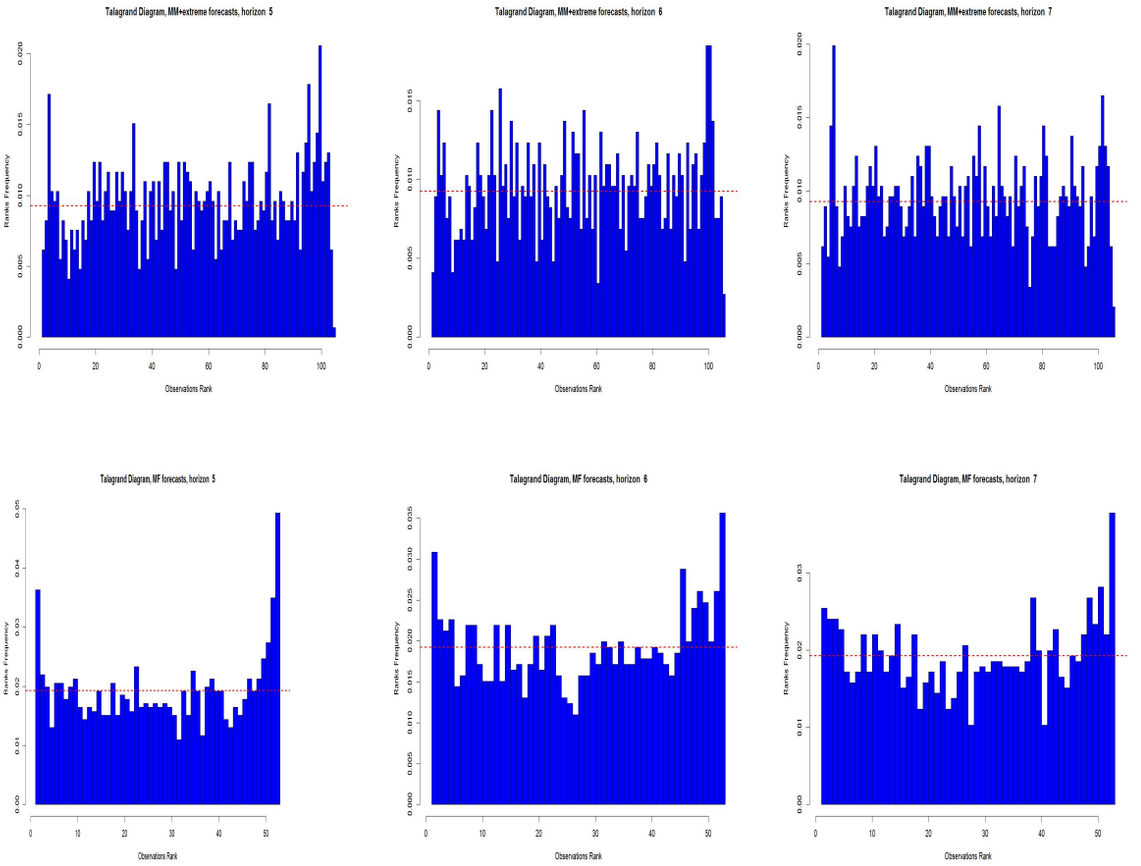


Figure 3.9: Talagrand Diagrams corresponding to the mixture model (with the GEV parameters estimated by tail and by season) on top and the initial predictions at the bottom, for the time-horizons 5, 6 and 7 respectively.

Horizon	CRPSpot	CRPSReli	CRPS	MAE	RMSE	R2
5	0.5843	0.0050	0.5893	1.259	1.695	0.967
6	0.7367	0.0038	0.7405	1.553	2.081	0.951
7	0.9046	0.0043	0.9089	1.895	2.510	0.931
8	1.0478	0.0038	1.0516	2.170	2.852	0.913
9	1.1761	0.0033	1.1794	2.417	3.153	0.895
10	1.2870	0.0040	1.2909	2.652	3.434	0.878
11	1.3727	0.0033	1.3760	2.812	3.634	0.866
12	1.4177	0.0045	1.4223	2.933	3.776	0.855
13	1.4716	0.0044	1.4760	3.021	3.870	0.846
14	1.4790	0.0047	1.4838	3.078	3.955	0.842

Table 3.8: The different scores by time-horizon, for the forecasts obtained when simulating with the mixture model that separates extreme values by tail and for the right tail, by time-horizon.

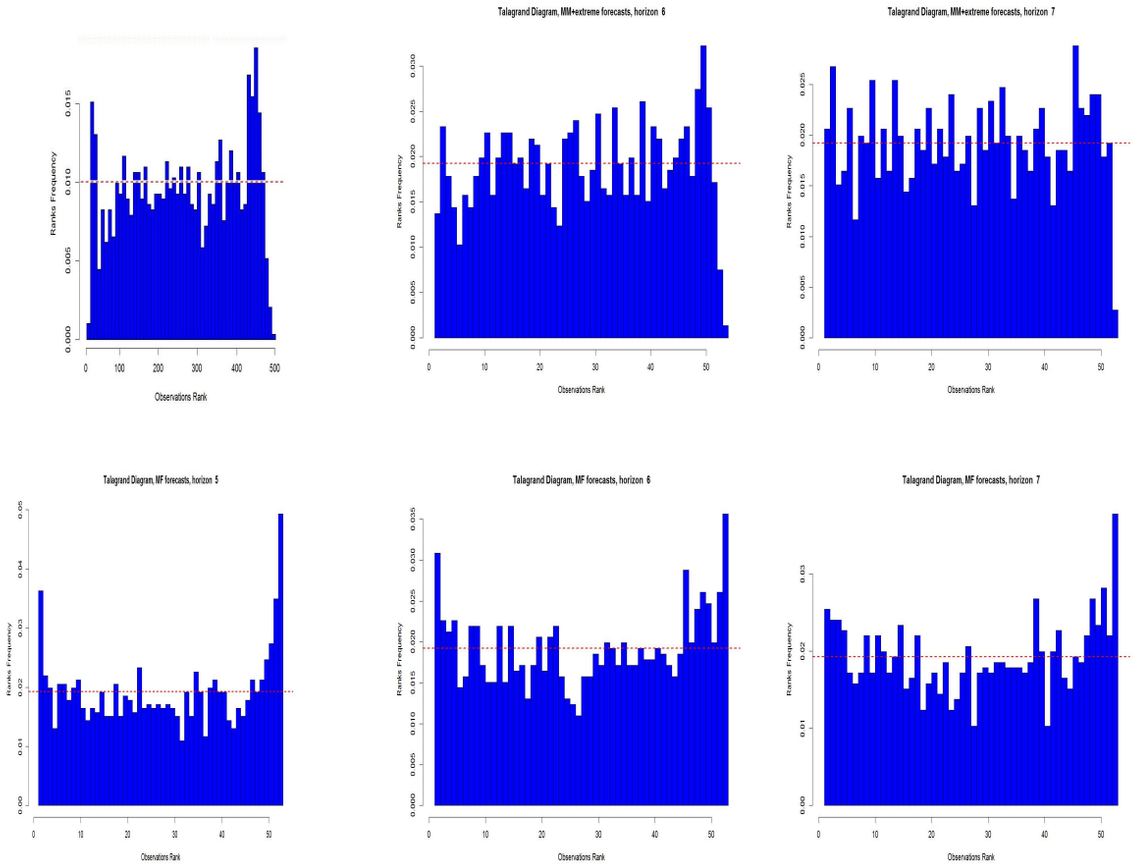


Figure 3.10: Talagrand Diagrams corresponding to the mixture model - that separates extreme values by tail and for the right tail, by time-horizon - on top and the initial predictions at the bottom, for the time-horizons 5, 6 and 7 respectively.

Mixture Models with the GEV parameters estimated by tail, by month and by package of time-horizon.

The third model we build is the one recommended by the AIC and we use it to produce another ensemble of forecasts. The scores for it are in the table 3.9 and show an improvement of the overall precision of the forecasts (CRPS) but also the existence of more large errors (RMSE).
We draw the Talagrand Diagram (we can see in the Figure 3.11 the ones for the horizons 5, 6, 7 with the corresponding diagrams of the initial MF forecasts) and we notice as for the two cases before the diagrams for the mixture model are slightly flatter the ones of the MF forecasts. But the estimation problem still exists for the extreme right values.

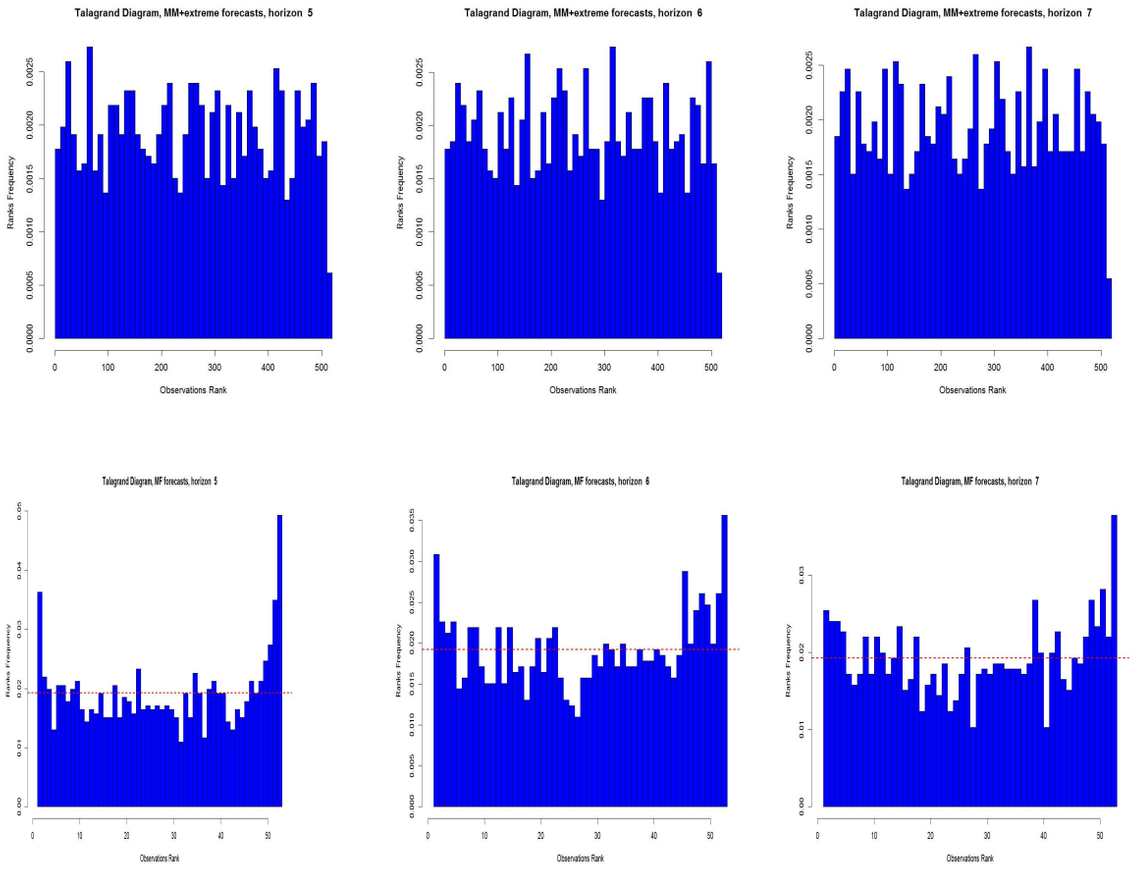


Figure 3.11: Talagrand Diagrams corresponding to the mixture model (GEV parameters estimated by tail, by month and by package of time-horizon) on top and the initial predictions at the bottom, for the time-horizons 5, 6 and 7 respectively.

This last model, recommended by AIC is improving the CRPS, and the Rank Diagram globally. It also gives (as we can see in the table 3.10) good estimations for the quantiles of the left side. But it has that imperfection on the extreme right side, which is rather annoying as we want to implement this method to improve the extreme values of the distribution.

Discussion extreme mixture models

We built three mixture models differing by class of values we used to estimate the GEV parameters. The implementation of the three mixture models on our data gave us three new EPS. Comparing the resulting scores for the three models we can see that they are not significantly different one from each other. The three of them give a good representation of the bulk of the distribution, they improve the CRPS (comparing to the initial EPS), meaning that the overall skill of the EPS is better (see the CRPS values in table 3.9). The ratio between the RMSE and MAE remains the same so we keep having a high error variance, but we loose a little bit in performance from the RMSE point of view meaning that there are more large errors in the three new EPS we built than in the initial one (see table 3.7). The estimation of the tails (more exactly for the right tail for $Q_{0.99}$ and $Q_{0.98}$) of the three new distributions is less good at the right side of the distribution (see Figure 3.11) and this can also be seen in the quantiles table (see table 3.10). The estimation becomes better starting with the $Q_{0.95}$, this could make us think that the problem is coming from a not larger enough number of extreme values, but the fact that we don't observe the same phenomena for the left tail (where the number of extreme values is comparable to the one of the right tail) is infirming this explanation. Another possible explanation might be that an extreme approach is not successfully connecting with an EPS or maybe the connection should be done differently than by a threshold corresponding to the statistical rank.

Finally as there is a matter of quantiles, it may be more justified to adopt a Quantile Regression for the study of the tails. This is what we propose in the next and last section of the thesis.

Horizon	CRPSpot	CRPSReli	CRPS	MAE	RMSE	R2
5	0.5859	0.0050	0.5909	1.260	1.698	0.967
6	0.7380	0.0038	0.7417	1.557	2.087	0.951
7	0.9068	0.0041	0.9109	1.892	2.508	0.931
8	1.0499	0.0031	1.0530	2.170	2.849	0.913
9	1.1801	0.0028	1.1829	2.419	3.151	0.895
10	1.2941	0.0027	1.2968	2.654	3.431	0.878
11	1.3784	0.0028	1.3812	2.817	3.636	0.866
12	1.4293	0.0042	1.4336	2.943	3.781	0.855
13	1.4757	0.0035	1.4791	3.020	3.867	0.846
14	1.4872	0.0035	1.4907	3.073	3.940	0.842

Table 3.9: The different scores by time-horizon, for the forecasts obtained when simulating with the mixture model selected by the AIC criterion: GEV parameters estimated by tail, by month and by horizon package. We have good scores, we improve the CRPS for all the period.

H	Model	Q1	Q99	Q2	Q98	Q5	Q95	Q10	Q90
5	Mixt	.0219	.0000	.0432	.0007	.0609	.0071	.0952	.0871
	MF	.0363	.0493	.0583	.0843	.0781	.1117	.1323	.1768
6	Mixt	.0185	.0027	.0398	.0069	.0562	.0240	.0947	.1063
	MF	.0309	.0357	.0535	.0617	.0748	.0816	.1276	.1564
7	Mixt	.0275	.0000	.0474	.0034	.0645	.0206	.1139	.0879
	MF	.0254	.0377	.0494	.0597	.0734	.0879	.1290	.1599
8	Mixt	.0165	.0021	.0282	.0117	.0543	.0288	.0934	.1078
	MF	.0192	.0357	.0405	.0536	.0659	.0769	.1312	.1477
9	Mixt	.0165	.0014	.0371	.0089	.0543	.0296	.1010	.1107
	MF	.0234	.0289	.0467	.0488	.0701	.0742	.1271	.1478
10	Mixt	.0131	.0021	.0351	.0158	.0543	.0337	.1018	.1072
	MF	.0282	.0296	.0509	.0502	.0702	.0729	.1279	.1472
11	Mixt	.0110	.0014	.0344	.0145	.0496	.0385	.1005	.1115
	MF	.0255	.0248	.0434	.0496	.0654	.0695	.1246	.1342
12	Mixt	.0103	.0021	.0289	.0145	.0544	.0406	.1033	.1247
	MF	.0269	.0310	.0455	.0475	.0654	.0723	.1198	.1357
13	Mixt	.0138	.0014	.0338	.0096	.0503	.0283	.1054	.1061
	MF	.0262	.0241	.0496	.0482	.0669	.0731	.1165	.1289
14	Mixt	.0172	.0007	.0407	.0138	.0634	.0345	.1097	.0917
	MF	.0221	.0283	.0434	.0524	.0593	.0731	.1214	.1345

Table 3.10: The quantiles 1%, 2%, 5%, 95%, 98%, 99%, for the initial MF forecasts and for the mixture models when we compute the parameters separately by month and by package of time-horizons.

Chapter 4

Improvement of Short-Term Extreme Temperature Density Forecasting using Best Member Method (NHESS Article)

This chapter is in the form of an article, which was submitted at the *Natural Hazards and Earth System Sciences* Journal, published by the Copernicus GmbH (Copernicus Publications) on behalf of the European Geosciences Union (EGU).

It contains the quantile regression method that we want to adopt in the study of the extreme parts of the distribution. Since we want to model the tails, it is important to take into account relative errors on quantiles. That is why we will use a χ^2 distance which allows explicit overweighting of the tails. We choose classes around the probability of interest for us, which is 1%: [0; 0.01], [0.01; 0.02] and [0.02; 0.05] for the lower tail, and the symmetric classes for the upper one. We will use this to measure all improvements and the results are positives even if there remain some biases in the tail representation.

Manuscript prepared for Nat. Hazards Earth Syst. Sci.
with version 4.2 of the LaTeX class copernicus.cls.
Date: 17 December 2012

Improvement of Short-Term Extreme Temperature Density Forecasting using Best Member Method

Adriana Gogonel[1,2], Jérôme Collet[1], and Avner Bar-Hen[2]

[1]EDF R&D Division, OSIRIS Department, 1, avenue du Général de Gaulle, 92141 Clamart cedex, France
[2]MAP5, UFR de Mathématiques et Informatique, Université Paris Descartes, 45 rue des Saints-Pères, 75270 Paris cedex 06

Abstract. Temperature influences electric demand and supply, so it may be a cause of blackout. That is why, as any electricity generator, Électricité de France (EDF) has to model the uncertainty about future temperatures, using ensemble prediction systems (EPS). Nevertheless, the probabilistic representation of the future temperatures provided by EPS suffers some lack of reliability. This lack of reliability becomes crucial for extreme temperatures, since they can result in a blackout. To solve this problem, a method of choice is the Best Member Method: it improves the whole representation, but there is still some room for improvements about the tails. We show that, in this case, using quantile regression to model the error distribution is more efficient than the usual two-stage OLS regression. To obtain further improvement, one may use extreme value distribution to model the error, when the realization is smaller or bigger than all forecasts. Another possibility would be to model the probability that a given forecast is the best one, using exogenous variables.

Keywords. quantile regression, probability integral transform, ensemble prediction system, density forecasting, energy

1 Introduction

The uncertainty of future temperatures is a major risk factor for an electricity utility such as Électricité de France: demand increases when temperature is lower than 18 ° C for heating, and larger than 18 ° C for cooling. Moreover, high temperatures also create cooling problems for thermal plants.
To fulfill the risk management needs, a compulsory source of information is the ensemble prediction systems (EPS), provided by weather forecasting institutes, such as ECMWF DOCUMENTATION (2002, 2006). Ensemble forecasting is a numerical prediction method that is used to attempt to generate a representative sample of the possible future states of a dynamical system. Ensemble forecasting is a form of Monte Carlo analysis: multiple numerical predictions are conducted using slightly different initial conditions that are all plausible given the past and current set of observations. One combine them to obtain estimates of future temperatures, estimates of uncertainty about future temperatures Whitaker and Loughe (1998), estimates of predictive density, ...
Nevertheless, the probabilistic representation of the future temperatures provided by EPS suffers some lack of reliability, especially for probabilities around 1%. Lack of reliability is prohibitive, since as a risk manager, EDF has to use the most reliable information available Diebold et al. (1998). This is emphasized by the size and the market power of EDF: EDF has to be able to prove to any stakeholder of the market it uses an unbiased representation of the risk . The 1% level is a constraint imposed by French technical system operator: the probability of using exceptional means (e.g. load shedding) has to be lower than 1% Sur (2004). The lack of smoothness issue is more technical, since it may cause some problems in generation management tools.
Many methods have been developed in order to get a smooth unbiased representation of the risk caused by temperature Hagedorn (2010), but for most of them the extremes are still difficult to predict. We will use here as a basis the best member method Fortin et al. (2006); Gogonel-Cucu et al. (2011b,a). The main point is that the fitting of the kernel dispersion is improved when using quantile regression, in place of a two-stage OLS regression.

Correspondence to: Jérôme Collet
(Jerome.Collet@edf.fr)

The outline of the paper is the following: we first present briefly a basic use of best member method. Then, we show the use of quantile regression to improve the error modeling needed for best member method. In a last part, we show some more tentative improvements.

2 A use of the best member method

We describe here briefly a simple use of best member method on ECMWF forecasts, with a « dressing » depending on the rank of the forecast. We study here all the horizons, but each one independently of the others. Then, in the following, we assume there is only one forecast horizon.

2.1 The data

The data consists of two different arrays. First one is the array of the forecasts, with 3 dimensions.

- The date of forecasting (the forecast is made at this date), in our set the dates are between 2007-03-27 and 2011-04-30 (which makes 1473 dates).
- The forecast horizon: ECMWF provides forecasts for 1 day ahead to 14 days ahead.
- The member: it is identified by a number between 0 and 50. For a given date, a member corresponds to a given initial state.

Second one is the array of the realizations, with 1 dimension, which is the date, approximately with the same extent as the forecasts.
The member 0 is a bit different of the other, since its initial state is exactly the one derived from the observations. About other members, for two different dates, there is no obvious link between the members denoted by the same number. Despite the little difference of member 0, we consider the member number is uninformative, and the members of the EPS can be considered as exchangeable as defined in Bernardo (unlike what happens for multi-model ensembles Gneiting et al. (2005)).
The temperature we study is an average of temperatures in some points of France. The weights are chosen in order to estimate the electricity load in France Dordonnat et al. (2008).

2.2 The best member method for ECMWF data

The best member method was proposed by V. Fortin (see Fortin et al. (2006)) and improves the studies previously led by Roulston and Smith (see Roulston and Smith (2002)) then by Wang and Bishop (see Wang and Bishop (2005)).
For each date and each horizon, the « best member » is the member closest to the realization (with the smallest absolute difference). The principle of the method is to model the probability distribution of the difference between the « best member » and the realization. Then, the probabilistic forecast will be the mixture of the modeled probability distribution. Since all members of the EPS are exchangeable, in a first step, all error models and weights will be the same.
An additional idea, proposed by Fortin, is to use the rank. In this case, we first rank the forecasts, for each date. Then, the error model and the weights will be a function of the rank.
More precisely, for each rank r, the weight linked to this rank is the number of best members having rank r, divided by the total number of forecasts. To model the error, we use the rank r as a discrete variable.
Since we prefer to use a parametric framework, the model of the error consists of:

1. a model for the mean of the error of the best member, noted ϵ,
2. a model for the variance of the error of the best member.

There is many possible variables to model the error mean and variance:

- variables summarizing features of the current EPS: mean m, square of the mean to take account of the effect of extreme temperatures, spread (here measured as the interquartile range `IQR`) of the ensemble;
- variables to take account of a smooth influence of the date (the date t itself, and $\cos(2t\pi/365)$, $\sin(2t\pi/365)$);
- the rank of the member, considered as a discrete variable;
- since central ranks behave similarly; a variable `edge`, equal to 0 when the rank is central, to the rank elsewhere.

Furthermore, we may assume interactions between all discrete and all continuous variables: for example, an interaction between mean temperature of the EPS and rank means that the model has a different slope for the mean temperature for each rank.
An additional degree of freedom lies in the definition of « central ranks ». Ranks 1 and 51 behave obviously differently than very central ranks: larger probability to be the best member and larger dispersion of the error. Nevertheless, the same is true, in a smaller extent, for ranks 2 and 50,, and even for ranks 3 and 49.

Furthermore, it is impossible to estimate all the parameters at once for all horizons, since the variance of the residuals is approximately 3 times bigger for horizon 14 than for horizon 1.

We can use an automated variable selection method, such as the « stepwise » selection method Institute (2006), it allowed to get some information about significant variables.

- The rank never appears, the information it carries is always provided by variable `edge`.
- The mean temperature is significant for the model of th error mean.
- The spread is significant for the error variance, but we also need the mean temperature and its square.
- It is useless to take account of a smooth influence of the date.

The impact of the definition of the « central ranks » has been assessed manually. The best choice, regarding AIC, or adjusted R^2, is to consider as non-central the ranks {1,2,50,51}.

Finally, the chosen models are the following:

$$\begin{aligned}\epsilon &= a(\texttt{edge}) + b(\texttt{edge}) \cdot m \\ (\epsilon - \hat{\epsilon})^2 &= \alpha(\texttt{edge}) + \beta(\texttt{edge}) \cdot m + \\ &\quad \gamma(\texttt{edge}) \cdot m^2 + \delta(\texttt{edge}) \cdot \texttt{IQR}\end{aligned}$$

In order to take account of some possible evolution in the data, we divided it in two equal parts, fitting the model on one part and using it on the other.

This results is an important improvement of the reliability of the temperature density forecast. To prove this, we use probability integral transform Diebold et al. (1998), and we plot on figure 1 the difference between the theoretical cumulated density function of the PITs, and its empirical counterpart. For the curve of « raw EPS », the point (0.32,0.083) means that the realized temperature is less than 32% of the simulations for 32+8.3=40.3% of the dates, instead of 32%. For the curve of « Usual BMM », the realized temperature is less than 32% of the simulations for 31.5% of the dates, instead of 32%: the reliability improvement is important.

Another argument is the comparison of reliability component of the CRPS Hersbach (2000): this indicator decreases from a factor 7 for horizon 1 to 20% for horizon 14, with a local minimum of 8% for horizon 5.

Furthermore, this reliability improvement results in important differences on practically useful quantities. For example, between raw EPS and simulations using BMM, the variance of the temperature increases from 100% for horizon 1 to 10% for horizon 14 (regarding the mean, the differences are significant but small).

Despite the important reliability improvement due to BMM, we see the tails are not well represented.

3 Improving dispersion estimation using quantile regression

Since we want to model tails, it is important to take account of **relative** errors on quantiles, we can not anymore use Kolmogorov-Smirnov distance. For example, it would be possible to use Jager and A. (2005): this paper proposes to compute a likelihood ratio between the theoretical and the empirical cumulative distribution function.

In this study, we know the range of probabilities which are of interest of us. We have to prove the supply-demand balance is positive in 99% of the cases, and the variability of demand is a bit bigger than the variability of supply. So, when the supply-demand balance is on its 1% quantile, the demand is close to its 1% quantile, but not exactly equal to. That is why we will use a χ^2 distance, with the following classes: [0;0.01], [0.01;0.02] and [0.02;0.05] for the lower tail, and the symmetric classes for upper one. We will use this to measure all improvements.

A possible explanation to the poor quality of the tail representation is that the dispersion measure heavily depends on distribution form. Yet, this distribution is very different for the smallest, biggest and central members. That is why we propose to model, in one stage, two different quantiles of the error, taking account of the same variables as OLS. Quantile regression Koenker and Bassett (1978) is a type of regression: whereas the method of least squares results in estimates of the conditional mean of the response variable, quantile regression aims at estimating any quantile of the response variable. A crucial point, here, is this estimation does not rely on distributional hypothesis.

When using BMM, we know that the errors are not exactly normally distributed, but we assume their shape (at least the errors for central members) has little influence on the whole model accuracy. Nevertheless, we know that their location and scale are highly variable and have huge impact on whole model accuracy, that is why we model it.

The purpose of this modeling is to simulate, and the accuracy of this simulation will be measured using mainly rank-based measures. Then, we are not interest in moments, we are interested in quantiles.

If our distributional assumption is wrong, but the mean and variance right, there will be intersections between the real error distribution and the theoretical one, but we do not know where: these intersections may be located anywhere. If we use quantiles to locate and rescale, we know where are the intersections, they are in the quantiles we chose. Then, this location and scaling method is more robust to distributional deviations than using two-step OLS. For example, if we model the quantiles $Q_{1/3}$ and $Q_{2/3}$, we know the Kolmogorov-Smirnov

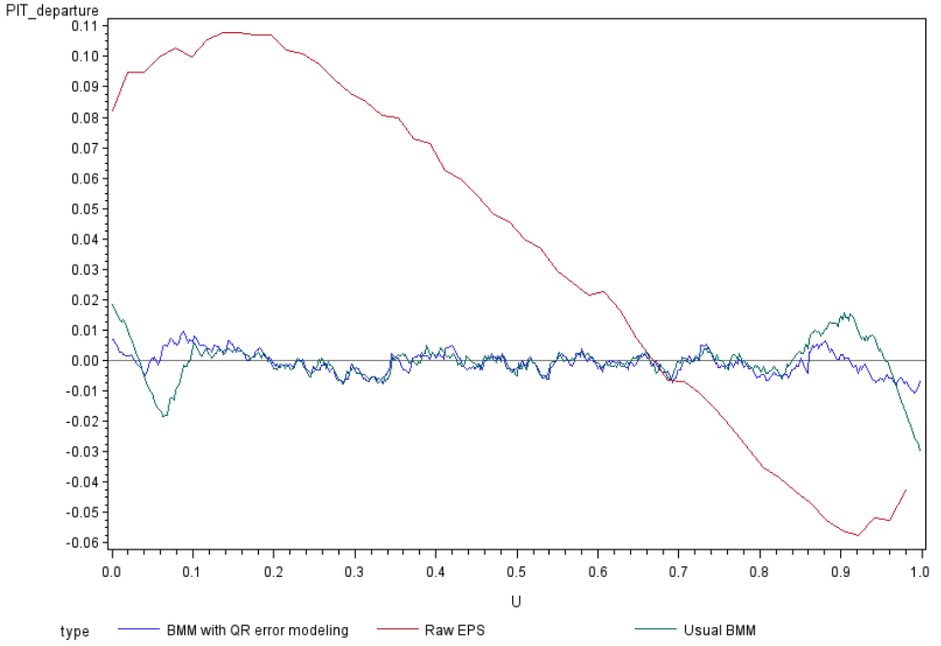


Fig. 1. Comparison of PIT anomalies

distance between the modeled distribution and the real one is bounded by 1/3, instead of 1/2 in case of two-stage OLS (let consider the case of the mixture of two normally distributed random variables: $(1-p)\times\mathcal{N}(0,1)+p\times\mathcal{N}(1/p,1)$, with $p\to 0$).

We modeled the quantiles $Q_{1/3}$ and $Q_{2/3}$, using the variable named `edge`, and the ensemble mean and spread. For both quantiles, the chosen model is:

$$Q_{x/3} = \alpha(\texttt{edge}) + \beta(\texttt{edge})\cdot m + \gamma(\texttt{edge})\cdot m^2 + \delta(\texttt{edge})\cdot \texttt{IQR}$$

Since we want to use a normal distribution to simulate the errors, its parameters will be the ones necessary to obtain the estimated third and first quartiles. In other words:

$$m = \frac{Q_{2/3}+Q_{1/3}}{2} \qquad \sigma = \frac{Q_{2/3}+Q_{1/3}}{2\times\phi^{-1}(2/3)}$$

where ϕ is the cumulative density function of the normal distribution, so $\phi^{-1}(2/3) = 0.431$. The rest of the model remains the same as in 2.2.

The results are the following. We may first look at the same PIT plot, in figure 1.More globally, we look at the χ^2 distances, for each horizon, in figure 2. We see the improvement is important, and that there is a little forecast degradation for only 2 horizons (among 14).

We know that the EPS, in ECMWF, are fitted on horizon 5. Furthermore, for horizon 1 and 2, the evolution of the numerical model can not diverge a lot from reality.

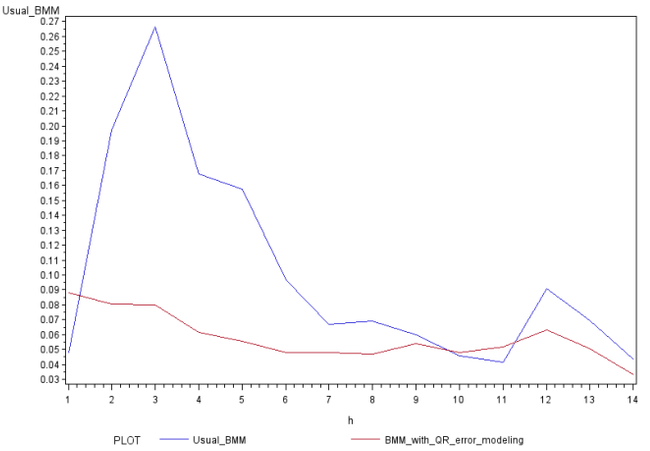


Fig. 2. χ^2 distances, for all horizons

This gives a possible explanation for the huge χ^2 distance of horizon 3.

4 Conclusions

We wanted to improve the representation of the tails, which is poor when using BMM, at least on ECMWF EPS. We state it is possible, using Quantile Regression.

This improvement makes possible a better sizing of the power supplies, resulting in important cost reductions. Nevertheless, there remain some biases in the tail representation.

To obtain further improvement, one may list the following possibilities:

- The extreme members are more frequently than others the best one, the error in this case is more dispersed (its standard deviation is approximately 5 times bigger than for central ranks, for each horizon), and its distribution is asymmetric. That is why one could try to model it using Extreme Value Distribution.

- In our modeling, the probability that a member of a given rank is the best one does not depend on any exogenous variable. Testing and possibly rejecting this independence assumption could help improving tail estimation.

References

Mémento de la sûreté du système électrique, édition 2004, Tech. rep., Réseau de Transport d'Électricité, www.rte-france.com/uploads/media/pdf_zip/publications-annuelles/memento_surete_2004_complet___.pdf, 2004.

Bernardo, J.-M.: The concept of exchangeability and its applications, Far East J. Mathematical Sciences, Special volume, 111–121.

Diebold, F. X., Gunther, T. A., and Tay, A. S.: Evaluating Density Forecasts with Applications to Financial Risk Management, Internationa1 Economic Review, 39, 863–883, 1998.

DOCUMENTATION, I.: The Ensemble Prediction System, Tech. rep., ECMWF, 2002.

DOCUMENTATION, I.: Part V: The Ensemble Prediction Systems, Tech. rep., ECMWF, 2006.

Dordonnat, V., Koopman, S.-J., Ooms, M., Dessertaine, A., and Collet, J.: An hourly periodic state space model for modelling French national electricity load, International Journal of Forecasting, 24, 566–587, 2008.

Fortin, V., Favre, A., and Said, M.: Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member, Q. J. R. Meteorol. Soc., 132, 1349 –1369, 2006.

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

Gogonel-Cucu, A., Collet, J., and Bar-Hen, A.: Implementation of two Statistic Methods of Ensemble Prediction Systems for Electric System Management, Case Studies in Business, Industry and Government Statistics (CSBIGS), Submitted, 2011a.

Gogonel-Cucu, A., Collet, J., and Bar-Hen, A.: Implementation of Ensemble Prediction Systems Post-Processing Methods, for Electric System Management, in: 58th Congress of the International Statistical Institute, 2011b.

Hagedorn, R.: Post-Processing of EPS Forecasts, http://www.ecmwf.int/newsevents/training/meteorological_presentations/pdf/PR/Calibration.pdf, 2010.

Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, Weather and Forecasting, 15, 559–570, 2000.

Institute, S.: The GLMSelect Procedure : Stepwise Selection(STEPWISE), http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_glmselect_a0000000241.htm, 2006.

Jager, L. and A., W. J.: A new goodness-of-fit test: the reversed Berk-Jones statistic, Tech. Rep. TR443, Department of Statistics, University of Washington, http://www.stat.washington.edu/www/research/reports/2004/tr443.pdf, 2005.

Koenker, R. and Bassett, G.: Regression Quantiles, Econometrica, 46, 33–50, 1978.

Roulston, M. and Smith, L.: Combining dynamical and statistical ensembles, Tellus, 55A, 16–30, 2002.

Wang, X. and Bishop, C.: Improvement of ensemble reliability with a new dressing kernel., Q. J. R. Meteorol. Soc., 131, 965–986, 2005.

Whitaker, J. and Loughe, A.: The Relationship between Ensemble Spread and Ensemble Mean Skill, Monthly Weather Review, 126, 3292 – 3302, 1998.

Conclusion

The aim of this thesis is to improve the short-term forecasts of temperatures (from 5 days ahead to 14 days ahead) - provided by Meteo-France as an ensemble prediction system (EPS). The improvement is to be useful for the electric system management, at EDF France and it can be an improvement from the spread or the skill point of view.

The uncertainty of future temperatures is a major risk factor for an electricity utility such as Électricité de France: demand increases when temperature is lower than 18°C for heating, and larger than 18°C for cooling. Moreover, high temperatures also create cooling problems for thermal plants.

To fulfill the risk management needs, a compulsory source of information is the EPS, provided by weather forecasting institutes, such as ECMWF (see [DGT98], [HER00]). Ensemble forecasting is a numerical prediction method that is used to attempt to generate a representative sample of the possible future states of a dynamical system. Ensemble forecasting is a form of Monte Carlo analysis: multiple numerical predictions are conducted using slightly different initial conditions that are all plausible given the past and current set of observations. One combine them to obtain estimates of future temperatures, estimates of uncertainty about future temperatures (see [WL98]), estimates of predictive density,... Nevertheless, the probabilistic representation of the future temperatures provided by EPS suffers some lack of reliability, especially for probabilities around 1%. Lack of reliability is prohibitive, since as a risk manager, EDF has to use the most reliable information available (see [DGT98]). This is emphasized by the size and the market power of EDF: EDF has to be able to prove to any stakeholder of the market it uses an unbiased representation of the risk . The 1% level is a constraint imposed by French Technical System Operator: the probability of using exceptional means (e.g. load shedding) has to be lower than 1% (see [Sur04]). The lack of smoothness issue is more technical, since it may cause some problems in generation management tools.

The EPS we are working on contains forecasts of daily temperature in France for a four years period, from March 2007 to March 2011, that we will compare with the daily mean of the observed temperatures. It is a 51-member EPS: we have 51 different values for every one of the 14 time-step (time-horizons). Those 51 values are obtained by running 51 times the same model with different initial conditions, only one forecast, the scenario 0 is run with a non-perturbed initial conditions.

The first method we propose, it is chosen as to improve the probability density function of the forecasts, preserving at the same time the quality of the mean forecasts. It is the best member method (see [FFS06]) and its principle is to design for each lead time in the data set, the best forecast among all the k forecasts provided by the temperature prediction system, to construct an error pattern using only the errors made by those "best members" and then to "dress" all the members of the initial prediction system with this error pattern. This method allows us to extend the number of simulated temperatures. The case of the method we implement is the one where in the simulation part the ensemble members are dressed and weighed differently by classes of statistical order. We obtain a new EPS with 10 time more members than the initial one and with a significant improvement of the distributions spread in its central part (Talagrand Diagram) with a small loss of precision (RMSE) but keeping the same global skill (CRPS).

We use a second method to compare the results we obtained with the BMM and this is the bayesian method proposed by Raftery (see [GRWG05]). It is a statistical method for post processing model outputs which allows providing calibrated and sharp predictive PDFs even if the output itself is not calibrated. The method allows to use a sliding- window training period to estimate new models parameters, instead of using all the database of past forecasts and observations. We obtain another system of forecasts with a very close spread (to the initial spread) but with less good skill qualities (CRPS, RMSE). This results confirm that the first method is a success: increasing the number of forecasts for improving the distribution of the EPS, without losing the precision of its mean forecast. The thing we could still improve is the distribution of the tails.

What we need now is to find a way to keep the BMM for the central part of the distribution and another method for the tails. This brings us to use a mixture model, which will have a gaussian distribution model for the heart of the distribution and a GEV distribution-type for the tails. Choosing the best GEV to fit the tails is finding the best way of estimating the parameters - shape, location, scale - among the most popular methods: maximum likelihood, bayesian method, method of moments. Finding the best GEV is also choosing the more adequate type of GEV (Weibull, Fréchet, Gumbell) function of the shape parameter (strictly negative, strictly positive, or null) and finding the best model that estimates the three parameters. We find that the most adequate way of estimating is the maximum likelihood method, the GEV comes out to be a Weibull type and we select three models (chosen by three different criteria) of estimating parameters. All the three models improve the global skill of the forecasting system (CRPS) but give a strange effect to the last rank of the right tail of the rank diagram confirmed by the quantile computations (.99, .98, .95) that are not well estimated by the forecasts given by the mixture model. We propose to implement a last method, and this time we are not interest in moments, we are interested in quantiles.

The method is the quantile regression method. Since we want to model tails, it is important to take account of relative errors on quantiles. That is why in this last part of the thesis we will use a χ^2 distance which allows explicit over-weighting of the tails. We choose classes around the probability of interest for us, which is 1%: [0; 0.01], [0.01; 0.02] and [0.02; 0.05] for the lower

tail, and the symmetric classes for upper one. We will use this to measure all improvements and the results are positives even if there remains some biases in the tail representation.

In the end all the methods we implement on the ensemble prediction system for a four years period provided by Meteo France, when trying to improve its skill and/or spread bring us to the conclusion that the optimum method is to use a best member method type for the heart of the distribution and to adapt a quantile regression for the tails.

A perspective of this work is to build a mixture model combining the best member model for the bulk of the distribution and extreme model for the tails but choosing differently the connection between the models, other than the statistical rank of the forecasts in the EPS. A study for finding the optimum threshold is to be done: either by bayesian inference or by bootstrap approach, using a Monte-Carlo method.

Another way of building the mixture model is by using a bayesian inference approach for the central part of the distribution. Study how a bayesian model is connecting with the extreme model for the tails, in the EPS case it may be interesting.

The data we work on is the mean of the daily forecasts of temperature. Another possible approach is to implement the same kind of method on the minima/maxima of the daily forecasts. This could give good estimations for the distribution tails.

The same kind of study, as in the thesis could be make on the hourly temperature in France. A study where the 26 French stations are treated separately is also conceivable. Making one of this two separations of the data (by geographic or time criteria) reduces the cell and may increase the chances of obtaining good results comparing with this study where we work on mean of means (of 26 stations of 24-hourly) which artificially increases the skill of the initial system.

A multivariate study is possible, from the time-horizon point of view, in order to obtain coherent trajectories of temperature that might be use a simulator for the generalized additive model used in the the management of the electricity consumption at the EDF R&D.

Appendix

Horizon	CRPSpot	CRPSReli	CRPS	MAE	RMSE	R2
5	0.6286	0.0035	0.6321	1.178	1.555	0.972
6	0.7841	0.0036	0.7877	1.485	1.958	0.956
7	0.9608	0.0047	0.9655	1.814	2.379	0.937
8	1.1080	0.0055	1.1135	2.104	2.740	0.918
9	1.2393	0.0048	1.2440	2.362	3.048	0.901
10	1.3540	0.0048	1.3588	2.586	3.314	0.885
11	1.4343	0.0040	1.4383	2.756	3.521	0.872
12	1.4897	0.0042	1.4939	2.879	3.662	0.861
13	1.5386	0.0042	1.5428	2.980	3.779	0.852
14	1.5645	0.0037	1.5681	3.036	3.845	0.847

Table 4.1: The different scores by time-horizon, for the MF (initial) forecasts.

Horizon	CRPSpot	CRPSReli	CRPS	MAE	RMSE	R2
5	0.6250	0.0073	0.6323	1.297	1.720	0.966
6	0.7784	0.0079	0.7863	1.598	2.109	0.950
7	0.9527	0.0080	0.9607	1.941	2.547	0.929
8	1.0970	0.0093	1.1063	2.231	2.902	0.909
9	1.2295	0.0091	1.2386	2.484	3.207	0.891
10	1.3504	0.0088	1.3592	2.720	3.486	0.874
11	1.4295	0.0091	1.4386	2.883	3.692	0.861
12	1.4856	0.0098	1.4954	3.013	3.849	0.849
13	1.5343	0.0097	1.5440	3.099	3.944	0.841
14	1.5566	0.0096	1.5662	3.161	4.019	0.835

Table 4.2: The different scores by time-horizon, for the forecasts obtained with the BMM for the entire distribution.

Season/H	**H 5**	**H 6**	**H 7**	**H 8**	**H 9**	**H 10**	**H 11**	**H 12**	**H 13**	**H 14**	**All Horizons**
Spring	60	66	54	36	56	46	48	60	58	72	556
Summer	116	92	84	88	100	100	92	100	96	96	964
Autumn	52	44	50	32	40	50	62	58	54	46	488
Winter	248	192	178	170	150	136	128	128	118	110	1558
All seasons	476	394	366	326	346	332	330	346	326	324	3566

Table 4.3: The number of values, by time-horizon and by season included in our study of the extreme values.

	$p-value$ KS test
"horizon 5"	0.725
"horizon 6"	0.301
"horizon 7"	0.16
"horizon 8"	0.096
"horizon 9"	0.357
"horizon 10"	0.069
"horizon 11"	0.013
"horizon 12"	0.21
"horizon 13"	0.006
"horizon 14"	0.618

Table 4.4: The p-values corresponding to the KS test where $H0$ is the hypothesis that the extreme values (for $board = 1$) of the j-time horizon can be consider as coming from the same distribution as the values from all the time-horizons other than j.

	$p-value$ KS test
"horizon 5"	0.0004
"horizon 6"	0.283
"horizon 7"	0
"horizon 8"	0.073
"horizon 9"	0.039
"horizon 10"	0.012
"horizon 11"	0.009
"horizon 12"	0.022
"horizon 13"	0.001
"horizon 14"	0.118

Table 4.5: The p-values corresponding to the KS test where $H0$ is the hypothesis that the extreme values (for $board = 51$) of the j-time horizon can be consider as coming from the same distribution as the values from all the time-horizons other than j. W notice all the $p-values$, excepting horizon 6 and 14, are smaller than 0.05 so we can reject $H0$: for the rank 51 values, the extreme values coming from different horizons can not be considered as coming from the same distribution.

Horizon	**CRPSpot**	**CRPSReli**	**CRPS**	**MAE**	**RMSE**	**R2**
5	0.5847	0.0026	0.5872	1.217	1.634	0.969
6	0.7367	0.0026	0.7393	1.518	2.028	0.954
7	0.9040	0.0028	0.9068	1.843	2.434	0.934
8	1.0484	0.0026	1.0510	2.133	2.796	0.916
9	1.1757	0.0029	1.1786	2.382	3.099	0.899
10	1.2871	0.0034	1.2905	2.618	3.383	0.882
11	1.3726	0.0029	1.3755	2.772	3.571	0.870
12	1.4185	0.0030	1.4215	2.882	3.692	0.861
13	1.4717	0.0033	1.4750	2.976	3.799	0.851
14	1.4784	0.0028	1.4812	3.006	3.840	0.849

Table 4.6: The different scores by time-horizon, for the forecast obtained when creating a mixture model: BMM for the central part of the distribution and EVT for the tails

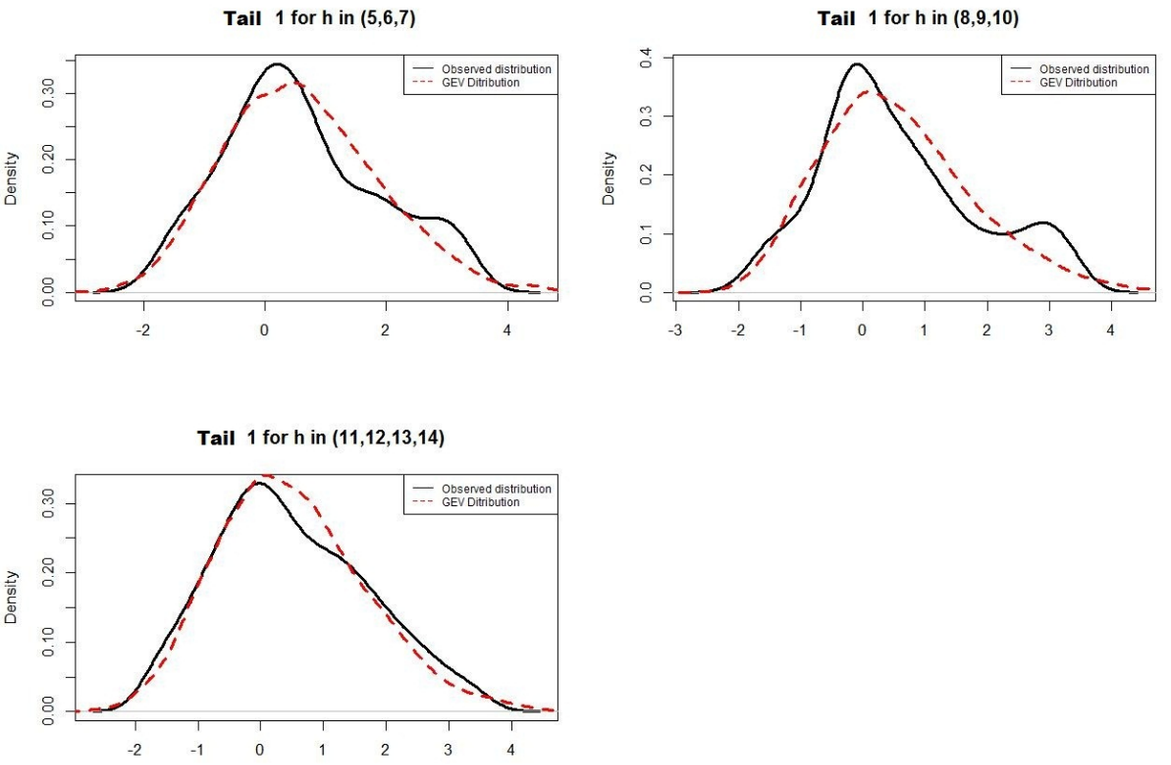


Figure 4.1: Errors densities and their corresponding GEV distributions for the left extreme values for three packages of horizons. We notice that the GEV fits good enough the observed distributions, less god for the 2nd package but in the adjustment test for the three cases $H0$ is not rejected ($H0$: assumes that the values in the two distributions come from a single distribution.)

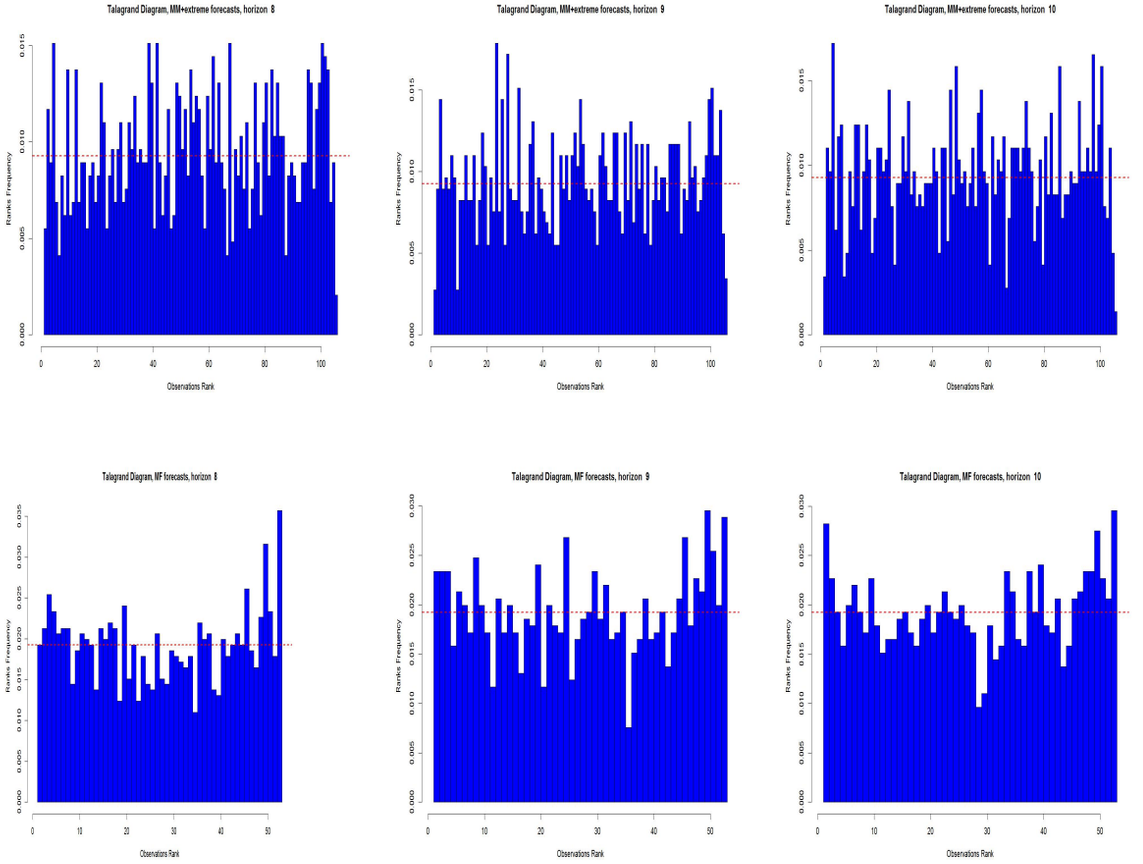


Figure 4.2: Talagrand Diagrams corresponding to the mixture model on top and the initial predictions at the bottom, for the time-horizons 8, 9 and 10 respectively.

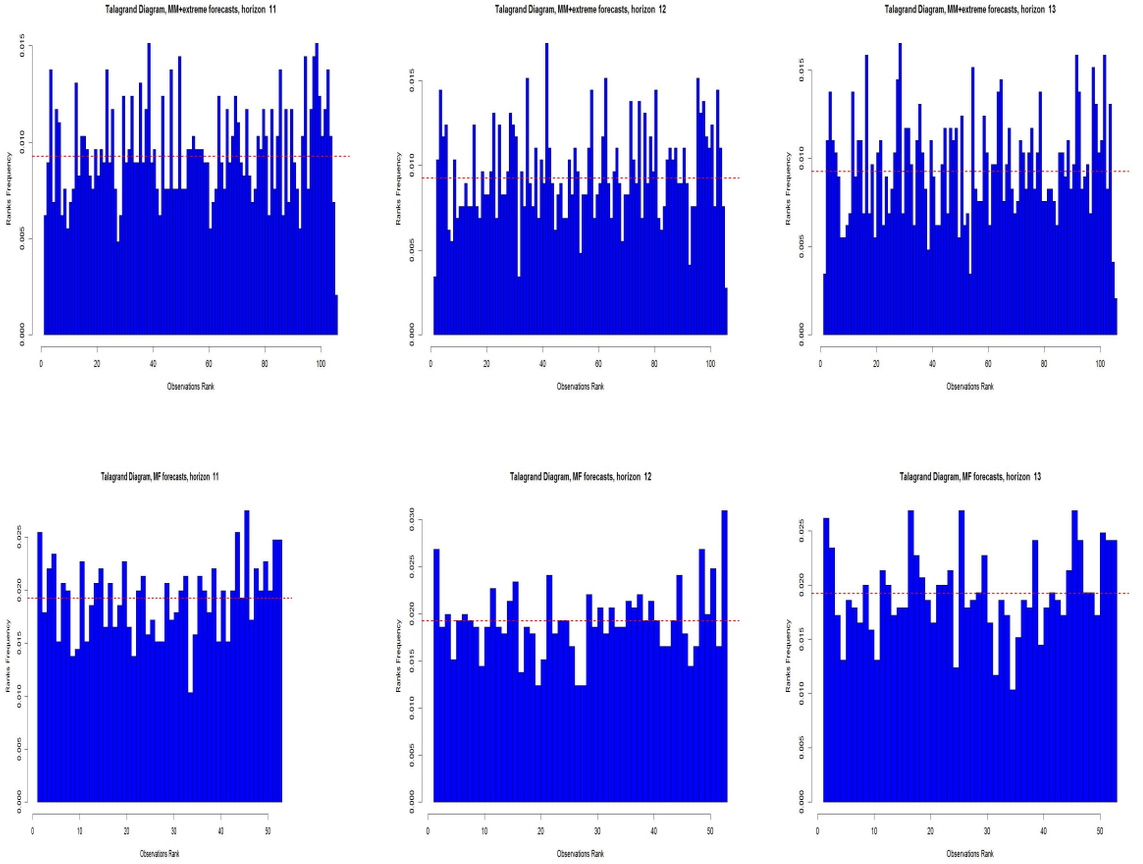


Figure 4.3: Talagrand Diagrams corresponding to the mixture model on top and the initial predictions at the bottom, for the time-horizons 11, 12 and 13 respectively.

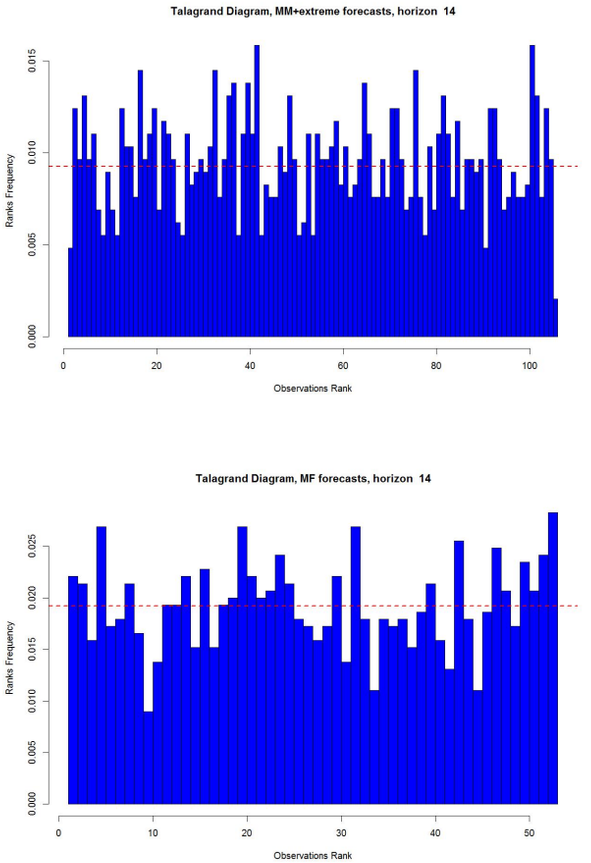


Figure 4.4: Talagrand Diagrams corresponding to the mixture model on top and the initial predictions at the bottom, for the 14th time-horizon.

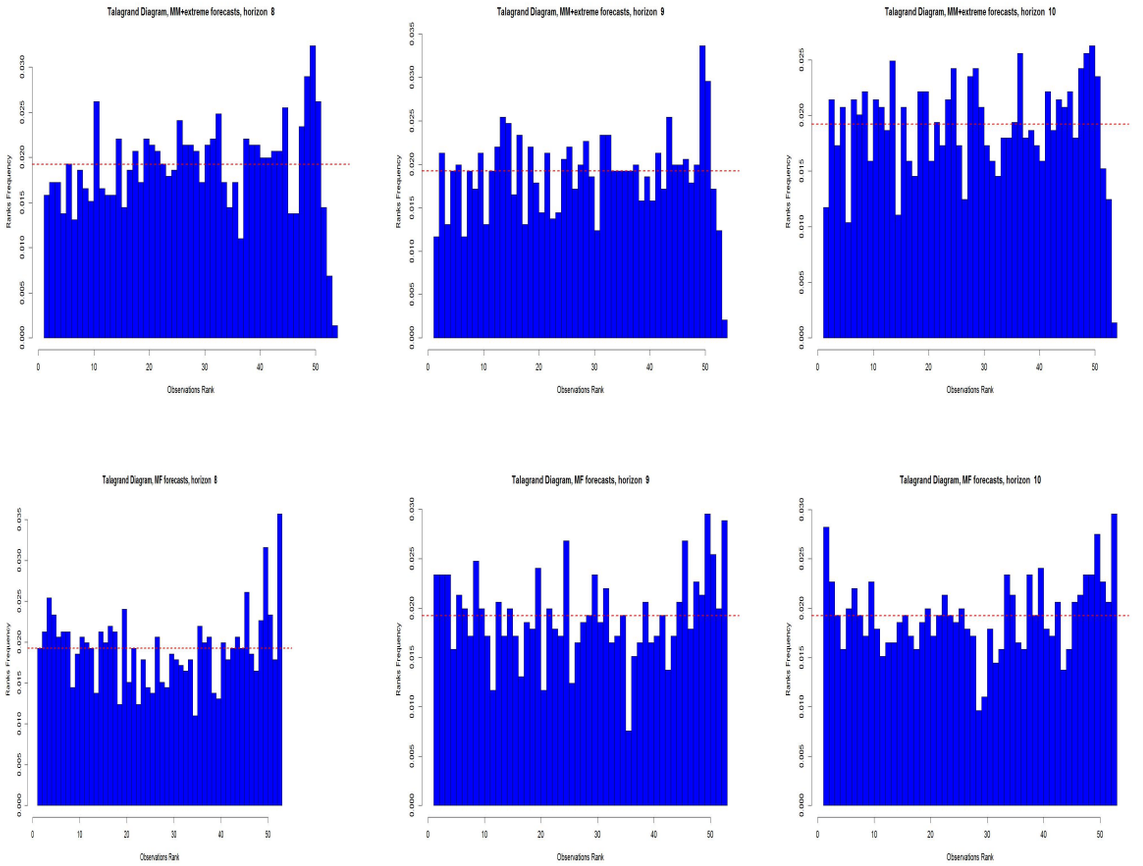


Figure 4.5: Talagrand Diagrams corresponding to the mixture model when separating extreme values by boards and at right also by horizon. On top and the initial predictions at the bottom, for the time-horizons 8, 9 and 10 respectively.

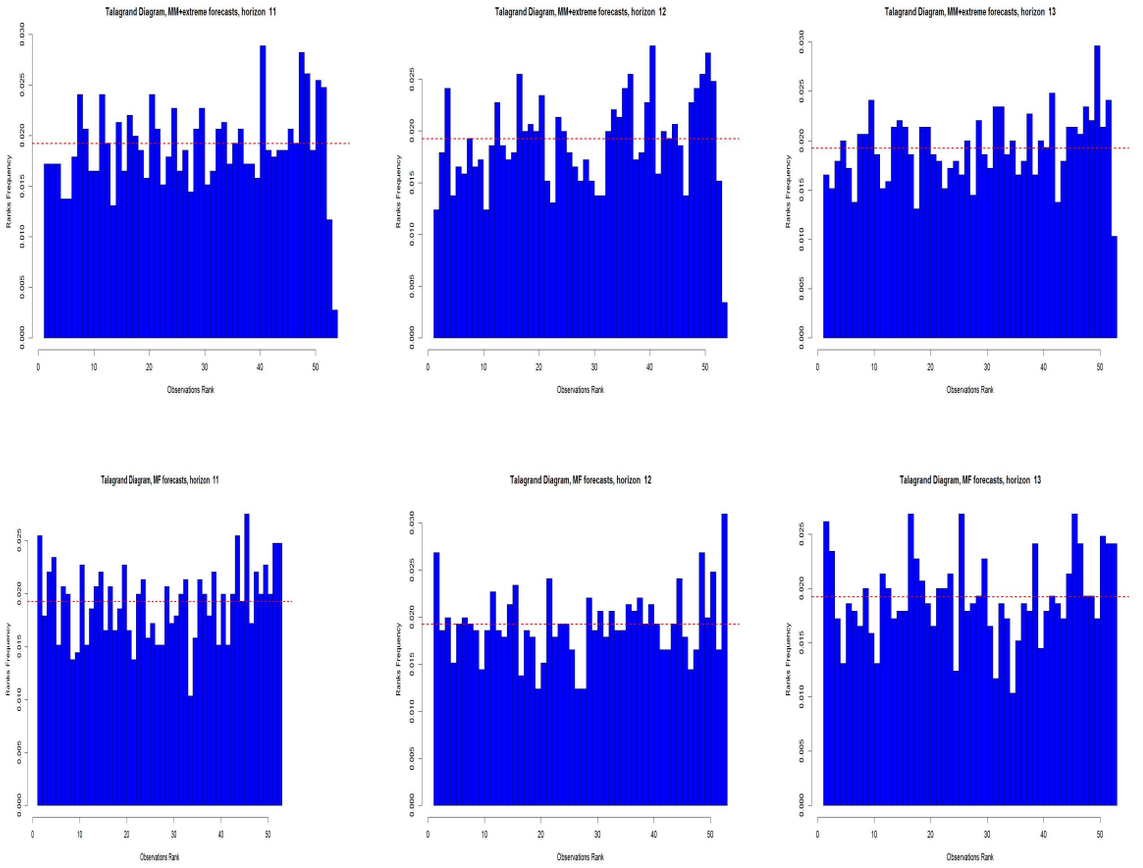


Figure 4.6: Talagrand Diagrams corresponding to the mixture model when separating extreme values by boards and at right also by horizon. On top and the initial predictions at the bottom, for the time-horizons 11, 12 and 13 respectively.

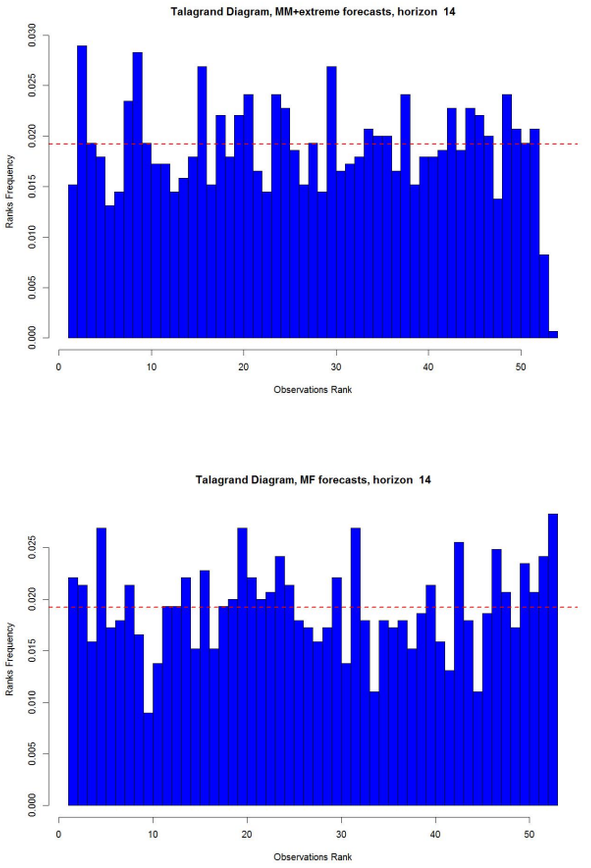


Figure 4.7: Talagrand Diagrams corresponding to the mixture model when separating extreme values by boards and at right also by horizon. On top and the initial predictions at the bottom, for the 14th time-horizon.

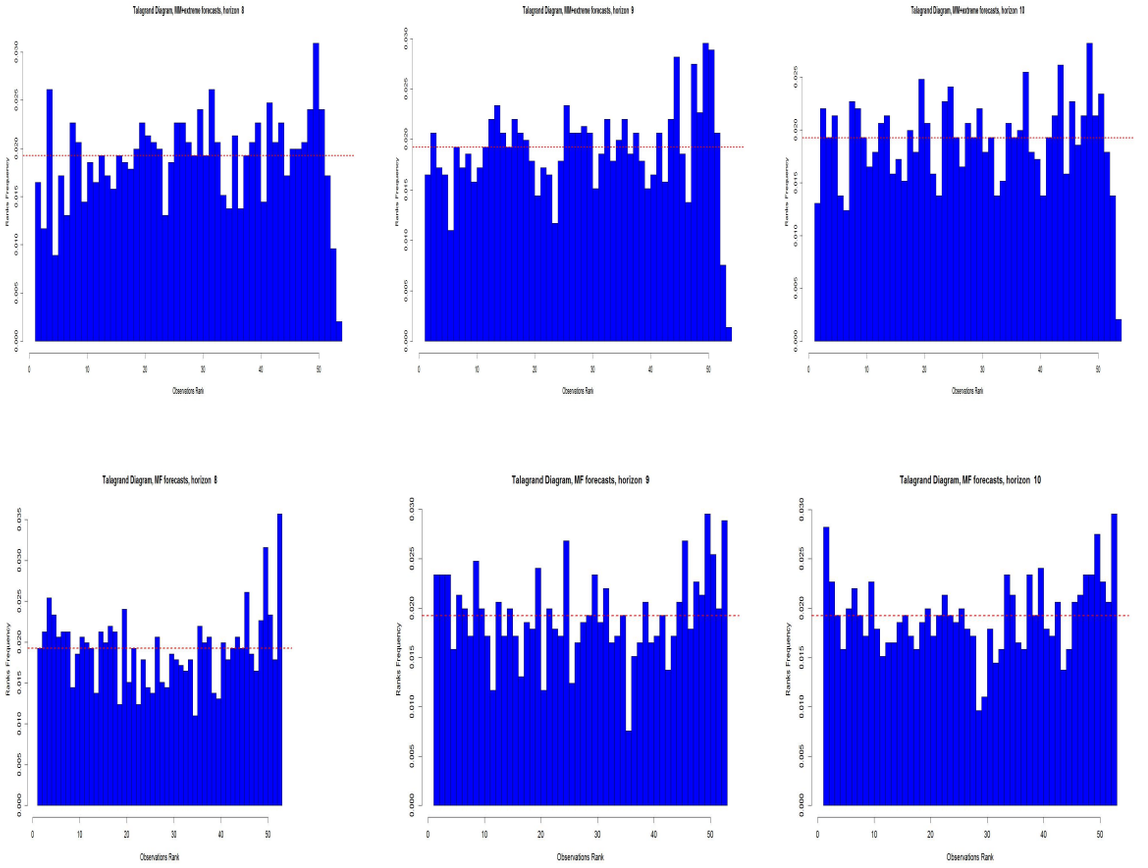


Figure 4.8: Talagrand Diagrams corresponding to the mixture model when separating extreme values by time-horizon packages and by month on top and the initial predictions at the bottom, for the time-horizons 8, 9 and 10 respectively.

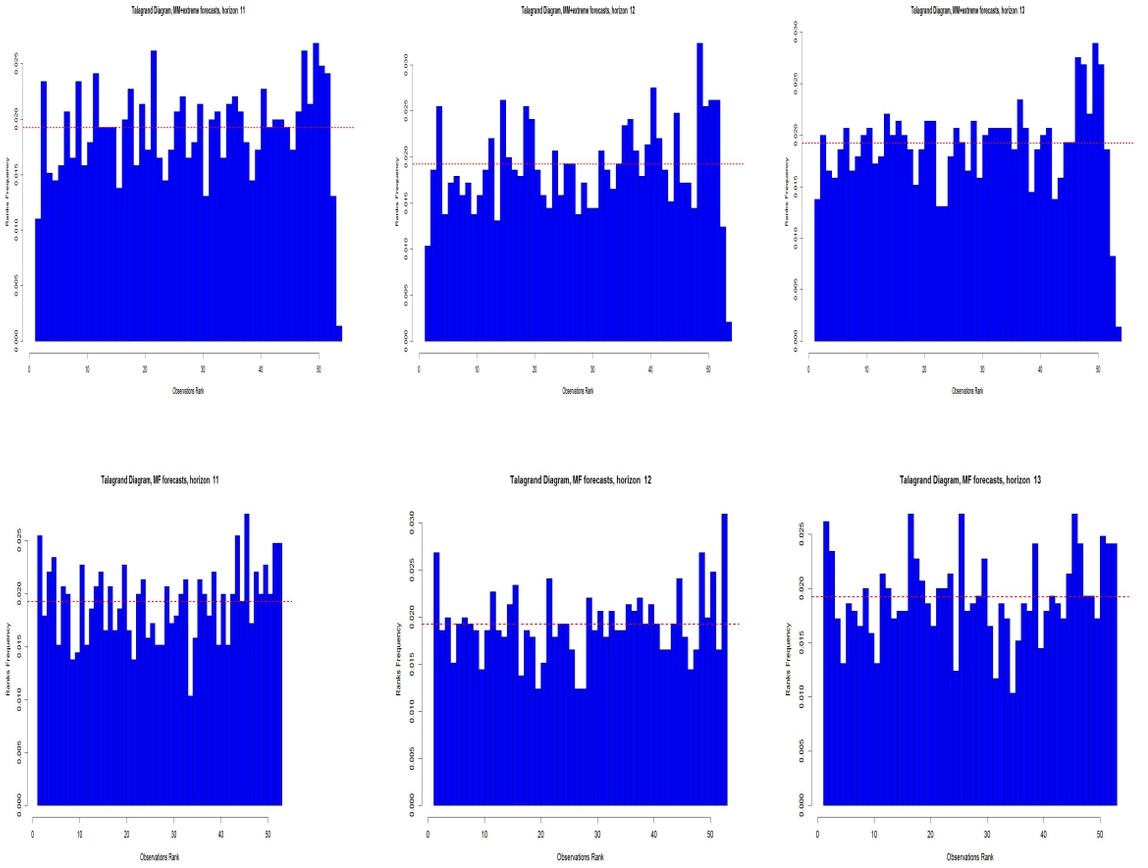


Figure 4.9: Talagrand Diagrams corresponding to the mixture model when separating extreme values by time-horizon packages and by month on top and the initial predictions at the bottom, for the time-horizons 11, 12 and 13 respectively.

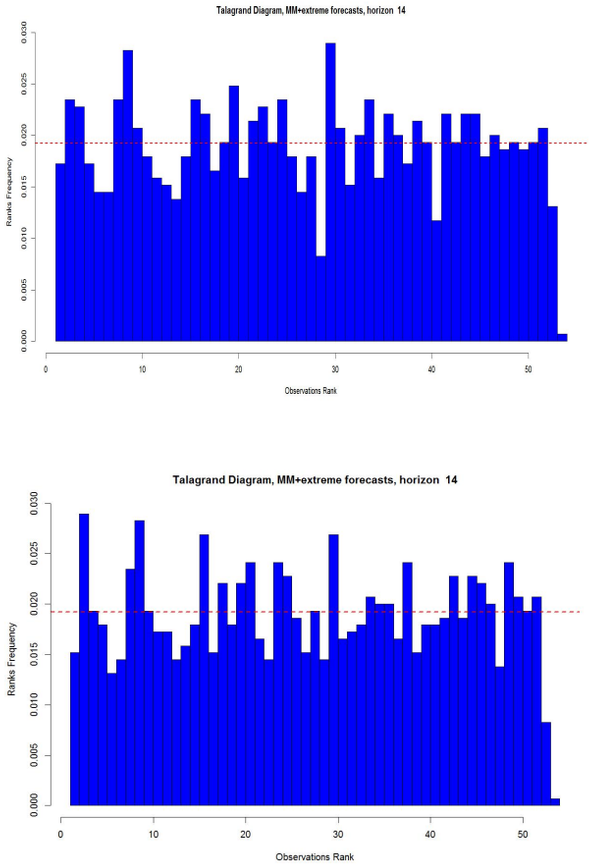


Figure 4.10: Talagrand Diagrams corresponding to the mixture model when separating extreme values by boards and at right also by horizon on top and the initial predictions at the bottom, for the 14th time-horizon.

Case of the Extreme Values given by the 1st and 51st rank

We will present here the results obtained when we consider as extreme values only the values issues of the Ranks 1 and 51. The most important thing changing is the number of the extreme values: 1129 against 3566 when we consider also the 2nd the 50th rank in the extreme values data.

Horizon	**Rank 1**	**Rank 51**	**Rank 1 + Rank 51**
5	67	93	160
6	58	68	126
7	52	64	116
8	42	65	107
9	46	51	97
10	54	57	111
11	51	52	103
12	52	56	108
13	51	49	100
14	46	55	101
All	519	610	1129

Table 4.7: If we consider as extreme values only the values coming from the rank 1 and 51 we will have the number of values in this table.

From the AIC point of view the best models is the 15th and the 16th. But even when we choose the one less separating our data (the 16th) - by months, by ranks and by packages of time-horizons- we still have certain classes with not enough data. We will have to go up to the model 23rd to have enough data in all the classes for estimating the parameters of the GEV distributions. The model 23 splits extreme values: by rank (1st and 51st) and by horizon packages (5, 6, 7 $\in PH1$, 8, 9, 10 $\in PH2$, 11, 12, 13, 14 $\in PH3$)
When we compute all the scores we studied in the thesis, we notice an improvement from a CRPS point of view and a loss from the RMSE point of view of the same order as from the case where we considered the rank 2 and 50 as well as extreme values
The Talagrand Diagram (see Fig 4) was the reason we studied this case of extreme values selection (1 and 51 ranks) but we notice the extreme right side of the diagram still has very little values represented - this is why the case is only presented in the Appendix. This is also confirmed by the quantile representation, Table 4.9.

Horizon	CRPSpot	CRPSReli	CRPS	MAE	RMSE	R2
5	0.5859	0.0040	0.5899	1.248	1.683	0.968
6	0.7383	0.0031	0.7413	1.545	2.074	0.952
7	0.9062	0.0037	0.9098	1.880	2.493	0.932
8	1.0498	0.0026	1.0523	2.159	2.843	0.913
9	1.1794	0.0022	1.1816	2.397	3.127	0.897
10	1.2937	0.0023	1.2960	2.630	3.406	0.879
11	1.3762	0.0024	1.3785	2.796	3.617	0.867
12	1.4285	0.0036	1.4321	2.924	3.763	0.856
13	1.4761	0.0025	1.4786	3.001	3.847	0.848
14	1.4864	0.0025	1.4889	3.043	3.903	0.845

Table 4.8: The different scores by time-horizon, for the forecasts obtained when creating a mixture model: BMM for the central part of the distribution and 6 different GEV functions for the two tails and three horizon packages. We notice an improvement from a CRPS point of view and a loss from the RMSE point of view of the same order as from the case where we considered the rank 2 and 50 as well as extreme values.

	AIC	rank(AIC)	BIC	rank(BIC)	sumlogliks	npars	horizon	temps	bords	nparmax
1	2884.676	22.0	3029.013	9.0	-1412.3378	30	10	405	2	24300
2	5448.000	33.5	18553.830	33.5	0.0000	2724	10	405	2	24300
3	2841.100	10.5	2869.967	2.5	-1414.5498	6	10	405	1	12150
4	5448.000	33.5	18553.830	33.5	0.0000	2724	10	12	2	720
5	2852.768	13.0	8698.429	26.0	-211.3838	1215	10	12	2	720
6	2875.914	21.0	3049.119	12.0	-1401.9570	36	10	12	1	360
7	2850.164	12.0	2907.899	6.0	-1413.0821	12	10	4	2	240
8	2841.100	10.5	2869.967	2.5	-1414.5498	6	10	4	2	240
9	5448.000	33.5	18553.830	33.5	0.0000	2724	10	4	1	120
10	2856.704	15.5	2900.005	4.5	-1419.3521	9	10	1	2	60
11	2853.981	14.0	2868.414	1.0	-1423.9904	3	10	1	2	60
12	5448.000	33.5	18553.830	33.5	0.0000	2724	10	1	1	30
13	5448.000	33.5	18553.830	33.5	0.0000	2724	3	405	2	7290
14	5448.000	33.5	18553.830	33.5	0.0000	2724	3	405	2	7290
15	2531.105	1.5	5461.153	24.5	-656.5527	609	3	405	1	3645
16	2531.105	1.5	5461.153	24.5	-656.5527	609	3	12	2	216
17	2675.473	3.0	4407.521	23.0	-977.7363	360	3	12	2	216
18	2726.410	4.5	3866.675	21.5	-1126.2050	237	3	12	1	108
19	2726.410	4.5	3866.675	21.5	-1126.2050	237	3	4	2	72
20	2943.426	27.0	3520.775	18.0	-1351.7129	120	3	4	2	72
21	2890.521	24.5	3179.196	13.5	-1385.2606	60	3	4	1	36
22	2890.521	24.5	3179.196	13.5	-1385.2606	60	3	1	2	18
23	2884.676	23.0	3029.013	10.0	-1412.3378	30	3	1	2	18
24	3552.000	29.0	12096.770	29.0	0.0000	1776	3	1	1	9
25	3552.000	29.0	12096.770	29.0	0.0000	1776	1	405	2	2430
26	3552.000	29.0	12096.770	29.0	0.0000	1776	1	405	2	2430
27	2826.971	6.5	3794.031	19.5	-1212.4857	201	1	405	1	1215
28	2826.971	6.5	3794.031	19.5	-1212.4857	201	1	12	2	72
29	2914.387	26.0	3434.001	17.0	-1349.1933	108	1	12	2	72
30	2867.141	19.5	3213.551	15.5	-1361.5707	72	1	12	1	36
31	2867.141	19.5	3213.551	15.5	-1361.5707	72	1	4	2	24
32	2865.711	18.0	3038.916	11.0	-1396.8555	36	1	4	2	24
33	2839.780	8.5	2926.383	7.5	-1401.8902	18	1	4	1	12
34	2839.780	8.5	2926.383	7.5	-1401.8902	18	1	1	2	6
35	2856.704	15.5	2900.005	4.5	-1419.3521	9	1	1	2	6
36	2858.768	17.0	8718.863	27.0	-211.3838	1218	1	1	1	3

Figure 4.11: We consider as extreme values only the values issues of the Ranks 1 and 51. From the AIC point of view the best models is the 15th and the 16th. But even when we choose the one less separating our data (the 16th) - by months, by ranks and by packages of time-horizons- we still have certain classes with not enough data. We will have to go up to the model 23rd to have enough data in all the classes for estimating the parameters of the GEV distributions. The model 23 splits extreme values: by rank (1st and 51st) and by horizon packages (5, 6, 7 $\in PH1$, 8, 9, 10 $\in PH2$, 11, 12, 13, 14 $\in PH3$)

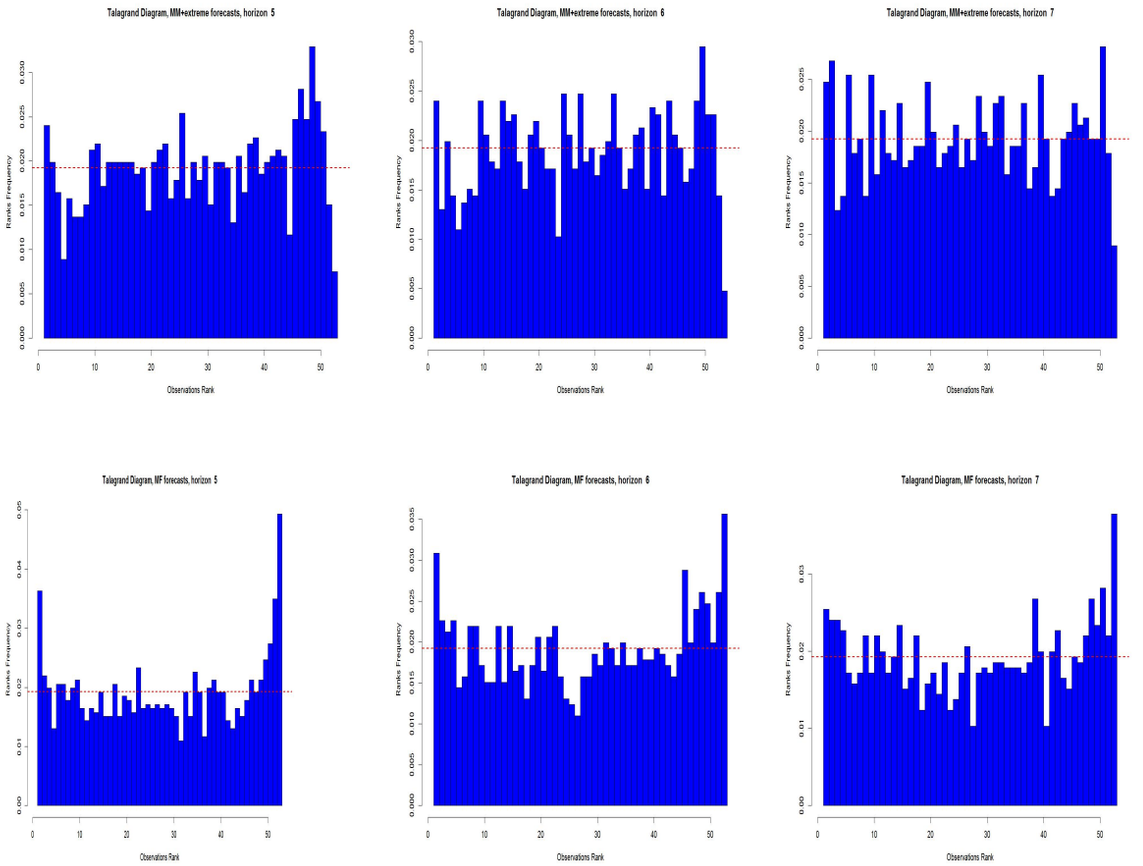


Figure 4.12: On top: Talagrand Diagrams corresponding to the mixture model (when the extreme values are those corresponding to the rank 1 and 51) indicated by the AIC criterion (the first where we have enough values for all the classes to estimate the GEV parameters). At the bottom, rank diagram for the initial predictions, for the time-horizons 5, 6 and 7 respectively. We remarque the same problem for the 51st rank, proofed also by the Table of the right quantiles bellow.

H	Model	Q1	Q99	Q2	Q98	Q5	Q95	Q10	Q90
5	Mixt	.0240	.0000	.0438	.0078	.0602	.0233	.0985	.1061
	MF	.0363	.0493	.0583	.0843	.0781	.1117	.1323	.1768
6	Mixt	.0240	.0048	.0370	.0192	.0569	.0418	.0960	.1180
	MF	.0309	.0357	.0535	.0617	.0748	.0816	.1276	.1564
7	Mixt	.0247	.0000	.0515	.0089	.0638	.0268	.1208	.0933
	MF	.0254	.0377	.0494	.0597	.0734	.0879	.1290	.1599
8	Mixt	.0151	.0014	.0309	.0206	.0467	.0412	.0968	.1120
	MF	.0192	.0357	.0405	.0536	.0659	.0769	.1312	.1477
9	Mixt	.0151	.0055	.0323	.0227	.0495	.0515	.1031	.1134
	MF	.0234	.0289	.0467	.0488	.0701	.0742	.1271	.1478
10	Mixt	.0144	.0069	.0330	.0275	.0509	.0453	.1025	.1093
	MF	.0282	.0296	.0509	.0502	.0702	.0729	.1279	.1472
11	Mixt	.0117	.0062	.0289	.0220	.0502	.0468	.0970	.1191
	MF	.0255	.0248	.0434	.0496	.0654	.0695	.1246	.1342
12	Mixt	.0110	.0090	.0289	.0227	.0482	.0468	.0964	.1274
	MF	.0269	.0310	.0455	.0475	.0654	.0723	.1198	.1357
13	Mixt	.0159	.0021	.0372	.0172	.0524	.0420	.1006	.1116
	MF	.0262	.0241	.0496	.0482	.0669	.0731	.1165	.1289
14	Mixt	.0200	.0014	.0490	.0207	.0648	.0414	.1145	.0931
	MF	.0221	.0283	.0434	.0524	.0593	.0731	.1214	.1345

Table 4.9: The quantiles 1%, 2%, 5%, 95%, 98%, 99%, for the initial MF forecasts and for the mixture model. We are in the case where the extreme values are those corresponding to the ranks 1, 51 and where the parameters of the GEV functions are computed separately by and by rank and by package of time-horizons.

List of Figures

1.1 Daily electricity loads versus the national average temperature from September 1, 1995 to August 31, 2004, at 9 AM. 23
1.2 The map of the Meteo France stations. 25
1.3 Figures corresponding to initial predictions. On top, the curve of realizations and the one of the average predicted temperature. At the bottom, the curves of temperatures for one day, for all the 14 horizons time. 26
1.4 Possible situations for the Talagrand Diagram 33
1.5 Each point on the ROC curve represents a sensitivity/specificity pair corresponding to a particular decision threshold. 35

3.1 Errors densities and empirical distributions of all the extreme values included, all ranks included (rank 1 and rank 2 with rank 50 and rank 51, and also rank 1 and rank 2 with -rank 50 and -rank 51) and left ranks and right ranks separated. . . 77
3.2 At the top: Errors densities by time-horizon, all ranks included. We notice the densities shapes look alike but theirs mean are divided between 0 and 1. At the bottom errors densities by time-horizon, by ranks, separating left and right. We notice that the densities of the left values, for different horizons are much more alike than the densities of the different horizons for the right values. 82
3.3 Errors densities and GEV distributions of the left extreme values for the four season. 85
3.4 Errors densities and GEV distributions of the right extreme values for the four season. 86
3.5 The GEV parameters for the values coming from the $B51$. We can see that the shape parameter (the most important as it decides of the case of the GEV function we are in) has a small variation from a season to another and it is always negative. 87
3.6 The GEV parameters for the values coming from the $B1$. We can see that the shape parameter (the most important as it decides of the case of the GEV function we are in) has a very small variation from a season to another and it is always negative. 88
3.7 The errors densities and their corresponding GEV distributions for the right extreme values for three packages of horizons. The first package, with the 5th, 6th and 7th horizon is the one fitting less good the distribution. 89

3.8 From the AIC point of view the best models is the 15th but it is corresponding to a very large number of parameters as it separates the extreme values by day. The estimation parameters classes does not respect the condition of the number of values large enough. The first model respecting it is the 18th and it separates the extreme values by Package of time-horizon, by month and by rank. 90
3.9 Talagrand Diagrams corresponding to the mixture model (with the GEV parameters estimated by tail and by season) on top and the initial predictions at the bottom, for the time-horizons 5, 6 and 7 respectively. 92
3.10 Talagrand Diagrams corresponding to the mixture model - that separates extreme values by tail and for the right tail, by time-horizon - on top and the initial predictions at the bottom, for the time-horizons 5, 6 and 7 respectively. 94
3.11 Talagrand Diagrams corresponding to the mixture model (GEV parameters estimated by tail, by month and by package of time-horizon) on top and the initial predictions at the bottom, for the time-horizons 5, 6 and 7 respectively. 95

4.1 Errors densities and their corresponding GEV distributions for the left extreme values for three packages of horizons. We notice that the GEV fits good enough the observed distributions, less god for the 2nd package but in the adjustment test for the three cases $H0$ is not rejected ($H0$: assumes that the values in the two distributions come from a single distribution.) 113
4.2 Talagrand Diagrams corresponding to the mixture model on top and the initial predictions at the bottom, for the time-horizons 8, 9 and 10 respectively. 114
4.3 Talagrand Diagrams corresponding to the mixture model on top and the initial predictions at the bottom, for the time-horizons 11, 12 and 13 respectively. . . . 115
4.4 Talagrand Diagrams corresponding to the mixture model on top and the initial predictions at the bottom, for the 14th time-horizon. 116
4.5 Talagrand Diagrams corresponding to the mixture model when separating extreme values by boards and at right also by horizon. On top and the initial predictions at the bottom, for the time-horizons 8, 9 and 10 respectively. 117
4.6 Talagrand Diagrams corresponding to the mixture model when separating extreme values by boards and at right also by horizon. On top and the initial predictions at the bottom, for the time-horizons 11, 12 and 13 respectively. 118
4.7 Talagrand Diagrams corresponding to the mixture model when separating extreme values by boards and at right also by horizon. On top and the initial predictions at the bottom, for the 14th time-horizon. 119
4.8 Talagrand Diagrams corresponding to the mixture model when separating extreme values by time-horizon packages and by month on top and the initial predictions at the bottom, for the time-horizons 8, 9 and 10 respectively. 120
4.9 Talagrand Diagrams corresponding to the mixture model when separating extreme values by time-horizon packages and by month on top and the initial predictions at the bottom, for the time-horizons 11, 12 and 13 respectively. 121

4.10 Talagrand Diagrams corresponding to the mixture model when separating extreme values by boards and at right also by horizon on top and the initial predictions at the bottom, for the 14th time-horizon. 122
4.11 We consider as extreme values only the values issues of the Ranks 1 and 51. From the AIC point of view the best models is the 15th and the 16th. But even when we choose the one less separating our data (the 16th) - by months, by ranks and by packages of time-horizons- we still have certain classes with not enough data. We will have to go up to the model 23rd to have enough data in all the classes for estimating the parameters of the GEV distributions. The model 23 splits extreme values: by rank (1st and 51st) and by horizon packages (5, 6, 7 $\in PH1$, 8, 9, 10 $\in PH2$, 11, 12, 13, 14 $\in PH3$) 125
4.12 On top: Talagrand Diagrams corresponding to the mixture model (when the extreme values are those corresponding to the rank 1 and 51) indicated by the AIC criterion (the first where we have enough values for all the classes to estimate the GEV parameters). At the bottom, rank diagram for the initial predictions, for the time-horizons 5, 6 and 7 respectively. We remarque the same problem for the 51st rank, proofed also by the Table of the right quantiles bellow. 126

List of Tables

3.1 The most common laws distributed by attraction domain 64
3.2 The number of values included in our study of the extreme values. 76
3.3 The p-values for the KS test, at the 5% significance level with $H0$: the values of the time-horizon i and the ones from the time-horizon j (with $i \neq j$) belong to the same distribution. 83
3.4 The KS $p-value$ tests, when taking out values from each horizon, one at the time and comparing the distribution of what is resting with the distribution of what we took out. For $B51$ $H0$ is rejected, at a 5% significance level for all the horizons excepting horizons 6, 8 and 14. For $B1$ $H0$ is not rejected and that is for all the horizons excepting horizons 11 and 13. This results confirm the test horizon by horizon. 84
3.5 The estimations parameters by season, to notice the shape parameter that is always negative, corresponding to a GEV distribution of Weibull type. 84
3.6 The different scores by time-horizon, for the forecasts obtained when creating a mixture model: BMM for the central part of the distribution and GEV functions for the tails, having their parameters estimated by tail and by season. 91
3.7 The different scores by time-horizon, for the initial MF forecasts 91
3.8 The different scores by time-horizon, for the forecasts obtained when simulating with the mixture model that separates extreme values by tail and for the right tail, by time-horizon. 93
3.9 The different scores by time-horizon, for the forecasts obtained when simulating with the mixture model selected by the AIC criterion: GEV parameters estimated by tail, by month and by horizon package. We have good scores, we improve the CRPS for all the period. 97
3.10 The quantiles 1%, 2%, 5%, 95%, 98%, 99%, for the initial MF forecasts and for the mixture models when we compute the parameters separately by month and by package of time-horizons. 98

4.1 The different scores by time-horizon, for the MF (initial) forecasts. 110
4.2 The different scores by time-horizon, for the forecasts obtained with the BMM for the entire distribution. 110
4.3 The number of values, by time-horizon and by season included in our study of the extreme values. 111

4.4 The p-values corresponding to the KS test where $H0$ is the hypothesis that the extreme values (for $board = 1$) of the j-time horizon can be consider as coming from the same distribution as the values from all the time-horizons other than j. 111
4.5 The p-values corresponding to the KS test where $H0$ is the hypothesis that the extreme values (for $board = 51$) of the j-time horizon can be consider as coming from the same distribution as the values from all the time-horizons other than j. W notice all the $p - values$, excepting horizon 6 and 14, are smaller than 0.05 so we can reject $H0$: for the rank 51 values, the extreme values coming from different horizons can not be considered as coming from the same distribution. 112
4.6 The different scores by time-horizon, for the forecast obtained when creating a mixture model: BMM for the central part of the distribution and EVT for the tails 112
4.7 If we consider as extreme values only the values coming from the rank 1 and 51 we will have the number of values in this table. 123
4.8 The different scores by time-horizon, for the forecasts obtained when creating a mixture model: BMM for the central part of the distribution and 6 different GEV functions for the two tails and three horizon packages. We notice an improvement from a CRPS point of view and a loss from the RMSE point of view of the same order as from the case where we considered the rank 2 and 50 as well as extreme values. 124
4.9 The quantiles 1%, 2%, 5%, 95%, 98%, 99%, for the initial MF forecasts and for the mixture model. We are in the case where the extreme values are those corresponding to the ranks 1, 51 and where the parameters of the GEV functions are computed separately by and by rank and by package of time-horizons. 127

Bibliography

[AF97] J. ALDRICH and R.A. FISHER, *The making of maximum likelihood*, Statistical Science **12** (1997), 162 –176.

[BBW+07] R. BUIZZA, J.R. BIDLOT, M. WEDI, N. FUENTES, M. HAMRUD, G. HOLT, and F. VITART, *The new ecmwf vareps (variable resolution ensemble prediction system)*, Q.J.R. Meterological Society **133** (2007), 681–695.

[BCG00] C. BIERNACKI, G. CELEUX, and G. GOVAERT, *Assessing a mixture model for clustering with the integrated completed likelihood*, IEEE transactions on pattern analysis and machine intelligence **22** (2000), 719–725.

[BCHY09] T. BENAGLIA, D. CHAVEAU, D.R. HUNTER, and D.S. YOUNG, *mixtools: An r package for analyzing finite mixture models*, Journal of Statistical Software **32** (2009).

[BDR05] A. BRUHNS, G. DEURVEILHER, and J-S. ROY, *A non-linear regression model for mid-term load forecasting and improvements in seasonality*, 15th PSCC, Liege„ 2005.

[BER] J.M. BERNARDO, *The concept of exchangeability and its applications*, Far East J. Mathematical Sciences **Special volume**, 111–121.

[BER72] L. Von. BERTALANFFY, *The history and satus of general system theory*, The Academy of Management Journal **15** (1972), 407– 426.

[BHG04] C.N. BEHRENS, LOPES H.F., and D. GAMERMAN, *Bayesian analysis of extreme events with threshold estimation.*, Statistical Modelling **4 (3)** (2004), 227–244.

[BLA09] P. BLANC, *Ensemble-based uncertainty prediction for deterministic 2 m temperature forecasts*, Master's thesis, Faculty of Science University of Bern, 2009.

[BRI50] G. W. BRIER, *Verification of forecasts expressed in terms of probability*, Monthly Weather Review **78** (1950), 1–3.

[BS07] William BUA and Bonnie SLAGEL, *La prévision d'ensemble expliquée*, 2007.

[BTV96] J. BEIRLANT, J. TEUGELS, and P. VYNCKIER, *Practical analysis of extreme values*, Leuven University Press (1996).

[CB09] J. CARREAU and Y. BENGIO, *A hybrid pareto model for asymmetric fat-tailed data: the univariate case.*, Extremes **12 (1)** (2009), 53–76.

[CD10] J. COLLET and V. DORDONNAT, *Aléas physiques perturbant la gestion offre-demande : état de l'art complément à la feuille de route du pôle aléas*, Tech. report, EDF R&D, 2010.

[COL01] S. COLES, *An introduction to statistical modeling of extreme values*, Springer, 2001.

[COL08] J. COLLET, *Note d'opportunité d'un partenariat et d'une thèse sur l'utilisation statistique de résultats de modèles physiques*, Tech. report, EDF R&D, 2008.

[CWS+08] B. CASATI, L.J. WILSON, D. B. STEPHENSON, P. NURMI, A. GHELLI, and M. POCERNICH, *Forecasts verification: current status and future directions*, Meteorological Applications **15** (2008), 3–18.

[DC10] V. DORDONNAT and J. COLLET, *Méthodes de prévision en loi : état de l'art*, Tech. report, EDF R&D, 2010.

[DDHDV01] J. DANIELSSON, L. DE HAAN, and C. G. DE VRIES, *Using a bootstrap method to choose the sample fraction in tail index estimation*, Journal of Multivariate Analysis **76** (2001), 226–248.

[DGT98] F. DIEBOLD, T. GUNTHER, and A. TAY, *Evaluating Density Forecasts with Applications to Financial Risk Management*, Internationa1 Economic Review **39** (1998), no. 4, 863–883.

[DKO+08] V DORDONNAT, S.J. KOOPMAN, M. OOMS, A. DESSERTAINE, and J. COLLET, *An hourly periodic state space model for modelling French national electricity load*, International Journal of Forecasting **24** (2008), no. 4, 566–587.

[DLR77] A. P. DEMPSTER, N. M. LAIRD, and D. B. RUBIN, *Maximum likelihood from incomplete data via the em algorithm*, Journal of the Royal Statistical Society **39** (1977), 1–38.

[DOC02] IFS DOCUMENTATION, *The ensemble prediction system*, Tech. report, ECMWF, 2002.

[DOC06] ———, *Part v: The ensemble prediction systems*, Tech. report, ECMWF, 2006.

[DOR09] V. DORDONNAT, *State-space modelling for high frequency data. three applications to french national electricity load*, Ph.D. thesis, Vrije Universiteit Amsterdam, 2009.

[DU07] J. DU, *Uncertainty and ensemble forecast*, Tech. report, SREF Development Team, EMC/NCEP/NOAA, 2007.

[EAT08] S. EL ADLOUNI and OUARDA T.B., *Comparaison des méthodes d'estimation des paramètres du modèle gev non stationnaire*, Journal of Water Science **21** (2008), 35–50.

[EKM97] P. EMBRECHTS, C. KLUPPELBERG, and T. MIKOSCH, *Modelling extremal events for insurance and finance*, 1997.

[EMP11] G. EVIN, J. MERLEAU, and L. PERREAULT, *Two-component mixtures of normal, gamma, and gumbel distributions for hydrological applications*, Water Resour. Res. **47** (2011).

[ENM]

[EPS69] E. S. EPSTEIN, *Stochadtic dynamic prediction*, Tellus **21** (1969), 739–759.

[FFS06] V. FORTIN, A.C. FAVRE, and M. SAID, *Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member*, Q. J. R. Meteorol. Soc. **132** (2006), 1349 –1369.

[FHR02] A. FRIGESSI, O. HAUG, and H. RUE, *A dynamic mixture model for unsupervised tail estimation without threshold selection*, Extremes **5 (3)** (2002), 219–235.

[FRGS09] C. FRALEY, A. E. RAFTERY, T. GNEITING, and J. M. SLOUGHTER, *ensemblebma: An r package for probabilistic forecasting using ensembles and bayesian model averaging*, Tech. report, Department of Statistics University of Washington, 2009.

[GAA10] COLLET J. GOGONEL A. and BAR-HEN A., *Utilisation des méthodes ensembliste pour la prévision d'électricité*, Congrès de la Société Marocaine de Mathématiques Appliquées, Rabat, 2010.

[GAA11] _____, *Ensemble prediction systems of temperature for the french electric system*, International Symposium on Forecasting, Prague, 2011.

[GAA12a] _____, *Improvement of extreme temperatures probabilistic short-terme forecasting*, International Conference on Computational Statistics, Limassol, 2012.

[GAA12b] _____, *Improvement of extreme temperatures probabilistic short-terme forecasting*, Workshop on Stochastic Weather Generators, Roscoff, 2012.

[GAAer] _____, *Implementation of ensemble prediction systems post-processing methods, for the electric system management*, The World Statistics Congress (ISI), Dublin, 2011, Poster.

[GBL+05] G. GIEBEL, J. BADGER, L. LANDBERG, H-A NIELSEN, T-S NIELSEN, H. MADSEN, K. SATTLER, H. FEDDERSEN, H. VEDEL, J. TOFTING, L. KRUSE, and L. VOULUND, *Wind power prediction using ensembles*, Tech. report, Risk National Laboratory Roskilde Denmark, 2005.

[GCAA12] COLLET J. GOGONEL-CUCU A. and BAR-HEN A., *Statistical post-processing methods implemented on the outputs of the ensemble prediction systems of temperature, for the french electric system management*, Journal Case Studies in Business, Industry and Government Statistics, Bentley University (CSBIGS) **5(2)** (2012).

[GDC97] E. GASSIAT and D. DACUHNA-CASTELLE, *Estimation of the number of components in a mixture.*, Bernoulli **3** (1997), 279–299.

[GIL09] E. GILLELAND, *An introduction to the analysis of extreme values using r and extremes*, Graybill VIII, 2009.

[GIL12] _____, *Package 'verification'*, Tech. report, NCAR - Research Application Program, 2012.

[GRWG05] T. GNEITING, A. E. RAFTERY, A.H. WESTVELD, and T. GOLDMAN, *Calibrated probabilistic forecasting using ensemble model output statistics and minimum crps estimation*, Monthly Weather Review **133** (2005), 1098–1118.

[GYH+11] M. GHIL, P. YIOU, S. HALLEGATTE, BD. MALAMUD, and P.NAVEAU, *Extreme events: dynamics, statistics and prediction*, Tech. report, Copernicus Publications, 2011.

[Hag10] Renate Hagedorn, *Post-processing of eps forecasts*, 2010.

[HER00] H. HERSBACH, *Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems*, Weather and Forecasting **15** (2000), no. 5, 559–570.

[HOU07] P. HOUTEKAMER, *Rudiments de la prévision d'ensemble*, Tech. report, Séminaires RPN, 2007.

[Ins06] SAS Institute, *The GLMSelect procedure : Stepwise selection(stepwise)*, 2006.

[JA05] L JAGER and WELLNER J A., *A new goodness-of-fit test: the reversed Berk-Jones statistic*, Tech. Report TR443, Department of Statistics, University of Washington, 2005.

[JMW09] S. JENSEN, B.B. McSHANE, and A.J. WYNER, *Hierarchical bayesian modelling of hitting performances in baseball*, Bayesian Analysis **4** (2009), 631–652.

[JS03] I. T. JOLLIFFE and D. B. STEPHENSON, *Forecast verification: A practitioner's guide in atmospheric science.*, Wiley (2003).

[JS07] ———, *Proper scores for probability forecasts can never be equitable*, American Meteorological Society (2007).

[KA09] D. KORTSCHAK and H. ALBRECHER, *Asymptotic results for the sum of dependent non-identically distributed random variables*, Methodology and Computing in Applied Probability **11** (2009), 279–306.

[KAT08] R.W. KATZ, *Background on extreme value theory with emphasis on climate applications*, Tech. report, Institute for Study of Society and Environment National Center for Atmospheric Research Boulder, CO USA, 2008.

[KB78] R. KOENKER and G. BASSETT, *Regression Quantiles*, Econometrica **46** (1978), no. 1, 33–50.

[LEI74] C. E. LEITH, *Theoretical skill of monte carlo forecasts*, Monthly Weather Review **102** (1974), 409–418.

[LLR83] R. LEADBETTER, G. LINDGREN, and H. ROOTZEN, *Extremes and related propertiesof random sequences and processes*, 1983.

[MAL05] V. MALLET, *Estimation de l'incertitude et prévision d'ensemble avec un modèle de chimie-transport*, Ph.D. thesis, Ecole nationale des ponts et chaussées, 2005.

[MAL08] ———, *Prévision d'ensemble*, Tech. report, INRIA Roquencourt, 2008.

[MBPP96] F. MOLTENI, R. BUIZZA, T. PALMER, and T. PETROLIAGIS, *The ecmwf ensemble prediction system*, Q.J.R. Meterological Society **122** (1996), 73–119.

[MF00] A. J. McNEIL and R. FREY, *Estimation of tail-related risk measures for heteroscedastic financial time series: An extreme value approach*, Journal of Empirical Finance **7** (2000), 271–300.

[ML04] B. MENDES and H.F LOPES, *Data driven estimates for mixtures*, Computational Statistics and Data Analysis **47 (3)** (2004), 583–598.

[MMP]

[MP07] LEUTBECHER M. and T.N. PALMER, *Ensemble forecasting*, Tech. report, ECMWF, 2007.

[MRT09] V. MARIMOUTOU, B. RAGGAD, and A. TRABELSI, *Extreme value theory and value at risk: Application to oil market*, Energy Economics **31** (2009), 519–530.

[MS97] A.J. MCNEIL and T. SALADIN, *The peaks over thresholds method for estimating high quantiles of loss distributions*, Tech. report, Departement Mathematik ETH Zentrum, Zurich, 1997.

[MSL+11] A. MACDONALD, C.J. SCARROTT, D. LEE, B. DARLOW, M. REALE, and G. RUSSELL, *A flexible extreme value mixture model*, Computational Statistics and Data Analysis **55** (2011), 2137–2157.

[NAS84] S.G. NASH, *Newton-type minimization via the lanczos method*, SIAM J. **21** (1984), 770–778.

[NJ11] P. NORTHROP and P. JONATHAN, *Threshold modelling of spatially dependent non-stationary extremes with application to hurricane-induced wave heights*, Environmetrics (2011).

[NPDC07] M. NOGAJ, S. PAREY, and D. DACUNHA-CASTELLE, *Non-stationary extreme models and a climatic application*, Nonlinear Processes in Geophysics **14** (2007), 305–316.

[NUR07] P. NURMI, *Mixture models*, Tech. report, Helsinki Institute for Information Technology, 2007.

[oYSS12] International Conference of Young Scientists and Students (eds.), *Statistical post-processing methods and their implementation on the ensemble prediction systems for forecasting temperature in the use of the french electric consumption.*, Department of Information Systems Sevastopol National Technical University, Sevastopol, Ukraine, 2012.

[PBM+07] T.N. PALMER, R. BUIZZA, LEUTBECHER M., R. HAGEDORN, T. JUNG, M. RODWELL, F. VITART, J. BERNER, E. HAGEL, A. LAWRENCE, F. PAPPENBERGER, Y-Y. PARK, L. VON BREMEN, and I. GILMOUR, *The ensemble prediction system - recent and ongoing developments*, Tech. report, Research Department of ECMWF and Hungarian Met Service and Budapest, Korea Meteorological Administration, Seoul ForWind, Center for Wind Energy Research, Germany, Merrill Lynch, London, 2007.

[PER03] A. PERSSON, *User guide to ecmwf forecast products*, Tech. report, ECMWF, 2003.

[PET08] T. PETIT, *Evaluation de la performance de prévisions hydrologiques logiques d'ensemble issues de prévisions météorologiques d'ensemble*, Ph.D. thesis, Faculté Des Sciences Et De Génie Université Laval Québec, 2008.

[PIC07] F. PICARD, *An introduction to mixture models*, Tech. report, Laboratoire Statistique et Genome, UMR CNRS 8071 - INRA 1152 - Univ. d'Evry, France, 2007.

[POC10] M. POCERNICH, *Verification package: examples using weather forecasts.*, Tech. report, National Center for Atmospheric Sciences (NCAR), 2010.

[RAG09] B. RAGGAD, *Fondements de la théorie des valeurs extrêmes, ses principales applications et son apport à la gestion des risques du marché pétrolier*, Mathematics and Social Sciences **186** (2009), 29–63.

[RGBP04] A.E. RAFTERY, T. GNEITING, F. BALABDAOUI, and M. POLAKOWSKI, *Using bayesian model averaging to calibrate forecast ensembles*, Physical Review (2004), 20.

[ROB] C. Y. ROBERT, *Automatic declustering of rare events*, Université Lyon 1 (ISFA).

[ROSG09] M. RIBATET, T.B. OUARDA, E. SAUQUET, and J.M. GRESILLON, *Modeling all exceedances above a threshold using an extremal dependence structure: Inferences on several flood characteristics*, Water Ressources Research **45** (2009).

[RS02] ROULSTON and SMITH, *Combining dynamical and statistical ensembles*, Tellus **55A** (2002), 16–30.

[SCH78] G. SCHWARTS, *Estimating the dimension of a model. annals of statistics*, Annals of Statistics **6** (1978), 461–464.

[SDH+07] J. SCHAAKE, J. DEMARGNE, R. HARTMAN, M. MULLUSKY, E. WELLES, L. WU, H. HERR, X. FAN, and D. J. SEO, *Precipitation and temperature ensemble forecasts from single-value forecasts*, Hydrology and Earth System Sciences Discussions **4** (2007), 655 – 717.

[STE12] A. STEPHENSON, *Package 'evd': Functions for extreme value distributions.*, Tech. report, R cran, 2012.

[Sur04] *Mémento de la sûreté du système électrique, édition 2004*, Tech. report, Réseau de Transport d'Électricité, 2004.

[SWa06] S. SAHA, W. WANG, and H.-L. PAN and, *The ncep climate forecast system*, American Meteorological Society (2006), 3483–3517.

[SWB89] H.R. STANSKI, L.H. WILSON, and W. R. BURROWS, *Survey of common verification methods in meteorology*, Tech. report, Environement Canada, Service de l'environement atmosphérique., 1989.

[TAO06] A. TANCREDI, C. ANDERSON, and A. O'HAGAN, *Accounting for threshold uncertainty in extreme value estimation.*, Extremes **9 (2)** (2006), 87–106.

[TL00] S. TADJUDIN and D.A. LANDGREBE, *Robust parameter estimation for mixture model*, IEEE Transactions on Geoscience and Remote Sensing **38** (2000), 439–455.

[VER06] J. VERDUN, *Introduction au filtrage de kalman*, Tech. report, Ecole Nationale des Sciences Géographiques, Département Positionnement Terrestre et Spatial, 2006.

[WB05] X. WANG and C.H. BISHOP, *Improvement of ensemble reliability with a new dressing kernel.*, Q. J. R. Meteorol. Soc. **131** (2005), 965–986.

[WL98] J.S. WHITAKER and A. F. LOUGHE, *The relationship between ensemble spread and ensemble mean skill*, Monthly Weather Review **126** (1998), 3292 – 3302.

[WZZ00] B. WANG, X. ZOU, and J. ZHU, *Data assimilation and its applications*, Proceedings of the National Academy of Sciences, 2000.