Forest-based and semi-parametric methods for the postprocessing of rainfall ensemble forecasting

Maxime Taillardat, Anne-Laure Fougères, Philippe Naveau, Olivier Mestre

Forest-based methods and ensemble model output statistics for rainfall ensemble forecasting

Maxime Taillardat[1,2,3], Anne-Laure Fougères[2], Philippe Naveau[3], and Olivier Mestre[1]

[1]CNRM UMR 3589 and DirOP/COMPAS, Météo-France/CNRS, Toulouse, France

[2]Univ Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, 43 blvd. du 11 novembre 1918, F-69622 Villeurbanne cedex, France

[3]Laboratoire des Sciences du Climat et de l'Environnement (LSCE-CNRS-CEA-UVSQ-IPSL), Gif-sur-Yvette, France

maxime.taillardat@meteo.fr

Abstract

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

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

1 Introduction

1.1 Post-processing of ensemble forecasts

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

1.2 Forecasting and calibration of precipitation

Not all meteorological variables are equal in terms of forecast and calibration. In particular, Hemri et al. (2014) highlighted that rainfall forecasting represents a steep hill. In this study, we will focus on 6-h rainfall amounts in France because this is the unit of interest of the ensemble forecast system of Météo-France. For daily precipitation, extended logistic regression was frequently applied (see, e.g. Hamill et al., 2008; Roulin and Vannitsem, 2012; Ben Bouallègue, 2013). Bayesian Model Averaging techniques (Raftery et al., 2005; Sloughter et al., 2007) were also used in rainfall forecasting, but we will not cover them here because a gamma fit is often applied to cube root transformed precipitation accumulations and this complex transformation may not be adapted to 6h rainfall. Concerning analogs and EMOS techniques, they have been applied to calibrate daily rainfall (see Hamill and Whitaker, 2006; Scheuerer, 2014; Scheuerer and Hamill, 2015). As the QRF method in Taillardat et al. (2016) performed better than EMOS for temperatures and wind speeds, one may wonder if QRF could favourably compete with EMOS and analogs techniques for rainfall calibration. This question is particularly relevant because recent methodological advances have been made concerning random forests and quantile regressions. In particular, Athey et al. (2016) proposed an innovative way, called gradient forests (GF), of using forests to make quantile regression. In this context, we propose to implement and test this quantile regression GF method for rainfall calibration and compare it with other approaches, see Section 2.

1.3 Parametric probability density functions (pdf) of precipitation

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

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

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

$$f_{CGEV}(y) \quad = \quad \begin{cases} (1-\pi) \cdot g(y; \mu, \sigma, \xi), & \text{if } y > 0 \\ \pi, & \text{if } y = 0, \end{cases} \tag{2}$$

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

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

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

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

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

$$f_{EGP}(y) = \begin{cases} (1-\pi) \cdot \frac{\kappa}{\sigma} \cdot \{H_\xi(x/\sigma)\}^{\kappa-1} \cdot h_\xi(y/\sigma), & \text{if } y > 0 \\ \pi, & \text{if } y = 0, \end{cases} \quad (3)$$

where $h_\xi(.)$ is the pdf associated with $H_\xi(.)$. In contrast to (1) and (2), the probability weight π is not obtained by censoring, and it is just a parameter independent of $(\kappa, \sigma, \xi)^T$.

At this stage, we have three parametric pdfs, see (1) and (2) and (3), to implement a EMOS approach to 6-hour rainfall data, see Section 3. Besides comparing these three EMOS models, it is natural to wonder if QRF and GF methods could take advantage of these three parametric forms.

1.4 Coupling parametric pdfs with random forest approaches

A drawback of data driven approaches like QRF and GF is that their intrinsic non parametric nature make them useless to predict beyond the largest recorded rainfall. To circumvent this limit, we also propose to combine random forest techniques with a EGP pdf defined by (3), see Section 2.3. Hence, random forest-based post-processing techniques will be in compliance with EVT and this should be an interesting path to improve prediction behind the largest values of the sample at hand.

1.5 Outline

This article is organized as follows. In Section 2, we recall the basic ingredients to create quantile regression forests and gradient forests. In particular, we review the calibration process of the GF method recently introduced by Athey et al. (2016) for quantile regression. Then, we explain how these trees are combined with the EGP pdf defined by (3).

In Section 3, we propose to integrate the EGP pdf within a EMOS scheme.

The different approaches are implemented in Section 4 where the test bed dataset of 87 French weather stations and the French ensemble forecast system of Météo-France called PEARP (Descamps et al., 2014) is described. Then, we assess and compare each method with a special interest for heavy rainfall, see Section 5. The paper closes with a discussion in Section 6.

2 Quantile regression forests and gradient forests

2.1 Quantile regression forests

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

A random forest is an aggregation of randomized trees based on bootstrap aggregation on the one hand, and on classification and regression trees (CART) (Breiman, 1996; Breiman et al., 1984) on the other hand. These trees are built on a bootstrap copy of the samples by recursively maximizing a

splitting rule. Let $\mathcal{D}_0$ denote the group of observations to be divided into two subgroups, say $\mathcal{D}_1$ and $\mathcal{D}_2$. For each group, we can infer its homogeneity defined by

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

where $\overline{Y}(\mathcal{D}_j)$ corresponds to the sample mean in $\mathcal{D}_j$. To determine if this splitting choice is optimal, the homogeneities $v(\mathcal{D}_1)$ and $v(\mathcal{D}_2)$ are compared to the one of $\mathcal{D}_0$. For example, if wind speed is one predictor in X and dividing low and large winds could better explain rainfall, then the cutting value, say s, will be the one that maximizes

$$\mathcal{H}(\mathcal{D}_1, \mathcal{D}_2) = \max_{s \in \mathcal{E}^*} [v(\mathcal{D}_0) - v(\mathcal{D}_1) - v(\mathcal{D}_2)] \tag{4}$$

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

$$\widehat{F}(y|x) = \sum_{i=1}^{n} \omega_i(x) \mathbf{1}(\{Y_i \leq y\}), \tag{5}$$

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

2.2 Gradient forests

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

$$\Delta(\mathcal{D}_1, \mathcal{D}_2) = \sum_{j=1}^{2} \frac{-1}{|\{i : Y_i \in \mathcal{D}_j\}|} \left(\sum_{\{i: Y_i \in \mathcal{D}_j\}} \rho_i \right)^2 \tag{6}$$

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

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

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

2.3 Fitting a parametric form to QRF and GF trees

As mentioned in Section 1.4, the predicted cdf defined by (5) cannot predict values which are not in the learning sample. This can be a strong limitation if the learning sample sample is small or rare events are of interest or both. The GF method has the same issue. To parametrically model rainfall, the EGP

pdf defined by (3) appears to be a good candidate. It allows more flexibility in the fitting than CSG or CGEV. This distribution has four parameters, π, κ, σ and ξ, it is in compliance with EVT for low and heavy rainfalls and works well in practice (see, e.g. Naveau et al., 2016). In terms of inference, a simple and fast method-of-moment can be applied. Basically, probability weighted moments (PWM) of a given random variable, say Y, with survival function $\overline{F}(y) = \mathbf{P}(Y > y)$, can be expressed as (see, e.g. Hosking and Wallis, 1987)

$$\mu_r = \mathbf{E}([Y\overline{F}^r(Y)]) = \int_0^1 F^{-1}(q)(1-q)^r dq. \tag{7}$$

If Y follows a EGP pdf defined by (3), then we have

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

where $B(.,.)$ represents the beta function. Knowing the PWM triplet $(\mu_0, \mu_1, \mu_2)^T$ is equivalent to know the parameter vector $(\kappa, \sigma, \xi)^T$. Hence, we just need to estimate these three PWMs. For any given forest, it is possible to estimate the distribution of $[Y|X = x]$ by the empirical cdf $\widehat{F}(y|X = x)$ defined by (5). Then, we can plug it in (7) to get

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

This leads to the estimates of $(\kappa(x), \sigma(x), \xi(x))^T$ and consequently of $f(y|X = x)$ via Equation (3). Note that the probability of no rain $\pi(x)$ is just inferred by counting the number of dry events in the corresponding trees. In the following, this technique is called "EGP TAIL", despite the fact that the whole distribution is fitted from QRF and GF trees.

3 Ensemble model output statistics and EGP

In Section 1.3, three definitions of parametric pdfs were recalled. By regressing their parameters on the ensemble values, different EMOS models have been proposed for the CSG and CGEV pdfs defined by (1) and by (2), respectively. More precisely, Baran and Nemoda (2016) used the CSG pdf by letting the mean $\mu = \kappa\theta$ and variance $\sigma^2 = \kappa\theta^2$ depend linearly as functions of the raw ensemble values and their mean, respectively. The coefficients of this regression were estimated by miminizing the continuous ranked probability score (CRPS) (see, e.g. Scheuerer and Hamill, 2015; Hersbach, 2000). The same strategy can be applied to fit the CGEV pdf (see, e.g. Hemri et al., 2014). Scheuerer (2014) modelled the scale parameter σ in (2) as an affine function of the ensemble mean absolute deviation rather than of the raw ensemble mean or variance. Another point to emphasize is that the shape parameter ξ was considered invariant in space in Hemri et al. (2014).

In this section, we basically explain how an EMOS approach can be built with the EGP pdf defined by (3) and we now highlight common features and differences between the two EMOS with CSG and CGEV. The scale parameter σ^2 in (3) is estimated in the same way than for CGEV. The presence of the parameter κ allows an additional degree of freedom. The expectation of our EGP is mainly driven by the product $\kappa\sigma$. Consequently, we model κ as an affine function of the predictors divided by σ. As France has a diverse climate, it is not reasonable to assume a constant shape parameter among all locations, see the map in Figure 1. In addition, minimizing the CRPS to infer different shape parameters may be inefficient (see, e.g. Friederichs and Thorarinsdottir, 2012). To estimate ξ at each location, we simply use the PWM inference scheme described in Section 2.3. To complete the estimation of the parameters in (3), the probability π is modeled as an affine function on $[0, 1]$ of the raw ensemble probability of rain and affine function parameters are also estimated by CRPS minimization. The table 1 sums up the optimal estimation strategies that we have found for each distribution.

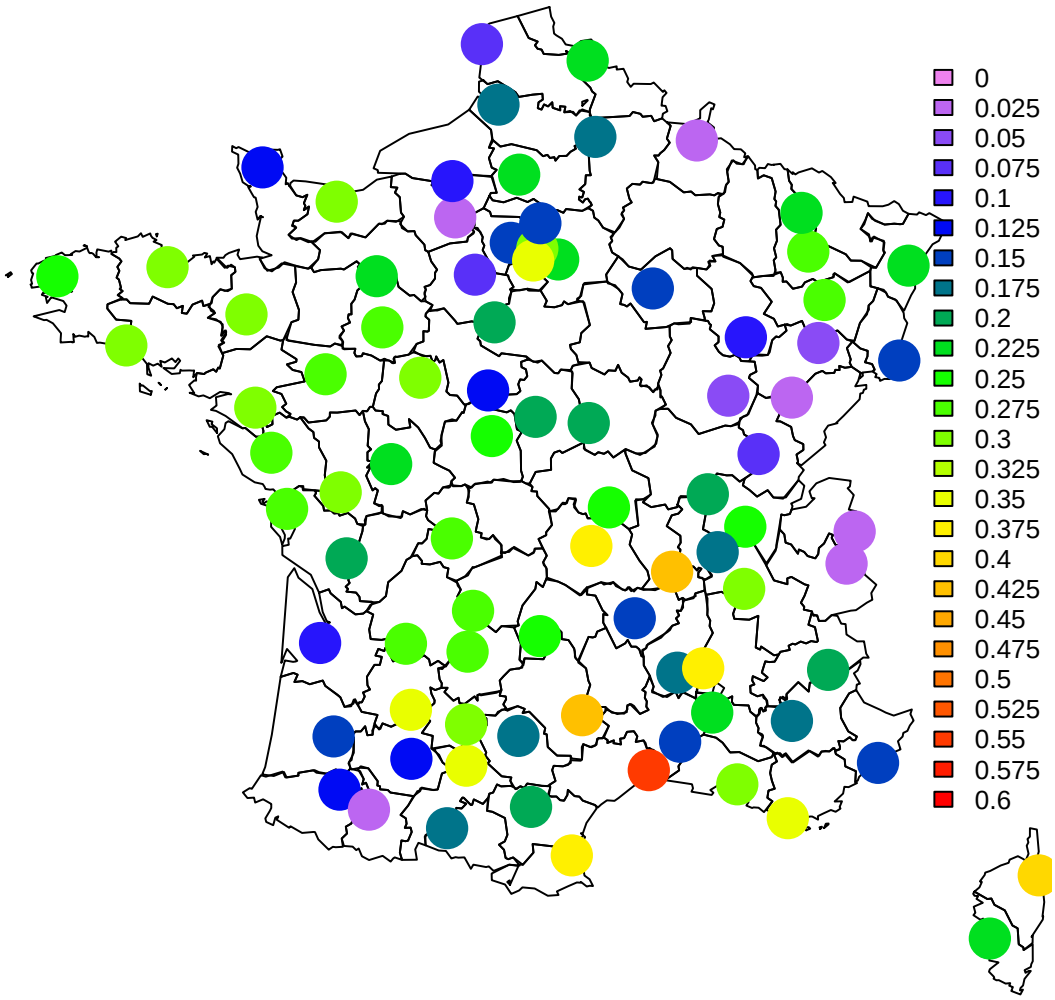

Figure 1: Spatial values of ξ among locations.

4 Case study on the PEARP ensemble prediction system

4.1 Data description

Our rainfall dataset corresponds to 6-h rainfall amounts produced by 87 French weather stations and the 35-member ensemble forecast system called PEARP (Descamps et al., 2014) at a 51-h lead time forecast. Our period of interest spans four years from 1 January 2012 to 31 December 2015.

4.2 Inferential details for EMOS and analogs

Verification has been made on this entire period. For a fair comparison each method has to be tuned optimally. EMOS uses all the data available for each day (4 years less the forecast day as a training period). The same strategy is used to fit the analogs method, see Appendix A for details on this approach.

Table 1: Optimal strategies for parameter estimation using CRPS minimization in the EMOS context.

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

QRF and GF employ a cross-validation method: each month of the 4 years is kept as validation data while the rest of the 4 years is used for learning. The tuning algorithm for EMOS is stopped after few iterations in order to avoid overfitting, as suggested in Scheuerer (2014) concerning the parameter estimations.

4.3 Sets of predictors used

We either use a subset of classical predictors (denoted by "C" in the rest of the paper) detailed in Table 2 or the whole set of available predictors as listed in Table 3.

Table 2: Subset "C" representing the most classical predictors.

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

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

4.4 Zooming on extremes

Finding a way to assess the quality of ensembles for extreme and rare events is quite difficult, as seen in Williams et al. (2014) in a comparison of ensemble calibration methods for extreme events. Weighted scoring rules can be adopted as done in Gneiting and Ranjan (2011); Lerch et al. (2017) but there are here two main issues. The ranking of compared methods depends on the weight function used, as already suggested in Gneiting and Ranjan (2011). Besides, giving a weight to such rare events avoid discriminant power of scoring rules, the same issue than for the Brier score (Brier, 1950). Moreover, reliability is not sound here since there are not enough extreme cases (by definition) to measure it. We have finally decided to focus on two ideas here, matching with forecasters' desires: first, what is the discriminant power of our forecasts for extreme events in terms of binary decisions ? Second, what is the potential risk of our ensemble to mismatch an extreme event ? The choice done in our study is discussed in Section 5.

Table 3: Set of all available predictors.

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

5 Results

Table 4 compares different metrics for all post-processing techniques which have been fitted to the 87 stations and averaged over 4 years of verification. Ten methods are competing: The raw ensemble, 4 analogs, 3 EMOS (3 different distributions using the set C), 2 forest-based methods (1 QRF and 1 GF) and 2 tail-extended forest-based methods (1 QRF and 1 GF). Scores used concern respectively (i) global performance (calibration and sharpness) measured by the CRPS; (ii) reliability performance, measured by the mean, the normalized variance and the entropy of the PIT histograms, denoted by Ω in the sequel; (iii) gain in CRPS compared to the raw ensemble, measured by the Skill of the CRPS using the raw ensemble as baseline. A brief summary about these measures is done in D, where references are also provided. And the boxplots showing rank histograms are in E.

According to Table 4, the raw ensemble is biased and underdispersive. The EMOS post-processed ensembles share with QRF and GF a good CRPS. Moreover, we can consider them as unbiased and mostly well-dispersed. The tail-extended methods get a lower CRPS, that can be explained by their skill for extreme events. Finally, the four analog methods show a quite poor CRPS compared to the raw ensemble, even if they exhibit reliability. Nevertheless we can notice that a weightning of the predictors, especially with a non-linear variable selection algorithm (Analogs_VSF), brings benefits to this method. This phenomenon can be explained by Figure 2, where the ROC curves are given for the event of rain. Consider a fixed threshold s and the contingency table associated to the predictor $\mathbf{1}\{rr6 > s\}$. Recall that the ROC curve then plots the probability of detection (or hit rate) as a function of the probability of false detection (or false alarm rate). A "good" prediction must maximize hit rates and minimize false

alarms (see, e.g. Jolliffe and Stephenson, 2012). Figure 2 explicitely shows the lack of resolution of the analogs technique. Incidently, we can also notice that the rain event discrimination is not improved by post-processed ensembles.

Table 4: Comparing performance statistics for different post-processing methods for 6-h rainfall forecasts in France. The mean CRPS estimations come from bootstrap replicates, the estimation error is under 6.1×10^{-3} for all methods.

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

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

Concerning extreme events, Figure 3 shows the benefit of the tail extension for forest-based methods. Note that we prefer to pay attention to the *value* of a forecast more than to its *quality*. According to Murphy (1993), the *value* can be defined as the ability of the forecast to help users to take better decisions. The *quality* of a forecast can be summarized by the area on the *modelled* ROC curve (classically denoted by AUC), with some potential drawbacks exhibited by Lobo et al. (2008); Hand (2009). Zhu et al. (2002) made a link between optimal decision thresholds, value and cost/loss ratios. In particular, they show that the value of a forecast is maximized for the "climatological" threshold and equals the hit rate minus the false alarm rate which is the maximum of the Peirce Skill Score (Manzato, 2007). This value corresponds to the upper left corner of ROC curves, which is of main interest in terms of extremes verification, as explained in Section 4.4. Several features already seen on Figure 2 can be observed on Figure 3: analogs lack resolution and the other post-processed methods compete more or less favourably with the raw ensemble. Nonetheless, the other post-processing techniques stay better than the raw ensemble even for methods that cannot extrapolate observed values such as QRF and GF. Note that QRF is rather surprisingly better than EMOS techniques. Tail extension methods show their gain in a binary decision context.

6 Discussion

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

Tail extension of these non-parametric techniques generates ensembles more tailored for extremes catchment. However, reliability as well as resolution remain quite stable when extending the tail, so that our paradigm about verification (good extreme discrimination subject to satisfying overall performance)

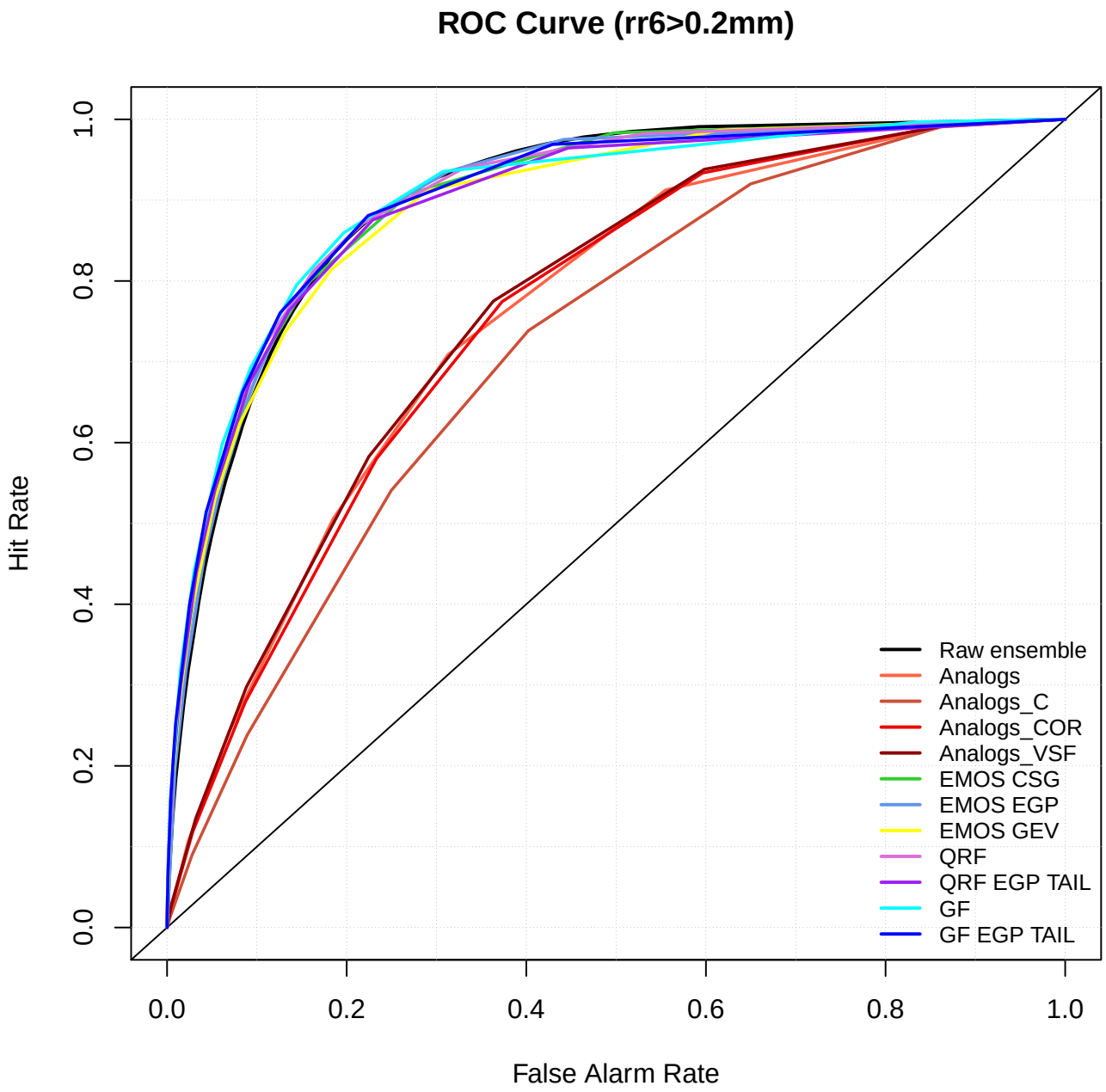

Figure 2: ROC Curves for the event of rain. A "good" prediction must maximize hit rate and minimize false alarms. The analogs method lacks resolution. We can notice that there is no improvement of post-processed methods compared to the raw ensemble.

remains.

One of the advantages of distribution-free calibration (analogs, QRF and GF) is that there is no assumption on the parameters to calibrate. This benefit is emphasized for rainfall amounts for which EMOS techniques have to be studied using different distributions. In this sense, the recent mixing method of Baran and Lerch (2016) looks appealing. A brand new alternative solution consists in working with (standardized) anomalies as done in Dabernig et al. (2016).

Another positive aspect of the forest-based methods is that there is no need of a predictor selection. Concerning the analogs method, our results suggest that the work of Genuer et al. (2010) could be a cheaper alternative to brute force algorithms like in Keller et al. (2017) for the weightning of predictors. For analogs techniques, we can notice that the complete set of predictors gives the best results. In

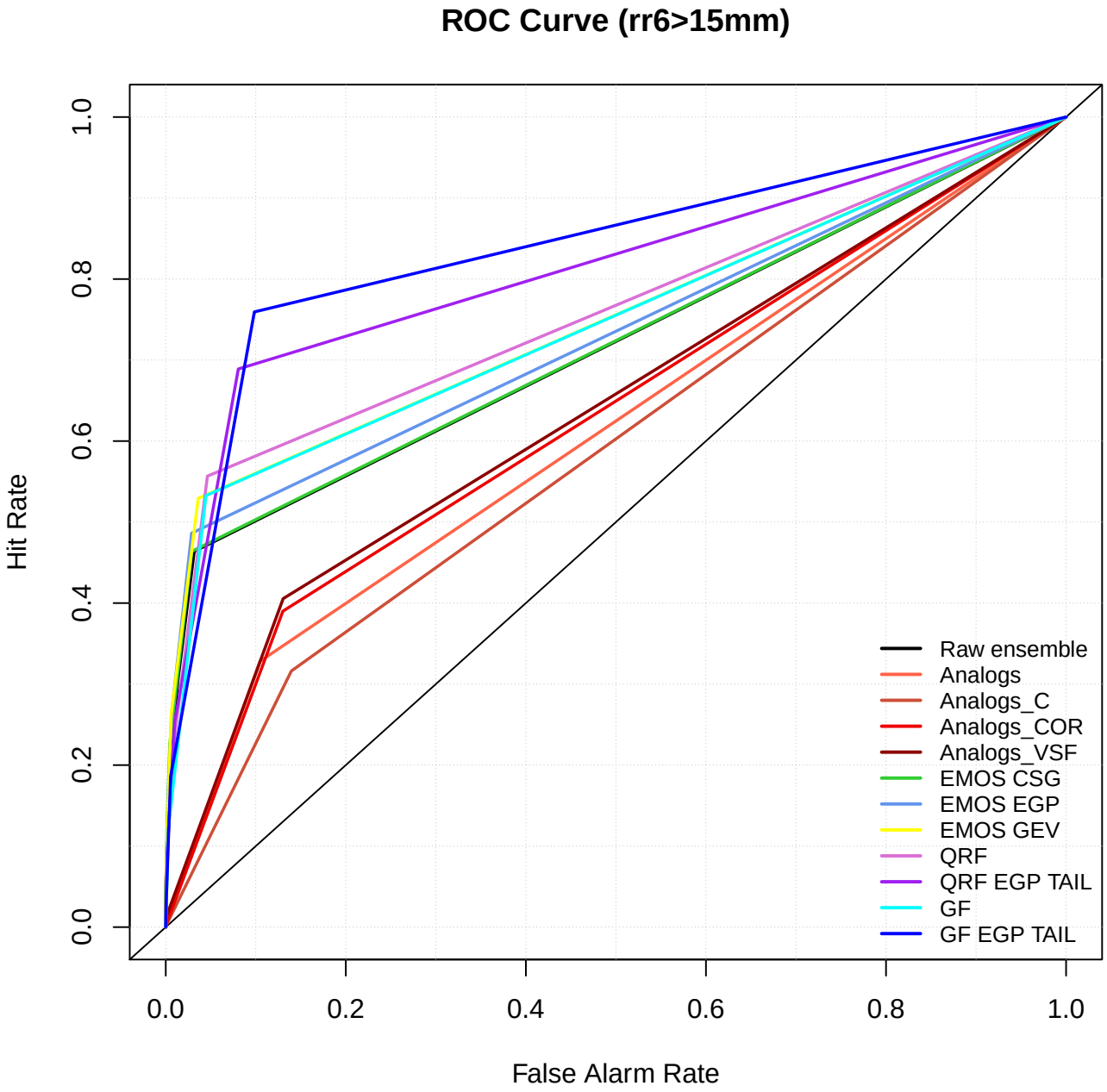

Figure 3: ROC Curves for the event of rain above 15mm. A "good" prediction must maximize hit rate and minimize false alarms. The analogs method lacks resolution. Tail extension methods show their gain in a binary decision context.

contrast, the choice of the set of predictors is still an ongoing issue for EMOS techniques regarding precipitation. For easier variables to calibrate, Messner et al. (2017) shows that some variable selection can be effective.

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

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

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

Acknowledgments

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

Funding information

LABEX MILYON, Investissements d'Avenir and DADA, ANR, Grant/Award Numbers: ANR-10-LABX-0070, ANR-11-IDEX-0007 and ANR-13-JS06-0007; EoCoE, Horizon2020, Grant/Award Number: 676629; A2C2, ERC, Grant/Award Number: 338965

References

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

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

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

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

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

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

— (2001) Random forests. *Machine learning*, **45**, 5–32.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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**, 9197–9205.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Naveau, P., Huser, R., Ribereau, P. and Hannart, A. (2016) Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection. *Water Resources Research*, **52**, 2753–2769. URL: `http://dx.doi.org/10.1002/2015WR018552`.

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

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

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

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

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

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

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

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

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

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**, 3209–3220.

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

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

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

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

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

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

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

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

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

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

Zhou, B. and Zhai, P. (2016) A new forecast model based on the analog method for persistent extreme precipitation. *Weather and Forecasting*, **31**, 1325–1341.

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

A Analogs method

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

analogs requires to choose both the set of predictors and the metric. Concerning the metric, several have been tried like the Euclidean or the Mahalanobis distance but they have been outperformed by the metric provided in Delle Monache et al. (2013):

$$\sum_{j=1}^{N_v} \frac{w_j}{\sigma_{f_j}} \sqrt{\sum_{i=-\tilde{t}}^{\tilde{t}} \left(F_{j,t+i} - A_{j,t'+i}\right)^2}, \tag{8}$$

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

B CRPS formula for EGP

The CRPS for the distribution F detailed in 3 is:

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

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

C Variable selection using random forests

We have seen that most parameters in EMOS and the distance used in analogs can be inferred using different sets of predictors. Contrary to the QRF and GF methods where the add of a useless predictor does not impact the predictive performance (since this predictor is never retained in the splitting rule), it can be misguiding for EMOS and analogs. We have therefore investigated some methods that keep the most informative meteorological variables and guarantee the best predictive performance. Our first choice was to use the well-known Akaike information criterion and the Bayesian information criterion (Akaike, 1998; Schwarz et al., 1978) but it resulted that the selection was not enough discriminant (too many predictors kept in our initial set). The algorithm of Genuer et al. (2010) has then been considered. Such an algorithm is appealing since it uses random forests (and we already have these objects from the QRF method) and it permits to keep predictors without redundancy of information. For example this algorithm eliminates correlated predictors even if they are informative. A reduced set of predictors (mostly 3 or 4) is thus obtained, which avoids misestimation generated by multicolinearity. The method of variable selection used here is one among plenty others. The interested reader in variable selection using random forests can refer to Genuer et al. (2010) for detailed explanations.

The variable selection algorithm is used to keep the first predictors (max 4 of them) that form the set of predictors for each location. Figure 4 shows the ranked frequency of each chosen predictor. Predictors never retained are not on this figure. We can see here that only one third of the predictors in A are retained at least in 10% of the cases. Moreover, predictors representing central and extreme tendencies are preferred. Some predictors appear that differ from rainfall amounts ; see CAPE, FX or HU. It is not

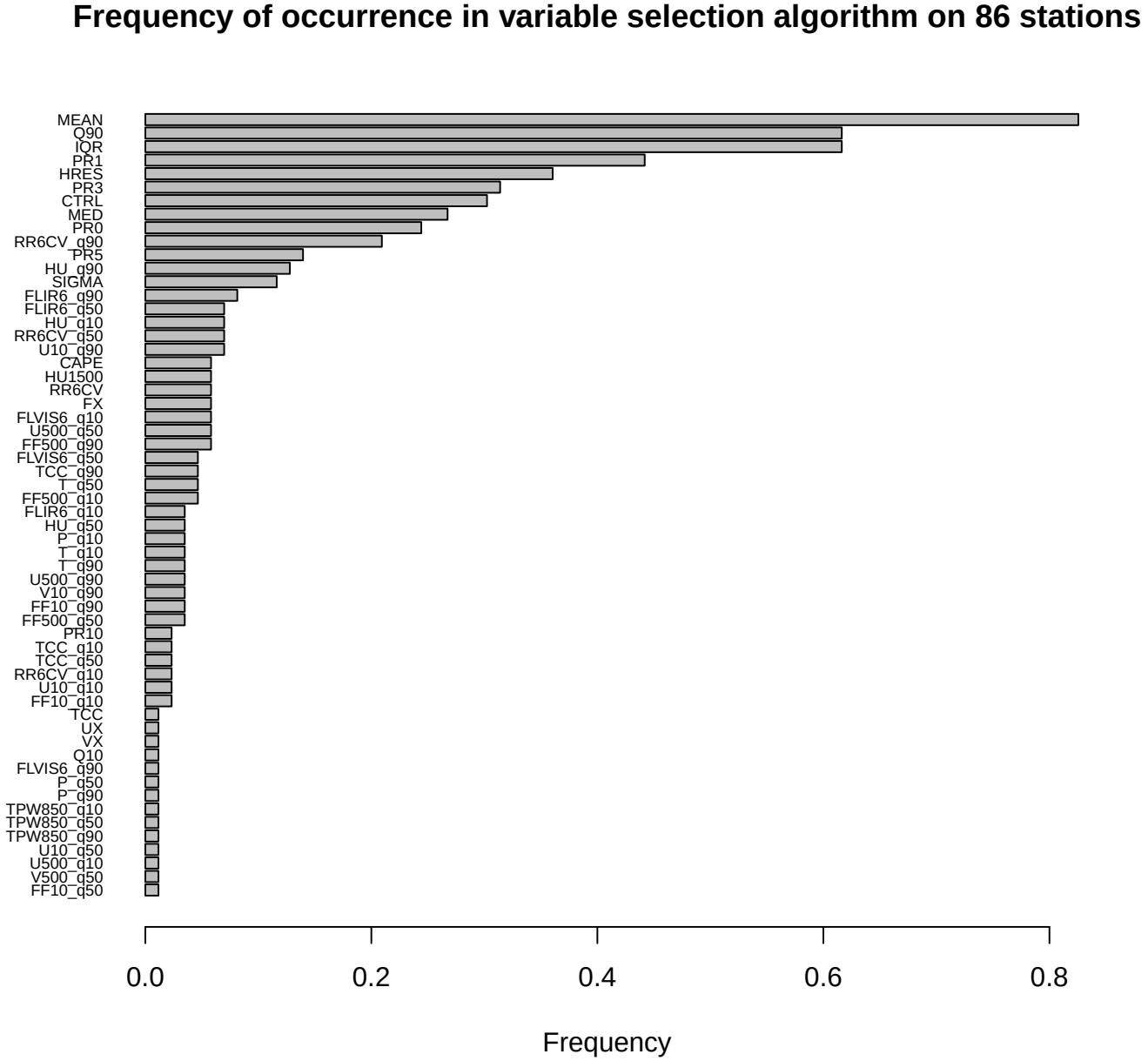

Figure 4: Frequency of predictors' occurrence in variable selection algorithm. Variables representing central and extreme tendencies are preferred. Some covariables like CAPE, FX or HU can be retained. It is interesting to see that only one third of the predictors of the set is taken more than in 10% of the cases.

surprising since these parameters are correlated with storms. It is not shown here but when the MEAN variable is not selected, either MED or CTRL stands in the set. This shows that the algorithm mostly selects just one information concerning central tendency and avoid potential correlations. So the results concerning the variable algorithm selection seem to be sound. Last but not least, one notices that the predictors of the set C are often chosen. This remark confirms both the robustness of the algorithm and the relevance of previous studies on precipitation concerning the choice of the predictors.

D Verification of ensembles

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

D.1 Reliability

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

Another tool used to assess calibration is the entropy:

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

For a calibrated system the entropy is maximum and equals 1. Tribus (1969) showed that the entropy is an indicator of reliability linked to the Bayesian psi-test. It is also a proper measure of reliability used in the divergence score described in Weijs et al. (2010); Roulston and Smith (2002).

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

D.2 Scoring rules

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

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

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

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

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

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

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

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

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

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

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

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

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

E Rank histograms boxplots

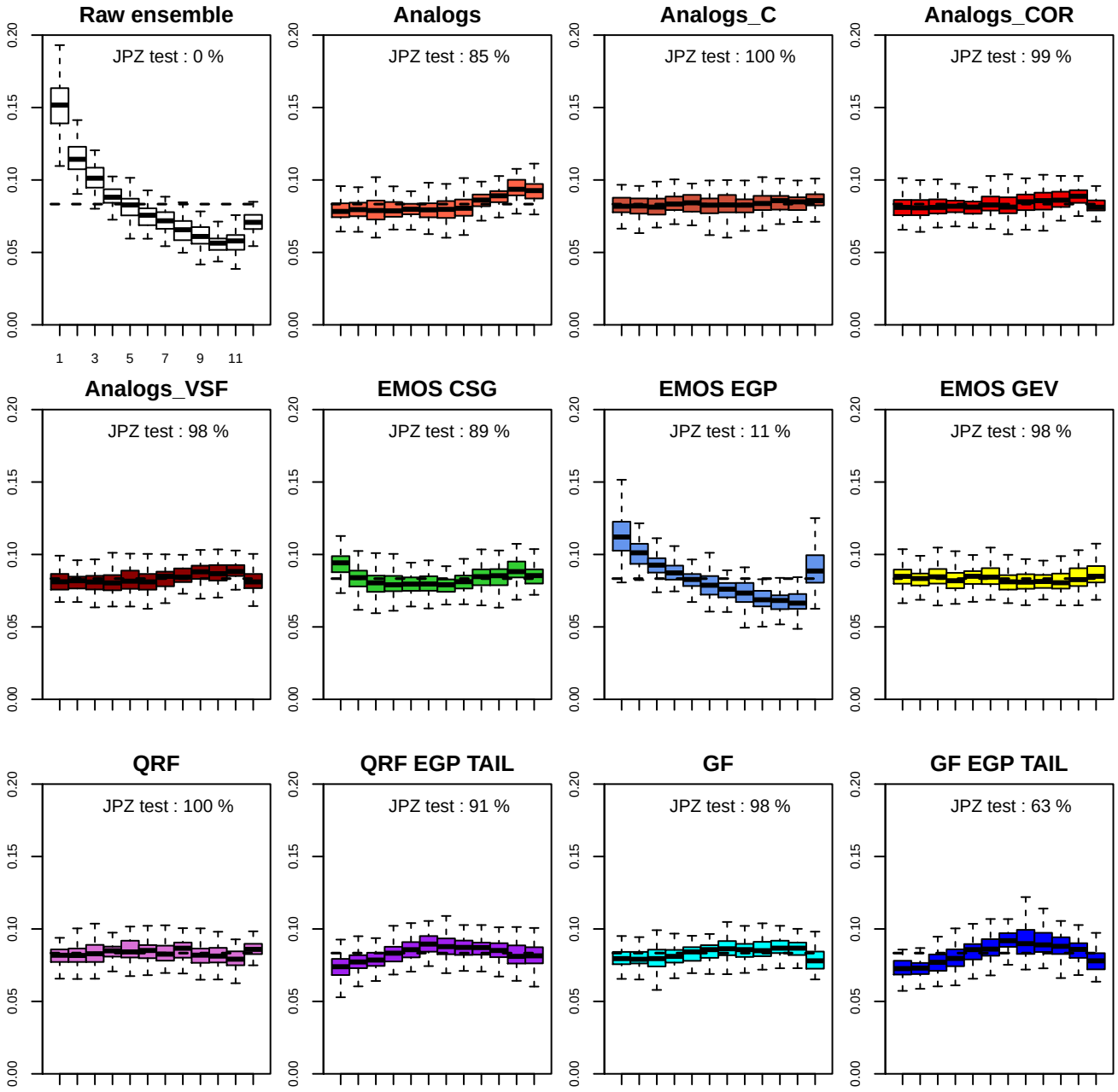

Figure 5: Boxplots of rank histograms for each technique according to the locations. The proportion of rank histograms for which the JPZ test does not reject the flatness hypothesis is also provided. The results confirm the Table 4.