Math Geosci (2018) 50:209–234
https://doi.org/10.1007/s11004-017-9709-7

SPECIAL ISSUE

Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts

Michaël Zamo[1] · **Philippe Naveau**[2]

Received: 25 November 2016 / Accepted: 8 October 2017 / Published online: 21 November 2017

Abstract The continuous ranked probability score (CRPS) is a much used measure of performance for probabilistic forecasts of a scalar observation. It is a quadratic measure of the difference between the forecast cumulative distribution function (CDF) and the empirical CDF of the observation. Analytic formulations of the CRPS can be derived for most classical parametric distributions, and be used to assess the efficiency of different CRPS estimators. When the true forecast CDF is not fully known, but represented as an ensemble of values, the CRPS is estimated with some error. Thus, using the CRPS to compare parametric probabilistic forecasts with ensemble forecasts may be misleading due to the unknown error of the estimated CRPS for the ensemble. With simulated data, the impact of the type of the verified ensemble (a random sample or a set of quantiles) on the CRPS estimation is studied. Based on these simulations, recommendations are issued to choose the most accurate CRPS estimator according to the type of ensemble. The interest of these recommendations is illustrated with real ensemble weather forecasts. Also, relationships between several estimators of the CRPS are demonstrated and used to explain the differences of accuracy between the estimators.

Keywords Continuous ranked probability score · Estimation · Forecast comparison · Ensemble weather forecasts

✉ Michaël Zamo
michael.zamo@meteo.fr

[1] Météo-France, 42, avenue Gaspard Coriolis, 31057 Toulouse Cedex 07, France

[2] Laboratoire des Sciences du Climat et l'Environnement (LSCE) CNRS, Orme des Merisiers / Bat. 701 C.E. Saclay, 91191 Gif-sur-Yvette, France

1 Introduction

Verifying the quality of forecasts expressed in a probabilistic form requires specific graphical or numerical tools (Jolliffe and Stephenson 2011), among them some numerical measures of performance such as the Brier score (Brier 1950), the Kullback–Leibler divergence (Weijs et al. 2010) and many others (Winkler et al. 1996; Gneiting and Raftery 2007). When the probabilistic forecast is a cumulative distribution function (CDF) and the observation is a scalar, the continuous ranked probability score (CRPS) is often used as a quantitative measure of performance. Classically (Matheson and Winkler 1976; Hersbach 2000), the instantaneous CRPS is defined as the quadratic measure of discrepancy between the forecast CDF, noted F, and $\mathbb{1}(x \geq y)$, the empirical CDF of the scalar observation y

$$\mathrm{crps}(F, y) = \int_{\mathbb{R}} [F(x) - \mathbb{1}(x \geq y)]^2 \, \mathrm{d}x, \tag{INT}$$

where $\mathbb{1}$ is the indicator function.

Analytic formulations of $\mathrm{crps}(F, y)$ can be derived for most classical parametric distributions, some of which are listed in Table 1. In some situations, the forecast CDF may not be fully known, such as for ensemble numerical weather prediction (NWP) or other types of Monte-Carlo simulations, or the forecast CDF may be known, but an analytic formulation of the CRPS may not be derivable. In the latter case, one may be able to sample values from F. In any case, in these two situations, the forecast CDF is summarized with a set of M values $x_{i=1,\ldots,M}$. Following the convention in meteorology, such a set will be called here an “ensemble”, and each value x_i will be called a “member”. The instantaneous CRPS must then be estimated with this ensemble. This may be problematic when using the CRPS to compare parametric forecasts, whose CRPS may be computed exactly, and forecasts whose CRPS is estimated based on the limited information about F contained in the ensemble. The unknown error in the CRPS estimation may lead to the wrong choice of the best forecast. Although meteorological vocabulary is used, this situation can occur in other fields of geosciences too, for instance when conditional simulations are used to sample from a probability distribution and choose between competing techniques or settings (Emery and Lantuéjoul 2006; Pirot et al. 2014; Yin et al. 2016).

Usually, the instantaneous CRPS is averaged in space and/or time over several pairs of forecast/observation. Candille (2003) and Ferro et al. (2008) showed that when the ensemble is a random sample from F, the usual estimator of the instantaneous CRPS based on Eq. (INT), introduced later, is biased: its expectation over an infinite number of forecast/observation pairs does not give the right theoretical value. This bias stems from the limited information about F contained in an ensemble with finite size M. Several solutions have been proposed to remove this bias. Ferro (2014) introduced the notion of fair score and a formula to correct the bias in the estimation of the averaged CRPS. Müller et al. (2005) proposed two solutions to the same problem of biased estimation of the ranked probability score (RPS), the version of the CRPS for ordinal random variables. Adapted to the CRPS, their first solution would be to use an absolute value instead of a square inside the integral in Eq. (INT). As demonstrated

Table 1 List of distributions whose closed-form CRPS exists and were used in this study

Distribution	Original reference
Beta: $Y \sim Beta(\alpha, \beta)$	Taillardat et al. (2016)
Gamma: $Y \sim Gamma(\alpha, \beta)$	Möller and Scheuerer (2013)
Gaussian mixture: $Y \sim \sum_{i=1}^{p} \omega_i \mathcal{N}(\mu_i, \sigma_i)$, with $\sum_{i=1}^{p} \omega_i = 1, \quad \omega_{i=1,\ldots,p} > 0$	Grimit et al. (2006)
Generalized extreme value: $Y \sim GEV(\mu, \sigma, \xi)$	Friederichs and Thorarinsdottir (2012)
Generalized Pareto: $Y \sim GPD(\mu, \sigma, \xi)$	Friederichs and Thorarinsdottir (2012)
Log-normal: $\ln(Y) \sim \mathcal{N}(\mu, \sigma)$	Baran and Lerch (2015)
Normal: $Y \sim \mathcal{N}(\mu, \sigma)$	Gneiting et al. (2005)
Square-root truncated normal: $\sqrt{Y} \sim \mathcal{N}^0(\mu, \sigma)$	Hemri et al. (2014)
Truncated normal: $Y \sim \mathcal{N}^0(\mu, \sigma)$	Thorarinsdottir and Gneiting (2010)

The reference of the original article where to find the formula is also given. Taillardat et al. (2016) gathers the closed form expression of the CRPS for these and other distributions

in Appendix A, this score for an ensemble is minimized if all the members x_i equal the median of F, which is obviously not the purpose of an ensemble. Their second solution is to compute the RPS skill score against some ensemble of size M whose RPS is estimated by bootstrapping past observations. Although interesting, this solution does not allow assessing the absolute performance of the ensemble, but only the performance relative to this bootstrapped ensemble.

This study aims at improving heuristically the estimation of the average CRPS of a forecast CDF under limited information. The information is limited in two ways: (i) the CDF is known only through an ensemble as defined above, and (ii) the average CRPS is computed over a finite number of forecast/observation pairs. The problem is not to estimate the unknown forecast CDF F, but to estimate the CRPS of F under limited information about F. To improve the estimation with this limited information, the usual strategy is to correct the empirical mean score, as in Ferro (2014) or Müller et al. (2005). Here the approach is to improve the estimation of each term of the average, that is, the estimation of the instantaneous CRPS crps(F, y).

The rest of this paper is organized as follows. Section 2 reviews several estimators of the instantaneous CRPS proposed in the literature and demonstrates relationships among them. In particular, it is shown that the four proposed estimators reduce to two only. In Sect. 3, synthetic data are used to study the variations in accuracy of these two CRPS estimators, with the size M of the ensemble and the way this ensemble is built. These simulations lead to recommendations on the best estimation of the CRPS. Section 4 illustrates issues in CRPS estimation with two real meteorological data sets. Improvements in the inference obtained by following the recommendations from Sect. 3 are shown on these data. Section 5 gives a summary of the recommendations to get an accurate estimation of the instantaneous CRPS, concludes and discusses the results.

2 Review of Available Estimators of the CRPS

The instantaneous CRPS is defined as a quadratic discrepancy measure between the forecast CDF and the empirical CDF of the observation

$$\mathrm{crps}(F, y) = \int_{\mathbb{R}} [F(x) - \mathbb{1}(x \geq y)]^2 \, \mathrm{d}x. \tag{INT}$$

Equation (INT) is called the integral form of the CRPS.

Gneiting and Raftery (2007) showed that, for forecast CDFs with a finite first moment, the CRPS can be written as

$$\mathrm{crps}(F, y) = \mathbb{E}_X|X - y| - \frac{1}{2}\mathbb{E}_{X,X'}|X - X'|, \tag{NRG}$$

where X and X' are two independent random variables distributed according to F, and $\mathbb{E}_A$ is the expectation according to the law of the random variable(s) A. This is called the energy form of the CRPS, since it is just the one-dimensional case of the energy score introduced by Gneiting and Raftery (2007), based on the energy distance of Székely and Rizzo (2013).

Taillardat et al. (2016) introduced a third expression of the CRPS, valid for continuous forecast CDFs

$$\mathrm{crps}(F, y) = \mathbb{E}_X|X - y| + \mathbb{E}_X X - 2\mathbb{E}_X X F(X), \tag{PWM}$$

which is called the probability weighted moment (PWM) form of the CRPS because its third term is a probability weighted moment (Greenwood et al. 1979; Rasmussen 2001; Furrer and Naveau 2007).

When F is known only through an M-ensemble $x_{i=1,\ldots,M}$, the above definitions lead to the following estimators of the instantaneous CRPS

$$\widehat{\mathrm{crps}}_{\mathrm{INT}}(M, y) = \int_{\mathbb{R}} \left[\frac{1}{M} \sum_{i=1}^{M} \mathbb{1}(x \geq x_i) - \mathbb{1}(x \geq y) \right]^2 \mathrm{d}x, \tag{eINT}$$

$$\widehat{\mathrm{crps}}_{\mathrm{NRG}}(M, y) = \frac{1}{M} \sum_{i=1}^{M} |x_i - y| - \frac{1}{2M^2} \sum_{i,j=1}^{M} |x_i - x_j|, \tag{eNRG}$$

$$\widehat{\mathrm{crps}}_{\mathrm{PWM}}(M, y) = \frac{1}{M} \sum_{i=1}^{M} |x_i - y| + \hat{\beta}_0 - 2\hat{\beta}_1, \tag{ePWM}$$

respectively, where $\mathbb{E}_X X$ is estimated by $\hat{\beta}_0 = \frac{1}{M}\sum_{i=1}^{M} x_i$, and $\mathbb{E}_X X F(X)$ is estimated by $\hat{\beta}_1 = \frac{1}{M(M-1)}\sum_{i=1}^{M}(i-1)x_i$. Without loss of generality, the members x_i

are supposed sorted in increasing order, and the size M of the ensemble is supposed greater than two.

Candille (2003) and Ferro et al. (2008) showed that the expectation of Eq. (eINT) over an infinite number of forecast/observation pairs is biased with, under conditions of stationarity of the observation and the ensemble, and exchangeability of the members

$$\mathbb{E}_Y \widehat{\mathrm{crps}}_{\mathrm{INT}}(M, Y) = \mathbb{E}_Y \mathrm{crps}(F, Y) + \frac{1}{M} \mathbb{E}_{X_1, X_2} \frac{|X_1 - X_2|}{2}, \tag{1}$$

where X_1 and X_2 are any two distinct members of one ensemble forecast. This relation holds only when the ensemble is a random sample from F. Ferro (2014) proposed the notion of fair score for an ensemble of random values, which leads to a fourth estimator of the instantaneous CRPS, the fair CRPS defined as

$$\widehat{\mathrm{crps}}_{\mathrm{Fair}}(M, y) = \frac{1}{M} \sum_{i=1}^{M} |x_i - y| - \hat{\lambda}_2, \tag{eFAIR}$$

where $\hat{\lambda}_2 = \frac{1}{2M(M-1)} \sum_{i,j=1}^{M} |x_i - x_j|$ estimates $\mathbb{E}_{X_1, X_2} \frac{|X_1 - X_2|}{2}$, and is unbiased when the members are independently sampled from F.

These four estimators reduce to only two since, as shown in Appendix B

$$\widehat{\mathrm{crps}}_{\mathrm{INT}}(M, y) = \widehat{\mathrm{crps}}_{\mathrm{NRG}}(M, y),$$
$$\widehat{\mathrm{crps}}_{\mathrm{PWM}}(M, y) = \widehat{\mathrm{crps}}_{\mathrm{Fair}}(M, y).$$

The properties of only two estimators have to be studied. In light of the second equality, the fair CRPS can be interpreted as a PWM-based estimator of the instantaneous CRPS, which explains why it is an unbiased estimator of the average CRPS of a random ensemble as proven by Ferro (2014). Indeed, the unbiasedness property of the mean for the first term and of the PWMs for the second term, in the case of a random sample, immediately proves that the two terms in Eq. (ePWM) are unbiased estimators of their population counterpart, if the members are randomly and independently drawn from F.

Moreover, the relationship

$$\widehat{\mathrm{crps}}_{\mathrm{INT}}(M, y) = \widehat{\mathrm{crps}}_{\mathrm{PWM}}(M, y) + \frac{\hat{\lambda}_2}{M} \tag{2}$$

holds for these two estimators, as shown in Appendix B. Equation (2) holds for a single forecast/observation pair, and requires no assumption on the nature or statistical properties of the ensemble.

3 Study with Simulated Data

The accuracy of the two instantaneous CRPS estimators presented above, $\widehat{\mathrm{crps}}_{\mathrm{PWM}}(M, y)$ and $\widehat{\mathrm{crps}}_{\mathrm{INT}}(M, y)$, is studied with synthetic forecast/observation pairs. The forecast CDF F is chosen such that the theoretical CRPS $\mathrm{crps}(F, y)$ can be exactly computed with a closed-form expression (see Table 1 for a list of such distributions). To mimic actual situations when F is not fully known, two types of ensembles are built from this forecast CDF. The two types of ensembles successively used in the remaining of this section are random ensembles and ensembles of quantiles, defined later. The estimators are then computed and compared to the theoretical value.

3.1 CRPS Estimation with a Random Ensemble

3.1.1 Methodology

A random ensemble is a sample of M independent draws from F. In actual applications, a random ensemble may be viewed as M members from an NWP ensemble model, or, more generally, as an M-sample from Monte-Carlo simulations. Protocol 1 describes the simulation plan.

Protocol 1: ESTIMATION OF THE CRPS WITH SIMULATED RANDOM ENSEMBLES

```
Input: $M$: number of members.
       $F$: forecast CDF.
       $G$: CDF of the observation.
       $N$: number of ensemble forecast/observation pairs.
Output: $N$ values of instantaneous CRPS for each estimator.
for $n \leftarrow 1$ to $N$ do
    Draw the observation $y$ from $G$.
    Compute the theoretical CRPS $\mathrm{crps}_{\mathrm{th}}(F, y)$ with its closed-form expression.
    Draw $x_{i=1,\ldots,M}$ from $F$.
    Compute and store $\widehat{\mathrm{crps}}_{\mathrm{INT}}(M, y)$ and $\widehat{\mathrm{crps}}_{\mathrm{PWM}}(M, y)$ with this ensemble.
```

3.1.2 Results

The results are presented for a standard normal forecast CDF F. For the sake of simplicity the CDF of the observation is also standard normal ($G = F$).

Since the ensemble is random, the estimated CRPS is also a random variable that depends on the observation y and the members $x_{i=1,\ldots,M}$. In order to study the variability of the estimated CRPS with the ensemble only, the observation is first held constant (with a value of -0.0841427, for each n in Protocol 1), while $N = 1000$ ensembles of M members are drawn from F. The impact of M on the accuracy of the estimated CRPS is assessed by observing Protocol 1 with different ensemble sizes M.

The point-wise 10, 50 and 95% intervals of the estimation error $\mathrm{crps}_{\mathrm{th}} - \widehat{\mathrm{crps}}$ (with $\mathrm{crps}_{\mathrm{th}} = 0.2365178$ here) are computed over these 1000 ensembles for each ensemble size M. The intervals contain the corresponding proportion of the 1000 computed CRPS errors for a given ensemble size. As shown in Fig. 1 (left) for $\widehat{\mathrm{crps}}_{\mathrm{INT}}$, the

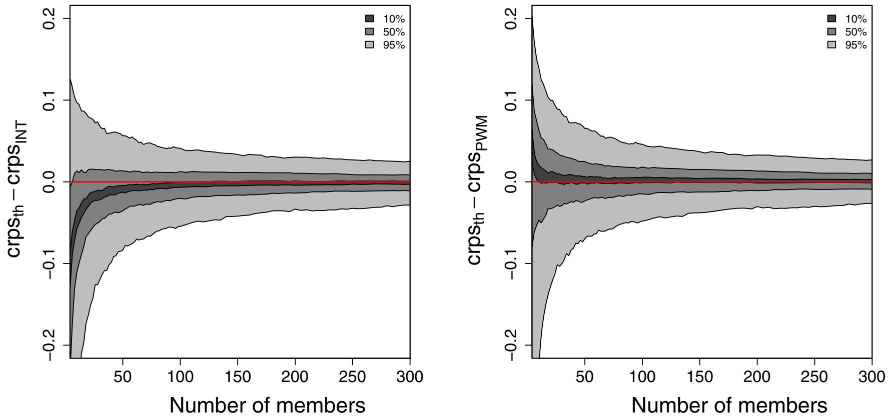


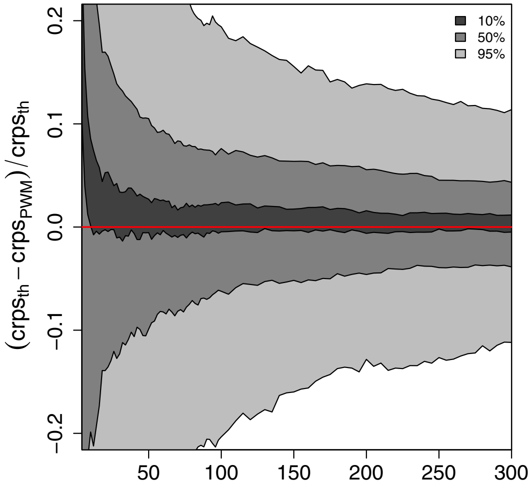
Fig. 1 Intervals of estimation error of $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ (left) or $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$ (right) for a random ensemble of varying size. Intervals are computed point-wise, with the 1000 CRPS of independently built random ensembles with the same observation. The observation and members come from a standard normal distribution

error tends toward 0 when the ensemble size increases. However, important errors (as high as ± 10% of $\mathrm{crps_{th}}$) can still occur even for very large ensembles of several hundreds of members. As shown in Fig. 1 (right), the estimator $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$ exhibits a similar behaviour for large random ensembles, as deduced from Eq. 2 if $M \rightarrow \infty$. But $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$ becomes unbiased for much smaller ensemble sizes than $\widehat{\mathrm{crps}}_{\mathrm{INT}}$. The unbiasedness of $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$ proven by Ferro (2014) holds only for ensembles with more than about 20 members. The variability of the estimation, as quantified by the half-width of the 50% central interval, may be important when the random ensemble contains less than 50 members (more than 10% of $\mathrm{crps_{th}}$, in Fig. 2). With increasing ensemble sizes, the variability of this estimation does not scale linearly with the number of members, as shown in Fig. 2. Tripling the ensemble size from $M = 100$ to about $M = 300$ decreases the half-width of the 50% central interval of the relative estimation error by only about 2% (from 7 to 4%).

Common practice is to average instantaneous CRPSs over several locations and/or times. Here, this is mimicked by taking the average of N instantaneous CRPSs generated according to Protocol 1, while no longer holding the observation constant. The number of forecast/observation pairs N is varied from 1 to 1000. The size M of the ensemble is also varied, with 10, 30, 50, 100 and 300 members. The average theoretical CRPS and average estimation are computed for each combination of N and M. As shown in the left of Fig. 3 for $\widehat{\mathrm{crps}}_{\mathrm{INT}}$, a stable estimation of the average CRPS is reached if the number of averaged estimations is large enough (more than 300 for a random ensemble of 10 members). But a large ensemble is required to get an accurate estimation of the true average CRPS. As shown in the right of Fig. 3, the averaged $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$ shows a better estimate than the averaged $\widehat{\mathrm{crps}}_{\mathrm{INT}}$, even for small ensembles and small numbers of averaged estimations.

These behaviours for the instantaneous and the averaged estimates remain true for every distribution listed in Table 1, every parameter value and even if the G and F are different (not shown).

Fig. 2 Same as in Fig. 1, but for the relative estimation error of $\widehat{\text{crps}}_{\text{PWM}}$



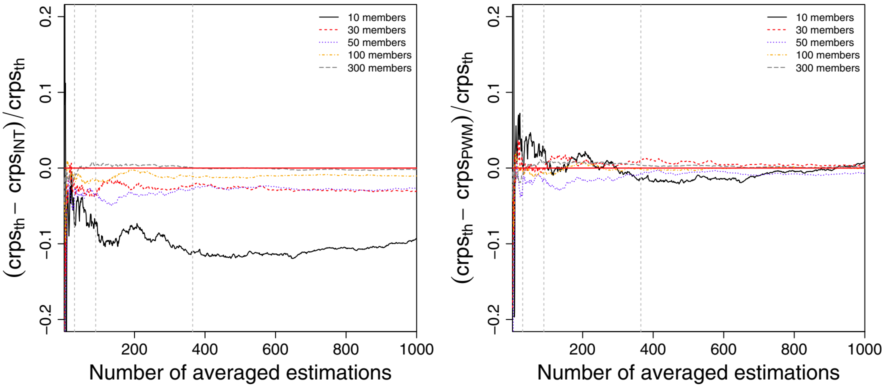


Fig. 3 Evolution of the relative estimation error of the averaged $\widehat{\text{crps}}_{\text{INT}}$ (left) or $\widehat{\text{crps}}_{\text{PWM}}$ (right) with the number of members for a random ensemble. The averaged CRPS is an arithmetic mean of the CRPS of several pairs of ensemble/observation among 1000. The vertical grey dashed lines correspond to an average computed with 30, 90 and 365 ensembles (to mimic a monthly, seasonal or yearly average CRPS)

The added value of these simulations to the results of Ferro (2014) is to show the behaviour of $\widehat{\text{crps}}_{\text{PWM}}$ for small ensemble sizes M and finite numbers of forecast/observation pairs. The poor scaling of this estimator's variability with the ensemble size has been empirically shown, which had never been done, to the best of our knowledge. Finding a formula for the variability of $\widehat{\text{crps}}_{\text{PWM}}$ would be interesting to quantify the estimation uncertainty for practical purposes. Theoretical error bounds have been demonstrated but are not usable in practice since they require to know the forecast distribution (not shown).

The conclusion of these simulations is that, for a random ensemble, the estimation of the instantaneous CRPS is not very accurate whatever estimator is used, but the averaged CRPS can be estimated with a good accuracy. The unbiasedness of $\widehat{\text{crps}}_{\text{PWM}}$ for random ensembles stems from the use of estimators that are unbiased for independent samples from the underlying distribution F. In practice, if one seeks to estimate the potential performance of an ensemble with an infinite number of members, one should use the PWM estimator of the CRPS. The integral estimator of the CRPS assesses the global performance of the actual ensemble, and should be used for actual performance verification.

3.2 CRPS Estimation with an Ensemble of Quantiles

3.2.1 Methodology

An ensemble of M quantiles of orders $\tau_{i=1,\ldots,M} \in [0; 1]$ is a set of M values $x_{i=1,\ldots,M}$ such that: $x_i = F^{-1}(\tau_i)\,\forall i \in \{1,\ldots,M\}$. Contrasting with a random ensemble, the orders τ_i associated to the members x_i are known.

In this case, the data are simulated according to Protocol 2. The two built ensembles of quantiles are defined as:

- *regular* ensemble (reg): it is the ensemble of the M quantiles of orders τ_i, with $\tau_i \in \{\frac{1}{M}, \frac{2}{M}, \ldots, \frac{M-1}{M}, \frac{M-0.1}{M}\}$ of F. The last order is not 1 to prevent infinite values.
- *optimal* ensemble (opt): it is the set of M quantiles of orders $\tau_i \in \{\frac{0.5}{M}, \frac{1.5}{M}, \ldots, \frac{M-0.5}{M}\}$ of F. This ensemble is called "optimal" because Bröcker (2012) showed that this set of quantiles minimizes the expectation of the CRPS of an ensemble over an infinite number of forecast/observation pairs, when using Eq. (eINT).

Protocol 2: ESTIMATION OF THE CRPS WITH SIMULATED ENSEMBLES OF QUANTILES

```
Input: $M$: number of quantiles.
       $F$: forecast CDF.
       $G$: CDF of the observation.
       $N$: number of ensemble forecast/observation pairs.
Output: $N$ values of the instantaneous CRPS for each estimator and type of quantile ensemble.
Compute the ensemble of $M$ regular quantiles of $F$.
Compute the ensemble of $M$ optimal quantiles of $F$.
for $n \leftarrow 1$ to $N$ do
    Draw $y$ from $G$.
    Compute and store the theoretical CRPS $\mathrm{crps}_{\mathrm{th}}(F, y)$ with this observation.
    Compute and store $\widehat{\mathrm{crps}}_{\mathrm{INT}}(M, y)$ and $\widehat{\mathrm{crps}}_{\mathrm{PWM}}(M, y)$ with this observation for the ensemble of regular quantiles.
    Compute and store $\widehat{\mathrm{crps}}_{\mathrm{INT}}(M, y)$ and $\widehat{\mathrm{crps}}_{\mathrm{PWM}}(M, y)$ with this observation for the ensemble of optimal quantiles.
```

3.2.2 Results

Relative estimation errors of $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ and $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$ have been computed for a fixed observation ($N = 1$, $y = -0.0841427$) and regular and optimal ensembles, all built

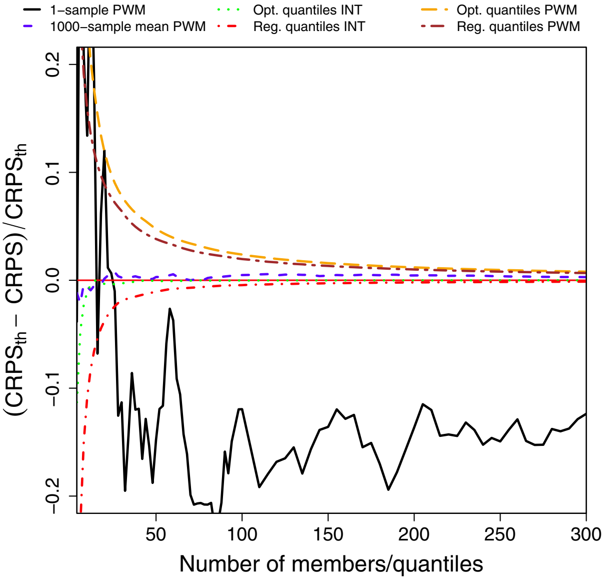


Fig. 4 Evolution with ensemble size of relative error of several estimations of CRPS, for different ensembles and different estimators of CRPS. All computation are done with the same observation for all forecasts. The ensembles and the observation come from a standard normal distribution

from a standard normal distribution ($G = F$ for the sake of simplicity). As shown in Fig. 4, the CRPSs estimated with quantile ensembles clearly outperform the $\widehat{\text{crps}}_{\text{PWM}}$ estimation with one random ensemble whatever the number of members. Averaging the $\widehat{\text{crps}}_{\text{PWM}}$ estimations of 1000 random ensembles gives a similar estimation accuracy to the one of the best estimation with quantile ensembles, namely $\widehat{\text{crps}}_{\text{INT}}$ with optimal quantiles. This configuration is not feasible in most applications, since it requires 1000 forecast/observation pairs with the same observation. Anyway, computing one set of quantiles may be much simpler and quicker than creating 1000 random ensembles. Among the estimation with ensembles of quantiles, the combination of $\widehat{\text{crps}}_{\text{INT}}$ and optimal quantiles exhibits a dramatic improvement in accuracy over the other combination, even for ensembles with less than 10 quantiles. Whatever the distribution F is used, $\widehat{\text{crps}}_{\text{INT}}$ computed with the optimal quantiles gives a much more accurate estimation, for all ensemble sizes, than the other combinations of estimator and type of ensemble of quantiles (not shown).

In order to assess the robustness of the remarks in the last paragraph in regards to the observation, data are simulated with Protocol 2 for several ensemble sizes M, with $N = 1000$ ensemble forecast/observation pairs for each ensemble size. Note that, at M fixed, the ensemble of quantiles is the same for all the forecast/observation pairs. From the point-wise intervals of the relative estimation errors represented in Fig. 5, it appears that computing $\widehat{\text{crps}}_{\text{INT}}$ with the optimal quantiles gives the most accurate estimation of $\text{crps}(F, y)$, whatever number of quantiles is used. With only a few tens of quantiles, this estimation achieves a much higher precision than the others with several hundreds of quantiles. Figure 5 also shows that, for finite ensembles of quantiles, the PWM

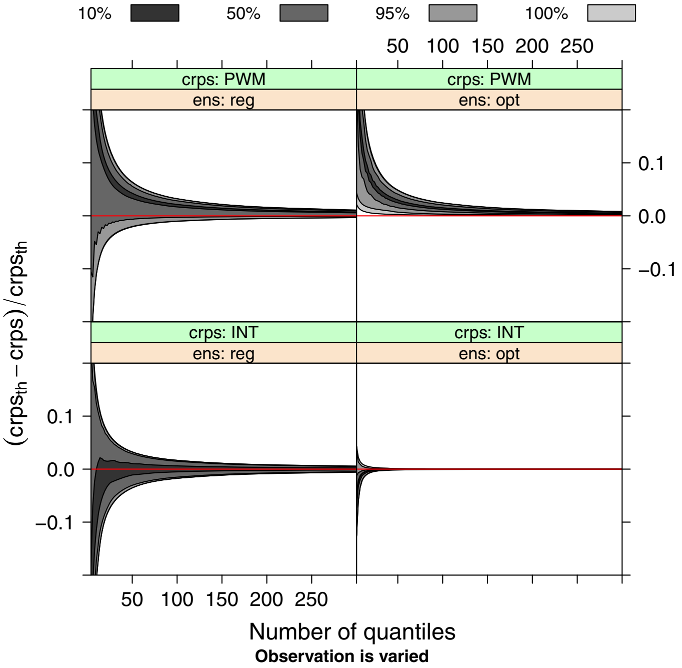


Fig. 5 10, 50, 95 and 100% point-wise intervals of relative error for several combination of quantile ensembles and CRPS estimator. Intervals are computed by drawing 1000 observations from a standard normal distribution. Ensembles are regular (left column) or optimal (right column) quantiles of a standard normal distribution. The CRPS is estimated with the PWM (top) or integral (bottom) estimator

estimator is biased, being too low (positive relative errors). Indeed, according to Eq. (2), since $\widehat{\text{crps}}_{\text{INT}}$ is an unbiased estimator of the average CRPS of an ensemble of quantiles as shown here, and since $\widehat{\lambda}_2$ is positive, $\widehat{\text{crps}}_{\text{PWM}}$ must be biased towards low values.

These conclusions hold for all the tried distributions and the set of parameters values for each distribution (not shown). As for the poor performance of $\widehat{\text{crps}}_{\text{PWM}}$ with an ensemble of quantiles, let us recall that $\widehat{\text{crps}}_{\text{PWM}}$ is a sum of terms that are unbiased estimators of their population counterpart when computed with a *random* sample, which is not the case of an ensemble of quantiles. The computation of $\widehat{\text{crps}}_{\text{INT}}$ uses the approximation of the forecast distribution as a step-wise CDF, with a fixed stair-step height $\frac{1}{M}$. The difference in estimation accuracy with the type of quantiles comes from the position of the stair steps. With regular quantiles, the step-wise CDF is always located below the forecast CDF. With optimal quantiles, the associated quantiles are shifted leftward, making the stair steps sometimes above F and sometimes below. This better approximates the forecast CDF F than with regular quantiles, thus improves the estimation of the CRPS.

3.2.3 Influence of Ties in an Ensemble of Quantiles

An ensemble of quantiles may be produced by statistical methods called quantile regression (White 1992; Koenker 2005; Meinshausen 2006; Takeuchi et al. 2006).

Some of these quantile regression methods can produce only a subset $\tau^{av}_{j=1,\ldots,N_\tau} \in [0;1]$ of N_τ orders. The quantiles associated to these available orders are called "available quantiles" hereafter and correspond to the abscissa of the black dots in Fig. 6. If one requires a quantile with an order τ outside of the subset of available orders, the quantile regression will not return the associated quantile of the forecast CDF (abscissa of the blue circles in Fig. 6), but the available quantile corresponding to the highest available order lower than τ (abscissa of the red triangles of Fig. 6). The set of different values returned by the quantile regression method when certain orders are requested is called the "unique quantiles" hereafter. It is a subset of the available quantiles. The quantile regression methods with this feature will introduce many ties in the produced ensembles of quantiles, as shown in Fig. 7 on real data. For the Canadian ensemble forecasts, although 1002 regular quantiles are required from a quantile regression method at one grid point and one lead time, the number of unique quantiles returned by the quantile regression function varies from a few tens of values to a few hundreds. On average, only about one hundred unique quantiles are produced in this example. Some implementations of quantile regression methods, such as the function `rq` in R package `quantreg`, have an option to produce the available orders τ^{av}_j and their associated quantiles. Other packages, such as `quantregForest`, have not yet implemented this possibility, and will return only forecast quantiles with (potentially many) ties.

In order to assess the impact of ties on the accuracy of the CRPS estimators for an ensemble of quantiles, ensembles of quantiles with ties are simulated with Protocol 3, with $N = 1000$ forecast/observation pairs. The left side of Fig. 8 shows that with only $N_\tau = 30$ available orders, the four estimates become inaccurate. The distribution of the estimated CRPS becomes clearly biased whatever ensemble size is considered. This bias is pessimistic (negative estimation errors) for most ensemble sizes, but may be optimistic (positive estimation errors).

A way to address this issue of equal quantiles is to remove the ties by interpolation. The first considered case is when the implementation of the quantile regression method do not propose to know the available quantiles. Protocol 3 is modified as follows at lines 3 and 4: after computing the quantiles with ties, linear interpolation is done between unique values to recover the number of required regular or optimal quantiles, as explained in Fig. 6. As shown on the right side of Fig. 8, this interpolation results in a better estimation accuracy, even though the curves are less smooth than when all orders are available (compare with Fig. 5). The best CRPS estimation is now obtained with $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ and regular quantiles, with at least $M = 30$ regular quantiles to get a sufficient accuracy. This behavior barely depends on the chosen distribution and parameter value, but requiring 100 regular quantiles seems to be the minimal number to get satisfactory accuracy, whatever the forecast distribution F is used (not shown). If the available quantiles and orders can be produced by the implementation of the quantile regression method, similar linear interpolation can be done relatively to the associated points, that is, the black dots in Fig. 6. Figure 9 shows that this linear interpolation nearly fully reproduces the good accuracy obtained when all orders are available. The best estimation strategy is again to use $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ with optimal quantiles, albeit with a slightly worst accuracy than the one reached without ties.

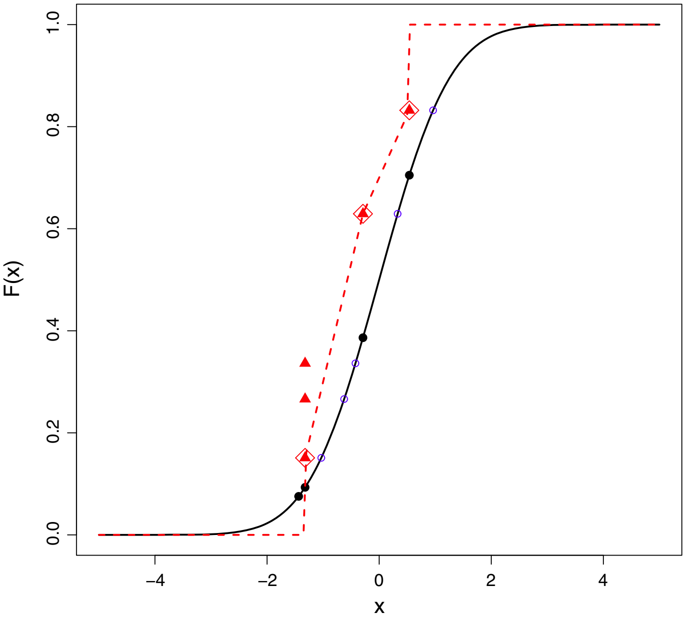


Fig. 6 Graphical illustration of the production of ties by quantile regression methods. The black continuous line is the forecast CDF. The abscissa (resp. ordinates) of the $N_\tau = 4$ black dots are the available quantiles (resp. orders), that can produce the quantile regression method. The empty blue dots are the five requested points. The red triangles are the five points actually obtained, due to the limited number of available quantiles and orders. Within each group of obtained points whose abscissa is the same, only the point with the lowest order is kept (three red diamonds) for removing the ties by interpolation. The interpolation function (dashed red line) is a linear interpolation between the red diamonds, and a constant order of 0 or 1 outside (left and right, respectively)

The influence of the number of available orders N_τ and the kind of post-processing on $\widehat{\text{crps}}_{\text{INT}}$ is crucial as shown in Fig. 10. If the number of available quantiles is too low, no matter the post-processing of the quantile ensemble, the estimated CRPS will not converge to the true value due to insufficient information about F. The number of available quantiles necessary to achieve a good accuracy depends on the complexity of the forecast distribution: a gaussian mixture with many different modes requires more available quantiles to be accurately described (not shown here).

Based on these simulations, several recommendations can be drawn to estimate the instantaneous CRPS of an ensemble of quantiles. First, if the quantile regression cannot yield enough available quantiles (less than about $N_\tau = 30$), the instantaneous CRPS should not be used whatsoever. Even the average CRPS should be used with care due to a (possibly large) estimation bias. However, if the number of available unique quantiles is sufficient (more than 30), the estimation of the instantaneous CRPS can be much improved by interpolating the quantiles and using of $\widehat{\text{crps}}_{\text{INT}}$. The best interpolation depends on the available information: if the whole set of available quantiles in the quantile regression method is not accessible, linear interpolation between the unique quantiles and their associ-

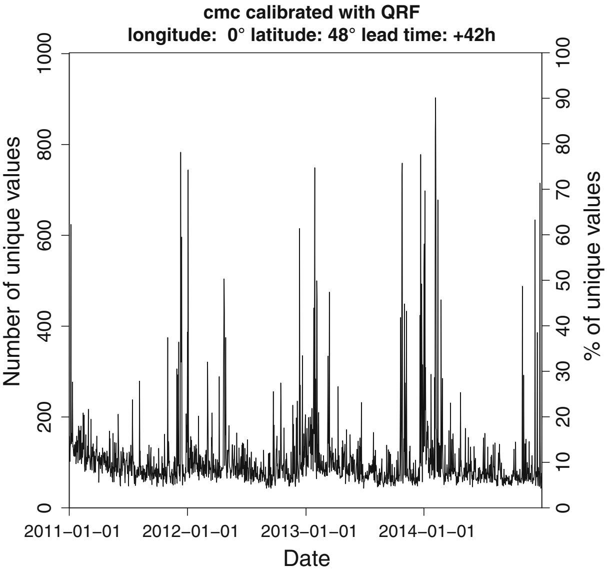


Fig. 7 Number and percentage of unique quantiles among 1002 regular quantiles requested from a quantile regression method applied to the Canadian ensemble model

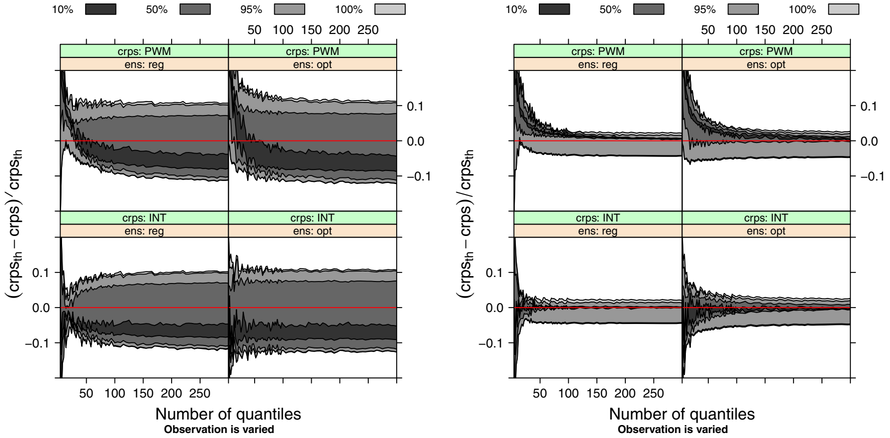


Fig. 8 Same as in Fig. 5 but with ties in the ensembles and only $N_\tau = 30$ available orders (left), and after removing ties by linear interpolation of the unique quantiles in the forecasts (right)

ated order toward regular quantiles is preferred. However, if the available quantiles and orders can be known, linear interpolation of those quantiles and orders toward optimal quantiles is the best approach. Table 2 sums up the recommendations to estimate the instantaneous CRPS for a random ensemble or an ensemble of quantiles.

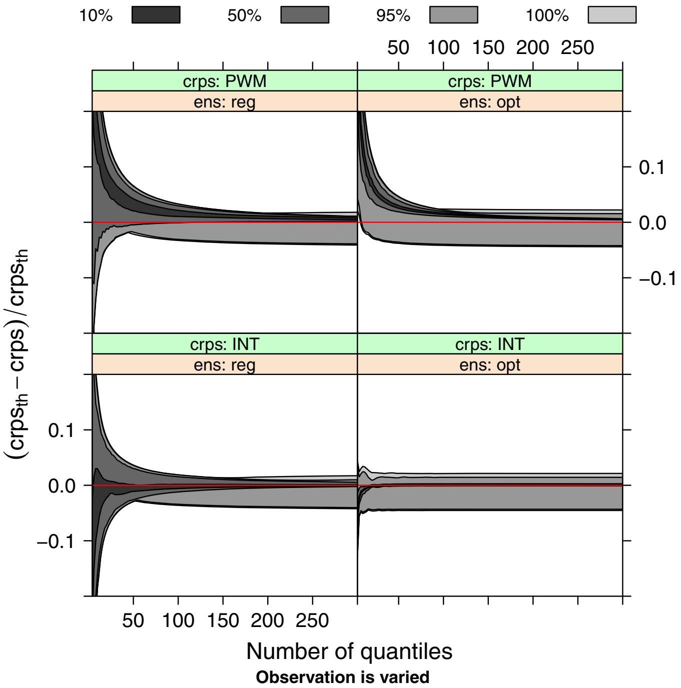


Fig. 9 Same as in Fig. 5, but with ties in the ensembles, only $N_\tau = 30$ available orders, and linear interpolation of the N_τ available quantiles

Protocol 3: ESTIMATION OF THE CRPS WITH SIMULATED ENSEMBLES OF QUANTILES, WITH TIES

```
Input: $M$: number of quantiles.
       $F$: forecast CDF.
       $G$: CDF of the observation.
       $N$: number of ensemble forecast/observation pairs.
       $N_\tau$: number of available quantiles.
Output: $N$ values of instantaneous CRPS for each estimator and kind of quantile ensemble.
Draw uniformly in [0; 1] the $N_\tau$ available orders $\tau_j^{av}$.
Compute the available quantiles of $F$: $F^{-1}(\tau_j^{av}) \forall j \in \{1, \dots, N_\tau\}$.
Compute the ensemble of $M$ regular quantiles of $F$. Make each regular quantile $x_i$ equal to the available quantile with order $\tau_j^{av}$ immediately inferior to $\tau_i$.
Compute the ensemble of $M$ optimal quantiles of $F$. Make each optimal quantile $x_i$ equal to the available quantile with order $\tau_j^{av}$ immediately inferior to $\tau_i$.
for $n \leftarrow 1$ to $N$ do
    Draw $y$ from $G$.
    Compute and store the theoretical CRPS $\mathrm{crps}_{\mathrm{th}}(F, y)$ with this observation.
    Compute and store $\widehat{\mathrm{crps}}_{\mathrm{INT}}(M, y)$ and $\widehat{\mathrm{crps}}_{\mathrm{PWM}}(M, y)$ with this observation for the ensemble of rounded regular quantiles.
    Compute and store $\widehat{\mathrm{crps}}_{\mathrm{INT}}(M, y)$ and $\widehat{\mathrm{crps}}_{\mathrm{PWM}}(M, y)$ with this observation for the ensemble of rounded optimal quantiles.
```

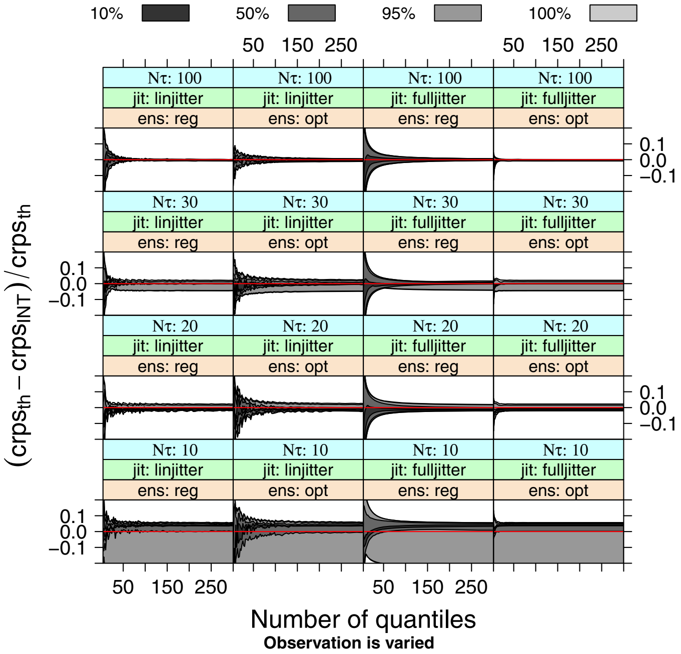


Fig. 10 Influence of post-processing. The ensemble quantiles are post-processed by linear interpolation between unique quantiles (*linjitter*) or between the N_τ available quantiles (*fulljitter*). Each panel represents the same intervals as in Fig. 5 for $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ computed from post-processed quantile ensembles with a varying number N_τ of available quantiles

Table 2 Summary of recommendations to estimate the CRPS

Type of ensemble	Condition	Recommendation
Random	The purpose is to assess the performance of an infinite ensemble	Use average $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$
	The purpose is to assess the performance of the actual ensemble	Use average $\widehat{\mathrm{crps}}_{\mathrm{INT}}$
Quantiles	All orders available	Use average $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ with optimal quantiles
	$N_\tau \lesssim 30$	Use average $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ with care
	$N_\tau \gtrsim 30$ and available quantiles unknown	Use average $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ with linearly interpolated regular quantiles between unique quantiles
	$N_\tau \gtrsim 30$ and available quantiles known	Use average $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ with linearly interpolated optimal quantiles between available quantiles

4 Real Data Examples

With two real data sets, issues resulting from the uncertainty in the estimation of the instantaneous CRPS are illustrated. The practical benefits of the recommendations listed in Table 2 are highlighted.

4.1 Raw and Calibrated Ensemble Forecast Data Sets

The first forecast data set consists in four NWP ensembles from the TIGGE project (Bougeault et al. 2010). Ten-meter high wind speed forecasts have been extracted from four operational ensemble models issued by meteorological forecast services: the US National Centers for Environmental Prediction (NCEP), the Canadian Meteorological Center (CMC), the European Center for medium-range weather forecasts (ECMWF) and Météo-France (MF). Those ensembles have respectively 21, 21, 51 and 35 members. The study domain is France with a grid size of 0.5° (about 50 km), for a total of 267 grid points. Available forecast lead-times are every 6 h. The period goes from 2011 to 2014.

The second forecast data set is composed of two versions of each ensemble calibrated with statistical post-processing methods. In order to improve the forecast performance, each ensemble has been post-processed thanks to two statistical methods: nonhomogeneous regression [NR, Gneiting et al. (2005)] and quantile regression forests [QRF, Meinshausen (2006)]. In NR, the forecast probability distribution F is supposed to be some known distribution: here the square root of forecast wind speed follows a truncated normal distribution whose mean and variance depend on the ensemble forecast. This is similar to the work of Hemri et al. (2014), who also gives the closed form expression of the instantaneous CRPS for this case. QRF is nonparametric and yields a set of quantiles x_i with chosen orders τ_i. This study uses a simplified version of the model proposed in Taillardat et al. (2016). Since QRF is nonparametric, the CRPS has to be estimated with limited information. Furthermore, QRF cannot yield every order and may lead to many ties among predicted quantiles, as seen in Fig. 7. To the best of our knowledge, no implementation of QRF in R allows knowing the available quantiles. Post-processing was done separately for each of the 267 grid points, each ensemble and each lead time. The regression was trained with cross-validation: 3 years were used as training data, the fourth one being used as test data. The four possible combinations of three training years and one test year were tested. The raw ensembles can be seen as random ensembles whereas the ensembles calibrated with QRF are ensembles of quantiles as defined above. The observation comes from a wind speed analysis made at Météo-France, presented in Zamo et al. (2016).

4.2 Issues Estimating the CRPS of Real Data

In Figs. 11 and 12, the CRPS is estimated with the first M members of the raw CMC ensemble at one grid point and for lead time +42 h. First, as shown in

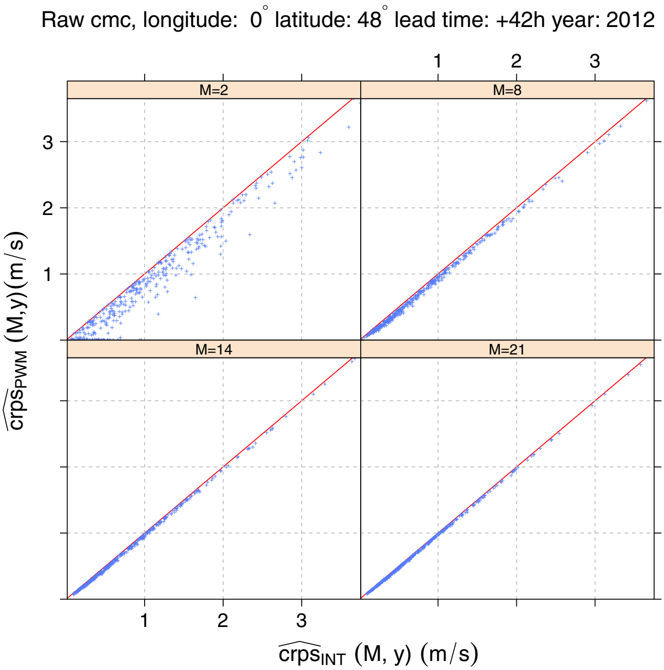


Fig. 11 Scatter plots of instantaneous CRPS computed for raw ensemble forecasts, with the two CRPS estimators. The forecasts are for one grid point of the Canadian ensemble forecast model. The number of members goes from 2 to 21 (the actual size of the ensemble). Each point corresponds to one forecast (one date and valid time)

Fig. 11, for very small ensemble sizes, differences between $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ and $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$ may be huge. With an increased ensemble size, both estimators get very similar values. Even for the largest number of members, $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ is systematically higher than $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$, in agreement with Eq. (2). These differences result in important differences on the averaged CRPS, as shown in Fig. 12, representing the evolution with M of the yearly averaged $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ and $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$. Whereas the yearly averaged $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$ is nearly independent of M, the average $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ requires a minimum ensemble size to yield a stable value. But even then, the two estimators do not yield the same average CRPS: for the year 2011, on average $\widehat{\mathrm{crps}}_{\mathrm{INT}}(M = 21) \simeq 0.75\,\mathrm{m/s}$ whereas $\widehat{\mathrm{crps}}_{\mathrm{PWM}}(M = 21) \simeq 0.7\,\mathrm{m/s}$, a difference of 7%. These conclusions from Fig. 12 are in agreement with those from Fig. 3, that shows that the average $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$ attains the true value with much smaller ensembles than $\widehat{\mathrm{crps}}_{\mathrm{INT}}$. The left side of Fig. 3 exhibits negative estimation errors, which is in agreement with the averaged $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ being higher than the averaged $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$ in Fig. 12 and in agreement with Eq. (2).

Figure 13 uses the version of the CMC ensemble calibrated with QRF. For each of the two sets of quantiles, $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ and $\widehat{\mathrm{crps}}_{\mathrm{PWM}}$ are computed for each forecast date and averaged over each test year. The number M of requested quantiles is varied from 2 to 50 and are either of regular or optimal orders. Figure 13 shows the evolution of the

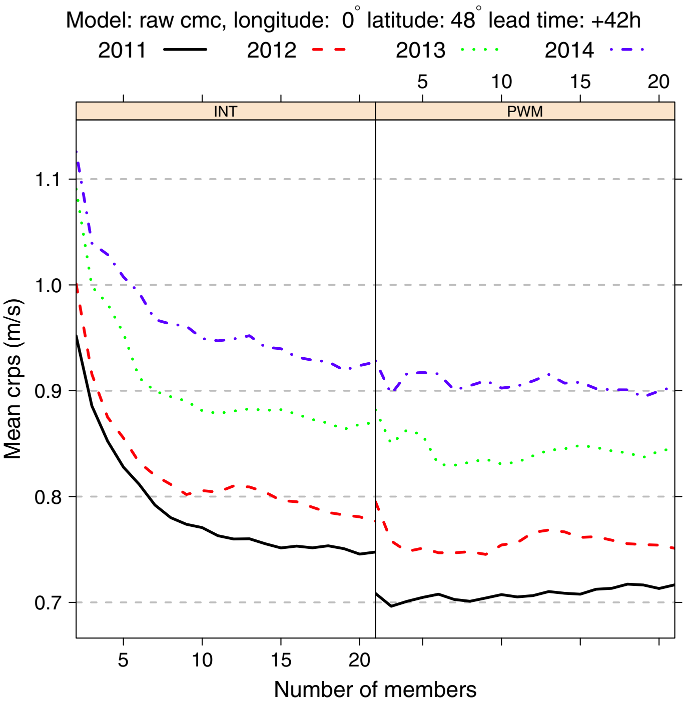


Fig. 12 Evolution of the yearly averaged CRPS with the number of members for the raw CMC ensemble. Each panel contains the average CRPS computed by averaging the instantaneous CRPS estimator, $\widehat{\text{crps}}_{\text{INT}}$ (left) or $\widehat{\text{crps}}_{\text{PWM}}$ (right). Each curve is computed by averaging the estimated instantaneous CRPS over one test year, for forecasts at one grid point and for one lead time

four estimated average CRPS with the number of quantiles, for the same grid point and lead time as above. First, the average $\widehat{\text{crps}}_{\text{INT}}$ decreases rapidly toward some value, whatever the type of quantile. Second, the yearly averaged $\widehat{\text{crps}}_{\text{PWM}}$ is not independent of the number of quantiles, as it was independent of the number of members in Fig. 12. Here, it slowly increases toward some value for a fixed type of quantile. Third, the limit values are on average $\widehat{\text{crps}}_{\text{PWM}}(50) \simeq 0.48$ m/s, $\widehat{\text{crps}}_{\text{INT}}(50) \simeq 0.47$ m/s a difference of only 2%. Last, the rate of evolution of the average CRPS with the ensemble size strongly depends on the choice of the CRPS estimator and of the type of required quantiles. For these data, removing ties in the forecast quantiles does not change the conclusions (not shown). In agreement with the recommendations from the simulated data, the fastest converging estimate is the average $\widehat{\text{crps}}_{\text{INT}}$ computed with optimal quantiles. Other ensembles, grid points and lead-times give similar results (not shown).

4.3 Issues on the Choice Between QRF and NR

For the real data set, the CRPS of QRF has been estimated with $\widehat{\text{crps}}_{\text{INT}}$ and $\widehat{\text{crps}}_{\text{PWM}}$ computed with optimal quantiles, and ties have been kept or removed by interpolation. Figure 14 shows the proportion of times QRF gets a lower CRPS than NR, out

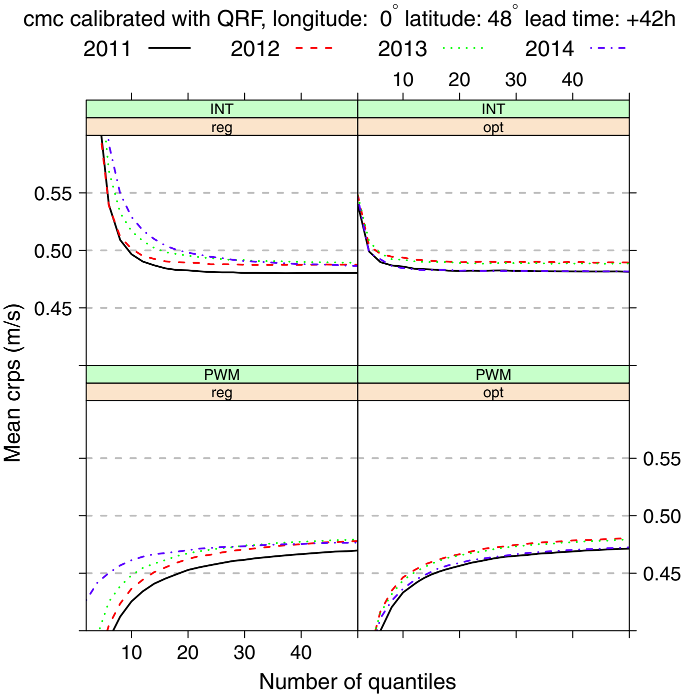


Fig. 13 Evolution with the number of members of the estimated CRPS averaged over 1 year for CMC ensemble forecast calibrated with quantile regression forests. Two sets of quantiles are requested: regular (left) and optimal (right). Ties between quantiles are not removed. Equations (eINT) (top) and (ePWM) (bottom) are used to estimate the instantaneous CRPS for each quantile sets. Each curve is then computed by averaging the estimated instantaneous CRPS over 1 year, for forecasts at one grid point and for one lead time

of the 365 forecasts during test year 2012, for one grid point and one lead time with calibrated CMC data. The proportion of times QRF outperforms NR strongly depends on the number of quantiles, but stabilizes at similar values when $\widehat{\text{crps}}_{\text{INT}}$ or $\widehat{\text{crps}}_{\text{PWM}}$ is used. In agreement with the conclusions on simulated data, the proportion stabilizes with less quantiles when $\widehat{\text{crps}}_{\text{INT}}$ is used. With too few quantiles (less than about 20), the difference of performance between QRF and NR may be deemed significant depending on the estimator. But in this specific case, after the curves have stabilized, the performance of QRF and NR are not statistically different to the level 0.01 for all the estimations. This shows that the choice of the best post-processed forecast may be misguided by poor performance estimates if the wrong estimator is used and/or not enough quantiles are required. The number of available quantiles is unknown, but has been estimated to be at least 52 for this test year. Based on the recommendations in Table 2, the best method to estimate the CRPS of QRF would be to use $\widehat{\text{crps}}_{\text{INT}}$ and at least 30 optimal quantiles, which is in agreement with the previous remarks.

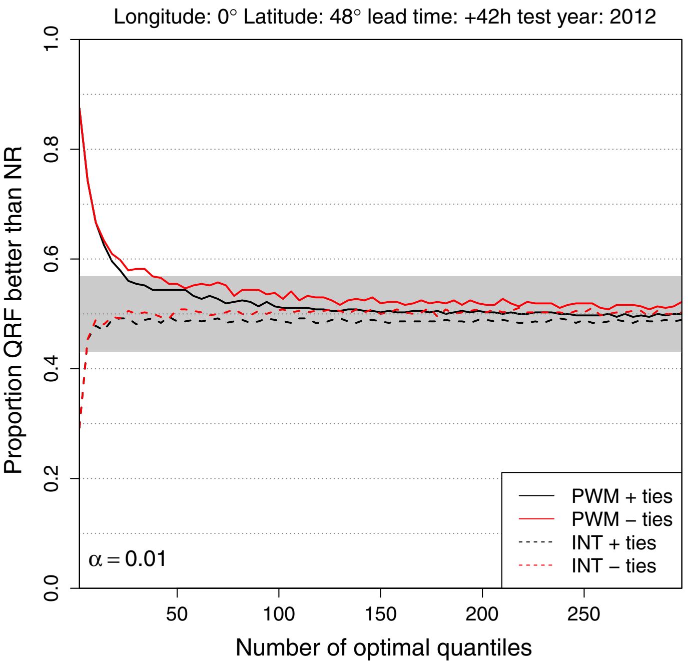


Fig. 14 Proportion of forecasts when QRF gets a lower CRPS than NR, for calibrated CMC ensemble, at one grid point, for one lead time and one test year. QRF yields M optimal quantiles and its CRPS is estimated with $\widehat{\text{crps}}_{\text{PWM}}$ (continuous line) or $\widehat{\text{crps}}_{\text{INT}}$ (dashed line), without removing ties (black curves) or after removing ties with linear interpolation between unique quantiles (red curves). NR's CRPS is computed with the closed form expression available in Hemri et al. (2014). The grey zone is the 0.01-confidence interval that the proportion is not significantly different from 0.5 (quantiles 0.995 and 0.005 of a binomial distribution with 365 tries)

5 Conclusions

A review of four estimators of the instantaneous CRPS when the forecast CDF is known through a set of values have been done. Among these four estimators proposed in the literature, only two, called the integral estimator and the probability weighted moment estimator, are not equal. Furthermore, a relationship between these two estimators has been demonstrated and generalizes to the instantaneous CRPS of any ensemble, a relationship established by Ferro et al. (2008) for the average CRPS of a random ensemble. With simulated data, the accuracy of the two estimators has been studied, when the forecast CDF is known with a limited information and the number of forecast/observation pairs is finite. The study leads to recommendations on the best CRPS estimator depending on the type of ensemble, whether random or a set of quantiles. For a random ensemble, the best estimator of the CRPS is the PWM estimator $\widehat{\text{crps}}_{\text{PWM}}$ if one wants to assess the performance of the ensemble of infinite size, whereas the integral estimator $\widehat{\text{crps}}_{\text{INT}}$ must be used to assess the performance of the ensemble with its current size. For an ensemble of quantiles, ties introduced by quantile regression methods strongly affect the estimation accuracy, and removing these ties by an interpolation step is paramount to allow a good estimation accuracy. If the number of available quantiles is too low (say, $N_\tau \leq 30$) all the studied estimators exhibit a strong bias. But if the number of available quantiles is larger, the best estimation is obtained by computing the integral estimator $\widehat{\text{crps}}_{\text{INT}}$ with linearly interpolated quantiles, between the available quantiles if they are known or between the unique quantiles otherwise.

The established relationships between the estimators proposed in the literature have been linked to previous results. These relationships also explain why an estimator is more accurate for one type of ensemble and not for the other. The PWM estimator performs better on random ensembles because it is based on estimators that are unbiased for independent samples from the true underlying distribution. On the other hand, the integral estimator gives a good estimate when computed with optimal quantiles. This is because regular weights are associated to the members in the estimator formula but, when using optimal quantiles, the associated quantiles are shifted to better approximate the underlying forecast CDF.

The important consequences on the choice of method of estimation of the CRPS has also been illustrated on real meteorological data with raw ensembles and calibrated ensembles. As an example, the comparison of several calibrated ensembles may be mislead by a poor estimate of the average CRPS of ensembles of quantiles.

Acknowledgements Both authors are grateful to Pr. Liliane Bel (AgroParisTech, France) and Pr. Tilmann Gneiting (Heidelberg Institute for Theoretical Studies, Germany) for their useful comments on this paper. Part of the work of Philippe Naveau has been supported by the ANR-DADA, LEFE-INSU-Multirisk, AMERISKA, A2C2, CHAVANA and Extremoscope projects. Part of the work was done when Philippe Naveau was visiting the IMAGE-NCAR group in Boulder, Colorado, USA.

A What is Elicited When the 1-Norm CRPS of an Ensemble is Minimized?

Let $\{x_i\}_{i=1,\ldots,M}$ be an ensemble of M values. Let $F_e(x) = \sum_{i=1}^{M} \omega_i \mathbb{1}(x \geq x_i)$ be the associated empirical CDF, with weights ω_i, such that $\omega_i \geq 0 \quad \forall i \in \{1, \ldots, M\}$ and $\sum_{i=1}^{M} \omega_i = 1$. Let y be the observation.

Following Müller et al. (2005), the 1-norm CRPS of this ensemble relative to this observation is defined as

$$\text{crps}_1(F_e, y) = \int_{\mathbb{R}} |F_e(x) - \mathbb{1}(x \geq y)| \text{d}x.$$

This can be rewritten in a more interpretable form.

$$
\begin{aligned}
\mathrm{crps}_1(F_e, y) &= \int_{-\infty}^{y} \left| \sum_{i=1}^{M} \omega_i \mathbb{1}(x \geq x_i) \right| \mathrm{d}x \\
&\quad + \int_{y}^{+\infty} \left| \sum_{i=1}^{M} \omega_i \left(\mathbb{1}(x \geq x_i) - 1 \right) \right| \mathrm{d}x \\
&= \int_{-\infty}^{y} \sum_{i=1}^{M} \omega_i \mathbb{1}(x \geq x_i) \mathrm{d}x \\
&\quad + \int_{y}^{+\infty} \sum_{i=1}^{M} \omega_i \left(1 - \mathbb{1}(x \geq x_i) \right) \mathrm{d}x \\
&= \sum_{i=1}^{M} \omega_i \Bigg[\int_{-\infty}^{y} \mathbb{1}(x \geq x_i) \mathrm{d}x \\
&\quad + \int_{y}^{+\infty} 1 - \mathbb{1}(x \geq x_i) \mathrm{d}x \Bigg].
\end{aligned}
$$

If $y \geq x_i$

$$\int_{-\infty}^{y} \mathbb{1}(x \geq x_i) \mathrm{d}x = \int_{-\infty}^{x_i} 0 \mathrm{d}x + \int_{x_i}^{y} 1 \mathrm{d}x = y - x_i,$$

and

$$\int_{y}^{+\infty} 1 - \mathbb{1}(x \geq x_i) \mathrm{d}x = \int_{y}^{+\infty} 0 \mathrm{d}x = 0.$$

If $y \leq x_i$

$$\int_{-\infty}^{y} \mathbb{1}(x \geq x_i) \mathrm{d}x = \int_{-\infty}^{y} 0 \mathrm{d}x = 0,$$

and

$$\int_{y}^{\infty} 1 - \mathbb{1}(x \geq x_i) \mathrm{d}x = \int_{y}^{x_i} 1 \mathrm{d}x + \int_{x_i}^{+\infty} 0 \mathrm{d}x = x_i - y.$$

Therefore, $\forall y$ and $\forall i$

$$\int_{-\infty}^{y} \mathbb{1}(x \geq x_i) \mathrm{d}x + \int_{y}^{+\infty} 1 - \mathbb{1}(x \geq x_i) \mathrm{d}x = |y - x_i|.$$

Finally

$$\mathrm{crps}_1(F_e, y) = \sum_{i=1}^{M} \omega_i |y - x_i|.$$

The 1-norm CRPS is just the weighted mean of the absolute error of each member. The average 1-norm CRPS is thus minimized if all the members are equal to the median of the observation CDF (Gneiting 2011).

B Relationships Between the Estimators of the CRPS

Without loss of generality, the forecast is an ensemble of M values $x_{i=1,\ldots,M}$ sorted in increasing order.

B.1 Equality of $\widehat{\mathbf{crps}}_{\mathbf{Fair}}$ and $\widehat{\mathbf{crps}}_{\mathbf{PWM}}$

Following the definition of L-moments and their relationship with PWMs (Wang 1996; Hosking 1990), one can rewrite

$$\begin{aligned}\hat{\lambda}_2 &= \frac{1}{2M(M-1)} \sum_{i,j=1}^{M} |x_i - x_j| \\ &= 2\hat{\beta}_1 - \hat{\beta}_0 \\ &= \frac{1}{M(M-1)} \sum_{i,j=1}^{M} (2i - M - 1) x_i, \end{aligned} \tag{3}$$

where $\hat{\lambda}_2$, $\hat{\beta}_1$ and $\hat{\beta}_0$ are estimators of the second linear moment, the PWM of order 1 and the PWM of order 0 (i.e. the average), respectively. These estimators are unbiased if the ensemble is a random sample.

Introducing these notations in Eq. (eFAIR) leads to

$$\begin{aligned}\widehat{\mathrm{crps}}_{\mathrm{Fair}}(M, y) &= \frac{1}{M} \sum_{i=1}^{M} |x_i - y| + \hat{\beta}_0 - 2\hat{\beta}_1 \\ &= \widehat{\mathrm{crps}}_{\mathrm{PWM}}(M, y).\end{aligned}$$

B.2 Equality of $\widehat{\mathbf{crps}}_{\mathbf{NRG}}$ and $\widehat{\mathbf{crps}}_{\mathbf{INT}}$

As Gneiting and Raftery (2007) showed, the representations in Eqs. (INT) and (NRG) are equivalent for forecast CDFs with a finite first moment. Since empirical distributions have a finite first moment, and since Eqs. (INT) and (NRG) reduce to Eqs. (eINT) and (eNRG) respectively, equality of $\widehat{\mathrm{crps}}_{\mathrm{INT}}$ and $\widehat{\mathrm{crps}}_{\mathrm{NRG}}$ follows immediately.

Thanks to Pr. Tilmann Gneiting for this proof, much more straightforward than the one initially proposed.

B.3 Relationship Between $\widehat{\mathbf{crps}}_{\mathbf{PWM}}$ and $\widehat{\mathbf{crps}}_{\mathbf{NRG}}$

Using Eq. (3) leads to

$$\begin{aligned}\widehat{\text{crps}}_{\text{NRG}}(M, y) &= \frac{1}{M}\sum_{i=1}^{M}|x_i - y| - \frac{2M(M-1)}{2M^2}\left(2\hat{\beta}_1 - \hat{\beta}_0\right)\\ &= \frac{1}{M}\sum_{i=1}^{M}|x_i - y| + \hat{\beta}_0 - 2\hat{\beta}_1 + \frac{\hat{\lambda}_2}{M}\\ &= \widehat{\text{crps}}_{\text{PWM}}(M, y) + \frac{\hat{\lambda}_2}{M}.\end{aligned}$$

References

Baran S, Lerch S (2015) Log-normal distribution based ensemble model output statistics models for probabilistic wind-speed forecasting. Q J R Meteorol Soc 141:2289–2299

Bougeault P, Toth Z, Bishop C, Brown B, Burridge D, Chen DH, Ebert B, Fuentes M, Hamill TM, Mylne K et al (2010) The THORPEX interactive grand global ensemble. Bull Am Meteorol Soc 91(8):1059

Brier G (1950) Verification of forecasts expressed in terms of probability. Mon Weather Rev 78(1):1–3

Bröcker J (2012) Evaluating raw ensembles with the continuous ranked probability score. Q J R Meteorol Soc 138(667):1611–1617

Candille G (2003) Validation des systèmes de prévisions météorologiques probabilistes. PhD thesis, Paris 6

Emery X, Lantuéjoul C (2006) Tbsim: a computer program for conditional simulation of three-dimensional gaussian random fields via the turning bands method. Comput Geosci 32(10):1615–1628

Ferro CAT (2014) Fair scores for ensemble forecasts. Q J R Meteorol Soc 140(683):1917–1923

Ferro CA, Richardson DS, Weigel AP (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores. Meteorol Appl 15(1):19–24

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

Furrer R, Naveau P (2007) Probability weighted moments properties for small samples. Stat Probab Lett 77(2):190–195

Gneiting T (2011) Quantiles as optimal point forecasts. Int J Forecast 27(2):197–207

Gneiting T, Raftery A (2007) Strictly proper scoring rules, prediction, and estimation. J Am Stat Assoc 102(477):359–378

Gneiting T, Raftery A, Westveld A III, Goldman T (2005) Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon Weather Rev 133(5):1098–1118

Greenwood JA, Landwehr JM, Matalas NC, Wallis JR (1979) Probability weighted moments: definition and relation to parameters of several distributions expressable in inverse form. Water Resour Res 15(5):1049–1054

Grimit E, Gneiting T, Berrocal V, Johnson N (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. Q J R Meteorol Soc 132(621C):2925–2942

Hemri S, Scheuerer M, Pappenberger F, Bogner K, Haiden T (2014) Trends in the predictive performance of raw ensemble weather forecasts. Geophys Res Lett 41(24):9197–9205

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

Hosking J (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. J Roy Stat Soc: Ser B (Methodol) 52(1):105–124

Jolliffe I, Stephenson D (2011) Forecast verification: a Practioner's guide in atmospheric science, 2nd edn. Wiley, London

Koenker R (2005) Quantile regression, vol 38. Cambridge University Press, Cambridge

Matheson JE, Winkler RL (1976) Scoring rules for continuous probability distributions. Manag Sci 22(10):1087–1096

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

Möller D, Scheuerer M (2013) Postprocessing of ensemble forecasts for wind speed over Germany. PhD thesis, Diploma thesis, Faculty of Mathematics and Computer Science, Heidelberg University. http://www.rzuser.uni-heidelberg.de/~kd4/files/Moeller2013.pdf

Müller W, Appenzeller C, Doblas-Reyes F, Liniger M (2005) A debiased ranked probability skill score to evaluate probabilistic ensemble forecasts with small ensemble sizes. J Clim 18(10):1513–1523

Pirot G, Straubhaar J, Renard P (2014) Simulation of braided river elevation model time series with multiple-point statistics. Geomorphology 214:148–156

Rasmussen PF (2001) Generalized probability weighted moments: application to the generalized pareto distribution. Water Resour Res 37(6):1745–1751

Székely GJ, Rizzo ML (2013) Energy statistics: a class of statistics based on distances. J Stat Plan Inference 143(8):1249–1272

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

Takeuchi I, Le Q, Sears T, Smola A (2006) Nonparametric quantile estimation. J Mach Learn Res 7:1231–1264

Thorarinsdottir TL, Gneiting T (2010) Probabilistic forecasts of wind speed: ensemble model output statistics by using heteroscedastic censored regression. J R Stat Soc: Ser A (Stat Soc) 173(2):371–388

Wang Q (1996) Direct sample estimators of L moments. Water Resour Res 32(12):3617–3619

Weijs SV, Van Nooijen R, Van De Giesen N (2010) Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. Mon Weather Rev 138(9):3387–3399

White H (1992) Nonparametric estimation of conditional quantiles using neural networks. In: Page C, LePage R (eds) Computing science and statistics. Springer, New York, NY, pp 190–199

Winkler R, Muñoz J, Cervera J, Bernardo J, Blattenberger G, Kadane J, Lindley D, Murphy A, Oliver R, Ríos-Insua D (1996) Scoring rules and the evaluation of probabilities. Test 5(1):1–60

Yin G, Mariethoz G, McCabe MF (2016) Gap-filling of landsat 7 imagery using the direct sampling method. Remote Sens 9(1):12

Zamo M, Bel L, Mestre O, Stein J (2016) Improved gridded windspeed forecasts by statistical post-processing of numerical models with block regression. Weather Forecast 31(6):1929–1945