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

MAXIME TAILLARDAT*

Centre National de Recherches Météorologiques, Météo-France, Toulouse, France

OLIVIER MESTRE AND MICHAËL ZAMO

Météo-France, Toulouse, France

PHILIPPE NAVEAU

Laboratoire des Sciences du Climat et de l'Environnement, CNRS, Saclay, France

(Manuscript received 24 July 2015, in final form 17 February 2016)

ABSTRACT

Ensembles used for probabilistic weather forecasting tend to be biased and underdispersive. This paper proposes a statistical method for postprocessing ensembles based on quantile regression forests (QRF), a generalization of random forests for quantile regression. This method does not fit a parametric probability density function (PDF) like in ensemble model output statistics (EMOS) but provides an estimation of desired quantiles. This is a nonparametric approach that eliminates any assumption on the variable subject to calibration. This method can estimate quantiles using not only members of the ensemble but any predictor available including statistics on other variables.

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

1. Introduction

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

Several techniques for the statistical postprocessing of ensemble model output have been developed to square up to these shortcomings. Local quantile regression and probit regression were used for probabilistic forecasts of precipitation by Bremnes (2004). Other techniques of regression like censored quantile regression have been applied to extreme precipitation (Friederichs and Hense 2007) and logistic regression was employed for probabilistic forecasts of precipitation (Hamill et al. 2008; Wilks 2009; Ben Bouallègue 2013). Two approaches are baseline in postprocessing techniques: the Bayesian model averaging (BMA; Raftery et al. 2005) and the ensemble model output statistics (EMOS; Gneiting et al.

Denotes Open Access content.

* Current affiliation: Laboratoire des Sciences du Climat et de l'Environnement, CNRS, Saclay, France.

Corresponding author address: Maxime Taillardat, Direction des Opérations/COMPAS, Météo-France, 42 avenue Gaspard Coriolis, Toulouse 31057, France.
E-mail: maxime.taillardat@meteo.fr

DOI: 10.1175/MWR-D-15-0260.1

2005). Whereas the BMA predictive distribution is a mixture of PDF depending on the variable to calibrate, the EMOS technique fits a single PDF from a raw ensemble. All parameters of theses PDFs are generally fitted on a sliding training period. In meteorology, BMA has been studied for many variables such as surface temperature (Raftery et al. 2005), quantitative precipitation (Sloughter et al. 2007), surface wind speed (Sloughter et al. 2010), or surface wind direction (Bao et al. 2010). Meanwhile EMOS techniques have been used for surface temperature (Gneiting et al. 2005; Hagedorn et al. 2008), quantitative precipitation (Scheuerer 2014), surface wind speed (Thorarinsdottir and Gneiting 2010; Baran and Lerch 2015), wind vectors (Pinson 2012; Schuhen et al. 2012), or peak wind (Friederichs and Thorarinsdottir 2012). More recently, Hemri et al. (2014) have applied EMOS to many variables.

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

The paper is organized as follows: in section 2 we describe the QRF technique in detail and we do a quick review of the EMOS technique. We explain how we verify ensemble forecasts. Guided by Gneiting et al. (2007) we apply tools like rank histograms and indices to quantify their behavior, in particular we introduce entropy for verification of reliability. Scoring rules like the continuous ranked probability score (CRPS) is also presented to assess both reliability and sharpness. Section 3 presents a case study comparing postprocessing techniques for surface temperature and surface wind speed over 87 French locations at 18 lead times using observations and the French ensemble forecast system of Météo-France called PEARP (Descamps et al. 2015). Data comprise 4 years between 1 January 2011 and 31 December 2014 using initializations at 1800 UTC. Section 4 shows general results of postprocessing techniques for studied variables. The QRF forecast and more particularly QRF forecasts based on multivariable predictors are better calibrated than EMOS forecasts and bring a real gain in comparison to this technique. The paper closes with a discussion in section 5.

2. Methods

a. Quantile regression forests

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

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

$$\Pr[X \leq x_\alpha] = \alpha. \tag{1}$$

While classical regression techniques allow us to estimate the conditional mean of a response variable, quantile regression allows us to estimate the conditional median or any other quantile of the response variable given a set of predictors (Koenker and Bassett 1978). Quantile regression such as QRF consists in building random forests from binary decision trees called classification and regression trees (CART), which are presented below. This is a nonlinear approach.

1) Decision trees (CART)

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

$$v(\mathscr{D}_i) = \sum_{y \in \mathscr{D}_i} [y - \overline{y}(\mathscr{D}_i)]^2. \tag{2}$$

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

$$H(t, \mathscr{D}_0) = v(\mathscr{D}_0) - [v(\mathscr{D}_1) + v(\mathscr{D}_2)] \geq 0. \tag{3}$$

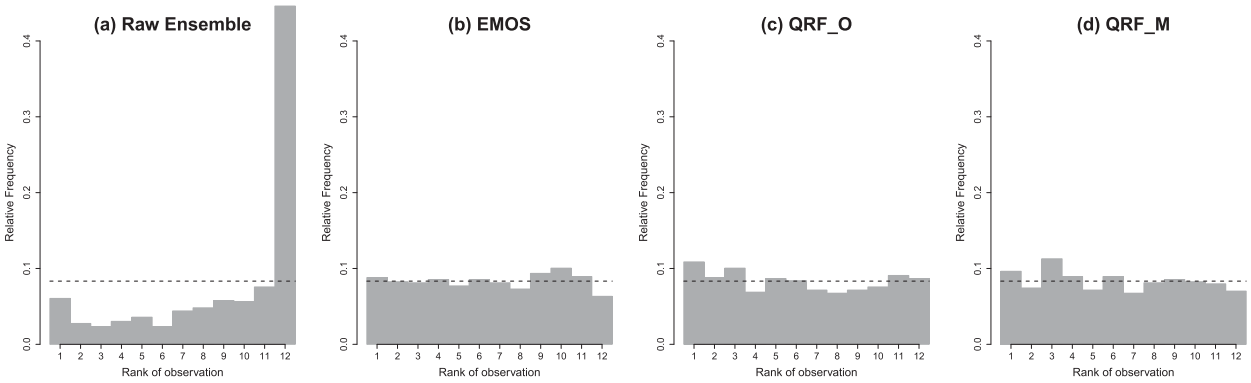

FIG. 1. Rank histograms for the Lyon airport for a 36-h forecast of surface temperature. The raw ensemble is clearly biased and underdispersed. QRF techniques are very efficient.

And we choose t such as

$$H(t, \mathscr{D}_0) = \max_{t \in \mathscr{E}}[H(t, \mathscr{D}_0)]. \quad (4)$$

Each resulting group is itself split into two, and so on until some stropping criterion is reached, which can be a minimum number of data or an insufficient decrease in the resulting groups' variance. Finally, for each final group (called leaves), the predicted value is the mean of observed values of the variable response belonging to the leaf. To avoid overfitting, binary trees are pruned at the splitting level that minimizes the squared error loss function estimated by cross validation. When one is faced with a new prediction situation, one follows the path in the tree with the value of the situation's predictors until a final leaf is reached. The forecast value is the mean of the predictand's values grouped in this leaf. Binary regression trees are easily interpretable because they can be represented by a decision tree, each node being the criterion used to split the data and each final leaf giving the predicted value. The interested reader can refer to Hastie et al. (2009, 305–312, 587–602) for detailed explanations.

2) BOOTSTRAP AGGREGATING (BAGGING)

According to the previous scheme, a tree can be a very unstable model (i.e., very dependent on the learning sample used for estimation). Breiman (1996) proposed to grow several trees and to average their predicted values to yield a more stable final prediction. This would require a lot of data in order to build enough independent trees. Since such a big amount of data is usually not available, bootstrap samples are usually used to build the trees. This means that artificial samples of data are simulated by randomly drawing with replacement among the original data. The complexity of the model is tuned with the number of bagged trees, and each individual tree is not pruned. The principle of bagging can be applied to other regression methods than binary trees.

3) RANDOM FORESTS

Since the binary trees used in bagging are built from the same data, they are not statistically independent and the variance of their mean cannot be indefinitely decreased. To make the bagged trees more independent, Breiman (2001) proposed to add another randomization

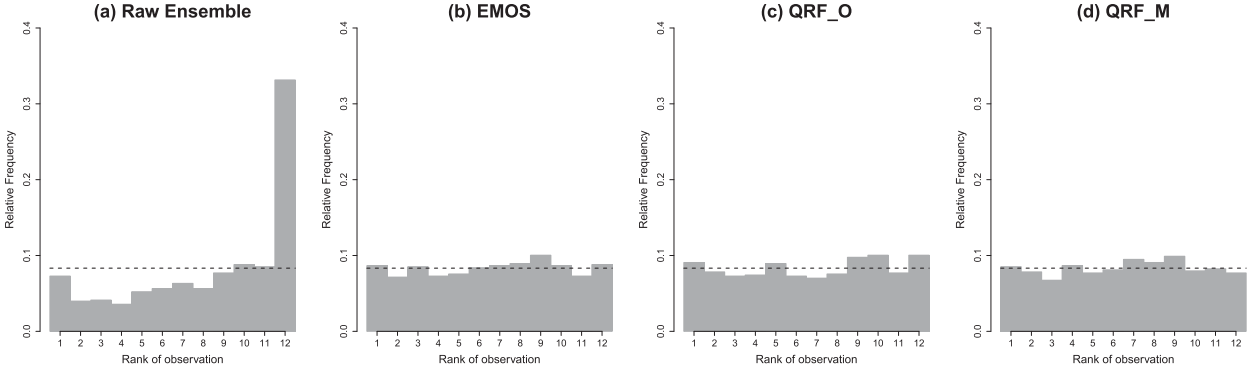

FIG. 2. Rank histograms for the Paris-Orly airport for a 36-h forecast of surface temperature. The raw ensemble is clearly biased and underdispersed. QRF techniques are very efficient.

TABLE 1. Results for surface temperature at two locations for a 36-h forecast. QRF_M performs better than other techniques and gives sharp ensembles.

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

step to bagging. Each split of each bagged tree is built on a random subset of the predictors. Hence, this method is called random forest. As in bagging, the overfitting problem is solved by tuning the number of trees.

4) QUANTILE REGRESSION FORESTS

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

5) MODEL FITTING

The QRF method is used with different inputs here. The first, called QRT_O, uses as predictors only those statistics on the variable. The second, called QRT_M, contains not only statistics on the variable to calibrate but also on other meteorological variables issued from the ensemble: this is a multivariable approach. The lists of predictors are given in appendix A. For these variants, one must fit the number of trees and the size of the leaves. For temperature, the final leaf size is set to 10 and the number of tree is set to 300, which is a good compromise between quality and computation speed. For

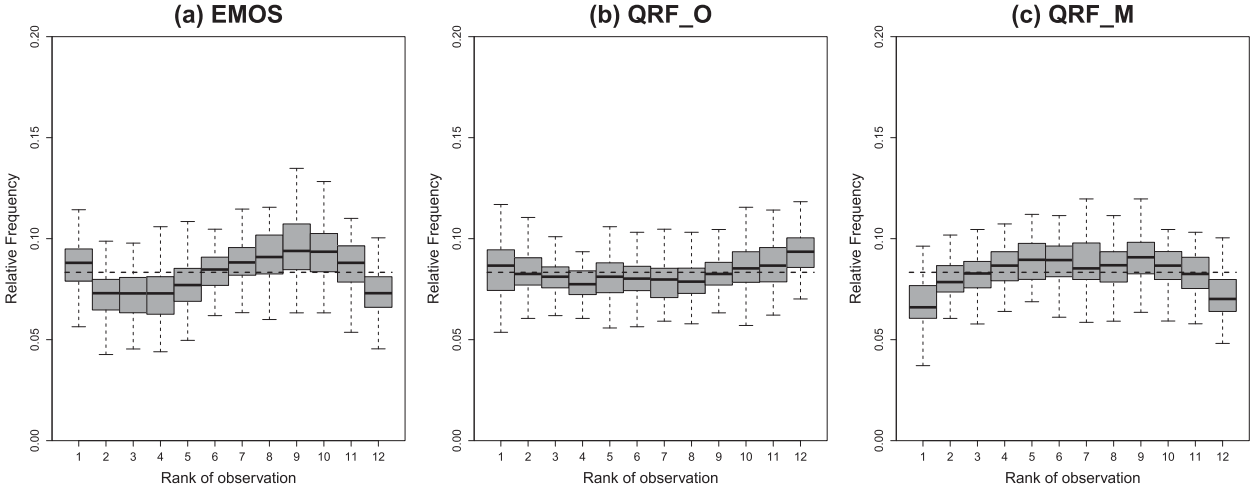

FIG. 3. Box plot of rank histograms for all locations for 36-h forecast of surface temperature. QRF_M tends to be a little overdispersed. There is a little overdispersion for EMOS.

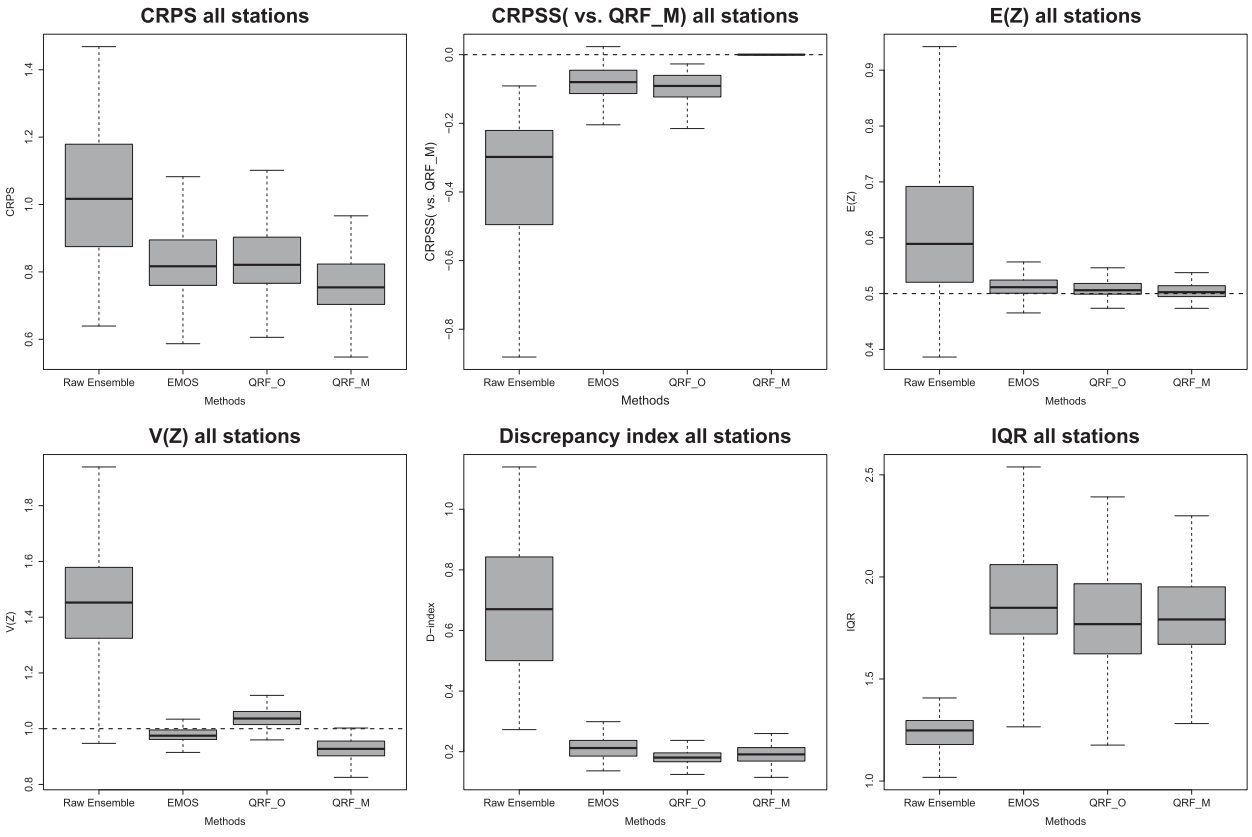

FIG. 4. Box plot of different scores for all locations for 36-h forecast of surface temperature. QRF_M technique has better CRPS for almost all stations according to the CRPS skill score. All calibrated ensembles are unbiased, reliable, and quite well dispersed.

wind speed, the final leaf size is 20 and the number of trees is set to 400. Note that these parameters are set empirically by means of cross validation (not shown here).

b. *Ensemble model output statistics*

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

MODEL FITTING

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

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

c. *Assessing sharpness and calibration*

Gneiting et al. (2007) propose to evaluate predictive performance based on *the paradigm of maximizing the sharpness of the predictive distributions subject to calibration*. Calibration refers to the statistical consistency between forecasts and observations. Also called reliability this is a joint property of predictions and events that materialize. Sharpness refers to the spread of predictive distributions and is a property of the forecasts

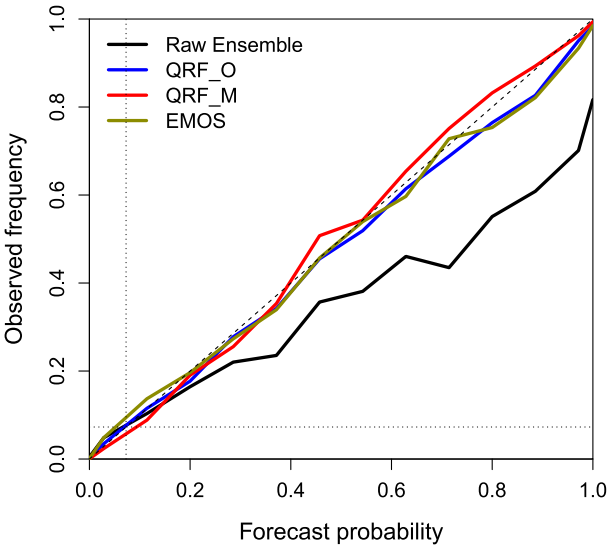

FIG. 5. Reliability diagram for probabilistic 36-h forecast of frost for all locations. Dotted lines represent climatology. Calibrated ensembles are almost perfect here.

only. For example, a climatological forecast would be reliable, but would have a poor sharpness.

1) SHARPNESS

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

2) THE RANK HISTOGRAM AND THE PIT HISTOGRAM

Rank histograms (RH), also called *Talagrand diagrams* were developed independently by Anderson (1996), Talagrand et al. (1997), and Hamill and Colucci (1997). We employ RH to check the reliability of an ensemble forecast or a set of quantiles. An RH is built by ranking observations according to associated forecasts. Reliability implies that each rank should be filled with the same probability. Calibrated ensemble prediction systems should result in a flat RH. The opposite is not true: a flat RH may not refer to a calibrated system (Hamill 2001). In a general way, a U-shaped histogram refers to underdispersion or conditional bias, a dome-shaped generally refers to overdispersion, while a non-symmetric histogram refers to bias. A PIT histogram is the continuous version of the RH and permits to check reliability between observations and a predictive distribution by calculating $Z' = F(Y)$, where Y is the observation and F is the CDF of the associated predictive distribution. Subject to calibration, the random variable Z' has a standard uniform distribution (Gneiting and Katzfuss 2014) and we can check ensemble bias by comparing $\mathbb{E}(Z')$ to 1/2 and ensemble dispersion by comparing the variance $\mathrm{var}(Z')$ to 1/12. We apply this approach to a RH with $K+1$ ranks using the discrete random variable $Z = [\mathrm{rank}(y) - 1]/K$. Subject to calibration, Z has a discrete standard uniform distribution with $\mathbb{E}(Z) = 1/2$ and a normalized variance of $\mathbb{V}(Z) = 12[K/(K+2)]\,\mathrm{var}(Z) = 1$.

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

$$\Delta = \sum_{i=1}^{K+1} \left| f_i - \frac{1}{K+1} \right| = \sum_{i=1}^{K+1} |\varepsilon_i| = \|\varepsilon\|_1, \tag{5}$$

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

We can complete this tool by checking $\|\varepsilon\|_2$ (quadratic index) or $\|\varepsilon\|_\infty$ (max index), which are more sensitive to bigger errors than Δ.

Another tool that we will use to assess calibration is the entropy, called ψ here:

$$\Omega = \frac{-1}{\log(K+1)} \sum_{i=1}^{K+1} f_i \log(f_i). \tag{6}$$

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

3) RELIABILITY DIAGRAM

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

d. *Scoring rules*

Following Gneiting et al. (2007), Gneiting and Raftery (2007), and Gneiting and Katzfuss (2014), scoring rules assign numerical scores to probabilistic forecasts and form attractive summary measures of predictive performance, since they address calibration and sharpness

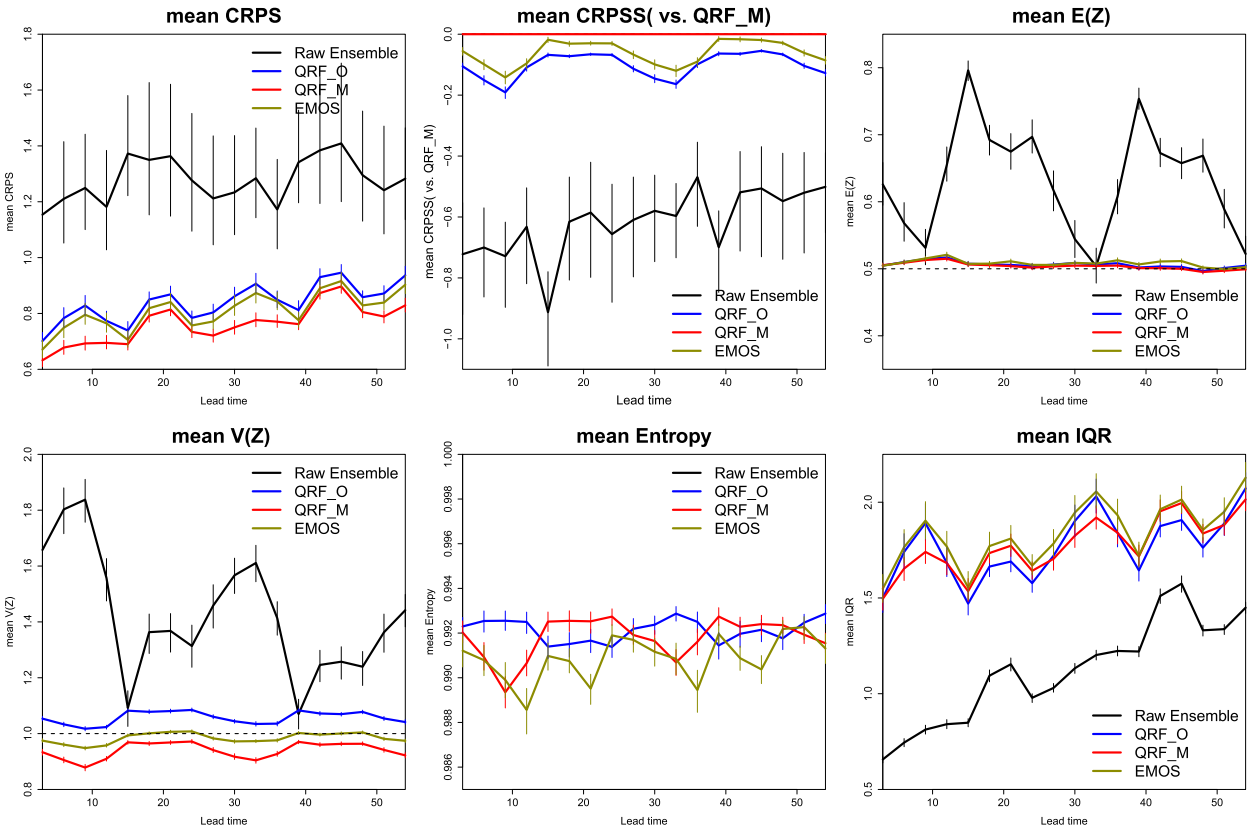

FIG. 6. Mean scores with 95% bootstrap confidence intervals for all locations across lead times for surface temperature. QRF_M is the best technique for CRPS and CRPSS. Calibrated ensembles are unbiased and in general better dispersed than raw ensembles. QRF techniques tend to provide more reliable forecasts than EMOS (the raw ensemble entropy is around 0.75). The raw ensemble is the sharpest, but it is not reliable.

simultaneously. These scores are usually taken to be negatively oriented and we wish to minimize them. A *proper* scoring rule is designed such that the expected value of the score is minimized when the observation is drawn from the same distribution than the predictive distribution.

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

$$\widehat{\mathrm{CRPS}}(F, y) = \frac{1}{K}\sum_{i=1}^{K}|x_i - y| - \frac{1}{2K(K-1)}\sum_{i=1}^{K}\sum_{j=1}^{K}|x_i - x_j|. \tag{7}$$

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

$$\mathrm{CRPSS}(A, B) = 1 - \frac{\mathrm{CRPS}_A}{\mathrm{CRPS}_B}. \tag{8}$$

The value of the continuous ranked probability skill score (CRPSS) will be positive if and only if system A is better than system B for the CRPS scoring rule.

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

3. Analysis of the French operational ensemble forecast system (PEARP)

We now compare QRF and EMOS techniques for lead times from 3 up to 54 h for forecasts of surface temperature and wind speed over 87 French stations using observations and the French ensemble forecast system of Météo-France called PEARP (Descamps et al. 2015). Data comprise 4 yr between 1 January 2011 and 31 December 2014 using initializations at 1800 UTC. Verification and results are made over the years 2013 and 2014. The aim of our study is to compare both techniques according to their specificities and advantages: on the one hand the QRF method is nonparametric so it needs a large data sample for learning, which is why we employed a cross-validation method (each month of years 2013 and 2014 are retained as validation data for testing the model, while all 4 years of data without the forecasted month are used for learning). On the other

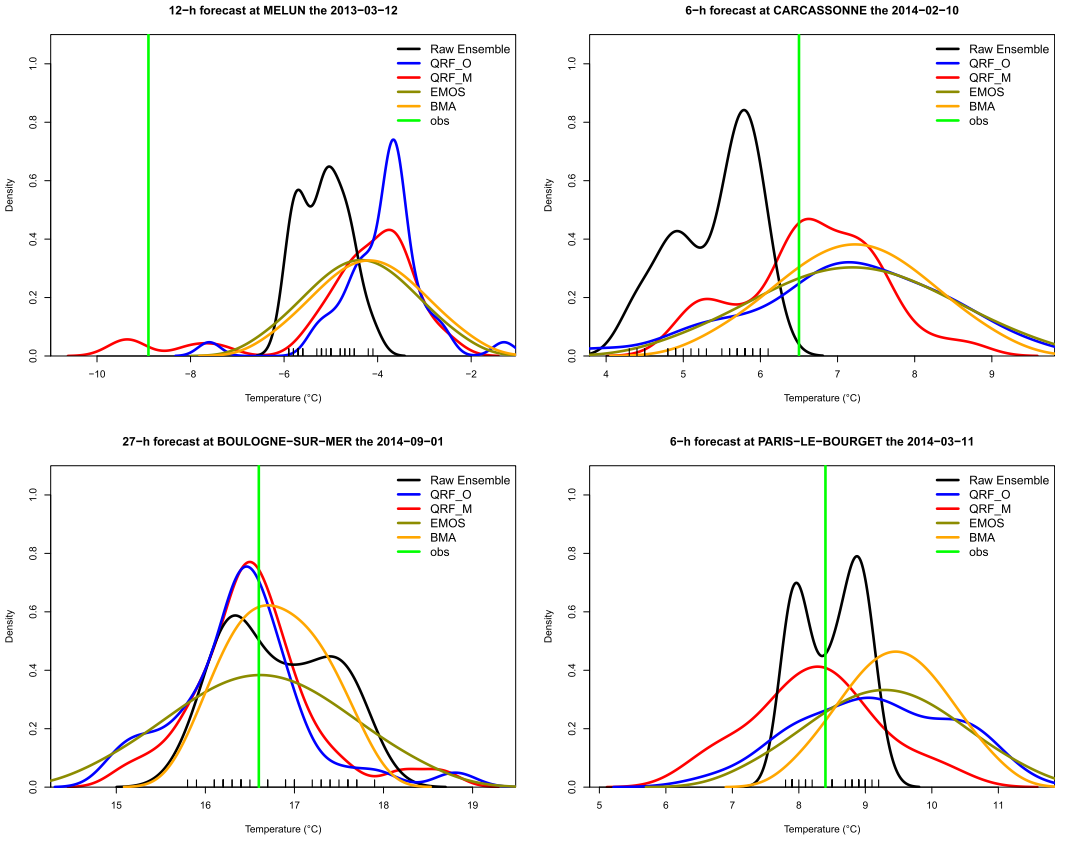

FIG. 7. Some forecasts for different meteorological situations where the QRF_M technique is useful for forecasters. (top left) QRF_M technique proposes cooler scenarios. (top right) The bimodality of raw ensemble is preserved. (bottom left) Bimodality is still conserved but a mode is preferred to the other. (bottom right) QRF_M technique proposes a unimodal PDF contrary to raw ensemble. The little segments on the x axis represents the 35 raw members: there are several members associated to the same temperature.

hand, a sliding period of the 40 last days prior the forecast output as in Gneiting et al. (2005), Schuhen et al. (2012), and Thorarinsdottir and Gneiting (2010) give good results for EMOS. But EMOS has to be tuned optimally for a fair comparison, which is why for temperature all the data available for each day (4 yr less the forecast day) with a seasonal dependence like in Hemri et al. (2014) are taken. For wind speed, a sliding period of 1 yr gives the best results for EMOS.

For verification, we choose for all methods to form a K-member ensemble from predictive CDFs by taking forecast quantiles at level $i/(K+1)$ for $i = 1, \ldots, K$,

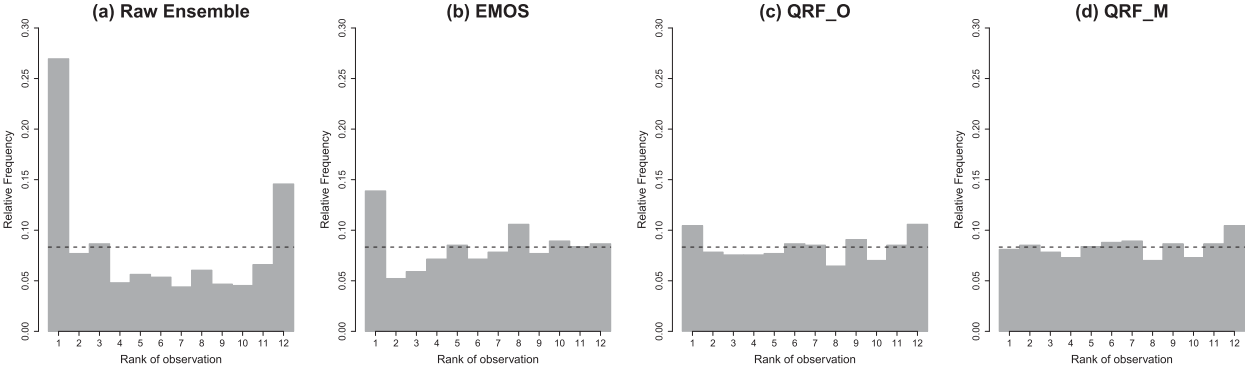

FIG. 8. Rank histograms for the Lyon airport for the 24-h forecast of surface wind speed. The raw ensemble is clearly biased and underdispersed. QRF techniques are very efficient.

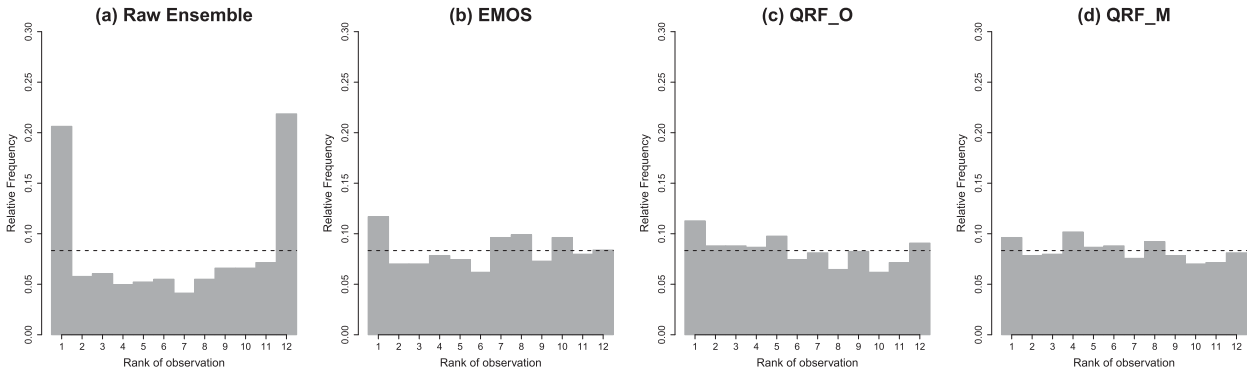

FIG. 9. Rank histograms for the Paris-Orly airport for the 24-h forecast of surface wind speed. This time, raw ensemble is not biased but still underdispersed.

respectively, to conciliate with PEARP raw ensemble here $K = 35$. So all scores are computed with 35 quantiles and rank histograms have 36 classes, but for graphical reasons we show RH computed on 12 ranks only (each group of 3 consecutive ranks are gathered as a single rank).

4. Results

a. Surface temperature

We now give results for surface temperature. We show an example for 36-h lead time (corresponding to 0600 UTC) at two locations which are Lyon and Paris-Orly airports in France. Figures 1 and 2 show RH for all presented methods. For both examples, the raw ensemble is biased and underdispersive whereas EMOS and QRF techniques show graphically good calibration. Table 1 confirms these first results. We can see that the raw ensemble is not reliable and has the worst CRPS. EMOS and QRF techniques are unbiased and dispersion is satisfying. In a general way, the lowest CRPS are for QRF_M. It is very interesting to notice that most of the time all indices of reliability (discrepancy index, quadratic index, max index, and entropy) exhibit the same rankings for the different models. Reliability for EMOS and QRF_O focuses only on the example of Paris-Orly. The discrepancy index shows a better reliability for QRF whereas other indexes penalize this. Thus, it is sometimes interesting to assess calibration with several tools.

Now let us focus on all stations for a 36-h lead time. Figure 3 shows RH for the three techniques where a box plot represents the distribution of a rank for all stations. Results are satisfying, all the RHs are unbiased, but we have a ''wavy'' RH for EMOS whereas the RH for QRF techniques seems to be better. Nevertheless we can assume a slightly U-shaped RH for QRF_O and a slightly dome-shaped for QRF_M to be signs of an unperfect dispersion. These first remarks are strengthened by Fig. 4 where we see that the three calibration techniques are unbiased and QRF techniques are a little more reliable than EMOS technique for the discrepancy index (we only show this index of reliability here according to our previous remarks on indices of reliability), but we can assume that results are quite mixed now. The diagnosis of spread ensembles exhibits a slight

TABLE 2. Results for surface wind speed at two locations for a 24-h forecast. QRF_M performs better than other techniques and gives sharp ensembles.

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

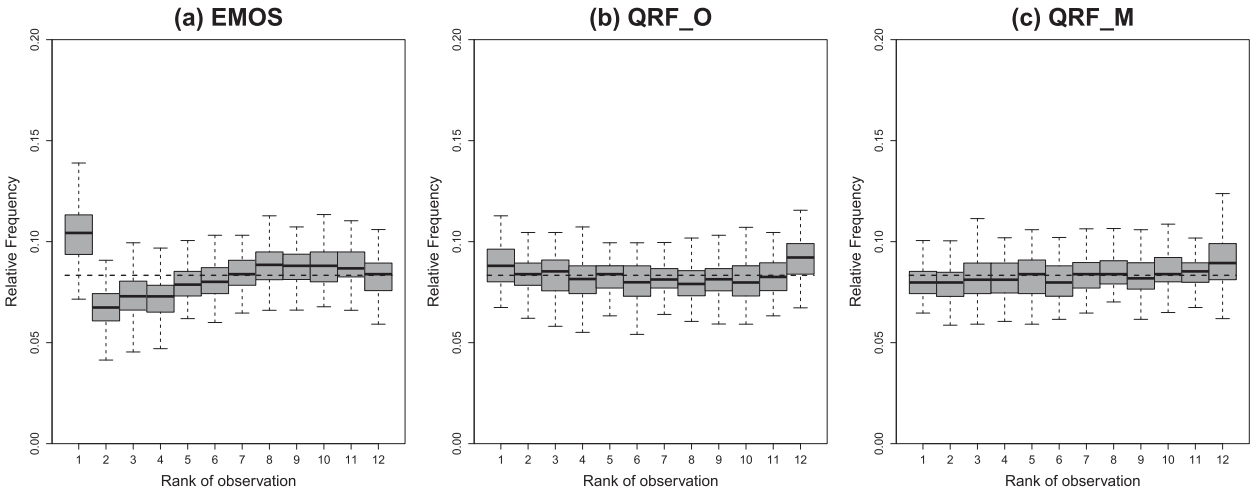

FIG. 10. Box plot of rank histograms for all locations for 24-h forecast of surface wind speed. QRF RH are almost flat whereas EMOS RH has a high first rank.

underdispersion for QRF_O and a little overdispersion for QRF_M even if the box plot is close to 1. There are three main remarks when we are looking at Fig. 4. First, we can assume that contrary to the raw ensemble, all box plots concerning reliability are quite small for the three techniques of calibration: we can say that performances of techniques of calibration for reliability do not depend on location or on time. In addition, we can see that the

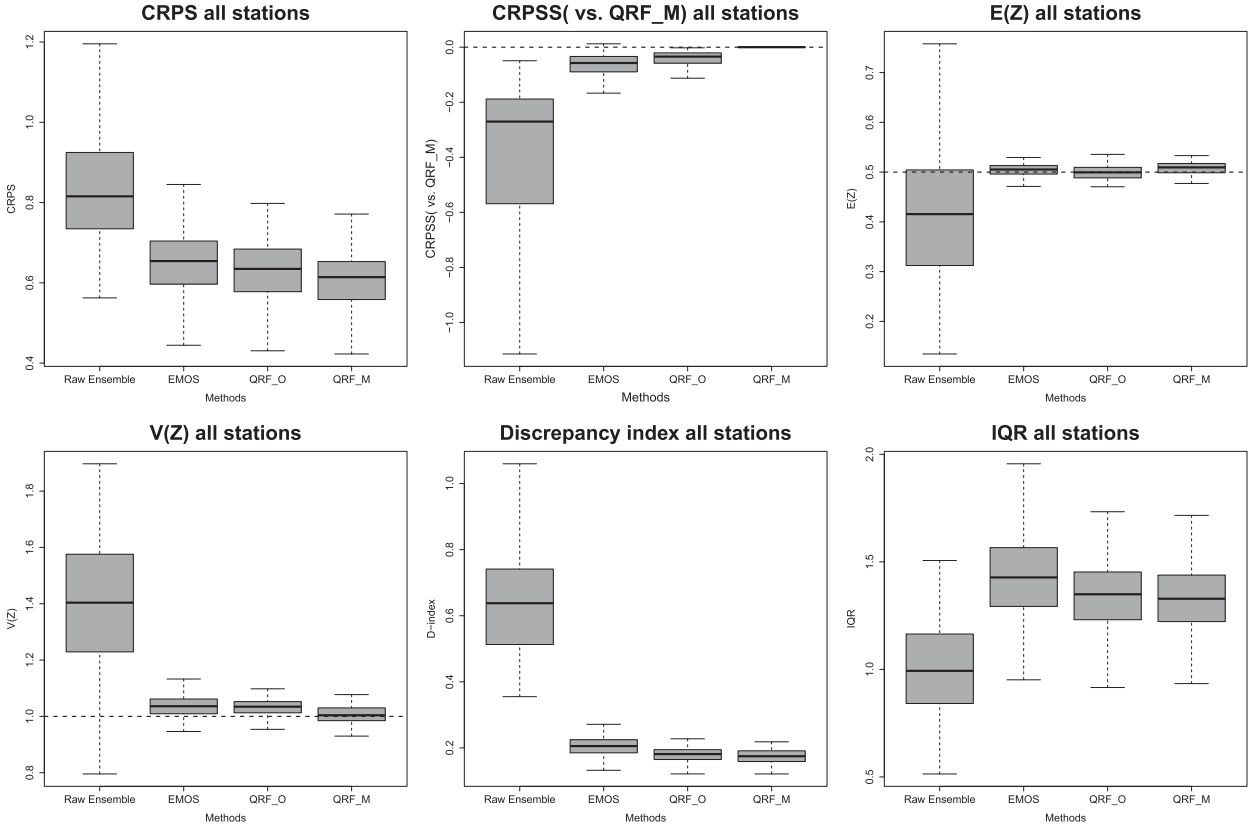

FIG. 11. Box plot of different scores for all locations for 24-h forecast of surface wind speed. QRF_M technique has better CRPS for almost all stations according to the CRPS skill score. All calibrated ensembles are unbiased, reliable, and well dispersed even if there is still a little bit of underdispersion for EMOS.

IQR box plots for calibrated ensembles are taller than the raw ensemble. And last but not least, when we focus on the CRPS skill score computed with regard to QRF_M for each station we see that almost all the values of the different box plots are under 0: not only does QRF_M have a better CRPS in general but QRF_M is better in CRPS than all other ensembles and this is true for almost all stations in this study.

We also investigate performances of probabilistic forecasts of frost for all stations for the 36-h lead time. Figure 5 shows reliability diagrams for all ensembles. We can see very good performances of calibrated ensembles whereas raw ensemble tends to overpredict frost. This is not surprising since in Fig. 4 we see that the raw ensemble is essentially cold biased.

We continue this study on surface temperature by showing results across lead times in Fig. 6. We note that raw ensemble follows a diurnal cycle for all scores. This phenomenon is not shared by calibration techniques concerning reliability but just for CRPS and IQR: we conclude that reliability is not influenced by lead time for calibrated ensembles, only IQR is concerned and thus the CRPS. In addition, the very good entropy of calibrated ensembles (the raw ensemble entropy is around 0.75), which causes us to think that the gain is mainly in reliability. It is interesting to see that raw ensemble does not manage to conciliate good dispersion with small bias. Moreover, reliability of the raw ensemble tends to increase among lead times: indeed predictions are less sharp so they can manage to catch the observation. Besides, we note that calibrated ensembles still remain unbiased and reliable with a preference for QRF techniques concerning entropy and are quite well dispersed. The QRF_O technique is a little bit underdispersed and QRF_M is a little bit overdispersed but both are quite close to 1. Last but not least, we see for CRPS that QRF_O and EMOS are very similar and the gap with QRF_M tends to remain the same across lead times. We can explain the gain in CRPS by the introduction of predictors from other variables than surface temperature and show all the interest of QRF_M method regarding QRF_O.

Now let us conclude by showing the interest of QRF techniques and in particular QRF_M technique for forecasters. In our opinion, the main issue of the EMOS technique is that it loses one of the main aims of ensemble forecasting, which is to assess different scenarios from different initial conditions (i.e., to build different trajectories that can converge or diverge in order to create meteorological scenarios). Indeed, EMOS technique fits a single and unimodal PDF and does not permit one to make alternative scenarios. In Fig. 7 we have four examples of meteorological situations where QRF_M can show all its interest: for Melun, France, in

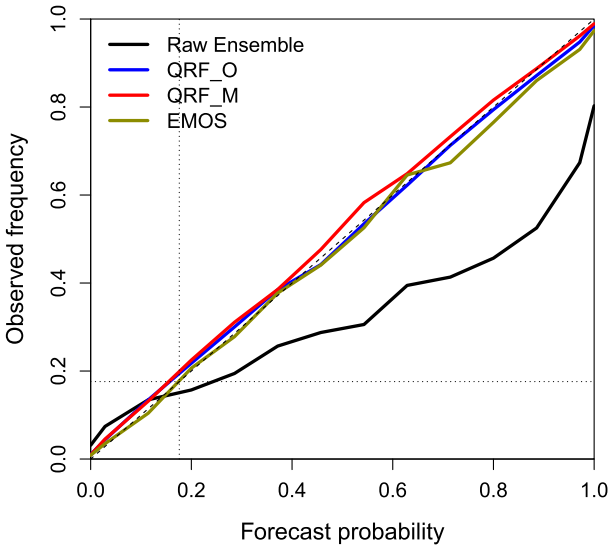

FIG. 12. Reliability diagram for probabilistic 24-h forecast of exceedance of the $5\,\mathrm{m\,s^{-1}}$ threshold in all locations. The dotted lines represent climatology. Calibrated ensembles give reliable probabilistic forecasts for this threshold.

Fig. 7 we have a situation with snowy ground and clear skies during the night causing a rapid cooling. Here, even if all calibrated ensembles give a mode around −4° we can see that QRF_M proposes cooler scenarios. The forecaster knowing this phenomenon of rapid cooling would choose this scenario to make a deterministic forecast for example. We can assume here that the combined predictors *snowfall amount* and *surface irradiation* permit one to detect a nonlinear phenomenon. For the forecast at Carcassonne, France, we see that the raw ensemble is bimodal. QRF_M technique is able to detect a situation conducting to a bimodality and so it fits a bimodal PDF (and if this bimodality is just an artifact it is an artifact now shared by the raw ensemble and the QRF_M ensemble). Moreover, observation corresponds to the first mode of QRF_M PDF whereas other calibrated ensembles are unimodal. It is the same case for Boulogne-sur-Mer, France: the bimodal raw ensemble leads to bimodal PDFs for QRF techniques (second modes are between 18° and 19°) and the first mode is preferred and almost corresponds to observation. EMOS technique here fits the PDF in order to avoid mistakes and put its mean between the two raw ensemble modes. It can happen that meteorological situations detected by QRF_M technique lead to a unimodal PDF whereas the raw ensemble sees two different scenarios. It is the case of the forecast at Paris-Le Bourget airport where QRF_M does not take into account the (misleading) raw ensemble bimodality and fits its mode between these modalities, and is consistent with the observation. Nevertheless, we remember that

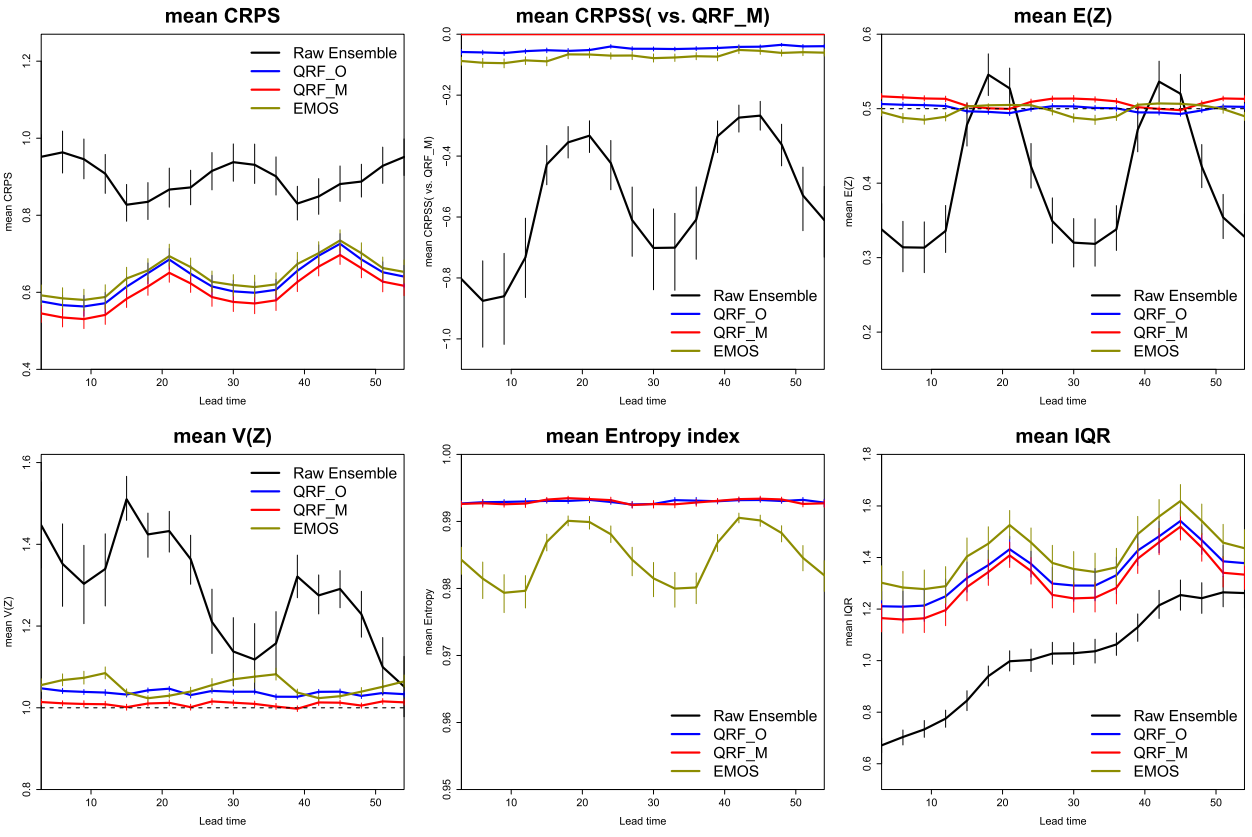

FIG. 13. Mean scores with 95% bootstrap confidence intervals for all locations across lead times for surface wind speed. QRF_M is the best technique for CRPS and CRPSS. Calibrated ensembles are unbiased and in general better dispersed than raw ensemble (the raw ensemble entropy is around 0.85). QRF techniques tend to provide sharper, more reliable, and better dispersed forecasts than EMOS.

we cannot evaluate ensemble forecasts on single cases. In Fig. 7, a BMA calibration was also made with the same learning sample than EMOS. If BMA technique permits bimodalities this is not the case here: we think that the deterministic forecast, the control member, and the mean of the raw ensemble are much too close to have bimodalities. BMA should be more convenient with ensembles made of several deterministic forecasts.

b. *Surface wind speed*

We now give results for surface wind speed. Like for surface temperature we choose to begin with an example for 24-h lead time (corresponding to 1800 UTC) at the same locations. Figures 8 and 9 and Table 2 show RH and scores for all presented methods. Mainly the commentaries are the same as for surface temperature. EMOS tends to be a little underdispersed.

Figure 10 showing RH for all stations confirms that there is still a little issue with the first rank for EMOS: this is likely due to a suboptimally chosen distribution type. The square root–truncated normal distribution used here minimizes the average CRPS on whole stations. The form of this distribution may not be optimal for calibrated ensemble forecasting little wind speed. This behavior is similar to the PIT histogram in the middle of Fig. 5 of Scheuerer and Möller (2015). At the same time we can note that QRF_M dispersion is almost perfect. Figure 11 confirms the good dispersion of QRF_M. We also note that calibrated ensembles seem unbiased and QRF techniques provide reliable ensembles. Last but not least, for temperatures the CRPS skill score shows that the QRF_M method is the best in terms of CRPS for almost all locations.

We can look at the performance of probabilistic forecast of threshold $5\,\mathrm{m\,s^{-1}}$ for all stations and 24-h lead time. Figure 12 reveals an overprediction of threshold exceedances by raw ensemble, and this feature is corrected by calibrated ensembles. It is not shown here but the results for $10\,\mathrm{m\,s^{-1}}$ are as good as for $5\,\mathrm{m\,s^{-1}}$. We have examined the threshold of $15\,\mathrm{m\,s^{-1}}$, but there are not enough observations and the reliability diagram is too noisy to be meaningful.

Figure 13 shows results across lead times for surface wind speed. If conclusions are strictly the same as for surface temperature, we can add here that sharpness and entropy of QRF ensembles are better than EMOS. Last, QRF techniques are very well dispersed and reliable and

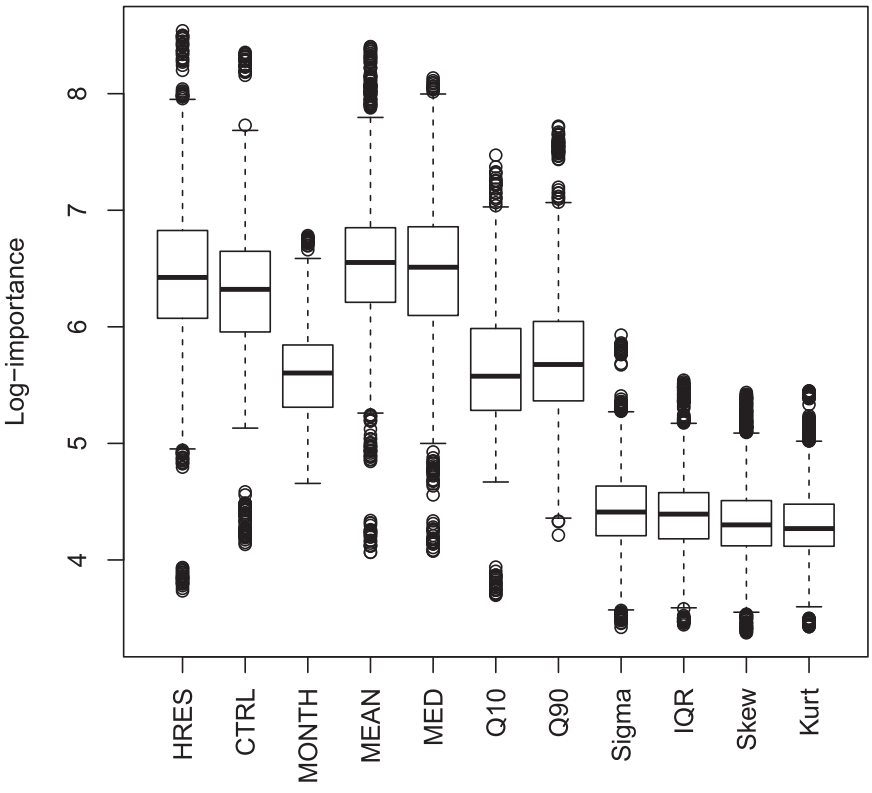

FIG. 14. Log importance of QRF_O predictors for the 24-h forecast of surface wind speed. A box plot is composed of measures of log importance of all the forests and all the stations (so 24 forests × 87 stations = 2088 measures of log importance per predictor). Most important predictors are those who represent central and extreme locations of the ensemble.

thus QRF_O (and QRF_M of course) has much better CRPS than EMOS. We can explain these differences with surface temperature by the fact that finding a good parametric distribution is a bit trickier for wind speed than for temperatures and so EMOS performs less well than QRF techniques in that case.

c. *Importance of the QRF predictors*

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

Figure 14 shows the importance of QRF_O predictors for the 24-h forecast of the surface wind speed. As expected, the most important predictors are those that give information on the center of the distribution. Next, we have the month (a seasonal information) and the first and the ninth decile. It is interesting to see that information on spread or other moments is quite useless, these predictors about spread and higher moments even have same importance that artificially generated random variables (not shown here). We can explain this by the fact that spread information is already contained in decile predictors (in addition to a value on the variable of interest), and it is easier for the model to split meteorological situations by their extreme quantiles rather than their predictability summarized by a statistic such as standard deviation. It is not shown here, but Fig. 14 also applies to another lead time and the other variable, which is surface temperature (with a slightly higher seasonal importance, however).

For the QRF_M method we focus on surface temperature to show that we can detect a meteorological consistency in the QRF model. Figures 15 and 16 show the importance of two different lead times (33 h for 0300 UTC and 42 h for 1200 UTC). We can assume that both figures have quite the same shape. Indeed, we find that central parameters, deciles, and the month are important. In addition, the predictor TPW850 is also important: there is a clear

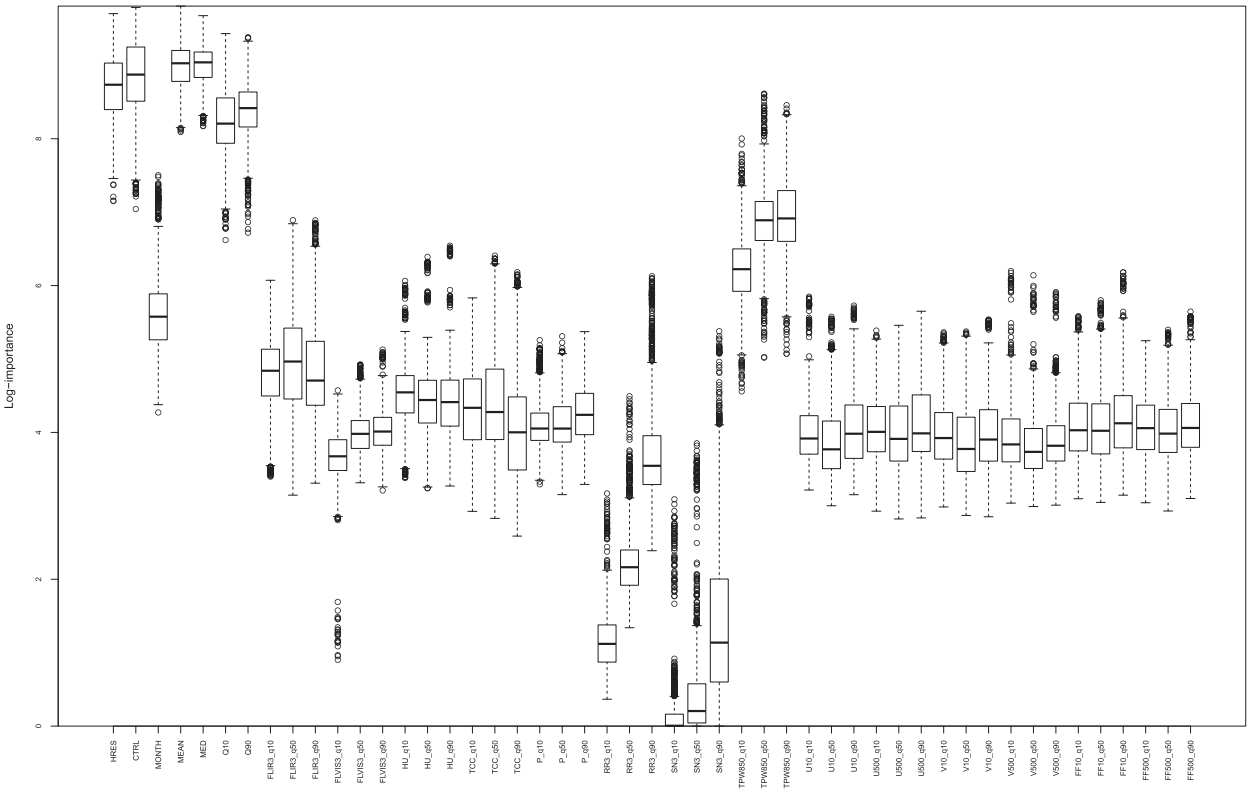

FIG. 15. Log importance of QRF_M predictors for 33-h forecast of surface temperature. A box plot is composed of measures of log importance of all the forests and all the stations (so 24 forests × 87 stations = 2088 measures of log importance per predictor). Temperature predictors are the most important with the month. Note the high importance of surface irradiation in infrared wavelengths.

link between surface temperature and potential wet-bulb temperature. Because most of the other predictors have the same importance as noisy predictors, we focus on FLIR3 and FLVIS3: for the 33-h forecast (day) FLIR3 is higher than FLVIS3 in importance but this is the contrary for the 42-h forecast (night). These differences show that the QRF model takes into account diurnal and nocturnal radiation (in terms of wavelengths). Last but not least, we note that RR3 and SN3 have small importance: these predictors are often zero and thus permuting zeros does not change anything, explaining their small importance. We can understand this phenomenon when we are looking at RR3_q90 and SN3_q90. Higher quantiles are less frequently zero and they have higher importance. Nevertheless we can keep them in the model since we remember that random forests do not choose these predictors during node splitting anyway.

5. Discussion

Through this article, we see that the QRF techniques and the QRF_M technique, which yields on multivariable predictors, give reliable and sharp ensembles compared to EMOS techniques. Moreover, we have noticed that the improvement is more consequent for a non-Gaussian variable like surface wind speed than for surface temperature. This improvement is quite the same among lead times showing that nonparametric calibration methods do not lose predictive performance compared to EMOS and can improve over this method. We also believe that nonparametric calibration is more useful for forecasters since output PDF is not constrained by the QRF technique. It allows us to keep the notion of the scenario for our calibrated ensembles and it can detect nonlinear phenomena. It is not just a correction of bias and dispersion for a given distribution. This nonparametric method is a data-driven technique. This may be viewed as a drawback, but the advent of big data and reforecast techniques let us think that nonparametric methods will be frequently used in order to calibrate forecast ensembles and more generally for ensemble output statistics in meteorology. The QRF technique is linked to the method of analogs (Hamill and Whitaker 2006; Delle Monache et al. 2013) in the sense that QRF is another way to find the closest observations given a set of predictors. The method of analogs consists in finding the closest past forecasts (the analogs) according to a given metric of the predictors' space to build an

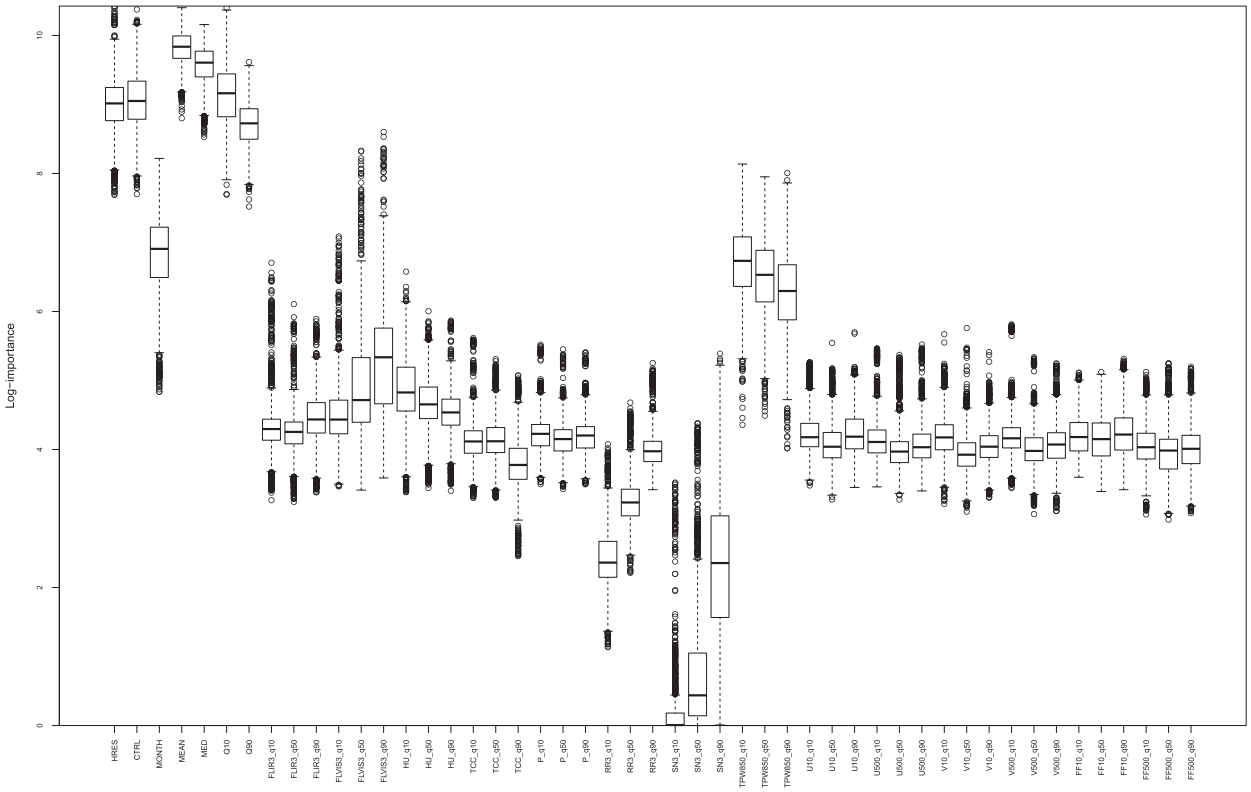

FIG. 16. Log importance of QRF_M predictors for 42-h forecast of surface temperature. A box plot is composed of measures of log importance of all the forests and all the stations (so 24 forests × 87 stations = 2088 measures of log-importance per predictor). Temperature predictors are the most important together with month. Note the high importance of surface irradiation in visible wavelengths for this lead time.

analog-based ensemble. The QRF technique proceeds by iterative dichotomies on the predictors' space to find the closest past forecasts. So both methods share many advantages (e.g., no parametric assumption, easily applicable to multipredictor settings, etc.) and drawbacks (large datasets). Moreover, Delle Monache et al. (2013) applied the method of analogs for surface temperature and wind speed on much smaller datasets (and with only three or four predictors) than in Hamill and Whitaker (2006) for rainfall: the size of the dataset is an issue depending on the weather variable under consideration, it will be interesting to check the performances of the analogs technique and QRF with smaller datasets than in Hamill and Whitaker (2006) for rainfall but with many more predictors (we remember that our QRF_M technique uses more than 40 predictors).

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

As a perspective we will apply QRF techniques to other parameters (rainfall as said above) and try regional calibration: we could add, for example, predictors like longitude, latitude, and altitude to make regional QRF regroup some stations/grid points in order to have fewer (but bigger) forests and model some spatial interactions. Some works in the same vein have been published recently for EMOS (Feldmann et al. 2015). We will also try techniques for trajectory recovery in ensemble forecasts by using the nonparametric technique of ensemble copula coupling (Bremnes 2007; Krzysztofowicz and Toth 2008; Schefzik et al. 2013). We are also interested in combining bias correction for deterministic forecasts to

TABLE A1. Lists of predictors for QRF_O.

For surface temperature and surface wind speed
High-resolution member
Control member
Mean of raw ensemble
Median of raw ensemble
First decile of raw ensemble
Ninth decile of raw ensemble
Std dev of raw ensemble
IQR of raw ensemble
Skewness of raw ensemble
Kurtosis of raw ensemble
Month of the year

TABLE A2. Lists of predictors for QRF_M.

Surface temperature	Both variables	Surface wind speed
HRES	High-resolution member	
CTRL	Control member	
MEAN	Mean of raw ensemble	
MED	Median of raw ensemble	
Q10	First decile of raw ensemble	
Q90	Ninth decile of raw ensemble	
Month	Month of the year	
Sigma	—	Std dev of raw ensemble
IQR	—	IQR of raw ensemble
Skew	—	Skewness of raw ensemble
Kurt	—	Kurtosis of raw ensemble
q10, 50, and 90 are the first decile, the median, and ninth decile, respectively, of the raw ensemble for the following variables:		
HU_q10, 50, 90	Surface humidity	
P_q10, 50, 90	Sea level pressure	
TCC_q10, 50, 90	Total cloud cover	
RR3_q10, 50, 90	3-h rainfall amount	
SN3_q10, 50, 90	3-h snowfall amount	
U10_q10, 50, 90	Surface zonal wind	
V10_q10, 50, 90	Surface meridional wind	
U500_q10, 50, 90	500-m zonal wind	
V500_q10, 50, 90	500-m meridional wind	
FF500_q10, 50, 90	500-m wind speed	
TPW850_q10, 50, 90	850-hPa potential wet-bulb temperature	
FLIR3_q10, 50, 90	3-h total surface irradiation in infrared wavelengths	
FLVIS3_q10, 50, 90	3-h total surface irradiation in visible wavelengths	
T_q10, 50, 90	—	Surface temperature
FF10_q10, 50, 90	Surface wind speed	—

the correction of ensemble forecasts. Last but not least, we could try to use multivariate regression forests (De'Ath 2002) directly to make multivariate calibrated forecasts.

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

APPENDIX A

List of Predictors for QRF_O and QRF_M

See Tables A1 and A2 for the list of predictors.

APPENDIX B

List of Theoretical Formulas and Analytic Formulas for the CRPS for Several Distributions

The continuous ranked probability score (CRPS; Matheson and Winkler 1976; Hersbach 2000) is defined directly in terms of the predictive CDF, F, as

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

Another representation (Gneiting and Raftery 2007) shows that

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

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

Another elegant representation that we found using the L-moments (Hosking 1989) is

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

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

a. Normal distribution

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

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

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

b. Truncated normal distribution

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

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

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

c. Square root–transformed truncated normal distribution

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

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

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

d. Lognormal distribution

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

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

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

e. Gamma distribution

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

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

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

f. Beta distribution

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

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

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

g. Logistic distribution

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

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

where $\omega = (y - \mu)/s$.

h. Truncated logistic distribution

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

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

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

i. Log-logistic distribution

For $X \sim \mathrm{Log} - \mathrm{Logis}(\alpha, \beta)$ and $\beta > 1$,

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

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

j. Truncated logistic distribution with a point mass in 0

Here X is a nonnegative random variable whose CDF is $F(x) = e^{a+bx}/(1 + e^{a+bx})$,where a is real and $b > 0$ [the PDF has a Dirac delta in 0: $\delta(x)F(0)$]:

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

k. Square root–transformed truncated logistic distribution with a point mass in 0

Here X is a nonnegative random variable whose CDF is $F(x) = e^{a+b\sqrt{x}}/(1 + e^{a+b\sqrt{x}})$,where a is real and $b > 0$ [the PDF has a Dirac delta in 0: $\delta(x)F(0)$]:

$$\mathrm{CRPS}(X, y) = \frac{1}{b^2}[4\mathrm{Li}_2(-e^{a+b\sqrt{y}}) - 2\mathrm{Li}_2(-e^{a}) - 2\log(1 + e^{a}) - 2b\sqrt{y}\log(1 + e^{a+b\sqrt{y}})] + \frac{a(a+2)}{b^2} - y,$$

where $\mathrm{Li}_2(z)$ is the dilogarithm function. These two last distributions are extracted from Wilks (2009).

l. Generalized Pareto distribution (GPD) and generalized extreme value (GEV) distribution

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

m. Von Mises distribution

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

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518–1530, doi:10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Bao, L., T. Gneiting, E. P. Grimit, P. Guttorp, and A. E. Raftery, 2010: Bias correction and Bayesian model averaging for ensemble forecasts of surface wind direction. *Mon. Wea. Rev.*, **138**, 1811–1821, doi:10.1175/2009MWR3138.1.

Baran, S., and S. Lerch, 2015: Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. *Quart. J. Roy. Meteor. Soc.*, **141**, 2289–2299, doi:10.1002/qj.2521.

Ben Bouallègue, Z., 2013: Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Wea. Forecasting*, **28**, 515–524, doi:10.1175/WAF-D-12-00062.1.

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

——, 2001: Random forests. *Mach. Learn.*, **45** (1), 5–32, doi:10.1023/A:1010933404324.

——, J. Friedman, C. J. Stone, and R. A. Olshen, 1984: *Classification and Regression Trees.* CRC Press, 368 pp.

Bremnes, J., 2004: Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Wea. Rev.*, **132**, 338–347, doi:10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2.

——, 2007: Improved calibration of precipitation forecasts using ensemble techniques. Part 2: Statistical calibration methods. Norwegian Meteorological Institute, Tech. Rep. 04/2007.

Bröcker, J., and L. A. Smith, 2007: Increasing the reliability of reliability diagrams. *Wea. Forecasting*, **22**, 651–661, doi:10.1175/WAF993.1.

Courtier, P., C. Freydier, J. Geleyn, F. Rabier, and M. Rochas, 1991: The Arpege project at Météo-France. *Proc. ECMWF Seminar*, Vol. 2, Reading, United Kingdom, ECMWF, 193–231. [Available online at http://www.ecmwf.int/sites/default/files/elibrary/1991/8798-arpege-project-meteo-france.pdf.]

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

Delle Monache, L., J. P. Hacker, Y. Zhou, X. Deng, and R. B. Stull, 2006: Probabilistic aspects of meteorological and ozone regional ensemble forecasts. *J. Geophys. Res.*, **111**, D24307, doi:10.1029/2005JD006917.

——, F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev.*, **141**, 3498–3516, doi:10.1175/MWR-D-12-00281.1.

Descamps, L., C. Labadie, A. Joly, E. Bazile, P. Arbogast, and P. Cébron, 2015: PEARP, the Météo-France short-range ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **141**, 1671–1685, doi:10.1002/qj.2469.

Feldmann, K., M. Scheuerer, and T. L. Thorarinsdottir, 2015: Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous gaussian regression. *Mon. Wea. Rev.*, **143**, 955–971, doi:10.1175/MWR-D-14-00210.1.

Ferro, C., 2014: Fair scores for ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **140**, 1917–1923, doi:10.1002/qj.2270.

——, D. S. Richardson, and A. P. Weigel, 2008: On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteor. Appl.*, **15**, 19–24, doi:10.1002/met.45.

Friederichs, P., and A. Hense, 2007: Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Wea. Rev.*, **135**, 2365–2378, doi:10.1175/MWR3403.1.

——, and T. L. Thorarinsdottir, 2012: Forecast verification for extreme value distributions with an application to probabilistic peak wind prediction. *Environmetrics*, **23**, 579–594, doi:10.1002/env.2176.

Gneiting, T., and A. E. Raftery, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.*, **102**, 359–378, doi:10.1198/016214506000001437.

——, and M. Katzfuss, 2014: Probabilistic forecasting. *Annu. Rev. Stat. Appl.*, **1**, 125–151.

——, A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, doi:10.1175/MWR2904.1.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc. Stat. Methodol.*, **69B**, 243–268, doi:10.1111/j.1467-9868.2007.00587.x.

Grimit, E. P., T. Gneiting, V. J. Berrocal, and N. A. Johnson, 2006: The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quart. J. Roy. Meteor. Soc.*, **132**, 2925–2942, doi:10.1256/qj.05.235.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.*, **136**, 2608–2619, doi:10.1175/2007MWR2410.1.

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.*, **129**, 550–560, doi:10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2.

——, and S. J. Colucci, 1997: Verification of ETA-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312–1327, doi:10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

——, and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, doi:10.1175/MWR3237.1.

——, R. Hagedorn, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Wea. Rev.*, **136**, 2620–2632, doi:10.1175/2007MWR2411.1.

Hastie, T., R. Tibshirani, and J. Friedman, 2009: The *Elements* of *Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed. Springer, 745 pp.

Hemri, S., M. Scheuerer, F. Pappenberger, K. Bogner, and T. Haiden, 2014: Trends in the predictive performance of raw ensemble weather forecasts. *Geophys. Res. Lett.*, **41**, 9197–9205, doi:10.1002/2014GL062472.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting*, **15**, 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hosking, J. R. M., 1989: *Some Theoretical Results Concerning L-Moments*. IBM Thomas J. Watson Research Center, 9 pp.

Juban, J., L. Fugon, and G. Kariniotakis, 2007: Probabilistic short-term wind power forecasting based on kernel density estimators. European Wind Energy Conference and Exhibition, Milan, Italy, EWEC 2007. [Available online at https://hal.inria.fr/file/index/docid/526011/filename/EWEC2007-juban.pdf.]

Koenker, R., and G. Bassett Jr., 1978: Regression quantiles. *Econometrica*, **46**, 33–50, doi:10.2307/1913643.

Krzysztofowicz, R., and Z. Toth, 2008: Bayesian processor of ensemble (BPE): Concept and implementation. *Fourth NCEP/NWS Ensemble User Workshop*, Laurel, MD, NCEP/NWS. [Available online at www.emc.ncep.noaa.gov/gmb/ens/ens2008/Krzysztofowicz_Presentation_Web.pdf.]

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.*, **22**, 1087–1096, doi:10.1287/mnsc.22.10.1087.

Meinshausen, N., 2006: Quantile regression forests. *J. Mach. Learn. Res.*, **7**, 983–999.

Pinson, P., 2012: Adaptive calibration of (u, v)-wind ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **138**, 1273–1284, doi:10.1002/qj.1873.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, doi:10.1175/MWR2906.1.

Schefzik, R., T. L. Thorarinsdottir, T. Gneiting, 2013: Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.*, **28**, 616–640, doi:10.1214/13-STS443.

Scheuerer, M., 2014: Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Quart. J. Roy. Meteor. Soc.*, **140**, 1086–1096, doi:10.1002/qj.2183.

——, and D. Möller, 2015: Probabilistic wind speed forecasting on a grid based on ensemble model output statistics. *Ann. Appl. Stat.*, **9**, 1328–1349, doi:10.1214/15-AOAS843.

Schuhen, N., T. L. Thorarinsdottir, and T. Gneiting, 2012: Ensemble model output statistics for wind vectors. *Mon. Wea. Rev.*, **140**, 3204–3219, doi:10.1175/MWR-D-12-00028.1.

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220, doi:10.1175/MWR3441.1.

——, T. Gneiting, and A. E. Raftery, 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.*, **105**, 25–35, doi:10.1198/jasa.2009.ap08615.

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

Thorarinsdottir, T. L., and T. Gneiting, 2010: Probabilistic forecasts of wind speed: Ensemble model output statistics by using heteroscedastic censored regression. *J. Roy. Stat. Soc.*, **173A**, 371–388, doi:10.1111/j.1467-985X.2009.00616.x.

Tribus, M., 1969: *Rational Descriptions, Decisions and Designs.* Pergamon Press, 500 pp.

Weijs, S. V., R. Van Nooijen, and N. Van De Giesen, 2010: Kullback–Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition. *Mon. Wea. Rev.*, **138**, 3387–3399, doi:10.1175/2010MWR3229.1.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

——, 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.*, **16**, 361–368, doi:10.1002/met.134.

Zamo, M., O. Mestre, P. Arbogast, and O. Pannekoucke, 2014a: A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part I: Deterministic forecast of hourly production. *Sol. Energy*, **105**, 792–803, doi:10.1016/j.solener.2013.12.006.

——, ——, ——, and ——, 2014b: A benchmark of statistical regression methods for short-term forecasting of photovoltaic electricity production. Part II: Probabilistic forecast of daily production. *Sol. Energy*, **105**, 804–816, doi:10.1016/j.solener.2014.03.026.