Predictability of Mediterranean heavy precipitation events using a 30-year hindcast dataset

Matteo Ponzano

▶ **To cite this version:**

Matteo Ponzano. Predictability of Mediterranean heavy precipitation events using a 30-year hindcast dataset. Meteorology. Université Paul Sabatier - Toulouse III, 2019. English. NNT : 2019TOU30329 . tel-03253755

HAL Id: tel-03253755

https://theses.hal.science/tel-03253755v1

Submitted on 8 Jun 2021

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 Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Matteo PONZANO

Le 12 décembre 2019

Prévisibilité des épisodes méditerranéens de pluies intenses à l'aide d'un jeu de données de 30 ans de prévisions rétrospectives

---

Ecole doctorale : **SDU2E - Sciences de l'Univers, de l'Environnement et de l'Espace**

Spécialité : **Océan, Atmosphère, Climat**

Unité de recherche :
CNRM - Centre National de Recherches Météorologiques

Thèse dirigée par
Laurent DESCAMPS

Jury

M. Romualdo ROMERO, Rapporteur
M. Florian PAPPENBERGER, Rapporteur
M. Silvio DAVOLIO, Rapporteur
M. Michaël ZAMO, Examinateur
M. Andrea MONTAGNI, Examinateur
Mme Evelyne RICHARD, Examinatrice
M. Laurent DESCAMPS, Directeur de thèse
M. Bruno JOLY, Co-directeur de thèse

Abstract

The French Mediterranean region is prone to very intense flash-flood events induced by heavy precipitation events (HPEs), which are responsible for considerable human and material damage. Quantitative precipitation forecasts have improved dramatically in recent years towards quasi-realistic rainfall estimations. Nevertheless, the proper estimation of the uncertainty associated with the physical processes representation remains a challenging issue.

In this thesis, we address the predictability of intense precipitation in the French Mediterranean region using a 30-year ensemble hindcast dataset based on the ensemble prediction system PEARP, operational at Météo-France. This reforecast system implements the same model error as PEARP, but initial and boundary conditions are differently assessed.

In order to assess the ability of the reforecast to represent the errors of the original model, we first verify this reforecast using some verification scores. The lack of initial condition perturbation makes the ensmble spread of the reforecast lower than the PEARP's one. Though probabilistic forecast scores are weak due to these set-up deficiencies, some skills are observed at 4-days lead time and for very large thresholds. However, the duration of the reforecast dataset and the resolution inherited from the operational model seem to provide enough complexity to the rainfall reforecast distributions. Two post-processing methods, based on quantile mapping and extended logistic regression techniques, are applied to the reforecast. The quantile mapping approach reduces the members biases, but the benefits in terms of probabilistic scores are lower than expected. The calibration procedure using the extended logistic regression approach leads to better probabilistic scores, both for low and large precipitation thresholds. The extended logistic regression fitted on the reforecast as a learning dataset is then applied on the operational ensemble system PEARP over a 4-month period. Though calibrated forecasts skills are not globally improved, some high probability thresholds are slightly improved, suggesting such methodology could be finally efficiently tuned.

The last part of this thesis further investigates systematic errors of intense precipitation forecasting using the feature-based metric SAL (Wernli et al., 2008). This spatial metric applied to the reforecast shows that both amplitude and structure components are controlled by deep convection parametrizations. Indeed, between the two main deep convection schemes implemented in the model, one scheme performs better, in particular for the most extreme events. A remarkable aptitude of the model is emphasised as the ranked distribution of the very intense integrated rainfall features is accurately represented by the model.

Résumé

Le sud-est de la France est une région particulièrement propice à l'occurrence de crues torrentielles associées à des événements de pluies très intenses. Ces événements peuvent causer des pertes matérielles et humaines considérables. Les techniques de prévision de ces pluies exceptionnelles ont nettement progressé et on parvient à représenter des cumuls de pluie très proches de ceux observés. Néanmoins, les incertitudes liées à la prévision de ces événements sont encore importantes et il reste nécessaire d'améliorer la connaissance des processus qui y contribuent.

Dans cette thèse, nous nous intéressons à la prévisibilité des épisodes intenses de pluie sur le sud-est de la France. Notre étude repose sur l'utilisation d'une base de prévisions rétrospectives par un système dérivé du modèle de prévision d'ensemble opérationnel PEARP, que l'on dénomme *reforecast* et d'une profondeur de 30 années. Cette version utilise plusieurs schémas physiques comme pour le système d'origine mais ne peut techniquement disposer des mêmes conditions initiales et de la technique utilisée pour les perturber.

Afin de vérifier la capacité du *reforecast* à représenter les incertitudes du système PEARP, une première partie de l'étude est consacrée à son évaluation. Le fait de ne pas avoir de conditions initiales perturbées entraîne un manque de dispersion du *reforecast* par rapport à celle de PEARP. On observe cependant une bonne qualité du *reforecast* pour des seuils de précipitation élevés et des échéances de prévision de quatre jours. Cela montre la possibilité d'extraire d'un tel système de l'information utile pour améliorer ses performances par des techniques de post-traitement ou calibrage. Deux expériences de calibrage sont ensuite menées, l'une basée sur une méthode de quantile mapping et la seconde sur une méthode de régression logistique étendue, appliquées chacune sur le reforecast. Avec la première méthode appliquée membre par membre, on améliore le biais de chacun des membres, mais on n'améliore pas les scores probabilistes. Dans la seconde expérience, le reforecast calibré donne de meilleurs scores quelque soit le seuil de définition de l'événement. Cette technique a donc été appliquée à la prévision opérationnelle, les résultats ne sont pas aussi convaincants que ceux obtenus avec le *reforecast* mais on observe tout de même une amélioration des prévisions pour les événements les plus intenses.

La dernière partie de l'étude a été consacrée à l'utilisation d'une métrique basée sur l'identification de structures cohérentes ou objets de pluie proposée par Wernli et al. (2008). On montre que le facteur prédominant de la performance du modèle réside dans le choix du schéma de convection profonde de la paramétrisation de chaque membre. Dans le cas de PEARP, ces schémas peuvent être regroupés en deux grandes familles, dont la dichotomotie se projète significativement sur la performance de la prévision. Le schéma donnant les meilleurs résultats montre la très bonne

capacité du modèle à reproduire la distribution du volume de pluie par objet pour les épisodes les plus intenses.

Acknowledgements

I want to thank my supervisors, Bruno and Laurent, for their guidance through each stage of the process. I particularly appreciate their moral and practical support, especially during the last period of this PhD. I would like to acknowledge also Philippe, who was instrumental in defining the first steps of my research.

I would like to thank the jury members: Silvio Davolio, Florian Pappenberger, Romualdo Romero for having examined this thesis and provided precious advices, Andrea Montagni and Evelyne Richard for their interest on this thesis and their suggestions during the thesis committees. I would to acknowledge Michael Zamo for some technical explanations and beneficial discussions about the verification and post-processing techniques.

I would like to thank all my colleagues at GMAP/RECYF. The coffee break was a special moment to relax and having stimulating conversations. A special thank goes to Pascal, with whom I shared the office during these three years. His presence not only provided me valuable technical support, but also made working days more pleasant.

I would like to thank all PhD students (or other) who I met along the way: Iris, Maxence, Quentin, Zied, Mary, Maxime, Léo, Thomas, Alexane, César, Clément, Rémy, Yann, Ivana, P-A, Francesca, Filipa, Michael ... I am sorry if I didn't mentioned someone, the list is very extensive. A special thank goes to Léo, Thomas and Remy: I share memorable moments with you, both at CNRM, and going out in Toulouse. This also let the last period of the PhD being less overwhelming.

I would like to thank Amélie for the encouragement given during the last period of the thesis and for bearing with me during stressful times. I would like to thank Jane for nor disturbing me and for giving me moral support during the moments of

writing of the thesis that lasted often until the small hours.

Last but not least, I would like to thank my family for their presence and their confidence in me during all the study period.

Acronyms

AEARP Météo-France ensemble data assimilation system. 42

ARPEGE Action de Recherche Petite Echelle Grande Echelle. 40

AUC Area Under the Curve. 82

B85 Bougeault (1985), scheme. 44

BS Brier Score. 74

BSS Brier Skill Score. 77

CAPE Convective Available Potential Energy, scheme. 44

CDF Cumulative Distribution Function. 30

CMC Canadian Meteorological Centre. 18

CRA Contiguous Rain Area. 28

CRPS Continuous Ranked Probability Score. 77

ECDF Empirical Cumulative Distribution Function. 99

ECMWF European Center Medium Weather Forecast. 18

ECUME Exchange Coefficients from Unified Multicampaigns Estimate. 45

EDA Ensemble-Data Assimilation. 19

EDKF Eddy-Diffusivity/Kain-Fritsch scheme. 44

EFI Extreme Forecast Index. 22

EPS Ensemble Prediction System. 18

F False Alarm Rate. 63

FAR False Alarm Ratio. 63

H Hit Rate. 63

HPE Heavy Precipitation Event. 2, 4

KFB Kain and Fritsch (1993), Bechtold et al.(2001), scheme. 44

L79 Louis, 1979, scheme. 44

LLJ Low-Level Jet. 8

MAE Mean Absolute Error. 64

MCS Mesoscale Convective System. 7

MECE Mutually Exclusive Collectively Exhaustive. 61

MODE Method for Object-based Diagnostic Evaluation. 28

NCEP National Centers for Environmental Prediction. 18

PCMT Prognostic Condensates Microphysics and Transport scheme. 44

PDF Probability Density Function. 17

PEARP Prévision d'Ensemble ARPEGE. xxx, 36

PMMC Pergaud, Masson, Malardel, Couvreux (2009), scheme. 44

QM Quantile Mapping. 98

QPF Quantitative Precipitation Forecast. 61

RMSE Root Mean Squared Error. 52, 64

ROC Relative Operating Characteristic. 82

RPS Ranked Probability Score. 115

SAL Structure-Amplitude-Location quality measure. 28

SEDI Symmetric Extremal Dependence Index. 63

SKEB Stochastic Backscatter Scheme. 20

SOT Shift of Tails. 22

SPPT Stochastically Perturbed Parameterization Tendencies. 20

SST Sea Surface Temperature. 12

SV Singular Vector. 18, 43

TKE Turbulent Kinetic Energy, scheme. 44

UKMO United Kingdom Met Office. 18

XLR Extended Logistic Regression. 98

Contents

List of Tables

List of Figures

Introduction

In autumn, very intense precipitation events can affect the French Mediterranean region. They may be responsible for considerable human and material damage, as the region is prone to torrential hydrographic watershed response. An iconic case was the Vaison-la-Romaine catastrophic flash flood (Sénési et al., 1996), which occurred the 22 September 1992 (Fig. 1). This extreme event caused 47 fatalities.

Figure 1: A dramatic picture showing two caravans carried away by the flash-flood occurred on 22 September 1992 in Vaison-La-Romaine, in the Vaucluse French departement. From: https://www.croix-rouge.fr/Actualite/Vaison-la-Romaine-Il-y-a-25-ans-2152

If long term territory adaptations should take into account learning from the climatology of these episodes, a more reliable and anticipating short-range weather warning would be immediately helpful. Improving the inclusion of all sources of forecast uncertainty in the model at early lead times would be beneficial for security procedures and help to reduce the societal impact.

In weather numerical modelling, the development of ensemble techniques is an

effective method to progress in the estimation of the errors. It is also an adapted methodology to address extreme events issues which correspond to low probability events. These events are better analysed with large sample sizes. In order to investigate the precipitation forecast predictability at a daily scale over the French Mediterranean region, one recommended approach is to use a large ensemble reforecast dataset. It corresponds to a hindcast database produced for a past period with a model close to the one devoted to the operational forecasts (Hamill et al., 2008). It has been shown that a large reforecast dataset can be beneficial for assessing and improving the performances of the model. This is particularly useful in the context of rare events, where only a few are usually observed during standard operational verification periods. This thesis takes advantage of a 30-year reforecast built with a simplified version of the Météo-France ensemble prediction system PEARP (Prévision d'Ensemble ARPEGE; Descamps et al., 2015).

The comparison between ensemble precipitation forecasts and observed precipitation over the whole period of the reforecast can be meaningful about the model skill. A further utilisation of the reforecast as a learning dataset to improve post-processing methods is also carried out. Issues in precipitation post-processing may derive from the complex statistical properties associated with this predictand variable. On another hand, they can be compensated as a large size training dataset may make the statistical estimation techniques more efficient.

In this study, a deterministic and a probabilistic calibration methods are applied on the reforecast dataset. The targeted events correspond to large daily rainfall amounts. One of the objectives of this study is to test the ability of the reforecast to be used as a training dataset to post-process the operational system PEARP. Another objective of this thesis is to analyse systematic errors in HPE forecasting, using an object oriented approach. Point-to-point verification may have inherent limitations, notably the so-called double penalty problem (Rossa et al., 2008). Based on the two-dimensional structure of the rainfall, feature-based metrics prevent from this effect. We analyse consistent errors of intense precipitation forecasts with such a metric from the reforecast dataset depending on the impacted region and on the physical parametrizations implemented in the model.

Chapter 1 presents a survey about the state of the art of intense precipitation ensemble forecasting. In chapter 2 a brief summary of the characteristics of the Operational Ensemble System PEARP is presented. A description of the reforecast and 24-hour precipitation reference datasets is also given in this chapter. In chapter 3 the analysis of deterministic and probabilistic scores applied to the reforecast is shown. Chapter 4 presents a deterministic (Quantile Mapping) and a probabilistic (Extended Logistic Regression) post-processing method applied both on the reforecast and on a 4-month period of the operational PEARP system. In chapter 5 the under-review article entitled "Systematic errors analysis of heavy precipitating events prediction using a 30-year hindcast dataset" is presented. Conclusions are given in chapter 6.

Introduction (Français)

Les événements de précipitations intenses du sud-est de la France peuvent occasionner des dégâts et pertes humaines considérables. En effet, ces événements sont souvent responsables de crues torrentielles en raison du relief particulier de la région. Un événement particulièrement marquant s'est produit à Vaison-la-Romaine le 22 septembre 1992 (Fig. 2), lors duquel on dénombra 47 victimes.

FIGURE 2: Une photo marquante de l'épisode de crues torrentielles qui toucha Vaison-la-Romaine le 22 Septembre 1992 dans le département du Vaucluse.

L'aménagement du territoire se doit de tenir compte au mieux des avancées récentes de la connaissance sur la climatologie de ces épisodes pour réduire la vulnérabilité des communes exposées. D'un autre côté, tout progrès réalisé pour améliorer l'alerte en temps réel de ces épisodes peut s'avérer très important pour la mise en sécurité des personnes et des biens.

La prévision météorologique de ces épisodes comporte encore une part d'incertitude très importante. Le développement des techniques de prévision d'ensemble

a montré que l'approche probabiliste permettra de progresser dans l'estimation des erreurs. Par ailleurs, c'est aussi une approche adaptée aux phénomènes extrêmes le plus souvent associés à des probabilités très faibles d'occurrence. Pour étudier la prévisibilité associée à ces phénomènes, nous utilisons un grand échantillon de prévisions rétrospectives, appelé reforecast. Ces reforecasts sont contruits à partir d'une version opérationnelle d'un système ensembliste en reproduisant au mieux les techniques de perturbation d'ensemble, et permettent de produire des bases de données de quelques décades destinées à apprendre et corriger les erreurs du modèle original (Hamill et al., 2008). Le fait de pouvoir évaluer un système proche du modèle opérationnel sur d'aussi longues périodes peut apporter beaucoup dans la connaissance du comportement de ce dernier. C'est aussi une grande opportunité d'obtenir des scores significatifs sur des épisodes intenses peu représentés habituellement dans les périodes courtes disponibles pour la vérification de la prévision numérique. Dans cette thèse, nous utilisons un reforecast produit sur une période de 30 années et dérivé du système de prévision d'ensemble PEARP (Prévision d'Ensemble ARPEGE) opérationnel à Météo-France (Descamps et al., 2015) comportant notamment 10 paramétrisations physiques différentes pour représenter l'erreur de modélisation.

La prévision de pluie dans le reforecast peut être évaluée soit en tant que prévision, soit comparée attentivement aux observations. Une estimation de la performance que le modèle original obtiendrait sur de telles durées, donc sur un grand nombre d'événements rares peut ainsi être étudiée. Ce même reforecast peut également être avantageusement utilisé comme échantillon d'apprentissage dans des techniques de post-processing utilisées pour améliorer le modèle orginal. Nous savons que ces techniques de post-processing ou calibrage s'appliquent difficilement à la pluie, paramètre aux propiétés statistiques particulièrement non homogènes. Nous pensons que, là aussi, la durée et la taille d'un échantillon d'apprentissage tel que celui du reforecast doit permettre de compenser cette difficulté.

Dans cette étude, deux expériences de calibrage sont effectuées. Les épisodes considérés correspondent à des cumuls de pluie quotidienne très élevés. Un des objectifs principaux est de montrer que le reforecast permet d'envisager d'améliorer la performance du modèle original grâce aux méthodes de calibrage. Un autre objectif

de l'étude est de diagnostiquer les erreurs systématiques de la prévision des épisodes de pluies intenses grâce à une métrique basée sur la structure spatiale de la pluie. En effet la prévision en points de grille, à la résolution du modèle, peut être pénalisée par des effets non désirés liés à la forte résolution, environ $10km$ dans notre cas. Cet effet est appelé « double-peine », et est décrit dans Rossa et al. (2008). Il peut être illustré par un cas d'erreur de localisation d'un maximum de pluie très localisé qui produirait une erreur négative là où la pluie a effectivement eu lieu, et positive là où le modèle la positionne. Une métrique liée à la structure spatiale de la pluie, elle, permet de compenser cet effet en considérant les deux dimensions du champs de pluie. En appliquant cette métrique nous pourrons étudier les erreurs systématiques attribuées à chaque groupe de paramétrisations physiques de PEARP et leur dépendance en fonction de la région concernée.

Le premier chapitre présente donc l'état de l'art en matière de prévision des phénomènes de précipitations intenses. Dans le chapitre 2, les caractéristiques techniques du modèle utilisé sont présentées ainsi que le traitement préalable des données pour la constitution des jeux de données du reforecast et de l'observation. Le chapitre 3 présente les résultats de l'évaluation du reforecast en tant que modèle avec les scores déterministes et probabilistes utilisés en vérification de modèle de prévision. Le chapitre 4 présente les deux expériences de calibrage, basées sur les techniques de quantile mapping et de régression logistique étendue appliquées au reforecast et au système opérationnel PEARP. Le chapitre 5 reprend l'article soumis dont le sujet est « l'analyse des erreurs systématiques de la prévision des événements pluvieux intenses à l'aide d'un reforecast sur une période de 30 ans ». Enfin les conclusions de la thèse sont présentées dans le chapitre 6.

Chapter 1

Introduction

Contents

In this introductory chapter, a description of main features related to the formation of Heavy Precipitation Events (HPEs) is proposed, followed by a focus on precipitation ensemble forecasting issues. The third section concerns model verification, scoring rules, and reviews some standard approaches in weather forecasting post-processing. Finally, the last part presents the objectives and the main addressed issues of this thesis.

1.1 The Heavy Precipitation Events

1.1.1 Characteristics of HPEs in the Mediterranean area

The Mediterranean basin is an area with peculiar geographical, morphological, historical and societal characteristics. It covers portions of three continents: Europe in the north, Asia in the east, and Africa in the south. The Mediterranean Sea is connected to the Atlantic Ocean trough the strait of Gibraltar (14.5 km wide and less than 300 m deep). Its narrow width makes the Mediterranean Sea an almost closed basin, unique of its kind.

The Mediterranean climate is extremely diverse. It is located in a transitional zone, at the interaction between mid-latitude and tropical variability (Lionello et al., 2006). On the basis of the Köppen climate classification, the southern part of the Mediterranean is characterised by a Desert climate, while its northern part is classified as Mediterranean (also known as dry summer climate). The Mediterranean climate is linked to mid-latitude variability, which strongly determines the seasonal precipitation regimes. The summer period is influenced by the subtropical ridge which keeps the atmospheric conditions very dry, with regular heat waves (Colacino and Conte, 1995). Winter is generally mild and wet, since the subtropical ridge migrates towards the equator. Precipitations are more intense and more frequent during the autumn-winter season (Mehta and Yang, 2008). At this time of the year a few cold waves can impact the basin, more likely the eastern side depending on the position of the Siberian high-pressure system.

The Mediterranean region (Fig. 1.1) has a complex morphology, due to the existence of peninsulas, islands, basins, gulfs, and mountain ranges at different eleva-

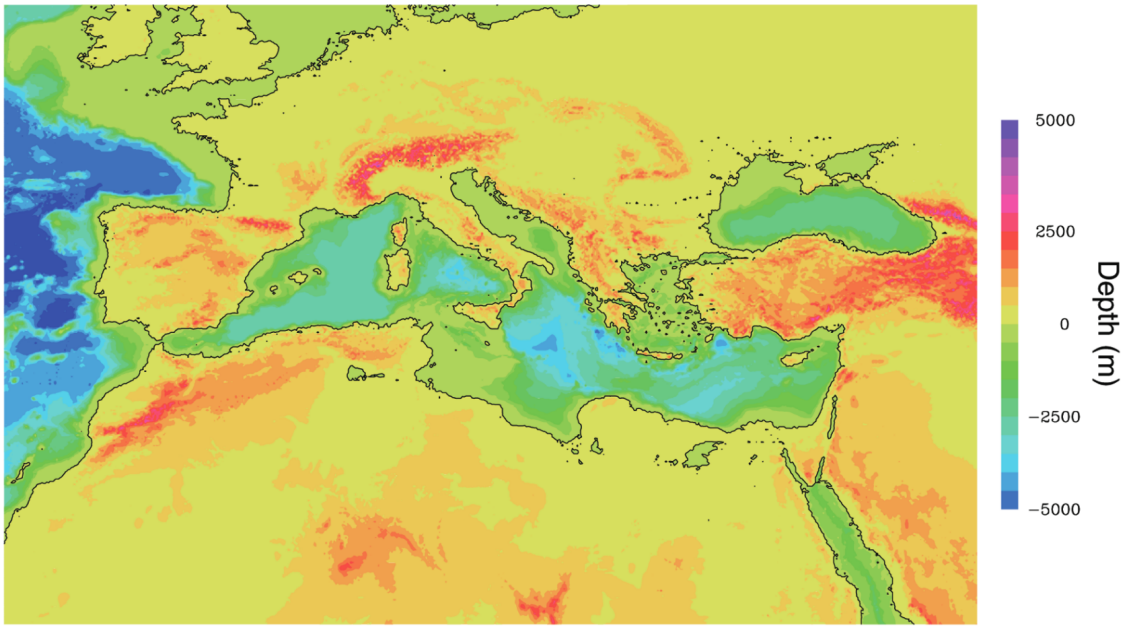


Figure 1.1: Mediterranean Basin with orography and sea-depth expressed in colors. From Lionello et al. (2006).

tions. These features influence the atmospheric circulation, as well as the mesoscale atmospheric processes, affecting the distribution and the intensity of precipitations.

Among the different classes of precipitation events, heavy precipitation regularly occurs over the Mediterranean region, during the autumn usually when the first cold cyclonic troughs from the North Atlantic enter the region. The interaction of such upper level dynamics with the warm sea and the complex morphology of the terrain makes this period propitious to the onset of these phenomena. The accumulated rainfall amounts during these events often exceed 200 mm, and rarely more than 500 mm in a single day (Ramis et al., 2013; Ricard et al., 2011). These events and their associated flash floods, are often responsible for large social and economic impacts (Llasat-Botija et al., 2007). The most severe damage occurs close to the coast coincidentally where a quite large number of densely populated cities are located. The exact location and intensity of such events are difficult to forecast due to the combination of many factors of different scales and physical processes that strongly interact. These factors, as well as a description of the main mesoscale features of Heavy Precipitation Events are detailed in this section.

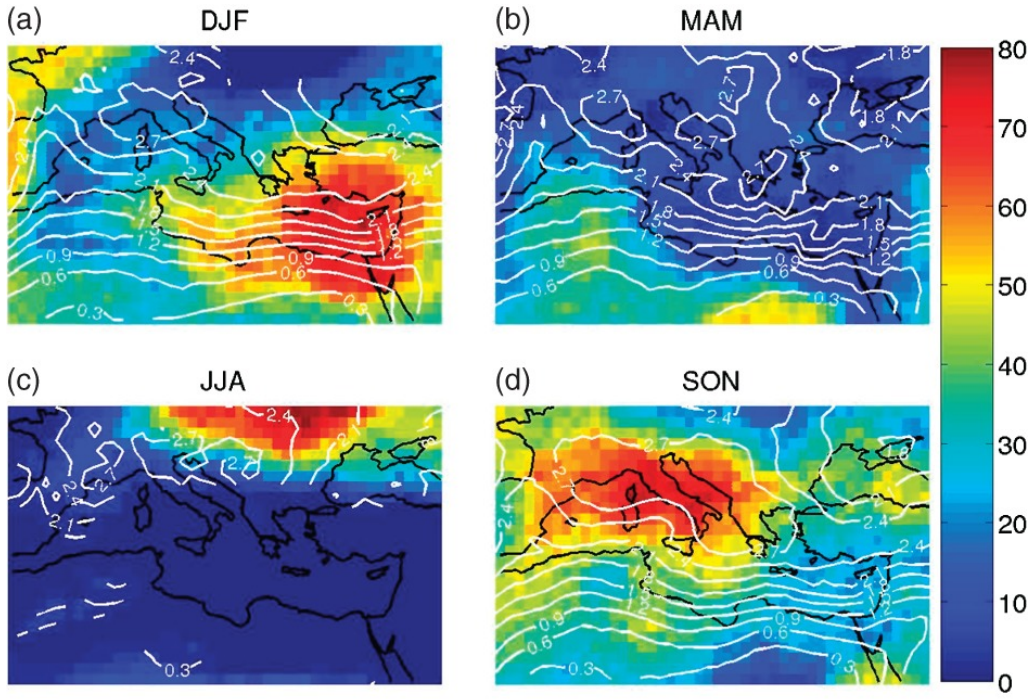


Figure 1.2: Percentage of top 50 large-scale precipitation events at each grid point in colour. The averaged precipitation of the events ($mm/6h$)) is shown in white contours; Results are presented for each season: winter (DJF), spring (MAM), summer (JJA) and autumn (SON) over the 1979–2012 period. From Raveh-Rubin and Wernli (2015).

HPEs (Heavy Precipitation Events) are often characterized by a rainfall amount that can only be reached throughout long and persistent rainfall conditions. The duration and the spatial extension directly influence the magnitude of the damage.

In the recent years, the Mediterranean basin has been prone to numerous cases of HPEs. Although there are several numerical case studies of HPEs, only a limited number of studies have addressed the climatology of these phenomena over the Mediterranean Sea. One main obstacle to HPE climatology is the necessity to have access to a dense observation network, encompassing remote mountainous regions and that this observation network homogeneous.

However, based on ERA-Interim Reanalysis, Raveh-Rubin and Wernli (2015) studied the classification of intense precipitation over the whole Mediterranean Sea. Precipitation is not extracted from in-situ observations, but integrated over a spatial scale of 1000 km and a temporal scale of three days. The detected events are found to be more extreme on the Western part of the Mediterranean and to occur pre-

dominantly during the autumn, while in the Eastern Mediterranean they occur in winter (Fig. 1.2). A limitation of ERA-Interim is that its resolution (approximately 80 km) cannot resolve heavy rainfall occurring at smaller scales. In the Spanish Mediterranean Area, a classification of 30-year rainfall events have been conducted by Romero et al. (1999) using a clustering classification method. They showed that these events impact the western part of the Spanish Mediterranean Area more frequently in winter and that they are associated with Rossby waves entering from the Atlantic. Over the eastern part, intense events dominate during the autumn in conjunction with torrential floods and are influenced by the Mediterranean dynamics. A ranking of daily precipitation records was performed by Ramis et al. (2013). This latter study showed that the most extreme rainfall cases ever recorded in Spain occurred in the Mediterranean coastlands, mainly in the Valencia region. Over Northwestern Italy, most heavy rainfall events also occurs in autumn (Pinto et al., 2013). The authors showed this period is related both with large-scale forcing (large-scale troughs) and regional forcing (higher sea surface and air temperatures), which often coexist. Precipitation intensity distribution can be also analyzed from remote sensing data, implying data retrieved from spatial satellite imagery. For instance a precipitation climatology for the Mediterranean was conducted by Mehta and Yang (2008) over a 10-year period. Based on such data source, the maximum rainfall is found over the mountain regions of Europe. Moreover, they observed that eastern Mediterranean is more rainy than the western Mediterranean, by a 20% ratio.

The role of convection in HPEs formation

During HPEs, rainfall formation is not exclusively related to convection (Anquetin et al., 2003; Ducrocq et al., 2002; Miniscloux et al., 2001). The heaviest convective rainfall usually occurs in regions of low level moisture, elevated instability, and slow movement systems (Doswell, 1987; Doswell et al., 1996).

The concept of instability is directly linked to the buoyancy in the atmosphere. Buoyancy (also called Archimedes's buoyant force) is a force exerted upon a parcel of fluid subjected to the gravitation field by virtue of the density difference between

Figure 1.3: Schematic diagram showing the motion components of an upwind-propagating mesoscale convective system. The propagation vector is directed into the low-level jet, while cells are advected by the mean cloud-layer wind. The combination of these two contributions nearly cancels the cell motion vector of the centroid of the MCS, represented by the cross symbol. From Corfidi (2003).

the parcel and that of the surrounding parcels in the atmosphere. Instability refers to the stability of the atmosphere with respect to the vertical displacement of an air parcel, controlled by the buoyancy force. Atmosphere is unstable when buoyancy force is directed upwards, resulting in an acceleration of the parcel to upper levels. Atmospheric convection then indicates these vertical motions of the atmosphere. Convection is not necessarily associated with moist processes. Specific conditions related only with horizontal termal gradients can lead to dry convection. Otherwise, upwards motions often result in condensation processes that enhance rainfall formation. It is commonly denoted as moist convection. HPEs are commonly associated with deep moist convection. The upper limit of the convection is a major factor influencing the intensity of the convection and of the associated rainfall.

Instability intensity is commonly assessed by examining the Convective Available Potential Energy (CAPE) (Doswell et al., 1998). This diagnostic is related to the acceleration rate of the vertical motion of an air parcel above the Level of Free Convection (LFC). Large CAPE values are generally associated with deep moist convection. Air masses characterized by high values of equivalent potential temperature are generally advected from the Mediterranean Sea, and they constitute a favorable condition for the enhancement of convection (Trapero et al., 2013a), because the equivalent potetial temperature of an air parcel increases with increasing temperature and increasing moisture content.

Another key aspect for the onset of convection is the low level moisture transport and convergence. Moisture convergence takes into account the effect of converging winds and moisture advection. At low level, convection tends to be formed downstream to maximum moisture transport and near where moisture convergence is high (Banacos and Schultz, 2005). Conversely, at upper level, divergence conditions, often related to a jet stream exit area, can favour upwards motions and can lead to a strengthening of convection (Maddox and Doswell, 1982).

Convection can produce scattered isolated convective cells, or, when a set of thermodynamic ingredients are brought together, can be organized to form Mesoscale Convective Systems (MCSs). An MCS can persist for several hours or more (e.g., Davolio et al., 2009; Fresnay et al., 2012; Nuissier et al., 2008; Romero et al., 2000; Trapero et al., 2013b). One organizational mode of MCSs is the squall line (Ogura and Liou, 1980), which consists in an elongated line of severe thunderstorms. In some cases, MCSs driven by large scale conditions can become stationary. It is an important factor for MCS deepening. These systems have been largely examined in many studies and often denoted as "quasi-stationary" (QS) MCS. QS MCSs are composed by following convective cells at different maturation stages. The maturation propagation direction is opposed to the cell advection vector, resulting in a quasi-cancellation effect (Corfidi, 2003; Doswell et al., 1996). A schematic diagram of this mechanism from Corfidi (2003) is shown in Fig. 1.3. This effect can induce long-duration heavy precipitation at the same location. QS MCS are very common in the Mediterranean.

Convection is not always a sufficient factor for the production of the most intense precipitation. Some mesoscale ingredients also partecipate to the onset and the enhancement of HPEs. Some of these mesoscale features are detailed in the following part.

Mesoscale ingredients associated with HPEs

A large number of HPE case studies have investigated the role of mesoscale mechanisms that contribute to the development of heavy precipitation. The mesoscale is an intermediary spatial scale used to describe certain atmospheric processes

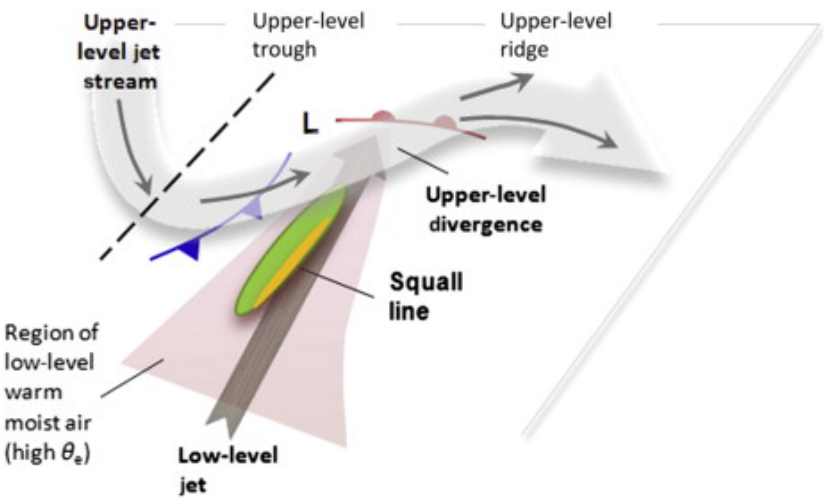


Figure 1.4: Schematic representation of the large-scale features associated with mid-latitude prefrontal squall lines. The low-level jet advects warm moist air in the region ahead of the cold front. MCSs are commonly observed in this prefontal area. From Laing (2015).

and weather systems smaller than the synoptic scale ($\approx 1000\,km$) and larger than the microscale ($\leq 1\,km$). One important difference from the synoptic scale is that at the meso-scale the vertical scale (often known as H) is no more negligible compared to the horizontal scale (often known as L). In this sense, processes at this scale may not be described using the hydrostatic approximation framework.

One mesoscale feature that is frequently involved in HPEs occurrence is the low-level jet (LLJ). This is a narrow air current found in the lower atmosphere, typically around the 850 hPa Geopotential height level. Moist low-level jets (LLJs) are important in the development of heavy precipitating systems (Buzzi and Foschini, 2000; Homar et al., 2002; Ricard et al., 2011; Romero et al., 2000), because they transport heat and moisture, and increase instability. A schematic diagram of a squall line formed in the area affected by the LLJ is given in Fig. 1.4. In most of the cases low-level winds are oriented southerly, so that moist and warm air particle are advected from the Mediterranean Sea towards the Northern lands of the Mediterranean basin.

The LLJ favours the moisture convergence in the initiation and evolution of convective systems (e.g., Delrieu et al., 2005; Ducrocq et al., 2008a; Fresnay et al., 2012). Convergence can be impacted by the presence of a complex terrain, which

can modify the circulation on the low levels (Khodayar et al., 2016).

The presence of a cold pool dynamics can take part to the triggering of the uplift and consequently of the convection. This mesoscale feature can develop, due to the HPEs dynamics themselves. For example, a vigorous precipitative downdraft can induce latent heat cooling due to the evaporation of precipitation within subsaturated air. This downwards flow both initiates or feeds the cold pool, whose location and extension can also be influenced by the presence of orography. Cold pool dynamics are documented in Bresson et al. (2009) or Ducrocq et al. (2008b). In the latter study, three case studies of HPEs were examined. Where a cold pool plays an important role on the intensity and the localization of intense precipitation. A synthetic drawing elaborated by the authors is shown in Fig. 1.5. The first case (Cévennes case) is a pure orographic precipitation case and no cold pool is formed. The Gard case is characterized by the formation of a cold pool ahead of the Cévennes chain and upstream of the low-level flow. This example shows how the convective cells are triggered in a flat area, forced by the lifting induced by the low-level cold pool with analogy to the orographic effect conditions. The Aude case shows simultaneous cold pool and orographic forcing conditions.

As it has been suggested in these processes analysis, heavy precipitation can also widely impacted by the orography. The interaction between precipitation and orography is described thereafter.

The interaction with the orography

The orography is a factor which plays a major role in the triggering of the convection by compelling upward motion of the air mass. This interaction with the relief occurs quite frequently in the Mediterranean, as a consequence of its complex morphology (see Fig. 1.1). Moreover, the presence of a steep slope raises the risk of occurrence of associated flash floods.

A precipitation climatology covering the European Alps was presented by Frei and Schär (1998). In this study the authors showed that the observed precipitation amounts in the Alps region do not necessarily increase as a function of the height. In fact, high rainfall accumulation are found along the primary flanks of the mountains

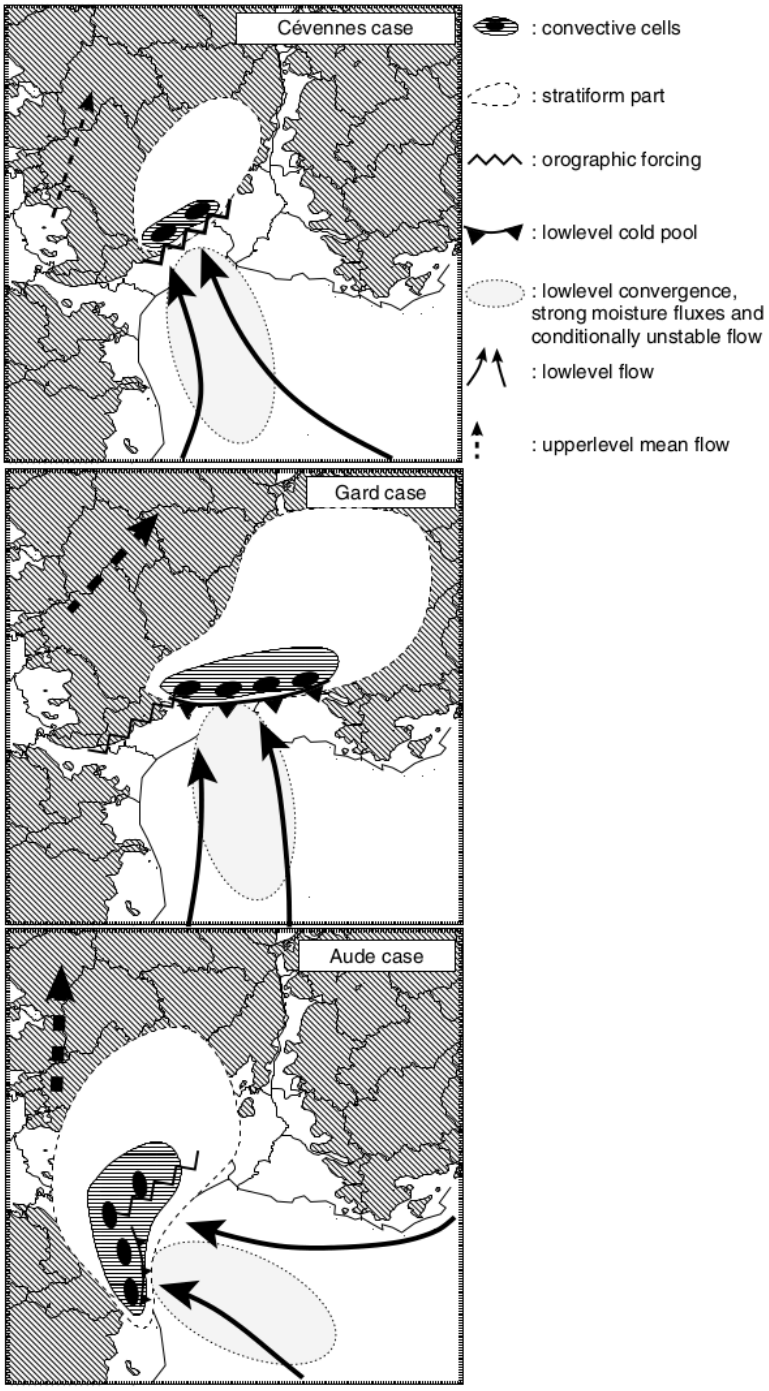


Figure 1.5: Schematic diagram of the precipitating structures of the MCSs and the mesoscale features of three flash-flood events. From Ducrocq et al. (2008a).

chains. This effect, known as orographic lifting is related to the interaction of a moist low-level streamflow with an orographic barrier (Buzzi et al., 1998). In this paper the role of the orography in the case of the Piedmont flood of 1994 is addressed. The authors show that, in addition to the blocking effect of a steep barrier, mountain chains play also a crucial role at a larger scale, by modifying the pressure field of the associated cyclone in the Mediterranean and by confining the southern prefrontal flow in the Po valley, upstream to the Alpine region.

In addition to the increase of precipitation amounts, the presence of a relief can play a role in the convection mechanisms. Houze (2012), shows that moist deep convective systems are affected by channeling of airflow near mountains. This is confirmed in (Ricard et al., 2011) with an example of a strengthening of the southern low-level flow in a channel between the Massif Central and the Alps .

The moist air flow lifting along the slope of a mountain chain, can enhance the convection, but it can also favour the stationarity of precipitation systems, supplied by a moist LLJ, as shown in the Cévennes case of Fig. 1.5 (Ducrocq et al., 2008a).

Several studies (e.g., Bresson et al., 2012, 2009; Corfidi, 2003; Miglietta and Rotunno, 2009, 2014) demonstrate that orographic effects interact with some aforementioned mesoscale features, resulting in a complex set of feedback mechanisms. In addition, mountains ranges could participate in the blocking of the cold pool (Ducrocq et al., 2008a) and can deflect the flow around them (Bresson et al., 2012; Buzzi and Foschini, 2000).

As already mentioned, the onset of intense precipitation is associated with warm and humid flows.

The interaction with the Mediterranean Sea

The interaction between the low-level flows and the Mediterranean Sea has an important impact on the HPEs. The precipitating systems are supplied in energy by the sensible and latent heat fluxes over the Sea. We present here how this complex interaction influences the location and the amplitude of the precipitating systems.

An experiment conducted by Buzzi et al. (1998) consisted in removing the surface heat and moisture fluxes in the 1994 Piedmont flood case simulation. It resulted in

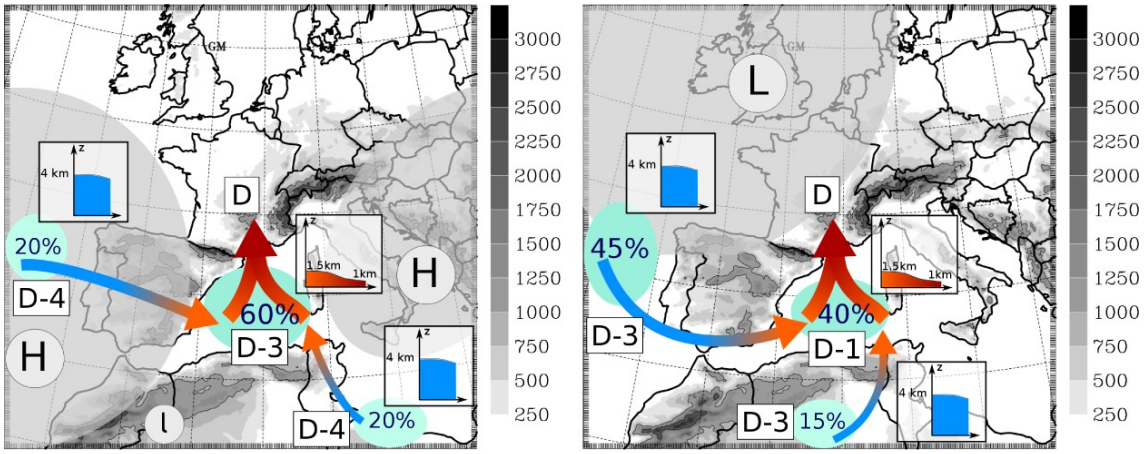


Figure 1.6: Conceptual schemes for the moisture supply to HPEs with both anticyclonic (on the left) and cyclonic (on the right) conditions prevailing a few days before the event. The contribution of the different moisture sources is indicated in percent, in blue. The arrows show the moisture transport. Their colour corresponds to the vertical extent of the flow. From Duffourg and Ducrocq (2013).

a depletion of 10% of precipitation over the Alps, a suppression of the convection over the sea and a reduction of 40% of precipitation over the Ligurian Apennines. The low depletion over the Alps is mainly related to the fact that air parcel that are subject to orographic uplift are so high that retrieving this effect has a little impact. Moreover, as suggested by the authors, the surface fluxes from the Mediterranean play a central role in preconditioning moist and warm conditions at low levels.

Lebeaupin et al. (2006) showed a sensitivity study of three torrential rainfall events over France to the sea surface temperature (SST) of the Mediterranean. Two cases are associated with QS MCS and one other case with a slow moving frontal perturbation. For the QS MCS cases, a higher (lower) SST increases (inhibits) the convection, resulting in larger (weaker) precipitation amounts. The increased precipitations are also related to larger values of equivalent potential temperature over the sea. Larger CAPE values produce a larger horizontal extension of the convection. Another factor which explains the increase of precipitation is that latent and sensible heat fluxes, in particular under the low-level jet, are significantly larger when STT is warm. For the third case, the sensitivity to the SST is less important.

Several studies have shown that the moisture sources involved in HPEs does not only originate from the Mediterranean Sea, but also from the Atlantic Ocean

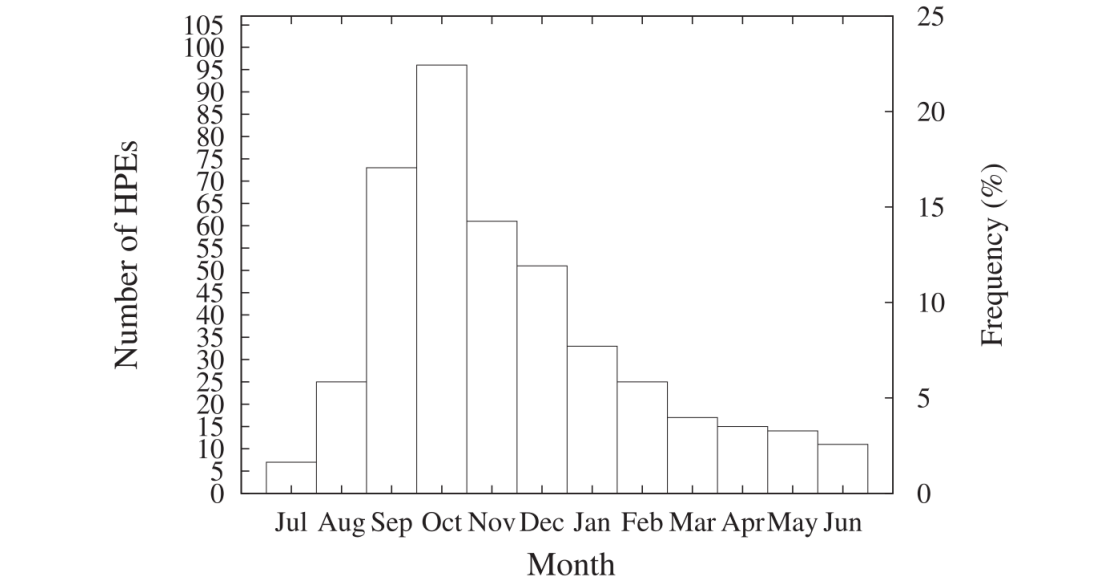
and from tropical Africa sources (Brossier et al., 2013; Duffourg and Ducrocq, 2011, 2013). The latter study shows that the oceanic contribution can reach 45% for HPE with cyclonic conditions prevailing before the event. If anticyclonic conditions the major contribution is provided by the Mediterranean Sea. More specifically, for both situations, moisture transport from the Atlantic has a larger vertical extension than in the Mediterranean. More details are given in Fig. 1.6.

1.1.2 Climatology of HPEs in France

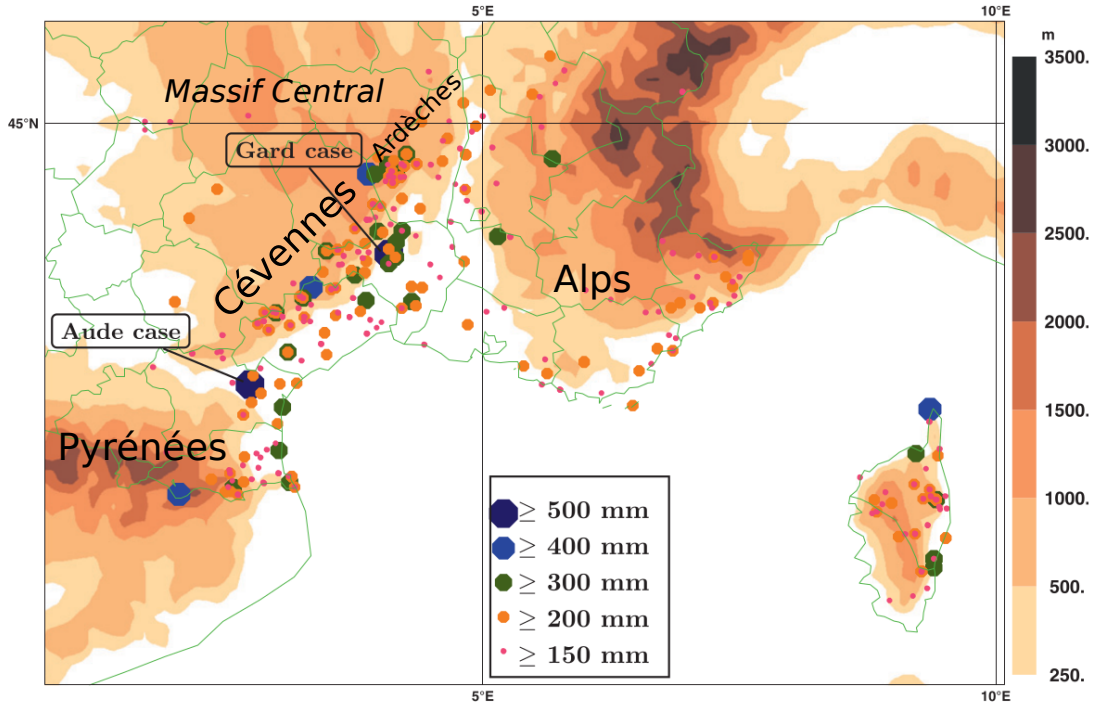
In this part details about the spatial distribution and temporal frequency of HPEs in France are reported. Some case studies about extreme events are also presented.

For the Southern France, including the Corsica region, Ricard et al. (2011) built a mesoscale reanalysis database, covering a 5-year period. The objective of their paper was to study mesoscale features associated with HPEs. They also computed the frequency of HPEs, defined as daily rainfall which exceeds 150 mm, for the period 1967-2006 based on the whole French rain-gauges network (Fig. 1.7(a)). They show that HPEs have a great inter annual variability with a maximum of occurrence in Autumn, from September to December. Figure 1.7(b) shows the spatial distribution of HPEs over southern France during the same period. The complex orography favours the creation of convergence zones and the triggering of convection, following some mechanisms described above. It can be seen that the major events are located on the southern and eastern sides of the mountain ranges, which face the moist flows advected from the Mediterranean Sea. Cévennes and Ardèche areas concentrate most of the HPEs, including some records. Blanchet and Creutin (2017) confirmed that extreme rainfall events in these specific sub regions are frequent. Rainfall events over the Alpine area tend to be less extreme, as they rarely exceed 200 mm.

One of the two most extreme episodes in the pre-Alpine region was the Vaison-La-Romaine event, with a very intense rainfall of 220 mm of rain in 3h. This event was studied by Sénési et al. (1996), with a 10-km resolution model. They show that a cold pool dynamics causes the quasi-stationarity of the two precipitation systems. A case impacting the Gard area (see Fig. 1.7(b)) has been examined by Delrieu



(a)



(b)

Figure 1.7: (a) Monthly distribution of HPEs from the period 1967-2006. (b) Locations of daily precipitation maxima above 150 mm for the HPEs occurring over southern France during the periods 1967–2006. The two figures are adapted from Ricard et al. (2011).

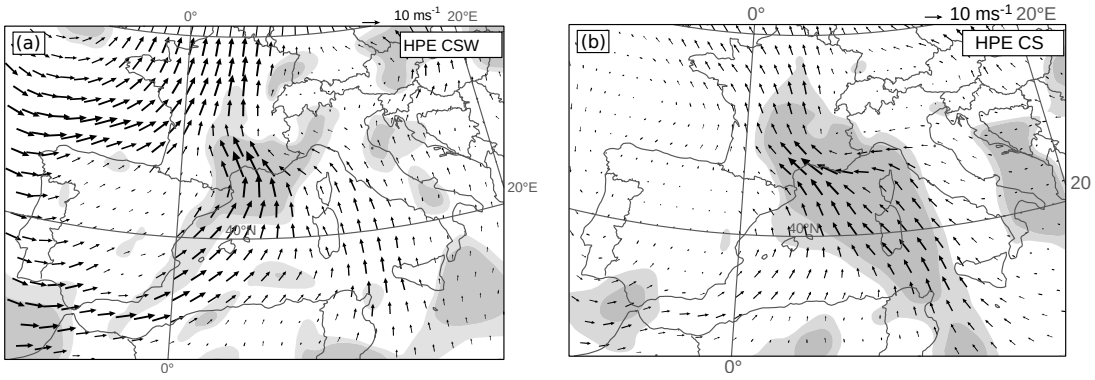
et al. (2005) using radar and rain gauge analyses in conjunction with hydological data, and a comprehensive sensitive study of the simulation of three torrential rain events in French Mediterranean has been conducted by Nuissier et al. (2008) and Ducrocq et al. (2008a).

Among the classical heavy precipitation situations associated with the orographic interaction, the "Cévenol" events are worth quoting. Although a large part of HPEs are associated with deep convection, some events are also associated with stratiform precipitation impinging the Massif Central, also known as Cévenol events.

1.1.3 HPEs and the large scale circulation

The mesoscale features, the convection, as well as the interaction with the orography require some specific large scale configurations to be set or maintained. Some HPE case studies have shown the role played by a trough or a cut-off positioned west of France (Fresnay et al., 2012; Nuissier et al., 2008; Ricard et al., 2011). This cyclonic circulation induces a southeasterly flow that reaches the French Mediterranean region. At the surface, a low pressure is often present over Spain and its anticyclonic counterpart is found over the Eastern or Central Europe. This circulation configuration favours the enhancement of a southerly low-level flow over western Mediterranean.

These results can be generalized as shown by Beaulant et al. (2011); Nuissier et al. (2011); Plaut et al. (2001); Plaut and Simonnet (2001), and Vrac and Yiou (2010). Among the cited articles, Nuissier et al. (2011) performed a classification of four synoptic types of large scale patterns for heavy rainfall in southern France, using a k-means clustering method during the period 1960-2001. Two specific circulations, the Cyclonic southwesterly (CSW) and the Cyclonic southerly (CS) represented 78% of the HPEs. For both these configurations, a large and strong upper-level low extends southward towards the Western Mediterranean. Figure 1.8 shows the composite of horizontal wind at 925 hPa for these two circulations associated with HPEs. For the CWS circulation, an intense low-level jet is present along the Spanish coast, with a maximum magnitude over the French Mediterranean coast. For the CS circulation, a southeasterly flow originates from southeastern Tunisia to reach



(a) (b)

Figure 1.8: Composites of horizontal wind (arrows) at 925 hPa for CSW (a) and CS (b) circulations associated with HPEs. The light and dark shaded areas indicate regions where the differences between the composite features associated with HPEs and non-HPEs (not shown) are significant for a Student t-test at confidence levels 90% and 95%, respectively. Adapted from Nuissier et al. (2011).

the northern Mediterranean coast, with maximum magnitude. Peters and Roebber (2014) confirmed the impact of synoptic-scale on the rainfall forecast error, in particular in terms of positioning.

In this section we have proposed an overview on the main interactions and mechanisms at different scales which contribute to the onset of heavy precipitations. These ingredients have a great impact on the predictive skills of HPEs. In the case of extreme events, taking into account all the various scales of the involved processes, ensemble global forecasting constitutes a proper way to address HPE forecast uncertainties.

1.2 The ensemble weather forecasting approach

In this section we describe the main key characteristics of an ensemble prediction system and some issues on heavy precipitation forecasting are outlined.

Weather prediction requires a sufficient accurate description of the initial conditions as well as a sufficient accurate representation of the physical laws of the atmosphere in order to produce a reliable forecast of the evolution of the atmosphere. Although operational systems are in constant evolution and improvement, HPEs deterministic quantitative precipitation forecasting is not likely to be solved

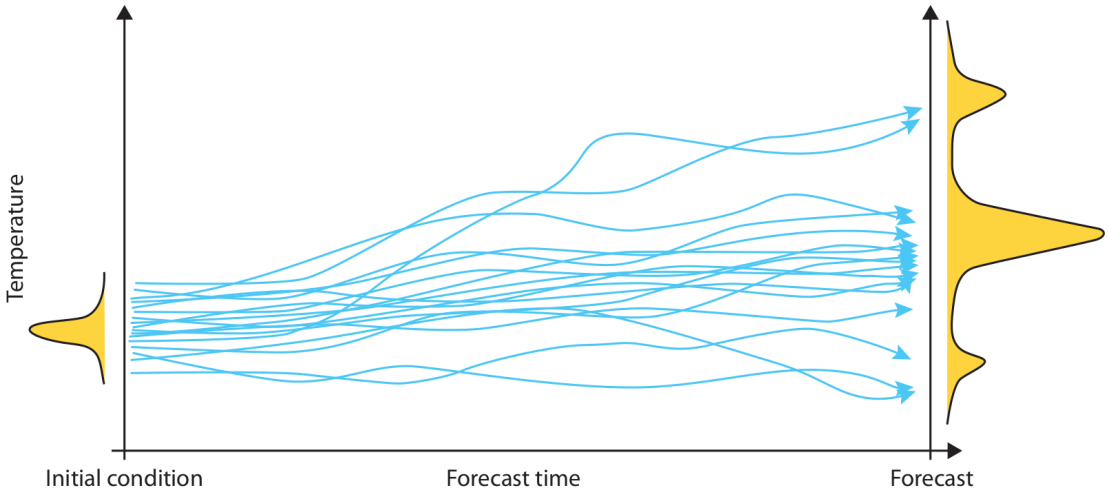


Figure 1.9: Schematic view of the concept of ensemble prediction. Each blue line corresponds to an individual forecast performed with a slightly different initial condition. The two curves on the left and right-hand sides of the figure represent the underlying PDF of the atmospheric state. From `https://www.ecmwf.int/sites/default/files/the_ECMWF_Ensemble_prediction_system.pdf`

within the next years. One source of error is the uncertainty related to the assimilation of observations in the analysis (Bauer et al., 2015). Since the pioneering works of Lorenz (1963) and Epstein (1969b) on the chaotic nature of the atmosphere, it is well known that errors in initial conditions and imperfections in the model formulation limit the skill range of a single deterministic forecast. It then appears essential to use another approach for weather forecasting, no longer based on a single prediction but on the representation of the PDF (Probability Density Function) of the atmospheric state, at any time of the forecast.

The first conceptual idea, which led to the foundation of a probabilistic approach of weather prediction, was proposed by Epstein (1969b). He derived the so-called stochastic-dynamic equations from the Liouville equation in order to represent the probability distribution of the state of the atmosphere. This approach is not sustainable with the current computing power. Leith (1974) introduced the *Monte Carlo* method, which consists in selecting a sufficiently large number of different initial conditions in order to sample the initial condition uncertainty. He showed that this approach is a practical approximation of the stochastic–dynamic equations.

The weather services ensemble prediction systems are still based on the *Monte*

Carlo approach. Figure 1.9 shows a schematic view of the ensemble forecasting technique. At the beginning of the forecast, an ensemble of initial states is built. It aims at sampling the underlying uncertainty of the atmospheric state (represented by the yellow PDF on the left-hand side of the figure). Each initial state is then evolved using a weather prediction model. A representation of model uncertainty should also be included in the forecast. At the final time of the forecast, the ensemble of states samples the underlying uncertainty of the predicted atmospheric state. The PDF does not necessarily assume normal values and it can be multi-modal, as in the example of Fig. 1.9.

1.2.1 The Ensemble Prediction Systems

The first Ensemble Prediction Systems (EPSs) were implemented in the early 1990s at the European Center for Medium Weather Forecast (ECMWF, in 1992; Molteni et al., 1996) and at the National Centers for Environmental Prediction (NCEP; Toth and Kalnay, 1993). Today global ensemble predictions are commonly used in most of the major operational weather prediction centers, including GEFS at the NCEP (Toth and Kalnay, 1997), EPS at ECMWF (Palmer, 2019), MOGREPS at the United Kingdom Met Office (UKMO Bowler et al., 2008a), EPS at the Canadian Meteorological Centre (CMC; Charron et al., 2009). Météo-France has developed its global short-range ensemble prediction system known as Prévision d'Ensemble ARPEGE (PEARP; Descamps et al., 2015).

Below the usual techniques for initial and model uncertainties representations are described.

Representing initial uncertainties

Most of the techniques used in operational ensemble forecasting systems to represent initial uncertainty are based on two main ideas. One is to span most of the fastest developing modes in the short-term forecast; the other is based on the proper representation of the uncertainties of the initial state of the forecast (Magnusson et al., 2008). The so-called Singular Vectors (SVs) approach is of the first type and it is built on the maximization of perturbations growth rate over a time

window around the initial time. A specific norm is used to measure the perturbation amplitudes (Buizza and Palmer, 1995; Buizza et al., 1993). The SVs approach is based on the idea that different directions of the phase space of the system are characterized by different amplification rates and that at least one part of the initial uncertainty will project onto the most unstable modes represented by the first SVs. Restrictions of this approach are linked with the underlying hypothesis of linearity of the evolution of the initial errors.

Trying to properly represent uncertainties in the initial state of the forecast is the aim of Ensemble-Data Assimilation approaches (EDA). Building a reliable analysis from various sources of observations is an optimization problem based on the discrepancy between a forecast and the observations. In these recent years, this challenging issue has been processed with ensemble techniques to better estimate observations co-variance errors terms in the assimilation. EDA can be based on the so-called variational framework of data assimilation (Desroziers et al., 2014) or within the Kalman filter framework (Evensen, 2003). In EDA systems that use perturbed observations, each member of the ensemble assimilates pools of observations that have been slightly perturbed, using for example their error measurement variances. Assimilating perturbed observations then produces perturbed analyses that produce perturbed backgrounds. The cycling process of ensemble data assimilation will then generates and propagates flow-dependent errors that represent the uncertainties of the atmospheric evolution. Unlike SVs, EDA does not require an hypotesis of linearity.

Both SVs and EDA approaches to represent initial uncertainties have drawbacks. Although theoretically appealing, the EDA technique may suffer from a lack of representation of all the sources of uncertainties that exists in the data assimilation process. This could result in a lack of dispersion of the ensemble. It should be mentioned that in practice, and to counteract this possible under-dispersion, ECMWF and Météo-France ensemble prediction systems are based on the use of the two approaches together.

Representing model uncertainties

Another source of uncertainties originates from the forecast model itself, which do not perfectly describe all the atmospheric processes. Sources of model error can arise from the dynamics (e.g., discretization, time-integration, transport, stabilization,...), others are related to the physical parametrizations (e.g., convection, clouds and microphysics, boundary layer,...) and coupled processes (e.g., land-surface, sea-ice,...). Physical parametrizations are developed in order to represent sub-grid processes. One approach is to use stochastic perturbations of one scheme inside the model. We cite random parameter perturbations (Bowler et al., 2008b), the stochastic backscatter scheme (SKEB; Shutts, 2005; Tennant and Beare, 2014) or the stochastically perturbed parametrization tendencies (SPPT; Buizza et al., 1999; Palmer et al., 2009; Sanchez et al., 2016). The random parameter perturbations approach "randomly" perturbs some selected parameters of the physical parametrization schemes implemented in the model. The SPPT does not perturb the parametrization itself, but the global tendencies provided by the scheme. The SKEB attempts to add a fraction of the dissipated kinetic energy back into the numerical model to compensate the excessive dissipation due to numerical diffusion and parametrized subgrid scale processes.

A different approach to take into account the model uncertainties is the multiphysics approach (Charron et al., 2009; Descamps et al., 2011). It is based on the use of several physical parametrization schemes. The stochastic methods and the multiphysics approach are conceptually different (Palmer, 2000). The multiphysics approach tries to estimate different mean values obtained from the parametrizations for any given atmospheric conditions. The stochastic approach perturbs a single mean value in order to increment the variability for the given atmospheric conditions. They are often considered mutually exclusive.

In the multiphysics approach the experimental selection of the physical schemes is not performed to select the best schemes, but to take into account all the possible sources of model error. Therefore, the selected parametrization schemes are commonly different in terms of formulation, or they can differ for some specific parameters to which the predictive variables are sensitive to. Météo-France ensemble

prediction system PEARP uses this kind of approach.

1.2.2 Intense precipitation prediction using the ensemble forecasting approach

The small spatial and temporal scales involved in the rainfall phenomenon can lead to poorly accurate deterministic forecasts (Cherubini et al., 2002; Fritsch and Carbone, 2004; Sukovich et al., 2014). Probabilistic forecasting using the ensemble forecast approach can outperform a deterministic forecast by providing an estimation of the spatial and temporal uncertainties, as well as the most probable rainfall amount. A set of predicted values is more informative and skillful than a single prediction generated by a deterministic forecast, because it provides a better estimation of the predictive value of the target variable. Therefore, the predictive skill can be extended to larger lead times than a deterministic forecast. Moreover, an ensemble forecast is able to assess the tails of a predictive PDF, a desired property when addressing to extreme events.

A review about the utilization of EPS for flood forecasting has been produced by Cloke and Pappenberger (2009). They concluded that the use of the probabilistic approach brings added value to the ensemble forecasting, but that forecasts skill need further improvement. Thielen et al. (2009) showed that a medium-range ensemble forecasts has been able to anticipate a risk of flood in Romania 9 to 11 days ahead. Schumacher and Davis (2010) examined heavy rainfall predictability over 5-day periods in the central and eastern United States using the ECMWF global ensemble forecasting system. In almost all cases, the ensemble provides very skillful 5-day forecasts. An inter comparison between different ensemble systems with varying configurations and spatial resolutions was performed by Herman and Schumacher (2016) for extreme precipitation forecast. They showed that the coarse Global Ensemble Forecast System performed as well as, or better than, the high-resolution system for extreme precipitation forecast.

The predictive PDF generated by a probabilistic forecast can be exploited to evaluate the probability of occurrence of a specific rare event over a selected area

(e.g., on a given model-grid point, a specific location or a selected region). Some indexes have been defined in order to measure how extreme a given probabilistic forecast is with respect to the model climatology. This approach requires the production of a sufficiently large sample of forecasts, including forecasts performed in the past, in order to have a proper estimation of the model climate. For example, the Extreme Forecast Index (EFI; Lalaurette, 2003) and the Shift of Tails (SOT; Zsótér, 2006) are designed to compare the estimated PDF of the issued forecasts to the PDF of the model climate.

It is worth mentioning that convection-permitting ensemble prediction system are developing since a few years (Clark et al., 2016; Hagelin et al., 2017; Hally et al., 2014; Schwartz et al., 2015; Vié et al., 2012; Vincendon et al., 2011). Frogner et al. (2019) discusses the scale-dependent predictability of precipitation produced by a convection-permitting ensemble and the added value of a higher resolution convection-permitting ensemble prediction systems with regard to a global EPS. Results showed that for high-precipitation thresholds high resolution ensemble systems outperform coarser resolution EPS, but only at short lead times. For this reason, convection-permitting ensemble prediction systems are considered as the state-of-the-art models for the probabilistic forecasting of precipitation at short ranges. However, at medium-range predictability is still mainly assessed using global EPSs, which are based on the hydrostatic approximation assumption.

1.3 Verification and post-processing methods

Forecast verification has always been included in the development of models. It is a necessary step to precisely learn about potential weaknesses that have to be improved, as well as to provide information as a support for decision-making or economic value.

The first part of this section is dedicated to the description of two categories of precipitation verification. The first considers the forecast probability at each point used for the comparison and the second assesses the spatial accuracy of the predicted precipitation fields. The second part introduces some traditional post-

processing methods used to calibrate raw ensemble precipitation forecasts with the aim of improving predictive skill.

1.3.1 Verification of probabilistic forecasts

Ensemble forecasts are generally used to produce probability distributions, whose quality depends on its absolute skill and its spread. The skill is related to the accuracy of the forecast, that is how the forecast is accurate with regard to observation, while the spread should represent the uncertainty of the forecast.

Following Murphy and Winkler (1987), the joint distribution of forecasts and observations provides the basis of a general framework for forecast verification. In practical settings, both the forecasts and observations are commonly discrete variables. Following the Murphy and Winkler (1987) framework, the joint distribution $p(y, o)$ can be expressed through the calibration-refinement factorization

$$p(y, o) = p(o|y)p(y), \tag{1.1}$$

where $p(o|y)$ is the conditional probability of the observation o, given the forecast y. $p(y)$ is the marginal distribution of the forecast, which specifies how often a given forecast is provided. The likelihood-base rate factorization is defined as follows

$$p(y, o) = p(y|o)p(o), \tag{1.2}$$

where $p(y|o)$ consists in the probability of the forecast y, given the observation o. The unconditional distribution $p(o)$ specifies how often a given observation is issued and is generally referred to as the sample climatological distribution. The characteristics and the relationships between all these distributions are investigated in the verification by means of classical tools or scoring rules.

The probabilistic verification (Casati et al., 2008; Murphy, 1990; Murphy and Winkler, 1992) relies on the evaluation of some attributes of a probabilistic forecast:

- The *Reliability* measures the statistical coherence between the reference and the associated forecast. A perfect probabilistic forecast is achieved when

$p(o|y) = y$, meaning that the average frequency of event occurrence equals the forecast probability for every probability category.

- The *Resolution* qualifies the ability of a forecast system to classify different observations depending on the forecast probability. It consists in the average of the squared difference between $p(o|y)$ and the sample climatology $p(o)$. A model that for each probability categories always predicts events that occur with a frequency equal to the sample climatology shows no resolution.

- The *Bias* corresponds to the difference between the expected value of the forecast and the observation. A forecast system is unbiased if $E(o) = E(y)$.

- The *Discrimination* represents the ability of a probability forecast system to vary depending on the observation occurrence. This property is related to the probability of the forecast, conditioned by the observation $p(y|o)$ and to the probability of the observation $p(o)$.

- The *Sharpness* is inherent to the variability of the forecast itself, more specifically to the unconditional distribution of the forecast $p(y)$. Forecasts that frequently deviates from the climatology are sharp. The sharpness is a necessary but not sufficient property for having a perfect forecast, as it has to be reliable at the same time.

- The *Uncertainty* is equal to the climatological distribution and it depends only on the $p(o)$ probability.

Since the number of forecasts and observations is always finite, it is not possible to completely explore the space of the joint distribution $p(y, o)$. As a general practice, forecasts and/or observations are sampled from a population which can be partitioned into sub populations, depending for example on the geographical area or the lead-time. This method is the so-called *stratification* process.

Based on the Murphy and Winkler (1987) framework, different scoring rules have been developed based on more or less complex combinations of the verification attributes previously described. Some scoring rules assess the whole forecast distribution with a continuous diagnostic, like the CRPS score (Candille and Talagrand,

2005). Other scores consider discrete events often defined as the exceeding of a specific threshold, like the Brier Score (Brier, 1950). In this latter case, the metric tends to degenerate to perfect values when the frequency of occurrence of the event tends to zero, making difficult the verification of rare events. The practice of selecting extreme observations implies a stratification of the observations before the verification. However, Lerch et al. (2017) warns about some unwanted effects associated with this specific approach, raising questions about the problematic defined as the "Forecaster's Dilemma". In fact, the restriction of the evaluation to subsets of the available observations (the most extreme ones, in this case) may discredit forecasts (Gneiting and Ranjan, 2011). A forecaster may give less importance to a potential predicted extreme event, when the verification for these events lead to inconsistent verification scores. One solution consists in using some specific weighted scoring rules, which focus on the tail of the distributions. However, research about weighted scoring rules is still ongoing. For example, Taillardat et al. (2019) showed that a weighted version of the CRPS for extreme events still generates undesirable effects on the quality of verification.

1.3.2 Precipitation verification using spatial approaches

It is common in probabilistic verification to estimate the attributes described above by applying dedicated scoring rules. This can be carried out over a river catchment (Roulin and Vannitsem, 2011), some grid points (Hamill, 2012) or some location of the observations (Roulston and Smith, 2003). These approaches, especially when applied to intense events, are subject to both timing or position errors leading to low scores (Gilleland, 2012; Mass et al., 2002). This combination of errors is also known as the double penalty problem (Rossa et al., 2008). Spatially aggregated verification techniques have been developed with the goal to evaluate forecast skill in a manner similar to a forecaster approach and to overcome the traditional grid-point to grid-point verification limitations.

Among spatial verification the neighbourhood technique methods (also known as fuzzy methods) (Ebert, 2009; Mittermaier, 2013; Roberts and Lean, 2008; Skok and Roberts, 2016; Zepeda-Arce et al., 2000) aim to upscale the constraint related to

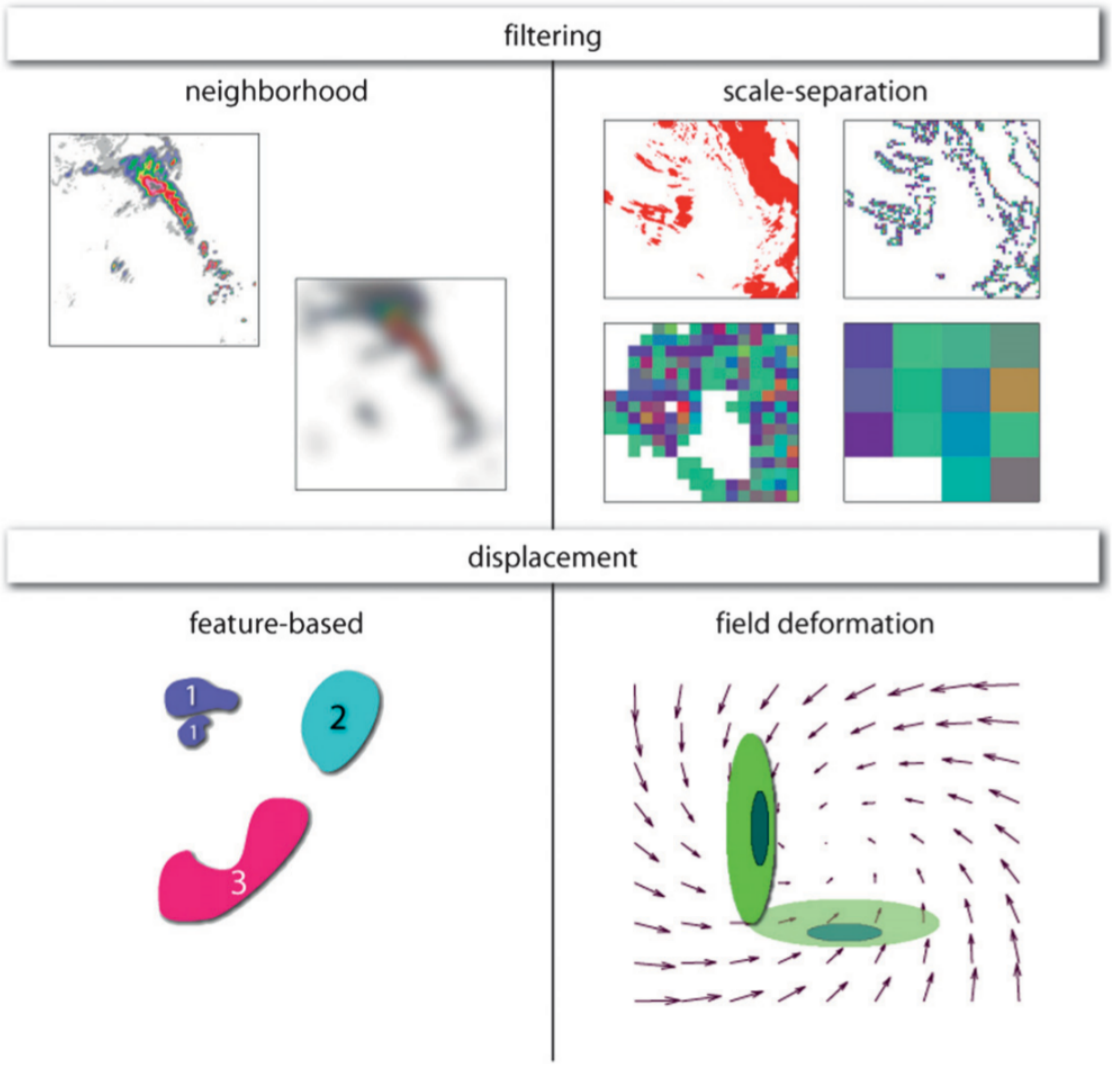


Figure 1.10: Schematic representations of the four categories of spatial verification methods: neighbourhood (top-left), scale separation (top-right), feature-based (bottom-left) and field deformation (bottom-right) methods. From Gilleland et al. (2009).

the point-to-point comparison to a larger space-time neighbourhood. This upscaling procedure results in an effect similar to the processing of a filter. The top-left panel of Fig. 1.10 gives an example of the neighbourhood approach applied to a rainfall field, giving a smoothed version of the original one. The spatial averaging filters out the smallest scales, reducing the time-position errors associated with the point-to-point verification.

Another spatial verification technique is the scale separation approach (Briggs and Levine, 1997; Lack et al., 2010; Weniger et al., 2017; Yano and Jakubiak, 2016). Similarly to the neighbourhood methods, this approach filters the spatial signal by selecting different scales and isolating the features at each scale of interest. The top-right panel of Fig. 1.10 shows an example of using the scale separation on a large-scale storm (red contours). In contrast to the neighbourhood approach, the scale separation allows to recombine the filtered fields to reproduce the original field.

Neighbourhood and scale separation methods are based on the estimation of errors at different spatial scales, but they do not consider displacement errors. This error contribution can be assessed using deformation methods (Gilleland, 2011; Keil and Craig, 2007, 2009; Venugopal et al., 2005). A forecast field is warped through the observation field on the basis of the minimization of a selected verification score. The transformed field is then evaluated, taking into account the displacement error. A graphical representation of a field deformation application is shown in the bottom-right panel of Fig. 1.10.

A fourth approach for spatial verification is the feature-based method that takes into account displacement errors and structure errors at a given scale (AghaKouchak et al., 2011; Davis et al., 2006, 2009; Ebert and McBride, 2000; Lack et al., 2010; Mittermaier et al., 2015; Nachamkin, 2009; Wernli et al., 2009, 2008). This spatial verification method is based on features identification (also called objects) from both the forecast and the observation fields. The following step consists in verifying (using a specific diagnostic or summary measure) the features at a given scale, depending on the process of selection of the feature. The bottom-left panel of Fig. 1.10 shows an example of identification of three features. It could be noted that the shape of object 1 results from the presence of two connected local maxima.

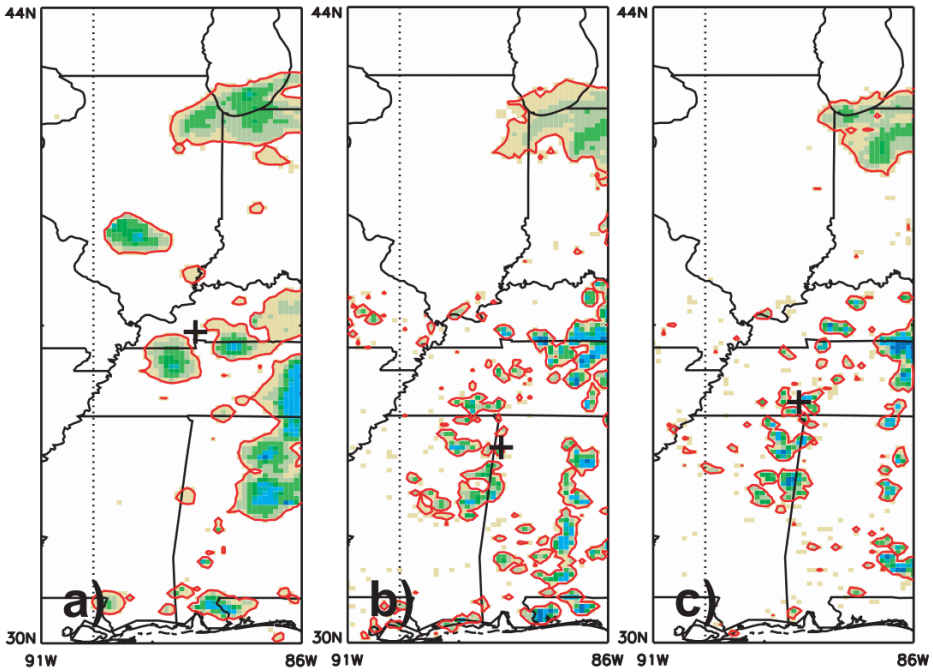


Figure 1.11: Example of feature definition used in the computation of the SAL quality measure for the observation (a) and two forecasted fields (b) and (c). The black plus signs denote the center of mass of the precipitation field in the domain. From Wernli et al. (2009).

Among the feature-based methods it is worth mentioning the Contiguous Rain Area (CRA; Ebert, 2001, 2008; Ebert and McBride, 2000; Weckwerth et al., 2004). The CRA considers a region bounded by a specific precipitation threshold. The displacement is evaluated as the translation of the forecasted feature to the observed feature until a pattern-matching criterion is met. The displacement error is then decomposed into three error metrics: location, rain volume and pattern errors. Another well-known verification metric is the Method for Object-based Diagnostic Evaluation (MODE; Davis et al., 2006, 2009; Gallus, 2010; Mittermaier et al., 2015). First, a convolution procedure with a smoothing process is applied to the field. Secondly, a threshold is applied to the convolved field in order to identify the features. The objects are then merged and matched on the basis of some attributes computed from the features such as the separation distance and the spatial orientation. A third method, often used for precipitation spatial verification is the Structure-Amplitude-Location quality measure (SAL; Leoncini et al., 2013; Wernli et al., 2009; Zacharov et al., 2013; Zimmer et al., 2008). In this approach, introduced by Wernli et al.

(2008), features are constituted as areas where a threshold equal to a fraction of the rainfall amount daily maximum (Wernli et al., 2008), or of the 95^{th} percentile of all grid-point values (Wernli et al., 2009) is exceeded. Figure 1.11 shows an example of feature definition using as threshold $1/15^{th}$ of the 95^{th} percentile. The SAL measure is condensed into three attributes: amplitude (A), structure (S) and location (L). The A component is associated with the total field, while the S and L components are feature-based. An important difference with respect to the other methods is that SAL approach does not require a one-to-one matching between the observed and forecasted features. More details about the SAL score will be given in chapter 5.

1.3.3 Post-processing methods for ensemble precipitation forecasts

Although some skill improvements occurred during the recent years (Haiden et al., 2015), medium-range ensemble forecasts can still suffer from underdispersion and bias errors. Post-processing methods are designed to statistically characterize the errors of a system using past forecasts and use these learning to calibrate the current ensemble prediction system. This goal can be achieved by building calibrated predictive distributions.

All calibration methods need to be learned on a training dataset to get robust estimations of forecast errors. An attractive alternative to forecast archives, which suffer from system instability, is to use a reforecast dataset, which covers a large period (Hamill, 2012; Hamill et al., 2013; Hamill and Whitaker, 2006; Hamill et al., 2004; Scheuerer and Hamill, 2015; Schmeits and Kok, 2010). Reforecasts are forecasts produced with a current operational version of the model, but integrated for a past period. These large-sized and computationally expensive forecasts are often produced for research applications, notably to verify model skill or to test calibration methods. Reforecasts can also be used to have a proper estimation of the model climate for the computation of some extreme indices as EFI or SOT, mentioned in section 1.2.2.

Before the application of some probabilistic calibration methods it is common practice to first attempt to reduce the ensemble bias. It can be done by adding a constant correction to each ensemble member of the raw forecast. Another alternative is to make use of the Quantile Mapping technique (QM; Hamill et al., 2017; Hamill and Scheuerer, 2018; Hopson and Webster, 2010a; Maraun, 2013; Voisin et al., 2010). This method computes differences between Cumulative Distribution Functions (CDFs) drawn from observations and forecasts to correct each ensemble member forecast. This procedure is considered as a deterministic calibration method because it is separately applied to each member of the ensemble.

Probabilistic calibration of precipitation forecast can be particularly challenging (e.g. in comparison to temperature) because of some specific attributes of the probability distribution of this variable. Precipitation statistical distributions cannot be approximated by Gaussian laws and gather some peculiar properties: 1) it is non-negative, 2) the null probability has to be prescribed separately from the rest of the distribution, and 3) forecast uncertainty typically increases with the magnitude of precipitation amounts, making the distribution right tailed.

A comprehensive review of the state-of-the-art of statistical postprocessing of ensemble forecasts is beyond the scope of this study. Some of the most frequently used univariate post-processing methods for precipitation are described hereafter: Bayesian Model Averaging (BMA; Raftery et al., 2005), Ensemble Dressing (ED; Roulston and Smith, 2002), Nonhomonegeous Gaussian Regression (NGR; Gneiting et al., 2005), Logistic Regression (LR; Hamill et al., 2004) and Analogue Method (AM; Hamill and Whitaker, 2006).

The BMA method models a probability distribution, also called kernel, constructed around each of the debiased values of the raw members. The calibrated distribution then is a weighted sum of these kernels. The estimation of the weights is achieved through a maximum likelihood technique (Raftery et al., 2005) on the training dataset. Kernels can be the same or they can be differently selected for each members. In this case the parameters estimation technique is performed to compute the standard deviation of each kernel, in addition to the weights. If members are exchangeable, weights and kernels are constrained to be equal from a member to

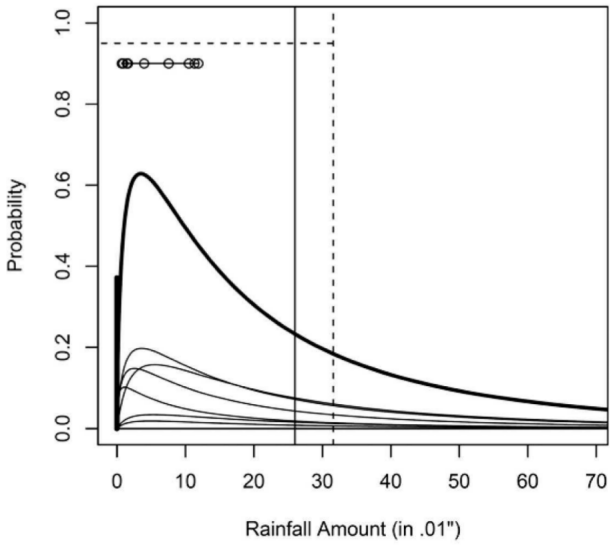


Figure 1.12: Example of BMA-fitted distribution. The thick vertical line at zero represents the BMA estimate of the probability of no precipitation and the upped solid curve represents the probability distribution for nonzero amounts, resulting in a contribution of the gamma distribution (lower curves) dressed around each member (dots). The dashed vertical line represents the q_{90} quantile upper bound of the BMA PDF; the dashed horizontal line is the respective prediction interval. From Sloughter et al. (2007).

another (Wilson et al., 2007). Original formulation of BMA considered kernels normally distributed. To take into account the non-Gaussianity of precipitation BMA is adapted by applying specific distribution laws to the kernels. Sloughter et al. (2007) and Schmeits and Kok (2010) use a gamma distribution for nonzero precipitation and the probability of non-precipitation is estimated through a logistic regression technique. Figure 1.12 shows an example of BMA predictive distributions for 24-h accumulated precipitation. It is interesting to observe that raw members range between zero and 15 mm, while the right tail of the calibrated distribution extends towards larger values. A calibrated forecast can provide a continuous probability distribution, more informative about the right tail of the precipitation probability.

Ensemble dressing method (Roulston and Smith, 2003; Wang and Bishop, 2005) is a kernel density smoothing technique (Wilks, 2009b), closely linked to the BMA method. In this approach, the kernels constructed around each members are the same as well as the weight used to compute the calibrated probability function. Therefore, the calibrated distribution reduces to an average of kernels. The standard

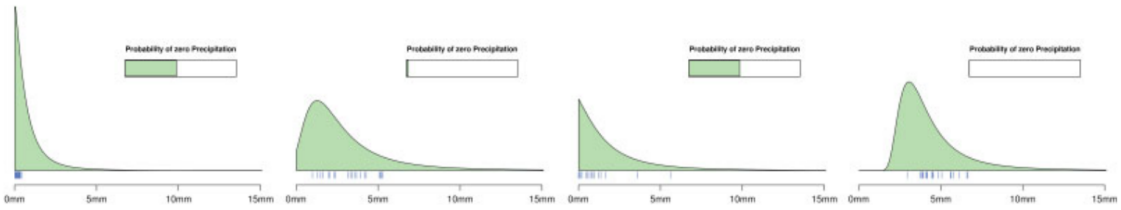


Figure 1.13: Predictive distributions of precipitation forecasts using censored nonhomogeneous regression. The short vertical lines below the densities represent the ensemble member forecasts. From Scheuerer (2014).

deviation assigned to the kernels can be estimated following different procedures, for example the "best" member dressing (Roulston and Smith, 2003), or the second-moment constraint dressing (Wang and Bishop, 2005; Wilks, 2006). In the "best" member dressing, the best member is defined as one with the lowest distance to the observations. Then, the standard deviation of the kernels is modelled using the RMSE between the observation and this best member in the training dataset. In the second-moment constraint dressing, the standard deviation of kernels is modeled from the training dataset using a combination of the RMSE computed between the ensemble mean and the observations together with the average of the ensemble variances. One limitation of the ensemble dressing, compared to the BMA, is that it is only suitable for underdispersive forecasts, because the sum of the kernels is only able to reduce the ensemble spread. Conversely, in BMA, the parametrization of the weights and the standard deviations associated to each different kernel allows to correct both an overdispersed and an underdispersed raw forecast.

The Nonhomonegeous Gaussian Regression (NGR) is a regression-based method (Gneiting et al., 2005; Jewson, 2003). The calibrated distribution of precipitation corresponds to a Gaussian probability function whose mean is defined as an optimized linear combination of the ensemble members or some corresponding statistics, which provide for one or several predictors. The standard deviation of the calibrated distribution is modeled using a linearly corrected value of the ensemble spread of the raw ensemble. The parameters used to model the mean and standard deviation of the calibrated distribution are estimated by minimization of a given score (CRPS or ignorance score) in the training dataset. Some specific nonhomogeneous regression methods have been developed to address the specific features of the distribution

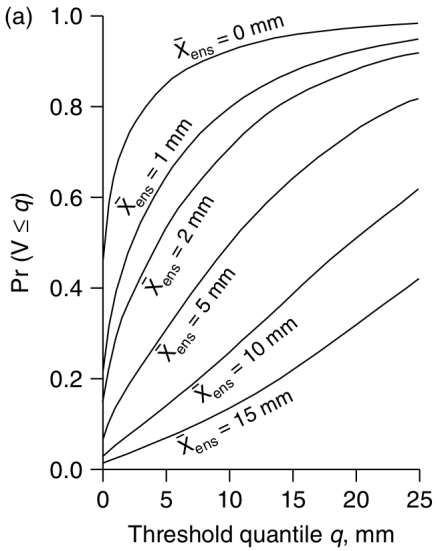


Figure 1.14: Predictive cumulative distributions from XLR method application. Each curve is evaluated at selected values of the ensemble precipitation mean predictor. From Wilks (2009a).

of precipitation, notably that strictly non-negative precipitation commonly exhibits large discontinuities in its probability density at zero. A truncated distribution can be used to describe the non-zero probability of precipitation (Hemri et al., 2014). Another further regression largely used for the post-processing of the precipitation forecast is the censored regression (Baran and Nemoda, 2016; Gebetsberger et al., 2017; Scheuerer, 2014; Scheuerer and Hamill, 2015; Stauffer et al., 2017). In censoring, unlike truncation, it is possible to model the non-precipitation probability by assigning to exactly zero any probability corresponding to negative precipitation values. Figure 1.13 shows some examples of the use of left-censored Generalized Extreme Value (GEV) distributions as calibrated distribution laws. The horizontal bar shows the probability of zero precipitation.

Another well-known regression-based post-processing method is the logistic regression (Applequist et al., 2002; Hamill et al., 2008, 2004; Lemcke and Kruizinga, 1988; Sohn et al., 2005; Sokol, 2003; Vislocky and Young, 1989; Wilks, 2006; Wilks and Hamill, 2007). It is a regression model from the generalized linear model framework to model the conditional probability of binary events. The binary event is defined as the exceeding of a specific precipitation threshold and regression coefficients are estimated using the maximum likelihood approach. The drawback of this

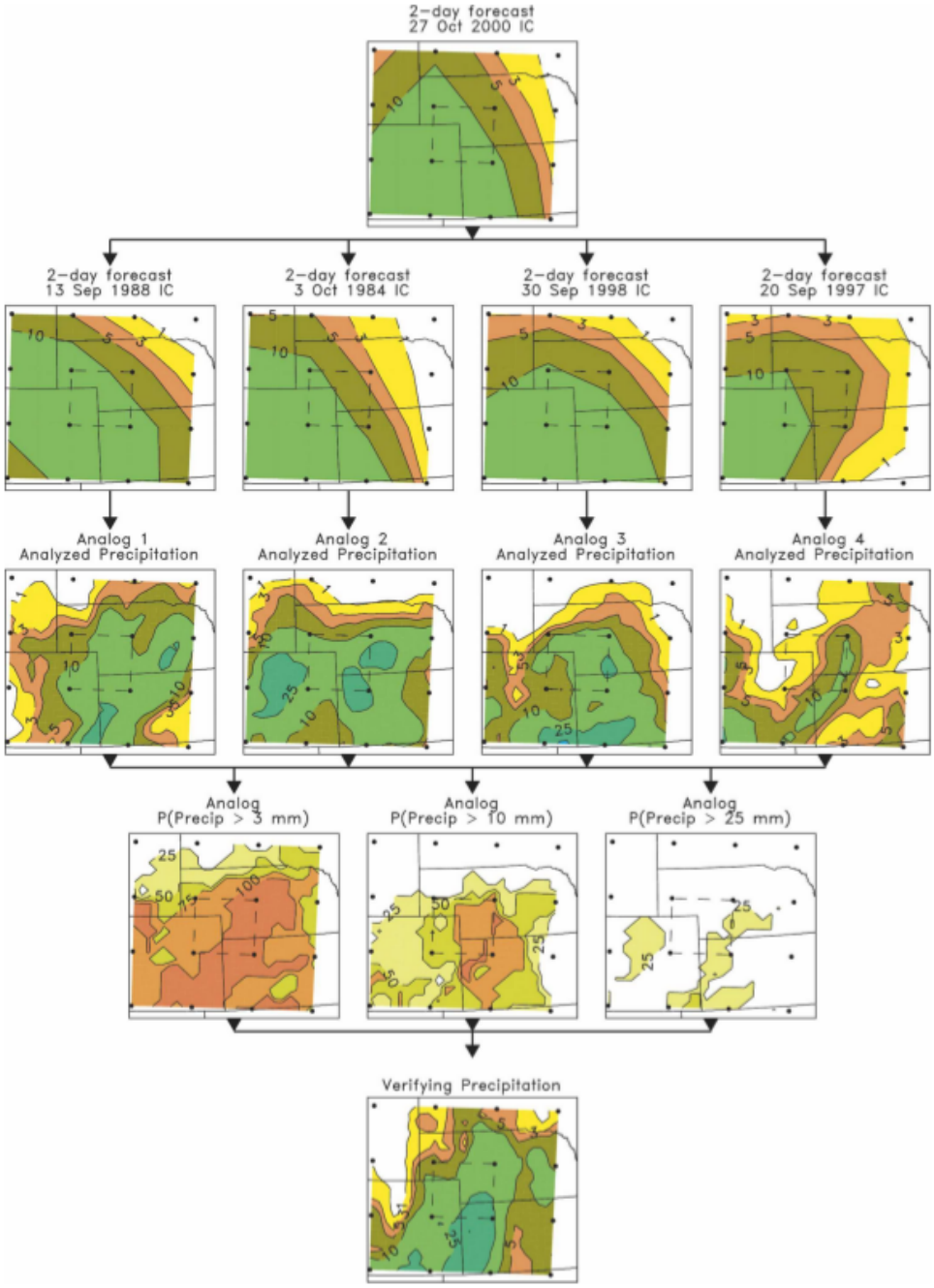


Figure 1.15: Illustration of the analog technique for precipitation forecast. From the top: first row represents the forecast to be calibrated (ensemble mean), the second row represents the "closest" archived forecasts, the third row the corresponding observations, the fourth row the predicted probability of threshold exceed, and the fifth row the verifying observation. From Hamill and Whitaker (2006).

procedure is that different regression coefficients have to be estimated for each given probability. Wilks (2009a) proposed an extension of the logistic regression method (called extended logistic regression, thereafter XLR) by including (a transformation of) the precipitation thresholds as an additional predictor variable. Figure 1.14 gives an example of the predictive cumulative distribution functions obtained from different predictor values using the XLR procedure. Roulin and Vannitsem (2011) and Hamill (2012) showed examples of applications of XLR to precipitation. Messner et al. (2014, 2013) proposed heteroschedastic formulations of the logistic regression for which the ensemble spread is directly used as a predictor for the dispersion of the predictive distribution. Ben Bouallègue (2012) defined a slightly modified version of XLR, introducing the interaction terms in the linear function of predictors.

Methods presented so far are generally carried out individually to each grid-point of the forecast. A calibration approach that attempts to calibrate the spatial precipitation fields is the analog method (Hamill and Whitaker, 2006). This technique is based on the similarity between the forecast and past historical forecast. The first step consists in taking the ensemble mean or another predictor and find the most similar past forecasts in the training dataset. The corresponding observations are then used to make a bias adjustment and to build well-calibrated probabilistic predictions. This process is illustrated in Fig. 1.15. Voisin et al. (2010) proposed an alternative, defined as the analog rank method for which members are matched individually on the basis of the ranks of forecast values in the forecast climatology. This approach presents some similarities with the QM method.

Some studies focused on the intercomparison between some of the aforementioned post-processing techniques for the precipitation variable. Schmeits and Kok (2010) compared XLR method and "modified" BMA version using a 20-yr ECMWF ensemble reforecast dataset over the Netherlands. The modified BMA version used an additive bias correction instead of an individual bias correction for each member and used a logistic regression technique to estimate the probability of precipitation for each dressing kernel. It was found that, for 24-hour area-mean and 24-hour area-maximum precipitation, both methods have the same performance. Ruiz and Saulo (2012) tested several calibration methods, including BMA and XLR techniques for

a 2-year period over South America. They showed that using the spread of the raw ensemble as an additional predictor of the ensemble mean does not improve the skill of the calibrated forecasts. It might be due to the strong correlation between the mean precipitation and the related prediction errors, so that the second predictor does not introduce additional information. Incidentally, BMA and XLR significantly improve the raw forecasts, in particular for large precipitation thresholds. No additional value was found using the XLR with interaction terms. Scheuerer (2014) implemented a NGR which makes use of a zero-censored GEV distribution model, using a 1-year forecast dataset of 6-hour accumulated precipitation over Germany. The method was compared to XLR and BMA (Sloughter et al., 2007) techniques. XLR yields a skill score similar to that of NGR method, while BMA improvement is less important. More specifically, the authors showed that the improvement of NGR over XLR was not significant.

1.4 Main questions addressed and objectives

Several studies have investigated the predictability of intense rainfall events (Collier, 2007; Walser et al., 2004; Walser and Schär, 2004). They show that the chaotic aspects of the moist dynamics of the small-scale phenomena involved in these events limit their predictability. The use of ensemble forecasting to assess the skill of forecasts and their associated uncertainty thereof represents an alternative approach. The underline theme of this study is the investigation of predictability of intense precipitation events over the Southeastern France by means of probabilistic tools.

One objective of this study is to investigate the predictive skill of probabilistic quantitative precipitation forecast for hydrostatic model, especially emphasizing large precipitation amounts. This task requires the availability of a large dataset that includes a significant number of HPEs. A 30-year ensemble reforecast is available with a 10-member global hydrostatic model inspired from the Météo-France operational ensemble forecasting system PEARP (Prévision d'Ensemble ARPEGE). Therefore, the main idea of the study is to use the reforecast dataset to explore the predictability of high precipitation events and to evaluate the link between the

forecast skill and the parametrizations used in the model. A gridded precipitation dataset has also been constructed from Météo-France rain-gauges network in order to dispose of an observation reference, at the same resolution of the model and over the same period.

The second main hypothesis of this thesis is to investigate the behaviour of standard post-processing methods for intense precipitation. In section 1.3.3, some univariate post-processing methods for precipitation have been presented. A large part of the presented studies have focused on the improvement provided by the calibration for low to moderate thresholds of rainfall amount, often lower than 20 mm. In the current study, we take advantage of a large reforecast dataset to test to which very high thresholds a post-processing method can result in a forecast improvement. The debiasing technique of Quantile Mapping is first tested, then XLR method is used. Tests with both methods are performed considering the reforecast as a model. Following Roulin and Vannitsem (2011), the XLR method is performed onto the operational system PEARP to assess whether the reforecast dataset can be useful to calibration for an operational scope.

The third issue raised in this study is to analyse the impact of physical parametrizations on the precipitation forecast. Multiphysics is designed to represent different physical properties of the model, with the objective of sampling its associated forecast uncertainty (see section 1.2.1). We aim to quantify the differences between the forecasts performed with different physical parametrizations, using a feature-based approach (see 1.3.2). A systematic analysis of the rainfall objects produced by each physical package is performed, and forecast precipitation is verified across the 30-year period using SAL quality-measure. The verification is carried out on the intense precipitation events of the dataset and targeted sub-regions, defined trough a clustering method applied to the observation reference dataset.

Chapter 2

Model set-up and observations

Contents

The reforecast dataset used in this study has been built from the set-up of the short-range ensemble prediction system PEARP (Descamps et al., 2015). The EPS version corresponds to its 4^{th} version, PEARP4, which operationally produced ensemble forecasts during the 2015-2017 period. A description of the set-up of the PEARP4 system is presented in this chapter, followed by a description of the procedure for the production of the reforecast dataset.

The last section of this chapter is devoted to the description of the method implemented for building the 24-hour precipitation reference dataset. Rainfall data collected from a large rain-gauge network are processed into a gridded rainfall field with the same resolution of the reforecast and covering the same period of time.

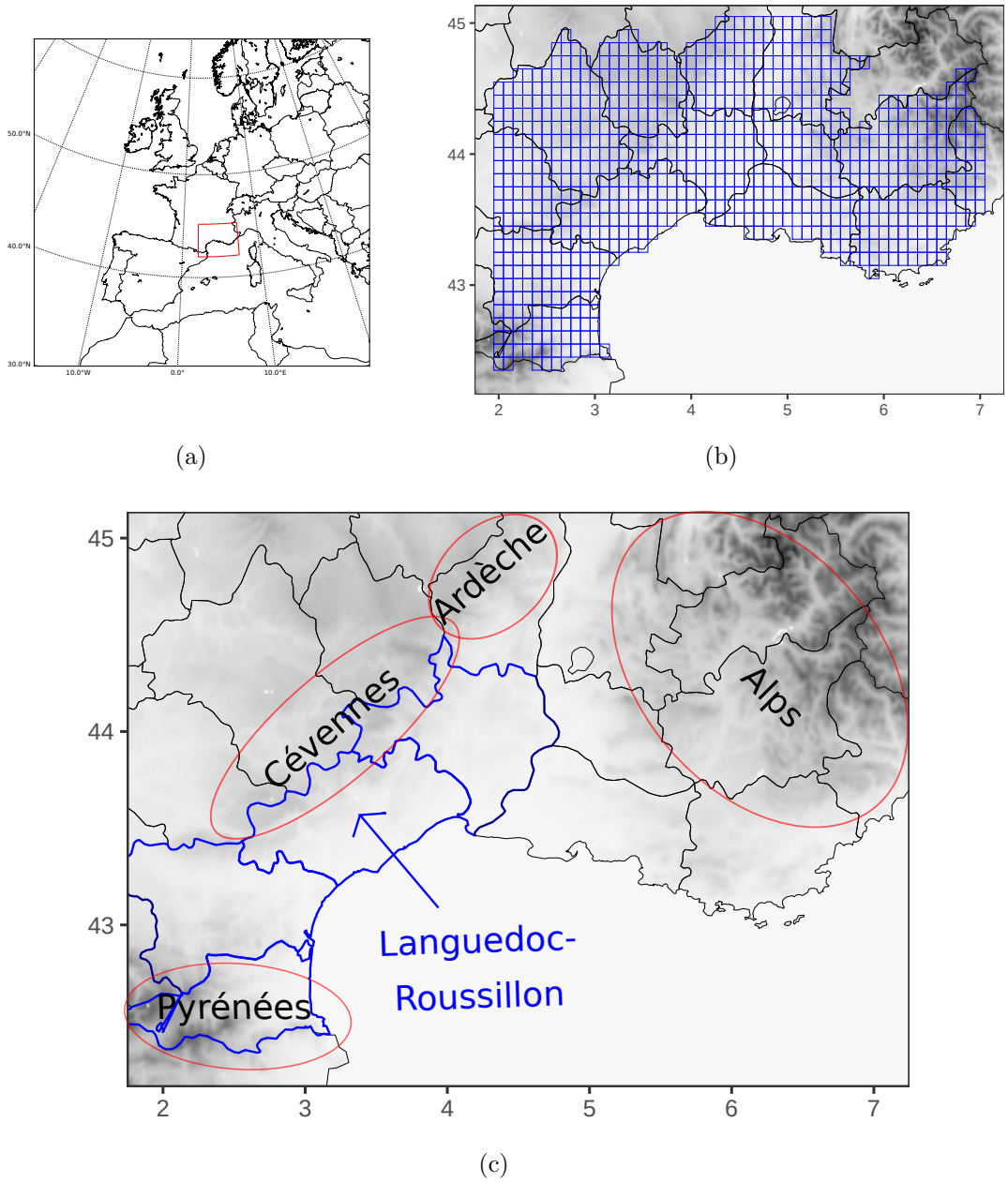
2.1 Domain of interest

All the analysis performed in this study are carried out over the French Mediterranean region (Fig. 2.1(a)). This area encompasses the Southeastern France. It is bounded by the Pyrénées, Massif Central, and Alps mountain chains and by the Mediterranean Sea. This particular graphical configuration, makes this region prone to HPEs, as already described in section 1.1.2.

Some geographic features that will be used in the discussion are labelled in Fig. 2.1(c). The red ellipses refer to the mountain chains, while the Languedoc-Roussillon region (blue borders) is indicated with regard to further commented results.

2.2 PEARP ensemble prediction system

The set-up summary given hereafter refers to PEARP4, which operated during a 2-year period between 2015 and 2017. PEARP is a short-range global ensemble prediction system, producing operational forecasts up to 4.5 days. It is based on the global deterministic model ARPEGE (Action de Recherche Petite Echelle Grande Echelle; Courtier et al., 1991) with a horizontal spectral truncation of T798 and a mapping factor of 2.4, with a stretching pole centered over France. The horizontal ensemble model resolution is variable and reaches a maximum of 10 km over France.



(a) (b)

(c)

Figure 2.1: a) Situation map showing the investigated area with respect to Western Europe and the Mediterranean Sea. b) Domain of concern of the study. The model grid is represented in blue. c) Location of major geographic features.

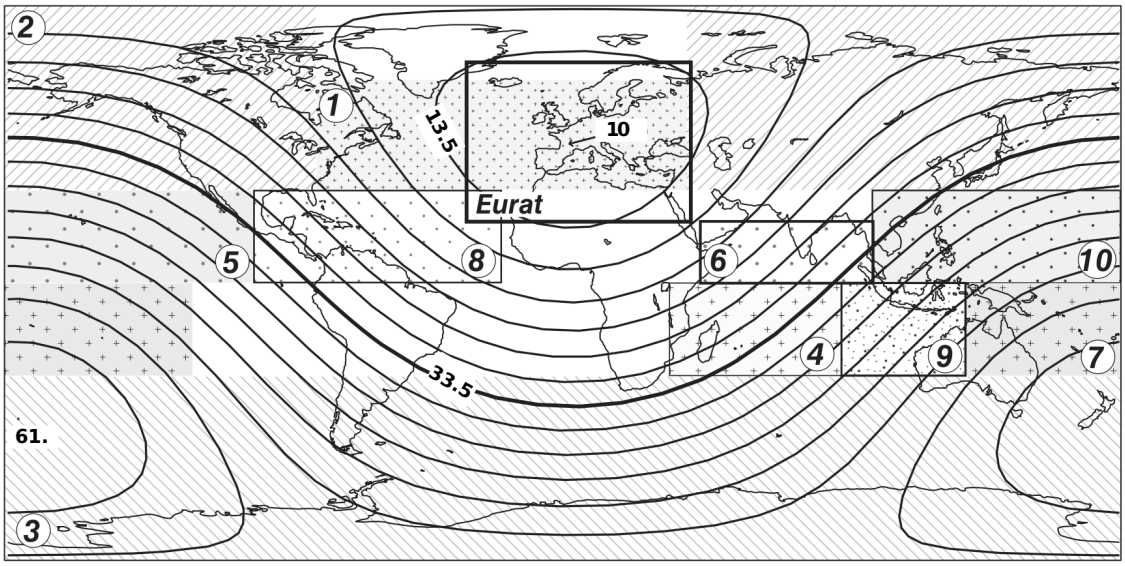


Figure 2.2: Location of the targeting areas used for the singular vectors. Contour intervals show the horizontal resolution in km of the PEARP4 system. Adapted from Descamps et al. (2015).

Even though the model is hydrostatic, this horizontal resolution could be related to the scale of meso-scale systems. There are 90 vertical levels from the ground to 50 km height. The ensemble size is set to 35 members, including a control "unperturbed" member, corresponding to a lower resolution version of the deterministic operational ARPEGE forecast (T1198). The remaining 34 perturbed members are centered around the control one at the initial step.

2.2.1 Initial condition perturbation

The initial condition perturbations of PEARP4 are built from a combination of Météo-France ensemble data assimilation system members (AEARP; Berre et al., 2007) and singular vectors. The ensemble data assimilation system is based on the 4D-Var approach and model error is accounted for through the introduction of the inflation method (Raynaud et al., 2012). 25 perturbed members are produced with a T399 spectral truncation. 17 members among the 25 perturbed members are randomly sampled and selected in the ensemble prediction system framework. They are interpolated to the PEARP resolution and then linearly combined with singular vectors by adding alternatively a positive and negative contribution with the same

a	Optimization time (h)	Norm	$n(a)$	Activation period
1	18	TE	16	Always
2	24	TE	10	Always
3	24	TE	10	Always
4	18	KE	7	1 Nov–31 May
5	18	KE	7	1 June–31 Oct
6	18	KE	7	16 Apr–15 Dec
7	18	KE	7	16 Dec–15 Apr
8	18	KE	7	1 June–30 Nov
9	18	KE	7	1 Dec–31 May
10	18	KE	7	Always

Table 2.1: Targeting areas used for singular vector computation. Locations of areas are shown in Fig. 2.2. $n(a)$ refers to the number of SVs computed on the area a.

magnitude.

Singular vectors (SVs) are computed over 10 different areas, shown in Fig. 2.2. They are computed using Total Energy Norm (TE) or Kinetic Energy Norm (KE), depending on the target area (see Table 2.1). TE is usually employed for extratropical SVs (Palmer et al., 1998), while KE is implemented in tropical SVs during the hurricane season (Barkmeijer et al., 2001). The optimization time is also dependent on the area.

We as note $\mathbf{e}_0(t_0)$ the deterministic analysis interpolated to the ensemble forecast resolution. The ensemble perturbations $\mathbf{e}_{l\pm}(t_0)$ are computed as follows

$$\begin{aligned} \mathbf{e}_{l+}(t_0) &= \mathbf{e}_0(t_0) + \mathbf{Q}\boldsymbol{\delta}\mathbf{e}_l(t_0) \\ \mathbf{e}_{l-}(t_0) &= \mathbf{e}_0(t_0) - \mathbf{Q}\boldsymbol{\delta}\mathbf{e}_l(t_0). \end{aligned} \tag{2.1}$$

The index l ranges between 1 and 17. State perturbations at time t_0 are denoted by $\boldsymbol{\delta}\mathbf{e}_l$. The perturbations are computed in a reduced phase space. The operator $\mathbf{Q}$ projects the perturbations from the reduced phase space into a full model phase space by adding the extra state variables, that are typically microphysics and turbulence variables, related to the physical parametrizations. The projection sets the corresponding perturbation fields to zero. The 34 analysis fields are obtained from the combination of the deterministic analysis state with a positive and a negative

contributions from the 17 state perturbations. The control run is added as a 35^{th} member. A given state perturbation results in a linear combination of one of the 17 assimilation ensemble perturbations sampled from AEARP and the SVs. The weighting coefficients of the SV are drawn from a Gaussian distribution.

2.2.2 Model error

In PEARP, the model error is represented by a multiphysics approach (the reader is referred to 1.2.1 to have some insight about this approach). Table 2.2 lists all the components of the ten different physical packages. All these packages are repeatedly assigned three times to randomly selected members of the ensemble. Five random physical parametrization packages are finally assigned to the four unassigned remaining members. Member 0 refers to the control member which uses the ARPEGE deterministic physical package.

Two turbulent diffusion schemes are considered: the Turbulent Kinetic Energy scheme (TKE; Bazile et al., 2012; Cuxart et al., 2000) and the Louis scheme (L79; Louis, 1979). TKE_{mod} is a slightly modified version of TKE, in which horizontal advection is ignored. For shallow convection different schemes are used: a mass flux scheme introduced by Kain and Fritsch (1993) and modified by Bechtold et al. (2001), thereafter the KFB approach, the Prognostic Condensates Microphysics and Transport scheme (PCMT; Piriou et al., 2007)), the Eddy-Diffusivity/Kain-Fritsch scheme (EDKF; Kain and Fritsch, 1993) and the PMMC scheme (Pergaud et al., 2009).

The deep convection component is parametrized by either the PCMT scheme or the Bougeault (1985) scheme (thereafter B85). The closing constraint of the equation systems used in these two schemes is based on a relationship between the bulk mass flux and the in-cloud vertical velocity with respect to the area covered by convection γ. Two closures relations are considered, the first one (C1) based on the convergence of humidity and the second one (C2) based on the CAPE (Convective Available Potential Energy). B85 scheme originally uses the C1 closure, while PCMT uses alternatively the closure (C1 or C2) which maximizes the γ parameter. In practice PCMT scheme uses most of the time the CAPE closure. The closure based on

	Turbulence	Shallow convection	Deep convection	Oceanic flux
00	TKE	KFB	B85	ECUME
01	TKE	KFB	B85	$\text{ECUME}_{\text{mod}}$
02	L79	KFB	B85_{mod}	ECUME
03	L79	KFB	CAPE	ECUME
04	TKE_{mod}	KFB	B85	ECUME
05	TKE	EDKF	B85	ECUME
06	TKE	PMMC	PCMT	ECUME
07	TKE	KFB	PCMT	ECUME
08	TKE	PCMT	PCMT	ECUME
09	TKE	KFB	B85	ECUME

Table 2.2: Physical parametrizations used in the ensemble reforecast.

the convergence of humidity is related to the moisture-based closure proposed by Kuo (Kuo, 1974), whereby the whole available humidity is directly distributed as precipitation or humidification of the environment. PCMT also differs from B85 as it includes a prognostic equation for the vertical velocity that allows to take into account the overshooting (air parcels in the updraft can rise beyond the level of neutral buoyancy.) In physics package 2, deep convection parametrization uses a modified version of the B85 scheme in which deep convection is triggered only if cloud top exceeds 3000 m (B85_{mod} in Table 2.2). The same trigger is used in physics package 3 in which deep convection is parametrized using the B85 scheme along with a CAPE closure (CAPE in Table 2.2).

The oceanic flux is solved by means of the ECUME scheme (Belamari, 2005). In $\text{ECUME}_{\text{mod}}$ oceanic fluxes are maximized. Control member and member 9 are characterized by the same parametrization set-up, but member 9 differs for the modelization of orographic waves.

2.2.3 PEARP forecast dataset

In this study, a 4-month forecast dataset generated during 2016 by PEARP4 is used. Forecasts are computed every day at 1800 UTC, over a period from the 1 September 2016 to the 31 December 2016. The number of days during this period is 122.

24-hour accumulated precipitation is extracted from the forecast over the domain

	Spectral Truncation	No. of levels	No. of members	EDA	SVs	Model Error	$t_0(t_{fcast})$ (h))
PEARP-2016	T798 (2.4)	90	35	AEARP	yes	multiphysics	1800 UTC(108) every day
Reforecast	T798 (2.4)	90	10	no	no	multiphysics	1800 UTC(108) every 4-day

Table 2.3: Summary of the main characteristics associated with the production of PEARP-2016 and the ensemble reforecast.

of interest on a $0.1° \times 0.1°$ grid. The model grid overlapped to the domain of interest is shown in Fig. 2.1(b). All values below 0.1 mm are set to zero because rain-gauges minimum collected value corresponds to that value. This forecast dataset is thereafter *PEARP-2016*.

2.3 Ensemble reforecast dataset

As hinted in section 1.3.3, a reforecast dataset can represent an opportunity to assess a great part of the quality range of an operational EPS and potentially test some post-processing methods.

In the current study, a reforecast based on PEARP4 is exploited. Reforecast horizontal and vertical resolutions are the same as PEARP4. Perturbation of the initial conditions has not been taken into account, because this task would be too computational and technically demanding at that time. The reforecast implements the same physical packages (see Table 2.2) as PEARP-2016. Lead time range is also the same, up to 108-hour. PEARP-2016 is available every day at 1800 UTC, while the reforecast has been generated every 4 days. The reforecast covers 4-month periods (from September to December), as PEARP-2016, but the period extends over 30 years, from 1981 to 2010. The dataset consists of $\approx$ 900 days of forecasts per lead time. Table 2.3 summarizes some experimental similarities and discrepancies between the reforecast and PEARP-2016.

The initialization strategy adopted for the production of the reforecast deserves to be looked more in detail. This operation follows the hybrid approach described in Boisserie et al. (2016). The atmospheric state variables at the initial step are extracted from ERA-Interim reanalysis (Dee et al., 2011), available at the ECMWF.

Second, the land-surface initialization parameters are interpolated from an offline simulation of the land-surface SURFEX model (Masson et al., 2013) driven by the 3-hourly atmospheric fields from ERA-Interim. As verified by (Boisserie et al., 2016), this procedure produces a reforecast having more skill compared to the one which is initialized with ERA-Interim land-surface fields, specially in terms of 2-m temperature and humidity budgets.

As for the production of PEARP-2016, 24-hour accumulated precipitation forecasts are extracted from the reforecast on a $0.1° \times 0.1°$ grid over the domain of interest. All values below 0.1 mm are set to zero.

2.4 Rainfall observation reference dataset

24-hour accumulated precipitation is derived from the in-situ Météo-France rain-gauges network, covering the same period as the reforecast dataset. Rainfall is collected from fourteen French departments over the domain shown in Fig. 2.1(b). The observation network passes a validation test which excludes bad quality measures.

The French Mediterranean region has been quite intensively observed, with about 700 rain gauge stations over the 1960-2010 period (Fig. 2.3). A moderate number of stations have been closed or opened during this period, resulting in a variable number of observations over time. Then, in order to maximize the quality of the rainfall analysis over the region, all rain-gauges available each day are included.

Figure 2.4 shows the whole rain-gauge network opened and closed during the reforecast period (1980-2010). We can observe that the mountain chains are particularly densely covered. Observational uncertainty can be certainly found among rain-gauges data. However, since it is extremely difficult to have a proper estimate of the uncertainty (depending on the location, instrument and accumulated precipitation), this information has been not taken into account for the construction of the observational reference.

In this study, in order to properly estimate the model at its resolution, we aim at providing the observation on the same grid. We intend to spatialize all available rainfall information as a gridded field. The spatialized dataset has to be constructed

Figure 2.3: Number of available Météo-France rain-gauges per year from 1960 to 2010.

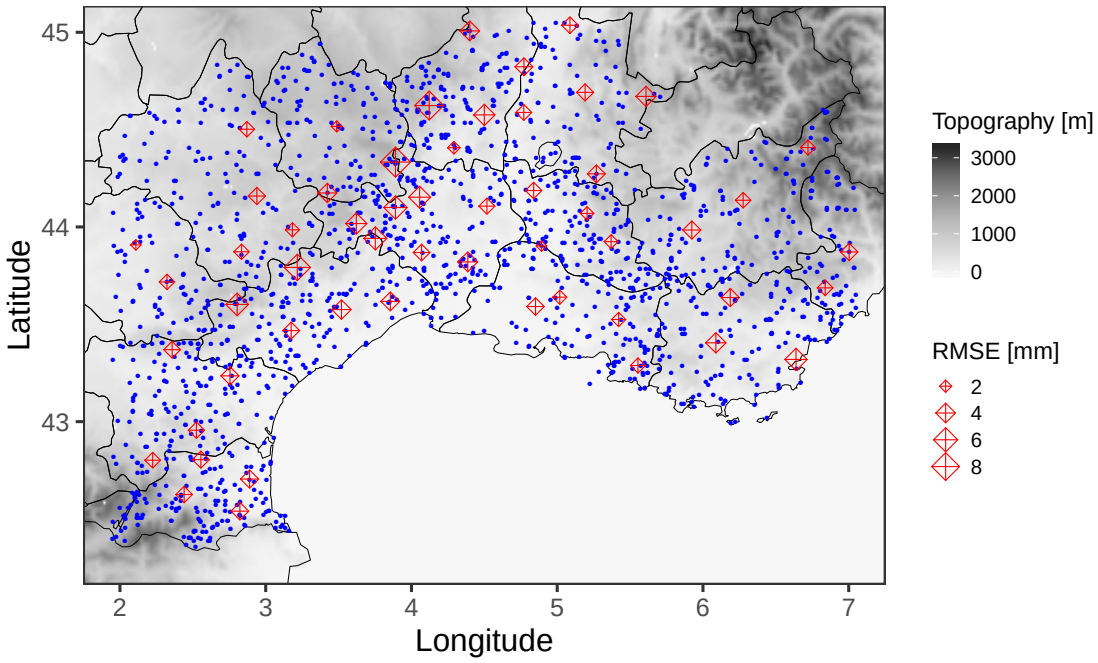


Figure 2.4: Rain-gauges network used for the study. Red diamonds represent the rain-gauges selected for cross-validation testing, the red diamond size is proportional to the RMSE computed over the whole period for each observation test using the daily best configuration. Blue dots represent the rain-gauges selected for cross-validation training.

by means of spatial interpolation techniques. For the interpolation, we use Ordinary Kriging (OK; Goovaerts and Goovaerts, 1997; Goudenhoofdt and Delobbe, 2009; Ly et al., 2011, 2013) and Inverse Distance Weighting methods (IDW; Chen and Liu, 2012; Shepard, 1968; Teegavarapu and Chandramouli, 2005). Kriging methods can be considered as a modeling method for which additional predictors can be included. The elevation would be a suitable predictor used to tackle the interpolation. The correlation between elevation and daily precipitation accumulation resulted in low values. Then, only ordinary kriging method was taken into account, using rainfall as unique predictor.

2.4.1 Ordinary Kriging

Ordinary kriging (OK) is a geostatistical interpolation method based on the use of a semivariogram, which describes the variability of the spatial fields. Semivariance is a measure of the degree of spatial dependence between values of the variable Z at two different locations separated by the distance $\mathbf{h}$:

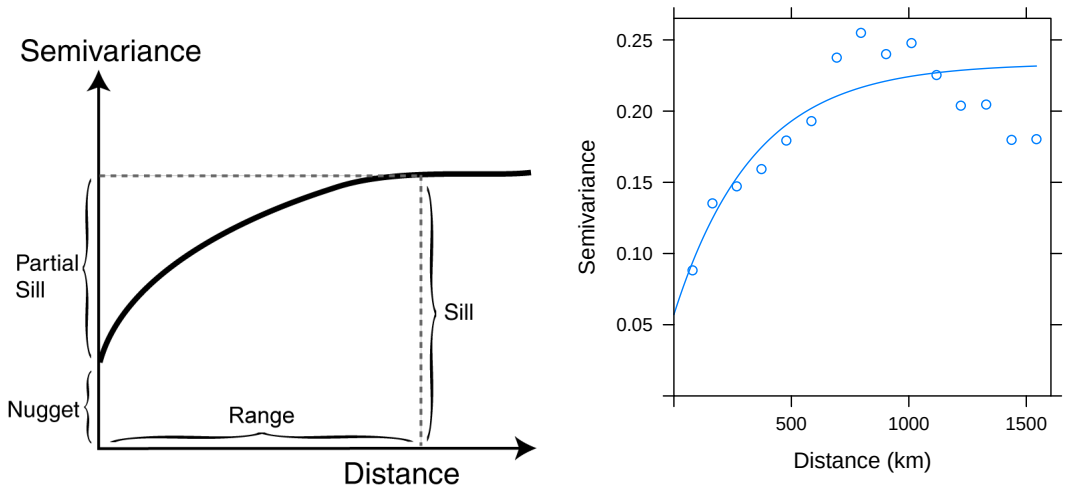
$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} \left[Z(\mathbf{u}_\alpha + \mathbf{h}) - Z(\mathbf{u}_\alpha) \right], \tag{2.2}$$

where $N(\mathbf{h})$ denotes the number of point pairs distanced by the lag $\mathbf{h}$. The vector h here reduces to a scalar, since we consider spatial isotropy over the spatial domain. A graphical example is given in Fig. 2.5. The empirical semivariogram is computed for different values of h. Then, these points, corresponding to the semivariance at fixed distances h, are fitted using a given theoretical model to obtain a continuous semivariogram function (see Fig. 2.5(b)) that is used for the interpolation procedure.

The estimation of Z at a point $\mathbf{u}$ is given by the linear equation:

$$Z(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}) Z(\mathbf{u}_\alpha) \tag{2.3}$$

where $\mathbf{u}_\alpha$ are the neighbouring data points used in the interpolation and $n(\mathbf{u})$ is their corresponding total number. The weights λ_α are estimated by solving a linear equation system which requires the spatial covariance values from the semivariogram



(a) (b)

Figure 2.5: (a) Graphical illustration of a semivariogram. The sill denotes the semivariance value at which the variogram levels off. The range is the lag distance h at which the semivariogram reaches the sill value. The autocorrelation is supposed to be zero beyond this range. The nugget represents the variability at distances smaller than the typical sample spacing. (b) Variogram of estimation points with an exponential function.

function.

The estimation of the weights used for the interpolation depends on the modelization of the semivariogram. In this study semivariograms are modeled using three different theoretical models: Exponential, Spherical, and Gaussian, whose expression are given below

– The Exponential semivariogram model

$$\gamma(h) = \begin{cases} 0, & \text{for } h = 0, \\ c_0 + c_1 \left[1 - \exp\left(-\frac{h}{a_0}\right)\right], & \text{for } h \neq 0. \end{cases} \tag{2.4}$$

– The Spherical semivariogram model

$$\gamma(h) = \begin{cases} 0, & \text{for } h = 0, \\ c_0 + c_1 \left[\frac{3}{2}\frac{h}{a_0} - \frac{1}{2}\left(\frac{h}{a_0}\right)^3\right], & \text{for } 0 < h \leq a_0, \\ c_0 + c_1, & \text{for } h > a_0. \end{cases} \tag{2.5}$$

– The Gaussian semivariogram model

$$\gamma(h) = \begin{cases} 0, & \text{for } h = 0, \\ c_0 + c_1 \left[1 - \exp\left(-\frac{h^2}{{a_0}^2}\right)\right], & \text{for } h \neq 0, \end{cases} \tag{2.6}$$

where c_0 is the nugget, c_1 the sill and a_0 the range parameters. For each day, a sample semivariogram is computed from the full rain-gauge network and it is fitted to the three theoretical semivariogram models. The fit is minimized by a weighted least squares method, in order to estimate the parameters c_0, c_1, and a_0 of the function $\gamma(h)$.

Since precipitation is known to follow a non-Gaussian distribution, it is a common practice for kriging operations to first transform the target variable, to ensure that the transformed distribution has statistical properties closer to a Gaussian law (Erdin et al., 2012). In this study, we test ordinary kriging both with the raw precipitation variable and with the square root transformation.

2.4.2 Inverse Distance Weighting

In a few number of cases, kriging can lead to unrealistic solutions. Then, we also apply a parallel interpolation method known as Inverse Distance Weighting (IDW). The interpolation weights (see eq. 2.7) depends on the inverse distance between the estimation point and the neighbouring data points. In contrast to kriging methods, IDW does not take into account the spatial covariance of the variable over the domain. This method assumes that the highest weight is attributed to the closest points. The weight is defined as follow

$$\lambda_\alpha(\mathbf{u}) = \frac{\frac{1}{(r_\alpha)^d}}{\sum_{\alpha=1}^{n(\mathbf{u})} \frac{1}{(r_\alpha)^d}}, \quad d > 0, \tag{2.7}$$

where r_α is the distance between the estimation point and one neighboring data point, corresponding to $|\mathbf{u} - \mathbf{u}_\alpha|$ in the projected space. The parameter d control the distance-decay effect. A large d value gives larger weights to closer neighbouring points. Several studies propose $d = 2$ (Goovaerts, 2000; Ly et al., 2011), or $d = 3$

(Burrough et al., 2015; Lu and Wong, 2008). The sensitivity to varying number of neighbouring points $n(\mathbf{u})$ can also be addressed.

Next paragraph presents the implementation of the interpolation algorithm, based on both Ordinary Kriging and Inverse Distance Weighting methods.

2.4.3 Interpolation algorithm implementation

In order to evaluate the best interpolation method for each day of the considered period cross-validation is performed. The procedure consists in retrieving a selected number of observations from the sample before running the interpolation method on the remaining observation points. We selected 55 observation points among the full dataset. These rain-gauges are selected in order to have a regular coverage over the domain (see Fig. 2.4), especially over the mountainous area. The interpolator estimators are then computed at each location of the 55 chosen test points.

The algorithm selects the best interpolation method in terms of error at the test points. The flowchart illustrating the procedure is presented in Fig. 2.6. Ordinary Kriging is performed over the 55 test observation points using the three different semivariogram models estimated from raw precipitation data (thereafter named as configuration 1.R, 2.R, and 3.R), and from the root square transformed precipitations (thereafter named as configuration 1.T, 2.T, and 3.T). The total number of Ordinary Kriging configurations is 6. At the same time, IDW is performed using four different configurations with various d values (2,3) and numbers of neighbouring points $n(\mathbf{u})$ (5,10, or all the points of the domain).

For a given day j from the 30-year dataset, the different outputs generated using the seven configurations are compared according to the Root Mean Squared Error (RMSE) test statistics applied to the 55 test points

$$RMSE[j] = \sqrt{\frac{1}{N^*}\sum_{i=1}^{N^*}\left[\hat{Z}(\mathbf{u}^*_{\mathbf{i,j}}) - Z(\mathbf{u}^*_{\mathbf{i,j}})\right]} \tag{2.8}$$

where $N^* = 55$, $\hat{Z}(\mathbf{u}^*_{\mathbf{i,j}})$ is the predicted value at the i^{th} test locations $\mathbf{u}^*$, and $Z(\mathbf{u}^*_{\mathbf{i,j}})$ is the corresponding observed value. Then for each day, the configuration which provides the lowest RMSE value is retained. Afterwards, the interpolation

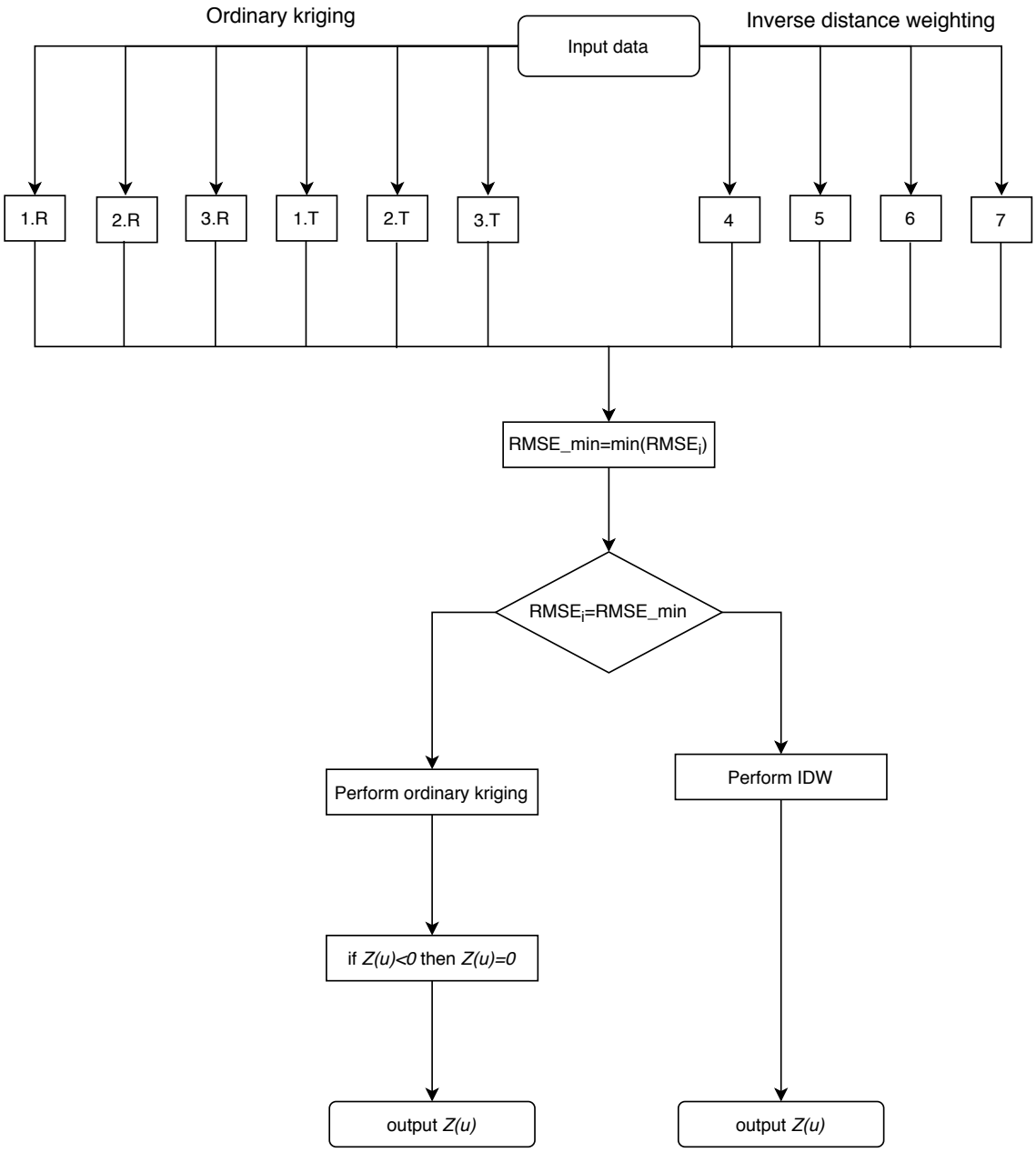


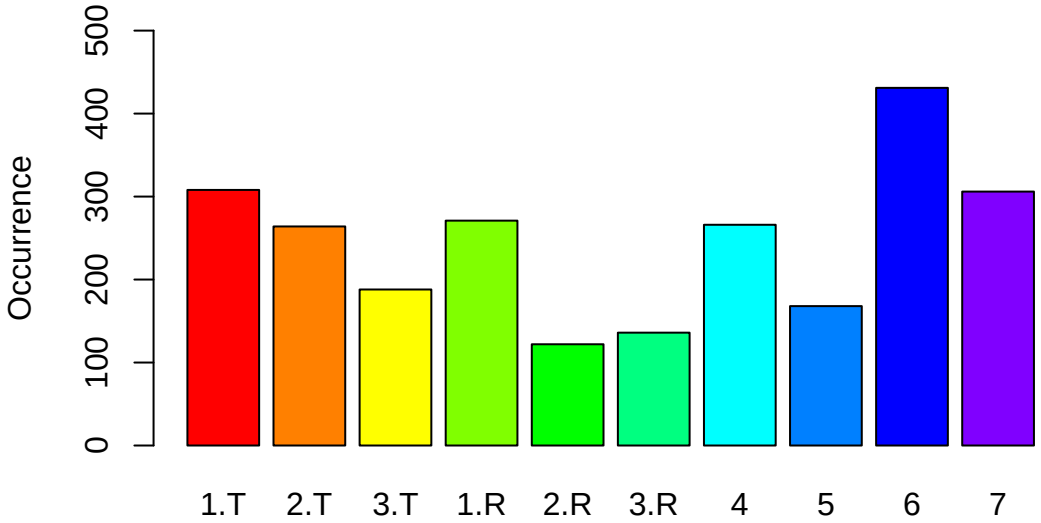
Figure 2.6: Flowchart of the interpolation algorithm developed for the production of the reference dataset.

is applied over the entire domain using this configuration. Since kriging equations do not prevent negative values of the weights λ_α, the estimated rainfall value might reach negative values. In these cases, values are automatically set to zero. For some dates, no precipitation can be observed from the test rain-gauges, while at least one of the training rain-gauges is stricly positive. For these cases, cross-validation is not feasible and IDW is applied. The use of IDW is preferred to kriging for these specific dates, because kriging can be unstable in certain particular conditions, leading to aberrant values.

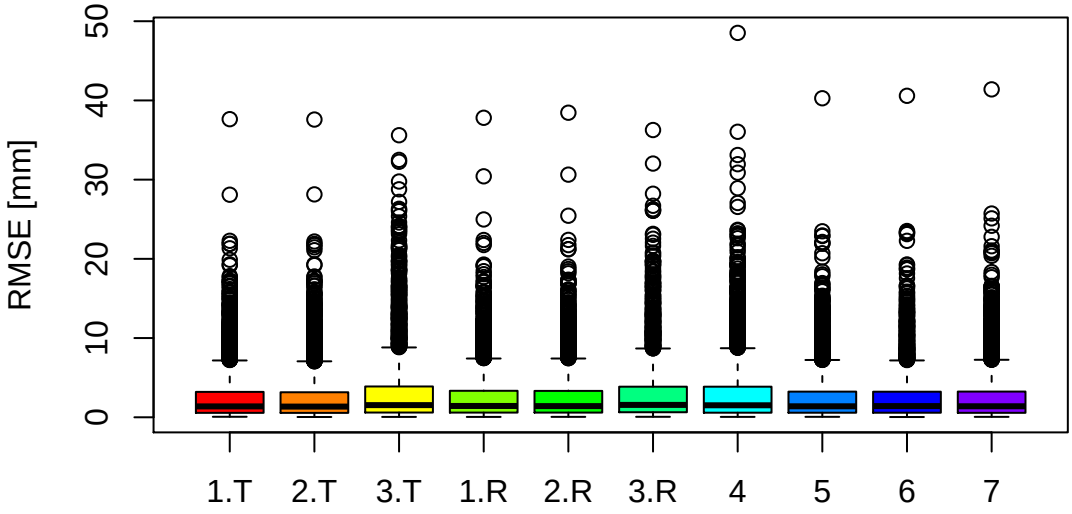
Figure 2.7 shows some statistics about which configuration is used for the daily interpolation. Fig. 2.7(a) shows the use histogram of each configuration across the 30-year dataset. The overall use of each interpolation method is 47% for OK and 53% for IDW. Exponential model fit (1.T and 1.R) is the most frequently used among the kriging configurations. The highest occurrence is associated with the configuration 6, implementing IDW, with $d = 2$ and $n = 5$. This shows that in our case IDW interpolation performs better using a reduced number of the closest neighbourhood points.

Figure 2.7(b) presents boxplots of the RMSE computed for the 55 test points over the whole period for each configuration. RMSE statistics between configurations is quite similar, but the largest errors are often associated with the Gaussian model fit semivariogram in conjunction with the square root transformation (configuration 3.R). For any configuration, upper quartile of the boxplot is always lower than 5 mm.

The quality of the estimated values obtained by means of the algorithm is assessed by computing the residuals on the test observation points. These set of 55 test points are used to individually compute the RMSE over the 30-year period, using for each day the best selected configuration. Figure 2.4 shows the obtained RMSE statistics for each test point, whose amplitude is proportional to the red diamond size. It is worth noting that the RMSE ranges between 2 mm and 8 mm. The largest RMSE values are found along the Cévennes and Ardèches mountain chains. Lower errors are present over the Pyrénees and the Alps. The magnitude of the RMSE tends to be higher in the areas prone to intense rainfall (reference to Fig. 1.7(b)).



(a)



(b)

Figure 2.7: a) Occurrence of each configuration in the choice for the final interpolation. b) Boxplot of daily RMSE (see eq. 2.8) for each configuration. Outliers are determined for a given value larger than $Q3 + 1.5 * IQR$, where $Q3$ is the upper quartile and IQR the interquartile range.

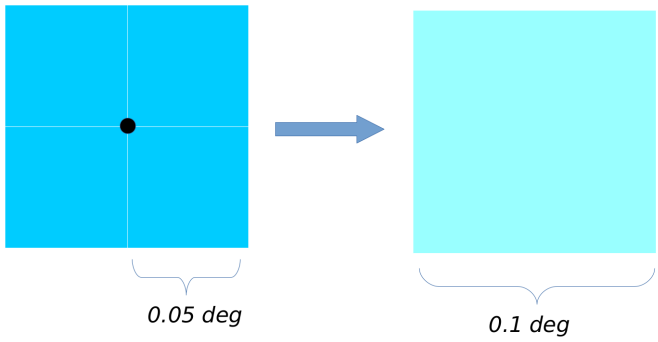


Figure 2.8: Symbolic illustration of the upscaling process.

2.4.4 Final set-up for the production of the gridded precipitation data

As described above, interpolation is performed for a specific day using the best interpolation configuration. We first applied the interpolation over a 0.05° resolution grid. Then the interpolated field is upscaled on the same grid as the model one (0.1°). This is done by performing a spatial average over the 4 grid cells surrounding the interpolated grid-point. A graphical representation of the process is given in Fig. 2.8. This up-scaling procedure is an attempt to reproduce the filtering effect produced by the parametrizations of the physical processes in the model that applies below the grid resolution. Finally, as performed on the forecasts, each value below 0.1 is set to zero. This final product defines the observation reference dataset.

One example of the interpolated rainfall field is given in Fig. 2.9. The left panel represents the interpolated rainfalls field at the 0.05° resolution, while the right panel shows the same field after the upscaling process. It can be noted that the precipitation maximum is partially smoothed after transformation.

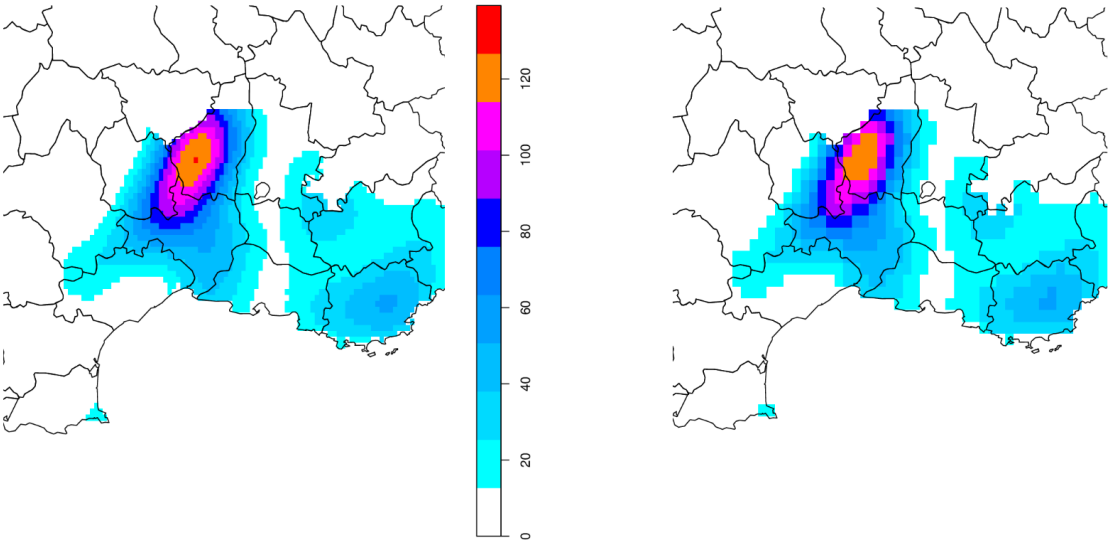


Figure 2.9: Left: Example of 0.05° resolution precipitation field (mm) estimated by interpolation for the 13 November 1986. Right: The same field after the upscaling process.

Chapter 3

Ensemble reforecast verification

Contents

The duration and experimental set-up of the reforecast may ensure to find substantial information for understanding the PEARP behaviour on HPEs. A cautious preliminary statistical analysis over the whole dataset is needed. For that purpose, deterministic or probabilistic scores are introduced and used, focusing on the verification of high precipitation forecast. With this long dataset, we have the benefit that scores for high precipitation thresholds ($\geq$ 100 mm) are still representative. As described in section 2.3, the reforecast is built using ten different physical parametrization packages. Therefore, the followed purpose is to analyse whether the 24-hour precipitation forecast skill depends on the parametrizations. This introductory exploration is followed by an evaluation of the reforecast as a probabilistic forecast system, in order to measure potential discrepancies with the PEARP system.

A verification procedure is based on the comparison between a probabilistic forecast and a reference. Different verification approaches can be used. Hamill and Whitaker (2006) use model interpolation on a finer grid that the original one, Wilks (2002) or Roulston and Smith (2003) choose to interpolate model data at observations locations, in Wilks and Hamill (2007) and Wilks (2009a) the closest model grid point is associated to the observation point, Scheuerer and Hamill (2015) propose a downscaling approach of model data by a calibration step, and Hamill et al. (2008), Hamill (2012), Zhu and Luo (2014) analyze model and observation data over a uniformed grid. The latter procedure relies on a remapping procedure, consisting in upscaling the observation data to the model grid resolution (e.i. Accadia et al. (2003)). In this study, the deterministic verification is based on the comparison between rain-gauges stations and the closest model grid point. On the other hand, the probabilistic verification, as well as the calibration methods (presented in the chapter 4), are based on the use of the gridded reference observation, whose interpolation procedure has been presented in section 2.4.

3.1 Deterministic forecast verification and comparison between physics schemes

The purpose of this sensitivity study is to depict the performance differences for 24-hour QPF (Quantitative Precipitation Forecast) related to the different physical parametrizations implemented in the reforecast, regardless of forecast range. In particular, we focus on the precipitation forecast skill at the extreme tail of the rainfall amount distribution.

For this exploratory verification, the reference dataset is not based on the interpolated rainfall reference, as described in section 2.4, but on 211 rain-gauges stations from the Météo-France network in southeastern France. Indeed, we wanted to comply with standard verification rules for this analysis. These observations ensure a temporal continuity at the same locations for the reference data during the 30-year period. All comparisons between model and observations are computed using the nearest-neighbour approach. Each observation point is associated to its nearest model grid point, as shown by the arrows in Fig. 3.1. We first describe the verification metrics used before analyzing the corresponding scores for the reforecast.

3.1.1 Forecast verification metrics

Forecast verification can be distinguished between nonprobabilistic and probabilistic verification. These two classes can be further subdivided in categorical verification and verification of continuous predictands. The categorical verification relates to a predictand belonging to one of a set of Mutually Exclusive Collectively Exhaustive (MECE) categories. Nonprobabilistic verification is applied to rainfall events corresponding to a threshold overrun.

Categorical verification

Conventionally, non probabilistic verification outcome is displayed in a $I \times J$ contingency table of observed/forecast frequencies. The most intuitive table is the

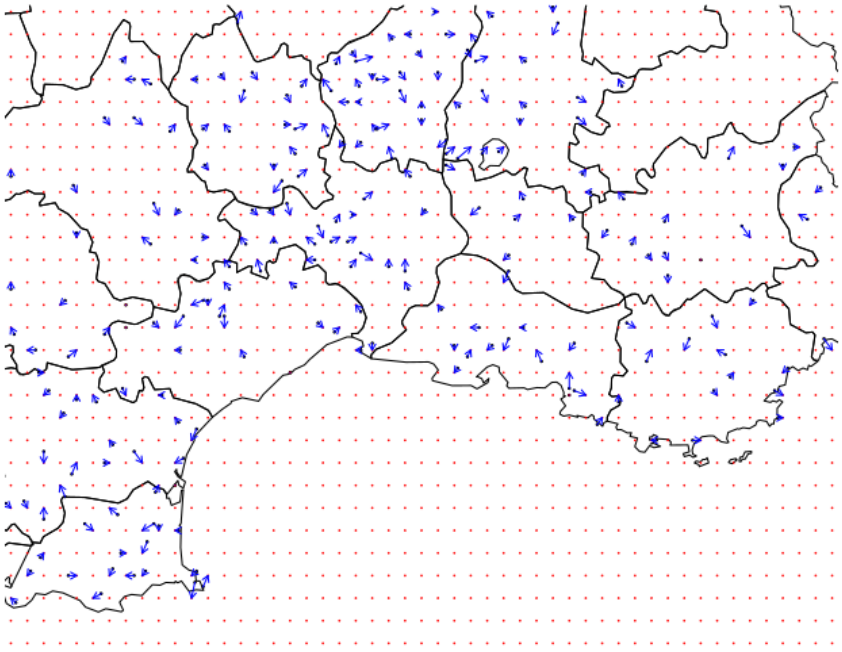


Figure 3.1: Model grid points (in red) and rain-gauges used for the verification. Blue arrows connect each rain-gauge observation to the nearest grid point.

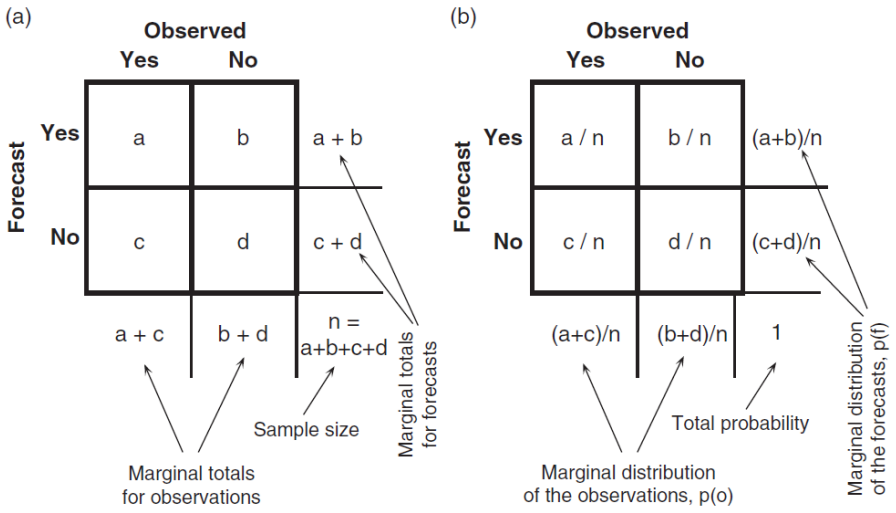


Figure 3.2: Contingency table (a) with a = *correct hit*, b = *false alarm*, c = *missed* and d = *correct negative*. The second table (b) include the corresponding joint distribution of the forecasts and of the observations $[p(f, o)]$ and the corresponding marginal distributions p(o) and p(f). From: Wilks (2009b).

2×2 contingency table (Fig. 3.2), providing the hit, false alarm, missed and correct negative categorical frequencies. The correct hit represents the cases when the event is both observed and forecasted. The false alarm corresponds to forecasted, but not observed events. The missed corresponds to the cases when the event occurs but is not forecasted. A correct negative represents an event that does not occurs and it is not forecasted. The sum of these categories is equal to the total number of forecast/observation pairs. The climatological frequency of the event is defined as the fraction of occurrences of the event from the observation, with regard to the total number of forecast/observation pairs.

Three scores (Doswell et al., 1990) can be computed from the contingency table:

- the Hit Rate ($H = \frac{a}{a+c}$) ranges between 0 (worst) and 1 (best). The score is sensitive to the hits, but it ignores the false alarms and is very sensitive to the climatological frequency of the event;
- the False Alarm Rate ($F = \frac{b}{b+d}$) is sensitive to the false alarms, but it ignores the misses. It ranges from 0 (best) to 1 (worst);
- the False Alarm Ratio ($FAR = \frac{b}{a+b}$) is sensitive to the false alarms, but it ignores the misses. It is very sensitive to the climatological frequency of the event. It ranges from 0 (best) to 1 (worst).

Low base-rate events (like HPEs) are difficult to verify because many traditional metrics tend to trivial, non-informative limits as the climatological frequency tends to zero. A specific metric that tends to compensate this effect is the SEDI (Symmetric Extremal Dependence Index) (Ferro and Stephenson, 2011; North et al., 2013), defined as:

$$SEDI = \frac{\log F - \log H - \log(1-F) + \log(1-H)}{\log F + \log H + \log(1-F) + \log(1-H)}. \tag{3.1}$$

SEDI is complement symmetric, i.e. it is invariant to a relabelling of the events as nonevents and the nonevents as the events (Stephenson, 2000). It has a fixed range of $[-1, 1]$ and is maximized when $H \to 1$ and $F \to 0$. It approaches its minimum value when $H \to 0$ and $F \to 1$. In addition, values above zero imply that

the forecast system is better than random, values below zero imply that the forecast system is worse.

Scalar Accuracy Measures

The two following scalar measures of forecast accuracy for continuous predictands are commonly used. The first one is the Mean Absolute Error:

$$MAE = \frac{1}{n}\sum_{k=1}^{n} \mid x_{f,k} - x_{o,k} \mid . \tag{3.2}$$

Here $(x_{f,k}, x_{o,k})$ is the k^{th} of n pairs of forecasts and observations. MAE is null if the forecast is perfect. We can interpret the MAE as the magnitude of the forecast error in a given verification dataset.

Another common scalar accuracy measure is the Root Mean Squared Error:

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(x_{f,k} - x_{o,k})^2}. \tag{3.3}$$

RMSE is an error metric that gives more weight to the largest errors.

3.1.2 Some deterministic verification scores on the reforecast

To study the impact of the physical parametrizations on HPEs rainfall forecasting we compare the scores of members differing only from one component of the physical parametrization schemes (for example, members with different vertical diffusion schemes, but with the same scheme for shallow convection, deep convection and oceanic fluxes). The reader is referred to Table 2.2 for the details about the physics set-up. The scores are computed at each observation locations. For graphical purposes, when needed, we interpolated the scores on a finer regular grid in order to illustrate the spatial distribution over the domain.

Rather than defining events using the same threshold at each point, we introduce a threshold that takes into account the spatial variability of the climatological distribution. A given quantile of the local climatological distribution is chosen at each

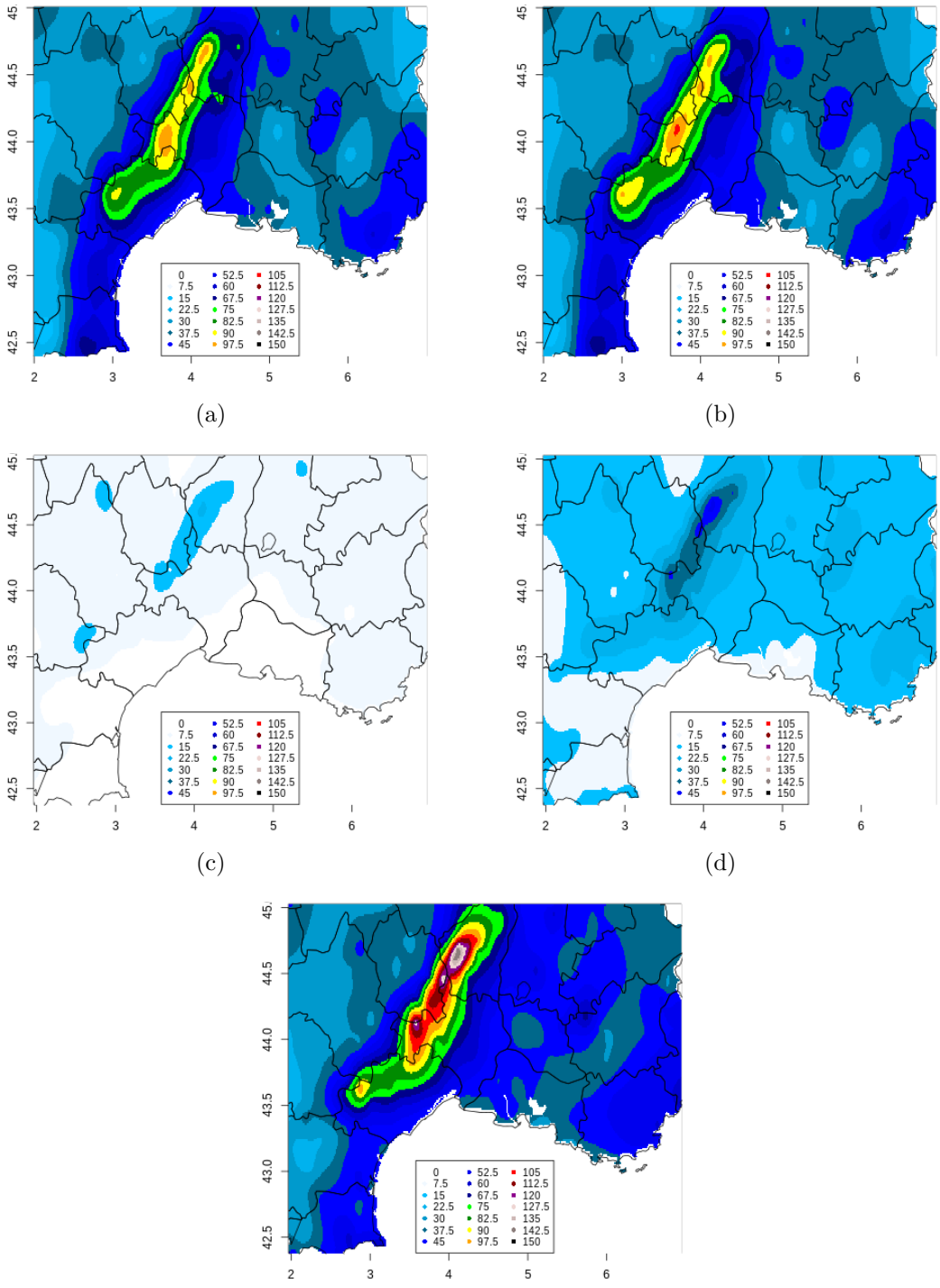


Figure 3.3: 24-hour rainfall amounts quantile q_{99} (mm) for member 0 (a) and 4 (b) of the ensemble reforecast. Member 0 implements the TKE turbulence scheme, member 4, the TKE_{mod}. Figures (c), (d) and (e) are drawn from the quantiles q_{90}, q_{95} and q_{99} of the observation climatology, respectively.

grid-point. This ensures that the climatological event distributions are the same everywhere. This also means that the considered thresholds are spatially dependent. Quantiles are then based on the whole 30-year period. Hereafter we note as $q_{n^{th}}$ the n^{th} quantile.

Figures 3.3(c), 3.3(d) and 3.3(e) shows some quantile values (q_{90}, q_{95}, q_{99}) of the 24-hour observed rainfall. For these three quantile thresholds, the highest values are located along the Cévennes mountains, with a maximum over the Ardèche mountains (see Fig. 2.1(c)). A peculiar observation comes out from the statistical distribution of rainfall over the southern part of Languedoc-Roussillon. q_{90} and q_{95} show relatively low values over this region, while q_{99} quantile is as high as over other mountainous areas. This suggests that only a few extreme events tend to affect this area, but they reach the same extreme values than the region for the highest rainfall.

Sensitivity of the rainfall forecast to the Turbulence scheme

TKE and TKE_{mod} diffusion parametrization schemes are compared. Figures 3.3(a) and 3.3(b) show the daily rainfall quantile q_{99} for members 0 and 4, and Fig. 3.3(e) the corresponding quantile from the observations. We observe that the rainfall quantile threshold q_{99} is slightly higher with TKE_{mod} scheme. It is worth noting that TKE_{mod} quantiles are also closer to the observation than the control member ones, while both members show underestimated q_{99} compared to the observation. It is not possible to compare the third turbulence scheme L79, as there are no two members that only differ by the implementation of L79. We may presume that turbulence schemes affect quite marginally rainfall forecast spatial distribution.

Sensitivity of the rainfall forecast to the Shallow Convection scheme

Shallow convection parametrization is used to represent the turbulent transport of heat and moisture by non precipitating cumulus clouds. However, this physical parametrization is also important as it can influence the variables involved in the deep convection scheme and, consequently, impact rainfall forecast.

Four different Shallow Convection schemes are implemented in the PEARP multi-physics. A comparison between KFB and EDKF shallow convection parametrization

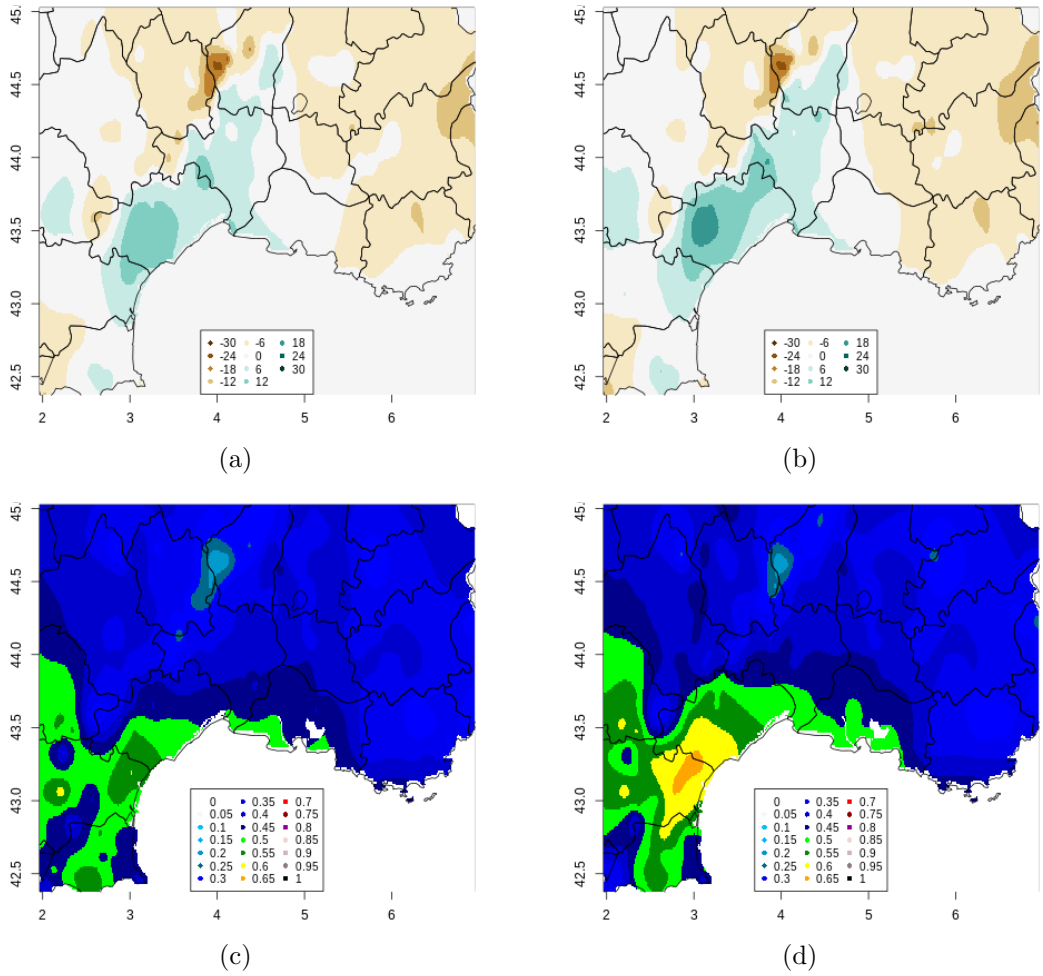
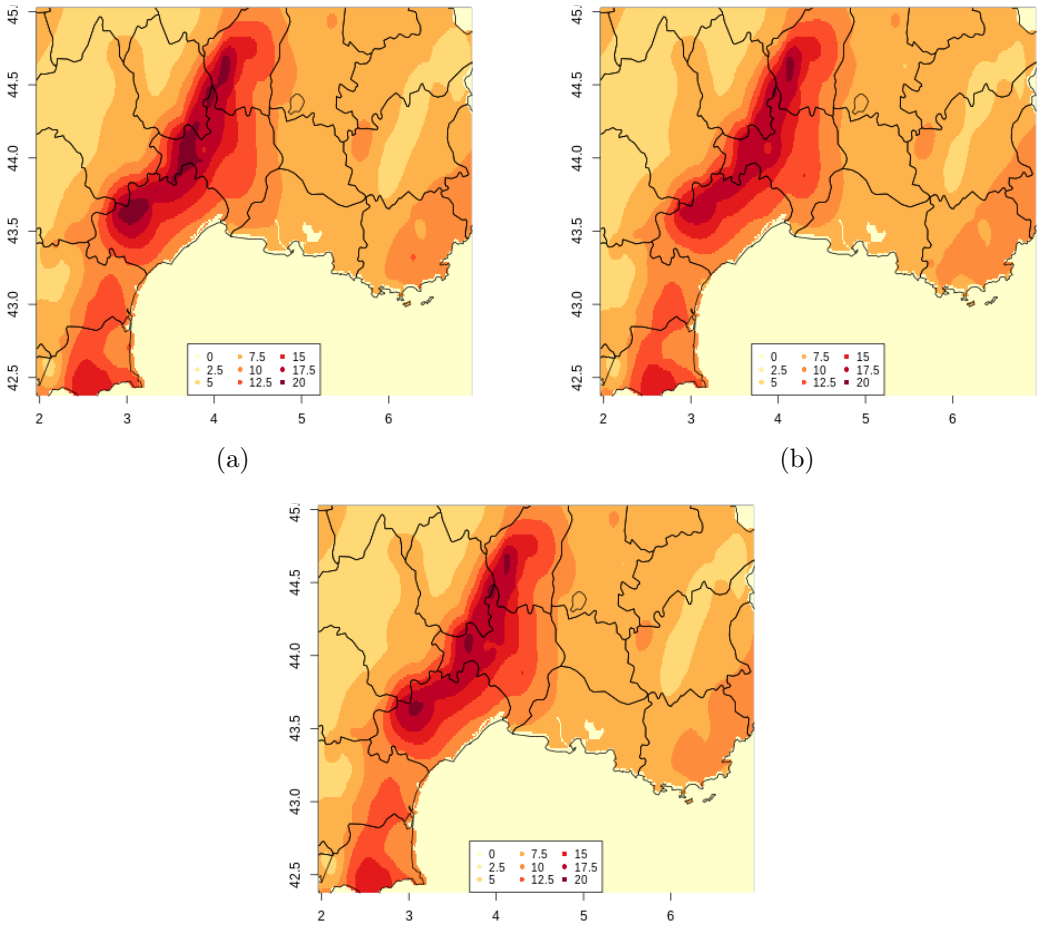


Figure 3.4: Top: Differences between the q_{95} 24-hour rainfall (mm) forecast and observation for the member 0 (a) and the member 5 (b) of the reforecast. Bottom: FAR metric scores for members 0 (c) and 5 (d) with quantile q_{90} threshold. The member 0 implements the KFB and member 5 the EDKF of the shallow convection schemes.



(a)

(b)

(c)

Figure 3.5: RMSE field of 24-hour rainfall amounts computed for members 6 (a), 7 (b) and 8 (c), implementing PMMC, KFB, and PCMT shallow convection schemes respectively.

schemes is done by comparing member 0 and member 5.

Figures 3.4(a) and 3.4(b) show the differences between these members and the observation for the rainfall quantile q_{95}. For both schemes, the quantile q_{95} is overestimated on the Languedoc-Roussillon region, typically upstream of the Massif Central foothills for most of the HPEs. This behaviour is observed also for the highest quantiles (not shown) and it is more significant for EDKF. The quantile q_{95} is underestimated over the mountain chain, with maximum underestimation on the Ardèche mountains. This discrepancy may be due to an excessive amount of precipitation produced by the model upstream of the flux impinging the Massif Central, resulting in a decrease of the potential precipitable water over the mountain chain. The same underestimation is present in the Alps region.

The False Alarm Ratio (FAR, see section 3.1.1) shows similar results using quantile q_{90} as threshold (Fig. 3.4, bottom panels). In Languedoc-Roussillon region, the FAR is maximum and it is larger for EDKF scheme (see panel (d)). The SEDI score analysis confirms this behaviour (not shown).

The three PMMC, KFB and PCMT shallow convection parametrizations, that are implemented in members 6, 7 and 8 are compared.

Figure 3.5 shows the RMSE scores computed for these three members. In general, the RMSE is higher over the Cévennes, the Ardèche mountains and the Pyrénées. Indeed, the daily rainfall climatological sample distributions tend to be more extreme in these areas compared to the plains. This error extends to the southeastern area of the Cévennes foothills. The comparison between the members reveals scarcely significant differences, showing a slightly worse performance concerning the PMMC scheme. KFB parametrization is associated to a slight higher value of SEDI score overall (not shown). PCMT shallow convection scheme gives intermediate results.

The inter comparison between the shallow convection schemes shows that model performances of 24-hour rainfall forecasts are barely sensitive to this kind of parametrization, although some differences are not completely negligible.

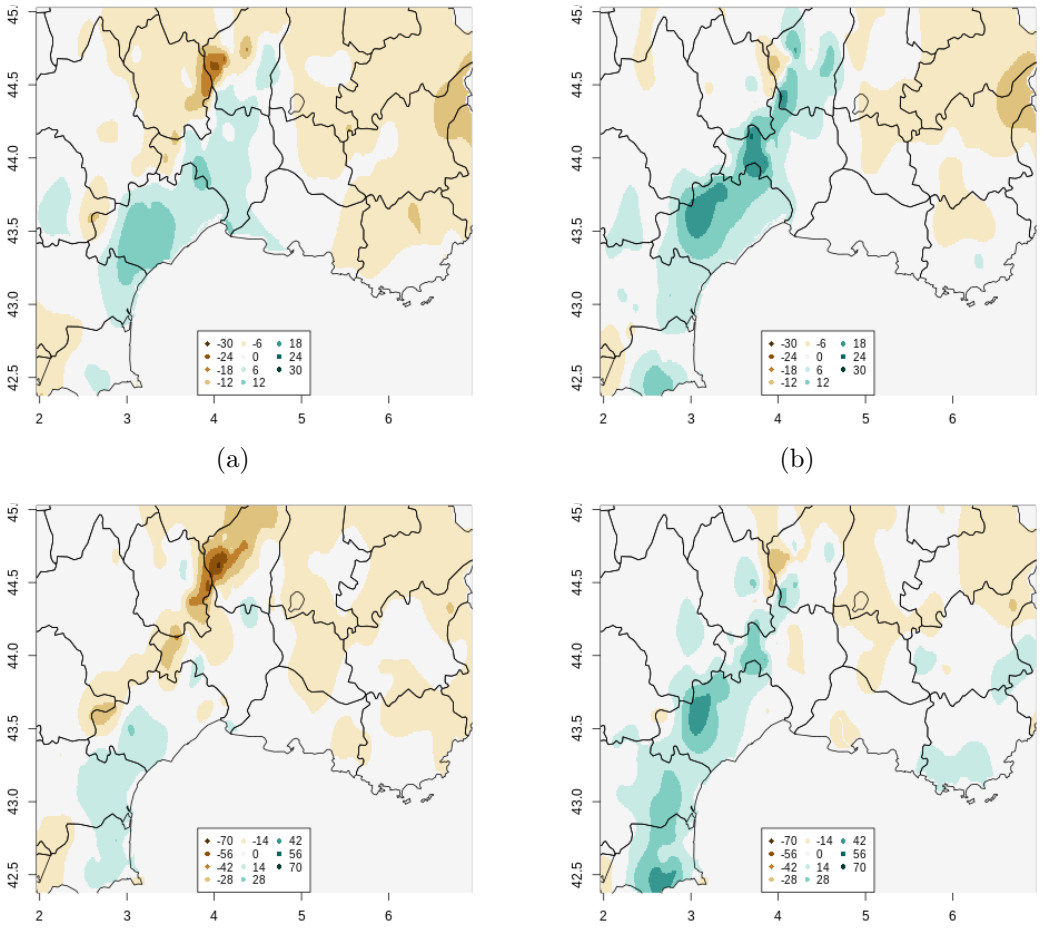


Figure 3.6: Difference between the observed 24-hour rainfall quantiles q_{95} (top) and q_{99} (bottom) and the model forecast for the member 0 (left) and the member 7 (right) implementing B85 and PCMT deep convection schemes, respectively (mm).

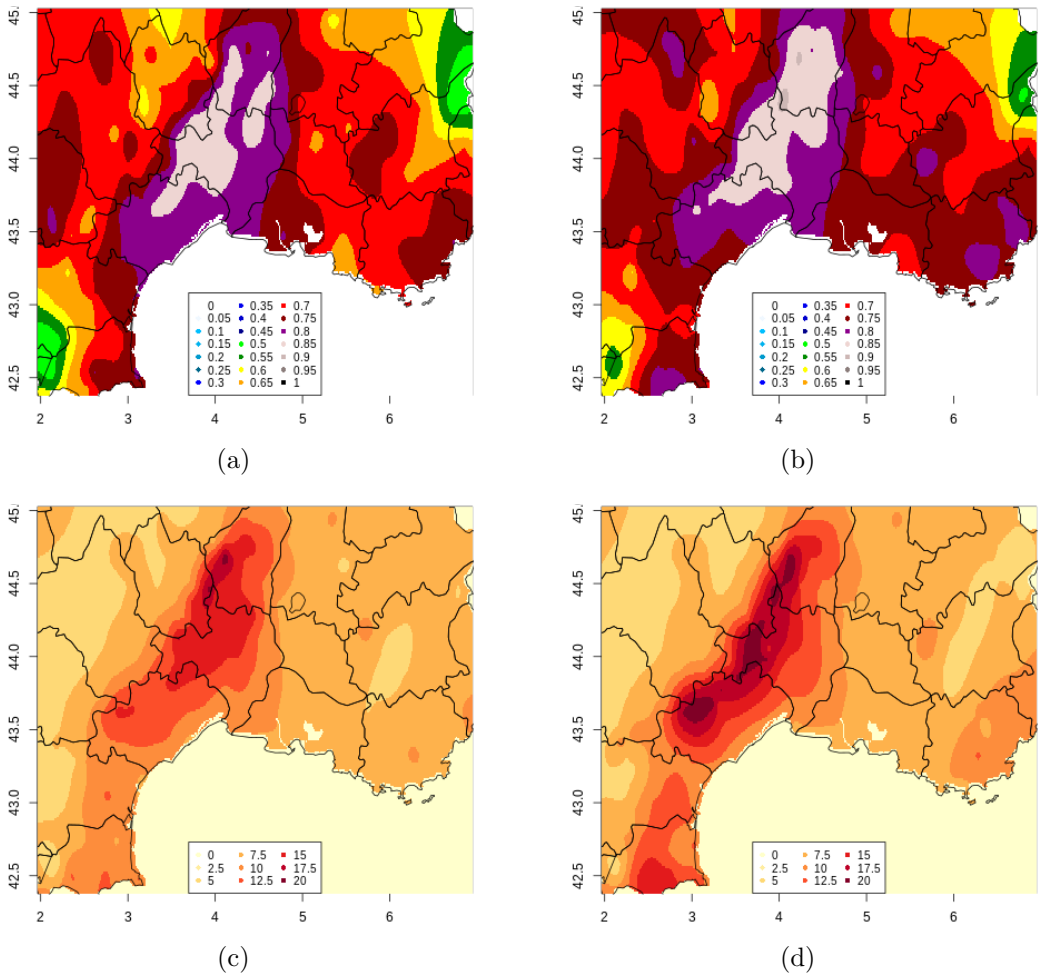


Figure 3.7: Top: SEDI score computed for members 0 (a) and 7 (b) with a threshold equal to the quantile q_{90}. Bottom: 24-hour rainfall amount RMSE computed for members 0 (c) and 7 (d), implementing B85 and PCMT deep convection schemes, respectively.

Sensitivity of the rainfall forecast to deep convection scheme

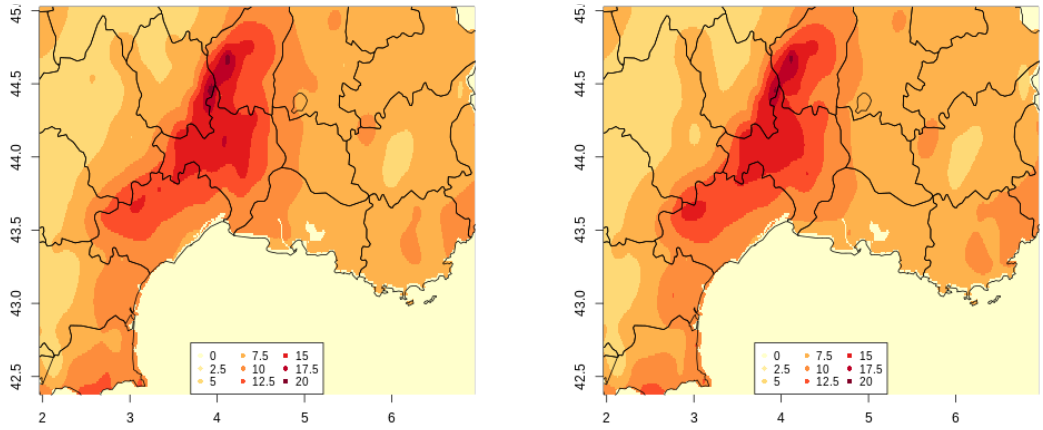
As described in section 2.2.2, our system implements two main deep convection frameworks, PCMT and B85. This is the only parametrization difference between members 0 and 7. Comparing them allows describing behaviour dissimilarities between B85 and PCMT deep convection parametrization schemes. For these two members, the shallow convection parametrization is KFB scheme.

An evaluation of the differences between these members of the reforecast and the observation for quantiles q_{95} and q_{99} is shown in figure 3.6. Differences in amplitude and spatial distributions of quantile biases are more significant for deep convection rather than for shallow convection. We can see that PCMT differences are larger than the B85 ones over Languedoc-Roussillon region. B85 quantiles are also significantly lower than the observed ones over the Ardèche, while PCMT quantiles are closer to the observations over this area. A moderate underestimation of quantile values is observed on the eastern side of the domain, especially for B85 scheme.

We analyse the SEDI scores for the q_{90} quantile threshold on figure 3.7(a,b). It is greater for the PCMT scheme. This may be related to the larger Hit Rates observed for PCMT, whereas False Alarms are similar for B85 and PCMT (not shown). We can also observe that the SEDI score reaches the maximum along the Cévennes chain. This suggests that the predictability tends to be higher over the mountains, where the convection is mainly triggered at a stationary location (Lin et al., 2001).

We also study the error magnitude through the RMSE score (Fig. 3.7, bottom). The non-linearity of the RMSE metric leads to large errors where the rainfall is higher. The RMSE values computed for PCMT scheme are high on the side of the Cévennes foothills positioned towards the Mediterranean Sea. RMSE are reduced for the B85 scheme.

We compare two other deep convection schemes: B85_{mod} and CAPE, which are implemented in members 2 and 3. As already detailed in section 2.3, these two deep convection parametrization schemes correspond to versions derived from B85 scheme. B85_{mod} corresponds to a convection trigger parameter based on the cloud top height, and CAPE scheme is based on the CAPE closure. Corresponding RMSE



(a) (b)

Figure 3.8: 24-hour rainfall amount RMSE for members 2 (a) and 3 (b), implementing $B85_{mod}$ and CAPE deep convection schemes, respectively.

results are shown in Fig. 3.8. Differences between these two modified versions of the same parametrization scheme appear to be negligible.

Oceanic flux

Precipitation forecasts obtained from the parametrization of oceanic flux $ECUME_{mod}$ exhibit daily rainfall spatial distributions and amounts similar to the standard ECUME scheme (not shown). The difference between the two schemes is related to a parameter that control the evaporation fluxes over the sea. This suggests that 24-rainfall forecast is barely sensitive to the two different parametrizations of oceanic flux. This result may suggest that the modification of this parameter in PEARP is too weak to produce significant difference in precipitation forecasts.

3.2 Probabilistic forecast verification

The previous section was dedicated to the exploration of the impact of the physical parametrization on the deterministic 24-hour precipitation forecast. In this part, the ten members of the reforecast are gathered and evaluated as a reduced version of the operational ensemble system PEARP. Even if we already think that the ensemble performance would be drastically weakened by a too small number of members, we assume that the long duration of the period, hence the meteorological variability

explored, would help to obtain robust probabilistic diagnostics. This hypothesis is a main keystone in this work and will also be the basic assumption assessed in the two calibration experiments described in the next chapter.

A first exploratory analysis of the reforecast ensemble scores is performed with some standard probabilistic verification measures.

The probabilistic forecast skills are evaluated following the recommendation proposed by Hamill and Juras (2006). The authors demonstrate that an evaluation can be erroneous if the score is computed for a fixed value on various samples spanning many locations and dates. Indeed, the corresponding samples event frequencies may vary significantly, and lead to a poor representation of the overall. This undesirable behaviour is prevented by considering events whose climatological frequency is not dependent on the sample choice. This is achieved by considering the quantile of the local climatological distribution (Zhu et al., 2002).

3.2.1 Probabilistic forecast verification metrics

A scoring rule for a probabilistic forecast is a summary of a measure that evaluates the probability distribution.

One desirable feature of scoring rules is to be proper. A score is strictly proper if it reaches its optimal value when the predicted distribution is identical to the verification one (Gneiting and Raftery, 2007). Only a tiny fraction of the classical scores used in forecast verification are proper. Practically, non-proper scores are also used but with the caution that high scores have to be interpreted carefully.

Hereinafter a survey of the scoring rules computed in this work is presented, specifying if the score is proper and showing some basic features thereof.

Brier Score

Brier Score (Brier, 1950) is a metric largely used in verification problems involving probability forecasts for dichotomous predictands. It corresponds to the mean squared error of the probability forecasts, considering that the observation $o = 1$ if the event occur and that it is 0 otherwise. The event occurrence is usually defined by threshold overrun. The score stands for the averages of the squared differences

between pairs of forecast probabilities and binary observations:

$$BS = \frac{1}{n}\sum_{k=1}^{n}(y_k - o_k)^2 \tag{3.4}$$

where the index k browses the n forecast-event pairs. y_k represents the forecast probability of the event, and o_k is the event occurrence (0,1). The analogy with RMSE metric (eq. 3.3) is clear, except that for RMSE, (y_k,o_k) represent the pairs of continuous values related to the forecasts and the observations. The Brier score is negatively oriented.

The Brier score is a proper score. In this respect, Bröcker (2009) demonstrated that the expected value of strictly proper rules allows for a decomposition into three terms: reliability, resolution and uncertainty. A brief summary of decomposition terms is given, following Wilks (2009b).

Given the discrete number of members of an ensemble, forecast probabilities can take a finite number I of allowable values. I corresponds to $m+1$ if m is the number of members of the ensemble. Sampled forecast values y_i can be described by means of a finite arithmetic progression as follows:

$$y_i = \frac{(i-1)}{m}, \quad 1 \leq i \leq I \tag{3.5}$$

If we define N_i as the number of times each forecast probability y_i is used, the total number of forecast-event pairs is the sum of the conditional sample sizes:

$$n = \sum_{i=1}^{I} N_i \tag{3.6}$$

The marginal distribution of forecasts is then easily evaluated for each forecast probability interval:

$$p(y_i) = \frac{N_i}{n} \tag{3.7}$$

It is possible to stratify the frequency of occurrence of the event according to the subsamples delineated by the I forecast probabilities. We may consider this

conditional average observation relative to the forecast value y_i:

$$\bar{o}_i = p(o_1|y_i) = \frac{1}{N_i} \sum_{k \in N_i} o_k, \tag{3.8}$$

where o_k correspond to the event paired with the forecast value y_k, $o_k = 1$ if the event occurs, $o_k = 0$ otherwise. The sample climatology is then given by:

$$\bar{o} = \frac{1}{n} \sum_{k=1}^{n} o_k = \frac{1}{n} \sum_{i=1}^{I} N_i \bar{o}_i \tag{3.9}$$

The Brier Score (eq. 3.4) can then be decomposed as the sum of three terms:

$$BS = \underbrace{\frac{1}{n} \sum_{i=1}^{I} N_i (y_i - \bar{o}_i)^2}_{\text{"Reliability"}} - \underbrace{\frac{1}{n} \sum_{i=1}^{I} N_i (\bar{o}_i - \bar{o})^2}_{\text{"Resolution"}} + \underbrace{\bar{o}(1 - \bar{o})}_{\text{"Uncertainty"}} . \tag{3.10}$$

The first term represents the reliability, which is the statistical coherence between the observation and the related forecast. A forecast is perfectly reliable if the forecast probability matches the conditional distribution of observations for each forecast probability category. This term is expected to be as low as possible and is generally called REL.

The second term is the resolution term. The resolution can be defined as the ability of a forecast system to classify the different observations according to the corresponding forecast. If the forecast sort the events into subsamples with relative frequencies similar to the sample climatology, the forecast system is poor in resolution. This term is then expected to be as high as possible and is usually referred as RES.

The third term is the uncertainty term. It depends only on the sample climatology. This term reaches its maximum for $\bar{o}$ equal to 0.5 and two minima are observed for $\bar{o} = 0$ and $\bar{o} = 1$, i.e. when the event almost always happens or almost never happens. In the case of HPEs, we are in the second situation and then scores can be artificially high due to the strength of $(1 - \bar{o})$ term.

Brier Skill Score

The Brier Skill Score refers to the evaluation of a prediction system with the Brier score but with respect to a reference one:

$$BSS = 1 - \frac{BS}{BS_{ref}}. \tag{3.11}$$

The simplest interpretation of the BSS is that, for negative values BS is larger than BS_{ref}, which means that the probabilistic forecast has less skill than the reference. The skill score ranges between $-\infty \leq BS \leq 1$. In this study, the reference forecast will be defined as the climatological relative frequency computed from the observation along the 30-year period covered by the reforecast during the September-December season.

The combination of equations 3.11 and 3.10 yields to the following decomposition of BSS relation:

$$BSS = 1 - \frac{BS}{BS_{ref}} = 1 - \frac{REL - RES + UNC}{UNC} = \frac{RES - REL}{UNC}, \quad \text{if } BS_{ref} = UNC \tag{3.12}$$

The relationship is verified only if BS_{ref} is computed from a sample climatology equal to the frequency of occurrence of the event in the verification sample dataset. This formulation of the Brier Skill Score infers that a forecast system skill is related to the proportion of REL and RES. When the resolution exceeds the reliability term, BSS becomes negative.

CRPS

The CRPS (Continuous Ranked Probability Score; Brown, 1974; Hersbach, 2000; Matheson and Winkler, 1976) is a scoring rule strictly proper to evaluate probabilistic forecast of a continuous predictand. The score is based on the integration of distance between the forecast and observation distributions:

$$CRPS = \int_{-\infty}^{+\infty} \left[C(x) - C_o(x)\right]^2 dx, \tag{3.13}$$

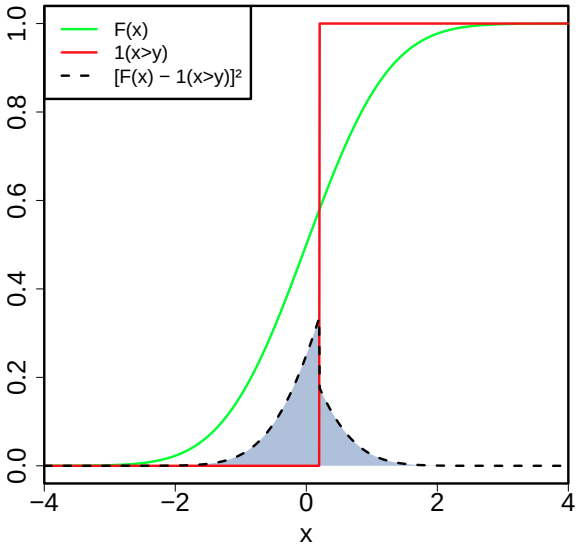


Figure 3.9: Illustration of the Continuous Ranked Probability Score. The green line corresponds to the CDF of the forecast, while the red line represents the Heaviside function corresponding to the observation. The dashed line corresponds to the squared difference between the two curves.

where $C(x)$ is the Cumulative Density Function (CDF) of the ensemble forecast of the variable x and $C_o(x)$ is given by:

$$C_o(x) = H(x - x_o), \tag{3.14}$$

where

$$H(x) = \begin{cases} 0 & \text{for } x < 0 \\ 1 & \text{for } x \geq 0. \end{cases} \tag{3.15}$$

is the Heaviside function. A perfect score is equal to zero and it can be obtained only in case of a deterministic forecast, or an ensemble forecast with identical members, which targets exactly the observed value.

A graphical illustration of the CRPS definition is shown in Fig. 3.9. The CRPS corresponds to the integrated area below the dashed curve, which is the squared difference between the CDF of the forecast $C(x)$ and the Heavyside function corresponding to the observation x_o. An heavily scattered ensemble forecast is associated with a stretched CDF, resulting in a high CRPS value. On the other hand, a sharp forecast will provide a low value of CRPS, as long as the forecasted values are close to the true value x_o.

CRPS score has the dimension of the parameter x used for the verification and it can be interpreted as the integration of a Brier Score over all possible thresholds or to the MAE for a deterministic forecast. CRPS is often used as it provides scores not related to predefined classes or thresholds and as it is sensitive to the whole range of values of the parameter of interest.

This metric can then be averaged over n points and/or cases during a given period:

$$\overline{CRPS} = \sum_{k=1}^{n} CRPS(C^k, x_o^k). \tag{3.16}$$

Such as for the BS (eq. 3.10), the $\overline{CRPS}$ can be decomposed as described in Candille and Talagrand (2005). An alternative decomposition is introduced in Hersbach (2000), who applied the decomposition for an ensemble system. In this case, a straightforward manner to estimate the CDF lies on computing the empirical CDF from the m members of the ensemble, assigning a probability of $1/m$ to each bin, sorting the forecast data in ascending order. Then the assigned probabilities up to and including each bin are summed to build the empirical CDF.

Using this method, the decomposition of $\overline{CRPS}$ over the n events yields to:

$$\overline{CRPS} = \underbrace{\sum_{i=0}^{m} \bar{g}_i(\bar{o}_i - p_i)^2}_{\overline{Reli}} + \underbrace{\sum_{i=0}^{m} \bar{g}_i \bar{o}_i(1 - \bar{o}_i)}_{CRPS_{pot}}, \tag{3.17}$$

where $\bar{g}_i$ is the average width of forecast bin, $\overline{x_{i+1} - x_i}$, or the distance between the observed value and the corresponding closer member for the distribution outliers. The $\bar{o}_i$ can be seen as the average frequency to which the observation is found below the middle of the i^{th} interval, and $p_i = \frac{i}{m}$ is the i^{th} cumulative probability of the ensemble system.

The first term is also called a reliability term, which does not have the same score meaning than in the Brier decomposition (see 3.10). Here this term provides an information similar to the rank histogram (Hamill, 2001), a verification tool that will be introduced hereinafter. $CRPS$ reliability tests whether, on average, the frequency o_i that the verifying observation is found below the middle of the i^{th} interval bin of the predictive CDF is proportional to the probability associated

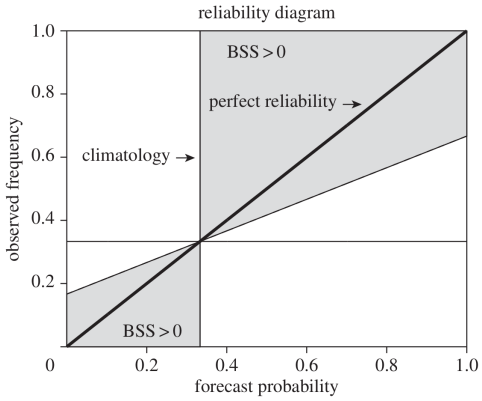
to this interval (i/m). One difference compared to the rank histogram is that the reliability term of CRPS takes into account the width of the bins, so more weight is given to large intervals in the predictive CDF.

The second term $CRPS_{pot}$, called the potential CRPS, represents the CRPS which would be obtained for a perfect reliable system ($\overline{Reli} = 0$). Hersbach (2000) shows that it is possible to further decompose $CRPS_{pot}$ into uncertainty and resolution terms. The $CRPS_{pot}$ is directly related to the spread of the ensemble. Indeed, a high ensemble spread would lead to greater $\bar{g}_i$. It is also sensitive to the outliers in the observations, corresponding to observed values that fall outside the extreme forecasted values. Outliers induce a growth of the $CRPS_{pot}$ term. This means that low $CRPS_{pot}$ is obtained when the compromise between a narrow ensemble spread and a low number of outliers is achieved.

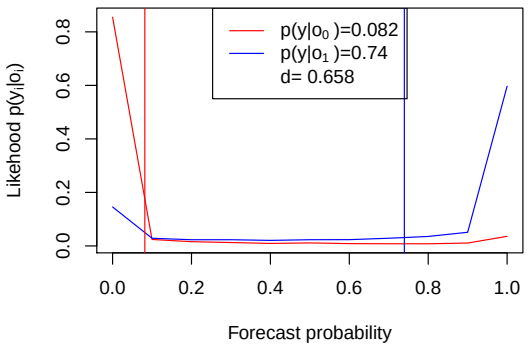
The $\overline{CRPS}$, the reliability and potential terms are negatively oriented and they do not have a superior limit. The score is perfect when $\overline{CRPS}$ is null. Hereafter, for practical reason the $CRPS$ acronym will refer to $\overline{CRPS}$.

The Reliability Diagram

The reliability diagram has been extensively used in verification literature (Atger, 2004; Jolliffe and Stephenson, 2012; Murphy and Winkler, 1987) where it stands for the reliability as the ability of the model to produce probabilities of an event corresponding to its observed frequency. In the decomposition terms of Brier Score previously presented, a system is perfectly reliable if $p(o_1|y_i) = y_i$ for all the probability intervals. In a reliability diagram (Fig. 3.10(a)) the horizontal axis shows the forecast probability y_i, whereas the ordinate axis corresponds to the conditional distribution of the observations $\bar{o}_i$. Therefore, the departure between the points and the bisector is a measure of the lack of reliability. Similarly, the departure between the points and the sample climatology lines corresponds to the square root of the resolution term of the Brier Score. From the equation 3.12, if the BS_{ref} is set equal to the sample climatology of the verification dataset, then the BSS is positive as long as the resolution RES term is larger than the reliability REL term. The area represented by the gray zone in the graph shows this area of skill.



(a) Reliability diagram. From: Weisheimer A. and Palmer T. N. (2014)



(b) Discrimination diagram

Figure 3.10: a) Example of a reliability diagram. The horizontal and vertical lines show the sample climatology of the event. The bisector shows a perfect forecast, while the gray area defines the area of positive skill of the BSS. b) Example of a discrimination diagram. Red line shows the conditional probability for dry events, and the blue line the conditional probability for wet events. Vertical lines summarize the mean of the distributions, whose values are displayed in the legend. The discrimination distance d is also shown.

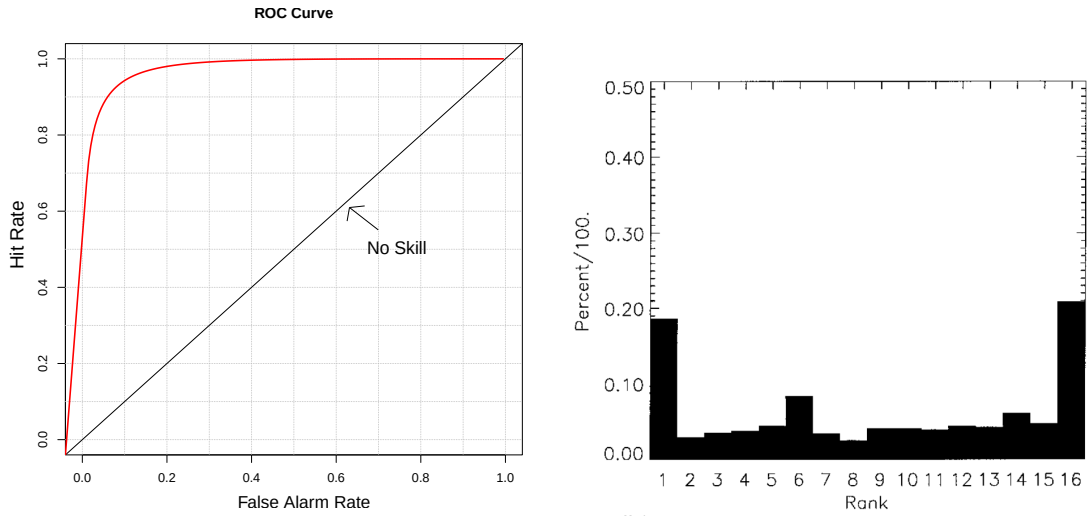
The Discrimination Diagram

The discrimination diagram measures the ability of the forecast to differentiate the observation categories. Whereas the reliability diagram is built on the conditional distribution $p(o_1|y_i)$, the discrimination diagram is based on both conditional likelihood distributions $p(y_i|o_j)$ for the two dichotomous events ($j = 0$, or $j = 1$, event non-occurrence and occurrence, respectively).

An example of discrimination diagram is shown in figure 3.10(b). The two distributions are drawn for a ten member ensemble ($I = 11$) and for a rain event. o_1 refers to the cases when the event occurs, o_0 when it does not occur. Then, the more distinct the tails of the distributions are, the more discriminant the forecast is. The magnitude of this distinction between the two distributions is diagnosed with the discrimination distance:

$$d = |\overline{p(y|o_1)} - \overline{p(y|o_0)}|, \tag{3.18}$$

which corresponds to the absolute difference between the means of the two distributions. A large value of d suggests high level of discrimination of the forecast



(a) ROC diagram

(b) Rank histogram

Figure 3.11: a) Example of a ROC diagram. On x-axis the False Alarm Rate and on the y-axis the Hit Rate. The diagonal denote the limit of no skills. b) Example of a rank histogram (from: Hamill and Colucci (1997)).

system.

The ROC curve

The ROC (Relative Operating Characteristic; Hanley and McNeil, 1982; Mason, 1982) scoring rule evaluates the ability of the forecast to discriminate between events and non-events. Like in the discrimination diagram, the ROC is conditioned by the observations. Considering a given threshold, the contingency table elements Hit Rate H and False Alarm Rate F are taken into account. The Hit Rate is plotted as a function of the False Alarm Rate for various forecast probabilities. Figure 3.11(a) shows an example of ROC diagram. The two extremes of the ROC curve are by construction set on the origin and on $(x = 1, y = 1)$ even when these cases are not sampled. The perfect forecast exhibits $F = 0$ and $H = 1$, and is associated with an area beneath the curve equal to 1. Conversely, random forecasts would result in the same value of F and H for all the probability thresholds, represented by the diagonal. Beneath this line, the ROC score depicts a no skill forecast. It is worth mentioning that, even if this scoring rule remains a basic method in meteorological verification, AUC (Area Under the Curve) is not a proper score (Byrne, 2016), and

it is sensitive to subsampling.

Rank Histogram

The rank histogram (Anderson, 1996; Hamill and Colucci, 1997; Talagrand et al., 1997) is another scoring rule aimed at evaluating the reliability of the ensemble system. More specifically, it tries to answer the question: are the ensemble members statistically indistinguishable from the verification data?

The histogram is built by sorting each observation among the potential $m+1$ ordered bin values of the forecast. It can be considered as a frequency histogram of counting the observations according to ascending members values. If the m ensemble members and the observations would have been drawn from the same distribution the histogram would be equally distributed. Figure 3.11(b) shows an example of a rank histogram drawn from a 15 ensemble-member forecast system. The U-shape indicates that more observations are classified outside the ensemble values, then the ensemble values are not spread enough. Although the rank histogram is widely used for reliability estimation, Hamill (2001) suggests it is not a sufficient condition to determine if an ensemble is reliable.

Evaluation of the ensemble spread

The spread of the ensemble is computed assuming exchangeability between forecasts and observations, as in Fortin et al. (2014). It is expressed as follows:

$$SPREAD = \sqrt{\left(\frac{m}{m-1}\right)\frac{1}{n}\sum_{k=1}^{n} s_k^2}, \tag{3.19}$$

where

$$s_k^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\bar{x}_{f,k} - x_{f,k,i})^2. \tag{3.20}$$

and n is the sample size, m the ensemble size, s_k^2 an unbiased estimator for the variance of the ensemble members, $x_{f,k,i}$ is the i^{th} ensemble member for the event k and $\bar{x}_{f,k}$ is the ensemble mean for the event k. The factor $\left(\frac{m}{m-1}\right)$ in equation 3.19 vanishes for large ensemble sizes. In this study, this term is retained for a 10-member

ensemble.

The ensemble reliability can be diagnosed through the gap between the spread and the root mean squared error of the ensemble mean. For a reliable system RMSE and Spread are close together (Whitaker and Loughe, 1998):

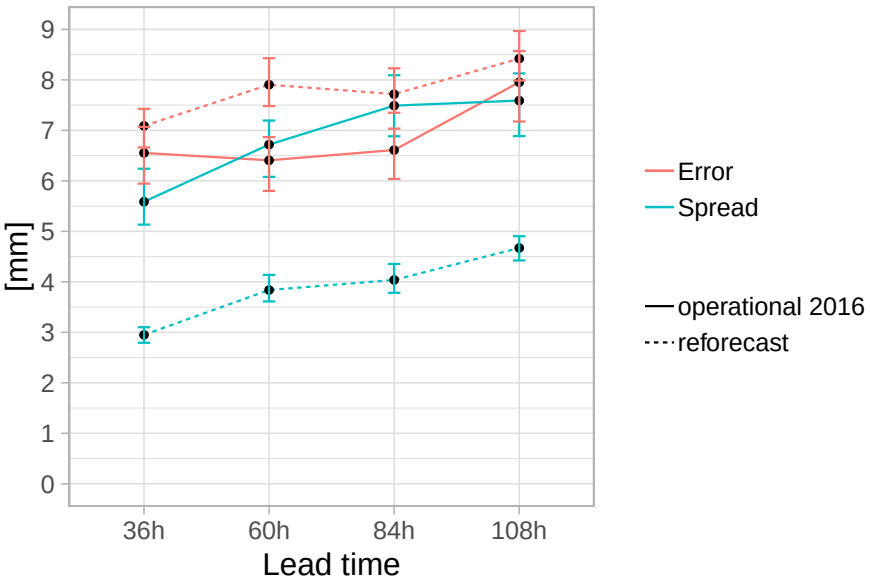
$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\bar{x}_{f,k} - x_{o,k}\right)^2} \tag{3.21}$$

3.2.2 Verification scores applied the reforecast

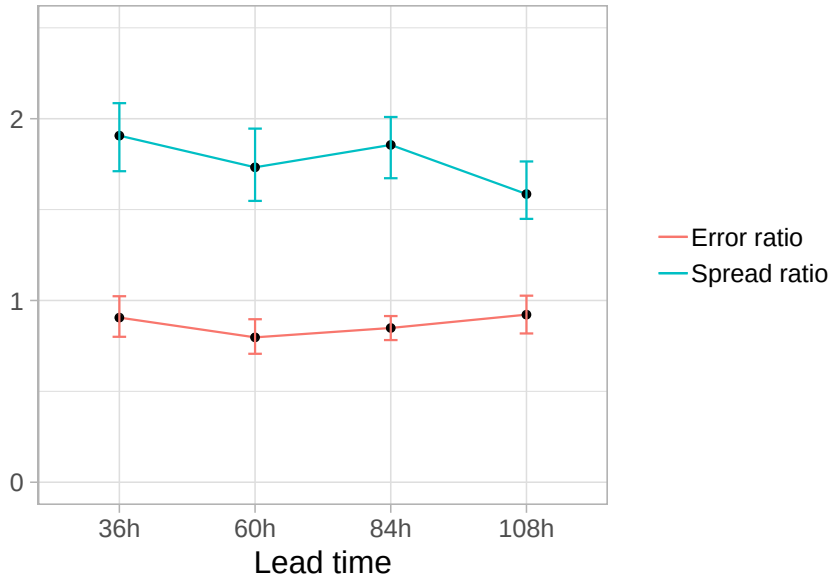
Some of the scores described in the previous section are applied to the whole reforecast dataset over a 4-month period of 30 years. All scores are averaged over the 4-month period, i.e. inner season trends of rainfall events are not taken into account in occurrence frequency.

First, considering the difference of the reforecast set-up with respect to the operational system, we think it is important to assess the spread difference. To allow this comparison, we had to process to a few data arrangements. Starting from the PEARP-2016, described in section 2.2.3, the 35-member PEARP is resized to 10 members. We name this adapted dataset 10M-PEARP. In this way, we will compare two ensembles with the same size, the same physics, but with only the initial conditions perturbations that are missing in the reforecast dataset. This procedure is carried out by sampling three sets of ten members from PEARP-2016. In each set, each of the 10 members uses a different physical package which corresponds to one of the physics of the reforecast. This sampling procedure is repeated three times in order to increase the size of the dataset used to get comparable scores. Bootstrapping sampling technique described in Efron and Tibshirani (1994) or in Candille et al. (2007) is applied to the Spread/Error diagnostic. A 90% confidence interval is chosen, based on 100 different samplings for the reforecast dataset as well as for the 10-M PEARP dataset. The results are averaged over the whole set of the domain.

Figure 3.12(a) shows the Spread-Error relationship for the 10-M PEARP and the reforecast datasets. First, the ensemble spread actually grows along with lead time. We can observe a notable gap between the 10-M PEARP spread and the



(a)



(b)

Figure 3.12: a) Spread-error relationship. Solid lines are referred to 10-M PEARP, dashed lines to the reforecast. In red the error, in blue the spread. b) Ratio between the errors associated with the 10-M PEARP and the reforecast (red line), and ratio between the spreads (blue line).

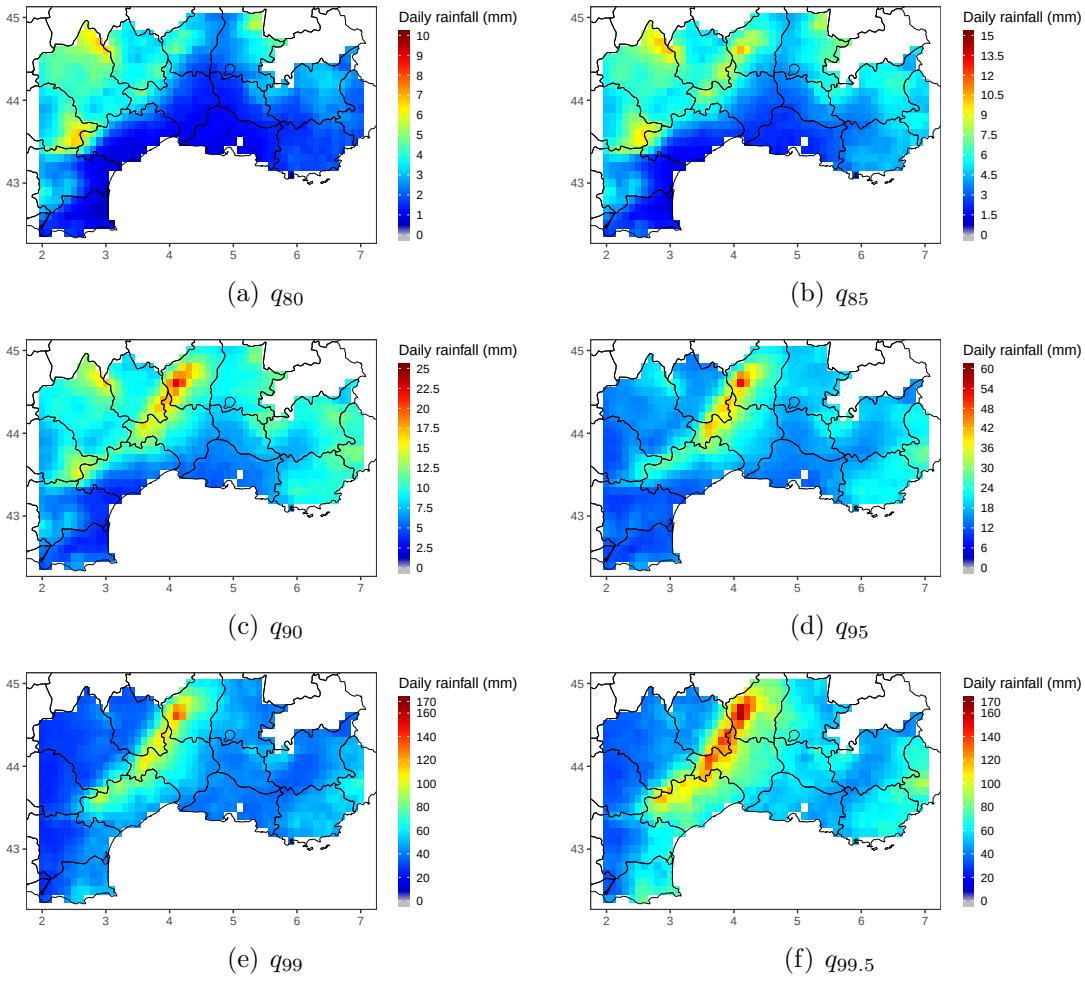
reforecast one which is much lower. It must be related to the lack of initial condition perturbations in the reforecast at the first lead time of the forecast. Then, this deficiency is never compensed along the forecast. For the 10-M PEARP ensemble, the spread seems to stop increasing after 92-hour lead time, and it is not the case for the reforecast. The spread-error relationship is close to 1. This shows that 10-M PEARP exhibits high reliability. Reforecast error is approximately twice compared to the spread, suggesting a lower reliability than 10-M PEARP.

The $RMSE$ is also depicted in Fig. 3.12(a). Reforecast values are larger than the operational ensemble ones, but with less difference than for the spread. We think this difference can be related to the quality of initial conditions employed for the reforecast. Indeed, PEARP ensemble system is built from AEARP data assimilation system, whereas the reforecast initial conditions come from the ERA-Interim reanalysis (Berrisford et al., 2011; Dee et al., 2011). The poorer resolution of the reanalysis can be source of larger errors in the reforecast.

The differences between the 10-M PEARP and the reforecast is further investigated in Fig. 3.12(b). We plotted the ratio between the 10-M PEARP and the reforecast $RMSE$ and $SPREAD$. The spread ratio is bounded between 1.5 and 2. The lowest value is achieved for the longest lead time, suggesting that the initial condition perturbation tends to have higher impact on the spread at the first lead times.

As already mentioned at the beginning of this chapter, probabilistic verification is performed on the model grid points. In the following, we present the result of the ensemble verification of the reforecast. In order to focus on the right tail of the rainfall distribution most of the scores will be computed on the rainfall observation reference (see section 2.4) for six quantiles thresholds: q_{80}, q_{85}, q_{90}, q_{95}, q_{99} and $q_{99.5}$.

The spatial dependence of these quantiles values from the interpolated rainfall fields is shown in Fig. 3.13. We can observe that orography significantly affects the rainfall highest events. Local maxima are located over the Cévennes and the Ardèche mountains, except for q_{80} threshold. We note that depending on the quantiles, two different rainfall regimes could be determined. Quantile thresholds below q_{90} exhibits

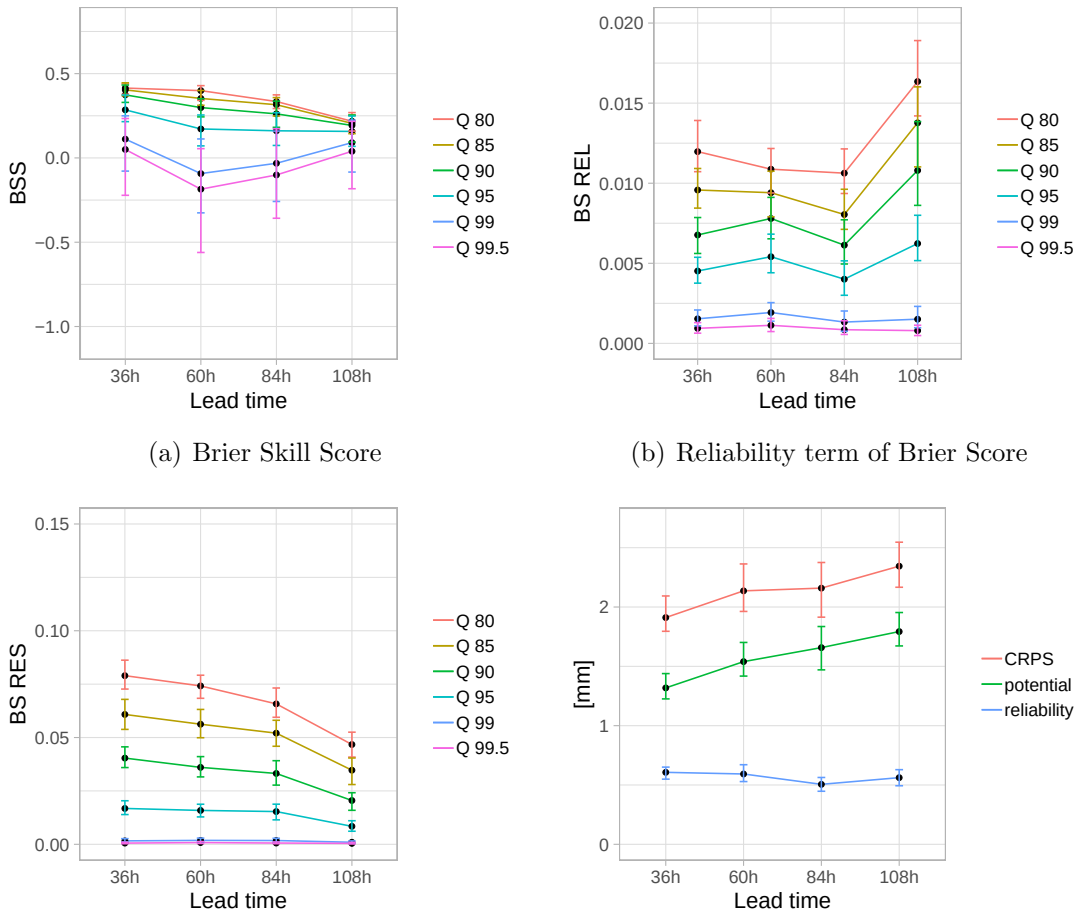


(a) q_{80} (b) q_{85}

(c) q_{90} (d) q_{95}

(e) q_{99} (f) $q_{99.5}$

Figure 3.13: Graphical illustration of the 24-hour rainfall amounts computed from the observation reference for the quantile thresholds q_{80}, q_{85}, q_{90}, q_{95}, q_{99} and $q_{99.5}$. For graphical reason the range values differ from one legend to another.



(a) Brier Skill Score

(b) Reliability term of Brier Score

(c) Resolution term of Brier Score

(d) Continuous Ranked Probability Score

Figure 3.14: BSS (a), terms of BS (b,c) and CRPS (d) computed for the reforecast dataset and for 24-hour rainfall. CRPS is decomposed in reliability and potential terms. Error bars are estimated using a bootstrapping sampling technique and cover the 90% interval.

an enhanced signal northwest of the Cévennes chain, while for the largest quantile thresholds the Languedoc-Roussillon region and the southeastern side of Cévennes chain are more affected. It is worth noting that an enhanced signal is also present over the Pyrénées and the Alps. The first regime seems to be associated with a westerly low-level flow, while the second one with a southerly low-level flow.

BSS and CRPS computed using a set of quantile thresholds are shown in Fig. 3.14. Scores are averaged over all the grid points of the domain. Error bars are estimated using a bootstrapping sampling technique, based on 100 resampling for each score. The bars cover the 90% confidence interval. First, the BSS (Fig. 3.14(a)) is positive for thresholds lower than q_{95}, and shows no skill for the largest thresholds

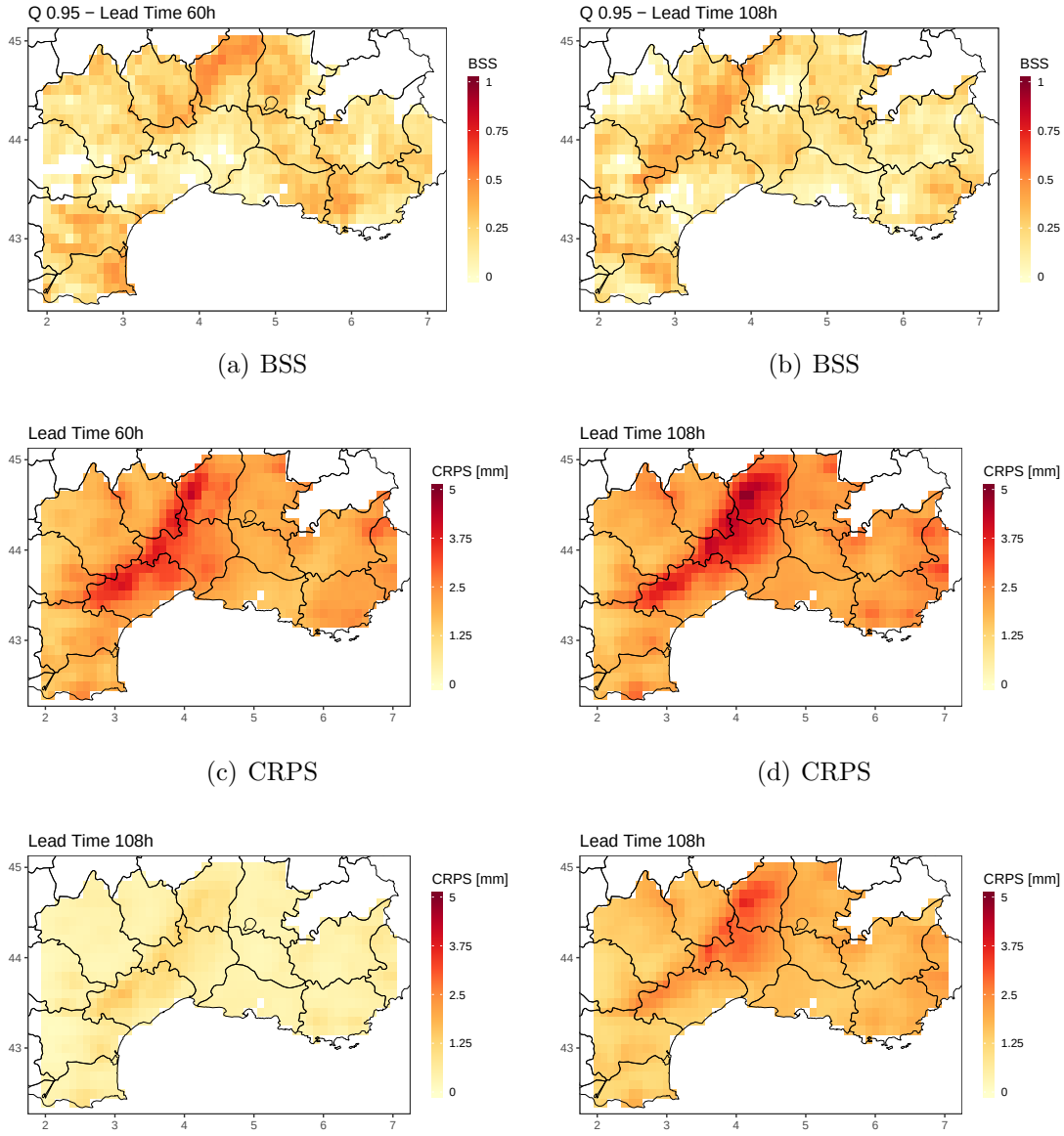
(q_{99}, $q_{99.5}$). As expected, the model is more skillful for low thresholds, and gets worse for the longest lead times. BSS variability (errors bars) is getting larger for q_{99} and $q_{99.5}$ because events for that specific thresholds become extremely rare, respectively counting only $\approx$ 36 and $\approx$ 18 per grid points over the 30-year period. For the highest thresholds, there is no a clear dependence of the score with lead time. It may be due to a misrepresentation of rare events among lead times over the whole period with respect to the 4-day sampling of the reforecast (reforecast is run every 4-days). The reliability term of BS is globally unchanged until 92-hour lead times (Fig. 3.14(b)). Then the reliability decreases at 108-hour lead time for almost all quantiles. Resolution decreases with the lead time (Fig. 3.14(c)).

The CRPS score shows increasing value with the lead time (Fig. 3.9), indicating worsening of the forecast skill. We can see that this is mainly related to the potential term of CRPS which increases with lead time, while the reliability term remains almost constant.

We already noted, that one main interest of the reforecast verification is that we still have significant scores at the grid-point scale. It is confirmed by the analysis of BSS and CRPS spatial scores (Fig. 3.14). BSS computed for the quantile q_{95} at 60-hours lead time is shown in Fig. 3.15(a) and at 108-hours lead time in Fig. 3.15(b). BSS shows more skill over the mountainous areas, while some negative values (white areas) of BSS are observed over some of the plain areas. The analysis of the corresponding reliability diagrams over some targeted points over these zones (not shown) reveals that forecasts tend to be more skillful on the Cévennes/Ardèche mountains, whereas poor resolution and wet bias are associated with the no-skill forecasts points.

BSS computed for the most extreme quantile thresholds (larger than q_{95}, not shown) are inconsistent due to the too small number of events. Indeed, since BS and $\mathrm{BS_{ref}}$ tend to zero for rare events, a small variation of the score can lead to a significant variation of the BSS.

CRPS is computed for the same lead times (Fig. 3.15(c) and 3.15(d)). Due to the sensitivity of the CRPS to the magnitude of the error, CRPS takes large values over the relief, where 24-hour rainfall distribution exhibits larger values. This effect



(a) BSS (b) BSS

(c) CRPS (d) CRPS

(e) CRPS reliability (f) CRPS potential

Figure 3.15: BSS of 24-hour precipitation computed on each grid-point for q_{95} at 60-hour (a), 108-hour (b) lead times. CRPS of 24-hour precipitation computed on each grid-point at 60-hour (c), 108-hour (d) lead times. Decomposition of the CRPS of figure (d) in reliability (e) and potential (f) terms at 108-hour lead time.

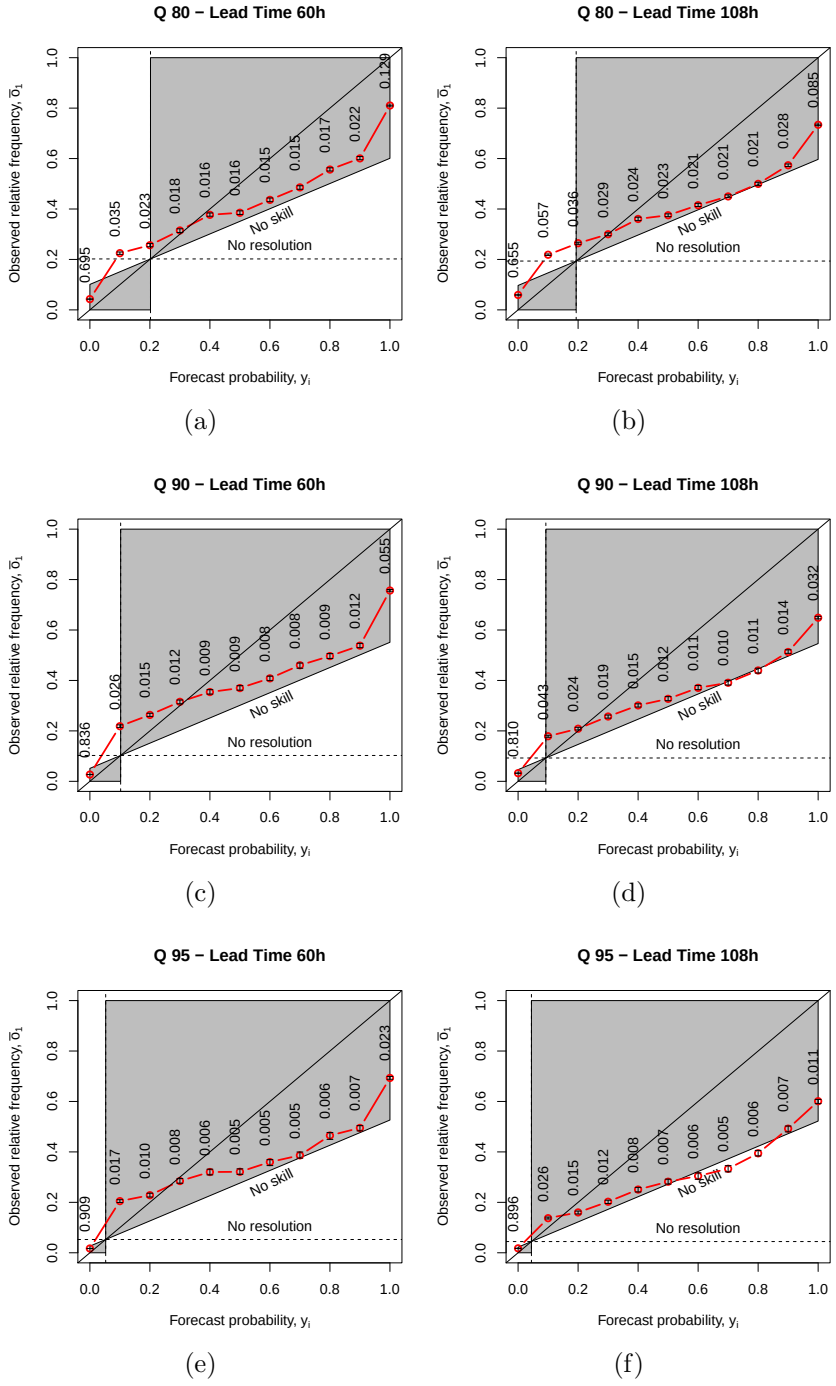


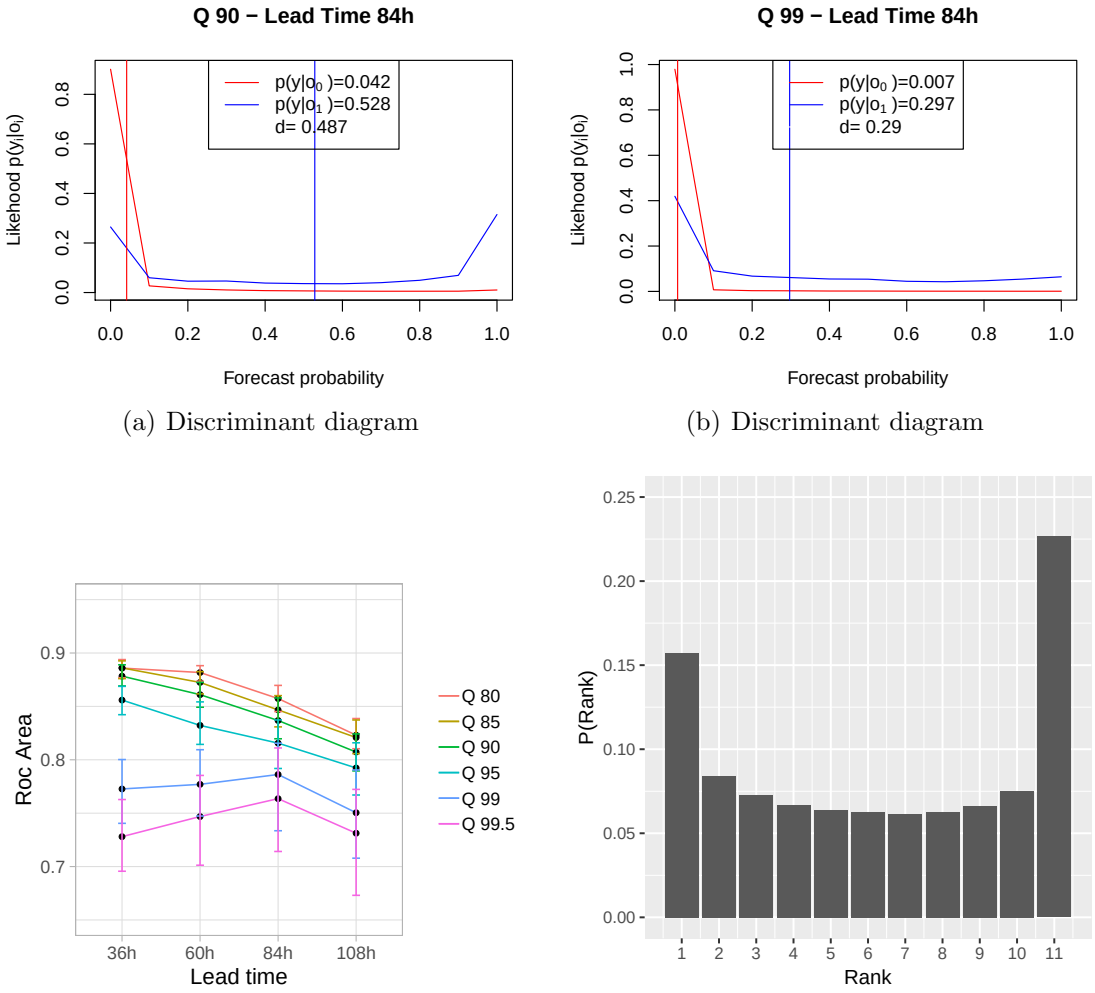
Figure 3.16: Reliability diagrams for q_{80} (top), q_{90} (middle) and q_{95} (bottom), for 2-days (left) and 4-days (right) forecasts. Values above red points indicate the marginal distribution of the forecasts $p(y_i)$. Error bars are estimated using a bootstrapping sampling technique and covers the 90% interval.

is also observed southeastward of the Cévennes. A similar behaviour is observed in the deterministic scores (section 3.1). We analyze the decomposed CRPS reliability and potential terms, at 108-hour lead times (Fig. 3.15(e) and 3.15(f)). The CRPS potential term is maximum on the same area impacted by a large CRPS while the reliability term gets worse there, which was not the case on the domain averaged CRPS.

We analyze reliability diagrams for q_{80}, q_{90} and q_{95} quantile thresholds and for 60-hour and 108-hour lead times (Fig. 3.16). The reliability gets worse at long lead times and for higher thresholds. If we focus on the position of the red line in the graph, we can observe it follows a nearly flatten line above the no skill area lower limit. It may indicate a conditional bias that could be related with the magnitude of the forecast biases. This implies that the resolution of the ensemble is poor. Moreover, most of the points are located below the diagonal (especially for q_{95} threshold). This means that the relative frequencies are small compared to the forecast probability, which indicates a wet bias of the probabilistic forecast. This is particularly true for high forecast probabilities except the highest one. Nevertheless, despite this bias, forecasts still have skill, even for high threshold and long lead times (except q_{95} at 108-hour).

The numerical values expressed above the points on the reliability diagram of Fig. 3.16 denote the frequency $p(y)$ of each forecast probability categories. Reforecast appears to be sharp because maxima of $p(y)$ are reached for low and high forecast probabilities. This indicates that the reforecast exhibits a good forecast confidence.

Discriminant diagrams at 84-hour lead times and for q_{90} and q_{99} thresholds are shown in Fig. 3.17(a) and 3.17(b). We note that the mean of the distribution $p(y|o_0)$ (vertical line) is concentrated around zero as these thresholds are associated with rare events. Conversely, when the event occurs, it appears more challenging for the forecast to issue a probability forecast close to 1. This effect is enhanced for q_{99} threshold; the forecast probability corresponding to the occurrence of the event is 0.297, meaning that a rare event is more difficult to discriminate from other events. As a result the discrimination distance diminishes for the highest thresholds. We also observe that this score decreases with lead times (not shown).



(a) Discriminant diagram

(b) Discriminant diagram

(c) Area under the ROC curve

(d) Rank histogram

Figure 3.17: Top: discrimination diagrams for 84-hour lead-time, for q_{90} (left) and q_{99} (right) quantile thresholds. Bottom: AUC computed for the same quantile thresholds as in Fig. 3.14 (left) and rank diagram (right) drawn for 84-hour lead time.

Another information about the forecasts conditioned by the observation can be drawn from the ROC diagram. We analyze the area under the ROC curve (AUC) (Fig. 3.17(c)). The AUC decreases for the larger lead times, except for the most extreme thresholds. The event discrimination is greater for the lowest thresholds. However, it is worth noting that, even for the $q_{99.5}$ threshold, discrimination is still better than random forecast (AUC>0.5).

Finally, reliability is examined also by means of the rank histogram tool (Fig. 3.17(d), 84-hour lead time), which shows a typical U-shape. Such a graph, which suggests an under-dispersed ensemble, corroborates the spread value observed in Fig. 3.12. As the 11-th rank is the most frequent, the ensemble reforecast has also a global tendency to underestimate rainfall (dry bias).

3.3 Summary and Conclusions

In this chapter the evaluation of the 24-hour precipitation forecast from the reforecast dataset has been presented. The deterministic verification member-by-member has shown that systematic errors and scores are different from a member to another. The main factor of these discrepancies is the deep convection parametrization, which strongly affects spatial distribution and intensity of the precipitation forecast. The PCMT scheme produces more intense precipitation than B85 scheme, which should be a positive compensation to the global underestimation, but it results in larger values of RMSE on the Cévennes and Ardèche chains, compared to B85. On the other hand, the most extreme events are better forecast by PCMT members as confirmed by the SEDI scores.

Combining all physical parametrizations, the deterministic evaluation shows that the members of the ensemble reforecast are not exchangeable. One member with a specific parametrization set-up exhibits errors that differ from another member differently implemented.

The second part of this chapter has been dedicated to the probabilistic verification of the ensemble reforecast. The comparison of the ensemble reforecast with the reduced size ensemble (10-M PEARP) obtained from the PEARP-2016 has revealed

a lack of dispersion of the reforecast, caused by the absence of the initial condition perturbations.

As it is common in forecast verification, scores get worse with increasing lead time and for increasing thresholds. The downgraded characteristics of the reforecast do not make it possible to produce high performing probabilistic forecasts. Reliability diagrams showed a quite low reliability and a wet bias. Indeed, the lack of initial condition perturbation and the small number of members are crucial in these deficiencies. However, it is remarkable that despite these factors, it is still possible to observe skill in forecasting very high rainfall thresholds over the mountains, when the q_{99} can reach values over 100 mm. We assume the sample size of the reforecast dataset and its relatively high horizontal resolution ($\approx$ 10 km) can be sources of positive skill.

Chapter 4

Postprocessing of 24-hour Ensemble Precipitation Forecasts

Contents

The reforecast has been evaluated in terms of probabilistic forecast verification. Though this assessment shows that the reforecast could not be considered as a potentially fully usable ensemble forecasting model, we also conclude that the reforecast has actual abilities to represent with some accuracy portions of the rainfall distributions, in particular for intense rainfall. In section 1.3.3, some univariate post-processing methods used for the calibration of precipitation have been described. Most of the post-processing methods described in chapter 1 are conceived for low precipitation values. In this chapter, two methods are tested on the reforecast with the aim of post-processing daily ensemble precipitation forecast: Quantile Mapping (QM) and Extended Logistic Regression (XLR). The second method is also applied to the operational ensemble system PEARP to see if the reforecast used as a learning dataset can lead to successful post-processing for PEARP precipitation forecasts.

Calibration methodologies are fitted through cross-validation which is basically testing the post-processing method over data that were not used for learning. Practically, one given year is tested with the calibration trained over all the other years of the period; for example, 1981 forecasts are calibrated using 1982-2010 as training dataset. Periods of one year have been selected because we suppose training dataset is still large enough ($\approx$ 97% of the whole dataset), and the test dataset can be considered as independent.

Deterministic calibration methods are designed to correct one given forecast. When applied to an ensemble forecast the procedure is performed individually on each member. On the other hand, ensemble calibration methods can operate with each single member as predictor, or some statistics computed from the ensemble forecast to model the predictive ensemble distributions. In this study, QM is applied on each individual ensemble member, as it is a deterministic technique. XLR method is tested using an ensemble statistics as predictor.

4.1 Quantile mapping

4.1.1 Method description

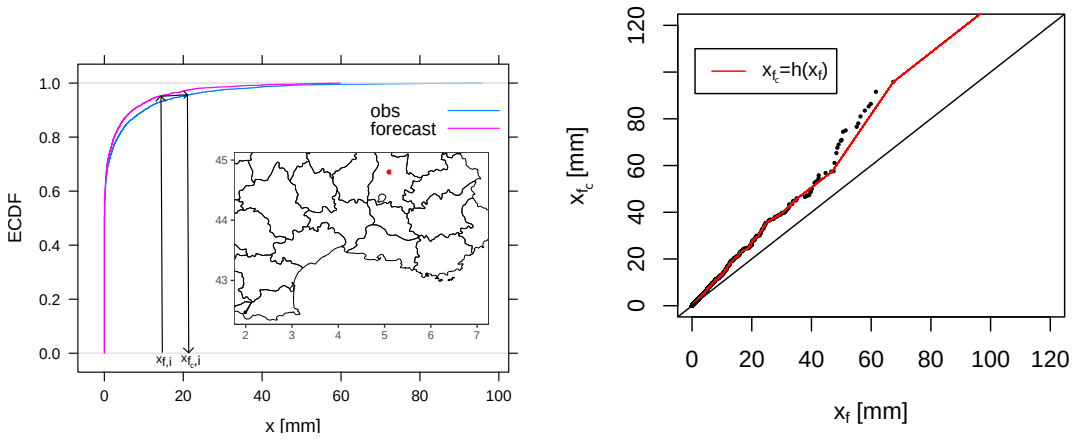
The QM method (Gudmundsson et al., 2012; Hamill et al., 2017; Hamill and Scheuerer, 2018; Hopson and Webster, 2010b; Piani et al., 2010a,b; Voisin et al., 2010; Wood et al., 2002) basically relies on the comparison analysis between the CDF (Cumulative Density Function) of raw forecast and the CDF of the observations. Accordingly to the notation previously introduced in Chapter 3, we denote x_o and x_f the observations and the forecasts respectively. x_{f_c} is the corrected forecasts at a given grid-point and for a given date. QM transformation is then defined as follows:

$$x_{f_c} = F_o^{-1}[F_f(x_f)], \tag{4.1}$$

where F_f is the CDF of x_f and F_o is the CDF of x_o. Hereafter we denote the transformation function $F_o^{-1}[F_f(x)]$ as $h(x)$. Practically, in the postprocessing procedure, x_{f_c} is the calibrated forecast and x_f the raw forecast from the test dataset, while the statistical transformation h is estimated from the training dataset.

The function h can be differently modeled depending on the approximation used for its estimation. In the current study, a nonparametric ECDF (Empirical Cumulative Distribution Function) of observed and forecasted values is used, like in Wood et al. (2002), or Hamill et al. (2017). The QM method is separately applied to each member of the ensemble reforecast and to each grid point. Lead times are pooled together in order to dispose of a larger sample for the estimation of the ECDFs. Therefore, the sample size reaches 3660 cases for each grid-point. Then, the ECDF is linearly interpolated on some selected point in order to obtain the transformation function h. The interpolation is computed on the selected points corresponding to the quantiles q_k, where $k \in \{1, 2, ..., 99, 100\}$.

If some forecasted value in the test dataset exceed the maximum value of the training dataset, a simple extrapolation is used, following Boé et al. (2007): outside the range of the correction function, a constant correction is applied. This means that the slope of the funtion h remains constant outside the given range. This basic



(a) ECDFs (b) Quantile mapping function

Figure 4.1: a) ECDFs of observations and forecast (control member) for 24-hour rainfall estimated above a grid-point for the 30-year period. The red point indicate the grid-point used for the estimation of the ECDFs. b) Transformation function $h(x)$ (red line) computed for the same grid-point as (a). Black points correspond to the values of the ECDFs drawn from the training dataset.

correction is applied in order to prevent the h function from a large slope value in the extrapolation that would lead to unusual remapped values. As the training dataset is 29-times larger than the test dataset, the estimation of the transformed values through extrapolation rarely occurs.

An example of the implementation of QM on the reforecast for one selected grid-point, for one member (the control member) and for all lead times is shown in Fig. 4.1(a). In this example we can see that rainfall forecast CDF is below the obserevd one. The black arrow describes the correction, which is the forecasted value remapped on the ECDF of the observations. The two ECDF curves show for this grid-point that dry days are more frequent than wet days, which is often the case over that region during this season. The corresponding percentage is equal to 51% for the observations and 58% for the forecast. In Fig. 4.1(b) the transformation function h for the same point as in Fig. 4.1(a) is shown. The red line represents the stepwise function $h(x)$. In this specific case, as we presented before, values above 63 mm, which is the maximum observed value, are remapped by adding a constant increment. Moreover, it can be observed that, despite the large size of the sample, the data points corresponding to the largest quantiles tend to be more

Member		00	01	02	03	04	05	06	07	08	09
BIAS	Raw reforecast	-0.20	-0.27	-0.25	-0.14	-0.19	-0.02	0.30	0.19	0.12	-0.21
	Corrected	0.01	0.01	0.01	0.02	0.02	0.02	0.03	0.02	0.03	0.01
MAE	Raw reforecast	2.77	2.76	2.85	2.86	2.78	2.86	3.02	2.89	2.96	2.78
	Corrected	2.83	2.83	2.93	2.90	2.84	2.86	2.86	2.79	2.89	2.84

Table 4.1: Bias and MAE before and after correction by means of QM. Grey columns refer to members implementing PCMT deep convection parametrization scheme. The remaining ones implement B85 scheme.

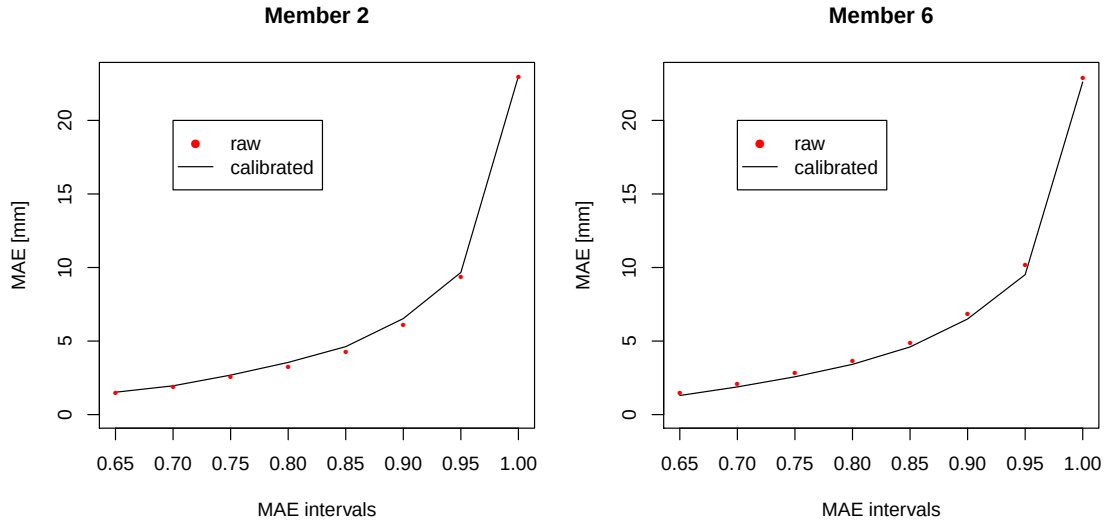
spaced, making the estimation of the transformation function less accurate.

4.1.2 Quantile mapping applied to the ensemble reforecast dataset

The QM method is first separately applied to each member, as in a deterministic calibration approach. The bias and Mean Absolute Error (MAE, as defined in Chapter 3) are computed before and after correction for each member considering all the grid-points together. These scores are presented in Table 4.1. The bias is estimated for one rainfall field by subtracting the spatial and temporal average of observations to the same average of the forecasts. We can see that B85 deep convection parametrization (members 0, 1, 2, 3, 4, 5, and 9) is characterized by a dry (negative) bias, and PCMT (members 6, 7, and 8) by a wet (positive) bias. The QM method properly corrects these biases, whose magnitudes are reduced to much lower values for all the members. A similar result has been found by Hamill et al. (2017) applying the QM procedure on global deterministic and 20-members ensemble 12-hour precipitation forecasts.

Table 4.1 also indicates that, for the raw reforecast, MAE is larger for members 6, 7 and 8. The QM correction performs uniformly well in correcting the biases, but the MAE is not homogeneous after the correction. Indeed, members whose raw forecasts are characterized by dry biases experience an increase of MAE after calibration. Conversely, wet biased members calibration leads to a reduction of MAE.

The MAE scores included in Table 4.1 take into account all dates from the



(a) (b)

Figure 4.2: MAE scores before (red points) and after (solid line) correction of QM. MAE is computed for different 24-hour observed rainfall quantile intervals. Results are shown for member 2 (left) and 6 (right). Member numbers refer to table 2.2.

reforecast period. Additional information can be obtained by discriminating the MAE for different classes of observed daily rainfall. Daily rainfall amounts are ranked and grouped between quantile intervals, with 5% steps. For example, the MAE interval for the quantile q_{95} stands for daily rainfall amounts ranging between q_{90} and q_{95}. The analysis for members 2 and 6 is shown in Fig. 4.2. These members are selected as they show the highest negative and positive variation of MAE before and after the QM correction. The greater the daily rainfall quantiles, the larger the absolute errors. Largest MAE variations after calibration are observed in the middle of the chosen quantile range, around q_{80} and q_{85} for both members.

The QM correction seems to apply differently on local bias signs. We analyze two points characterized by opposite biases in order to study this behaviour of QM correction. Point A, located in Ardèche, is evaluated for member 2 (Fig. 4.3(a)). It has a negative bias equal to -1.2 mm. Point B, located in Languedoc-Roussillon, refers to member 6 (Fig. 4.3(b)). It shows a positive bias equal to +1.7 mm. These two points and two members have been selected in order to analyze the effect of QM correction on forecasts characterized by different signed biases. The MAE before correction are 2.7 mm and 6.5 mm, respectively. After QM calibration they become

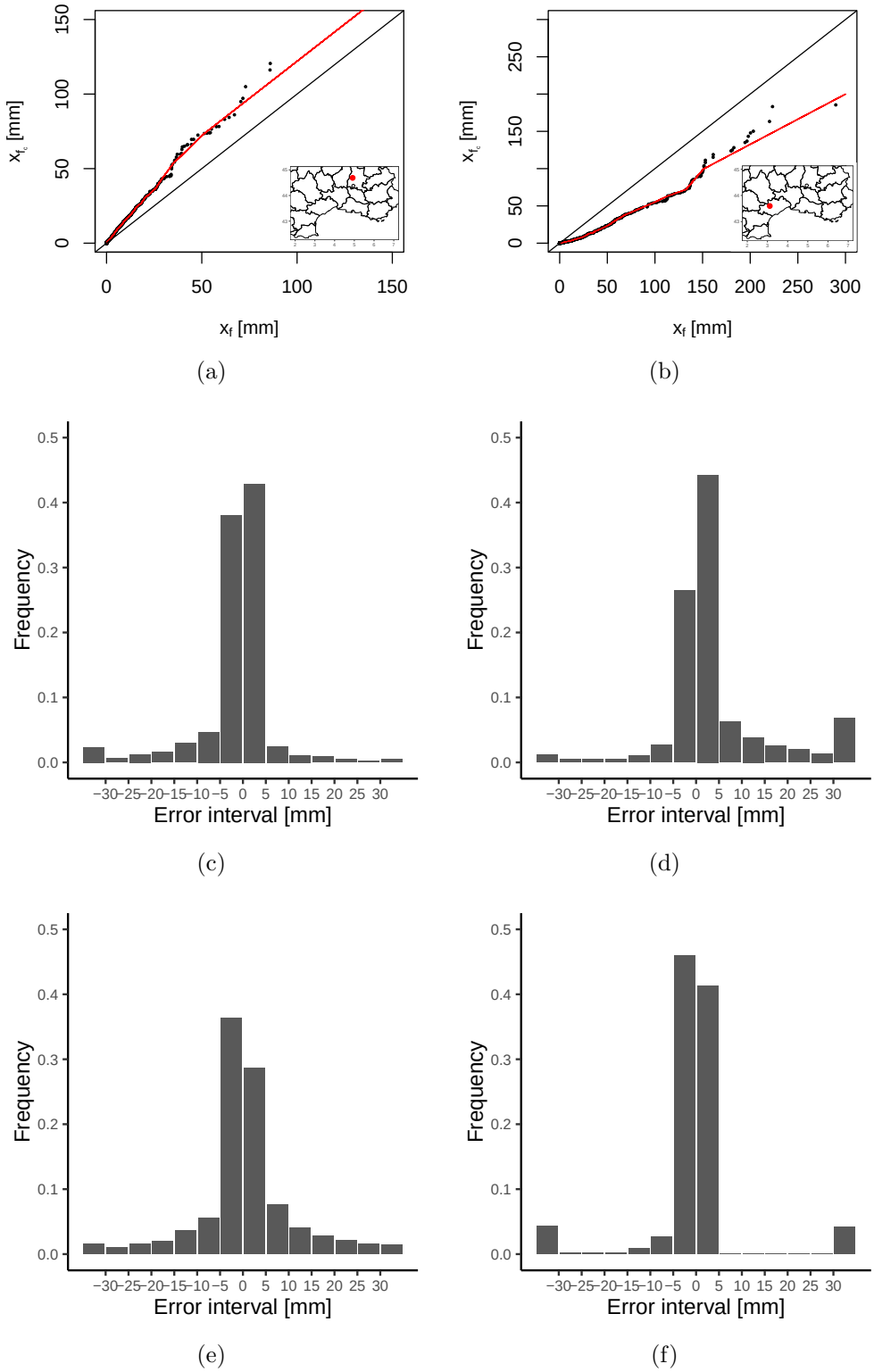


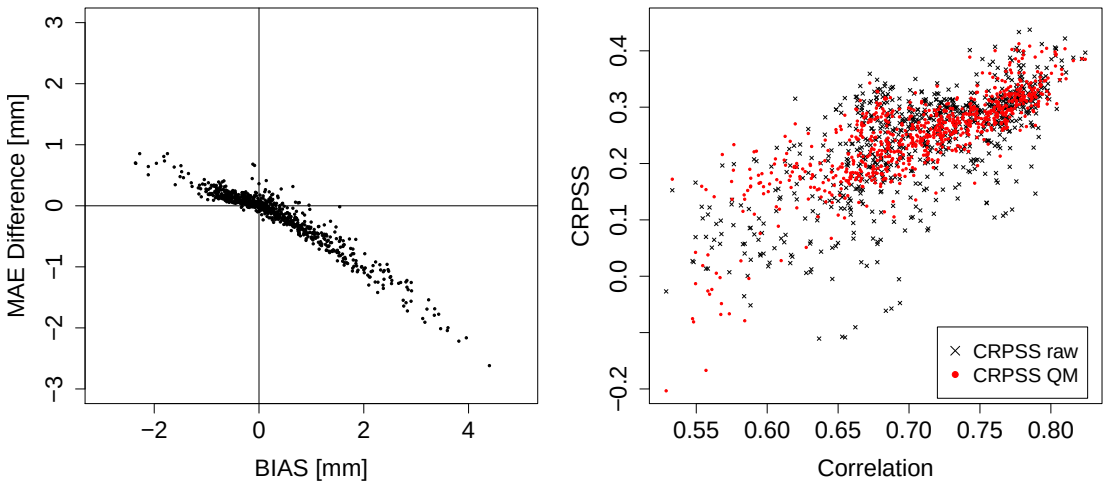
Figure 4.3: Application of QM on two selected points. Results refer to member 2 (left column) and to member 6 (right column) and are designed as point A and point B. The panel is composed by the transformation functions (top line), the histogram of errors before (center line) and after the QM correction (bottom line).

3.1 mm and 3.9 mm, respectively. This implies that, QM correction is worsening the MAE in the first case and improves the MAE in the second one. The two transformation functions h are shown in Fig. 4.3(a) and 4.3(b). The biases are characterized by departure between the stepwise function $h(x)$ and the bisector. This departure is growing as a function of the rainfall amount. The distribution of the errors before the QM correction is negatively skewed for point A and positively for point B (Fig. 4.3(c) and 4.3(d)). Most of the errors are centered around zero, meaning that small errors prevail. A high occurrence of strong positive errors ($\geq$ 30 mm) for member 2 in point B is observed. Actually, beyond this value positive errors spreads towards larger values (not shown). The effects of QM correction are presented in Fig. 4.3(e) and 4.3(f). For point A the distribution of errors after QM is flatter and slightly positively skewed. For point B most of the positive errors associated with the raw forecast are reduced.

The results presented in Table 4.1 and Fig. 4.3 suggest that the performance of the QM method applied to the estimation of QPFs is related to the initial bias. This hypothesis is further explored and extended to all the grid-points by representing the MAE difference before and after the application of QM as a function of the bias before calibration. The corresponding scatter plot for member 6 is shown in Fig. 4.4(a). The QPF is improved for large positive biases and gets worse for negative biases, showing the existence of a relationship between MAE variation and the original bias.

One possible factor for this behaviour is related to the remapping procedure. We look at the distribution functions for points A and B, whose biases are negative and positive respectively (Fig. 4.3(a) and 4.3(b)). We can see that for point A, raw values ranging within a given interval becomes more scattered after correction, while the opposite effect is observed for point B. This effect can influence the MAE after correction, because more dispersed forecast values after correction may induce larger errors.

Some studies have emphasized other effects of QM correction. Zhao et al. (2017) and Hamill and Whitaker (2006) have shown that the ensemble forecating skill can be impacted by the application of QM, with regard to the relationship between



(a) (b)

Figure 4.4: a) Difference between the MAE computed after and before the QM procedure against raw bias computed for each grid-point. b) CRPSS for each grid-point for the raw ensemble reforecast (black), and the calibrated one (red). Results are computed pooling all the lead times together.

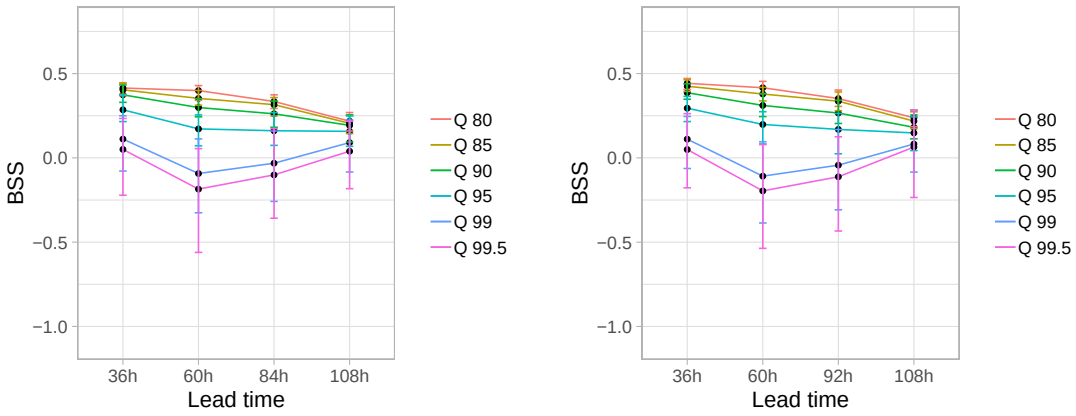
raw forecasts and the observations. The stronger this relationship, the more the QM correction improves the forecast skill. This hypothesis is tested computing an ensemble diagnostics, the Spearman's rank correlation between the daily ensemble mean of raw reforecast members and the observation at a given grid-point. The rank correlation assesses how well the relationship between two variables can be described using a monotonic function (whether linear or not). Then we use the CRPSS score to evaluate the reforecast as an ensemble forecasting system. The CRPSS corresponds to the skill score of the CRPS. It is evaluated by comparing the CRPS computed from a given forecast dataset to the CRPS computed from a given reference. CRPSS is obtained by replacing in eq. 3.11 the BS by the $CRPS$ and the BS_{ref} by the $CRPS_{ref}$. $CRPS_{ref}$ term is computed as the CRPS obtained using a number of members equal to ensemble size, which are sampled from the 30-year rainfall climatology reference.

In Fig. 4.4(b) the CRPSS values are shown for the raw (black) and the corrected (red) ensemble reforecast against the Spearman's correlation, all lead times pooled together. We can observe that ensemble forecasts are more skillful than the climatology (CRPSS>0) for most of the points. Globally, the CRPSS is larger when a

forecast-observation relationship leads to high correlation. If we compare the two CRPSS for correlation above 0.6, we can see that the number of low CRPSS values reduces, meaning that the skill is improved after calibration. On the other hand, correlation values below 0.6 lead to a decrease of the CRPSS values after calibration. This result reveals some limitations related to the QM procedure. In particular, the improvement of skill depends on the quality of the raw forecast. In that sense, we could expect that a forecast at long lead times is generally less correlated than at short lead times, resulting in a less efficient correction.

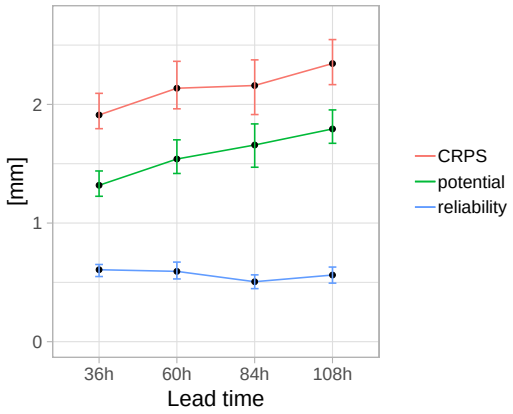
We apply some probabilistic score, in order to evaluate the effect of the reduction of the bias of each member on the probabilistic skill of the ensemble reforecast for high thresholds. This analysis is performed for each lead time. First we analyze the impact on the reliability looking at the so-called reliability term of the Brier Score (not shown). A slight improvement of reliability is observed for the q_{80} and the q_{85} quantile thresholds, but not for the highest values (not shown). Similarly, the resolution term gets better for thresholds values between q_{80} and q_{90} (not shown). These adjustments have a low impact on the modification of the BSS after calibration (Fig. 4.5(a) for the raw reforecast and Fig. 4.5(b) for the calibrated one). More specifically, no improvements are achieved for the most extreme quantile thresholds. No significant improvements are observed with the CRPS neither (Fig. 4.5(c) (raw) and Fig. 4.5(d) (calibrated)). This result suggests that the negative and positive modifications of CRPSS presented in Fig. 4.4(b) tend to compensate each other, resulting in a negligible variation of the CRPS computed across all grid points.

Scores are also examined point-by-point, in order to evaluate the effect of calibration depending on the location. Spatial BSS for q_{95} at 60-hour lead time is computed for the reforecast corrected with QM (Fig. 4.6(b)). Figure 4.6(a) shows the same score, but for the BSS computed from the raw reforecast. The no skill area for the raw reforecast in the Languedoc-Roussillon region exhibits larger BSS values after calibration. On the Cévennes chain no significant improvements are observed. On the opposite, scores get worse on the areas circled in red that correspond to high mountainous areas. The bias of the raw reforecast in these areas is negative (Fig. 4.6(c)). This negative bias is associated with a low Spearman's correlation, as shown

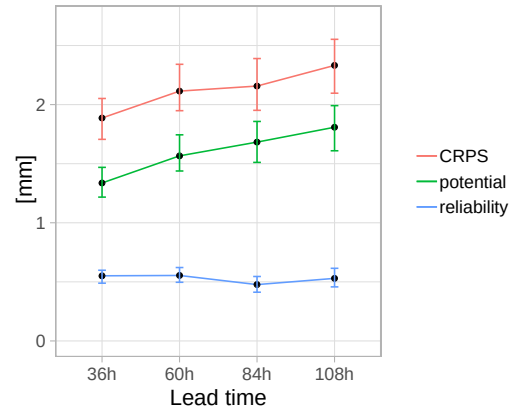


(a) BSS (raw)

(b) BSS (QM)

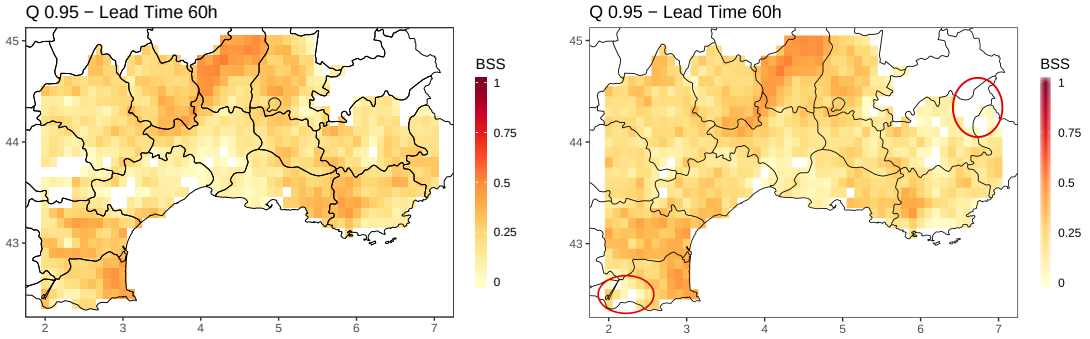


(c) CRPS (raw)



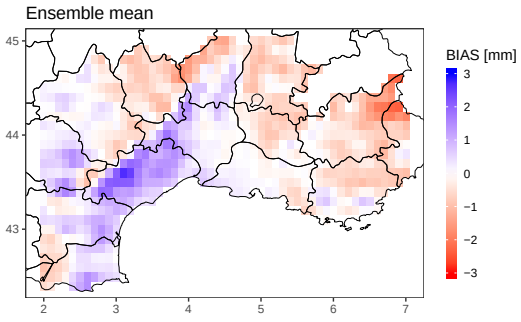
(d) CRPS (QM)

Figure 4.5: BSS (top) and CRPS (bottom) computed on the reforecast dataset for the 24-hour rainfall ensemble forecast using different quantile thresholds. Results are computed from the raw ensemble reforecast (left) and from the calibrated one using the QM method (right).

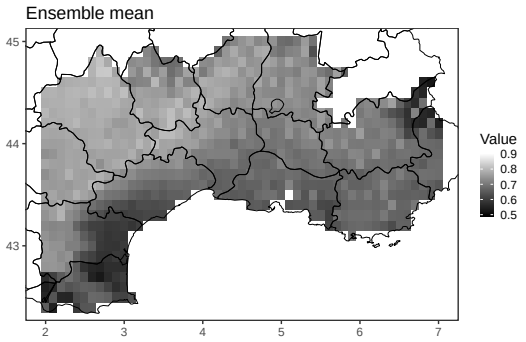


(a) BSS (raw)

(b) BSS (calibrated)



(c) Bias



(d) Rank correlation

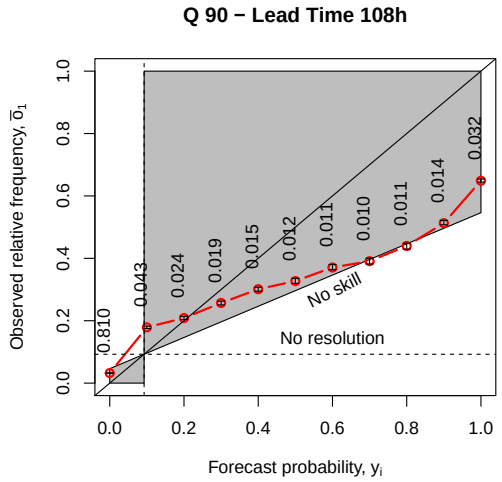
Figure 4.6: a) BSS of 24-hour precipitation computed on each grid-point for q_{95} at leads 60-hour for the raw reforecast. b) BSS of 24-hour precipitation computed on each grid-point for q_{95} at leads 60-hour for the calibrated reforecast. c) Bias of 24-hour precipitation computed on each grid-point using the daily ensemble mean of raw reforecast. d) Rank correlation of 24-hour precipitation computed on each grid-point using the daily ensemble mean of raw reforecast.

in Fig. 4.6(d). Then the poor correlation combined with a negative bias should produce an overcorrection. In this sense, the same effect observed member-by-member is observed considering the ensemble forecast.

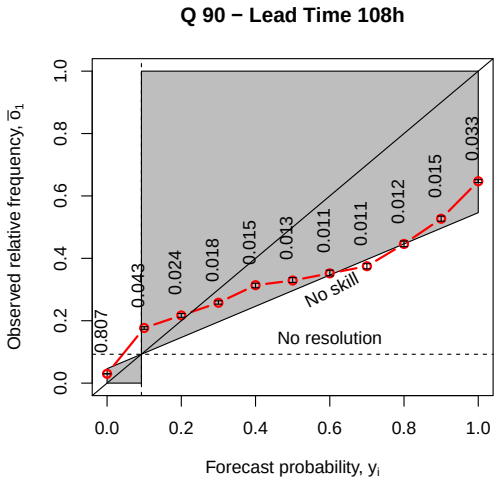
Reliability diagrams are computed at the longest available lead time (4-days) for q_{90} and shown in Fig. 4.7(a) (raw reforecast) and 4.7(b) (calibrated reforecast). We observe that reliability for q_{90} quantile remains unchanged after calibration. This result corroborates the conclusion that QM procedure has a limited impact on the reliability of the ensemble reforecast. The marginal distribution of the forecast, represented by the values above the diagram points, is not impacted by the QM application.

The conditional distribution $p(y_i|o_j)$ is assessed, and small modifications are observed for the QM corrected reforecast. Figure 4.7(c) and 4.7(d) show the discrimination diagrams at 84-hour lead time for the raw and for the corrected reforecast, respectively. The mean forecast probability conditioned by the non occurrence of the event remains unmodified, whereas the mean forecast probability conditioned by the occurrence of the event is larger for the corrected forecast. Consequently, the discrimination distance is increased showing that the calibrated reforecast better discriminates than the raw reforecast. A similar result is achieved by observing the AUC values (4.7(f)), compared to the raw reforecast (Fig. 4.7(e)). The improvement is greater for low threshold, even though a small AUC increment is detected also for the largest thresholds. The rank histogram is still U-shaped after calibration (not shown). This should indicate that forecasts could be underdispersed, which suggests that QM procedure has a negligible effect on the spread of the reforecast.

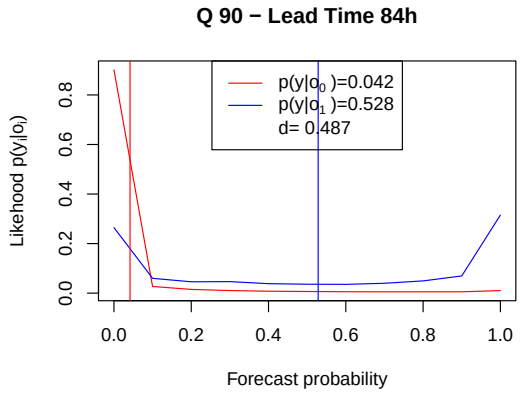
It is interesting to investigate if a method that is supposed to take into account ensemble statistics can produce more skillful probabilistic forecasts than the QM procedure. This question is addressed in the next section.



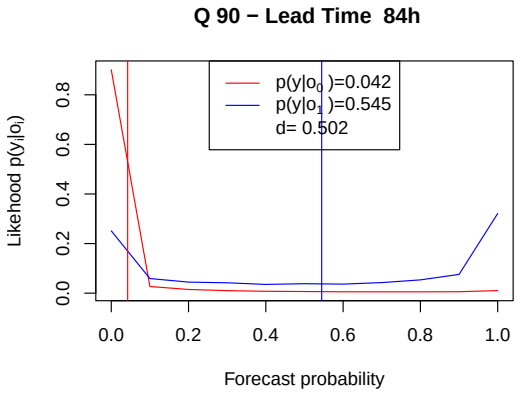
(a) Reliability diagram (raw)



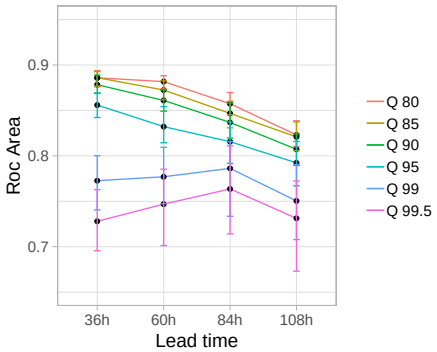
(b) Reliability diagram (QM)



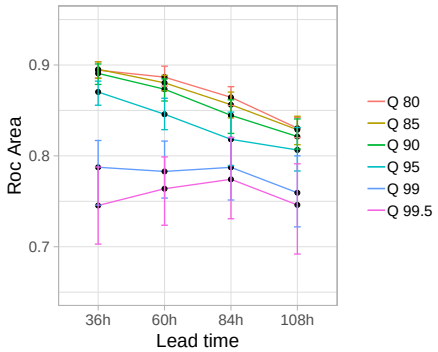
(c) Discrimination diagram (raw)



(d) Discrimination diagram (QM)



(e) AUC (raw)



(f) AUC (QM)

Figure 4.7: Top: reliability diagrams for q_{90} for 4-days forecasts are drawn. Middle: discrimination diagrams for 84-hour lead time, for q_{90} quantile thresholds. Bottom: AUC diagrams using different quantile thresholds. All graphs are drawn using the raw (left) and the QM corrected reforecast (right).

4.2 Logistic regression

4.2.1 Method description

Logistic regression is a nonlinear statistical model designed to estimate the probability of a definite binary event. For precipitation, the event is defined with respect to a defined precipitation threshold. If R represent the rainfall variable, the binary event assumes different values on the basis of the following relation $R \leq q_j$, given the quantile threshold q_j. This probability is a function of a set of predictor variables $\mathbf{x}_f$, that can be represented as a conditional probability $p(R \leq q_j|\mathbf{x}_f)$. This latter quantity can be expressed through the logistic regression formulation:

$$p = \frac{\exp\left(f(\mathbf{x}_f)\right)}{1 + \exp\left(f(\mathbf{x}_f)\right)}, \tag{4.2}$$

where $f(\mathbf{x}_f)$ is a linear function of the predictors $\mathbf{x}_f$:

$$f(\mathbf{x}_f) = \beta_0 + \beta_1 x_{f,1} + \beta_2 x_{f,2} + ... + \beta_N x_{f,N} = \beta_0 + \sum_{i=1}^{N} \beta_i x_{f,i}. \tag{4.3}$$

The terms β_i correspond to the regression coefficients and β_0 is the intercept of the linear function. N is the number of predictor variables involved in the regression. The logistic regression becomes linear in a logarithmic scale:

$$\ln\left(\frac{p}{1-p}\right) = f(\mathbf{x}_f). \tag{4.4}$$

The term on the left-hand side of the equation is the logit function.

The parameters $\boldsymbol{\beta}$ are generally estimated using an iterative maximum likelihood procedure (Wilks, 2009b). This procedure is based on the log-likelihood function that in the logistic regression framework leads to the function:

$$L = \sum_{k=1}^{n} \log(\Gamma_k(\boldsymbol{\beta})), \tag{4.5}$$

which has to be maximized and where Γ_k is the predicted probability for the k^{th}

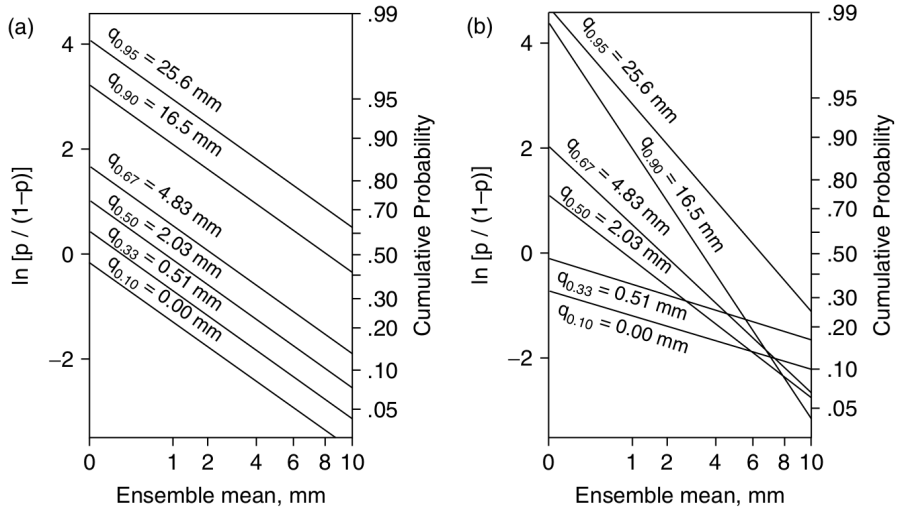


Figure 4.8: Graphical example of extended logistic regression (left panel) and logistic regressions (right panel) plotted on the log-odds scale. Regression lines estimated for different quantile thresholds using XLR are parallel (left panel), while if the standard logistic regression is implemented they cross, leading to inconsistent results (right panel). From Wilks (2009a).

event:

$$\Gamma_k = \begin{cases} p(R \leq q_j | \mathbf{x}_{f,k}) & x_{o,k} \leq q_j \\ 1 - p(R \leq q_j | \mathbf{x}_{f,k}) & x_{o,k} > q_j. \end{cases} \tag{4.6}$$

$x_{o,k}$ is the k^{th} observed value and $\mathbf{x}_{f,k}$ is the corresponding set of predictor values associated with the k^{th} event. In this case, one event corresponds to a given day at one given lead time and one grid point of the test dataset. n is the total number of events.

In the standard logistic regression, also called separate logistic regression, each modelled probability for a given quantile threshold $p(R \leq q_j | \mathbf{x}_f)$ has to be separately computed. Each regression, relative to a given quantile q_j, provides a different estimation of the parameters β. This separate estimation can lead to nonsense results like $p(R \leq q_1 | \mathbf{x}_f) > p(R \leq q_2 | \mathbf{x}_f)$ while $q_1 < q_2$. A graphical example of this undesirable effect can be visualized in the log-odds space considering one single predictor variable, where the vertical axis is expressed in a logarithmic scale given by $\ln\left(\frac{p}{1-p}\right)$ (the logit function) against the predictor (see right panel of Fig. 4.8). In this example the predictor is the ensemble mean. The β parameters are different for each logistic function. Therefore different values of the slopes, controlled by the β_1

parameter, result in a crossing of the regression lines leading to inconsistent results.

The extended logistic regression (XLR, suggested by Wilks, 2009a) is conceived to avoid this behaviour. The influence of the threshold q_j on the regression is taken into account by adding a function $g(q_j)$ as an additional predictor variable:

$$f(\mathbf{x}_f) = \beta_0 + \alpha g(q_j) + \sum_{i=1}^{N} \beta_i x_{f,i}, \tag{4.7}$$

where α is an additional coefficient and the transformation $g()$ is a monotonic function in order to preserve the consistency of the calibrated probabilities. Compared to the separate logistic regression, the advantage is that all the estimated parameters (α,$\boldsymbol{\beta}$) are the same for all the probability thresholds. This means that in the example of Fig. 4.8(a), referred to extended logistic regression, regression lines do not cross because the slope parameter β_1 is constant. Lines are spaced depending on $g(q_j)$, which added to the β_0 parameter determines the intercept value.

As for the separate logistic regression, the extended logistic regression is fitted using an iterative maximum likelihood procedure on the function $L(\alpha, \boldsymbol{\beta})$. This operation is performed by using a set of J threshold values q_j. The Γ_k function is then redefined for the k^{th} event

$$\Gamma_k = \begin{cases} p(R \leq q_1 | x_{f,k}) & x_{o,k} \leq q_1, \\ p(R \leq q_j | x_{f,k}) - p(R \leq q_{j-1} | x_{f,k}) & q_{j-1} < x_{o,k} \leq q_j, \\ 1 - p(R \leq q_J | x_k) & x_{o,k} > q_J. \end{cases} \tag{4.8}$$

The probability p corresponds to eq. 4.2, using the definition of $f(\mathbf{x}_f)$ of eq. 4.7.

The $g()$ function is also used to apply a transformation function to the precipitation threshold values q_j and then compensate its irregular distribution properties. In this study, we tested a few different functions and obtained best results with $g(q_j) = \alpha\sqrt{q_j}$. This square root transformation has been largely adopted in literature (Hamill et al., 2008; Messner et al., 2014, 2013; Roulin and Vannitsem, 2011; Scheuerer, 2014; Schmeits and Kok, 2010; Wilks, 2009a).

The set of predictors employed for the regression can be drawn from statistics of

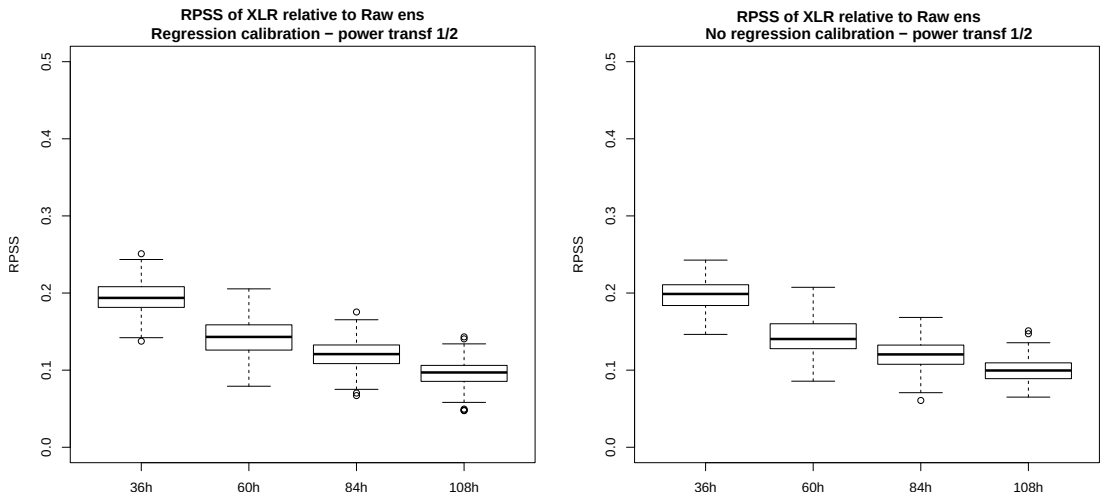
the ensemble forecast. Some studies suggest using a transformed value of the ensemble mean of precipitation forecasts (Roulin and Vannitsem, 2011; Schmeits and Kok, 2010; Wilks, 2009a). Messner et al. (2014) included, as additional predictors, the variance of the ensemble and Hamill (2012) included the product between the variance and the ensemble precipitation mean. In the current study, the ensemble mean is the only predictor we used. The other ensemble predictor statistics, notably the ensemble spread, are withdrawn because the reforecast ensemble and the PEARP set-ups are very different. The ensemble mean of square root transformed precipitation has been selected, since the corresponding corrected forecasts showed better scores compared to the ones generated using the non transformed mean. Schmeits and Kok (2010) proposed the same transformation. As a result of this experimental framework, equation 4.7 becomes:

$$f(x_f) = \beta_0 + \alpha(q_j^{1/2}) + \beta_1(\overline{x_f^{1/2}}). \tag{4.9}$$

In this study, the chosen precipitation thresholds are q_{80}, q_{85}, q_{90}, q_{95}, q_{99} and $q_{99.5}$. This choice corresponds to the same quantile used in the forecast verification (see section 3.2.2). The maximum log-likelihood is performed by means of a quasi-Newton method based on the Broyden–Fletcher–Goldfarb–Shanno algorithm (Fletcher, 2013). XLR method is performed for each grid point and lead time, whose total size is 915. The cross-validation technique is also adopted here, so that the calibration is carried out for a sliding year to year window over the 30-year period.

4.2.2 Extended logistic regression calibration applied to the ensemble reforecast dataset

The reduced number of members in the ensemble reforecast has an impact on the estimation of the ensemble mean. The possible misestimation of the ensemble mean might induce a bias in the estimation of the slope of the regression in the log-odds scale, an effect called *attenuation* (Carroll and Stefanski, 1990). In the case of an additive error in a simple linear model, this bias can be accounted by using the so-called reliability ratio (the ratio between the variance of the predictors and



(a) (b)

Figure 4.9: RPSS computed using the XLR method with (left) and without (right) regression calibration implementation.

the sum of this variance and the variance of the error). In the present study, we test this method, also called regression calibration. A formulation of this method is presented in Roulin and Vannitsem (2011) and it has also been used by Rosner et al. (1990, 1989). The procedure is based on replacing the misestimated predictor with a fitted linear approximation.

In order to evaluate the impact of the regression calibration, some experimental statistical tests are carried out, using synthetic ensembles with varying sizes, biases and spread. Results show that the ensemble mean can be better estimated using regression calibration instead of the raw ensemble mean, except when the raw ensemble shows a significant bias or an heterogeneous distribution.

However, we decide to test the application of the regression calibration method in conjunction with XLR calibration on the reforecast dataset. The ensemble mean of square root transformed precipitation is used as predictor. The Ranked Probability Score (RPS; Epstein, 1969a; Wilks, 2009b) is used as a metric for the evaluation of the results

$$RPS = \sum_{j=1}^{N} \left[P(R < q_j | \overline{x_f^{1/2}}) - I(R < q_j) \right]^2 , \tag{4.10}$$

where q_j are the quantile thresholds used in eq. 4.9 and $I(.)$ is equal to 1 if argument

in brackets is true and 0 if it is not. This score corresponds to the discrete definition of CRPS (eq. 3.13). Here a mean RPS, $\overline{RPS}$ is computed for a given lead time aggregating all the grid-points for all the days. A bootstrap technique is used in order to evaluate the uncertainty of the score. The skill scores (RPSS) is computed:

$$RPSS = 1 - \frac{\overline{RPS}}{\overline{RPS}_{ref}}, \tag{4.11}$$

where $\overline{RPS}_{ref}$ denotes here the score computed from the raw reforecast. Results are shown in Fig. 4.9. Skill scores are positive for all the lead times, showing a significant improvement given by the XLR method on the raw reforecast. Skills diminishes at longest lead times. XLR calibration applied in conjunction with regression calibration (Fig. 4.9(a)) exhibits lower skill than the same calibration without regression calibration implementation (see Fig. 4.9(b)). The results obtained with synthetic ensembles are confirmed with this test on a true case. For this reason, we implement the XLR procedure using the raw predictors, discarding the regression calibration implementation.

Hereafter a description of the set-up of the XLR calibration procedure is given. The approach proposed by Roulin and Vannitsem (2011) is followed. The XLR procedure applied to the reforecast provides continuous probability density function for the 24-hour precipitation for each day and each grid point. Calibrated members are then sampled from this distribution. The transition from continuous probability values to probabilities drawn from a discrete number of members of the ensemble requires the assignment of a certain number of probabilities to be sampled from the calibrated CDF. Since the reforecast is by construction composed by 10 members, the number of probabilities to be assigned to remap the new members is equal to 10. The preservation of the same number of calibrated members as in the raw reforecast is justified by two reasons: 1) this enables to dispose of a mutual correspondence between each raw and calibrated member to evaluate the effect of the correction member-by-member, 2) probabilistic scores may be not artificially improved since the ensemble size does not change after correction.

As in Katz and Ehrendorfer (2006); Roulin and Vannitsem (2011), and Roulston

and Smith (2002) the uncertainty of this sampling due to the relatively small number of members is taken into account by adding a fictitious member distributed equally between the occurrence of the event and the nonoccurrence, by assigning "half" of the fictitious ensemble to the frequency of occurrence of the event, the other half to nonoccurrence. This procedure eliminates the possibility of obtaining an estimated probability of zero or one, which is considered an undesirable property. Indeed, 0 or 1 probability values are theoretical values that could be achieved with an ensemble of infinite size. Therefore, the set of probability values are assigned as follows:

$$p(l) = p(R \leq x_f(l)) = \frac{l + 1/2}{m + 1}, \tag{4.12}$$

where R is the rainfall variable, $x_f(l)$ is the precipitation forecast issued by the l^{th} member and m is the number of members of the ensemble.

Once the probabilities $p(l)$ are determined, the value of the precipitation $x'_f(l)$ corresponding to the l^{th} ranked member, can be obtained by inverting the logistic function (eq. 4.2):

$$x'_f(l) = \left\{ \frac{\ln\left(\frac{1-p(l)}{p(l)}\right) - \beta_0 - \beta_1(\overline{x_f^{1/2}})}{\alpha} \right\}. \tag{4.13}$$

Figure 4.10(a) illustrates the remapping procedure, through the transformation of eq. 4.13. The arrow shows the correction of the raw sixth ranked member, associated to the assigned probability p(6), to the calibrated value $x'_f(6)$. Once the set of x'_f are computed for all the assigned probabilities, precipitation values are sorted using the same rank order as the one of the raw ensemble, in order to preserve the rank order of the members between the raw and the calibrated forecasts. In this manner, each raw ensemble member can be related to its corresponding calibrated one. The whole procedure is applied separately for each lead time and grid-point.

The practical processing of zero precipitation is detailed. The probability of

non-precipitation using the logistic regression formulation is given by

$$p_0 = p(R = 0|x_f) = \frac{\exp\left(\beta_0 + \beta_1(\overline{x_f^{1/2}})\right)}{1 + \exp\left(\beta_0 + \beta_1(\overline{x_f^{1/2}})\right)}. \tag{4.14}$$

Then, if the probability of a null precipitation with the raw ensemble is lower than p_0, each raw member whose assigned probability $p(l)$ is lower than p_0 is calibrated to zero (see Fig. 4.10(d), details are given below). Otherwise, if the probability of a null amount of precipitation in the raw ensemble is higher than p_0, only a fraction (corresponding to the probability $p(l) \leq p_0$) of the zero precipitation members, are randomly selected and set to zero. The remaining raw forecasts that predict zero precipitation are remapped (see Fig. 4.10(c), details are given below) to the non-zero values obtained from eq. 4.13.

The remapping procedure

The remapping procedure is illustrated in Fig. 4.10. Four example cases that enlighten the remapping procedure have been selected. As an example, Fig. 4.10(a) shows how a member is remapped onto the fitted logistic cumulative distribution computed for 108-hour lead time. This example makes reference to one of the most extreme dates (in terms of precipitation) of the reforecast. In this example, the forecasted value is transformed into a larger value than the raw forecast. This correction shifts all the members of the raw forecast to larger values.

Figure 4.10(b) shows another example associated with a 42 mm rainfall ensemble mean forecast. In this example, the lower half of ensemble members are weakly modified by the remapping procedure, while for the highest probabilities a large positive correction occurs. In this sense, the most intense raw members are corrected to large precipitation values, since the fitted CDF is tailed for high probabilities. This kind of correction leads to an increase of the dispersion of the calibrated ensemble.

Figure 4.10(c) illustrates a case of no precipitation forecast, for which the ensemble mean is zero. This means that the probability of zero precipitation in the forecast is higher than with the calibrated probability p_0. Only a fraction of members of the raw forecast are remapped to non-zero precipitation values. This case shows that for

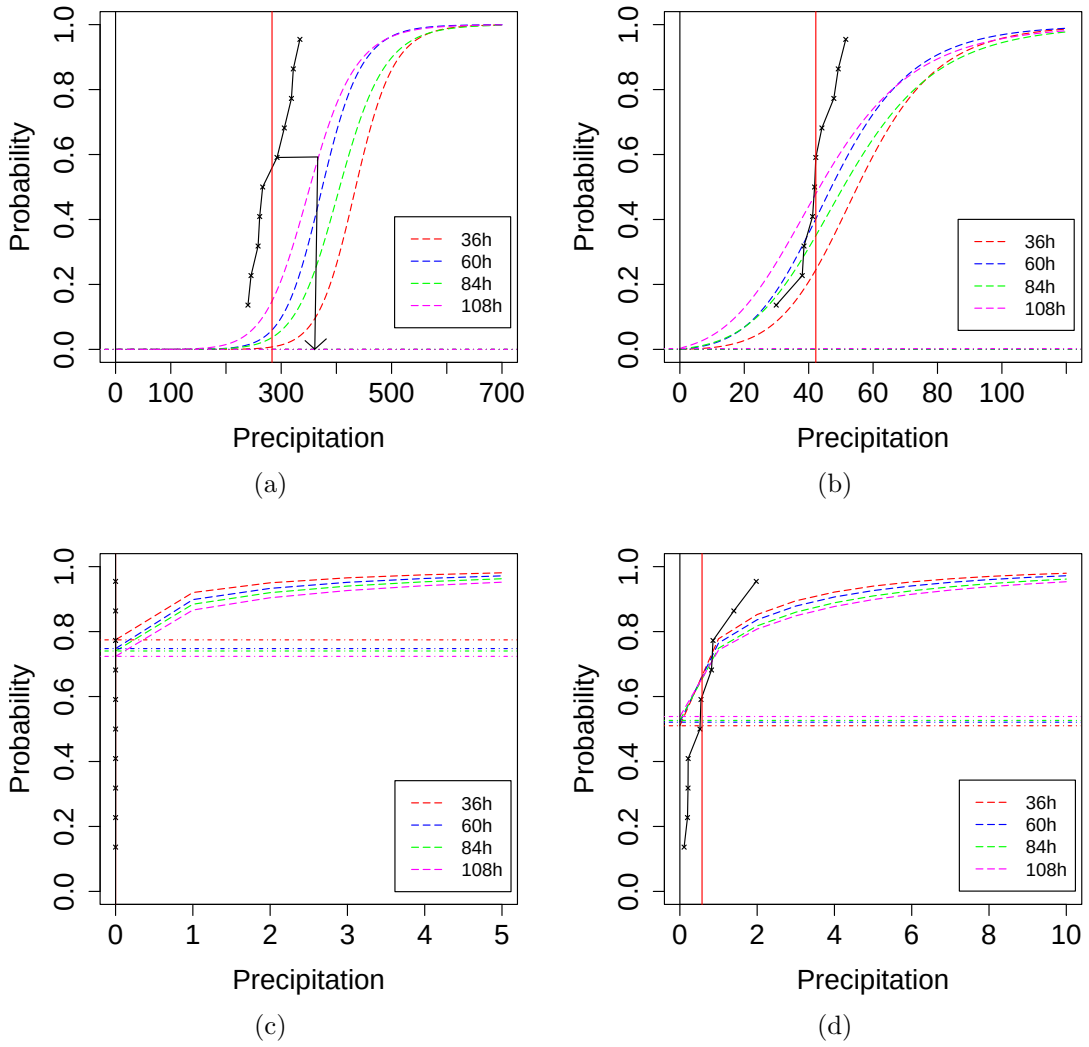


Figure 4.10: Remapping for daily precipitation (mm) of the members of four raw reforecast cases. The raw reforecast ensemble is presented as an ECDF in black solid line. The ensemble cumulative distributions fitted using the logistic regression, and drawn for the predictor issued by the raw reforecast, are represented by the dashed curves. Each coloured curve represents the cumulative density function drawn using the parameters estimated for different lead times. Horizontal dotted lines show the non-precipitation probability p_0 estimated for different lead times, using the same colour specification as for distribution curves. The vertical red line corresponds to the raw ensemble mean. Each panel is referred to different dates and grid-points, selected as examples cases. The discontinuity of the curves in (c) are due to the graphical discretization of the x-axis to 1 mm steps.

this specific grid-point, every raw ensemble forecasts with zero mean precipitation are remapped to non-zero mean precipitation. This correction can have an impact on the verification scores based on low threshold values or on the verification of rain/no rain events. The probability p_0 takes different values depending on the lead time. This probability is lower for the longest lead time than for the first ones.

The last example (Fig 4.10(d)) presents a situation where the probability for non-zero precipitation in the raw reforecast is lower than the probability p_0. In this case, the members associated with the assigned probabilities lying below p_0 are set to zero. The remaining ones are remapped to the fitted cumulative density function. It is interesting to note that for all the presented cases, and for every class of precipitation intensity, the most intense members of a given daily ensemble forecast are always remapped towards higher values.

Effect of the calibration on the spread

The precipitation values reassignment for the ensemble members can modify the ensemble spread. This hypothesis is addressed by computing the ensemble spread across all the grid-points and dates for each different lead time (Fig. 4.11(a)). We can see that the spread of the calibrated reforecast is inflated, compared to the raw reforecast. The increase of the spread with the lead time seems to reach a limit at 108-hour lead time. Though the calibrated reforecast is still underdispersed compared to the operational ensemble forecast (see Fig. 3.12). This spread modification is found to positively impact the skill of the calibrated reforecast.

Some example cases that shows the spread modification through XLR procedure are given. Figure 4.11(b) shows an example of the daily spread of the raw ensemble against the daily raw ensemble mean for a given grid-point at different lead times. Spread increases with the ensemble mean, but no distinction can be made between the lead times. Figure 4.11(c) shows how the spread is modified after the XLR calibration. First, the spread is shown to be directly dependent on the parameters estimated in the extended logistic regression. The spread increases monotonically against the ensemble mean precipitation. Since the regressions are fitted separately for each lead time, the corresponding spread-mean relationships are different. In this

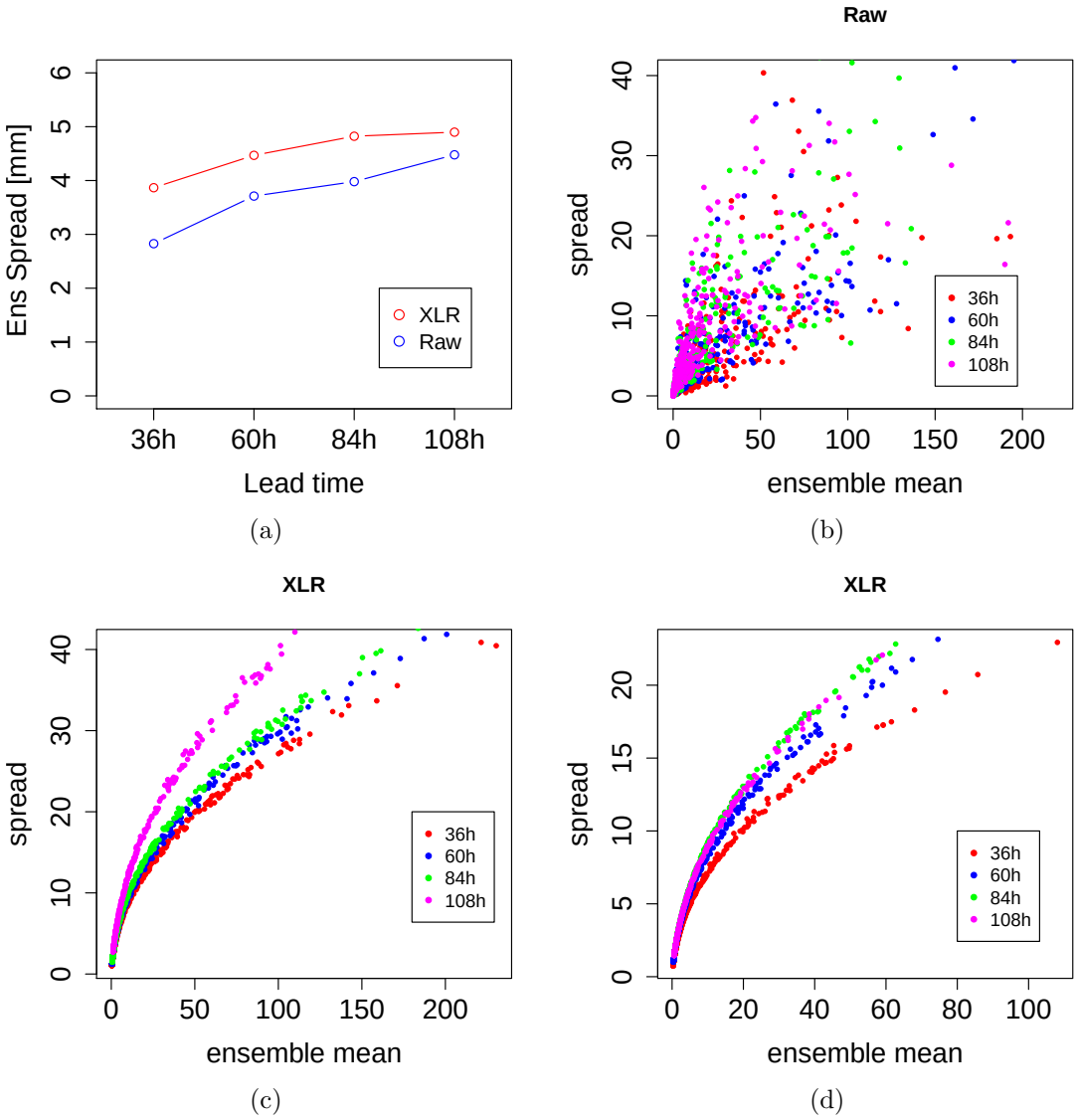
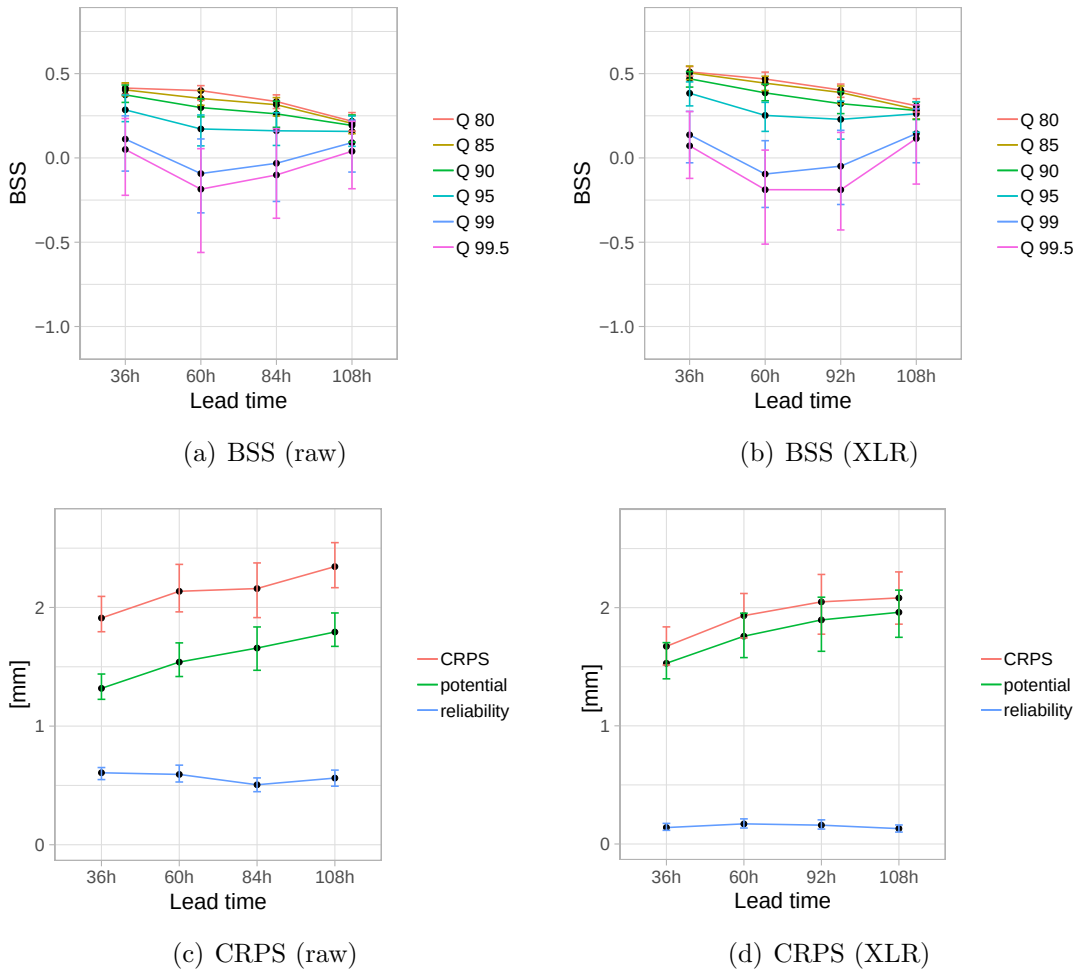


Figure 4.11: (a) Ensemble spread computed for all grid-points, before (blue line) and after (red line) calibration. (b) Ensemble spread (mm) from the raw reforecast as a function of ensemble mean (mm) for a given grid-point. Each point corresponds to a daily value, differently coloured depending on the lead time. (c) The same as (b), but after the application of XLR. (d) As in (c), but for a different grid-point.



(a) BSS (raw)

(b) BSS (XLR)

(c) CRPS (raw)

(d) CRPS (XLR)

Figure 4.12: BSS (top) and CRPS (bottom) computed on the reforecast dataset for the 24-hour rainfall ensemble forecast using different quantile thresholds. Results are computed from the raw ensemble reforecast (left) and from the calibrated one using the XLR method (right).

example, spread at 108-hour is significantly larger than the calibrated spread at short lead times. This behaviour is not always verified. For instance, Fig. 4.11(d) shows that for another grid-point, 84-hour lead time exhibits a larger calibrated spread than at 108-hour lead time. This shows that the parameters estimated through XLR method can significantly vary depending on the grid-point and lead time.

Probabilistic scores for the calibrated reforecast

In the following part, the assessment of the XLR calibrated ensemble reforecast is carried out through a comparison with the raw reforecast using probabilistic scores. The focus is set on the high precipitation thresholds.

Figures 4.12(a) and 4.12(b) show the BSS, computed for different quantile thresholds, for raw and calibrated reforecast. We can observe better scores at all lead times compared to the raw reforecast, except for q_{99} and $q_{99.5}$. As already observed for BS scores, this is related to the insufficient number of cases for these extreme events. The reliability and the resolution terms of Brier Score are both improved (not shown). The CRPS is slightly improved (Fig. 4.12(d)). More specifically, the reliability term gets better. As the reliability term of CRPS is related to the rank histogram as explained by Hersbach (2000), then the increase of spread may explain the reliability improvement.

All the probabilistic scores previously presented are computed using the 10-members calibrated forecast, as the raw reforecast was verified for the same ensemble-size. The number of members of an ensemble forecast affects the probabilistic scores. An experimental test is performed in order to evaluate the sensitivity of the CRPS score to the ensemble-size. Calibrated ensemble forecasts of varying size between 10 and 100 are sampled from the same fitted predictive cumulative functions. One could expect that the forecasted probability distribution is better sampled with a larger ensemble size. Figure 4.13 shows the CRPS score computed by aggregating all lead times and grid points for different calibrated reforecast ensemble sizes. The score gets better with the ensemble size and it tends to an asymptote for large sizes. However, relative variations are quite small ($\approx 1\%$). This result shows that the skill of ensemble forecasting system can be improved if calibrated with a large number of members, but weak improvements would be balanced by the computational and technical issues related to the construction of a very large-sized ensemble.

Spatial BSS and CRPS scores

The same scores, but computed point-by-points, are then presented. If we look at the spatial BSS maps computed for q_{95}, shown in Fig. 4.14(a) (raw) and 4.14(b) (XLR calibrated) at 108-hour lead time, we note that BSS is increased everywhere and grid-points with no skill do not appear.

This improvement is also observed for the lower thresholds (not shown). The spatial CRPS is also globally improved as we can see in the comparison of raw

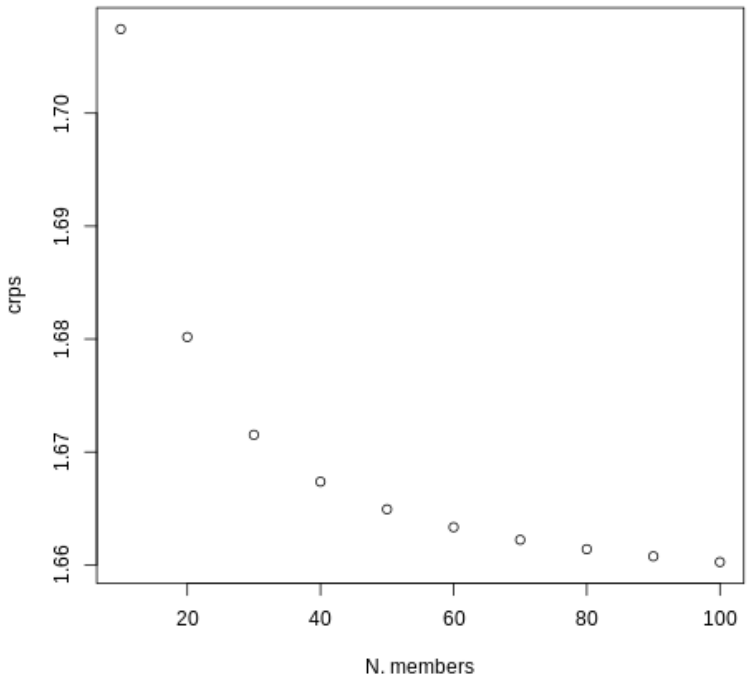
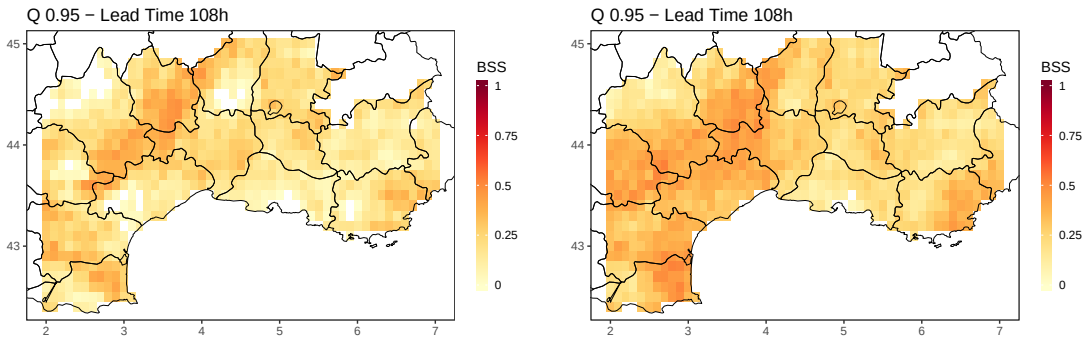


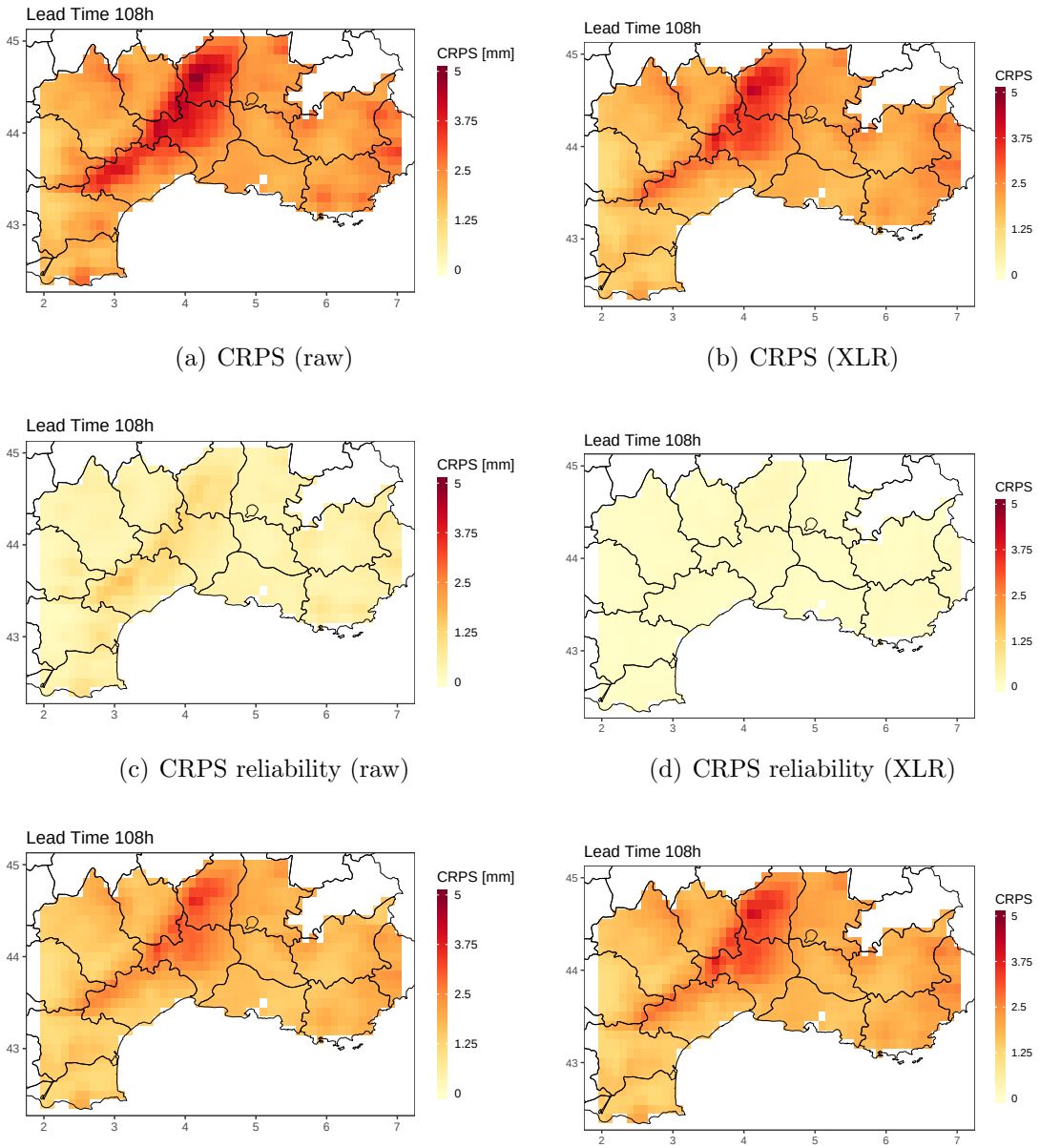
Figure 4.13: CRPS computed aggregating all lead times and grid-points using the calibrated ensemble adapted for different number of members.



(a) BSS (raw)

(b) BSS (XLR)

Figure 4.14: Top: BSS of 24-hour precipitation computed on each grid-point for q_{95} threshold at 108-hour lead time for the raw (a) and the XLR calibrated reforecast (b).



(a) CRPS (raw)

(b) CRPS (XLR)

(c) CRPS reliability (raw)

(d) CRPS reliability (XLR)

(e) CRPS potential (raw)

(f) CRPS potential (XLR)

Figure 4.15: Top: CRPS of 24-hour precipitation computed on each grid-point at 108-hour lead time. Middle: reliability term of the CRPS at 108-hour lead time. Bottom: resolution term of the CRPS at 108-hour lead time. Results refer to the raw (left) and the calibrated (right) ensemble.

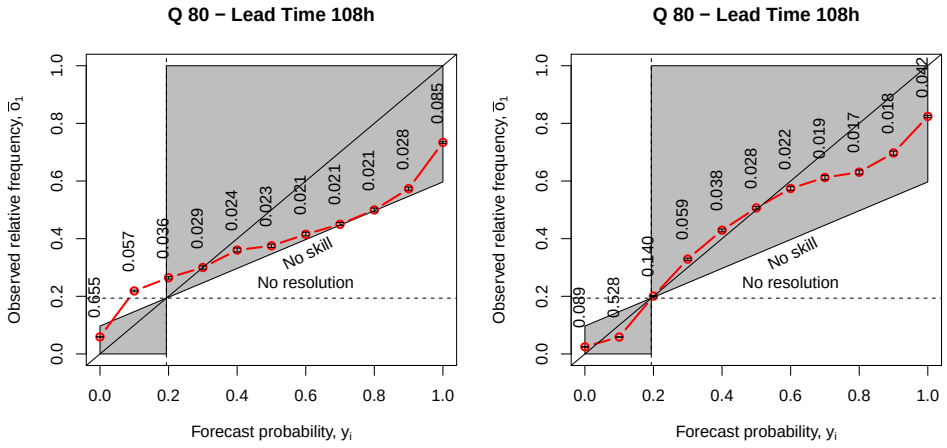
and calibrated reforecast (Fig. 4.15(a) and Fig. 4.15(b)), except for the eastern part of the Cévennes mountains. This can be explained by the reliability and the potential terms of the CRPS (Fig. 4.15(d) and 4.15(f)). Compared with the raw reforecast reference (Fig. 4.15(c) and 4.15(e)), the XLR correction properly corrects the reliability term and large CRPS values affecting especially the Cévennes area are reduced. The potential term is mostly unchanged, except over the eastern area of Cévennes where the potential CRPS is larger after calibration.

Reliability diagrams

The impact of the XLR correction on the reliability of the reforecast is also evaluated by means of reliability diagrams. Reliability diagrams drawn for the raw (left) and the calibrated (right) ensemble are shown in Fig. 4.16. They show an important increase of reliability for q_{80}. Although a weak wet bias is observed, significant reliability improvement is observed for q_{95}. Reliability diagrams provides also the frequencies of probabilities $p(y_i)$ issued by the reforecast, that represent the refinement distribution of the forecasts. Values are given above the points of the reliability diagram for each probability (Fig. 4.16). These values are modified by the XLR calibration compared to the raw reforecast. The forecast probability $y_i = 0$ is reduced compared to the raw, and the corresponding frequency of occurrence is redistributed towards larger forecast probabilities. This means that the forecaster confidence is partially reduced for low probabilities. In the same way, the frequency for the maximum forecast probability ($y_i = 1$) diminishes compared to the raw. This result reveals that the calibrated forecasts are less sharp than the raw ones.

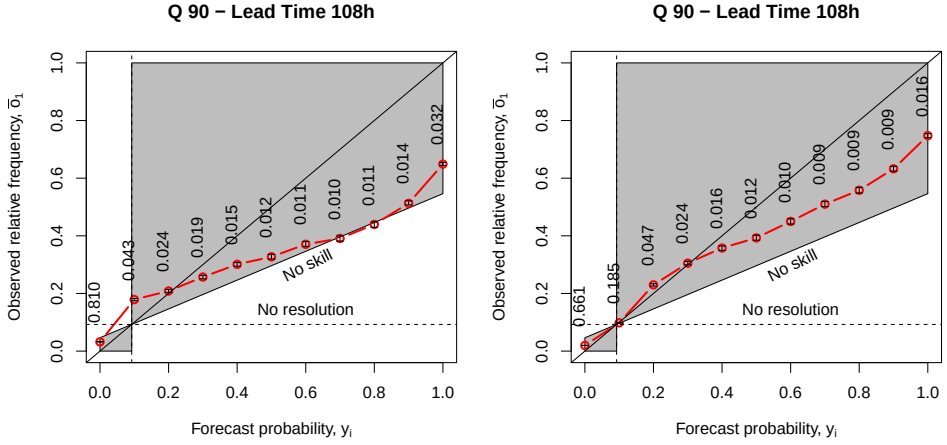
Discrimination diagrams and other diagnostic tools

The forecast probability conditioned by the observed values is explored using the discrimination diagrams. Fig. 4.17(a) and 4.17(b) show the discrimination diagrams for quantile q_{80} for the raw and the XLR calibrated reforecast. It can be useful to remind that this quantile threshold ranges between 0.5 mm and 7 mm, depending on the grid-points (see Fig. 3.13(a)), and corresponds to weak rainfall amounts. The red line, which refers to the probability of the forecast conditioned by the no event,



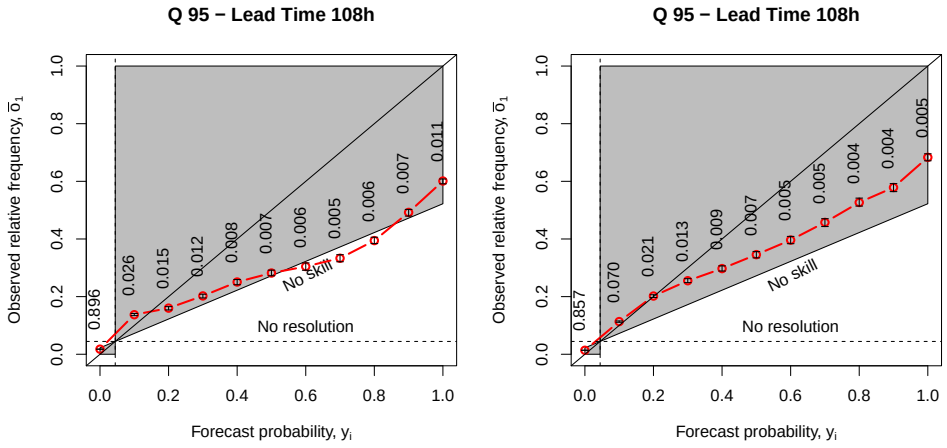
(a) raw

(b) XLR



(c) raw

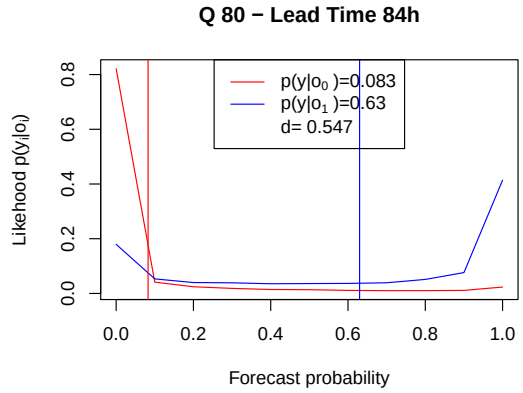
(d) XLR



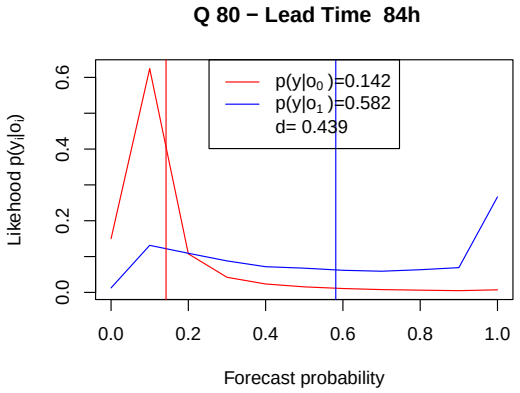
(e) raw

(f) XLR

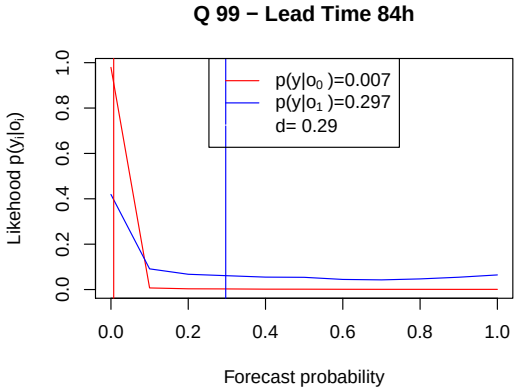
Figure 4.16: Reliability diagrams for q_{80} (top), q_{90} (middle) and q_{95} (bottom), for raw (left) and XLR calibrated (right) reforecast. Values above points indicate the marginal distribution of the forecasts $p(y_i)$. Error bars are estimated using a bootstrapping sampling technique and covers the 90% interval.



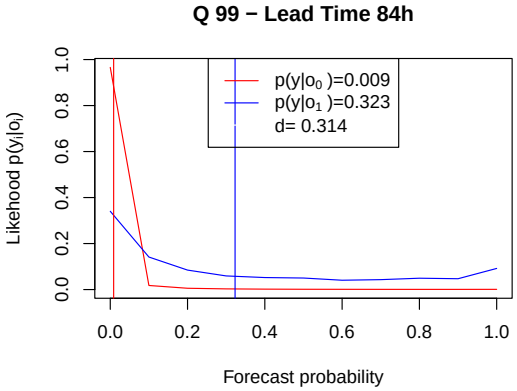
(a) Discrimination diagram (raw)



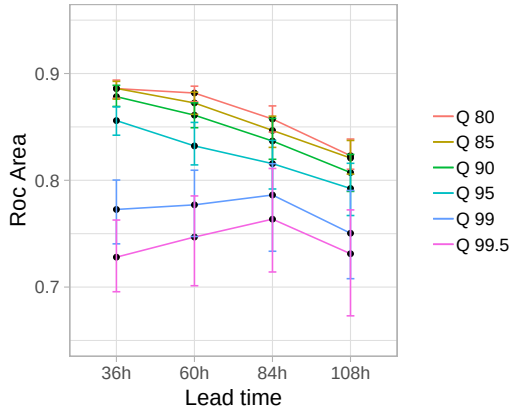
(b) Discrimination diagram (XLR)



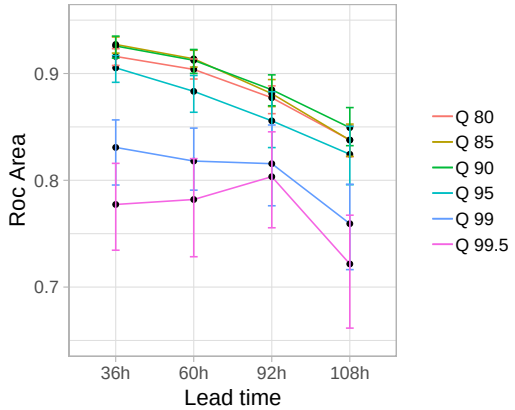
(c) Discrimination diagram (XLR)



(d) Discrimination diagram (XLR)

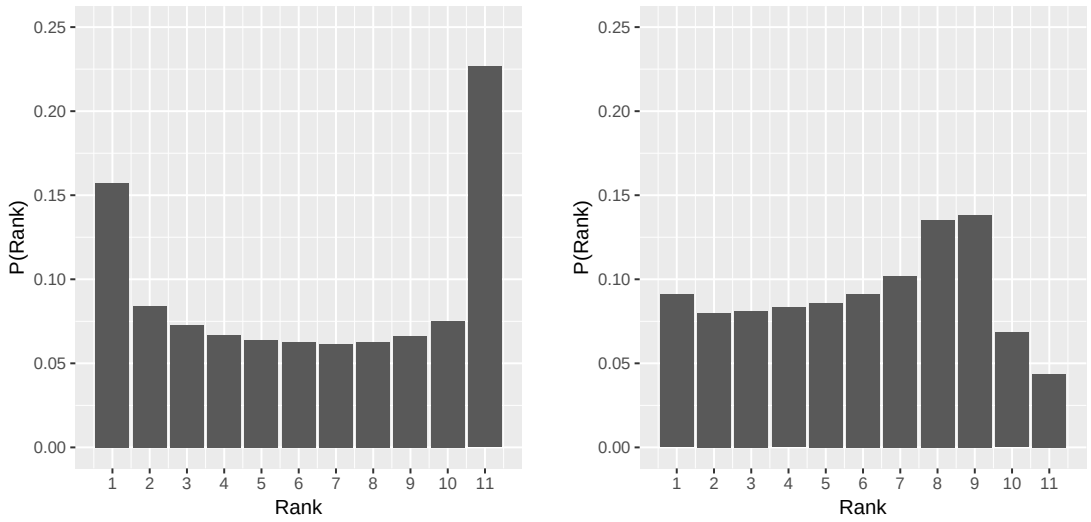


(e) AUC (raw)



(f) AUC (XLR)

Figure 4.17: Top: Discrimation diagrams for 84-hour lead time and for q_{80} threshold computed from the raw (a) and calibrated (b) reforecast. Center: Discrimination diagrams for q_{99} quantile threshold generated from the raw (c) and the XLR calibrated (d) ensemble. Bottom: AUC diagrams using different quantile thresholds generated from the raw (e) and the XLR calibrated (f) ensemble.



(a) Rank histogram (raw)

(b) Rank histogram (XLR)

Figure 4.18: Rank histograms for 84-hour lead time. Panels refers to the raw (left) and the XLR calibrated (right) reforecast.

shifts its maximum from 0 to 0.1 probability. This reveals the presence of a wet bias for low precipitations induced by the XLR procedure. In other words, although the observed values does not exceed the threshold, it often happens that at least one member systematically exceed this value. This bias can also be observed on the rank histogram of the calibrated ensemble. In Fig. 4.18(b) it can be seen that ranks 10 and 11 are less populated than the others. Rank histograms are significantly improved if compared to the raw ones because the U-shape is no more established (see examples of Fig. 4.18(b) and 4.18(a), respectively). Discrimination diagram for a larger observed quantile threshold reveals a weak increase of the discrimination distance compared to the raw (Fig. 4.17(d) and 4.17(c)). AUC values (Fig. 4.17(f)) get generally better, even for the most extreme thresholds (except at 108-hour lead time).

So far the skill of the XLR calibrated ensemble has been assessed in terms of probabilistic scores. Since the correspondence between the members of the raw and the corrected forecast is known from the remapping procedure, this allows to compute deterministic scores for each member before and after correction. For each date the remapping of a given member depends on its rank with respect to the others. This

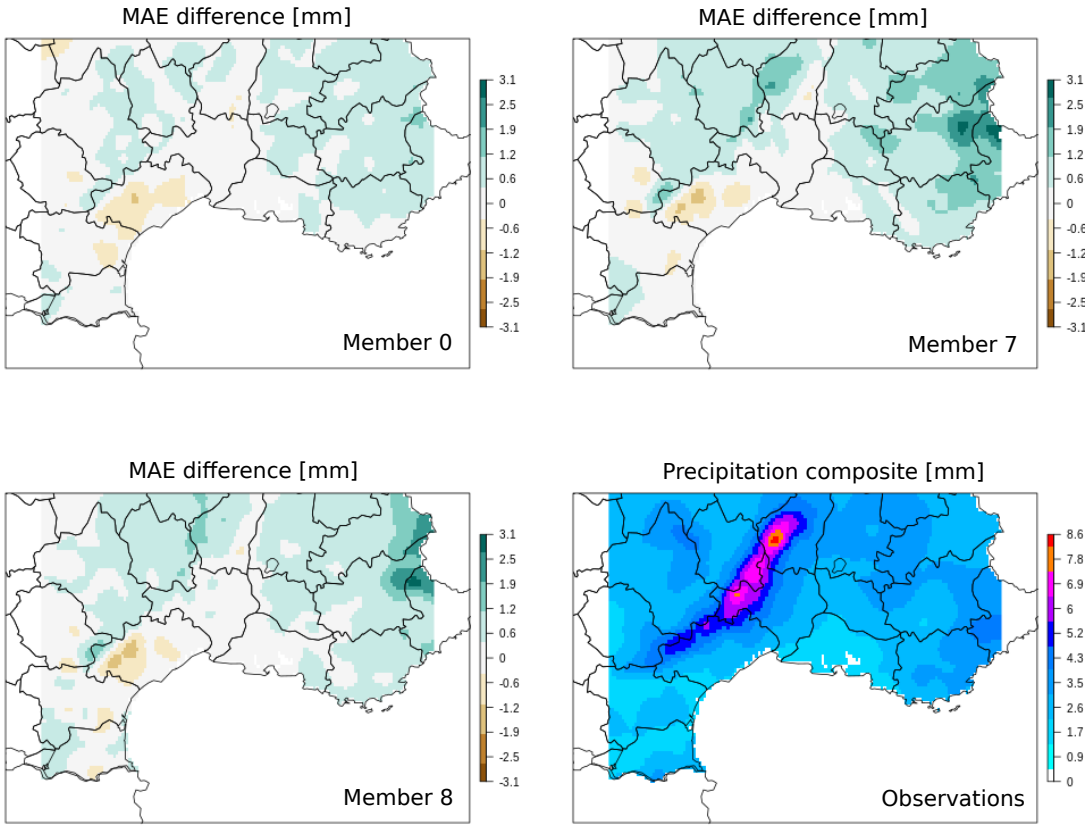


Figure 4.19: MAE difference between the XLR calibrated and raw reforecast for the members 0 (top-left), 7 (top-right) and 8 (bottom-left). Composite of 24-hour precipitation is shown on the bottom-right panel. Results are presented for 36-hour lead time.

rank, for a given member, should, in theory change for every grid-point and every day. However the use of a multiphysics approach leads to ensemble members that are distinguishable, and the rank, for a given member may not be randomly distributed. This effect can lead to some differences from a member to another one concerning correction of the systematic errors. Figure 4.19 shows the MAE difference between the XLR calibrated and raw reforecast computed for member 0, 7 and 8 for 36-hour lead time over the 30-year period. The precipitation composite (bottom-right) corresponds to the averaged 24-hour observed precipitation across the 30-year period for each grid-point. The MAE increases after calibration everywhere, except for a subarea of Languedoc-Roussillon (brown colour). For the 108-hour lead time the brown area is larger and it extends towards the Pyrénées chain (not shown). Score values are similar for member 7 and 8, which implement the same deep convection parametrization scheme (PCMT). This similarity reveals that the remapping tends to transform the members depending on the physical schemes, because the rank of the members tends to depend on the deep convection precipitation scheme. It is worth noting that the MAE difference per grid-point is only partially related to the magnitude of the corresponding observed daily rainfall mean.

4.2.3 Extended logistic regression applied to PEARP-2016

This last part of the calibration section focuses on the application of the XLR procedure on the PEARP-2016 dataset. The logistic regression parametrized distribution that is used for the calibration procedure corresponds to the one fitted on the reforecast dataset at each grid point and lead time.

An example of the procedure for the 35-members resampling is shown in Fig. 4.20. Compared to the resampling applied to the ensemble reforecast (Fig. 4.10), it is possible to observe a more accurate probability sampling of the probability distribution related to the larger number of members in PEARP-2016. In Fig. 4.20(a) the ECDF drawn from the raw ensemble shows strong similarities with the XLR functions, so that the correction is weak. In this case, the remapping procedure has a limited impact on the individual members. Conversely, in the example of Fig. 4.20(b) the correction leads to a significant shift towards more intense values. In

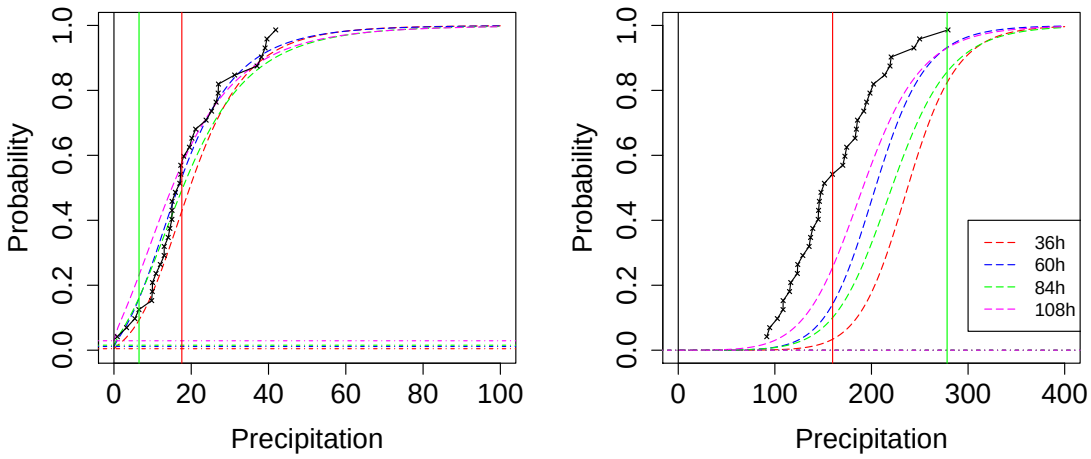


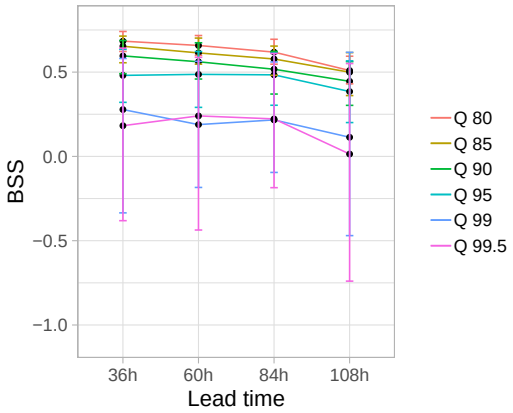
Figure 4.20: As in Fig. 4.10, but for different grid points and for two forecasts at 36-hour lead time produced by PEARP-2016. The vertical green line corresponds to the observed value, the red one to the predicted ensemble mean.

this latter case, only one member of the raw ensemble exceeds the observed value, while after correction 6 members within 35 are assigned to larger values than the observed one. The correction tends to make this specific precipitation forecast more extreme.

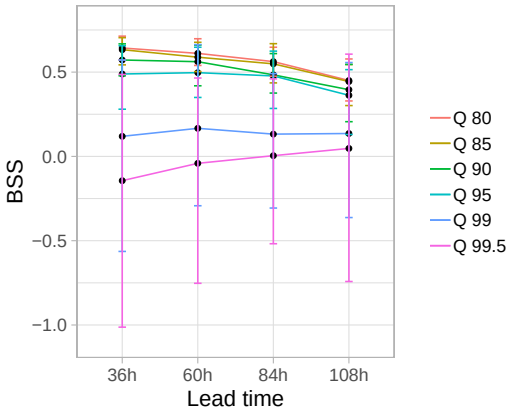
The BSS is used for more general probabilistic verification. The score is computed for the raw and the calibrated PEARP-2016 forecast dataset using the relevant long-term climatology computed from the observation reference over the 30-year period as for the reforecast verification (Juras, 2000; Wilks, 2009b). The post-processed PEARP-2016 ensemble gives similar results, except a weak worsening for the largest thresholds (Fig. 4.21(b)).

Comparing to the raw ensemble, the reliability term of the Brier Score gets worse after calibration (Fig. 4.21(d)), especially for the lowest quantile thresholds. This result is contradictory with the same score for the reforecast, which was improved. The resolution term (not shown) shows small modifications. On another hand we observe that, despite the CRPS score shows less skill overall after calibration, 108-hour lead time are improved.

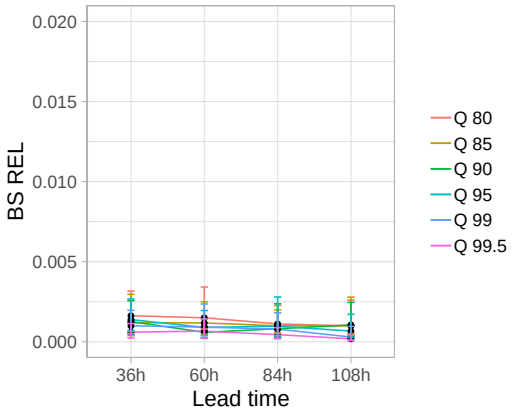
The approach used here is the same as the one developed by Roulin and Vannitsem (2011). They calibrated the daily rainfall precipitation integrated over two small



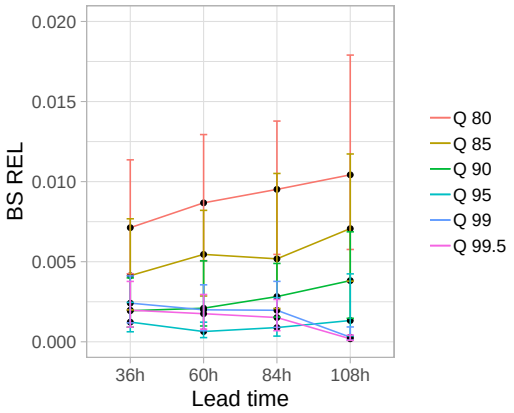
(a) BSS (raw PEARP-2016)



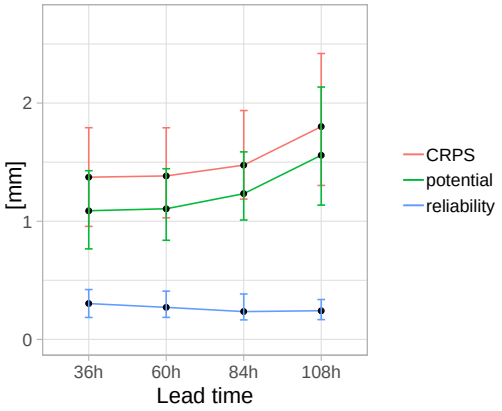
(b) BSS (XLR PEARP-2016)



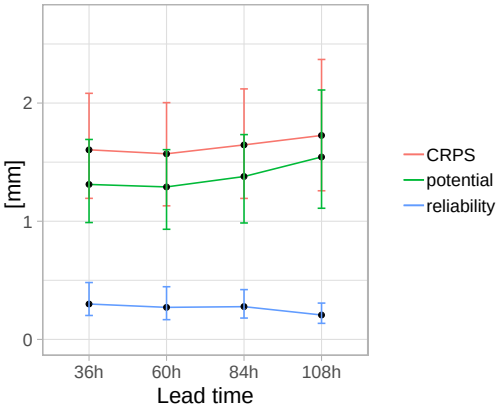
(c) Reliability BSS (raw PEARP-2016)



(d) Reliability BSS (XLR PEARP-2016)



(e) CRPS (raw PEARP-2016)



(f) CRPS (XLR PEARP-2016)

Figure 4.21: BSS (top), reliability term of BSS (middle), and CRPS (bottom) computed on the reforecast dataset for the 24-hour rainfall ensemble forecast using different quantile thresholds. Results are computed for the raw PEARP-2016 (left) and for the calibrated one using the XLR method (right).

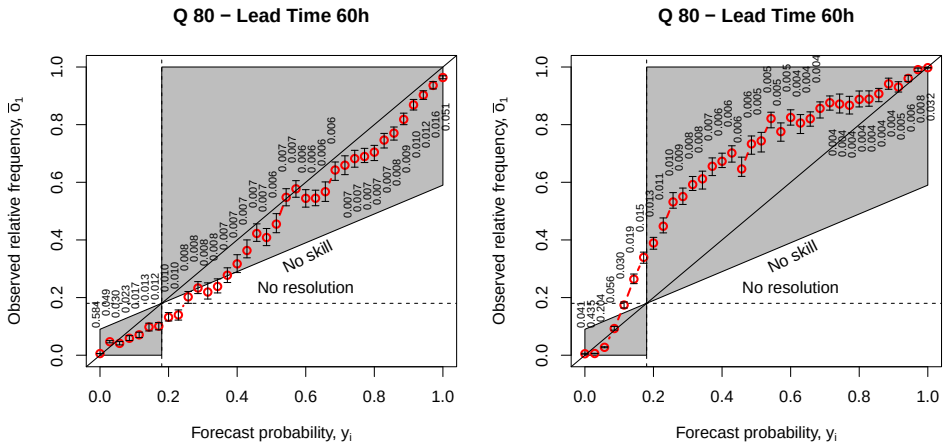
test catchments in Belgium, using a 18-year ensemble reforecast based on ECMWF EPS. Authors show that CRPSS is improved for at least 7-days lead time forecasts. In the present study, the consequences of the calibration are reversed, and a weak benefit is observed only at 108-hour lead time. This comparison reveals how the results can be strongly conditioned by the operational ensemble system adopted as well as the reforecast framework (model resolution, observation resolution, domain of interest...).

The calibrated PEARP-2016 reliability is also investigated trough the reliability diagrams (Fig. 4.22). The raw PEARP-2016 proves to be particularly skillful in terms of reliability (left column). Only a small wet bias is observed for some limited classes of probabilities. We observe that after calibration reliability is decreased. More specifically, a dry bias is added in the middle part of the forecast probability range. The forecast probabilities are consistently too small relative to the corresponding conditional event relative frequencies given by $p(o_1|y_i)$. For the q_{90} threshold this bias is less important and also restricted to probabilities below 0.6. For the q_{95} threshold, we observe a slight improvement after calibration. The better reliability for quantile q_{95} is appreciated, because it shows that forecasts for heavy precipitation benefit from the XLR application.

This resulting dry bias for some quantile thresholds can be related to the significant differences between raw reforecast and PEARP-2016 biases. Since the XLR function is fitted using the raw reforecast, the remapping procedure acts as reducing the wet bias observed in the reforecast. When applying XLR method to the raw PEARP-2016, whose basic reliability is better, the remapping may lead to an overcorrection that could cause the observed bias.

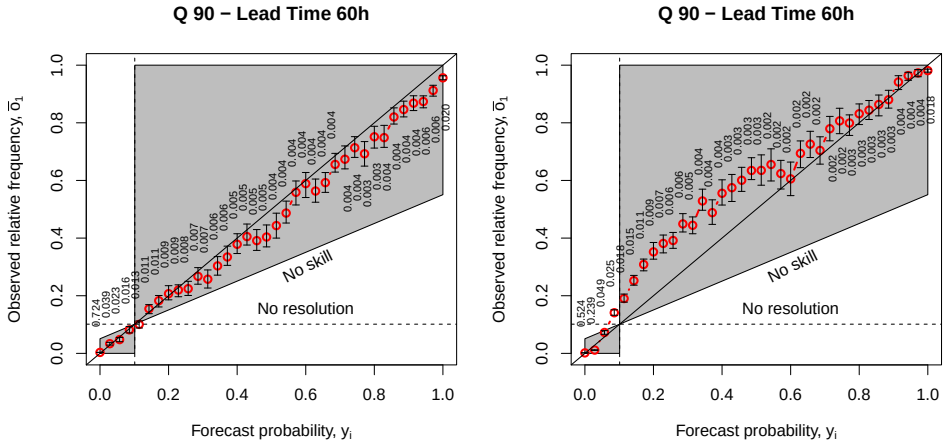
XLR calibrated forecasts tend to better discriminate large precipitation thresholds (q_{99} in this example) compared to the raw ones (Fig. 4.23(a) and 4.23(b)).

The calibration proves to slightly increase the AUC, except for 4-days forecasts (Fig. 4.23(d)). This improvement is more significant for the largest thresholds. The calibration degrades the rank histogram shape (Fig. 4.23(f)), introducing a shape similar to the one observed in the rank histogram drawn from the XLR calibrated reforecast (see Fig. 4.18(b)).



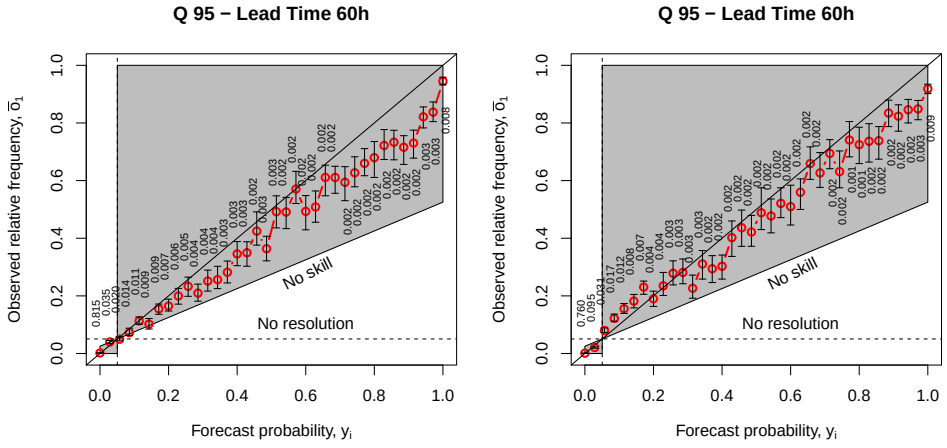
(a) raw

(b) XLR



(c) raw

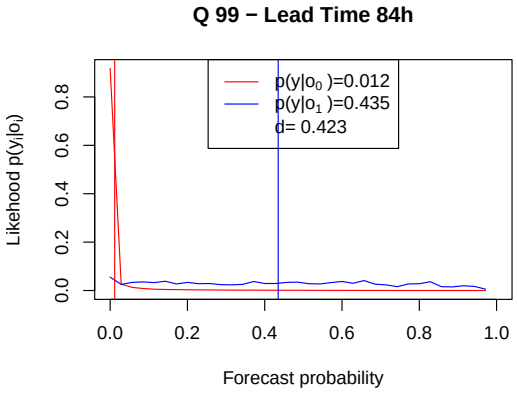
(d) XLR



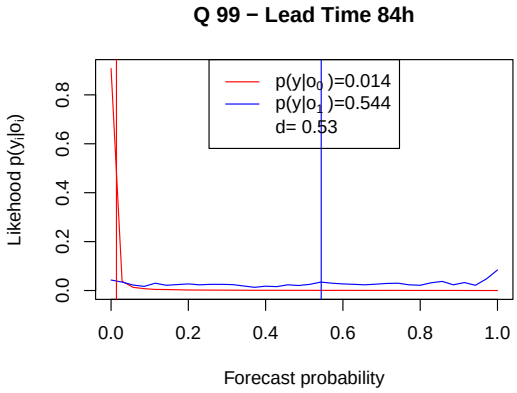
(e) raw

(f) XLR

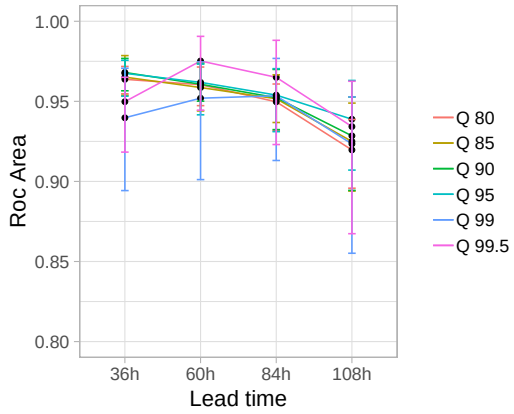
Figure 4.22: Reliability diagrams for q_{80} (top), q_{90} (middle) and q_{95} (bottom), for 2-day forecasts. Left panels are referred to the raw PEARP-2016, right panel to the calibrated version.



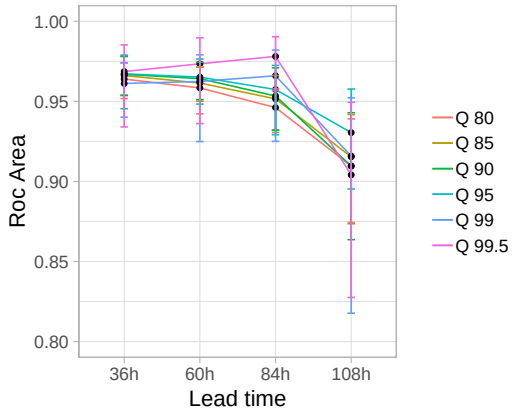
(a) Discr. diagram (PEARP-2016)



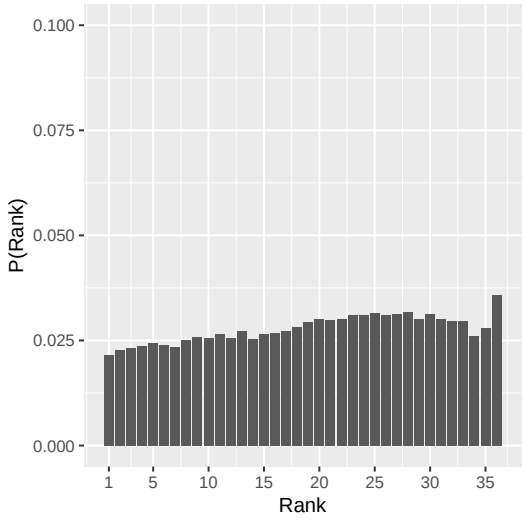
(b) Discr. diagram (XLR PEARP-2016)



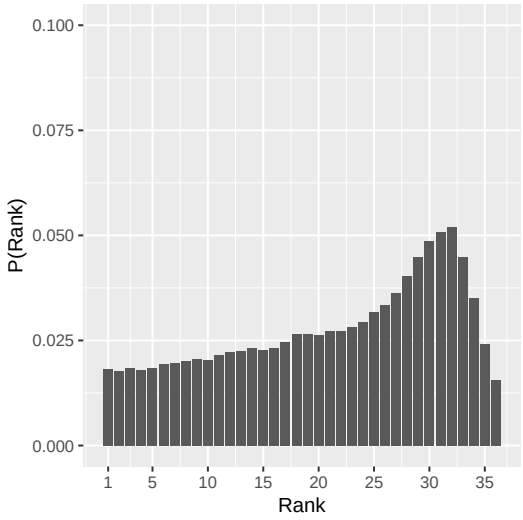
(c) AUC (PEARP-2016)



(d) AUC (XLR PEARP-2016)



(e) Rank histogram (PEARP-2016)



(f) Rank histogram (XLR PEARP-2016)

Figure 4.23: Top: Discrimation diagrams for 84-hour lead time for q_{99} computed from the raw (a) and calibrated (b) PEARP-2016. Center: As in Fig. 3.17(c), but for the raw (c) and calibrated (d) PEARP-2016 . Bottom: Rank histogram for 84-hour lead time computed from the raw (e) and calibrated (f) PEARP-2016.

4.3 Summary and Conclusions

In the current chapter QM and XLR have been applied to the ensemble reforecast dataset. The XLR procedure has been also applied to the PEARP-2016 ensemble, using the reforecast as training dataset.

The results analysis showed that QM method is prone to correct biases of the raw reforecast. The verification member-by-member reveals that MAE is reduced only for forecasts characterized by a positive bias, in other cases MAE get worse. QM does not explicitly address spread errors, because it is primarily meant for the calibration of a deterministic forecast. In fact, probabilistic scores are only tightly impacted by the QM correction. In particular, reliability and resolution are only partially modified by the reduction of the bias. Some small improvement are found for the discriminant ability of the calibrated reforecast.

Conversely, probabilistic scores get generally better with the application of the XLR method on the ensemble reforecast. The correction is designed to remap the ensemble forecasts into the fitted regression functions. After the calibration, the ensemble spread is increased, especially for the first lead times. The benefit of the correction is limited but significant for the largest quantile thresholds. Deterministic scores are not improved by XLR method, demonstrating that this approach is better for probabilistic ensemble forecast. Larger-sized ensembles can be constructed with this post-processing method. A test based on the CRPS score has shown that forecasts are more skillful with larger ensemble sizes.

The application of the XLR procedure on the PEARP-2016 essentially lead to a slight degradation of the scores. The primary reason for this result relies on some original differences between the reforecast and the operational PEARP-2016 ensemble forecast, like the conditional biases shown in the reliability diagrams. A slight improvement for the discrimination skill, as well as the reliability, is observed for the most extreme thresholds.

Chapter 5

Systematic errors analysis of heavy precipitating events prediction using a 30-year hindcast dataset

Contents

5.1 Introduction

In the previous chapter, the characteristics of the reforecast in terms of QPF have been explored using grid-point based approaches. These techniques, especially when applied to intense events, are subject to timing or position errors leading to low scores (Mass et al., 2002). In the current chapter, reforecast is analyzed member-by-member by means of an object-oriented approach, which adresses in particular to the HPEs.

The role of the parametrizations with regard to the intense events is investigated. The statistical analysis, which considers the spatial properties of the predicted and observed objects from the 24-hour precipitation, is carried out on the basis of the SAL measure (Wernli et al., 2009, 2008).

5.2 Article

Systematic errors analysis of heavy precipitating event prediction using a 30-year hindcast dataset

Matteo Ponzano[1], Bruno Joly[1], Laurent Descamps[1], and Philippe Arbogast[1]
[1]CNRM, Météo-France, Toulouse, France
Correspondence: Matteo Ponzano (matteo.ponzano@umr-cnrm.fr)

Abstract. The western Mediterranean region is prone to devastating flash floods induced by heavy precipitation events (HPEs), which are responsible for considerable human and material losses. Quantitative precipitation forecasts have improved dramatically in recent years to produce realistic accumulated rainfall estimations. Nevertheless, there are still challenging issues which must be resolved to reduce uncertainties in the initial conditions assimilation and the modeling of physical processes. In this study, we analyze the HPE forecasting ability of the multi-physics based ensemble model operational at Météo-France Prévision d'Ensemble ARPEGE (PEARP). The analysis is based on 30-year (1981-2010) ensemble hindcasts which implement the same 10 physical parametrizations, one per member, run every 4 days. Over the same period a 24-hour precipitation dataset is used as the reference for the verification procedure. Furthermore, regional classification is performed in order to investigate the local variation of spatial properties and intensities of rainfall fields, with a particular focus on HPEs. As gridpoint verification tends to be perturbed by the double penalty issue, we focus on rainfall spatial pattern verification thanks to the feature-based quality measure SAL that is performed on the model forecast and reference rainfall fields. The length of the dataset allows to sub-sample scores for very intense rainfall at a regional scale and still get significant analysis demonstrating that such a procedure is consistent to study model behaviour in HPE forecasting. In the case of PEARP, we show that the amplitude and structure of the rainfall patterns are basically driven by the deep convection parametrization. Between the two main deep convection schemes used in PEARP, we qualify that the PCMT parametrization scheme performs better than the B85 scheme. A further analysis of spatial features of the rainfall objects to which the SAL metric pertains shows the predominance of large objects in the verification measure. It is for the most extreme events that the model has the best representation of the distribution of object integrated rain.

Copyright statement. TEXT

1 Introduction

Episodes of intense rainfall in the Mediterranean affect the climate of western Europe and can have important societal impact. During these events, daily rainfall amounts associated with a single event can reach annual equivalent values. These rainfall events coupled with a steep orography are responsible for associated torrential floods, which may cause considerable human

and material losses. In particular, Southern France is prone to devastating flash fiood events such as the Aude case (Ducrocq et al., 2003), Gard (Delrieu et al., 2005), and Vaison-La-Romaine (Sénési et al., 1996), which occurred on 12–13 November 1999, 22 September 1992 and 8-9 September 2002, respectively. For instance, in the Gard case more than 600 mm were observed locally during a two-day event and 24 people were killed during the associated flash flooding. Extreme rainfall events generally occur in a synoptic environment favourable for such events (Nuissier et al., 2011).

A detailed list of the main atmospheric factors which contribute to the onset of HPEs are reported by Lin et al. (2001): 1) a conditionally or potentially unstable airstream impinging on the mountains, 2) a very moist low-level jet, 3) a steep mountain, and 4) a quasi-stationary convective system that persists over the threat area. However, not all these factors necessarily need to be present at the same time to produce HPEs. In Southeastern France, the Mediterranean Sea acts as a source of energy and moisture which is fed to the atmospheric lower levels over a wide pronounced orography above the Massif Central, Pyrenees, and South Alps areas (Delrieu et al., 2005). Extreme rainfall amounts are enhanced especially along the Southern and Eastern foothills of mountainous chains (Frei and Schär, 1998; Nuissier et al., 2008), in particular the Southeastern part of the Massif Central (Cévennes). Ehmele et al. (2015) emphasized the important role played by complex orography, the mutual interaction between two close mountainous islands in this case, on heavy rainfall under strong synoptic forcing conditions. Nevertheless, other regions are also affected by rainfall events with a great variety of intensity and spatial extension. Ricard et al. (2011) studied this regional spatial distribution based on a composite analysis and showed the existence of mesoscale environments associated with heavy precipitating events. Considering four sub-domains, they found that the synoptic and mesoscale patterns can greatly differ as a function of the location of the precipitation.

Extreme rainfall events are generally associated with coherent structures slowed down and enhanced by the relief, whose extension is often larger than a single thunderstorm cell. At some point, this mesoscale organization can turn into a self-organization process leading to a mesoscale convective system (MCS) when interacting with their environment, which in turn leads to high intensity rainfall (Nuissier et al., 2008).

Among the list of factors contributing to HPE creation, some are clearly only within the scope of high resolution convection permitting models. Indeed, vertical motion and moisture processes need to be explicitly solved to get realistic representation of convection. On the other hand, as we have just highlighted, some other factors linked with synoptic circulations or orography representations can be well estimated in global models, in particular when horizontal resolution gets close to 15-20 km. Consequently, the corresponding predictability of such factors can reach advantageous lead times for early warnings, i.e. longer than the standard 48 hours that the limited area model may be expected to achieve. Indeed, if long term territorial adaptations are necessary to mitigate the impact of HPEs, a more reliable and earlier alert would be beneficial in the short term. Weather forecasting coupled with hydrological impact forecasting is the main source of information for triggering of weather warnings. Severe weather warnings are issued for the 24-hour forecast only. However, in some cases, the forecast process could be issued some days prior to the severe weather warnings. A better understanding of the sources of model uncertainty at such time-range may provide a major source of improvement for early diagnosis.

Forecast uncertainties can be related to initialization data (analysis) or lateral boundary conditions, and it has been investigated with both deterministic models (Argence et al., 2008) and ensemble models (Vié et al., 2010). Several previous studies

showed that predictability associated with intense rainfall and flash-floods decreases rapidly with the event scale (Walser et al., 2004; Walser and Schär, 2004; Collier, 2007). Several studies based on ensemble prediction systems have shown the general ability of such models to sample the sources of uncertainty in HPE probabilistic forecasting (Du et al., 1997; Petroliagis et al., 1997; Stensrud et al., 1999; Schumacher and Davis, 2010; World Meteorological Organization, 2012). In ensemble forecasting, the uncertainty associated with the forecast is usually assessed by taking into account initial and model error propagation. As for the initial uncertainty, major meteorological centers implement different methods: the most common of which are singular vectors (Buizza and Palmer, 1995; Molteni et al., 1996) , bred vectors (Toth and Kalnay, 1993, 1997) and perturbed observation in analysis process (Houtekamer et al., 1996; Houtekamer and Mitchell, 1998). The model error is related to grid-scale unsolved processes in the parametrization scheme and is assessed in the models with two main techniques. Some models use stochastic perturbations of the inner-model physics scheme (Palmer et al., 2009), others use different parametrization schemes in each forecast member (Charron et al., 2009; Descamps et al., 2011).

The global ensemble model implemented at Météo-France Prévision d'Ensemble ARPEGE (PEARP; Descamps et al., 2015) is based on the second technique, also known as a multi-physics approach. Compared to the stochastic perturbation, the error model distribution cannot be explicitly formulated in the multi-physics approach. It is then difficult to know *a priori* the influence of the physics scheme modifications on the forecast ability of the model. This is even more the case when highly non-linear physics with high order of magnitude processes are considered. In order to improve the understanding and interpretation of ensemble forecasts in tense decision-making situations as well as for model development and improvement purposes, it would be of great interest to have a full and objective analysis of the model behaviour in terms of HPE forecasting. This is one of the main aims of this study.

In order to achieve such a systematic analysis, standard rainfall verification methods can be used. They are usually based on grid-point based approaches. These techniques, especially when applied to intense events, are subject to time or position errors leading to low scores (Mass et al., 2002) also known as the double penalty problem (Rossa et al., 2008). To counteract this problem, spatial verification techniques have been developed with the goal of evaluating a forecast quality from a forecaster standpoint. Some of these techniques are based on object-oriented verification methods (Ebert and McBride, 2000; Davis et al., 2006a; Wernli et al., 2008; Davis et al., 2009; AghaKouchak et al., 2011; Mittermaier et al., 2015). The feature-based quality measure SAL (Wernli et al., 2008, 2009) is used in this study. Another element required to achieve such an analysis is the availability of forecast datasets long enough to get a proper sampling of the events to verify.

In our study, we profit from a reforecast dataset based on a simplified version of the PEARP model available over a 30 year period. Such reforecast datasets have been previously shown to be relevant for calibrating operational models in various ways. In Hamill and Whitaker (2006), Hamill et al. (2008), Hamill (2012) and Boisserie et al. (2015), the reforecast is used as a learning dataset to fit statistical models to calibrate forecast error corrections that are then applied on operational forecasting outputs. Boisserie et al. (2015) and Lalaurette (2003) have shown the possibility of using a reforecast dataset as a statistical reference of the model to which the extremeness of a given forecast is compared. In this paper, we analyze the ensemble model PEARP forecast predictability at lead times between day 2 and day 4 of daily rainfall amounts. This analysis is performed on the long reforecast 30-year dataset. One aim is to determine whether a multi-physics approach could be considered as a model

error sampling technique appropriate for a good representation of HPEs in the forecast at such lead times. In particular, the behaviour of the different physics schemes implemented in PEARP have to be estimated individually. One main side aspect of this work focuses on developing a methodology suitable for evaluating the performances of an ensemble reforecast in a context of intense precipitation events using an object oriented approach. In particular, we focus on intense precipitation over the French Mediterranean region. In addition to the analysis of diagnostics from the SAL-metric, a statistical analysis of 24-hour rainfall objects identified in the forecasts and the observations is performed in order to explore the spatial properties of the rainfall fields.

The data and the methodology are presented in Section 2. Section 2.1 describes the reforecast ensemble dataset and Section 2.2 details the creation of the daily rainfall reference, the HPEs statistical definition, and the regional clustering analysis. Results arising from the spatial verification of the overall reforecast dataset are presented in Section 3.1. Section 3.2 presents SAL diagnostics divided into all different physical parametrization schemes of the ensemble reforecast, and for the spatial properties of individual objects. Conclusions are given in Section 4.

2 Data and methodology

2.1 PEARP hindcast

The PEARP reforecast dataset consists of a 10-member ensemble computed daily from 1800 UTC initial conditions, covering four months (from September to December), every year of a 30-year period (1981-2010). This period has been chosen since HPE occurrence in the considered region is largest during the autumn season (see Fig. 3 from Ricard et al., 2011). It uses ARPEGE (Action de Recherche Petite Echelle Grande Echelle, Courtier et al. (1991)), the global operational model of Météo-France with a spectral truncation T798, 90 levels on the vertical, and a variable horizontal resolution (mapping factor of 2.4 with a highest resolution of 10 km over France). One ensemble forecast is performed every 4 days of the four-month period up to 108-hour lead time. Our initialization strategy follows the hybrid approach described in Boisserie et al. (2016), in which first the atmospheric initial conditions are extracted from the ERA-Interim reanalysis (Dee et al., 2011) available at the European Center for Medium-range Weather Forecasts. Second, the land-surface initialization parameters are interpolated from an offline simulation of the land-surface SURFEX model (Masson et al., 2013) driven by the 3-hourly near-surface atmospheric fields from ERA-Interim. 24-hour accumulated precipitation forecasts are extracted on a $0.1^\circ \times 0.1^\circ$ grid, that defines the domain D (see Fig. 1c), which encompasses Southeastern France (Fig. 1a). The reforecast dataset does not have any representation of initial uncertainty, but it implements the same representation of model uncertainties (multiphysics approach) as in the PEARP operational version of 2016.

Nine different physical parametrizations (see Table 1) are added to the one that corresponds to the ARPEGE deterministic physical package. This set of parametrizations is the same as the one implemented in PEARP. Two turbulent diffusion schemes are considered: the Turbulent Kinetic Energy scheme (TKE; Cuxart et al., 2000; Bazile et al., 2012) and the Louis scheme (L79; Louis, 1979). TKE_{mod} is a slightly modified version of TKE, in which horizontal advection is ignored. For shallow convection, different schemes are used: a mass flux scheme introduced by Kain and Fritsch (1993) and modified by Bechtold et al. (2001),

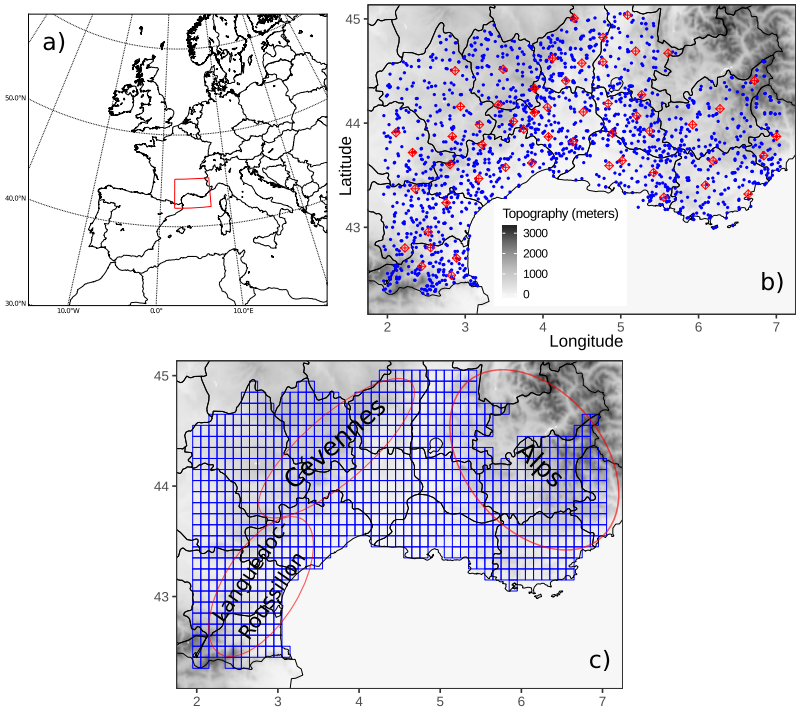


Figure 1. Panel **a** shows a situation map of the investigated area (rectangle with red edges) with respect to Western Europe and the Mediterranean Sea. Panel **b** shows the rain-gauges network used for the study. Red diamonds represent the rain-gauges selected for cross-validation testing, blue dots represent the rain-gauges selected for cross-validation training. Panel **c** shows the $0.1^\circ \times 0.1^\circ$ model grid (in blue), along with the location of three key areas. The domain D is located within the borders of the model grid (panel **c**).

thereafter the KFB approach, the Prognostic Condensates Microphysics and Transport scheme (PCMT; Piriou et al., 2007)), the Eddy-Diffusivity/Kain-Fritsch scheme (EDKF) and the PMMC (Pergaud, Masson, Malardel, Couvreux) scheme (Pergaud et al., 2009). The deep convection component is parametrized by either the PCMT scheme or the Bougeault (1985) scheme (thereafter B85). Closing the equation system used in these two schemes means relating the bulk mass flux to the in-cloud vertical velocity through a quantity γ qualifying the convection area coverage. Two closures are considered: the first one (C1) is based on the convergence of humidity and the second one (C2) is based on the CAPE (Convective Available Potential Energy). B85 scheme originally uses the C1 closure, while PCMT alternatively uses the closure (C1 or C2) which maximizes the γ parameter. Physics package 2 uses a modified version of the B85 scheme in which deep convection is triggered only if cloud top exceeds 3000 m (B85_{mod} in Table 1). The same trigger is used in physics package 3 in which deep convection is parametrized using the B85 scheme along with a CAPE closure (CAPE in Table 1). Finally the oceanic flux is solved by means of the ECUME

Table 1. Physical parametrizations used in the ensemble reforecast.

	Turbulence	**Shallow convection**	**Deep convection**	**Oceanic flux**
Ref	TKE	KFB	B85	ECUME
1	TKE	KFB	B85	$ECUME_{mod}$
2	L79	KFB	$B85_{mod}$	ECUME
3	L79	KFB	CAPE	ECUME
4	TKE_{mod}	KFB	B85	ECUME
5	TKE	EDKF	B85	ECUME
6	TKE	PMMC	PCMT	ECUME
7	TKE	KFB	PCMT	ECUME
8	TKE	PCMT	PCMT	ECUME
9	TKE	KFB	B85	ECUME

(Exchange Coefficients from Unified Multicampaigns Estimate) scheme (Belamari, 2005). In $ECUME_{mod}$ evaporation fluxes above sea surfaces are enhanced. Control member and member 9 are characterized by the same parametrization set-up, but member 9 differs for the modelization of orographic waves.

2.2 Daily Rainfall Reference

24-hour accumulated precipitation is derived from the in-situ Météo-France rain-gauge network, covering the same period as the reforecast dataset. 24-hour rainfall amounts collected from fourteen French departments within the reforecast domain D are used (Fig. 1b). In order to maximize the rain-gauge network density within the region, all daily available validated data covering the period have been used.

Rain-gauge observations are used to build gridded precipitation references by a statistical spatial interpolation of the observations. The aim of this procedure is to ensure a spatial and temporal homogeneity of the reference, as well as the same spatial resolution as the reforecast dataset. Ly et al. (2013) provided a review of the different methods for spatial interpolation of rainfall data. They showed that kriging methods outperform deterministic methods for the computation of daily precipitation. However, both types of methods were found to be comparable in terms of hydrological modelling results. For the interpolation, we use a mixed geo-statistical and deterministic algorithm, which implements Ordinary Kriging (OK; Goovaerts et al., 1997) and Inverse Distance Weighting methods (IDW; Shepard, 1968). For the kriging method, three semi-variogram models (Exponential, Gaussian and Spherical) are fitted to daily sample semi-variogram drawn from raw and square root transformed data (G. Gregoire et al., 2008; Erdin et al., 2012). This configuration involves the use of six different geo-statistical interpolation models. In addition, four different IDW versions are used, by varying the geometric form parameter d used for the estimation of the weights (see Eq. (2) in Ly et al., 2011) and the maximum number n of neighbour stations involved in the IDW compu-

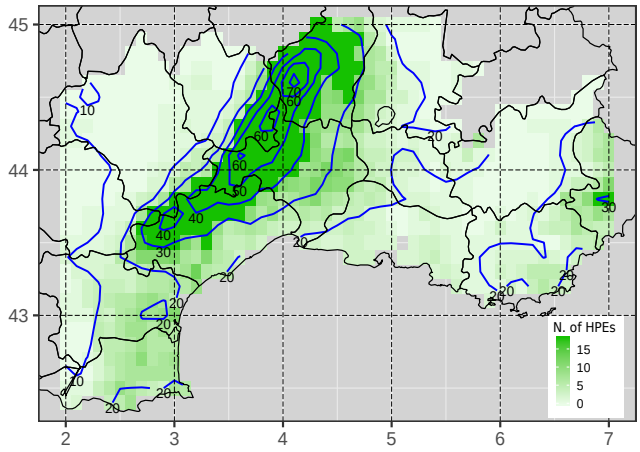


Figure 2. Annual average of HPE occurrence per grid point (in green). The composite of daily rainfall amounts (mm/day) of the HPE dataset is represented by the blue isohyets.

tation. Three versions are defined by fixing parameter $d = 2$ and alternatively assigning n values equal to 5, 10 and N (with N being the total number of stations available for that specific day). In the fourth version we set $n = N$ and $d = 3$. For each day, a different interpolation method is used and its selection is based on the application of a cross validation approach. We select 55 rain-gauges as a training dataset (see the red diamonds in Fig. 1c) in order to have sufficient coverage over the domain, especially on the mountainous area. Root Mean Square Error (RMSE) is used as a criterion of evaluation. For each day, the method which minimizes the RMSE computed within the rain-gauges of the training dataset is selected and the spatial interpolation is then performed on a regular high resolution grid of 0.05°. The highest resolution estimated points are then up-scaled to the 0.1°grid resolution of domain D, by means of a spatial average. This up-scaling procedure aims at reproducing the filtering effect produced by the parametrizations of the model on the physical processes that occur below the grid resolution.

2.2.1 HPE database

We implement a methodology in order to select the HPEs from the daily rainfall reference. Anagnostopoulou and Tolika (2012) have examined parametric and non-parametric approaches for the selection of rare events sampled from a dataset. Here we adopt a non-parametric peak-over-threshold approach, on the basis of WMO guidelines (World Meteorological Organization, 2016). The aim is to generate a set of events representative of the tail of the rainfall distribution for a given region and season. Following the recommendation of Schär et al. (2016), an all-day percentile ($P_{0 \leq n \leq 1}$) formulation is applied. A potential weakness of the research methodology based on the gridded observation reference is that a few extreme precipitation

events affecting a smaller area than the grid resolution may not be identified. However, this approach has been preferred to a classification using rain-rauges because spatial and temporal homogeneity are ensured.

We proceed as follows: first the domain is split into two sub-regions based on the occurrence of climatological intense precipitations during the 30 year period. The sub-region A includes all the points whose climatological 99.5 percentile is lower or equal to a threshold T, subregion B includes all the other points. Threshold T, after several tests, has been set to 85 mm. This choice was made in order to separate the domain into two regions characterized by different frequency and intensity of HPEs. Subregion A designates a geographical area where a large number of cases of intense precipitation are observed. Subregion B primarily covers the plain area, where HPE frequency is lower. For this reason, two different level thresholds values are selected to define an event, depending on the subregion. More specifically, a day is classified as an HPE if one point of sub-region A accumulated rainfall is greater than 100 mm or if one point of sub-region B rainfall is greater than its 99.5 percentile. The selection led to a classification of 192 HPEs, corresponding to a climatological frequency of 5% over the 30-year period. The 24-hour rainfall amount maxima within the HPE dataset ranges from 100 mm to 504 mm. It is worth mentioning that since we consider daily rainfall, rainfall events that would have high 48 hour or 72 hour accumulated rainfall may be disregarded. Figure 2 shows for each point of the domain the number of HPE, as well as the composite analysis of HPEs. The composite analysis involves computing the grid point average from a collection of cases. The signal is enhanced along the Cévennes chain and on the Alpine region. It should be noted that some points are never taken into account for the HPE selection (white points of Fig. 2), because the required conditions have not been met. The analysis of the rainfall fields across the HPE database exhibits the presence of patterns of different shape and size, revealing potential differences in terms of the associated synoptic and mesoscale phenomena (not shown).

2.2.2 Clustering analysis

Clustering analysis methods can be applied to daily rainfall amounts in order to identify emergent regional rainfall patterns. This classification is largely used for assessing the between-day spatial classification of heavy rainfall (Romero et al., 1999; Peñarrocha et al., 2002; Little et al., 2008; Kai et al., 2011). We applied a cluster analysis, as an exploratory data analysis tool, in order to assess geographical properties of the precipitation reference dataset. The size of the dataset is first reduced and the signal is filtered out by means of a principal component analysis (Morin et al., 1979; Mills, 1995; Teo et al., 2011). The first 13 Principal Components (PCs), whose projection explains 90% of the variance, are retained. Then the *K*-means clustering method is applied. It is a non-hierarchical method based on the minimization of the intraclass variance and the maximization of the variance between each cluster. A characteristic of *k*-means method is that the number of clusters (K) into which the data will be grouped has to be *a priori* prescribed. Consequently, we first have to implement a methodology to find the number of clusters which leads to the most classifiable subsets.

The analysis is applied to the full reference dataset, including rainy and dry days. We run 2000 tests for a range of *a priori* cluster numbers K that lie between 3 and 13, by varying a random initial guess each time. Then, for a given K, an evaluation of the stability of the assignment into each cluster is performed. The number of clusters is considered stable if each cluster size is almost constant from one test to another. $K = 5$ is retained as the most stable number of clusters and because it suggests a

Table 2. Classification of days computed from 24-hour rainfall amounts in southern France (1981-2010), percentage of HPEs and fraction of HPEs. HPEs(%) refers to the ratio between the number of HPEs within the cluster and the total number of HPEs. Fraction of HPEs (%) refers to the ratio between the number of HPEs within the cluster and the total number of dates included in the corresponding cluster.

Cluster	Total (%)	HPEs (%)	Fraction of HPEs (%)
1	14.5	11.4	4.3
2	5.3	24.0	24.6
3	1.8	30.7	92.2
4	75.8	2.6	0.2
5	2.6	31.3	65.2
Total number of days	3660	192	

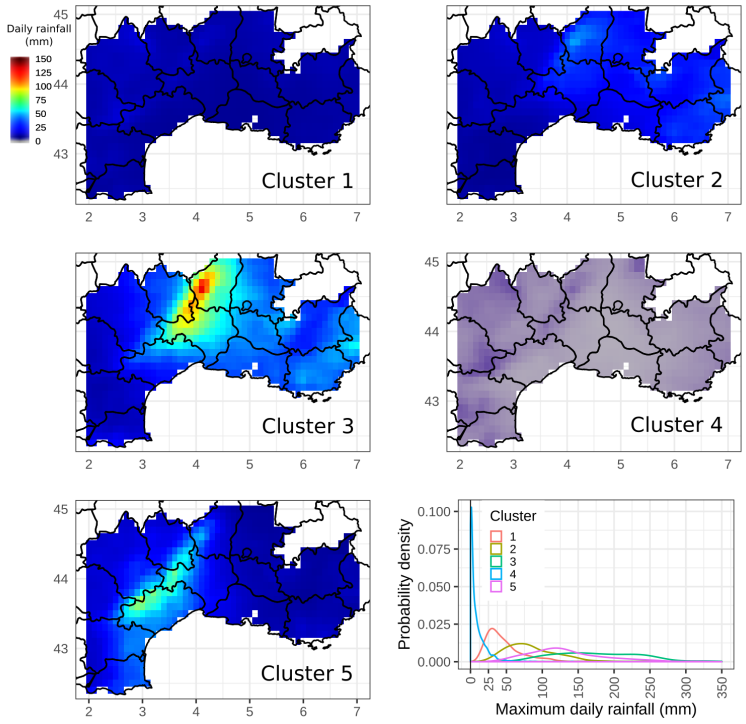


Figure 3. Rainfall composites (mm/day) for the 5 clusters selected by the *K*-means algorithm. The bottom-right panel shows the probability density distribution of the maximum daily rainfall (mm) for each cluster class.

coherent regional stratification of the daily rainfall data. The final classification within the 2000 tests is selected by minimizing the sum of the distance between the cluster centroids from each test and the geometric medians of cluster centroids computed from all the tests. The test which minimizes this quantity has been selected as the reference classification. The results from the cluster classification are summarized in Table 2. The clusterization shows large differences in term of cluster size, more than 3/4 of the dataset is grouped in cluster 4, which mostly collects the days characterized by weak precipitation amounts or dry days. The percentage of HPEs within the clusters shows that the most intense events are represented in clusters 2, 3 and 5, among which cluster 5 shows largest proportion of HPE (65% of HPEs within this cluster). Clusters 2,3 and 5 together account for 86% of the HPEs.

The same composite analysis as the one previously applied to HPE class, is now computed for each cluster class (Fig. 3). It shows significant differences between clusters. Not only the relative intensity of events is different for each of the clusters, but also the location differs. Rainfall range is weak for cluster 1 and close to zero for cluster 4. Cluster 2 includes some moderate 24-hour rainfall amounts related to generalized precipitation events and a few HPEs. For cluster 1, composite values are slightly higher on the northwestern area of the domain, while for cluster 2, rainfall amounts values are more significant on the eastern side of the domain D. Clusters 3 and 5 together account for 63% of the HPEs of the whole period, but rainfall events seem to affect different areas. Cluster 3 includes most of the events impacting the Cévennes mountains and the eastern departments on the southern side of the Alps. Cluster 5 average rainfall is enhanced along the southern side of the Cévennes, especially the Languedoc-Roussillon region.

The bottom-right panel of Fig. 3 shows the density distributions computed from the maximum daily rainfall for each cluster. It is worth noting that cluster rainfall distributions cover different intervals of maximum daily rainfall amounts. Cluster 4 includes all the dry days. As this paper focuses on the most severe precipitation events, results will only be shown for clusters 2, 3 and 5 for the remainder of the paper.

2.3 The SAL verification score

2.3.1 The SAL score definition

The SAL score is an object-based quality measure introduced by Wernli et al. (2008) for the spatial verification of numerical weather prediction (NWP). It consists in computing three different components: structure **S** is a measure of volume and shape of the precipitations patterns, amplitude **A** is the normalized difference of the domain-averaged precipitation fields, and location **L** is the spatial displacements of patterns on the forecast/observation domains.

Different criteria for the identification of the precipitation objects could be implemented: a threshold level (Wernli et al., 2008, 2009), a convolution threshold (Davis et al., 2006a, b), or a threshold level conditioned to a cohesive minimum number of contiguous connected points (Nachamkin, 2009; Lack et al., 2010). The threshold level approach needs only one estimation parameter, so it has been preferred to the other methods for its simplicity and interpretability. Since we focus on the patterns associated with the HPEs, we decided to adapt the threshold definition given by $T_f = x_{max} \times f$, where x_{max} is the maximum precipitation value of the points belonging to the domain and f is a constant factor (=1/15, in the paper of Wernli et al., 2008).

Here the coefficient f has been raised to $1/4$, because a smaller value results in excessively large objects spreading out over most of the domain D. Choosing a higher f factor enables to obtain more realistic features within the domain considered. Threshold levels T_f are computed daily for the reforecast and the reference dataset. Although objects are smaller than the domain for most of the situations, a few objects extending outside the domain are consequently limited by the boundaries of the region concerned.

If we consider the domain D, the amplitude A is computed as follows:

$$A = \frac{\langle R_{\rm for} \rangle_D - \langle R_{\rm obs} \rangle_D}{0.5(\langle R_{\rm for} \rangle_D + \langle R_{\rm obs} \rangle_D)} \in [-2,2]\,, \tag{1}$$

where $\langle \rangle_D$ denotes the average over the domain D. $R_{\rm for}$ and $R_{\rm obs}$ are the 24-hour rainfall amounts over D associated with the forecast and the observation, respectively. A perfect score is achieved for $A = 0$. The domain-averaged rainfall field is overestimated by a factor 3 if $A = 1$, similarly it is underestimated by a factor 3 if $A = -1$. The amplitude is maximal ($A = 2$) if $\frac{\langle R_{\rm for} \rangle_D}{\langle R_{\rm obs} \rangle_D} \to +\infty$ and minimal ($A = -2$) if $\frac{\langle R_{\rm for} \rangle_D}{\langle R_{\rm obs} \rangle_D} \to 0$.

The two other components require the definition of precipitation objects (thereafter $\{Obj\}$), also called features, which represent contiguous grid points belonging to the domain D, characterized by rainfall values exceeding a given threshold. The location L is a combined score defined by the sum of two contributions, $L1$ and $L2$. $L1$ measures the magnitude of the shift between the center of mass of the whole precipitation field for the forecast ($\bar{x}_{\rm for}$) and observation ($\bar{x}_{\rm obs}$):

$$L1 = \frac{|\bar{x}_{\rm for} - \bar{x}_{\rm obs}|}{d} \in [0,1]\,, \tag{2}$$

where d is the largest distance between two boundary points of the considered domain D. The second metric $L2$ takes into account the spatial distribution of the features inside the domain, that is the scattering of the objects:

$$r = \frac{\sum_{n=1}^{N} M_n |\bar{x} - x_n|}{\sum_{n=1}^{N} M_n}, \tag{3}$$

where M_n is the integrated mass of the object n, x_n is the center of mass of the object n, N is the number of objects and $\bar{x}$ is the center of mass of the whole field.

$$L2 = 2\frac{|r_{\rm for} - r_{\rm obs}|}{d} \in [0,1]\,, \tag{4}$$

$$L = L1 + L2 \in [0,2]\,. \tag{5}$$

$L2$ aims at depicting object differences between observed and forecasted scattering of the precipitation objects. We can notice that the scattering variable (Eq. (3)) is computed as the weighted distance between the center of total mass and the center of mass of each object. Therefore L is a combination of the information provided by the global spatial distribution of the fields (L1) and the difference in scattering of the features over the domain (L2). The location score is perfect if $L1 = L2 = 0$, so if $L = 0$ all the centers of mass match each other.

The S-component is based on the computation of the integrated mass M_k of one object k, scaled by the maximum rainfall amount of the object k:

$$V_k = \frac{M_k}{\max R(x; x \in Obj_k)}. \tag{6}$$

Then, the weighted average V of all features is computed, in order to obtain a scaled, weighted total mass:

$$V = \frac{\sum_{n=1}^{N} M_n V_n}{\sum_{n=1}^{N} M_n}, \tag{7}$$

$$S = \frac{V_{\mathrm{for}} - V_{\mathrm{obs}}}{0.5(V_{\mathrm{for}} + V_{\mathrm{obs}})} \in [-2, 2]\,. \tag{8}$$

Then, S represents the difference of both forecasted and observed volumes, scaled by their half-sum. It is important to scale the volume so that the structure is less sensitive to the mass, meaning that it relates more to the shape and extension of the features rather than their intensities. In particular $S < 0$ means that the forecast objects are large and/or flat compared to the observations. Inversely, peaked and/or smaller objects in the forecast give positive values of S. We refer to Wernli et al. (2008) for the exploration of the behaviour of SAL for some idealized examples.

On the basis of the definition of the score, it can be noticed that A and $L1$ components are not affected by the object identification and depend only on the total rainfall fields.

2.3.2 A selected example of the application of SAL

An example of the SAL score applied to an HPE, that occurred on the 28 Oct 2004, is shown in Fig. 4 (60-hour lead time forecast run using the physical package n.8). For the rainfall reference, a 24-hour rainfall maximum value (121.3 mm), was registered in the southeastern coastal region. Therefore the threshold level T_f is set to 30.3 mm. For the forecast, the maximum value is 123.1 mm ($T_f = 30.8$ mm) and, in contrast with the reference, it is located on the Cévennes. The number of objects, three, is equivalent in both fields. The value of A is 0.08, which means that the domain-averaged precipitation field of the forecast is nearly similar to the reference one. The structure S-components is positive (0.28), which could be explained by the larger forecast object over the Cévennes area, while the object along the southeastern coast is smaller and less intense. The contribution of the third object is negligible for the computation of S. The L-component is equal to 0.23, with $L1$=0.13 and $L2$=0.10. The location error $L1$ means that the distance between the centers of total mass (see diamonds in Fig. 4) is $13/100$ of the largest distance between two boundary points of the considered domain. This error is mostly due to the fact that the most intense rainfall patterns are far apart from each other in the observations and the forecast.

3 Analysis of the reforecast HPE representation

An SAL verification score has been applied to the reforecast dataset to perform statistical analysis of QPF (Quantitative Precipitation Forecast) errors. The reforecast dataset is considered as a testbed model in order to study sources of systematic errors in

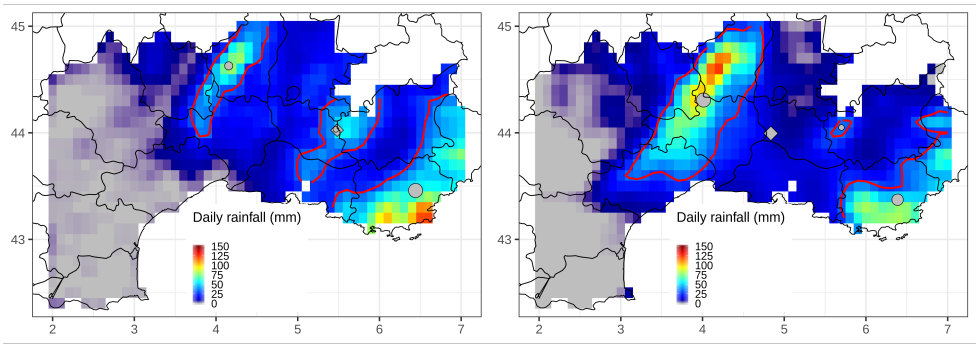


Figure 4. SAL pattern analysis for the case of 28 October 2004, applied on the observation data (left panel), and one 60-hour lead time forecast (right panel). Base contour of the identified objects are in red lines. Gray points stand for the rain barycenter of each pattern, gray diamond depicts the rain barycenter for the whole field. The size of the barycenter points is proportional to the integrated mass of the associated object.

Table 3. Contingency table computed for rainy and dry days.

Contingency table	Obs rainy day	Obs dry day
Model rainy day	3258	84
Model dry day	226	62

the forecast. The overall reforecast performance is first examined for HPE/non-HPE, then according to the clusters. In a second step, the behaviour of the different physics schemes is analyzed by separately considering the SAL results of each reforecast member. Similarly, the analysis is again allocated to HPE/non-HPEs and subsequently to each cluster.

For both the reforecast and the reference, we set all the days with at least one grid point beyond 0.1 mm as a rainy day. In order to facilitate the comparison between the parametrizations, SAL verification is only performed when all the members and the reference are classified as rainy day. Table 3 shows the contingency table of the rainy and dry days. Therefore 84 false alarms, 226 missed cases, and 62 correctly forecast dry days are not involved in the SAL analysis. No HPE belong to the misses and no simulated HPE belong to the false alarms. The SAL measure is then applied to the 3258 rainy days.

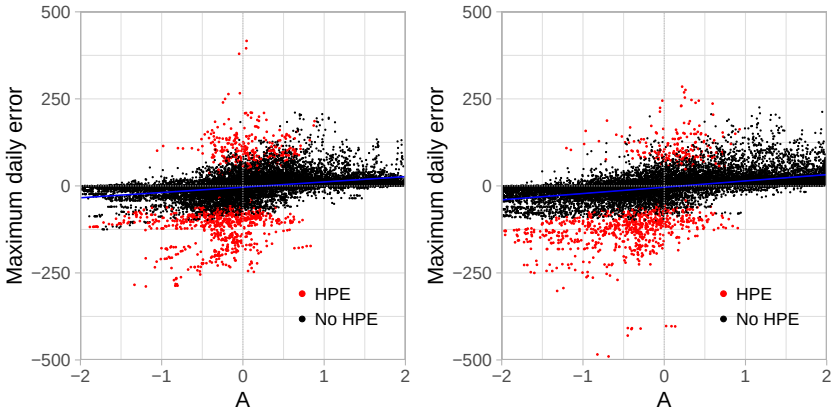


Figure 5. Relationship between the daily rainfall gridpoint maximum algebric error and the A-component of the SAL score. HPEs days are plotted in red, while other days are in black. Left panel is for LT12 lead time, right panel shows LT34 lead time. Linear regression analysis is added to the plot.

3.1 SAL Evaluation of the HPE forecast

3.1.1 HPE/non-HPE

First the relationship between the A-component of SAL and the maximum grid-point error is investigated (Fig. 5). 36-hour and 60-hour lead times (LT12 hereafter) and 84-hour and 108-hour lead times (LT34 hereafter) are grouped together. Maximum daily absolute errors range between -250 mm and 250 mm. Rare higher values are observed, which are likely related to strong double penalty effects that often occur in gridpoint-to-gridpoint verification. Points are mostly scattered along the amplitude axis showing that the error dependence on A-component is weak. Concerning HPEs, the scatter plot shows A-component values under 1, which means that the scaled average precipitation in the forecast never exceeds three times the observation. In contrast, A-component negative values are predominant, in particular at LT34, in relation with strong underestimations of the domain-averaged rainfall field. Some cases of significant maximum grid-point errors in conjunction with moderate negative A-component must be related to strong location errors. In these cases, the domain-averaged field may be similar to the observed one while the maximum rainfall is spatially deviated. For the non-HPE, we can see that, especially for LT34, the model could significantly overestimate both the A-component and the maximum grid-point error.

The relationship between the different SAL components might help to understand sources of model error. In Fig. 6 the S and A components are drawn for the HPEs only. Perfect scores are reached for the points located on the origin O of the diagram. Very few points are located on the top left-hand quadrant. This indicates that an overestimation of precipitation amplitude associated with too small rainfall objects is rarely observed. The points, especially for LT34, are globally oriented from the bottom left-hand corner to the top right-hand corner. This suggests a linear growth of the A-component as a function of the S-

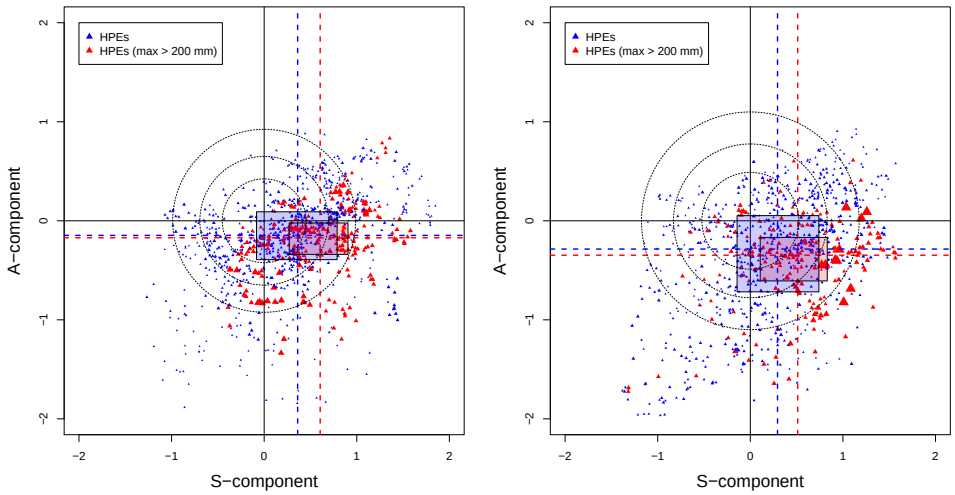


Figure 6. Relationship between the A-component and the S-component of the SAL score (SAL diagrams) for HPEs only, for lead times LT12 (left) and LT34 (right). Blue triangles represent HPEs with gridpoint maximum rainfall under 200 mm/day, and red triangles for rainfall amounts beyond 200 mm. Triangles are proportional to the rainfall value. Some main characteristics of the component distribution are plotted, the median value (dashed lines), percentile 25% and 75% delimitate the boxes. Circles represent the limits 25%, 50% and 75% percentiles to the best score (A=0, S=0).

component, which means that the average rainfall amount is roughly related to the structure of the spatial extension. For the two diagrams, it can also be noticed that many of the points are situated in the lower-right quadrant, suggesting the presence of too large and/or flat rainfall objects compared to the reference while the corresponding A-component is negative. This is supported by the values of the medians of the distribution of the two components (dashed lines) and the quartile values (respective limits of the boxes). The positive bias in the S-component is even stronger for the most extreme HPEs (red triangles). The distortion of S-component error compared to A-component shows that the model has more difficulties reproducing the complex spatial structure than simulating the average volume of a heavy rainfall. This deficiency may be related to the convection part not represented in the parametrization scheme. It may also be related to the representation of orography at a coarse resolution. As shown by Ehmele et al. (2015), an adequate representation of topographic features and local dynamic effects are required to correctly describe the interaction between orography and atmospheric processes. Furthermore, initial conditions have been shown to have a significant influence on rainfall forecasting (Kunz et al., 2018; Khodayar et al., 2018; Caldas-Álvarez et al., 2017).

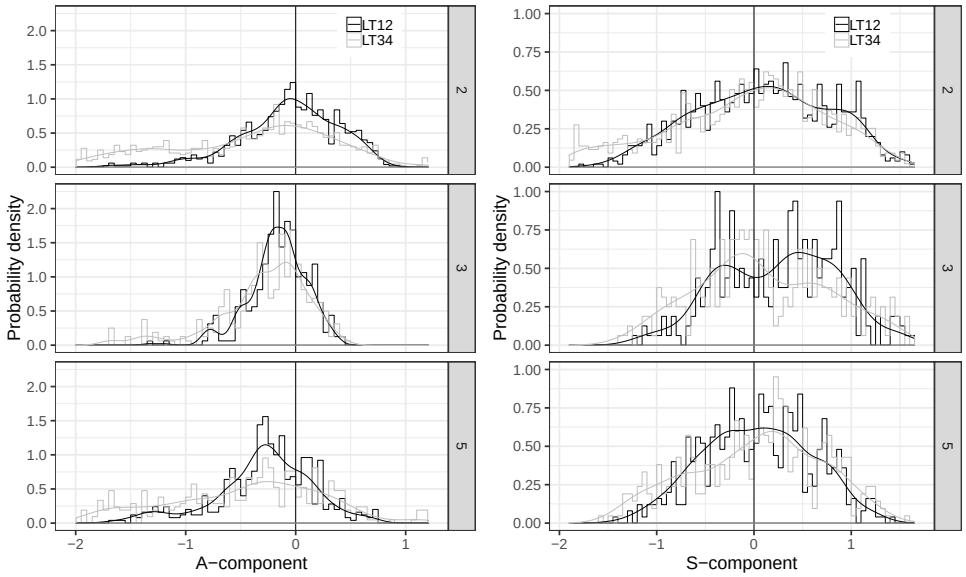


Figure 7. A-component (left column) and S-component (right column) normalized histograms and probability density functions for clusters 2, 3 and 5. Results for lead time LT12 are plotted in black lines and results for lead times LT34 are in grey.

For each point of the diagram in Fig. 6 we compute its distance from the origin (perfect score (A=0; S=0)). The dotted circles respectively contain the 25%, 50% and 75% points with the smallest distance. The radius of the circles are much larger for LT34, confirming a degradation of the scores for longer lead times.

3.1.2 Clusters

We use our clustering procedure (as defined in Section 2.2.2) to analyze the characteristics of the forecast QPF errors along with the regional properties. SAL components are stated for each day of each cluster associated with HPEs, i.e. C2, C3 and C5. In Fig. 7, PDFs (Probability Density Functions) are drawn from the corresponding normalized histograms for the two lead times LT12 and LT34. The distributions of the A-component are negatively-skewed for all the clusters. This shows that the model tends to produce too weak domain-averaged rainfall in the case of heavy rainfall. This is even more important for clusters 3 and 5. For long lead times, the distributions are flatter, showing that the left tail of the A-component PDF spreads far away from the perfect score.

The distributions of the S-component (right panels) are positively skewed in cluster 2 and 3, while they are more centered for cluster 5. For all the clusters, the spread of the S-component distributions is less dependent on the lead time, compared to the A-component distributions. It is interesting to examine whether a relationship between the S-component and the intensity of the rainfall can be identified. A Pearson correlation coefficient is computed between the daily mean of S-component estimated

Table 4. Pearson correlation between the daily mean S-component and the maximum daily rainfall for the three cluster classifications. A t-test is applied to the individual correlations. For the three clusters, the null hypothesis (true correlation coefficient is equal to zero) is rejected.

Cluster	LT12	LT34
2	**0.50**	**0.44**
3	**0.59**	**0.50**
5	**0.37**	**0.46**

within the ten members of the reforecast and the maximum observed daily rainfall for each cluster class (Table 4). A positive correlation is found for all three clusters, which corroborates the results from Fig. 6 where HPEs correspond to the highest S-component values. Maximum correlation is found for cluster 3. Although correlations are statistically significant, it is worth noting that values are quite weak (in particular for cluster 5).

3.2 Sensitivity to physical parametrizations

The SAL measure is analysed separately for the ten different physical packages to study corresponding systematic errors. More specifically, we raise the following questions: Do the errors based on an object-quality measure and computed for the different physics implemented in an ensemble system show different rainfall structure properties? Which physical packages are more sensitive to the intense rainfall forecast errors? As in Section 3.1, we first distinguish the results for the HPE group before the cluster ones.

3.2.1 HPEs

Probability density distributions for each SAL component are separately computed for each physics reforecast (Fig. 8), considering only the HPEs. Colours correspond to four categories, depending on the parametrization of the deep convection. The figure highlights that members from each of the two main parametrization schemes (B85 and PCMT) have similar behaviours. Considering the A-component, PCMT members are more centered around zero than B85 at LT12. This effect is higher at LT34, for which B85 and PCMT density distributions are more shifted. At LT34, more events with a positive A-component are associated with PCMT, whereas negative values are more recurrent in B85. The A-component never exceeds +1, but significant underestimations are observed. This range of values stems from the fact that the forecast verification is applied to a subsample of the observation limited to the most extreme events. For these specific events, a model underestimation is more frequent than an overestimation. At short lead times, the separation between the two deep convection schemes is also well established for the S-component (middle left panel), but it becomes mixed up at LT34 (middle right panel). One reason for this behaviour could be that predictability decreases at LT34, so that discrepancies in spatial rainfall structure assigned to the physics families become less identifiable. The S-component is positively skewed in all cases (in particular for the B85 physics at LT12 lead time). This supports the previous analysis of the S-component (Fig. 6 and 7), showing that for intense rainfall, the model mostly produces larger and flatter rainfall signal. The results for the S-component also highlight better skills for PCMT schemes for

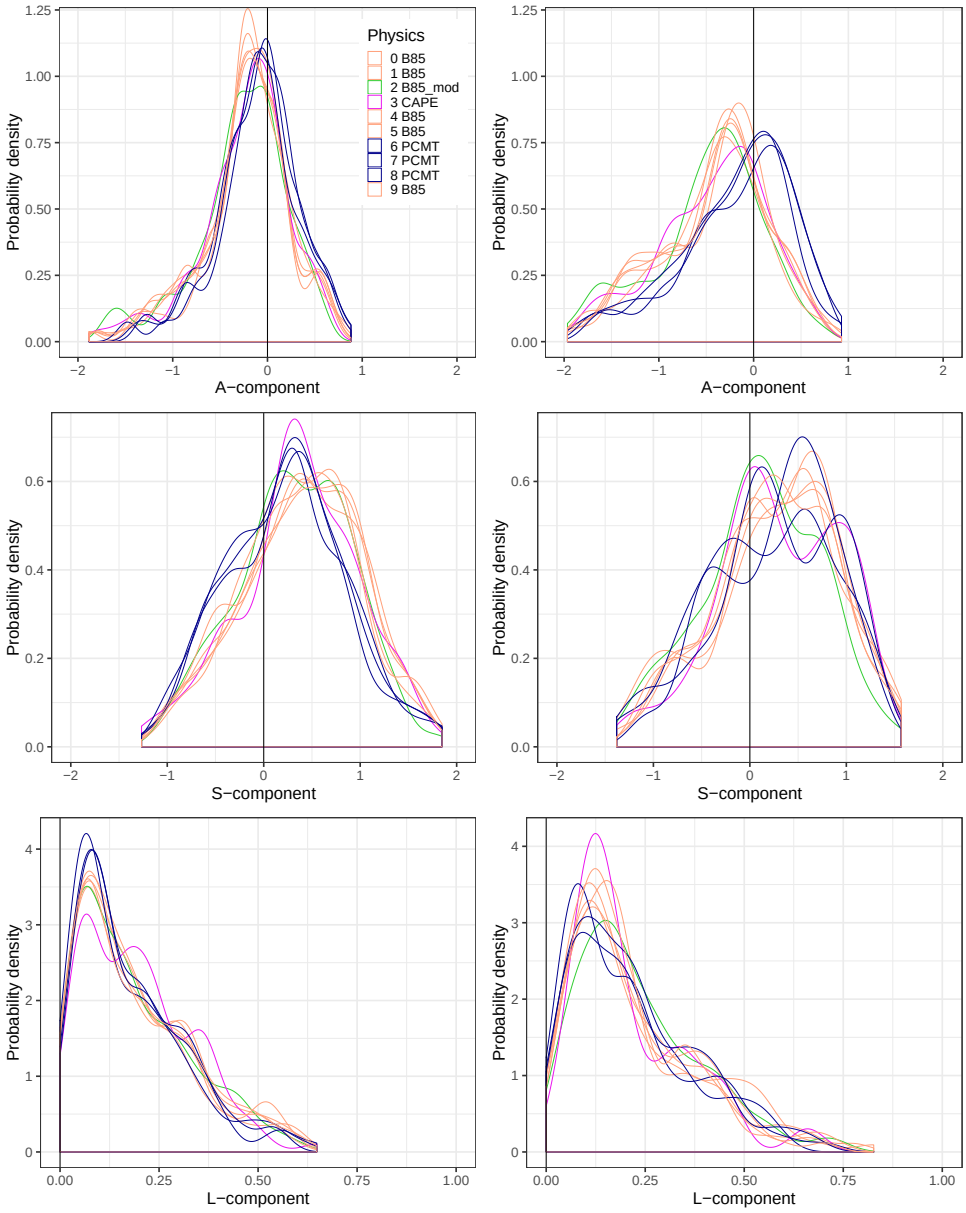


Figure 8. Probability density functions of the three SAL components for the HPEs and for each physics of the reforecast system (colored lines). Physics scheme are gathered in four categories depending on the parametrization of the deep convection: PCMT (blue), B85 (orange), $B85_{mod}$ (green), CAPE (purple). Left column corresponds to lead time LT12, and right column relates to lead time LT34.

HPEs, especially at short lead times. Focusing on high values of S, B85 exhibits a stronger distribution tail at LT12, while both schemes seem comparable for LT34.

For the L-component, the maxima of the density distributions are higher for PCMT at lead time LT12, implying a more significant number of good estimations of pattern location. Regarding the tail of the L-component PDF, it is globally more pronounced at LT34 than LT12. This means that the location of HPEs is poorly forecasted at long lead times. Concerning the behaviour of the forecasts that use the CAPE or $B85_{mod}$ schemes, their A-component PDFs are close to the B85 PDFs. This is not observed for the other components. For the S-component, the CAPE distribution follows the PCMT one at LT12. For the L-component, $B85_{mod}$ PDF is close to the B85 ones, while CAPE shows different behaviour from all the other physics. The use of a closure based on CAPE, rather than on the convergence of humidity seems to modulate the location of precipitation produced by this deep convection parametrization scheme. Moreover, at LT34 CAPE is characterized by a lower number of strong location errors, compared to the other physics.

3.2.2 Clusters

According to the results of the previous Section, which show that the predictability of intense rainfall events is sensitive to the parametrization of the deep convection, we have continued to analyze the model behaviour for the four different deep convection schemes: B85, $B85_{mod}$, CAPE, and PCMT. The link between the behaviour of the physical schemes and belonging to a particular cluster is statistically assessed through the SAL component differences between the schemes.

Any parametric goodness-of-fit tests, which assume normality, have been discarded, because SAL values are not normally distributed. We choose the *k*-sample Anderson–Darling (AD) test (Scholz and Stephens, 1987; Mittermaier et al., 2015), in order to evaluate whether differences between two given distributions are statistically significant. It is an extension of the two-sample test (Darling, 1957), originally developed starting from the Classic Anderson-Darling test (Anderson and Darling, 1952). The *k*-sample AD test is a non parametric test designed to compare continuous or discrete sub-samples of the same distribution. In this case the test is implemented for the evaluation of the pairs of distributions.

The tests are performed for the comparison of each pair of PDFs combined from the four deep convection families and from the three clusters classification. For the A-component, PCMT physics distributions depart significantly from B85 schemes at all lead times, while $B85_{mod}$ and CAPE perform as B85, meaning that the modified versions of B85 weakly affect physics behaviour (not shown).

With respect to the S-component distributions, k-sample AD tests show significant differences between B85 and PCMT physics for LT12, but not for the longest lead times (not shown). At LT34 we observe a convergence of the physics scheme towards a homogeneous distribution, meaning that the differences between physics are negligible.

The test applied to the location component does not reveal significant differences between the PDFs. We suppose that the limited dimensions of the domain employed in this study, as well as its irregular shape, may lead to a less coherent estimation of the location, resulting in a degradation of the score significance. Since the L-component result is not informative about HPEs, it is ignored hereafter.

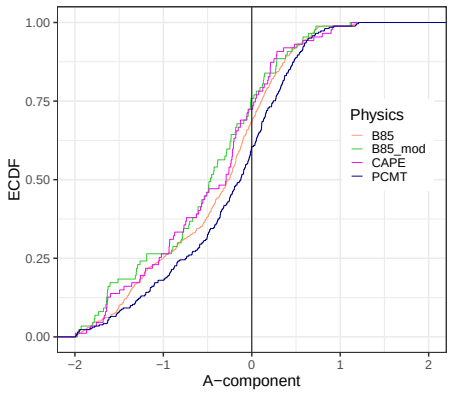


Figure 9. Empirical cumulative distribution function of the A-component computed from cluster 2 at lead time LT34 for the four classes of physics schemes.

Once the statistical differences between the PDFs of the physics have been examined, it is interesting to compare the relative error on the amplitude and structure components. S and A component errors are estimated by comparing the shapes of their distributions. Empirical Cumulative Density Functions (ECDF) of S and A components are computed separately for each cluster and lead time (LT12 and LT34). We show an example of an ECDF for cluster 2 at LT34 (Fig. 9). Forecasts are perfect when the ECDF tends towards a Heaviside step function, which means that the distribution tends towards the Dirac delta function centered on zero. These functions are estimated over a bounded interval, corresponding to the finite range of S and A components. The deviation from the perfect score was quantified, by estimating the area under the ECDF curve on the left side, and the area above the ECDF curve on the right side:

$$err_{-} = \int_{-2}^{0} F(x)dx - \int_{-2}^{0} H(x)dx = \int_{-2}^{0} F(x)dx - 0 = \int_{-2}^{0} F(x)dx, \tag{9}$$

$$err_{+} = \int_{0}^{2} H(x)dx - \int_{0}^{2} F(x)dx = 2 - \int_{0}^{2} F(x)dx, \tag{10}$$

$$err = err_{-} + err_{+} = 2 - \int_{0}^{2} F(x)dx + \int_{-2}^{0} F(x)dx, \tag{11}$$

where $F(x)$ is the ECDF computed for A or S, $H(x)$ is the Heaviside step function and err is the forecast error for a given component. The lower and upper boundaries of the integrals are equal to -2 and +2, because A and S components range between these two values by construction. Since the previous *k*-sample AD test highlighted significant differences within the two main classes B85 and PCMT, the evaluation of the errors is limited to these two specific classes.

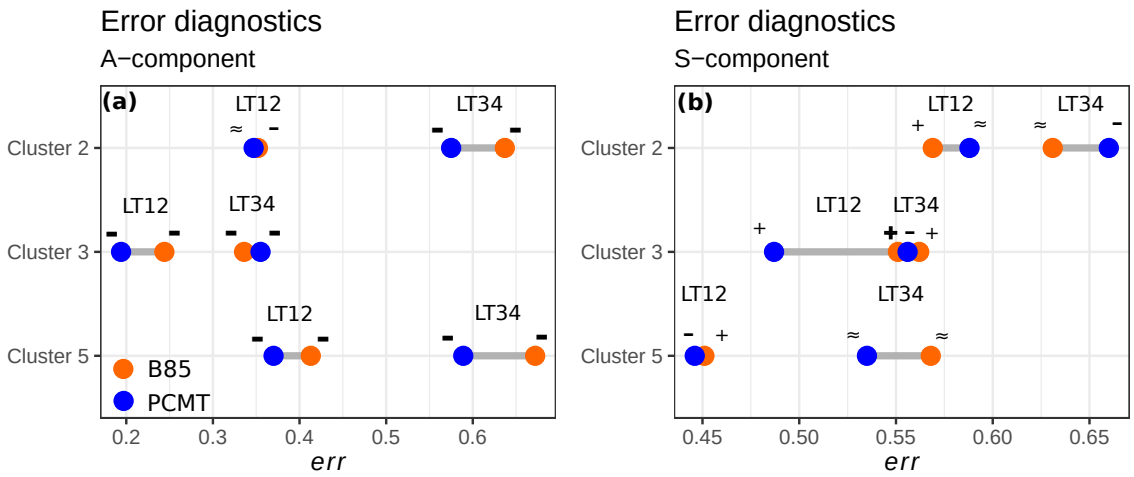


Figure 10. Dumbbell plot of integrated error diagnostics computed using Eq. 11. Colours refer to B85 (orange) and PCMT (blue) deep convection parametrization schemes. Results are stratified on the basis of the clusters and lead times. Symbols denote whether positive or negative errors dominate. These signs are defined using the following definition: $-$ (bold) if $\frac{err_-}{err_+} \geq 2$; $-$ if $1.1 \leq \frac{err_-}{err_+} < 2$; $\approx$ if $0.9 < \frac{err_-}{err_+} < 1.1$; $+$ if $0.5 < \frac{err_-}{err_+} \leq 0.9$; $+$ (bold) if $\frac{err_-}{err_+} \leq 0.5$.

The results of the error diagnostic err for the the A-component are shown in Fig. 10a. Errors increase with lead time. We note that the negative errors are always at least twice as large as the positive ones. Forecasted averaged rainfall amounts are almost always underestimated. PCMT produces overall better A-component statistics, except for cluster 3 at LT34. It is interesting to observe that the weakest errors are associated with cluster 3, which is the most extreme one. Since cluster 3 collects a large number of precipitation events impacting the Cévennes chain, we may suppose that the domain averaged rainfall amounts are more predictable in situations of precipitation driven by the orography. Concerning the S-component evaluation (see Fig. 10b), structures of rainfall patterns are better forecasted for heavy rainfall events (clusters 3 and 5), than for the remaining classes of events. In contrast to the A-component, the S-component exhibits the highest err_+ for B85 scheme for most of the cases (majority of + sign in Fig. 10(b)), whereas this trend is not systematic for PCMT physics. PCMT globally performs better than B85, except for cluster 2. As with the amplitude A, the S-component gets worse for longer lead times, resulting in a shift to larger err_- for both B85 and PCMT physics (more - sign for LT34 in Fig. 10a, b). The lowest errors of S-component are achieved for cluster 5. Cluster 5 HPEs are known to have specific regional properties whose influence on S-component results should be studied with further diagnostics.

3.2.3 Rainfall object analysis

We now analyze the physical properties of the objects, i.e. the number of objects from a rainfall field and the object integrated volumes, according to the different clusters. All the statistics are applied separately to the B85, PCMT physics, and observations. For each day of the dataset period, the thresholds defined in subsection 2.3.1 lead to the identification of a certain number of precipitating objects. The frequency of the number of objects per day is plotted by means of normalized histograms for the three clusters (Fig. 11). Clusters 2 and 3 show maximum frequency for one and three object range, whereas cluster 5 is dominated by one object per day. This specific property of cluster 5 can explain the best result obtained for S-component (Section 3.2.2). Indeed, we may assume that S-component estimation is more accurate for a one-to-one object comparison. The other clusters frequently display rainfall accumulated bands split over the domain, typically over the Cévennes and Alpine regions. Object identification for PCMT forecast shows that there is an overestimation of single object days compared to the observation and to B85 physics scheme, a behaviour emphasized in clusters 3 and 5.

More details about the magnitude of the objects can be produced by computing the integrated mass per object, M_k (see subsection 2.3.1). First, for each day, objects are sorted from the largest to the smallest integrated mass. Integrated mass distribution of the two heaviest objects (noted O_1 and O_2) are then dispatched as a function of the number of objects for each cluster on Fig. 12. First, the range value of M is highly variable from one cluster to another. Maximum values are observed for cluster 3, while the magnitude for clusters 2 and 5 is comparable. The decrease of the mass for O_1 is clearer for cluster 3, meaning that a high number of objects over the domain leads to a natural decrease of the M value of the heaviest ones. We think that a part of the total integrated mass is then redistributed to the other objects. This is confirmed by O_2 curves since its mass increases with the number of objects. Conversely, for cluster 5, O_1 mass increases with the number of the objects, while O_2 is almost stable. The gap between O_1 and O_2 masses is maximum in the most extreme clusters (3 and 5). This suggests that when computing the volume V (see Eq. 7) and L2 (see Eq. 4), the weighted average is dominated by the object O_1. This implies that the verification could be considered as a single to single object metric.

We now examine the ratio between the daily maximum rainfall of objects O_1 and O_2. This ratio ranges between 1.5 and 3 which means that O_1 represents the essential contribution of the daily rainfall peak. Since O_1 base area tends to be significantly larger that O_2, the information related to the inner object maximum rainfall is diluted in the large base area, resulting in a flat weak mean intensity of the object. This last result appears to support the fact that SAL metric gives more weight to the object that contains the most intense rainfall.

The comparison between the model reforecast physics and the observations is addressed using the whole distribution of daily mass M from the objects O_i identified across the full reforecast dataset, where i ranges between 1 and the total number N of objects. We proceed separately for each physical package. For a given scheme and cluster, the quantile values corresponding to the selected dataset are sorted in ascending order, and then plotted versus the quantiles calculated from observations (Fig. 13). Half of the quantile distributions are not visible as they correspond to very weak pattern masses. For cluster 2 and PCMT physics most of the distribution of object mass is close to the observations, however all other physics distributions are skewed to the right compared to the observations for values below 10000 mm. This behaviour is also observed for cluster 5 and it involves

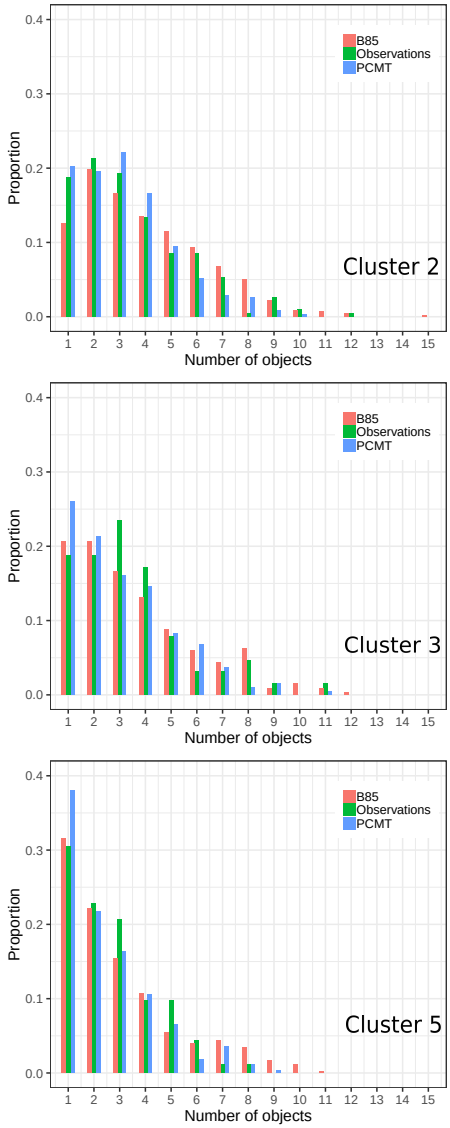


Figure 11. Normalized histograms of the daily number of SAL patterns, for B85 physics scheme (red), PCMT (blue), observation (green). Panels correspond to the 3 clusters classification.

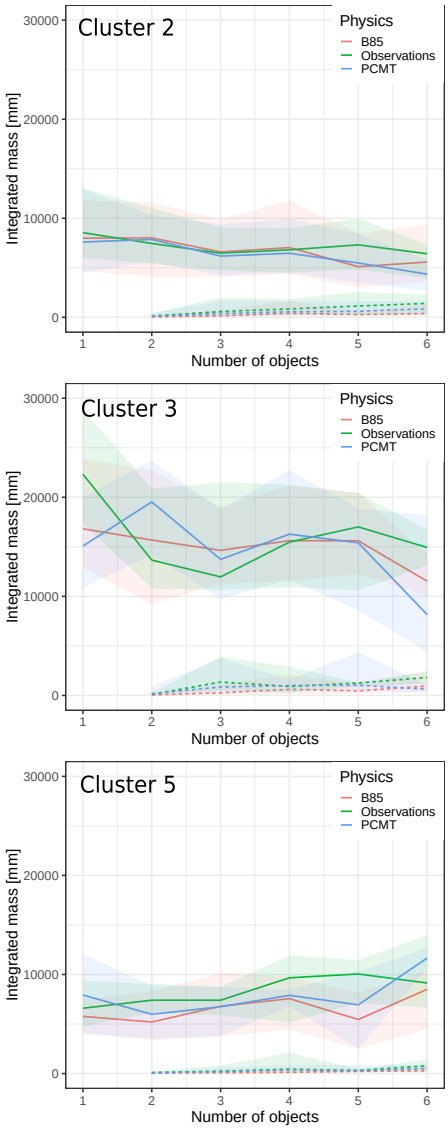


Figure 12. Distribution of SAL first pattern O_1 rain amount according to the number of patterns per day. Curves stand for the median of the distribution, shaded areas range between 25% and 75% percentiles. The dashed lines correspond to the second ranked SAL pattern O_2 rain amount.

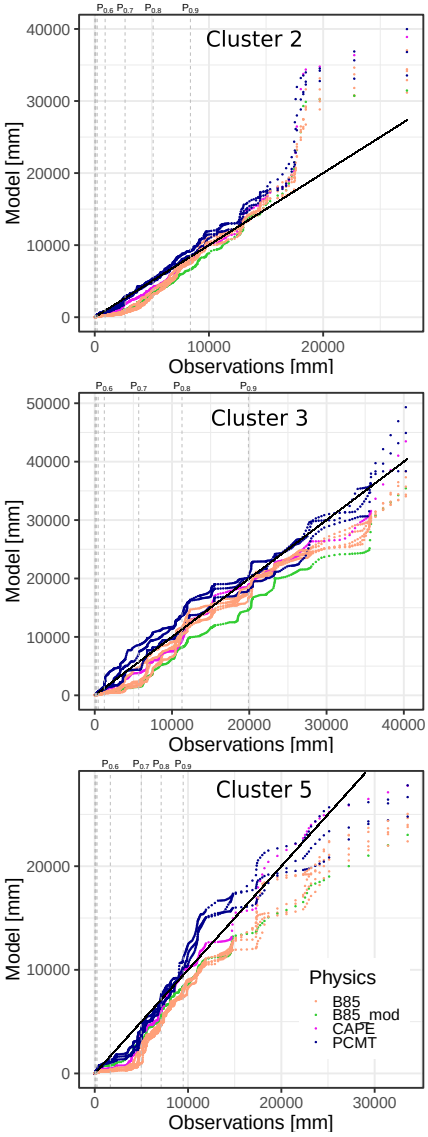


Figure 13. Quantile-quantile plot between SAL pattern rain amounts from the model (Y-axis) and from the observation (X-axis). Physics schemes are gathered into 4 classes (B85, PCMT, $B85_{mod}$, CAPE). Observation deciles correspond to the vertical dashed lines.

PCMT physics as well, for values between percentile 0.5 and percentile 0.7. Overall, in the quantile-quantile plot for cluster 5, the PCMT outperforms B85. In cluster 3, discrepancies between PCMT, B85, and the observations are of opposite sign, with PCMT being slightly above the observations, while B85 showing a weak underestimation. CAPE physics distribution is left skewed compared to the observations and to the other physics. These results highlight some interesting properties of the models in predicting the rainfall objects. Except for some deviation concerning a few extreme cases of cluster 2 and a small portion of distributions of cluster 5, object mass distribution of physics is similar to the distribution drawn from the observation, especially for cluster 3. This means that the forecast is able to reproduce the same proportion of rainfall amounts inside a feature as the observations, even concerning the extreme right tail of the distributions, which corresponds to the major events of the series.

4 Summary and conclusions

In this study we have characterized the systematic errors of 24-hour rainfall amounts from a reforecast ensemble dataset, covering a 30-year fall period. A 24-hour rainfall observation reference has been produced on a regular grid with a resolution identical to the model in order to run point-to-point verification. We applied an object-based quality measure in order to evaluate the performance of the forecasts of any kind of HPE. Then, we took advantage of a rainfall clustering to analyze the dependence of systematic errors on clusters.

The selection of the HPEs within the reference dataset was based on a peak-over-threshold approach. The spatial regional discrepancies between HPEs are highlighted on the basis of the k-means clustering of the 24-hour rainfall. Finally, we analyzed the rainfall object properties respectively in the model and in the observation to underline the rainfall field object properties for which the model acts distinctly.

The peak-over-threshold criterion lead to the selection of 192 HPEs, confirming that the most impacted regions are the Cévennes area and part of the Alps. The composite analysis for the five clusters shows that each cluster is associated with a specific class and location of 24-hour precipitation events. It was found that 86% of the number of HPEs are included in clusters 2, 3 and 5. Cluster 2 and 3 HPEs predominantly impact the Cévennes and Alps area, while cluster 5 HPEs are located over the Languedoc-Roussillon region. Moreover clusters 3 and 5 include the most extreme ones. Only diagnostics for clusters 2, 3 and 5 are considered.

The SAL object-quality measure has been applied distinctly to the ten physics schemes (one per member) of the reforecast dataset and compared to the rainfall reference. It shows that the model's overall behaviour for HPE forecasting is characterized by negative A-components and positive S-components. As in grid-point rainfall verification, all the SAL components get worse as a function of lead time. The model HPE rainfall objects tend to be more extended and less peaked. Even though their corresponding domain-average amplitude is weaker, it does not mean that the event maximum intensity is always weaker. This result is important showing to modelers that even for intense rainfall events when orography interaction and quasi-stationarity meso-scale systems play a great role, the model tends to reproduce rainfall patterns with greater extension, rather than both smaller extension and weaker intensity patterns.

In order to show regional disparities in the model behaviour, the SAL diagnostics have been divided according to the clusters and it shows interesting results. First, the A component negative contribution for the whole sample is higher, showing that in average more underestimation than overestimation is observed for the Amplitude SAL-component. It is notably the case for the most extreme clusters (over the Cévennes and over the Languedoc-Roussillon). However, when considering both positive and negative contributions to the integrated A-component, the most extreme cluster (cluster 3) leads to better scores. This could mean that the variability of the A-component is postively reduced for the most intense events. This is quite surprising and could reinforce the role of orography in this error decrease. As for the S-component distribution, we showed it is slightly positively skewed for cluster 2 and 3, while for cluster 5 the distribution of the S-component is more centered. Likewise for the A-component the integrated balance of positive and negative S-component contributions lead to better results for cluster 3 and 5. It is even more remarkable for cluster 5, for which the S-component reaches the best score. Though it is difficult at this point to determine whether this characterizes an actual contrast in the model behaviour or if it is due to the physical properties of the cluster 5 events. One hypothesis could be related to the large number of single objects characterizing this cluster.

The impact of the different physics schemes has also been investigated, and it mostly emphasized the role of the deep convection physical parameterization. Considering the SAL diagnostics, the two main deep convection schemes, B85 and PCMT, clearly determine the behaviour of the model in HPE forecasting until lead time ranges longer than three days, after which no significant differences appear. This difference is clearly in favour of the PCMT scheme which performs better than B85 for both SAL A and S components and in the majority of the subsampled scores considering the HPEs or the regional clusters. However, this PCMT asset is not huge, and both physics schemes can contribute to good or bad forecasts. The main significant difference is for the S-component for the most intense rainfall, which shows that PCMT better approximates the structure of the rainfall patterns in these cases.

In light of the ability of our method to produce significant results even after several subsampling steps, we decided to study further statistical characterization of the SAL rainfall objects. It has been shown that in most cases, one large object stands out among other smaller objects, which often gathers the most part of the rain signal. For cluster 5, characterized by the Languedoc-Roussillon HPEs, the rainfall distribution could even be considered as a single object rainfall field. Then we focused on the ranked distributions (quantile-quantile analysis) of the object masses to compare the overall rainfall climatology of the model with the reference. First, this analysis showed that in particular the weakest precipitation are overestimated by all physics schemes. However, looking at the object mass distributions for the whole period, we find they are relatively close between all the physics schemes and the observation for most extreme rainfall events, especially for the PCMT deep convection scheme. This statistical result implies that a global model should be able to reproduce a reliable distribution of rainfall objects along a long time period, e.g. the climate of the model and of the observations are close to each other. Therefore, in the case of PEARP, most of the forecast errors are mainly related to a low consistency between observed and forecasted fields, rather than to an inability of the prediction system to produce intense precipitation amounts.

This last result, objectively quantified for high rainfall event thresholds (around 100 mm to 500 mm) on a long enough period, is important for two reasons. The first one concerns atmospheric modelers, showing that the physics schemes are able to reproduce climatological distributions of the most challenging rainfall events. On this basis, future research could investigate

other sources of uncertainties like from the analysis setup and implement ensuing model improvements. The model physics perturbation technique should then play a greater role in the control of the ensemble dispersion. In this perspective, the novel reanalysis ERA5 would be interesting to use, in particular its perturbed members, to improve the uncertainty from initial conditions in the reforecast. The second lesson to be learned from this study is that it is worth focusing on the study of a model behaviour on intense events forecasting as it provides important learning to ensemble model end-users, in particular in the context of decision making based on weather forecast. Quantifying systematic errors could also be used to favorably improve their inclusion in nested forecast tools processes.

In terms of methodology, this study also highlights that the combination of SAL verification and clustering is a relevant approach to show systematic errors associated with regional features for intense precipitation forecasting. This achievement is only enabled by the availability of a long reforecast dataset. This methodology could be further extended to a different model and another geographic region, on the condition of sampling a large number of HPEs.

The inter-comparison between some model physics deep convection schemes and their role in HPEs predictability shows it is of course very sensitive for designing multi-physics type of ensemble forecasting systems. While the sensitivity to the initial perturbations was not studied in this work, the forecast of intense rainfall seems to be mainly driven by the classes of deep convection parametrizations. Since physical parametrization set-up is built by replicated schemes, the model error representation might lack an exhaustive sampling of the forecasted trajectories. Using more than two deep convection parametrization schemes may improve the representation of model errors, at least for heavy precipitation events.

Data availability. Research data can be accessed by contacting Matteo Ponzano at his e-mail address matteo.ponzano@meteo.fr and the other authors.

Author contributions. MP, BJ, and LD conceived and designed the study. MP carried out the formal analysis, wrote the whole paper, made the literature review, and produced the observation reference dataset. BJ built the hindcast dataset. BJ, LD, and PA reviewed and edited the original draft.

Competing interests. The authors declare that they have no conflict of interest.

References

AghaKouchak, A., Behrangi, A., Sorooshian, S., Hsu, K., and Amitai, E.: Evaluation of Satellite-Retrieved Extreme Precipitation Rates across the Central United States, Journal of Geophysical Research: Atmospheres, 116, https://doi.org/10.1029/2010JD014741, 2011.

Anagnostopoulou, C. and Tolika, K.: Extreme Precipitation in Europe: Statistical Threshold Selection Based on Climatological Criteria, Theoretical and Applied Climatology, 107, 479–489, https://doi.org/10.1007/s00704-011-0487-8, 2012.

Anderson, T. W. and Darling, D. A.: Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes, The Annals of Mathematical Statistics, 23, 193–212, https://doi.org/10.1214/aoms/1177729437, 1952.

Argence, S., Lambert, D., Richard, E., Chaboureau, J.-P., and Söhne, N.: Impact of Initial Condition Uncertainties on the Predictability of Heavy Rainfall in the Mediterranean: A Case Study, Quarterly Journal of the Royal Meteorological Society, 134, 1775–1788, https://doi.org/10.1002/qj.314, 2008.

Bazile, E., Marquet, P., Bouteloup, Y., and Bouyssel, F.: The Turbulent Kinetic Energy (TKE) scheme in the NWP models at Meteo France, in: Workshop on Workshop on Diurnal cycles and the stable boundary layer, 7-10 November 2011, pp. 127–135, ECMWF, ECMWF, Shinfield Park, Reading, 2012.

Bechtold, P., Bazile, E., Guichard, F., Mascart, P., and Richard, E.: A Mass-Flux Convection Scheme for Regional and Global Models, Quarterly Journal of the Royal Meteorological Society, 127, 869–886, https://doi.org/10.1002/qj.49712757309, 2001.

Belamari, S.: Report on uncertainty estimates of an optimal bulk formulation for surface turbulent fluxes, MERSEA IP Deliverable 412, pp. 1–29, 2005.

Boisserie, M., Descamps, L., and Arbogast, P.: Calibrated Forecasts of Extreme Windstorms Using the Extreme Forecast Index (EFI) and Shift of Tails (SOT), Weather and Forecasting, 31, 1573–1589, https://doi.org/10.1175/WAF-D-15-0027.1, 2015.

Boisserie, M., Decharme, B., Descamps, L., and Arbogast, P.: Land surface initialization strategy for a global reforecast dataset, Quarterly Journal of the Royal Meteorological Society, 142, 880–888, https://doi.org/10.1002/qj.2688, http://rmets.onlinelibrary.wiley.com/doi/full/10.1002/qj.2688, 2016.

Bougeault, P.: A Simple Parameterization of the Large-Scale Effects of Cumulus Convection, Monthly Weather Review, 113, 2108–2121, https://doi.org/10.1175/1520-0493(1985)113<2108:ASPOTL>2.0.CO;2, 1985.

Buizza, R. and Palmer, T. N.: The Singular-Vector Structure of the Atmospheric Global Circulation, Journal of the Atmospheric Sciences, 52, 1434–1456, https://doi.org/10.1175/1520-0469(1995)052<1434:TSVSOT>2.0.CO;2, 1995.

Caldas-Álvarez, A., Khodayar, S., and Bock, O.: GPS – Zenith Total Delay assimilation in different resolution simulations of a heavy precipitation event over southern France, Advances in Science and Research, 14, 157–162, https://doi.org/10.5194/asr-14-157-2017, https://www.adv-sci-res.net/14/157/2017/, 2017.

Charron, M., Pellerin, G., Spacek, L., Houtekamer, P. L., Gagnon, N., Mitchell, H. L., and Michelin, L.: Toward Random Sampling of Model Error in the Canadian Ensemble Prediction System, Monthly Weather Review, 138, 1877–1901, https://doi.org/10.1175/2009MWR3187.1, 2009.

Collier, C. G.: Flash Flood Forecasting: What Are the Limits of Predictability?, Quarterly Journal of the Royal Meteorological Society, 133, 3–23, https://doi.org/10.1002/qj.29, 2007.

Courtier, P., Freydier, C., Geleyn, J., Rabier, F., and Rochas, M.: The ARPEGE project at Météo-France, ECMWF Seminar proceedings, vol. II. ECMWF Reading, UK, pp. 193–231, 1991.

Cuxart, J., Bougeault, P., and Redelsperger, J.-L.: A turbulence scheme allowing for mesoscale and large-eddy simulations, Quarterly Journal of the Royal Meteorological Society, 126, 1–30, https://doi.org/10.1002/qj.49712656202, 2000.

Darling, D. A.: The Kolmogorov-Smirnov, Cramer-von Mises Tests, The Annals of Mathematical Statistics, 28, 823–838, https://doi.org/10.1214/aoms/1177706788, 1957.

Davis, C., Brown, B., and Bullock, R.: Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas, Monthly Weather Review, 134, 1772–1784, https://doi.org/10.1175/MWR3145.1, 2006a.

Davis, C., Brown, B., and Bullock, R.: Object-Based Verification of Precipitation Forecasts. Part II: Application to Convective Rain Systems, Monthly Weather Review, 134, 1785–1795, https://doi.org/10.1175/MWR3146.1, 2006b.

Davis, C. A., Brown, B. G., Bullock, R., and Halley-Gotway, J.: The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program, Weather and Forecasting, 24, 1252–1267, https://doi.org/10.1175/2009WAF2222241.1, 2009.

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F.: The ERA-Interim Reanalysis: Configuration and Performance of the Data Assimilation System, Quarterly Journal of the Royal Meteorological Society, 137, 553–597, https://doi.org/10.1002/qj.828, 2011.

Delrieu, G., Nicol, J., Yates, E., Kirstetter, P.-E., Creutin, J.-D., Anquetin, S., Obled, C., Saulnier, G.-M., Ducrocq, V., Gaume, E., Payrastre, O., Andrieu, H., Ayral, P.-A., Bouvier, C., Neppel, L., Livet, M., Lang, M., du-Châtelet, J. P., Walpersdorf, A., and Wobrock, W.: The Catastrophic Flash-Flood Event of 8–9 September 2002 in the Gard Region, France: A First Case Study for the Cévennes–Vivarais Mediterranean Hydrometeorological Observatory, Journal of Hydrometeorology, 6, 34–52, https://doi.org/10.1175/JHM-400.1, 2005.

Descamps, L., Labadie, C., and Bazile, E.: Representing model uncertainty using the multiparametrization method, in: Workshop on Representing Model Uncertainty and Error in Numerical Weather and Climate Prediction Models, 20-24 June 2011, pp. 175–182, ECMWF, ECMWF, Shinfield Park, Reading, https://www.ecmwf.int/node/9015, 2011.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., and Cébron, P.: PEARP, the Météo-France Short-Range Ensemble Prediction System, Quarterly Journal of the Royal Meteorological Society, 141, 1671–1685, https://doi.org/10.1002/qj.2469, 2015.

Du, J., Mullen, S. L., and Sanders, F.: Short-Range Ensemble Forecasting of Quantitative Precipitation, Monthly Weather Review, 125, 2427–2459, https://doi.org/10.1175/1520-0493(1997)125<2427:SREFOQ>2.0.CO;2, 1997.

Ducrocq, V., Aullo, G., and Santurette, P.: The extreme flash flood case of November 1999 over Southern France, La Météorologie, 42, 18–27, 2003.

Ebert, E. E. and McBride, J. L.: Verification of Precipitation in Weather Systems: Determination of Systematic Errors, Journal of Hydrology, 239, 179–202, https://doi.org/10.1016/S0022-1694(00)00343-7, 2000.

Ehmele, F., Barthlott, C., and Corsmeier, U.: The influence of Sardinia on Corsican rainfall in the western Mediterranean Sea: A numerical sensitivity study, Atmospheric Research, 153, 451 – 464, https://doi.org/https://doi.org/10.1016/j.atmosres.2014.10.004, http://www.sciencedirect.com/science/article/pii/S0169809514003731, 2015.

Erdin, R., Frei, C., and Künsch, H. R.: Data Transformation and Uncertainty in Geostatistical Combination of Radar and Rain Gauges, Journal of Hydrometeorology, 13, 1332–1346, https://doi.org/10.1175/JHM-D-11-096.1, 2012.

Frei, C. and Schär, C.: A Precipitation Climatology of the Alps from High-Resolution Rain-Gauge Observations, International Journal of Climatology, 18, 873–900, https://doi.org/10.1002/(SICI)1097-0088(19980630)18:8<873::AID-JOC255>3.0.CO;2-9, 1998.

G. Gregoire, T., Lin, Q. F., Boudreau, J., and Nelson, R.: Regression Estimation Following the Square-Root Transformation of the Response, Forest Science, 54, 597–606, 2008.

Goovaerts, P. et al.: Geostatistics for natural resources evaluation, Oxford University Press on Demand, 1997.

Hamill, T. M.: Verification of TIGGE Multimodel and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous United States, Monthly Weather Review, 140, 2232–2252, https://doi.org/10.1175/MWR-D-11-00220.1, 2012.

Hamill, T. M. and Whitaker, J. S.: Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application, Monthly Weather Review, 134, 3209–3229, https://doi.org/10.1175/MWR3237.1, 2006.

Hamill, T. M., Hagedorn, R., and Whitaker, J. S.: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation, Monthly Weather Review, 136, 2620–2632, https://doi.org/10.1175/2007MWR2411.1, 2008.

Houtekamer, P. L. and Mitchell, H. L.: Data Assimilation Using an Ensemble Kalman Filter Technique, Monthly Weather Review, 126, 796–811, https://doi.org/10.1175/1520-0493(1998)126<0796:DAUAEK>2.0.CO;2, 1998.

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H., and Mitchell, H. L.: A System Simulation Approach to Ensemble Prediction, Monthly Weather Review, 124, 1225–1242, https://doi.org/10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2, 1996.

Kai, T., Zhong-Wei, Y., and Yi, W.: A Spatial Cluster Analysis of Heavy Rains in China, Atmospheric and Oceanic Science Letters, 4, 36–40, https://doi.org/10.1080/16742834.2011.11446897, 2011.

Kain, J. S. and Fritsch, J. M.: Convective Parameterization for Mesoscale Models: The Kain-Fritsch Scheme, in: The Representation of Cumulus Convection in Numerical Models, edited by Emanuel, K. A. and Raymond, D. J., Meteorological Monographs, pp. 165–170, American Meteorological Society, Boston, MA, https://doi.org/10.1007/978-1-935704-13-3_16, 1993.

Khodayar, S., Czajka, B., Caldas-Alvarez, A., Helgert, S., Flamant, C., Di Girolamo, P., Bock, O., and Chazette, P.: Multi-scale observations of atmospheric moisture variability in relation to heavy precipitating systems in the northwestern Mediterranean during HyMeX IOP12, Quarterly Journal of the Royal Meteorological Society, 144, 2761–2780, 2018.

Kunz, M., Blahak, U., Handwerker, J., Schmidberger, M., Punge, H. J., Mohr, S., Fluck, E., and Bedka, K. M.: The severe hailstorm in southwest Germany on 28 July 2013: characteristics, impacts and meteorological conditions, Quarterly Journal of the Royal Meteorological Society, 144, 231–250, 2018.

Lack, S. A., Limpert, G. L., and Fox, N. I.: An Object-Oriented Multiscale Verification Scheme, Weather and Forecasting, 25, 79–92, https://doi.org/10.1175/2009WAF2222245.1, 2010.

Lalaurette, F.: Early detection of abnormal weather conditions using a probabilistic extreme forecast index, Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 129, 3037–3057, 2003.

Lin, Y.-L., Chiao, S., Wang, T.-A., Kaplan, M. L., and Weglarz, R. P.: Some Common Ingredients for Heavy Orographic Rainfall, Weather and Forecasting, 16, 633–660, https://doi.org/10.1175/1520-0434(2001)016<0633:SCIFHO>2.0.CO;2, 2001.

Little, M. A., Rodda, H. J. E., and McSharry, P. E.: Bayesian Objective Classification of Extreme UK Daily Rainfall for Flood Risk Applications, Hydrology and Earth System Sciences Discussions, 5, 3033–3060, https://doi.org/https://doi.org/10.5194/hessd-5-3033-2008, 2008.

Louis, J.-F.: A Parametric Model of Vertical Eddy Fluxes in the Atmosphere, Boundary-Layer Meteorology, 17, 187–202, https://doi.org/10.1007/BF00117978, 1979.

Ly, S., Charles, C., and Degré, A.: Geostatistical Interpolation of Daily Rainfall at Catchment Scale: The Use of Several Variogram Models in the Ourthe and Ambleve Catchments, Belgium, Hydrol. Earth Syst. Sci., 15, 2259–2274, https://doi.org/10.5194/hess-15-2259-2011, 2011.

Ly, S., Charles, C., and Degré, A.: Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale: a review, BASE, 2013.

Mass, C. F., Ovens, D., Westrick, K., and Colle, B. A.: Does Increasing Horizontal Resolution Produce More Skillful Forecasts?, Bulletin of the American Meteorological Society, 83, 407–430, https://doi.org/10.1175/1520-0477(2002)083<0407:DIHRPM>2.3.CO;2, 2002.

Masson, V., Le Moigne, P., Martin, E., Faroux, S., Alias, A., Alkama, R., Belamari, S., Barbu, A., Boone, A., Bouyssel, F., Brousseau, P., Brun, E., Calvet, J.-C., Carrer, D., Decharme, B., Delire, C., Donier, S., Essaouini, K., Gibelin, A.-L., Giordani, H., Habets, F., Jidane, M., Kerdraon, G., Kourzeneva, E., Lafaysse, M., Lafont, S., Lebeaupin Brossier, C., Lemonsu, A., Mahfouf, J.-F., Marguinaud, P., Mokhtari, M., Morin, S., Pigeon, G., Salgado, R., Seity, Y., Taillefer, F., Tanguy, G., Tulet, P., Vincendon, B., Vionnet, V., and Voldoire, A.: The SURFEXv7.2 land and ocean surface platform for coupled or offline simulation of Earth surface variables and fluxes, Geoscientific Model Development, 6, 929–960, https://doi.org/10.5194/gmd-6-929-2013, https://hal.archives-ouvertes.fr/hal-00968042, 2013.

Mills, G. F.: Principal Component Analysis of Precipitation and Rainfall Regionalization in Spain, Theoretical and Applied Climatology, 50, 169–183, https://doi.org/10.1007/BF00866115, 1995.

Mittermaier, M., North, R., Semple, A., and Bullock, R.: Feature-Based Diagnostic Evaluation of Global NWP Forecasts, Monthly Weather Review, 144, 3871–3893, https://doi.org/10.1175/MWR-D-15-0167.1, 2015.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF Ensemble Prediction System: Methodology and Validation, Quarterly Journal of the Royal Meteorological Society, 122, 73–119, https://doi.org/10.1002/qj.49712252905, 1996.

Morin, G., Fortin, J.-P., Sochanska, W., Lardeau, J.-P., and Charbonneu, R.: Use of Principal Component Analysis to Identify Homogeneous Precipitation Stations for Optimal Interpolation, Water Resources Research, 15, 1841–1850, https://doi.org/10.1029/WR015i006p01841, 1979.

Nachamkin, J. E.: Application of the Composite Method to the Spatial Forecast Verification Methods Intercomparison Dataset, Weather and Forecasting, 24, 1390–1400, https://doi.org/10.1175/2009WAF2222225.1, 2009.

Nuissier, O., Ducrocq, V., Ricard, D., Lebeaupin, C., and Anquetin, S.: A Numerical Study of Three Catastrophic Precipitating Events over Southern France. I: Numerical Framework and Synoptic Ingredients, Quarterly Journal of the Royal Meteorological Society, 134, 111–130, https://doi.org/10.1002/qj.200, 2008.

Nuissier, O., Joly, B., Joly, A., Ducrocq, V., and Arbogast, P.: A Statistical Downscaling to Identify the Large-Scale Circulation Patterns Associated with Heavy Precipitation Events over Southern France, Quarterly Journal of the Royal Meteorological Society, 137, 1812–1827, https://doi.org/10.1002/qj.866, 2011.

Palmer, T., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G., Steinheimer, M., and Weisheimer, A.: Stochastic parametrization and model uncertainty, ECMWF Technical Memorandum, p. 42, https://doi.org/10.21957/ps8gbwbdv, https://www.ecmwf.int/node/11577, 2009.

Peñarrocha, D., Estrela, M. J., and Millán, M.: Classification of Daily Rainfall Patterns in a Mediterranean Area with Extreme Intensity Levels: The Valencia Region, International Journal of Climatology, 22, 677–695, https://doi.org/10.1002/joc.747, 2002.

Pergaud, J., Masson, V., Malardel, S., and Couvreux, F.: A Parameterization of Dry Thermals and Shallow Cumuli for Mesoscale Numerical Weather Prediction, Boundary-Layer Meteorology, 132, 83, https://doi.org/10.1007/s10546-009-9388-0, 2009.

Petroliagis, T., Buizza, R., Lanzinger, A., and Palmer, T. N.: Potential Use of the ECMWF Ensemble Prediction System in Cases of Extreme Weather Events, Meteorological Applications, 4, 69–84, https://doi.org/10.1017/S1350482797000297, 1997.

Piriou, J.-M., Redelsperger, J.-L., Geleyn, J.-F., Lafore, J.-P., and Guichard, F.: An Approach for Convective Parameterization with Memory: Separating Microphysics and Transport in Grid-Scale Equations, Journal of the Atmospheric Sciences, 64, 4127–4139, https://doi.org/10.1175/2007JAS2144.1, 2007.

Ricard, D., Ducrocq, V., and Auger, L.: A Climatology of the Mesoscale Environment Associated with Heavily Precipitating Events over a Northwestern Mediterranean Area, Journal of Applied Meteorology and Climatology, 51, 468–488, https://doi.org/10.1175/JAMC-D-11-017.1, 2011.

Romero, R., Ramis, C., and Guijarro, J. A.: Daily Rainfall Patterns in the Spanish Mediterranean Area: An Objective Classification, International Journal of Climatology, 19, 95–112, https://doi.org/10.1002/(SICI)1097-0088(199901)19:1<95::AID-JOC344>3.0.CO;2-S, 1999.

Rossa, A., Nurmi, P., and Ebert, E.: Overview of Methods for the Verification of Quantitative Precipitation Forecasts, in: Precipitation: Advances in Measurement, Estimation and Prediction, edited by Michaelides, S., pp. 419–452, Springer Berlin Heidelberg, Berlin, Heidelberg, https://doi.org/10.1007/978-3-540-77655-0_16, 2008.

Schär, C., Ban, N., Fischer, E. M., Rajczak, J., Schmidli, J., Frei, C., Giorgi, F., Karl, T. R., Kendon, E. J., Tank, A. M. G. K., O'Gorman, P. A., Sillmann, J., Zhang, X., and Zwiers, F. W.: Percentile Indices for Assessing Changes in Heavy Precipitation Events, Climatic Change, 137, 201–216, https://doi.org/10.1007/s10584-016-1669-2, 2016.

Scholz, F. W. and Stephens, M. A.: K-Sample Anderson-Darling Tests, Journal of the American Statistical Association, 82, 918–924, https://doi.org/10.2307/2288805, 1987.

Schumacher, R. S. and Davis, C. A.: Ensemble-Based Forecast Uncertainty Analysis of Diverse Heavy Rainfall Events, Weather and Forecasting, 25, 1103–1122, https://doi.org/10.1175/2010WAF2222378.1, 2010.

Sénési, S., Bougeault, P., Chèze, J.-L., Cosentino, P., and Thepenier, R.-M.: The Vaison-La-Romaine Flash Flood: Mesoscale Analysis and Predictability Issues, Weather and Forecasting, 11, 417–442, https://doi.org/10.1175/1520-0434(1996)011<0417:TVLRFF>2.0.CO;2, 1996.

Shepard, D.: A Two-Dimensional Interpolation Function for Irregularly-Spaced Data, in: Proceedings of the 1968 23rd ACM National Conference, ACM '68, pp. 517–524, ACM, New York, NY, USA, https://doi.org/10.1145/800186.810616, 1968.

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S., and Rogers, E.: Using Ensembles for Short-Range Forecasting, Monthly Weather Review, 127, 433–446, https://doi.org/10.1175/1520-0493(1999)127<0433:UEFSRF>2.0.CO;2, 1999.

Teo, C.-K., Koh, T.-Y., Chun-Fung Lo, J., and Chandra Bhatt, B.: Principal Component Analysis of Observed and Modeled Diurnal Rainfall in the Maritime Continent, Journal of Climate, 24, 4662–4675, https://doi.org/10.1175/2011JCLI4047.1, 2011.

Toth, Z. and Kalnay, E.: Ensemble Forecasting at NMC: The Generation of Perturbations, Bulletin of the American Meteorological Society, 74, 2317–2330, https://doi.org/10.1175/1520-0477(1993)074<2317:EFANTG>2.0.CO;2, 1993.

Toth, Z. and Kalnay, E.: Ensemble Forecasting at NCEP and the Breeding Method, Monthly Weather Review, 125, 3297–3319, https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2, 1997.

Vié, B., Nuissier, O., and Ducrocq, V.: Cloud-Resolving Ensemble Simulations of Mediterranean Heavy Precipitating Events: Uncertainty on Initial Conditions and Lateral Boundary Conditions, Monthly Weather Review, 139, 403–423, https://doi.org/10.1175/2010MWR3487.1, 2010.

Walser, A. and Schär, C.: Convection-Resolving Precipitation Forecasting and Its Predictability in Alpine River Catchments, Journal of Hydrology, 288, 57–73, https://doi.org/10.1016/j.jhydrol.2003.11.035, 2004.

Walser, A., Lüthi, D., and Schär, C.: Predictability of Precipitation in a Cloud-Resolving Model, Monthly Weather Review, 132, 560–577, https://doi.org/10.1175/1520-0493(2004)132<0560:POPIAC>2.0.CO;2, 2004.

Wernli, H., Paulat, M., Hagen, M., and Frei, C.: SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts, Monthly Weather Review, 136, 4470–4487, https://doi.org/10.1175/2008MWR2415.1, 2008.

Wernli, H., Hofmann, C., and Zimmer, M.: Spatial Forecast Verification Methods Intercomparison Project: Application of the SAL Technique, Weather and Forecasting, 24, 1472–1484, https://doi.org/10.1175/2009WAF2222271.1, 2009.

World Meteorological Organization, ed.: Guidelines on Ensemble Prediction Systems and Forecasting, 1091, WMO, 2012.

World Meteorological Organization, ed.: Guidelines on the definition and monitoring of extreme weather and climate events, Task Team on definitions of Extreme Weather and Climate Events (TT-DEWCE), 2016.

5.3 Summary and conclusions

In this chapter, the feature-based quality measure SAL has been applied to the reforecast dataset using the observation gridded dataset as reference. This dataset has been stratified using a peak-over threshold approach and a clustering method. Results showed that Amplitude and Structure components of the rainfall objects are basically driven by the deep convection parametrization. Between the two main deep convection schemes used in PEARP, PCMT parametrization scheme performs better than the B85 scheme. A statistical comparison of A and S component distributions between the four deep convection parametrization schemes implemented in the reforecast, showed the presence of two classes of parametrizations with two distinct behaviours.

A further analysis of spatial features of the rainfall objects has shown the predominance of large objects and a better representation of objects distribution of the most extreme events.

Chapter 6

Conclusions and perspectives

Contents

6.1 Conclusions

In this study, we addressed the predictability of extreme rainfall events in the French Mediterranean region. A 30-year reforecast dataset has been used with a model set-up the closest possible to the operational ensemble prediction system PEARP. The reforecast and the operational PEARP are based on a mutliphysics scheme to assess the same model error, but the reforecast does not implement perturbed initial conditions. The number of ensemble members reduces from 35 members in PEARP to 10 members in the reforecast. We assumed that a long reforecast dataset would enable to make a robust and significant statistical analysis, regarding the significant number of intense events across the 30-year period. We assessed the overall model skill, following a point-to-point verification and also a feature-based approach, and we applied two post-processing methods for the calibration of 24-hour accumulated precipitation forecast. The comparison between the reforecast and the observations, as well as the post-processing procedure, were based on a unique grid resolution for both field datasets.

The main assumption that the reforecast could properly sample a significant portion of the operational model phase space, in particular around the HPEs spectrum window is assessed by verification scores. The reforecast is proved to have some skill in HPEs forecasting, even for the highest thresholds and at the longest lead times. The reduced member size and the lack of initial condition perturbation should specifically question the ensemble spread of the reforecast against the operational ensemble forecast. Indeed, it is significantly lower. Based on these assessments, two post-processing experiments have been carried out.

A first deterministic post-processing method has been applied to the reforecast, based on a quantile mapping technique. This calibration technique is based on different non-parametric cumulative distribution functions to remap each member. We showed that this deterministic calibration correctly reduces biases member-by-member, but does not systematically reduce the rainfall mean absolute error. Indeed, the mean absolute error is reduced when the member is associated with a positive bias, and conversely increased if the member shows a negative bias. The probabilistic

scoring rules after calibration are marginally affected. We found slight better scores where the correlation between forecast-observation is stronger. Probabilistic scores are hardly affected by Quantile Mapping correction. These results show that the correction of the error from the ranked distribution of rainfall performed member-by-member may not be sufficient for complete a calibration of rainfall forecast.

Alternatively, we applied the extended logistic regression as an ensemble probabilistic calibration approach on the reforecast dataset. The calibration is different at each lead time and grid-point of the domain. A set of calibrated members are sampled from the parametrized Logistic Regression functions that models the calibrated probability density function. The effect of the calibration on the ensemble spread is larger for the first lead times and reduced at the end of the forecast. The calibrated ensemble gets better especially for low rainfall thresholds but some improvement is observed for high precipitation thresholds too. This shows that Logistic Regression is an appropriate method that provides a fair estimation of the calibrated predictive distribution. A further experiment has been held by applying the same Logistic Regression on PEARP operational forecasts over a 4-month period. The PEARP calibrated ensemble forecast scores do not show consistent improvements. Some skills are still observed for the highest thresholds. The main reason for this poor performance of the calibration of the operational ensemble using a downgraded version of the model may be related to the differences in terms of ensemble properties. Some reforecast biases are overcorrected by the post-processing and lead to miscorrection of the operational ensemble.

Probabilistic verification and post-processing methods are often conceived to be performed on model grid-points. A limitation related to these approaches is that their spatial dependency restrains the potential to tackle the spatial structure of the precipitation fields. Therefore, in the second part of this study we investigated the analysis of the systematic errors of the reforecast using a feature-based approach (SAL metric). First, the sub-regional properties of the HPEs have been highlighted by clusters obtained from a k-means classification. This classification shows that HPEs, defined from a peak-over-threshold approach are grouped into three of the five clusters that connect to local well-known rainfall events types. The SAL metrics,

separately applied to the different clusters, reveals specific error dependencies on regional attributes. The most intense precipitation events are related to negative error in the feature amplitude, indicating an overall underestimation of the mean precipitation field. Rainfall clusters related to the Cévennes and the Alps areas are characterized by a positive bias of structure SAL-component, which correspond to forecast fields that are too flat and extended compared to the observed ones. On the other hand, this behaviour has not been observed for cluster events affecting the Languedoc-Roussillon region. When we apply the SAL metric to the different members of the reforecast, significant different scores are obtained depending on the two different deep convection parametrization schemes. In particular, PCMT scheme performs better than B85 scheme at first lead times for both amplitude and structure SAL components. Further, we analyze some features properties involved in the computation of the SAL metric. First, the analysis of the object integrated precipitation amounts has shown that for the highest rainfall events the ratio between the rainfall volume and spatial extension is high. The highest density objects are found for HPEs. Then, if we consider the ranked distribution of object rainfall amount, it revealed that the model is able to reproduce the distribution of the object rainfall for the cluster that collect the most extreme HPEs.

One main relevant point raised in this study has been the impact of the model error representation using the multiphysics approach on the HPEs predictability. The comparison between physics schemes using deterministic forecast verification has shown that deep convection parametrization strongly affects spatial distribution and intensity of the precipitation fields. The effects induced by the other physical parametrizations (i.e., turbulence, shallow convections, and oceanic flux) led to negligible differences in the scores. Similarly, the SAL-metric showed that the quality of the forecast for intense rainfall in terms of spatial structure and averaged amplitude is also mainly driven by the deep convection parametrizations. PCMT scheme leads to lower point-by-point forecast skill compared to B85 scheme, but it outperforms B85 scheme when considering an object-oriented approach. These results could mean that a multiphysics approach for the model error representation could have some limitations. Since physical parametrization set-up is only based on these

two main deep convection schemes, the multiphysic approach implemented in the PEARP model might lack of a more exhaustive sampling of the model uncertainty for heavy precipitation forecast.

6.2 Perspectives

In this study we analysed some challenging issues about ensemble prediction of intense precipitation. Experiments with two post-processing methods are scarcely beneficial for the improvement of intense precipitation forecasts, despite the large size of the learning dataset. Indeed, the frequency occurrence of these events is so small that a regular representation in the training dataset is not guaranteed. On another hand, we found that the forecast skill for such rare events is difficult to assess. Some scores tend to degenerate towards non-informative values (e.g., Brier Score) for such lightly subsampled events. Scores based on a continuous representation variable (e.g., CRPS) are not specifically informative for the extreme events. Then, we underline some perspectives that can be drawn out of this work.

We showed that the lack of initial condition perturbation in the reforecast dataset has a significant impact on the predictive distributions. We suggest that recent ensemble reanalysis, like ERA5 (i.e.,ERA5: Hersbach, 2016), which contains 10 perturbed members could be considered. The use of this ensemble reanalysis system in the PEARP reforecast may be a potential solution. Although it would not exactly be a replication of the operational ensemble system framework, it would offer an opportunity to sample the uncertainties due to the initialization. The combination of 10-member EDA with the ten physical packages implemented in PEARP could potentially enlarge the size of the reforecast to 100 members.

The inter comparison among the physical packages used in PEARP has shown that HPEs forecast is mostly controlled by two deep convection parametrizations, which are replicated in several packages. Model error might be better represented by introducing additional physical paremetrization to the ensemble system. One issue related to this implementation is related to the necessity of development and update of a large number of different parametrization schemes. An alternative so-

lution could be to replace the multiphysics approach with other methods, such as random parameter perturbations or as the stochastically perturbed parameterization tendencies.

The SAL analysis has emphasized that the model forecast includes meaningful information at the spatial scale of the event. The evaluation of the reforecast as well as the post-processed ensemble using point-to-point methodologies have shown some deficiencies. Some studies have suggested to take into account the event spatial variability through a neighbourhood approach (e.g., Scheuerer, 2014; Scheuerer and Hamill, 2015). It is based on a spatial resampling of the ensemble which, as a consequence, enlarges the sample dataset to provide a better estimation of predictors in post-processing. Different weights can be given to the neighbourhood grid points as a function of the distance from the targeted grid-point involved in the calibration. A further implementation should be to assign weights to the spatiotemporal neighborhood depending on the errors forecast of the neighbourhood points in the training dataset with respect to the observed ones over the targeted point (e.g., Scheuerer and Hamill, 2018; Scheuerer et al., 2017).

The quantile mapping procedure has shown potential skill in reducing the biases of the ensemble members, while the extended logistic regression enhanced probabilistic skill of ensemble forecasts. It could be practicable to combine these two methods, or to apply quantile mapping as a debiasing procedure to further perform a one selected post-processing method, as proposed by Hamill et al. (2017).

Diagnosis of errors associated to the object-based SAL metric analysis pointed out that intense precipitation forecast depends on the area considered and on physical parametrizations categories. We think it would be possible to extend the scope of the relationship between HPEs probabilistic quantitative precipitation forecasts and the large scale circulation. Nuissier et al. (2011) have shown that some specific large scale circulation patterns are correlated with the HPEs on the French domain used in the current study. Model predictability could be enhanced by assuming the existence of large scale atmospheric predictors favouring intense precipitation, that would likely be more predictable. Large scale variable could be introduced in post-processing methods as predictors which may enhance the predictability. This

procedure should be regarded as a downscaling approach.

Conclusions et perspectives (Français)

Contents

Conclusions

Dans ce travail de thèse on a étudié la prévisibilité des fortes précipitations dans le sud-est de la France à l'aide d'une approche probabiliste de la prévision. Un jeu de prévisions rétrospectives, ou re-jeu, couvrant une période de 30 ans, avec une version du modèle la plus proche possible du système opérationnel PEARP, a été utilisé. Le re-jeu de prévisions a été construit avec la même représentation de l'erreur de modélisation, basée sur une approche dite « multi-physiques », que PEARP mais sans représentation de l'incertitude initiale. Il comporte 10 membres contre 35 pour le système opérationnel. Le parti pris de cette étude est que la longue période temporelle que couvre le re-jeu de prévisions permet une meilleure analyse statistique des épisodes de pluies intenses. La qualité générale du re-jeu a été évaluée à la fois par des scores en points de grille, et par des scores prenant en compte la structure spatiale de la pluie. Nous avons aussi appliqué deux méthodes de calibrage sur la prévision de cumul quotidien de pluie. Il est également à noter qu'un travail a été effectué pour que toutes les données manipulées, observations ou prévisions, aient la même résolution spatiale.

Dans une première partie la capacité du re-jeu à représenter le climat du modèle, au sens d'un espace de phase physique, et en particulier dans le domaine des fortes pluies a été évaluée par des méthodes standard de vérification de modèles de prévision. Nous avons montré que le re-jeu répond positivement à cette hypothèse par des prévisions de dépassement de seuil significativement performantes y compris aux plus longues échéances. La dispersion de l'ensemble rejoué est cependant plus faible que celle du modèle opérationnel, ceci en raison du nombre plus faible de membres et de l'absence de perturbations de conditions initiales.

Dans un second temps, et nonobstant les lacunes observées du re-jeu de prévisions, deux méthodes de correction a posteriori ont été testées dans le but d'évaluer les possibilités de calibrer le système opérationnel à l'aide du re-jeu. Ici encore on a fait l'hypothèse que la longue période temporelle couverte par le re-jeu de prévisions était un atout dans l'application de ces méthodes.

Une première expérience a consisté à calibrer le re-jeu de prévisions à l'aide d'une

méthode de calibrage déterministe. La méthode dite du « quantile mapping » a été appliquée au re-jeu de prévisions en utilisant une approche de validation croisée. La correction a été basée sur l'utilisation de plusieurs fonctions de répartitions non-paramétriques pour recaler chaque membre en chaque point de grille du domaine. Il a été montré que ce calibrage déterministe permet de réduire les biais individuels mais que cela n'impliquait pas obligatoirement une baisse de l'erreur absolue moyenne. Il a été observé que l'erreur absolue moyenne n'était diminuée que dans les cas où les prévisions comportaient un biais positif. Au contraire, l'erreur absolue moyenne était dégradée dans le cas d'un biais initial négatif des prévisions. Les évaluations à l'aide de scores probabilistes ont montré un faible impact du calibrage avec une faible amélioration (dégradation) en lien avec une forte (faible) corrélation entre observations et prévisions. Les résultats obtenus avec cette approche « quantile mapping »montrent qu'elle doit plutôt être considérée comme une technique de correction déterministe de biais plutôt que comme un moyen de calibrer complètement un système probabiliste. Une autre approche, connue sous le nom de « régression logistique étendue » a également été appliquée. Avec cette méthode, le calibrage s'effectue de façon différenciée en chaque point de grille et pour chaque échéance de prévision. Les prévisions brutes sont corrigées à l'aide de fonctions de régression paramétriques, déterminées en utilisant une approche par validation croisée. Contrairement au « quantile mapping », la régression logistique a un fort impact sur la dispersion de l'ensemble qu'elle tend à augmenter. Cet impact apparaît comme très fort sur les premières échéances puis tend à décroître. Ceci peut s'expliquer par une compensation, sur les premières échéances, de la non-prise en compte de l'incertitude initiale par le re-jeu. Globalement les scores obtenus avec cette approche sont meilleurs, particulièrement pour les faibles seuils de précipitations mais aussi parfois pour des seuils plus élevés. De façon assez surprenante, les scores individuels de chaque membre de l'ensemble ne sont que peu améliorés après calibrage. Ces résultats montrent que l'approche par « régression logistique » peut être vue principalement comme un outil pour le calibrage probabiliste. Cette même technique a ensuite été appliquée, sur une période de quatre mois, au système opérationnel PEARP. Les fonctions de régression apprises en utilisant le re-jeu ont été utilisées pour calibrer PEARP. Les résultats obtenus

sont moins consistants dans ce cas et plutôt orientés vers une faible amélioration due au calibrage. Ce résultat montre que le succès du calibrage dépend fortement du fait que le système utilisé pour l'apprentissage est le plus proche possible du système à calibrer.

Les scores d'évaluation des prévisions probabilistes et les techniques de correction a posteriori sont généralement appliqués point de grille par point de grille. Une limite connue de cette approche est qu'elle ne tient pas compte des propriétés spatiales du champ que l'on évalue ou que l'on cherche à calibrer. Dans la seconde partie de ce travail on a donc utilisé une approche « objet » (au travers du score SAL) pour caractériser les erreurs des prévisions du re-jeu. Une technique de clustering (approche « k-means ») a tout d'abord été utilisée pour mettre en lumière le caractère régional des épisodes de fortes pluies. Cette classification a montré que les épisodes de pluies intenses se retrouvaient dans trois des cinq classes de pluies obtenues et que leur localisation était représentative de la régionalisation de ces épisodes. Le score SAL a ensuite été appliqué pour mieux mettre en avant la dépendance des erreurs de prévision à la localisation des épisodes. Les épisodes les plus intenses sont caractérisés par une sous-estimation globale de la valeur moyenne du cumul de précipitation. Sur les Cévennes et les Alpes les structures obtenues sont généralement trop étalées par rapport à celles détectées dans l'observation. Cette caractéristique n'a cependant pas été observée sur les épisodes touchant le Languedoc-Roussillon. Les résultats de l'application du score SAL ont ensuite été analysés sous l'angle des différentes physiques du re-jeu de prévisions. Il a été observé que les scores étaient très sensibles au choix du schéma de représentation de la convection profonde. L'utilisation du schéma PCMT permet d'obtenir de meilleurs résultats que ceux obtenus avec le schéma B85, en particulier aux premières échéances et pour les erreurs de structure et d'amplitude. La grande taille de notre échantillon de situations a par ailleurs permis une étude du score SAL appliqué aux propriétés des structures pluvieuses. Cette analyse a notamment montré que ceux associés à la classe de pluie comprenant les événements les plus intenses présentaient un ratio volume/extension très élevé.

Un point très important de cette étude a été l'évaluation de la représentation de l'erreur modèle à l'aide de l'approche multi-physiques. Il a été montré que l'impact

le plus important dans le choix des physiques concernait le schéma de convection profonde. L'application du score SAL a lui aussi montré que la qualité des prévisions des structures de pluie était d'abord affectée par le choix du schéma de convection profonde. Si le schéma PCMT a obtenu de moins bons scores que le schéma B85 dans une évaluation déterministe, il a montré de meilleurs résultats lors de l'évaluation par objets précipitants. Ces résultats semblent indiquer que l'approche multi-physiques a des limites. Dans la mesure où les différentes physiques ne dupliquent que deux schémas de convection profonde, la PEARP peut souffrir d'un manque de représentation de l'ensemble des erreurs de modélisation.

Perspectives

Dans ce travail de thèse, des points importants de la problématique de la prévision probabiliste des épisodes de pluies intenses ont été étudiés. Les résultats obtenus ont montré que les méthodes classiques de calibrage, même basées sur une longue période d'apprentissage, n'apportaient que peu d'amélioration à la prévision des épisodes intenses. La rareté de ces derniers implique sans doute des faiblesses dans la représentation fidèle de leur distribution statistique même sur de longues périodes d'apprentissage, ce qui peut expliquer en partie ces résultats. Un autre point à mettre en lumière est la difficulté à valider des prévisions pour ces épisodes intenses et rares. La plupart des scores classiques sont inutilisables ou non-indicatifs sur des cas aussi rares (que ce soit par exemple le score de Brier ou la CRPS). Quelques perspectives intéressantes de l'ensemble du travail effectué se dégagent cependant.

Nous avons montré que la non prise en compte des erreurs de prévision liées à l'incertitude des conditions initiales était préjudiciable. La récente disponibilité de réanalyses telles que ERA5 (Hersbach, 2016) qui contient 10 membres d'analyses perturbées est une réelle opportunité pour palier ce problème. Même si cela ne serait toujours pas un moyen de dupliquer fidèlement la version complète du système de prévision opérationnel, cela permettrait, en combinant ces réanalyses perturbées et les 10 schémas de paramétrisation physique différents, d'envisager la création d'un ensemble atteignant potentiellement 100 membres.

L'inter-comparaison des schémas de paramétrisation physiques a montré que la prévision des épisodes de pluies très intenses est surtout sensible au composant dédié à la représentation de la convection profonde dont deux versions différentes seulement sont dupliquées au sein de l'ensemble. Nous pensons qu'ajouter d'autres schémas permettrait de mieux prendre en compte les sources d'erreurs de prévision de fortes pluies, même si cela alourdirait de façon notoire la mise au point et la maintenance d'un tel système. Une alternative serait de remplacer l'approche multi-physiques par d'autres méthodes, comme les perturbations aléatoires des paramètres des schémas physiques, ou encore les perturbations stochastiques des tendances.

La vérification avec la métrique SAL a montré que la prise en compte de la structure spatiale de la pluie permet d'obtenir une information plus pertinente dans la prévision. Certaines études ont montré, que ce soit pour des méthodes de vérification ou de calibrage, que la prise en compte d'un voisinage autour d'un « point de grille » considéré permettait une meilleure prise en compte de la variabilité spatiale du phénomène. Des études comme (e.g., Scheuerer, 2014; Scheuerer and Hamill, 2015) ont démontré que l'augmentation de l'échantillonnage par utilisation des plus proches voisins donne de bons résultats. Grâce au ré-échantillonnage du signal contenu dans ce voisinage, les prédicteurs utilisés dans les méthodes de calibrage sont mieux estimés. Des poids peuvent être alloués aux points voisins en fonction de leur distance au point de grille considéré. D'autres méthodes suggèrent l'utilisation de poids liés à la performance générale de la prévision en ces points estimés sur la période d'apprentissage comme dans Scheuerer et al. (2017) et Scheuerer and Hamill (2018). Nous avons montré que la méthode « quantile mapping » permet une réduction correcte des biais par membre. De même la régression logistique permet d'améliorer la représentation de la distribution statistique de l'ensemble. Il semblerait intéressant de concevoir l'utilisation en série de ces deux méthodes combinant ces deux qualités majeurs. Hamill et al. (2017) par exemple a montré en particulier tout l'intérêt d'utiliser le « quantile mapping » pour débiaiser l'ensemble avant d'utiliser une méthode de calibrage.

L'étude des résultats obtenus avec le score SAL ont montré que la qualité des prévisions des épisodes intenses de pluie dépendait de la zone géographique impactée.

Il a également été montré le rôle crucial du schéma de convection profonde dans la qualité des prévisions. Une perspective du présent travail pourrait être d'étudier plus en profondeur, pour les épisodes de fortes pluies, l'impact de la circulation de grande échelle sur la qualité des prévisions probabilistes. Nuissier et al. (2011) ont montré la corrélation qui existe entre le type de circulation à grande échelle et la survenue des épisodes de fortes pluies en France. La prévision des ces épisodes pourrait donc être améliorée en prenant mieux en compte les environnements de grande échelle favorables à leur survenue.

References

Accadia, C., Mariani, S., Casaioli, M., Lavagnini, A., and Speranza, A. (2003). Sensitivity of Precipitation Forecast Skill Scores to Bilinear Interpolation and a Simple Nearest-Neighbor Average Method on High-Resolution Verification Grids. *Weather and Forecasting*, 18(5):918–932. 60

AghaKouchak, A., Behrangi, A., Sorooshian, S., Hsu, K., and Amitai, E. (2011). Evaluation of satellite-retrieved extreme precipitation rates across the central United States. *Journal of Geophysical Research: Atmospheres*, 116(D2). 27

Anderson, J. L. (1996). A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations. *Journal of Climate*, 9(7):1518–1530. 83

Anquetin, S., Minsicloux, F., Creutin, J.-D., and Cosma, S. (2003). Numerical simulation of orographic rainbands. *Journal of Geophysical Research: Atmospheres*, 108(D8). 5

Applequist, S., Gahrs, G. E., Pfeffer, R. L., and Niu, X.-F. (2002). Comparison of Methodologies for Probabilistic Quantitative Precipitation Forecasting. *Weather and Forecasting*, 17(4):783–799. 33

Atger, F. (2004). Estimation of the reliability of ensemble-based probabilistic forecasts. *Quarterly Journal of the Royal Meteorological Society*, 130(597):627–646. 80

Banacos, P. C. and Schultz, D. M. (2005). The use of moisture flux convergence in forecasting convective initiation: Historical and operational perspectives. *Weather and Forecasting*, 20(3):351–366. 7

Baran, S. and Nemoda, D. (2016). Censored and shifted gamma distribution based EMOS model for probabilistic quantitative precipitation forecasting. *Environmetrics*, 27(5):280–292. 33

Barkmeijer, J., Buizza, R., Palmer, T. N., Puri, K., and Mahfouf, J.-F. (2001). Tropical singular vectors computed with linearized diabatic physics. *Quarterly Journal of the Royal Meteorological Society*, 127(572):685–708. 43

Bauer, P., Thorpe, A., and Brunet, G. (2015). The quiet revolution of numerical weather prediction. *Nature*, 525(7567):47–55. 17

Bazile, E., Marquet, P., Bouteloup, Y., and Bouyssel, F. (2012). The turbulent kinetic energy (tke) scheme in the nwp models at meteo france. In *Workshop on Workshop on Diurnal cycles and the stable boundary layer, 7-10 November 2011*, pages 127–135, Shinfield Park, Reading. ECMWF, ECMWF. 44

Beaulant, A.-L., Joly, B., Nuissier, O., Somot, S., Ducrocq, V., Joly, A., Sevault, F., Deque, M., and Ricard, D. (2011). Statistico-dynamical downscaling for Mediterranean heavy precipitation. *Quarterly Journal of the Royal Meteorological Society*, 137(656):736–748. 15

Bechtold, P., Bazile, E., Guichard, F., Mascart, P., and Richard, E. (2001). A mass-flux convection scheme for regional and global models. *Quarterly Journal of the Royal Meteorological Society*, 127(573):869–886. 44

Belamari, S. (2005). Report on uncertainty estimates of an optimal bulk formulation for surface turbulent fluxes. *MERSEA IP Deliverable 4.1.2*, pages 1–29. 45

Ben Bouallègue, Z. (2012). Calibrated Short-Range Ensemble Precipitation Forecasts Using Extended Logistic Regression with Interaction Terms. *Weather and Forecasting*, 28(2):515–524. 35

Berre, L., Pannekoucke, O., Desroziers, G., Stefanescu, S. E., Chapnik, B., and Raynaud, L. (2007). A variational assimilation ensemble and the spatial filtering of its error covariances: Increase of sample size by local spatial averaging. In *ECMWF Workshop on Flow-Dependent Aspecyts of Data Assimilation, 11-13 June 2007*, pages 151–168, Shinfield Park, Reading. ECMWF, ECMWF. 42

Berrisford, P., Dee, D. P., Poli, P., Brugge, R., Fielding, M., Fuentes, M., Kållberg,

P. W., Kobayashi, S., Uppala, S., and Simmons, A. (2011). The ERA-Interim archive Version 2.0. Technical Report 1, ECMWF, Shinfield Park, Reading. 86

Blanchet, J. and Creutin, J.-D. (2017). Co-Occurrence of Extreme Daily Rainfall in the French Mediterranean Region. *Water Resources Research*, 53(11):9330–9349. 13

Boé, J., Terray, L., Habets, F., and Martin, E. (2007). Statistical and dynamical downscaling of the Seine basin climate for hydro-meteorological studies. *International Journal of Climatology*, 27(12):1643–1655. 99

Boisserie, M., Decharme, B., Descamps, L., and Arbogast, P. (2016). Land surface initialization strategy for a global reforecast dataset. *Quarterly Journal of the Royal Meteorological Society*, 142(695):880–888. 46, 47

Bougeault, P. (1985). A Simple Parameterization of the Large-Scale Effects of Cumulus Convection. *Monthly Weather Review*, 113(12):2108–2121. 44

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B., and Beare, S. E. (2008a). The MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 134(632):703–722. 18

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson, K. B., and Beare, S. E. (2008b). The MOGREPS short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 134(632):703–722. 20

Bresson, E., Ducrocq, V., Nuissier, O., Ricard, D., and de Saint-Aubin, C. (2012). Idealized numerical simulations of quasi-stationary convective systems over the Northwestern Mediterranean complex terrain. *Quarterly Journal of the Royal Meteorological Society*, 138(668):1751–1763. 11

Bresson, R., Ricard, D., and Ducrocq, V. (2009). Idealized mesoscale numerical study of Mediterranean heavy precipitating convective systems. *Meteorology and Atmospheric Physics*, 103(1):45–55. 9, 11

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3. 25, 74

Briggs, W. M. and Levine, R. A. (1997). Wavelets and Field Forecast Verification. *Monthly Weather Review*, 125(6):1329–1341. 27

Bröcker, J. (2009). Reliability, Sufficiency, and the Decomposition of Proper Scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519. 75

Brossier, C. L., Drobinski, P., Béranger, K., Bastin, S., and Orain, F. (2013). Ocean memory effect on the dynamics of coastal heavy precipitation preceded by a mistral event in the northwestern Mediterranean. *Quarterly Journal of the Royal Meteorological Society*, 139(675):1583–1597. 13

Brown, T. A. (1974). *Admissible Scoring Systems for Continuous Distributions*. The Rand Corporation, 1700 Main Street, Santa Monica, California 90406 (P-5235, $1. 77

Buizza, R., Milleer, M., and Palmer, T. N. (1999). Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 125(560):2887–2908. 20

Buizza, R. and Palmer, T. N. (1995). The Singular-Vector Structure of the Atmospheric Global Circulation. *Journal of the Atmospheric Sciences*, 52(9):1434–1456. 19

Buizza, R., Tribbia, J., Molteni, F., and Palmer, T. (1993). Computation of optimal unstable structures for a numerical weather prediction model. *Tellus A*, 45(5):388–407. 19

Burrough, P. A., McDonnell, R., McDonnell, R. A., and Lloyd, C. D. (2015). *Principles of Geographical Information Systems*. OUP Oxford. 52

Buzzi, A. and Foschini, L. (2000). Mesoscale Meteorological Features Associated with Heavy Precipitation in the Southern Alpine Region. *Meteorology and Atmospheric Physics*, 72(2):131–146. 8, 11

Buzzi, A., Tartaglione, N., and Malguzzi, P. (1998). Numerical Simulations of the 1994 Piedmont Flood: Role of Orography and Moist Processes. *Monthly Weather Review*, 126(9):2369–2383. 11

Byrne, S. (2016). A note on the use of empirical AUC for evaluating probabilistic forecasts. *Electronic Journal of Statistics*, 10(1):380–393. 82

Candille, G., Côté, C., Houtekamer, P. L., and Pellerin, G. (2007). Verification of an Ensemble Prediction System against Observations. *Monthly Weather Review*, 135(7):2688–2699. 84

Candille, G. and Talagrand, O. (2005). Evaluation of probabilistic prediction systems for a scalar variable. *Quarterly Journal of the Royal Meteorological Society*, 131(609):2131–2150. 24, 79

Carroll, R. J. and Stefanski, L. A. (1990). Approximate Quasi-likelihood Estimation in Models With Surrogate Predictors. *Journal of the American Statistical Association*, 85(411):652–663. 114

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A., Pocernich, M., Damrath, U., Ebert, E. E., Brown, B. G., and Mason, S. (2008). Forecast verification: Current status and future directions. *Meteorological Applications*, 15(1):3–18. 23

Charron, M., Pellerin, G., Spacek, L., Houtekamer, P. L., Gagnon, N., Mitchell, H. L., and Michelin, L. (2009). Toward Random Sampling of Model Error in the Canadian Ensemble Prediction System. *Monthly Weather Review*, 138(5):1877–1901. 18, 20

Chen, F.-W. and Liu, C.-W. (2012). Estimation of the spatial rainfall distribution using inverse distance weighting (IDW) in the middle of Taiwan. *Paddy and Water Environment*, 10(3):209–222. 49

Cherubini, T., Ghelli, A., and Lalaurette, F. (2002). Verification of Precipitation Forecasts over the Alpine Region Using a High-Density Observing Network. *Weather and Forecasting*, 17(2):238–249. 21

Clark, P., Roberts, N., Lean, H., Ballard, S. P., and Charlton-Perez, C. (2016). Convection-permitting models: A step-change in rainfall forecasting. *Meteorological Applications*, 23(2):165–181. 22

Cloke, H. L. and Pappenberger, F. (2009). Ensemble flood forecasting: A review. *Journal of Hydrology*, 375(3):613–626. 21

Colacino, M. and Conte, M. (1995). Heat waves in the central Mediterranean. A synoptic climatology. *Il Nuovo Cimento C*, 18(3):295–304. 2

Collier, C. G. (2007). Flash flood forecasting: What are the limits of predictability? *Quarterly Journal of the Royal Meteorological Society*, 133(622):3–23. 36

Corfidi, S. F. (2003). Cold Pools and MCS Propagation: Forecasting the Motion of Downwind-Developing MCSs. *Weather and Forecasting*, 18(6):997–1017. xvii, 6, 7, 11

Courtier, P., Freydier, C., Geleyn, J.-F., Rabier, F., and Rochas, M. (1991). The arpege project at meteo france. In *Seminar on Numerical Methods in Atmospheric Models, 9-13 September 1991*, volume II, pages 193–232, Shinfield Park, Reading. ECMWF, ECMWF. 40

Cuxart, J., Bougeault, P., and Redelsperger, J.-L. (2000). A turbulence scheme allowing for mesoscale and large-eddy simulations. *Quarterly Journal of the Royal Meteorological Society*, 126(562):1–30. 44

Davis, C., Brown, B., and Bullock, R. (2006). Object-Based Verification of Precipitation Forecasts. Part I: Methodology and Application to Mesoscale Rain Areas. *Monthly Weather Review*, 134(7):1772–1784. 27, 28

Davis, C. A., Brown, B. G., Bullock, R., and Halley-Gotway, J. (2009). The Method for Object-Based Diagnostic Evaluation (MODE) Applied to Numerical Forecasts from the 2005 NSSL/SPC Spring Program. *Weather and Forecasting*, 24(5):1252–1267. 27, 28

Davolio, S., Mastrangelo, D., Miglietta, M. M., Drofa, O., Buzzi, A., and Malguzzi, P. (2009). High resolution simulations of a flash flood near Venice. *Natural Hazards and Earth System Sciences*, 9(5):1671–1678. 7

Dee, D. P., Uppala, S. M., Simmons, A. J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M. A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A. C. M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A. J., Haimberger, L., Healy, S. B., Hersbach, H., Hólm, E. V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A. P., Monge-Sanz, B. M., Morcrette, J.-J., Park, B.-K., Peubey, C., de Rosnay, P., Tavolato, C., Thépaut, J.-N., and Vitart, F. (2011). The ERA-Interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137(656):553–597. 46, 86

Delrieu, G., Nicol, J., Yates, E., Kirstetter, P.-E., Creutin, J.-D., Anquetin, S., Obled, C., Saulnier, G.-M., Ducrocq, V., Gaume, E., Payrastre, O., Andrieu, H., Ayral, P.-A., Bouvier, C., Neppel, L., Livet, M., Lang, M., du-Châtelet, J. P., Walpersdorf, A., and Wobrock, W. (2005). The Catastrophic Flash-Flood Event of 8–9 September 2002 in the Gard Region, France: A First Case Study for the Cévennes–Vivarais Mediterranean Hydrometeorological Observatory. *Journal of Hydrometeorology*, 6(1):34–52. 8, 13

Descamps, L., Labadie, C., and Bazile, E. (2011). Representing model uncertainty using the multiparametrization method. 20

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., and Cébron, P. (2015). PEARP, the Météo-France short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 141(690):1671–1685. xx, xxx, xxxiv, 18, 40, 42

Desroziers, G., Camino, J., and Berre, L. (2014). 4denvar: Link with 4d state formulation of variational assimilation and different possible implementations. *Quarterly Journal of the Royal Meteorological Society*, 140. 19

Doswell, C. A. (1987). The Distinction between Large-Scale and Mesoscale Contribution to Severe Convection: A Case Study Example. *Weather and Forecasting*, 2(1):3–16. 5

Doswell, C. A., Brooks, H. E., and Maddox, R. A. (1996). Flash Flood Forecasting: An Ingredients-Based Methodology. *Weather and Forecasting*, 11(4):560–581. 5, 7

Doswell, C. A., Davies-Jones, R., and Keller, D. L. (1990). On Summary Measures of Skill in Rare Event Forecasting Based on Contingency Tables. *Weather and Forecasting*, 5(4):576–585. 63

Doswell, C. A., Ramis, C., Romero, R., and Alonso, S. (1998). A Diagnostic Study of Three Heavy Precipitation Episodes in the Western Mediterranean Region. *Weather and Forecasting*, 13(1):102–124. 6

Ducrocq, V., Nuissier, O., Ricard, D., Lebeaupin, C., and Thouvenin, T. (2008a). A numerical study of three catastrophic precipitating events over southern France. II: Mesoscale triggering and stationarity factors. *Quarterly Journal of the Royal Meteorological Society*, 134(630):131–145. xviii, 8, 10, 11, 15

Ducrocq, V., Nuissier, O., Ricard, D., Lebeaupin, C., and Thouvenin, T. (2008b). A numerical study of three catastrophic precipitating events over southern France. II: Mesoscale triggering and stationarity factors. *Quarterly Journal of the Royal Meteorological Society*, 134(630):131–145. 9

Ducrocq, V., Ricard, D., Lafore, J.-P., and Orain, F. (2002). Storm-Scale Numerical Rainfall Prediction for Five Precipitating Events over France: On the Importance of the Initial Humidity Field. *Weather and Forecasting*, 17(6):1236–1256. 5

Duffourg, F. and Ducrocq, V. (2011). Origin of the moisture feeding the Heavy Precipitating Systems over Southeastern France. *Natural Hazards and Earth System Sciences*, 11(4):1163–1178. 13

Duffourg, F. and Ducrocq, V. (2013). Assessment of the water supply to Mediterranean heavy precipitation: A method based on finely designed water budgets. *Atmospheric Science Letters*, 14(3):133–138. xviii, 12, 13

Ebert, E. E. (2001). Ability of a Poor Man's Ensemble to Predict the Probability

and Distribution of Precipitation. *Monthly Weather Review*, 129(10):2461–2480. 28

Ebert, E. E. (2008). Fuzzy verification of high-resolution gridded forecasts: A review and proposed framework. *Meteorological Applications*, 15(1):51–64. 28

Ebert, E. E. (2009). Neighborhood Verification: A Strategy for Rewarding Close Forecasts. *Weather and Forecasting*, 24(6):1498–1510. 25

Ebert, E. E. and McBride, J. L. (2000). Verification of precipitation in weather systems: Determination of systematic errors. *Journal of Hydrology*, 239(1):179–202. 27, 28

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC Press. 84

Epstein, E. S. (1969a). A Scoring System for Probability Forecasts of Ranked Categories. *Journal of Applied Meteorology*, 8(6):985–987. 115

Epstein, E. S. (1969b). Stochastic dynamic prediction. *Tellus*, 21(6):739–759. 17

Erdin, R., Frei, C., and Künsch, H. R. (2012). Data Transformation and Uncertainty in Geostatistical Combination of Radar and Rain Gauges. *Journal of Hydrometeorology*, 13(4):1332–1346. 51

Evensen, G. (2003). The ensemble kalman filter : theoritical formulation and practical implementation. *Ocean Dynamics*, 53:343–367. 19

Ferro, C. A. T. and Stephenson, D. B. (2011). Extremal Dependence Indices: Improved Verification Measures for Deterministic Forecasts of Rare Binary Events. *Weather and Forecasting*, 26(5):699–713. 63

Fletcher, R. (2013). *Practical Methods of Optimization*. John Wiley & Sons. 114

Fortin, V., Abaza, M., Anctil, F., and Turcotte, R. (2014). Why Should Ensemble Spread Match the RMSE of the Ensemble Mean? *Journal of Hydrometeorology*, 15(4):1708–1713. 83

Frei, C. and Schär, C. (1998). A precipitation climatology of the Alps from high-resolution rain-gauge observations. *International Journal of Climatology*, 18(8):873–900. 9

Fresnay, S., Hally, A., Garnaud, C., Richard, E., and Lambert, D. (2012). Heavy precipitation events in the Mediterranean: Sensitivity to cloud physics parameterisation uncertainties. *Natural Hazards and Earth System Sciences*, 12(8):2671–2688. 7, 8, 15

Fritsch, J. M. and Carbone, R. E. (2004). Improving Quantitative Precipitation Forecasts in the Warm Season: A USWRP Research and Development Strategy. *Bulletin of the American Meteorological Society*, 85(7):955–966. 21

Frogner, I.-L., Singleton, A. T., Køltzow, M. Ø., and Andrae, U. (2019). Convection-permitting ensembles: Challenges related to their design and use. *Quarterly Journal of the Royal Meteorological Society*, 145(S1):90–106. 22

Gallus, W. A. (2010). Application of Object-Based Verification Techniques to Ensemble Precipitation Forecasts. *Weather and Forecasting*, 25(1):144–158. 28

Gebetsberger, M., Messner, J. W., Mayr, G. J., and Zeileis, A. (2017). Fine-Tuning Nonhomogeneous Regression for Probabilistic Precipitation Forecasts: Unanimous Predictions, Heavy Tails, and Link Functions. *Monthly Weather Review*, 145(11):4693–4708. 33

Gilleland, E. (2011). Spatial Forecast Verification: Baddeley's Delta Metric Applied to the ICP Test Cases. *Weather and Forecasting*, 26(3):409–415. 27

Gilleland, E. (2012). Testing Competing Precipitation Forecasts Accurately and Efficiently: The Spatial Prediction Comparison Test. *Monthly Weather Review*, 141(1):340–355. 25

Gilleland, E., Ahijevych, D., Brown, B. G., Casati, B., and Ebert, E. E. (2009). Intercomparison of Spatial Forecast Verification Methods. *Weather and Forecasting*, 24(5):1416–1430. xix, 26

Gneiting, T. and Raftery, A. E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association*, 102(477):359–378. 74

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133(5):1098–1118. 30, 32

Gneiting, T. and Ranjan, R. (2011). Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics*, 29(3):411–422. 25

Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228(1):113–129. 51

Goovaerts, P. and Goovaerts, D. o. C. a. E. E. P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press. 49

Goudenhoofdt, E. and Delobbe, L. (2009). Evaluation of radar-gauge merging methods for quantitative precipitation estimates. *Hydrology and Earth System Sciences*, 13(2):195–203. 49

Gudmundsson, L., Bremnes, J. B., Haugen, J. E., and Engen-Skaugen, T. (2012). Technical Note: Downscaling RCM precipitation to the station scale using statistical transformations – a comparison of methods. *Hydrol. Earth Syst. Sci.*, 16(9):3383–3390. 99

Hagelin, S., Son, J., Swinbank, R., McCabe, A., Roberts, N., and Tennant, W. (2017). The Met Office convective-scale ensemble, MOGREPS-UK. *Quarterly Journal of the Royal Meteorological Society*, 143(708):2846–2861. 22

Haiden, T., Bidlot, J., Ferranti, L., Bauer, P., Dahoui, M., Janousek, M., Prates, F., Vitart, F., and Richardson, D. (2015). *Evaluation of ECMWF Forecasts, Including 2014-2015 Upgrades*. European Centre for Medium-Range Weather Forecasts. 29

Hally, A., Richard, E., Fresnay, S., and Lambert, D. (2014). Ensemble simulations with perturbed physical parametrizations: Pre-HyMeX case studies. *Quarterly Journal of the Royal Meteorological Society*, 140(683):1900–1916. 22

Hamill, T. M. (2001). Interpretation of Rank Histograms for Verifying Ensemble Forecasts. *Monthly Weather Review*, 129(3):550–560. 79, 83

Hamill, T. M. (2012). Verification of TIGGE Multimodel and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous United States. *Monthly Weather Review*, 140(7):2232–2252. 25, 29, 35, 60, 114

Hamill, T. M., Bates, G. T., Whitaker, J. S., Murray, D. R., Fiorino, M., Galarneau, T. J., Zhu, Y., and Lapenta, W. (2013). NOAA's Second-Generation Global Medium-Range Ensemble Reforecast Dataset. *Bulletin of the American Meteorological Society*, 94(10):1553–1565. 29

Hamill, T. M. and Colucci, S. J. (1997). Verification of Eta–RSM Short-Range Ensemble Forecasts. *Monthly Weather Review*, 125(6):1312–1327. xxii, 82, 83

Hamill, T. M., Engle, E., Myrick, D., Peroutka, M., Finan, C., and Scheuerer, M. (2017). The U.S. National Blend of Models for Statistical Postprocessing of Probability of Precipitation and Deterministic Precipitation Amount. *Monthly Weather Review*, 145(9):3441–3463. 30, 99, 101, 182, 190

Hamill, T. M., Hagedorn, R., and Whitaker, J. S. (2008). Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation. *Monthly Weather Review*, 136(7):2620–2632. xxx, xxxiv, 33, 60, 113

Hamill, T. M. and Juras, J. (2006). Measuring forecast skill: Is it real skill or is it the varying climatology? *Quarterly Journal of the Royal Meteorological Society*, 132(621C):2905–2923. 74

Hamill, T. M. and Scheuerer, M. (2018). Probabilistic Precipitation Forecast Postprocessing Using Quantile Mapping and Rank-Weighted Best-Member Dressing. *Monthly Weather Review*, 146(12):4079–4098. 30, 99

Hamill, T. M. and Whitaker, J. S. (2006). Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. *Monthly Weather Review*, 134(11):3209–3229. xix, 29, 30, 34, 35, 60, 104

Hamill, T. M., Whitaker, J. S., and Wei, X. (2004). Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts. *Monthly Weather Review*, 132(6):1434–1447. 29, 30, 33

Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36. 82

Hemri, S., Scheuerer, M., Pappenberger, F., Bogner, K., and Haiden, T. (2014). Trends in the predictive performance of raw ensemble weather forecasts. *Geophysical Research Letters*, 41(24):2014GL062472. 33

Herman, G. R. and Schumacher, R. S. (2016). Extreme Precipitation in Models: An Evaluation. *Weather and Forecasting*, 31(6):1853–1879. 21

Hersbach, H. (2000). Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems. *Weather and Forecasting*, 15(5):559–570. 77, 79, 80, 123

Hersbach, H. (2016). The ERA5 Atmospheric Reanalysis. *AGU Fall Meeting Abstracts*, 33. 181, 189

Homar, V., Romero, R., Ramis, C., and Alonso, S. (2002). Numerical study of the October 2000 torrential precipitation event over eastern Spain: Analysis of the synoptic-scale stationarity. *Annales Geophysicae*, 20(12):2047–2066. 8

Hopson, T. M. and Webster, P. J. (2010a). A 1–10-Day Ensemble Forecasting Scheme for the Major River Basins of Bangladesh: Forecasting Severe Floods of 2003–07. *Journal of Hydrometeorology*, 11(3):618–641. 30

Hopson, T. M. and Webster, P. J. (2010b). A 1–10-Day Ensemble Forecasting Scheme for the Major River Basins of Bangladesh: Forecasting Severe Floods of 2003–07. *Journal of Hydrometeorology*, 11(3):618–641. 99

Houze, R. A. (2012). Orographic effects on precipitating clouds. *Reviews of Geophysics*, 50(1). 11

Jewson, S. (2003). Moment based methods for ensemble assessment and calibration. *arXiv:physics/0309042*. 32

Jolliffe, I. T. and Stephenson, D. B. (2012). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. John Wiley & Sons. 80

Juras, J. (2000). Comments on "Probabilistic Predictions of Precipitation Using the ECMWF Ensemble Prediction System". *Weather and Forecasting*, 15(3):365–366. 132

Kain, J. S. and Fritsch, J. M. (1993). Convective Parameterization for Mesoscale Models: The Kain-Fritsch Scheme. In Emanuel, K. A. and Raymond, D. J., editors, *The Representation of Cumulus Convection in Numerical Models*, Meteorological Monographs, pages 165–170. American Meteorological Society, Boston, MA. 44

Katz, R. W. and Ehrendorfer, M. (2006). Bayesian Approach to Decision Making Using Ensemble Weather Forecasts. *Weather and Forecasting*, 21(2):220–231. 116

Keil, C. and Craig, G. C. (2007). A Displacement-Based Error Measure Applied in a Regional Ensemble Forecasting System. *Monthly Weather Review*, 135(9):3248–3259. 27

Keil, C. and Craig, G. C. (2009). A Displacement and Amplitude Score Employing an Optical Flow Technique. *Weather and Forecasting*, 24(5):1297–1308. 27

Khodayar, S., Raff, F., Kalthoff, N., and Bock, O. (2016). Diagnostic study of a high-precipitation event in the Western Mediterranean: Adequacy of current operational networks. *Quarterly Journal of the Royal Meteorological Society*, 142(S1):72–85. 9

Kuo, H. L. (1974). Further Studies of the Parameterization of the Influence of Cumulus Convection on Large-Scale Flow. *Journal of the Atmospheric Sciences*, 31(5):1232–1240. 45

Lack, S. A., Limpert, G. L., and Fox, N. I. (2010). An Object-Oriented Multiscale Verification Scheme. *Weather and Forecasting*, 25(1):79–92. 27

Laing, A. (2015). MESOSCALE METEOROLOGY | Mesoscale Convective Systems. In North, G. R., Pyle, J., and Zhang, F., editors, *Encyclopedia of Atmospheric Sciences (Second Edition)*, pages 339–354. Academic Press, Oxford. xviii, 8

Lalaurette, F. (2003). Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Quarterly Journal of the Royal Meteorological Society*, 129(594):3037–3057. 22

Lebeaupin, C., Ducrocq, V., and Giordani, H. (2006). Sensitivity of torrential rain events to the sea surface temperature based on high-resolution numerical forecasts. *Journal of Geophysical Research: Atmospheres*, 111(D12). 12

Leith, C. E. (1974). Theoretical Skill of Monte Carlo Forecasts. *Monthly Weather Review*, 102:409. 17

Lemcke, C. and Kruizinga, S. (1988). Model Output Statistics Forecasts: Three Years of Operational Experience in the Netherlands. *Monthly Weather Review*, 116(5):1077–1090. 33

Leoncini, G., Plant, R. S., Gray, S. L., and Clark, P. A. (2013). Ensemble forecasts of a flood-producing storm: Comparison of the influence of model-state perturbations and parameter modifications. *Quarterly Journal of the Royal Meteorological Society*, 139(670):198–211. 28

Lerch, S., Thorarinsdottir, T. L., Ravazzolo, F., and Gneiting, T. (2017). Forecaster's Dilemma: Extreme Events and Forecast Evaluation. *Statistical Science*, 32(1):106–127. 25

Lin, Y.-L., Chiao, S., Wang, T.-A., Kaplan, M. L., and Weglarz, R. P. (2001). Some Common Ingredients for Heavy Orographic Rainfall. *Weather and Forecasting*, 16(6):633–660. 72

Lionello, P., Malanotte-Rizzoli, P., Boscolo, R., Alpert, P., Artale, V., Li, L., Luterbacher, J., May, W., Trigo, R., Tsimplis, M., Ulbrich, U., and Xoplaki, E. (2006). The Mediterranean climate: An overview of the main characteristics and issues. In Lionello, P., Malanotte-Rizzoli, P., and Boscolo, R., editors, *Developments in Earth and Environmental Sciences*, volume 4 of *Mediterranean*, pages 1–26. Elsevier. xvii, 2, 3

Llasat-Botija, M., Llasat, M. C., and López, L. (2007). Natural Hazards and the press in the western Mediterranean region. *Advances in Geosciences*, 12:81–85. 3

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences*, 20:130–141. 17

Louis, J.-F. (1979). A parametric model of vertical eddy fluxes in the atmosphere. *Boundary-Layer Meteorology*, 17(2):187–202. 44

Lu, G. Y. and Wong, D. W. (2008). An adaptive inverse-distance weighting spatial interpolation technique. *Computers & Geosciences*, 34(9):1044–1055. 52

Ly, S., Charles, C., and Degré, A. (2011). Geostatistical interpolation of daily rainfall at catchment scale: The use of several variogram models in the Ourthe and Ambleve catchments, Belgium. *Hydrol. Earth Syst. Sci.*, 15(7):2259–2274. 49, 51

Ly, S., Charles, C., and Degré, A. (2013). Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale: a review. *BASE*. 49

Maddox, R. A. and Doswell, C. A. (1982). An Examination of Jet Stream Configurations, 500 mb Vorticity Advection and Low-Level Thermal Advection Patterns During Extended Periods of Intense Convection. *Monthly Weather Review*, 110(3):184–197. 7

Magnusson, L., Leutbecher, M., and Källén, E. (2008). Comparison between Singular Vectors and Breeding Vectors as Initial Perturbations for the ECMWF Ensemble Prediction System. *Monthly Weather Review*, 136(11):4092–4104. 18

Maraun, D. (2013). Bias Correction, Quantile Mapping, and Downscaling: Revisiting the Inflation Issue. *Journal of Climate*, 26(6):2137–2143. 30

Mason, I. (1982). A model for assesment of weather forecasts. *Australian Meteorological Magazine*, 30. 82

Mass, C. F., Ovens, D., Westrick, K., and Colle, B. A. (2002). Does increasing horizontal resolution produce more skillful forecasts? *Bulletin of the American Meteorological Society*, 83(3):407–430. 25, 140

Masson, V., Moigne, P. L., Martin, E., Faroux, S., Alias, A., Alkama, R., Belamari, S., Barbu, A., Boone, A., Bouyssel, F., Brousseau, P., Brun, E., Calvet, J.-C., Carrer, D., Decharme, B., Delire, C., Donier, S., Essaouini, K., Gibelin, A.-L., Giordani, H., Habets, F., Jidane, M., Kerdraon, G., Kourzeneva, E., Lafaysse, M., Lafont, S., Lebeaupin Brossier, C., Lemonsu, A., Mahfouf, J.-F., Marguinaud, P., Mokhtari, M., Morin, S., Pigeon, G., Salgado, R., Seity, Y., Taillefer, F., Tanguy, G., Tulet, P., Vincendon, B., Vionnet, V., and Voldoire, A. (2013). The SURFEXv7.2 land and ocean surface platform for coupled or offline simulation of earth surface variables and fluxes. *Geoscientific Model Development*, 6(4):929–960. 47

Matheson, J. E. and Winkler, R. L. (1976). Scoring Rules for Continuous Probability Distributions. *Management Science*, 22(10):1087–1096. 77

Mehta, A. V. and Yang, S. (2008). Precipitation climatology over Mediterranean Basin from ten years of TRMM measurements. In *Advances in Geosciences*, volume 17, pages 87–91. Copernicus GmbH. 2, 5

Messner, J. W., Mayr, G. J., Wilks, D. S., and Zeileis, A. (2014). Extending Extended Logistic Regression: Extended versus Separate versus Ordered versus Censored. *Monthly Weather Review*, 142(8):3003–3014. 35, 113, 114

Messner, J. W., Mayr, G. J., Zeileis, A., and Wilks, D. S. (2013). Heteroscedastic Extended Logistic Regression for Postprocessing of Ensemble Guidance. *Monthly Weather Review*, 142(1):448–456. 35, 113

Miglietta, M. M. and Rotunno, R. (2009). Numerical simulations of conditionally unstable flows over a mountain ridge. *Journal of the Atmospheric Sciences*, 66(7):1865–1885. 11

Miglietta, M. M. and Rotunno, R. (2014). Numerical simulations of sheared conditionally unstable flows over a mountain ridge. *Journal of the Atmospheric Sciences*, 71(5):1747–1762. 11

Miniscloux, F., Creutin, J. D., and Anquetin, S. (2001). Geostatistical Analysis of Orographic Rainbands. *Journal of Applied Meteorology*, 40(11):1835–1854. 5

Mittermaier, M., North, R., Semple, A., and Bullock, R. (2015). Feature-Based Diagnostic Evaluation of Global NWP Forecasts. *Monthly Weather Review*, 144(10):3871–3893. 27, 28

Mittermaier, M. P. (2013). A Strategy for Verifying Near-Convection-Resolving Model Forecasts at Observing Sites. *Weather and Forecasting*, 29(2):185–204. 25

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T. (1996). The ECMWF Ensemble Prediction System: Methodology and validation. *Quarterly Journal of the Royal Meteorological Society*, 122(529):73–119. 18

Murphy, A. H. (1990). Forecast verification: Its Complexity and Dimensionality. *Monthly Weather Review*, 119(7):1590–1601. 23

Murphy, A. H. and Winkler, R. L. (1987). A General Framework for Forecast Verification. *Monthly Weather Review*, 115(7):1330–1338. 23, 24, 80

Murphy, A. H. and Winkler, R. L. (1992). Diagnostic verification of probability forecasts. *International Journal of Forecasting*, 7(4):435–455. 23

Nachamkin, J. E. (2009). Application of the Composite Method to the Spatial Forecast Verification Methods Intercomparison Dataset. *Weather and Forecasting*, 24(5):1390–1400. 27

North, R., Trueman, M., Mittermaier, M., and Rodwell, M. J. (2013). An assessment of the SEEPS and SEDI metrics for the verification of 6 h forecast precipitation accumulations. *Meteorological Applications*, 20(2):164–175. 63

Nuissier, O., Ducrocq, V., Ricard, D., Lebeaupin, C., and Anquetin, S. (2008). A numerical study of three catastrophic precipitating events over southern France. I: Numerical framework and synoptic ingredients. *Quarterly Journal of the Royal Meteorological Society*, 134(630):111–130. 7, 15

Nuissier, O., Joly, B., Joly, A., Ducrocq, V., and Arbogast, P. (2011). A statistical downscaling to identify the large-scale circulation patterns associated with heavy precipitation events over southern France. *Quarterly Journal of the Royal Meteorological Society*, 137(660):1812–1827. xviii, 15, 16, 182, 191

Ogura, Y. and Liou, M.-T. (1980). The Structure of a Midlatitude Squall Line: A Case Study. *Journal of the Atmospheric Sciences*, 37(3):553–567. 7

Palmer, T. (2019). The ECMWF ensemble prediction system: Looking back (more than) 25 years and projecting forward 25 years. *Quarterly Journal of the Royal Meteorological Society*, 145(S1):12–24. 18

Palmer, T., Buizza, R., Doblas-Reyes, F., Jung, T., Leutbecher, M., Shutts, G., Steinheimer, M., and Weisheimer, A. (2009). Stochastic parametrization and model uncertainty. (598):42. 20

Palmer, T. N. (2000). Predicting uncertainty in forecasts of weather and climate. *Reports on Progress in Physics*, 63(2):71–116. 20

Palmer, T. N., Gelaro, R., Barkmeijer, J., and Buizza, R. (1998). Singular Vectors, Metrics, and Adaptive Observations. *Journal of the Atmospheric Sciences*, 55(4):633–653. 43

Pergaud, J., Masson, V., Malardel, S., and Couvreux, F. (2009). A Parameterization of Dry Thermals and Shallow Cumuli for Mesoscale Numerical Weather Prediction. *Boundary-Layer Meteorology*, 132(1):83. 44

Peters, J. M. and Roebber, P. J. (2014). Synoptic Control of Heavy-Rain-Producing Convective Training Episodes. *Monthly Weather Review*, 142(7):2464–2482. 16

Piani, C., Haerter, J. O., and Coppola, E. (2010a). Statistical bias correction for daily precipitation in regional climate models over Europe. *Theoretical and Applied Climatology*, 99(1-2):187–192. 99

Piani, C., Weedon, G. P., Best, M., Gomes, S. M., Viterbo, P., Hagemann, S., and Haerter, J. O. (2010b). Statistical bias correction of global simulated daily precipitation and temperature for the application of hydrological models. *Journal of Hydrology*, 395(3):199–215. 99

Pinto, J. G., Ulbrich, S., Parodi, A., Rudari, R., Boni, G., and Ulbrich, U. (2013). Identification and ranking of extraordinary rainfall events over Northwest Italy: The role of Atlantic moisture. *Journal of Geophysical Research: Atmospheres*, 118(5):2085–2097. 5

Piriou, J.-M., Redelsperger, J.-L., Geleyn, J.-F., Lafore, J.-P., and Guichard, F. (2007). An Approach for Convective Parameterization with Memory: Separating Microphysics and Transport in Grid-Scale Equations. *Journal of the Atmospheric Sciences*, 64(11):4127–4139. 44

Plaut, G., Schuepbach, E., and Marut, D. (2001). Heavy precipitation events over a few Alpine sub-regions and the links with large-scale circulation, 1971-1995. *Climate Research*, 17:285–302. 15

Plaut, G. and Simonnet, E. (2001). Large-scale circulation classification, weather regimes, and local climate over France, the Alps and Western Europe. *Climate Research - CLIMATE RES*, 17:303–324. 15

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, 133(5):1155–1174. 30

Ramis, C., Homar, V., Amengual, A., Romero, R., and Alonso, S. (2013). Daily precipitation records over mainland Spain and the Balearic Islands. *Natural Hazards and Earth System Sciences*, 13(10):2483–2491. 3, 5

Raveh-Rubin, S. and Wernli, H. (2015). Large-scale wind and precipitation extremes in the Mediterranean: A climatological analysis for 1979–2012. *Quarterly Journal of the Royal Meteorological Society*, 141(691):2404–2417. xvii, 4

Raynaud, L., Berre, L., and Desroziers, G. (2012). Accounting for model error in the Météo-France ensemble data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 138(662):249–262. 42

Ricard, D., Ducrocq, V., and Auger, L. (2011). A Climatology of the Mesoscale Environment Associated with Heavily Precipitating Events over a Northwestern Mediterranean Area. *Journal of Applied Meteorology and Climatology*, 51(3):468–488. xviii, 3, 8, 11, 13, 14, 15

Roberts, N. M. and Lean, H. W. (2008). Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Monthly Weather Review*, 136(1):78–97. 25

Romero, R., Doswell, C. A., and Ramis, C. (2000). Mesoscale Numerical Study of Two Cases of Long-Lived Quasi-Stationary Convective Systems over Eastern Spain. *Monthly Weather Review*, 128(11):3731–3751. 7, 8

Romero, R., Ramis, C., and Guijarro, J. A. (1999). Daily rainfall patterns in the Spanish Mediterranean area: An objective classification. *International Journal of Climatology*, 19(1):95–112. 5

Rosner, B., Spiegelman, D., and Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: The case of multiple covariates measured with error. *American Journal of Epidemiology*, 132(4):734–745. 115

Rosner, B., Willett, W. C., and Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8(9):1051–1069. 115

Rossa, A., Nurmi, P., and Ebert, E. (2008). Overview of methods for the verification of quantitative precipitation forecasts. In Michaelides, S., editor, *Precipitation: Advances in Measurement, Estimation and Prediction*, pages 419–452. Springer Berlin Heidelberg, Berlin, Heidelberg. xxx, xxxv, 25

Roulin, E. and Vannitsem, S. (2011). Postprocessing of Ensemble Precipitation Predictions with Extended Logistic Regression Based on Hindcasts. *Monthly Weather Review*, 140(3):874–888. 25, 35, 37, 113, 114, 115, 116, 132

Roulston, M. S. and Smith, L. A. (2002). Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review*, 130(6):1653–1660. 30, 116

Roulston, M. S. and Smith, L. A. (2003). Combining dynamical and statistical ensembles. *Tellus A: Dynamic Meteorology and Oceanography*, 55(1):16–30. 25, 31, 32, 60

Ruiz, J. J. and Saulo, C. (2012). How sensitive are probabilistic precipitation forecasts to the choice of calibration algorithms and the ensemble generation method? Part I: Sensitivity to calibration methods. *Meteorological Applications*, 19(3):302–313. 35

Sanchez, C., Williams, K. D., and Collins, M. (2016). Improved stochastic physics schemes for global weather and climate models. *Quarterly Journal of the Royal Meteorological Society*, 142(694):147–159. 20

Scheuerer, M. (2014). Probabilistic quantitative precipitation forecasting using Ensemble Model Output Statistics. *Quarterly Journal of the Royal Meteorological Society*, 140(680):1086–1096. xix, 32, 33, 36, 113, 182, 190

Scheuerer, M. and Hamill, T. M. (2015). Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions. *Monthly Weather Review*, 143(11):4578–4596. 29, 33, 60, 182, 190

Scheuerer, M. and Hamill, T. M. (2018). Generating Calibrated Ensembles of Physically Realistic, High-Resolution Precipitation Forecast Fields Based on GEFS Model Output. *Journal of Hydrometeorology*, 19(10):1651–1670. 182, 190

Scheuerer, M., Hamill, T. M., Whitin, B., He, M., and Henkel, A. (2017). A method for preferential selection of dates in the Schaake shuffle approach to constructing spatiotemporal forecast fields of temperature and precipitation. *Water Resources Research*, 53(4):3029–3046. 182, 190

Schmeits, M. J. and Kok, K. J. (2010). A Comparison between Raw Ensemble Output, (Modified) Bayesian Model Averaging, and Extended Logistic Regression Using ECMWF Ensemble Precipitation Reforecasts. *Monthly Weather Review*, 138(11):4199–4211. 29, 31, 35, 113, 114

Schumacher, R. S. and Davis, C. A. (2010). Ensemble-Based Forecast Uncertainty Analysis of Diverse Heavy Rainfall Events. *Weather and Forecasting*, 25(4):1103–1122. 21

Schwartz, C. S., Romine, G. S., Sobash, R. A., Fossell, K. R., and Weisman, M. L. (2015). NCAR's Experimental Real-Time Convection-Allowing Ensemble Prediction System. *Weather and Forecasting*, 30(6):1645–1654. 22

Sénési, S., Bougeault, P., Chèze, J.-L., Cosentino, P., and Thepenier, R.-M. (1996). The Vaison-La-Romaine Flash Flood: Mesoscale Analysis and Predictability Issues. *Weather and Forecasting*, 11(4):417–442. xxix, 13

Shepard, D. (1968). A Two-dimensional Interpolation Function for Irregularly-spaced Data. In *Proceedings of the 1968 23rd ACM National Conference*, ACM '68, pages 517–524, New York, NY, USA. ACM. 49

Shutts, G. (2005). A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Quarterly Journal of the Royal Meteorological Society*, 131(612):3079–3102. 20

Skok, G. and Roberts, N. (2016). Analysis of Fractions Skill Score properties for

random precipitation fields and ECMWF forecasts. *Quarterly Journal of the Royal Meteorological Society*, 142(700):2599–2610. 25

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., and Fraley, C. (2007). Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging. *Monthly Weather Review*, 135(9):3209–3220. xix, 31, 36

Sohn, K. T., Lee, J. H., Lee, S. H., and Ryu, C. S. (2005). Statistical prediction of heavy rain in South Korea. *Advances in Atmospheric Sciences*, 22(5):703–710. 33

Sokol, Z. (2003). MOS-Based Precipitation Forecasts for River Basins. *Weather and Forecasting*, 18(5):769–781. 33

Stauffer, R., Mayr, G. J., Messner, J. W., Umlauf, N., and Zeileis, A. (2017). Spatio-temporal precipitation climatology over complex terrain using a censored additive regression model. *International Journal of Climatology*, 37(7):3264–3275. 33

Stephenson, D. B. (2000). Use of the "odds ratio" for diagnosing forecast skill. *Weather and Forecasting*, 15(2):221–232. 63

Sukovich, E. M., Ralph, F. M., Barthold, F. E., Reynolds, D. W., and Novak, D. R. (2014). Extreme Quantitative Precipitation Forecast Performance at the Weather Prediction Center from 2001 to 2011. *Weather and Forecasting*, 29(4):894–911. 21

Taillardat, M., Fougères, A.-L., Naveau, P., and de Fondeville, R. (2019). Extreme events evaluation using CRPS distributions. *arXiv:1905.04022 [math, stat]*. 25

Talagrand, O., Vautard, R., and Strauss, B. (1997). Evaluation of probabilistic prediction systems. 83

Teegavarapu, R. S. V. and Chandramouli, V. (2005). Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology*, 312(1):191–206. 49

Tennant, W. and Beare, S. (2014). New schemes to perturb sea-surface temperature and soil moisture content in MOGREPS. *Quarterly Journal of the Royal Meteorological Society*, 140(681):1150–1160. 20

Thielen, J., Bogner, K., Pappenberger, F., Kalas, M., del Medico, M., and de Roo, A. (2009). Monthly-, medium-, and short-range flood warning: Testing the limits of predictability. *Meteorological Applications*, 16(1):77–90. 21

Toth, Z. and Kalnay, E. (1993). Ensemble Forecasting at NMC: The Generation of Perturbations. *Bulletin of the American Meteorological Society*, 74(12):2317–2330. 18

Toth, Z. and Kalnay, E. (1997). Ensemble Forecasting at NCEP and the Breeding Method. *Monthly Weather Review*, 125(12):3297–3319. 18

Trapero, L., Bech, J., Duffourg, F., Esteban, P., and Lorente, J. (2013a). Mesoscale numerical analysis of the historical November 1982 heavy precipitation event over Andorra (Eastern Pyrenees). *Natural Hazards and Earth System Sciences*, 13(11):2969–2990. 6

Trapero, L., Bech, J., Duffourg, F., Esteban, P., and Lorente, J. (2013b). Mesoscale numerical analysis of the historical November 1982 heavy precipitation event over Andorra (Eastern Pyrenees). *Natural Hazards and Earth System Sciences*, 13(11):2969–2990. 7

Venugopal, V., Basu, S., and Foufoula-Georgiou, E. (2005). A new metric for comparing precipitation patterns with an application to ensemble forecasts. *Journal of Geophysical Research: Atmospheres*, 110(D8). 27

Vié, B., Molinié, G., Nuissier, O., Vincendon, B., Ducrocq, V., Bouttier, F., and Richard, E. (2012). Hydro-meteorological evaluation of a convection-permitting ensemble prediction system for Mediterranean heavy precipitating events. *Natural Hazards and Earth System Sciences*, 12:2631– 2645. 22

Vincendon, B., Ducrocq, V., Nuissier, O., and Vié, B. (2011). Perturbation of convection-permitting NWP forecasts for flash-flood ensemble forecasting. *Natural Hazards and Earth System Sciences*, 11(5):1529–1544. 22

Vislocky, R. L. and Young, G. S. (1989). The Use of Perfect Prog Forecasts to

Improve Model Output Statistics Forecasts of Precipitation Probability. *Weather and Forecasting*, 4(2):202–209. 33

Voisin, N., Schaake, J. C., and Lettenmaier, D. P. (2010). Calibration and Downscaling Methods for Quantitative Ensemble Precipitation Forecasts. *Weather and Forecasting*, 25(6):1603–1627. 30, 35, 99

Vrac, M. and Yiou, P. (2010). Weather regimes designed for local precipitation modeling: Application to the Mediterranean basin. *Journal of Geophysical Research: Atmospheres*, 115(D12):D12103. 15

Walser, A., Lüthi, D., and Schär, C. (2004). Predictability of Precipitation in a Cloud-Resolving Model. *Monthly Weather Review*, 132(2):560–577. 36

Walser, A. and Schär, C. (2004). Convection-resolving precipitation forecasting and its predictability in Alpine river catchments. *Journal of Hydrology*, 288(1):57–73. 36

Wang, X. and Bishop, C. H. (2005). Improvement of ensemble reliability with a new dressing kernel. *Quarterly Journal of the Royal Meteorological Society*, 131(607):965–986. 31, 32

Weckwerth, T. M., Parsons, D. B., Koch, S. E., Moore, J. A., LeMone, M. A., Demoz, B. B., Flamant, C., Geerts, B., Wang, J., and Feltz, W. F. (2004). An Overview of the International H2O Project (IHOP_2002) and Some Preliminary Highlights. *Bulletin of the American Meteorological Society*, 85(2):253–278. 28

Weisheimer A. and Palmer T. N. (2014). On the reliability of seasonal climate forecasts. *Journal of The Royal Society Interface*, 11(96):20131162. 81

Weniger, M., Kapp, F., and Friederichs, P. (2017). Spatial verification using wavelet transforms: A review. *Quarterly Journal of the Royal Meteorological Society*, 143(702):120–136. 27

Wernli, H., Hofmann, C., and Zimmer, M. (2009). Spatial Forecast Verification Methods Intercomparison Project: Application of the SAL Technique. *Weather and Forecasting*, 24(6):1472–1484. xix, 27, 28, 29, 140

Wernli, H., Paulat, M., Hagen, M., and Frei, C. (2008). SAL—A Novel Quality Measure for the Verification of Quantitative Precipitation Forecasts. *Monthly Weather Review*, 136(11):4470–4487. i, iii, 27, 28, 29, 140

Whitaker, J. S. and Loughe, A. F. (1998). The Relationship between Ensemble Spread and Ensemble Mean Skill. *Monthly Weather Review*, 126(12):3292–3302. 84

Wilks, D. S. (2002). Smoothing forecast ensembles with fitted probability distributions. *Quarterly Journal of the Royal Meteorological Society*, 128(586):2821–2836. 60

Wilks, D. S. (2006). Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorological Applications*, 13(3):243–256. 32, 33

Wilks, D. S. (2009a). Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorological Applications*, 16(3):361–368. xix, xxiv, 33, 35, 60, 112, 113, 114

Wilks, D. S. (2009b). *Statistical Methods in the Atmospheric Sciences*. Number 91 in International Geophysics Series. Elsevier [u.a.], Amsterdam, 2. ed., [nachdr.] edition. OCLC: 845720508. xxi, 31, 62, 75, 111, 115, 132

Wilks, D. S. and Hamill, T. M. (2007). Comparison of Ensemble-MOS Methods Using GFS Reforecasts. *Monthly Weather Review*, 135(6):2379–2390. 33, 60

Wilson, L. J., Beauregard, S., Raftery, A. E., and Verret, R. (2007). Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging. *Monthly Weather Review*, 135(4):1364–1385. 31

Wood, A. W., Maurer, E. P., Kumar, A., and Lettenmaier, D. P. (2002). Long-range experimental hydrologic forecasting for the eastern United States. *Journal of Geophysical Research: Atmospheres*, 107(D20):ACL 6–1–ACL 6–15. 99

Yano, J.-I. and Jakubiak, B. (2016). Wavelet-based verification of the quantitative precipitation forecast. *Dynamics of Atmospheres and Oceans*, 74:14–29. 27

Zacharov, P., Rezacova, D., and Brozkova, R. (2013). Evaluation of the QPF of convective flash flood rainfalls over the Czech territory in 2009. *Atmospheric Research*, 131:95–107. 28

Zepeda-Arce, J., Foufoula-Georgiou, E., and Droegemeier, K. K. (2000). Space-time rainfall organization and its role in validating quantitative precipitation forecasts. *Journal of Geophysical Research: Atmospheres*, 105(D8):10129–10146. 25

Zhao, T., Bennett, J. C., Wang, Q. J., Schepen, A., Wood, A. W., Robertson, D. E., and Ramos, M.-H. (2017). How Suitable is Quantile Mapping For Postprocessing GCM Precipitation Forecasts? *Journal of Climate*, 30(9):3185–3196. 104

Zhu, Y. and Luo, Y. (2014). Precipitation Calibration Based on the Frequency-Matching Method. *Weather and Forecasting*, 30(5):1109–1124. 60

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Mylne, K. (2002). The economic value of ensemble-based weather forecasts. *Bulletin of the American Meteorological Society*, 83(1):73–84. 74

Zimmer, M., Wernli, H., Frei, C., and Hagen, M. (2008). Feature-based verification of deterministic precipitation forecasts with SAL during COPS. In *Proceedings from the MAP D-Phase Scientific Meeting in Bologna, Italy*, pages 116–121. 28

Zsótér, E. (2006). Recent developments in extreme weather forecasting. (107):8–17. 22

Prévisibilité des épisodes méditerranéens de pluies intenses à l'aide d'un jeu de données de 30 ans de prévisions rétrospectives

Doctorat de l'Université de Toulouse III - Paul Sabatier

Soutenue le 12 décembre 2019

Auteur : Matteo PONZANO

Directeur de thèse : Laurent DESCAMPS

Co-directeur de thèse : Bruno JOLY

Résumé : Le sud-est de la France est une région particulièrement propice à l'occurrence de crues torrentielles associées à des événements de pluies très intenses. Cette thèse se focalise sur la prévisibilité des ces évènements, en reposant sur l'utilisation d'une base de prévisions rétrospectives (*reforecast*) par un système dérivé du modèle de prévision d'ensemble opérationnel PEARP. Une première partie de l'étude est consacrée à son évaluation du système PEARP et à l'utilisation des techniques de post-traitement ou calibrage pour améliorer ses performances. Une technique est basée sur une méthode de quantile mapping et la seconde sur une méthode de régression logistique étendue, appliquées chacune sur le *reforecast*. La deuxième technique est ensuite appliquée au système opérationnel. La dernière partie de l'étude a été consacrée à l'utilisation d'une métrique de vérification basée sur l'identification de structures cohérentes ou objets de pluie.

Mots-clés : *pluies intenses, prévisibilité, PEARP, prévision d'ensemble, reforecast, calibration, vérification spatiale, SAL*

Discipline : Océan, Atmosphère, Climat

Unité de Recherche : Centre National de Recherche Météorologiques

Groupe de Modélisation et d'Assimilation pour la Prévision